Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Many-Server Heavy-Traffic Limits for Queueing Systems with Perfectly Correlated Service and Patience Times

Lun Yu, Ohad Perry

To cite this article:
Lun Yu, Ohad Perry (2023) Many-Server Heavy-Traffic Limits for Queueing Systems with Perfectly Correlated Service and Patience Times. Mathematics of Operations Research 48(2):1119-1157. https://doi.org/10.1287/moor.2022.1300

https://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 48, No. 2, May 2023, pp. 1119–1157
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Many-Server Heavy-Traffic Limits for Queueing Systems with Perfectly Correlated Service and Patience Times

Lun Yu,[a,*] **Ohad Perry**[b]

[a] Department of Industrial Engineering, Tsinghua University, Beijing 100084, P. R. China; [b] Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208
*Corresponding author
Contact: yl109376@tsinghua.edu.cn, http://orcid.org/0000-0003-0044-2514 (LY); ohad.perry@northwestern.edu (OP)

Received: December 29, 2020
Revised: December 29, 2021
Accepted: July 1, 2022
Published Online in Articles in Advance: September 8, 2022

MSC2020 Subject Classifications: 60K25, 90B22

https://doi.org/10.1287/moor.2022.1300

Abstract. We characterize heavy-traffic process and steady-state limits for systems staffed according to the square-root safety rule, when the service requirements of the customers are perfectly correlated with their individual patience for waiting in queue. Under the usual many-server diffusion scaling, we show that the system is asymptotically equivalent to a system with no abandonment. In particular, the limit is the Halfin-Whitt diffusion for the $M/M/n$ queue when the traffic intensity approaches its critical value 1 from below, and is otherwise a transient diffusion, despite the fact that the prelimit is positive recurrent. To obtain a refined measure of the congestion due to the correlation, we characterize a lower-order fluid (LOF) limit for the case in which the diffusion limit is transient, demonstrating that the queue in this case scales like $n^{3/4}$. Under both the diffusion and LOF scalings, we show that the stationary distributions converge weakly to the time-limiting behavior of the corresponding process limit.

Funding: This work was supported by the National Natural Science Foundation of China [Grant 72188101] and the Division of Civil, Mechanical and Manufacturing Innovation [Grants 1763100 and 2006350].

Keywords: many-server queues • square-root staffing • correlated service and patience • fluid limits • diffusion limits

1. Introduction

Service systems often experience abandonment due to customer impatience for waiting in queue. The significant impacts that abandonment has on the queueing dynamics are clear from the fact that stability— the most fundamental performance measure of a queueing system—is guaranteed to hold under weak regularity conditions on the system's primitives, regardless of the value of the traffic intensity; see Kang and Ramanan [10, section 4]. To model customer abandonment, it is typically assumed that the patience of the customers are independent and identically distributed (i.i.d.) random variables, that are also independent of all other random variables and processes in the model. However, it stands to reason that, in practice, the patience of customers depends on their individual service requirement, as was indeed empirically demonstrated to be the case in contact centers (Reich [16]) and restaurants (De Vries et al. [4]).

A heuristic fluid model developed in Wu et al. [23] (see also Wu et al. [24]) suggests that a positive dependence between the service and patience times of customers has large impacts on steady-state performance measures, such as the expected steady-state queue length and waiting times, when the system is overloaded (in the sense that the arrival rate exceeds the maximum service capacity). However, in overloaded systems, practically all the customers are delayed in queue, and their waiting times are, asymptotically (under fluid scaling), of the same order as the service time. It is therefore not immediately clear whether the insights in Wu et al. [23] extend to systems that are not overloaded, so that a significant proportion of the customers are not delayed at all, and the waiting times of those customers that are delayed are asymptotically negligible.

In this paper, we carry out asymptotic analysis in this latter setting, by considering systems that are staffed according to the square-root rule, whose aim is to put the systems in the Halfin-Whitt limiting regime. This regime, which was first characterized in the seminal paper by Halfin and Whitt [7] for the $M/M/n$ (Erlang-C) queue, and was later extended in Garnett et al. [6] to the $M/M/n+M$ (Erlang-A) model, which includes exponentially distributed customer patience, is also known as the *quality-and-efficiency* (QED) regime, as it achieves both efficient utilization, while simultaneously providing high quality of service. In particular, under standard independence assumptions of the system's primitives, the square-root staffing rule guarantees that almost all the service capacity is utilized at all times, as is the case in the conventional heavy-traffic regime, yet the probability

that arrivals are delayed in queue is smaller than 1 in the limit, and waiting times of delayed customers are asymptotically negligible; see, for example, van Leeuwaarden et al. [19] and Whitt [20]. It is significant that the Erlang-A model operates in the QED regime even if the traffic intensity approaches 1 from above, namely, if the service capacity in the system is smaller than the demand for service by an $O(\sqrt{n})$ term. We elaborate in Section 3.3.

1.1. The Impact of the Correlation

When the service and patience times are positively correlated, one expects the system to be more congested than when the two times are independent, because delayed customers that do not abandon tend to spend more time in service than a "generic" customer. On the other hand, the waiting times and the proportion of abandonment in the QED regime are asymptotically negligible, and so the extent to which correlation impacts the queueing dynamics is not a priori clear.

Our results show that, in the perfect-correlation case, abandonment has an asymptotically diminishing impact on the queues under diffusion scaling, in that the system behaves much like a system that has no abandonment at all. Thus, unlike in the typical independent models (which assume that all the primitive processes are mutually independent), the diffusion limit can be transient, despite the fact that the prelimit is always stable. The exact extent to which the correlation impacts congestion follows from limits for the queue process and for its steady-state distribution that are achieved under an $n^{3/4}$ spatial scaling.

Specifically, we prove the following functional weak limit theorems. The diffusion limit, which is achieved under the usual many-server diffusion scaling (see Section 3.1), is the same limit that is obtained for the Erlang-C model under the square-root staffing rule. Thus, if the traffic intensity approaches 1 from below as $n \to \infty$, then the diffusion limit is the Halfin-Whitt diffusion in Halfin and Whitt [7]. On the other hand, if the traffic intensity approaches 1 from above, then the limit is a transient diffusion, having a positive drift. To obtain the exact order of congestion in the latter case, we derive a lower-order fluid (LOF) limit, and a corresponding weak limit for the stationary distributions, both obtained under a spatial scaling of $n^{3/4}$ (with the former limit being obtained under a time scaling of $n^{1/4}$). Given that the Erlang-A model operates in the QED regime under the square-root staffing rule, those latter limit theorems imply that the correlation causes an increase of order $O(n^{1/4})$ in congestion relative to the independent case.

1.2. Implications

Even though perfect correlation between the service and patience times of customers is unlikely to be encountered in practice, this case is worth studying because the limits we obtain for the queues are simple one-dimensional Markov processes that are easy to interpret, despite the non-Markovian nature of the prelimit queue. That one-dimensional characterization is achieved by decomposing the service times of served customers into two phases, exploiting the perfect-correlation assumption together with the memoryless property of the exponential distribution; see (4). More general dependence structures will necessarily require complex (e.g., measured-valued) process descriptors, which will in turn lead to more complex, infinite-dimensional limiting processes; see Puha and Ward [13] for background. On the other hand, the diminishing impact of the abandonment on the system's dynamics, and the resulting congestion, are likely to hold in much greater generality than the special case we study. (Much like the QED regime, which was initially developed for systems with exponentially distributed service times, and was only later shown to hold in greater generality (Gamarnik and Goldberg [5], Puhalskii and Reiman [14], Reed [15]).)

We further remark that a certain martingale property, that is key to deriving measure-valued limits for a non-Markovian many-server queues with abandonment, relies heavily on having the service and patience times be independent; see Kang and Ramanan [9, proposition 5.1]. In the special case we consider, we circumvent this issue by employing an intricate representation of the state descriptors, exploiting submartingale properties of certain two-parameter processes. See the state descriptors in Section 5.2 and Lemma 7.

1.3. Notation

All the random elements are defined on a complete probability space $(\Omega, \mathcal{F}, P)$; expectation with respect to P is denoted by E. We let $\mathbb{R}$ and $\mathbb{Z}$ denote the sets of real numbers and integers, respectively, with $\mathbb{R}_+ := [0, \infty)$ and $\mathbb{Z}_+ := \mathbb{Z} \cap \mathbb{R}_+$. For $k \in \mathbb{N}$, we let $\mathbb{R}^k$ denote the space of k-dimensional vectors with real components. We let D^k denote the space of right-continuous $\mathbb{R}^k$-valued functions with left limits on $\mathbb{R}_+$, endowed with the usual Skorokhod J_1 topology; see Billingsley [1]. We let $D := D^1$ and $D_0 := \{x \in D : x(0) \geq 0\}$. We use C^k (and $C := C^1$) to denote the subspace of D^k of continuous functions, and $C_0 := D_0 \cap C$. It is well known that the J_1 topology relativized to

C^k coincides with the uniform topology on C^k, which is induced by the norm

$$\|x\|_t := \sup_{0\le u\le t} \|x(u)\|,$$

where $\|x\|$ denotes the usual Euclidean norm of $x \in \mathbb{R}^k$. We use $\eta : \mathbb{R} \to \mathbb{R}$ for the identity map, that is, $\eta(t) = t$ for $t \ge 0$, so that, in particular, 0η is the zeroth function in D.

Let $\Rightarrow$ denote convergence in distribution of a sequence of random elements in a metric space (see, e.g., Whitt [21, section 11.3.2]). For a sequence of processes $\{Y^n : n \ge 1\}$ and a sequence of scalars $\{a^n : n \ge 1\}$, we write (i) $Y^n = o_P(a^n)$, if for any $t \ge 0$ we have $\|Y^n/a^n\|_t \Rightarrow 0$ in $\mathbb{R}$, as $n \to \infty$; (ii) $Y^n = O_P(a^n)$, if Y^n is stochastically bounded, that is, $\{\|Y^n/a^n\|_t : n \ge 1\}$ is a tight sequence in $\mathbb{R}$ for any $t \ge 0$; (iii) $Y^n = \Theta_p(a^n)$ if $Y^n = O_P(a^n)$ but not $o_P(a^n)$. We write $\stackrel{\mathrm{d}}{=}$ to denote equality in distribution, and $\le_{st.}$ to denote the usual stochastic order. Namely, for two random variables X and Y, we write $X \le_{st.} Y$ if $P(X > x) \le P(Y > x)$ for all $x \in \mathbb{R}$. For a random variable with values in $\mathbb{Z}_+$, and a sequence of random variables $\{X_i : i \ge 1\}$, we define $\sum_{i=1}^N X_n := 0$ on the event $\{N = 0\}$.

We let $x^+ := \max\{x, 0\}$ and $x^- := -\min\{x, 0\}$ for $x \in \mathbb{R}$. For $x, y \in \mathbb{R}$ we let $x \wedge y := \min\{x, y\}$ and $x \vee y := \max\{x, y\}$. Moreover, we let the latter min and max operators $\wedge$ and $\vee$ have higher precedence than multiplication, so that $xy \wedge z = x(y \wedge z)$, and in particular, $x + y \wedge z = x + (y \wedge z)$, for $x, y, z \in \mathbb{R}$.

1.4. Background

Consider a sequence of systems, in which the nth element has a pool of n statistically homogeneous agents serving a single class of statistically homogeneous customers. Let λ^n denote the arrival rate to system n and μ denote the service rate of a customer (the latter does not scale with the system). The square-root staffing rule stipulates that the number of agents and the arrival rate satisfy the relation

$$\lim_{n\to\infty} \sqrt{n}(1 - \rho^n) = \beta, \tag{1}$$

for some $\beta > 0$, where $\rho^n := \lambda^n/(n\mu)$ is the traffic intensity to system n. In particular, the square-root rule implies that $\lambda^n = n\mu - O(\sqrt{n})$ as $n \to \infty$.

Now, consider the special case of Poisson arrivals and exponentially distributed service times, namely, the Erlang-C queue. Let $X_C^n := \{X_C^n(t) : t \ge 0\}$ denote the number-in-system process, and let $\widehat{X}_C^n := \{\widehat{X}_C^n(t) : t \ge 0\}$ denote its diffusion-scale version,

$$\widehat{X}_C^n(t) := n^{-1/2}(X_C^n(t) - n), \quad t \ge 0.$$

Theorem 2 in Halfin and Whitt [7] states that, if (1) holds, and in addition $\widehat{X}_C^n(0) \Rightarrow X_0$ in $\mathbb{R}$, then $\widehat{X}_C^n \Rightarrow \widehat{X}_C$ uniformly on compact (time) intervals as $n \to \infty$, where $\widehat{X}_C := \{\widehat{X}_C(t) : t \ge 0\}$ is the unique strong solution (e.g., see Revuz and Yor [17]) to the stochastic differential equation (SDE)

$$d\widehat{X}_C(t) = m_C(\widehat{X}_C(t))dt + \sqrt{2\mu}dB(t),\ \widehat{X}_C(0) = X_0, \tag{2}$$

for

$$m_C(x) := \begin{cases} -\mu\beta & \text{if } x \ge 0; \\ -\mu(\beta + x) & \text{if } x < 0, \end{cases}$$

and $B := \{B(t) : t \ge 0\}$ denoting a standard Brownian motion.

If, in addition, customers are assumed to have finite patience that is exponentially distributed with mean $1/\theta$ that is independent of all other random variables in the model, namely, if the Erlang-A queue is considered, then the square-root staffing rule can be generalized by allowing β in (1) to be nonpositive. In particular, let $X_A^n(t)$ denote the number-in-system process in a system with abandonment, and let

$$\widehat{X}_A^n(t) := n^{-1/2}(X_A^n(t) - n), \quad t \ge 0.$$

Theorem 2 in Garnett et al. [6] proves that, if (1) holds with $\beta \in (-\infty, \infty)$, and in addition, $\widehat{X}_A^n(0) \Rightarrow X_0$ in $\mathbb{R}$ for some random variable X_0, then $\widehat{X}_A^n \Rightarrow \widehat{X}_A$ uniformly over compact time intervals as $n \to \infty$, where

$$d\widehat{X}_A(t) = m_A(\widehat{X}_A(t))dt + \sqrt{2\mu}dB(t),\ \widehat{X}_A(0) = X_0. \tag{3}$$

Here, B denotes a standard Brownian motion as before, and

$$m_A(x) := \begin{cases} -(\mu\beta + \theta x) & \text{if } x \geq 0; \\ -\mu(\beta + x) & \text{if } x < 0. \end{cases}$$

We observe that both the diffusion limit in (2) and the limit in (3) imply that the stochastic fluctuations of X_C^n and X_A^n about n (the number of agents) are $O_P(\sqrt{n})$, namely, are of order $\sqrt{n}$. Therefore, both the number of idle agents and the number of customers waiting in queue are $O_P(\sqrt{n})$ as well, as $n \to \infty$. Moreover, both diffusion processes achieve values in $\mathbb{R}$, implying that a nonnegligible proportion of the customers do not wait at all, whereas the waiting times of those customers who are delayed in queue are $O_P(n^{-1/2})$, and so are asymptotically negligible, as $n \to \infty$.

1.5. Organization

The rest of the paper is organized as follows: We introduce the model in Section 2. The main results—the diffusion and LOF limits, as well as the corresponding weak limits for the stationary distributions—appear in Section 3. To simplify the exposition, we first introduce the stochastic-process limit theorems under a simplifying assumption on the initial conditions; we weaken that assumption significantly in Section 3.3. We summarize the results in Section 4. The following sections are dedicated to proving the main results: In Section 5.1, we provide a characterization of the system's dynamics that is key to establishing the main results, whose proofs appear in Section 5. Proofs of supporting results are given in Section 6–Appendix A.

2. The Model

We consider a sequence of systems denoted by $M/M_{pc}/n + M_{pc}$, indexed by the number of agents n; the subscript "pc" is mnemonic for "perfect correlation." Each of the systems along the sequence consists of a single service pool with statistically homogeneous agents, and an infinite buffer in which customers wait for their service. Customers arrive to system n according to a Poisson process with rate λ^n, where $\lambda^n/n \to \lambda$ as $n \to \infty$, for some $\lambda > 0$. Customers begin service with an agent immediately upon arrival, if an idle agent is available, and otherwise waits in the queue for their turn to enter service. We assume that customers are served in accordance with the first-come-first-serve (FCFS) discipline, namely, in the order of arrival, and that each customer has finite patience for waiting in queue: customers who run out of patience before their turn to enter service abandon the queue without returning. We further assume that the service requirement and the patience time of each customer are (marginally) exponentially distributed with respective means $1/\mu$ and $1/\theta$, $\mu, \theta > 0$, and that these two exponential random variables are independent from the arrival process and from the service and patience times of all other customers. Without loss of generality, we measure time in service-time units, taking $\mu = 1$. We further assume that n, λ^n and μ are related via the limit (1) (so that $\lambda = \mu = 1$), for some $\beta \in (-\infty, \infty)$.

Unlike the standard $M/M/n + M$ queue, we assume that the service requirement of a customer is perfectly correlated with the customer's patience. In particular, let (S, T) denote a random variable in $\mathbb{R}^2$, such that T is exponentially distributed with mean $1/\theta$ and S is exponentially distributed with mean $1/\mu = 1$. The assumption that S and T are perfectly correlated indicates that $T = S/\theta$ with probability 1 (w.p.1). We assume that the service requirement and patience of each customer is a draw from the joint distribution of S and T, independently of all other customers and of the arrival process.

Due to the assumed correlation, the service-time distribution of a *served customer* is different from the service-time distribution of a generic customer. For $w \geq 0$, temporarily let $S(w)$ denote a generic service time of a customer who waited w time units in queue. Utilizing the memoryless property of the exponential distribution, we have that

$$S(w) \stackrel{\mathrm{d}}{=} S_b + \theta w, \tag{4}$$

where S_b is an exponentially distributed random variable with mean 1. (We emphasize that our analysis hinges on the decomposition of $S(w)$ in (4), which holds only when the service and patience times are perfectly correlated.) Thus, the service time of each customer can be thought of as having two independent phases: conditional on the waiting time of the customer being w, phase 1 takes θw units of time, and phase 2 is distributed like S_b. Observe that the waiting time in queue completely determines the length of phase 1, whereas the length of phase 2 does not depend on the waiting time.

For $t \geq 0$ and $n \geq 1$, let $Z^n(t)$ denote the number of customers in service at time t, and let $Z_i^n(t)$ denote the number of customers in phase I at time t, $i = 1, 2$, so that $Z^n(t) = Z_1^n(t) + Z_2^n(t)$. We denote by $Q^n(t)$ the number of

customers waiting in queue, and by $X^n(t)$ the total number of customers in the system at time t, so that $X^n(t) := Z^n(t) + Q^n(t)$.

2.1. Preliminary: Stationarity of the $M/M_{pc}/n+M_{pc}$ System

Clearly, X^n is not a Markov process, because its evolution depends on the waiting times of the customers in Z_1^n. However, it is a regenerative process—a fact we employ to prove the following theorem.

Theorem 1. *The process X^n possesses a unique steady-state distribution, which is also its limiting distribution, as $t \to \infty$.*

Proof. First, note that, due to the arrival process being Poisson, and the fact that all customers entering service immediately upon arrival have i.i.d. exponential service times, X^n is a regenerative process, with state 0 being a regeneration point. Note that a regenerative cycle of X^n may consist of only one interarrival and one service time, both of which can be arbitrarily short. In particular, for τ^n denoting a generic cycle length of X^n, we have $P(\tau^n \le t) > 0$ for all $t > 0$, implying that X^n is nonlattice. By Sigman and Wolff [18, theorem 2.1(b)], we only need to demonstrate that X^n is a positive recurrent regenerative process. We prove this result by bounding the sample paths of X^n from above with a positive recurrent process via coupling the $M/M_{pc}/n + M_{pc}$ with an infinite-server queue. To this end, we give the two systems the same initial number of customers, and the same Poisson arrival process, letting the service time of each arrival to the infinite-server queue be equal to the service plus patience time of the corresponding customer in the $M/M_{pc}/n + M_{pc}$ system. In particular, with (S_i, T_i) denoting the service-patience times bivariate corresponding to the ith arrival to the $M/M_{pc}/n + M_{pc}$ system, we take $S_i + T_i$ to be the service time of the same arrival to the infinite-server system. Note that $S_i + T_i$ is exponentially distributed with rate $\theta/(1+\theta)$ because $S_i = \theta T_i$.

If $X^n(0) = K > 0$, then we endow each initial customer k, $1 \le k \le K$, with a bivariate (S_k, T_k), such that S_k is exponentially distributed with mean 1, $T_k = S_k/\theta$, and these K bivariates are i.i.d. We let the remaining service time of each such customer k in the infinite-server queue be $S_k + T_k$ (so that it is exponentially distributed with rate $\theta(1+\theta)^{-1}$), and the remaining service time in the $M/M_{pc}/n + M_{pc}$ system be an arbitrary number that is no larger than S_k; the remaining time to abandon of customer k that is waiting in the $M/M_{pc}/n + M_{pc}$ queue is no larger than T_k.

Under this construction, the infinite-server queue is an $M/M/\infty$ system. Because the time that a customer with patience T and service requirement S spends in the $M/M_{pc}/n + M_{pc}$ is smaller than $S + T$ w.p.1, the kth initial customer and the ith arrival after time 0 depart the $M/M_{pc}/n + M_{pc}$ system (either via service completion or abandonment) before they depart the infinite-server system, implying that the sample path of the queue in the latter system is no smaller than in the former w.p.1. In turn, whenever the $M/M/\infty$ system is empty, so is the $M/M_{pc}/n + M_{pc}$ system. Now, the $M/M/\infty$ queue is an ergodic continuous-time Markov chain (CTMC), regardless of the values of the arrival and service rates, and so its expected busy cycle is finite. This immediately implies that the regenerative cycle length is finite w.p.1 in the $M/M_{pc}/n + M_{pc}$ system as well. □

Henceforth, we let $X^n(\infty)$ denote a random variable having the unique stationary (and limiting) distribution of the process X^n.

3. Main Results

In this section, we present the main results of the paper, namely the diffusion and LOF limit, and the corresponding weak limits for the stationary distributions. Throughout, we assume that (1) holds; the specific range of values that β achieves is specified in the formal statements.

3.1. Limit Theorems Under Diffusion Scaling

The diffusion limit is achieved under the usual many-server diffusion scaling for the scaled number-in-system process:

$$\widehat{X}^n := n^{-1/2}(X^n - n).$$

We note that because X^n is not a Markov process, the value of $X^n(0)$ does not determine the law of X^n. Nevertheless, we can characterize the dynamics of X^n without resorting to infinite-dimensional (measure-valued) Markov representation for a special class of natural initial conditions. In particular, we can consider the case in which the system has started operating before time 0, such that all of the customers at time 0 are in service, and none of them experienced any wait before entering service. (For example, the system can be initialized empty.) In this case, the remaining service times of all the customers in the system at time 0 are i.i.d. exponentially distributed

random variables with mean 1. We can slightly generalize this initial condition by allowing $X^n(0)$ to be larger than n, but require that the waiting time of each customer in queue at time 0 is equal to 0.

To simplify the exposition, we first state the stochastic-process limit theorems under the previous assumption on the initial condition (see Assumption 1). However, we remark that we must consider much more general initial conditions in order to prove the limit theorems for the stationary distributions. Thus, we substantially generalize Assumption 1 in Section 3.3 (see (Ia) and (Ib) there), and prove the process limit theorems in the generalized setting.

Recall that $Z_1^n(0)$ is the number of customers in phase 1 service at time 0. Let $\ell_i^n = \{\ell_i^n(t) : t \geq 0\}$ be the elapsed waiting time of the ith customer (labeled in descending order of their arrival times) in queue at time t, $i \geq 1$, where $\ell_i^n(t) := 0$ for $i > Q^n(t)$.

Assumption 1 (Initial Condition). *$Z_1^n(0) = 0$ and $\sum_{i=1}^{Q^n(0)} \ell_i^n(0) = 0$ w.p.1.*

In particular, the condition $Z_1^n(0) = 0$ implies that the remaining service times of all the customers in service at time 0 are exponentially distributed with mean 1.

The following functional central limit theorem (FCLT) shows that, for large n, the $M/M_{pc}/n + M_{pc}$ system behaves much like the Erlang-C model. We remark that the asymptotic relation between the two systems is more intricate than what the diffusion limit reveals, as the LOF limit in Theorem 4 will show.

Theorem 2 (Diffusion Limit). *Assume that (1) holds with $\beta \in \mathbb{R}$. If Assumption 1 holds and, in addition, $\widehat{X}^n(0) \Rightarrow X_0$ in $\mathbb{R}$, then $\widehat{X}^n \Rightarrow \widehat{X}_C$ in D as $n \to \infty$, for $\widehat{X}_C$ in (2).*

It is well known that the solution to the SDE (2) has a unique steady-state distribution when $\beta > 0$, which is exponential on the positive real line, and normal on the negative real line; see theorem 1 and corollary 2 in Halfin and Whitt [7]. In particular, let $\widehat{X}_C(\infty)$ denote a random variable with that steady-state distribution, and let Φ denote the cumulative distribution function (cdf) of the standard normal random variable. Then,

$$P(\widehat{X}_C(\infty) > x | \widehat{X}_C(\infty) > 0) = e^{-\beta x}, \quad \text{for } x > 0, \tag{5}$$

$$P(\widehat{X}_C(\infty) \leq x | \widehat{X}_C(\infty) \leq 0) = \Phi(\beta + x)/\Phi(\beta), \quad \text{for } x \leq 0, \tag{6}$$

where

$$P(\widehat{X}_C(\infty) > 0) = [1 + \sqrt{2\pi}\beta\Phi(\beta)e^{\beta^2/2}]^{-1}. \tag{7}$$

On the other hand, when $\beta \leq 0$, the diffusion process $\widehat{X}_C$ is either null recurrent (when $\beta = 0$) or transient (when $\beta < 0$). This follows easily from the fact that $\widehat{X}_C$ is distributed like an ergodic Ornstein–Uhlenbeck process on $(-\infty, 0)$, and like a Brownian motion on $(0, \infty)$, which is driftless in the case $\beta = 0$, and has a positive drift when $\beta < 0$.

We next characterize the limits of the stationary distributions of $\widehat{X}^n$ for the two cases in which (i) the time-limiting behavior of $\widehat{X}_C$ exists, namely, when $\beta < 0$, and (ii) when $\beta > 0$. To this end, we say that a sequence of random variables Y^n converges in distribution to infinity, and write $Y^n \Rightarrow \infty$, if $P(Y^n > M) \to 1$ as $n \to \infty$ for any $M > 0$.

Theorem 3. *The following hold for the sequence $\{\widehat{X}^n(\infty) : n \geq 1\}$ as $n \to \infty$:*

i. *If $\beta > 0$, then $\widehat{X}^n(\infty) \Rightarrow \widehat{X}_C(\infty)$.*

ii. *If $\beta < 0$, then $\widehat{X}^n(\infty) \Rightarrow \infty$.*

When $\beta > 0$, Theorems 1, 2, and 3 imply the following interchange of limits:

$$\lim_{t\to\infty} \lim_{n\to\infty} P(\widehat{X}^n(t) > x) = P(\widehat{X}_C(\infty) > x) = \lim_{n\to\infty} \lim_{t\to\infty} P(\widehat{X}^n(t) > x), \quad \text{for all } x \in \mathbb{R}. \tag{8}$$

Note that an analogous limit-interchangeability result holds for the Erlang-C system (see theorem 1 and corollary 2 in Halfin and Whitt [7]). Roughly speaking, Theorems 2 and 3 suggest that the perfect correlation between service time and patience time "removes" the effect of abandonment when $\beta > 0$ for sufficiently large systems. Given that the diffusion limit when $\beta \leq 0$ is transient, it stands to reason that an analogous result to assertion (ii) of Theorem 3 holds when $\beta = 0$; this can be proved in the special case $\theta < \mu = 1$.

Proposition 1. *Let $\beta = 0$. If $\theta < 1$, then $\widehat{X}^n(\infty) \Rightarrow \infty$ as $n \to \infty$.*

3.2. Limit Theorems Under the LOF Scaling When $\beta \le 0$

Theorem 2 shows a discrepancy between the diffusion limit and the prelimit when $\beta < 0$, as the process X^n is ergodic for all $n \ge 1$, whereas the diffusion limit $\widehat{X}_C$ is transient. Theorem 3 further emphasizes this discrepancy by showing that $\{\widehat{X}^n(\infty) : n \ge 1\}$ converges weakly to infinity. In turn, this latter result implies that $\widehat{X}^n$ needs to be spatially scaled in order to achieve a nontrivial limit as $t \to \infty$. The LOF stated in Theorem 4 below identifies the exact additional spatial scaling of $\widehat{X}^n$ to be $n^{1/4}$.

However, the spatial scaling of $n^{3/4}$ by itself is not sufficient to obtain a nontrivial process limit. To see why, note that, for an unstable $M/M/n$ (Erlang-C) system with arrival rate λ^n and service rate 1, such that $\lambda^n = n - \beta\sqrt{n}$ for some $\beta < 0$, it takes $\Theta(n^{1/4})$ units of time for the queue to grow by $\Theta(n^{3/4})$. Because, for the $M/M_{pc}/n + M_{pc}$ sequence, the diffusion limit of $\{\widehat{X}^n : n \ge 1\}$ is the same as that of a sequence of Erlang-C systems with the same arrival and service rates by Theorem 2, it stands to reason that the two sequences of systems share the same growth rate of the queues, which is indeed the case as we show next. Thus, consider the following process:

$$\widetilde{X}^n(t) := \frac{\widehat{X}^n(n^{1/4}t)}{n^{1/4}} = \frac{X^n(n^{1/4}t) - n}{n^{3/4}}, \quad t \ge 0.$$

The next theorem characterizes the weak limit of $\widetilde{X}^n$ as the unique solution to an initial-value problem (IVP), which is why we refer to that limit as a fluid limit. (It is an LOF limit due to the spatial scaling, which is of lower order than the typical spatial scaling by n that gives rise to functional weak laws.)

Theorem 4 (LOF limit). *Assume that (1) holds with $\beta \le 0$. If Assumption 1 holds and in addition, $\widetilde{X}^n(0) \Rightarrow x_0$ in $\mathbb{R}$, where $x_0 \ge 0$ is deterministic, then $\widetilde{X}^n \Rightarrow x_F$ in D as $n \to \infty$, where x_F is the unique solution to the IVP*

$$\dot{x}_F = -\beta - \frac{\theta^2}{2}x_F^2, \quad x_F(0) = x_0. \tag{9}$$

Remark 1 (The Necessity of $x_0 \ge 0$). Although the IVP (9) has a unique solution for all $x_0 \in \mathbb{R}_+$, the only relevant solutions are for $x_0 \ge 0$. To see why, note that $x_0 < 0$ implies that there are idle agents initially, and in particular, that $X^n(0) = n - \Theta_P(n^{3/4}) < n$ for all n large enough. Now, for $t_0^n := \inf\{t \ge 0 : X^n(t) = n\}$ (namely, t_0^n is the first time in which all agents are busy), the departure rate from the system is $n - X^n(t) = O_P(n^{3/4})$ for all $t \in [0, t_0^n)$, whereas the arrival rate is λ^n, so that the idleness decreases at rate $\lambda^n - X^n(t) = O_P(n^{3/4})$ over $[0, t_0^n)$. In turn,

$$\lim_{n\to\infty} P\left(\sup_{t\in(0,\epsilon]} \widetilde{X}^n(t) \ge -\epsilon\right) = 1, \quad \text{for all } \epsilon > 0.$$

Thus, if $x_0 < 0$, a limit process x of $\widetilde{X}^n$ is not right-continuous at $t = 0$ because $x(t) \ge 0$ for all $t > 0$, implying that $x \notin D$. As a result, $\widetilde{X}^n$ does not converge in D when $\widetilde{X}^n(0) \Rightarrow x_0 < 0$ in $\mathbb{R}$.

For $x_0 \ge 0$, one can check that (9) has a closed-form expression, depending on the value of β.

Corollary 1 (Closed-form Solution). *The unique solution x_F to (9) is*

$$x_F(t) = \frac{\sqrt{-2\beta}}{\theta} \frac{\left(\sqrt{-2\beta} + \theta x_0\right)\left(1 - e^{-\sqrt{-2\beta}\theta t}\right) + 2\theta x_0 e^{-\sqrt{-2\beta}\theta t}}{\left(\sqrt{-2\beta} + \theta x_0\right)\left(1 - e^{-\sqrt{-2\beta}\theta t}\right) + 2\sqrt{-2\beta} e^{-\sqrt{-2\beta}\theta t}}, \text{ when } \beta < 0, \tag{10}$$

and

$$x_F(t) = \frac{2x_0}{2 + \theta^2 x_0 t} \text{ when } \beta = 0. \tag{11}$$

A point $a \in \mathbb{R}_+$ is a stationary point of x_F if $x_F(t) = a$ for all $t \ge 0$ whenever $x_F(0) = a$; it is $\mathbb{R}_+$-globally asymptotically stable (and then also the unique stationary point), if $x_F(t) \to a$ as $t \to \infty$, for any $x_0 \in \mathbb{R}_+$. Let

$$x^* := \sqrt{-2\beta}/\theta \tag{12}$$

The next corollary follows immediately from Corollary 1.

Corollary 2 (Stability of the IVP). *x^* is an $\mathbb{R}_+$-globally asymptotically stable stationary point of* (9).

In fact, any solution x_F to (9) approaches x^* monotonically, as can be seen from (10) and (11), or alternatively, from the fact that $\dot{x}_F(t) < 0$ whenever $x_F(t) > x^*$, and $\dot{x}_F(t) > 0$ whenever $x_F(t) < x^*$ (the latter being relevant only when $\beta < 0$).

Due to the $n^{1/4}$ time scaling, the LOF captures how the dynamics of an $M/M_{pc}/n + M_{pc}$ system differ from that of an Erlang-C system when $\beta < 0$; the latter is characterized by the limit in the following statement. Recall that η is the identity process, $\eta(t) = t$, $t \geq 0$, and that X_C^n denotes an the queue in an $M/M/n$ system. We use $\widetilde{X}_C^n$ to denote the LOF-scaled queue process.

Proposition 2. *If $\beta \leq 0$ and $\widetilde{X}_C^n(0) \Rightarrow x_0$ in $\mathbb{R}$ for $x_0 \geq 0$, then $\widetilde{X}_C \Rightarrow x_C := x_0 - \beta\eta$ in D as $n \to \infty$.*

Note that the solution x_F to (9) is equal to x_C in Proposition 2 when $\theta = 0$. Observe also that, when $\beta < 0$ and $x_F(0) = x_C(0)$, $x_F(t) < x_C(t)$ for all $t > 0$. Further, $x_C(t) \to \infty$ while $x_F \to x^*$, as $t \to \infty$. Indeed, the divergence of $x_C(t)$ to infinity as $t \to \infty$ corresponds to the fact that X_C^n is transient when $\beta < 0$. In contrast, an $M/M_{pc}/n + M_{pc}$ system is always stable due to the abandonment.

Analogously to Theorem 3, we can prove that x^* is the weak limit for the stationary random variables

$$\widetilde{X}^n(\infty) := \frac{\widehat{X}^n(\infty)}{n^{1/4}} = \frac{X^n(\infty) - n}{n^{3/4}}.$$

Theorem 5. *If $\beta \leq 0$, then $\widetilde{X}^n(\infty) \Rightarrow x^*$ in $\mathbb{R}$ as $n \to \infty$.*

Theorems 1, 4, and 5 again imply the following interchangeability of limits:

$$\lim_{t\to\infty}\lim_{n\to\infty} P(\widetilde{X}^n(t) > x) = 1\{x^* > x\} = \lim_{n\to\infty}\lim_{t\to\infty} P(\widetilde{X}^n(t) > x), \quad \text{for all } x \in \mathbb{R}_+\backslash\{x^*\}. \tag{13}$$

Let $Q^n(\infty)$ denote a random variable with the steady-state distribution of the queue process; $Q^n(\infty) = (X^n(\infty) - n)^+$. Because $x^* > 0$ when $\beta < 0$, Theorem 5 implies that $Q^n(\infty)$ is $\Theta_P(n^{3/4})$.

In ending, we remark that the time scaling in $\widetilde{X}^n$ implies that the relaxation time of X^n, namely, the time it takes to X^n to converge to its steady state (under any metric), is increasing in n when $\beta \leq 0$.

3.3. Generalizing the Initial Condition

The process limit results in Theorems 2 and 4 are both achieved under Assumption 1. However, to prove the limit theorems for the stationary distributions, we need to allow for more general initial conditions. (In particular, Theorems 3 and 5 will be proved by initializing the corresponding processes according to their stationary distribution.) To this end, we assume that, similarly to the customers arriving after time 0, the service time of each customer that is in service at time 0 also has two phases: phase 1 (corresponding to the delay that that customer experienced in queue before entering service) and phase 2, which is exponentially distributed with mean 1.

Observe that Theorems 2 and 4 do not necessarily hold if there are customers in phase 1 service initially. For example, if for some $c > 0$, all customers in service present at time 0 have at least c time units of remaining phase 1 service time, then there will be no departures from service in the first c time units, so that the queue will grow at rate λ^n in system n, and the system will have an initial period of overload. In that initial period, an FCLT clearly cannot hold. Hence, to generalize the initial condition in Theorems 2 and 4, we must enforce regularity conditions that prohibit such overload incidents.

Let $r_j^n(t)$ be the remaining phase 1 service time of the customer with server j at time t, $1 \leq j \leq n$, for all $t \in \mathbb{R}_+$, or $r_j(t) = 0$ if the customer is in phase-2 or server j is idle at time t. Recall also that $\ell_i^n(t)$ is the elapsed waiting time of the ith customer in the queue at time t. Let

$$L^n(t) := \sum_{j=1}^{Z^n(t)} r_j^n(t) + \sum_{i=1}^{Q^n(t)} \ell_i^n(t), \quad t \geq 0. \tag{14}$$

Proposition 3. *L^n possesses a unique stationary distribution, which is also the limiting distribution of $L^n(t)$ as $t \to \infty$.*

Proof. Similarly to X^n, L^n is a regenerative process, regenerating when X^n hits state 0, namely, when the system empties. It follows from Theorem 1 that L^n is nonlattice and the expected cycle length of L^n is finite, so that L^n is positive recurrent, implying the result. □

The proof of the next proposition appears in Section 5.7. Let $L^n(\infty)$ denote a random variable that has the stationary distribution of L^n.

Proposition 4. *The following hold:*
a. *If* $\beta > 0$, *then* $E[L^n(\infty)] = O(1)$.
b. *If* $\beta \le 0$, *then* $E[L^n(\infty)] = O(n^{1/2})$.

Let

$$\widehat{L}^n(0) := \frac{L^n(0)}{n^{1/2}} \quad \text{and} \quad \widetilde{L}^n(0) := \frac{L^n(0)}{n^{3/4}},$$

and consider the families of initial conditions satisfying the following: for a random variable X_0,

$$(\widehat{X}^n(0), \widehat{L}^n(0)) \Rightarrow (X_0, 0) \text{ in } \mathbb{R}^2 \text{ as } n \to \infty, \tag{Ia}$$

$$(\widetilde{X}^n(0), \widetilde{L}^n(0)) \Rightarrow (X_0, 0) \text{ in } \mathbb{R}^2 \text{ as } n \to \infty, \text{ where } X_0 \ge 0 \text{ w.p.1.} \tag{Ib}$$

Theorem 6. *Assume that* (1) *holds with* $\beta \in \mathbb{R}$. *If* (Ia) *holds, then* $\widehat{X}^n \Rightarrow \widehat{X}_C$ *in* D *as* $n \to \infty$.

Theorem 7. *Assume that* $\beta \le 0$. *If* (Ib) *holds, then* $\widetilde{X}^n \Rightarrow x_F$ *in* D *as* $n \to \infty$, *where, conditional on* $\{X_0 = x_0\}$, *for a positive scalar* x_0, x_F *is the unique solution to the IVP* (9).

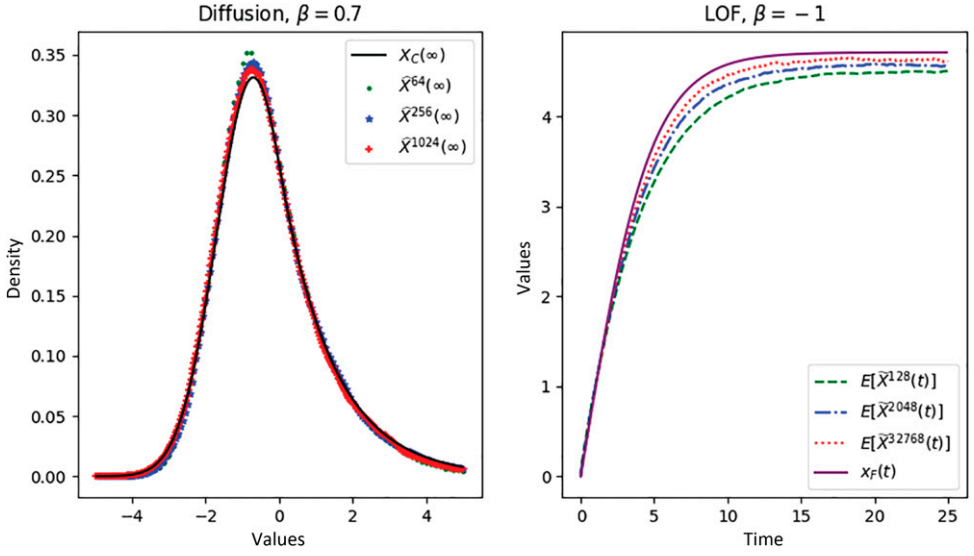
Both Conditions (Ia) and (Ib) hold trivially if Assumption 1 holds. As a result, Theorems 6 and 7 immediately imply the statements in Theorems 2 and 4, respectively. Due to Proposition 4, (Ia) and (Ib) also hold when the system is initialized at stationarity, that is, when $L^n(0) \overset{\mathrm{d}}{=} L^n(\infty)$, a result needed to prove Theorems 3 and 5.

3.4. A Numerical Example

In this section, we demonstrate the convergence of the properly scaled version of X^n to the diffusion limit X_C and the LOF limit x_F. We fix $\mu = 1$, $\theta = 0.3$, and vary n for different system sizes. For the diffusion limit, we take $\beta = 0.7$ and choose $n \in \{64, 256, 1,024\}$ for three different $M/M_{pc}/n + M_{pc}$ systems. To estimate the probability density function (pdf) of $\widetilde{X}^n(\infty)$, we generate 512 independent sample paths. Each was ran for 100 time units, with a warm-up period of 50 time units. We compare the estimated pdf of the three simulations with the pdf of the diffusion limit $X_C(\infty)$, given in Garnett et al. [6, theorem 2*]. This comparison is depicted in the left panel of Figure 1.

To demonstrate the convergence to the LOF limit, we fix $\beta = -1.0$ and estimate $\{E[\widetilde{X}^n(t)] : t \in [0, 25]\}$ for three different $M/M_{pc}/n + M_{pc}$ systems with different values of n. We initialize each of the three systems such that, at

Figure 1. (Color online) Convergence of $\hat{X}^n(\infty)$ to $X_C(\infty)$ when $\beta > 0$ (left panel), and $E[\widetilde{X}^n]$ to x_F when $\beta < 0$ (right panel).



time 0, there are n customers in the system, all of which are in phase 2 of their service, while no customer is waiting in the queue. For $n \in \{128, 2,048, 32,768\}$, we generate {8,128, 2,048, 512} independent sample paths for each system. We need more sample paths for smaller systems because for each t, $\widetilde{X}^n(t)$ has a higher coefficient of variation than for larger n. The systems are run for $25n^{1/4}$ time units and we compute $\{E[\widetilde{X}^n(t)] : t \in [0,25]\}$ by averaging over the independent sample paths. We plot $\{x_F(t) : t \in [0,25]\}$ by using the closed-form solution (10). The comparison of the simulations to x_F is depicted in the right panel of Figure 1.

4. Summary

In this paper, we consider the $M/M_{pc}/n + M_{pc}$ model, in which the patience and service times of each customer are perfectly correlated. We prove that, under the usual square-root staffing rule and diffusion scaling, the queue process of the $M/M_{pc}/n + M_{pc}$ system is asymptotically equivalent to the queue process of the Erlang-C system in the sense that both queues have same diffusion limit. Furthermore, when the sequence of $M/M/n$, $n \geq 1$, is stable for all n large enough (namely, when the traffic intensity is smaller than 1), the sequences of stationary distributions for both models have the same limit as well. When the traffic intensity is larger than 1 for all n large enough, the sequence of stationary distributions of the $M/M_{pc}/n + M_{pc}$ system converges weakly to infinity (a result that also holds in the critical case, when $\beta = 0$, provided the abandonment rate is smaller than the service rate). These results demonstrate the diminishing impact of abandonment on the queueing dynamics $M/M_{pc}/n + M_{pc}$ as n increases.

However, unlike the Erlang-C model, which is not always stable, the $M/M_{pc}/n + M_{pc}$ model is stable, and in particular, converges weakly with time to a unique stationary distribution, regardless of the traffic intensity. To approximate the stationary distribution of the $M/M_{pc}/n + M_{pc}$ system, we consider the LOF limit and its unique stationary point x^* in (12), which is proved to be the many-server heavy-traffic limit of the stationary queue process.

Even though all the results in this paper hold for the $M/M_{pc}/n + M_{pc}$ system, it stands to reason that similar results hold under less restrictive assumption on the correlation between service and patience times. In particular, we expect that positive correlation between these two random variables causes a system to be more heavily loaded than when the service and patience are independent (or negatively correlated). We leave proving this open problem for the future; see the discussion in Moyal and Perry [11].

5. Proofs of Main Results

To help navigate the proofs, we provide the following roadmap: First, Theorems 2 and 4 follow immediately from Theorems 6 and 7. The proofs of Theorems 6 and 7 appear in Sections 5.3 and 5.4, respectively, after providing sample-path and martingale representations in Sections 5.1 and 5.2. Theorem 5, whose proof relies on Theorem 7, is proved in Section 5.5, where we also prove Theorem 3 by utilizing both Theorem 6 and Theorem 5. The proof of Proposition 1 appears in Section 5.6, building on a coupling result between two $M/M_{pc}/n + M_{pc}$ systems. Utilizing the same coupling, Proposition 4 is proved in Section 5.7. Many proofs in this section build on auxiliary results, the proofs of which are relegated to Section 6. Finally, the proof of Proposition 2 appears in Appendix B.

5.1. Sample-Path Representation

Let

$$\beta^n := n^{-1/2}(n - \lambda^n), \tag{15}$$

and note that, due to the square-root staffing rule in (1), $\beta^n \to \beta$ as $n \to \infty$.

Recall that $Z_i(t)$ ($i = 1, 2$) are the number of customers in phase I service at time t and $Z(t) = Z_1(t) + Z_2(t)$. We can write

$$Z_1^n(t) = \sum_{j=1}^{Z^n(t)} 1\{r_j^n(t) > 0\}, \tag{16}$$

which provides a representation of $Z_1 \in D$ and $Z_2 = Z - Z_1 \in D$.

Let A, S, and R be three independent unit-rate Poisson processes. We represent the Poisson arrival process in system n via $A^n(t) := A(\lambda^n t)$, $t \geq 0$, and exploit the memoryless property of the exponential distribution to characterize the departures from service and abandonment. In particular, for $D^n(t)$ and $R^n(t)$ denoting the number of

departures from service and number of abandonment by time t in system n, respectively, we have

$$D^n(t) = S\left(\int_0^t Z_2^n(s)ds\right) \quad \text{and } R^n(t) = R\left(\theta\int_0^t Q^n(s)ds\right), \quad t \geq 0.$$

Then,

$$X^n(t) = X^n(0) + A(\lambda^n t) - S\left(\int_0^t Z_2^n(s)ds\right) - R\left(\theta\int_0^t Q^n(s)ds\right), \quad t \geq 0. \tag{17}$$

Notice that the following basic equalities hold:

$$Q^n = (X^n - n)\vee 0, \quad Z^n = X^n \wedge n, \quad Z^n = Z_1^n + Z_2^n. \tag{18}$$

To fully characterize X^n, we need to characterize Z_2^n, or equivalently, Z_1^n. Let $Z_0^n(t)$ denote the number of customers who were in the system initially (at time 0), and are in their phase 1 service at time t. Let T_0^n be the time in which the last customer from the initial queue leaves the queue, either by entering service or by abandoning the queue; in particular, at any $t < T_0^n$ there are customers in queue that were waiting in the queue at time 0, and there are no such customers in the queue at any time $t \geq T_0^n$. For any $t \geq 0$, let $w^n(t)$ be the minimum between t and the waiting time of the head-of-line customer. We set $w^n(t) := 0$ if $Q^n(t) = 0$.

Now, if a departure from service occurs at time $s \in [T_0^n, t]$ and $Q^n(s-) > 0$, then the customer at the head of the line begins the phase 1 service, and that customer is still in phase 1 at time t if and only if $\theta w^n(s-) + s > t$. Note that the latter statement holds trivially if $Q^n(s-) = 0$, because then $w^n(s-) = 0$. We can therefore characterize Z_1^n via the departure process as follows:

$$Z_1^n(t) = Z_0^n(t) + \int_{T_0^n\wedge t}^t 1\{\theta w^n(s-) + s > t\}dD^n(s), \quad t \geq 0. \tag{19}$$

To characterize the process w^n, we number the customers that arrive after time 0 by the order of their arrival, and denote by E_k^n the arrival time of the kth customer to system n, that is, $E_k^n := \inf\{t : A^n(t) = k\}$. Let T_k^n denote the patience time of the kth arrival to system n, so that $\{T_k^n : k \geq 1\}$ is a sequence of independent exponential random variable with mean $1/\theta$ for each $n \geq 1$. Under the FCFS policy, the arrival time of any customer that is in queue at time t is no less the arrival time of the head-of-line customer at that time, the latter being equal to $t - w^n(t)$. Hence, if the kth customer arrives during the time interval $[t - w^n(t), t)$, then that customer is still in the system (waiting in queue) at time t if and only if $E_k^n + T_k^n > t$. This gives

$$Q^n(t) = \int_{t-w^n(t)}^t 1\{E^n_{A^n(s)} + T^n_{A^n(s)} > t\}dA^n(s) + Q_0^n(t), \text{ for all } t \geq 0, \tag{20}$$

where $Q_0^n(t)$ is the number of customers that were waiting in queue at time 0 and are still waiting in queue at time t. Note that, due to abandonment, $Q_0^n(t) \leq (Q^n(0) - D^n(t))^+$, and that there are no waiting customers at time t if there are idle agents, so that

$$(Z^n - n)w^n = 0\eta. \tag{21}$$

If we assume that $L^n(0) = 0$, so that $Q_0^n = Z_0^n = 0\eta$, then (17)–(20) characterize the system's dynamics via the primitives A^n, S^n, and $\{T_k^n : k \in \mathbb{Z}_+\}$. When $L^n(0) > 0$, the dynamics of the nth system depend also on $\{\ell_i^n(0)\}$ and $\{r_j^n(0)\}$. However, as will be proved later, the impact of these two sequences is asymptotically negligible, in that they do not alter the diffusion limit and LOF limit under our assumed initializations in (Ia) and (Ib).

5.2. A Martingale Representation

Let $\mathcal{F}_0^n$ be the σ-algebra generated by

$$\{X_0^n, \ell_i^n(0), r_j^n(0) : 1 \leq i \leq Q^n(0), 1 \leq j \leq Z^n(0)\},$$

augmented by including all P-null sets. For $t \geq 0$ and $n \geq 1$, let $\mathcal{F}^n := \{\mathcal{F}_t^n : t \geq 0\}$, where $\mathcal{F}_t^n$ is the right-continuous σ-algebra associated to the σ-algebra generated by

$$(\mathcal{F}_0^n, \ell_i^n(s), r_j^n(s), A^n(s), D^n(s), R^n(s) : 1 \leq i \leq Q^n(t), 1 \leq j \leq Z^n(t), s \in [0,t]).$$

Note that the processes $X^n, Q^n, Q_0^n, Z^n, Z_i^n$, $i = 0, 1, 2$, and w^n have sample paths in D by construction. Now, $w^n(t) = 1\{Q^n(t) > 0\}\ell_1^n(t)$, so that w^n is $\mathcal{F}^n$-adapted, and it therefore follows from (16)–(19) that X^n, Q^n, Z^n, and Z_i^n,

$i = 0, 1, 2$, are also $\mathcal{F}^n$-adapted. Finally, noting that $T_0^n = \inf\{t \geq 0 : w^n(t) > t\}$ shows that T_0^n is an $\mathcal{F}^n$-stopping time.

Consider the following processes:

$$M_A^n(t) := A^n(t) - \lambda^n t, \qquad M_S^n(t) := D^n(t) - \int_0^t Z_2^n(s)ds,$$

$$M_R^n(t) := R^n(t) - \theta \int_0^t Q^n(s)ds, \qquad t \geq 0.$$

Because $Z_2^n \leq n$ and $D^n(t) \leq S(nt)$, we have $E[|M_i^n(t)|] < \infty$ and $E[|M_i^n(t)|^2] < \infty$, for $i = A$ and S. Therefore, M_A^n and M_S^n are square-integrable $\mathcal{F}^n$-martingales. Note that R^n and Q^n have nonnegative sample paths that are bounded pathwise by the sample paths of $A^n + X^n(0)$. For $\tau_k^n := k1\{|X^n(0)| < k\}$, $M_R^n(\cdot \wedge \tau_k^n)$ is a square-integrable $\mathcal{F}^n$-martingale, and because $\tau_k^n \to \infty$ w.p.1 as $k \to \infty$, M_R^n is an $\mathcal{F}^n$-local martingale. Thus, (17) admits the following martingale representation:

$$X^n(t) = X^n(0) + \lambda^n t - \int_0^t Z_2^n(s)ds - \theta \int_0^t Q^n(s)ds + M_A^n(t) - M_S^n(t) - M_R^n(t).$$

Next, for

$$U_1^n(t) := \int_{T_0^n \wedge t}^t 1\{\theta w^n(s-) + s > t\} dM_S^n(s), \quad t \geq 0, \tag{22}$$

we can rewrite (19) to obtain

$$Z_1^n(t) = \int_{T_0^n \wedge t}^t 1\{\theta w^n(s) + s > t\} Z_2^n(s)ds + U_1^n(t) + Z_0^n(t), \tag{23}$$

so that

$$\begin{aligned}\int_0^t Z_1^n(s)ds &= \int_0^t (U_1^n(s) + Z_0^n(s))ds + \int_0^t \int_{T_0^n \wedge s}^s 1\{\theta w^n(u) + u > s\} Z_2^n(u)duds \\ &= \int_0^t (U_1^n(s) + Z_0^n(s))ds + \int_{T_0^n \wedge t}^t (\theta w^n(u)) \wedge (t-u) Z_2^n(u)du. \end{aligned} \tag{24}$$

The last integral in (24) follows from Fubini's theorem together with the fact that

$$\int_a^b 1\{s < c\}ds = b \wedge c - a \wedge c, \quad \text{for } a \leq b,$$

so that

$$\int_u^t 1\{\theta w^n(u) + u > s\}ds = (\theta w^n(u) + u) \wedge t - u = \theta w^n(u) \wedge (t-u).$$

Finally, let

$$F^n(s,t) := \int_0^s 1\{E_{A^n(u)}^n + T_{A^n(u)}^n > t\} dA^n(u) + \theta^{-1}\lambda^n(e^{-\theta t} - e^{-\theta(t-s)}). \tag{25}$$

Then, for

$$U_2^n(t) := F^n(t,t) - F^n(t - w^n(t), t), \tag{26}$$

we can rewrite (20) as follows:

$$Q^n(t) = \theta^{-1}\lambda^n(1 - e^{-\theta w^n(t)}) + U_2^n(t) + Q_0^n(t), \; t \geq 0. \tag{27}$$

Plugging (24) and (27) in (17), and using the equality $Z = Z_1 + Z_2$, give the following modified martingale representation:

$$X^n(t) = X^n(0) + \lambda^n t - \int_0^t Z^n(s)ds + V^n(t) + \int_0^t (Z_0^n(t) - \theta Q_0^n(s))ds$$
$$+ \int_0^t (U_1^n(s) - \theta U_2^n(s))ds + M_A^n(t) - M_S^n(t) - M_R^n(t), \quad t \geq 0, \tag{28}$$

where

$$V^n(t) := \int_{T_0^n \wedge t}^t (\theta w^n(s)) \wedge (t-s) Z_2^n(s)ds - \int_0^t \lambda^n (1 - e^{-\theta w^n(s)})ds. \tag{29}$$

5.3. Proof of Theorem 6

We consider the following diffusion-scaled random variables and processes:

$$\widehat{Q}^n := n^{-1/2}Q^n, \quad \widehat{Z}^n := n^{-1/2}(Z^n - n) \quad \widehat{L}^n(0) := n^{-1/2}L^n(0), \quad \widehat{w} := n^{1/2}w^n,$$
$$\widehat{Z}_1^n := n^{-1/2}Z_1^n, \quad \widehat{Z}_2^n := n^{-1/2}(Z_2^n - n), \quad \widehat{Q}_0^n := n^{-1/2}Q_0^n, \quad \widehat{Z}_0^n := n^{-1/2}Z_0^n.$$

We similarly consider the diffusion-scaled processes in the martingale representation

$$\widehat{M}_i^n := n^{-1/2}M_i^n, \ i = A, S, R, \quad \widehat{U}_1^n := n^{-1/2}U_1^n, \quad \widehat{U}_2^n := n^{-1/2}U_2^n, \quad \widehat{V}^n := n^{-1/2}V^n.$$

Using the diffusion scaling in (28) gives

$$\widehat{X}^n(t) = \widehat{X}^n(0) - \beta^n t - \int_0^t \widehat{Z}^n(s)ds + \widehat{V}^n(t) + \int_0^t \left(\widehat{Z}_0^n(s) - \theta \widehat{Q}_0^n(s)\right)ds$$
$$+ \int_0^t \left(\widehat{U}_1^n(s) - \theta \widehat{U}_2^n(s)\right)ds + \widehat{M}_A^n(t) - \widehat{M}_S^n(t) - \widehat{M}_R^n(t). \tag{30}$$

The proof of Theorem 6 is a straightforward application of the continuous mapping theorem, given the following key result, whose proof appears in Section 6.4.

Proposition 5. *Assume that* (Ia) *holds. Then, as* $n \to \infty$,

a. $\left(\int_0^\cdot \widehat{Q}_0^n(s)ds, \int_0^\cdot \widehat{Z}_0^n(s)ds\right) \Rightarrow (0\eta, 0\eta)$ *in* D^2;

b. $\left(\int_0^\cdot \widehat{U}_1^n(s)ds, \int_0^\cdot \widehat{U}_2^n(s)ds, \widehat{V}^n\right) \Rightarrow (0\eta, 0\eta, 0\eta)$ *in* D^3; *and*

c. $(\widehat{M}_A^n, \widehat{M}_S^n, \widehat{M}_R^n) \Rightarrow (B_1, B_2, 0\eta)$ *in* D^3, *where* B_1 *and* B_2 *are two independent standard Brownian motions.*

Proof of Theorem 6. Using the equality $\widehat{Z}^n = \widehat{X}^n \wedge 0$ and (15) in (30), we have

$$\widehat{X}^n(\cdot) - \widehat{X}^n(0) + \beta^n \eta(\cdot) - \int_0^\cdot \widehat{X}^n(s) \wedge 0ds \Rightarrow \sqrt{2}B(\cdot) \quad \text{in } D \text{ as } n \to \infty, \tag{31}$$

where B is a standard Brownian motion.

By Pang et al. [12, theorem 4.1], there exists a unique solution $x \in D$ to the integral equation

$$x(t) = x(0) - \beta t - \int_0^t x(s) \wedge 0ds + y(t), \text{ for all } t \geq 0, \tag{32}$$

and the mapping $\phi : D \to D$, which maps the function y in (32) to the solution x, is continuous in the J_1 topology. Further, if y is continuous, then so is x. Hence, the statement of the theorem follows from (31) and the continuous mapping theorem, by noting that

$$\widehat{X}^n = \phi(\widehat{X}^n(\cdot) - \widehat{X}^n(0) + \beta^n \eta(\cdot) - \int_0^\cdot \widehat{X}^n(s) \wedge 0ds),$$

and that $\widehat{X}_C = \phi(\sqrt{2}B)$. □

5.4. Proof of Theorem 7

To establish the LOF limit, we consider the following scaled processes:

$$\widetilde{Q}^n(t) := n^{-3/4}Q^n(n^{1/4}t), \qquad \widetilde{Z}^n(t) := n^{-3/4}(Z^n(n^{1/4}t) - n),$$

$$\widetilde{Z}_1^n(t) := n^{-3/4}Z_1^n(n^{1/4}t), \qquad \widetilde{Z}_2^n(t) := n^{-3/4}(Z_2^n(n^{1/4}t) - n),$$

$$\widetilde{Q}_0^n(t) := n^{-3/4}Q^n(n^{1/4}t), \qquad \widetilde{Z}_0^n(t) := n^{-3/4}Z_0^n(n^{1/4}t),$$

$$\widetilde{U}_1^n(t) := n^{-3/4}U_1^n(n^{1/4}t), \qquad \widetilde{U}_2^n(t) := n^{-3/4}U_2^n(n^{1/4}t),$$

$$\widetilde{V}^n(t) := n^{-3/4}V^n(n^{1/4}t), \qquad \widetilde{L}^n(t) := n^{-3/4}L^n(n^{1/4}t),$$

$$\widetilde{w}^n(t) := n^{1/4}w^n(n^{1/4}t), \qquad \widetilde{T}_0^n := n^{-1/4}T_0^n,$$

and $\widetilde{M}_i^n(t) := n^{-3/4}M_i^n(n^{1/4}t)$, for $i = A, S, R$. Then the corresponding scaled process in (17) is represented via

$$\widetilde{X}^n(t) = \widetilde{X}^n(0) - \beta^n t - n^{1/4}\int_0^t \widetilde{Z}^n(s)ds + \widetilde{V}^n(t) + n^{1/4}\int_0^t \left(\widetilde{Z}_0^n(s) - \theta\widetilde{Q}_0^n(s)\right)ds$$
$$+ n^{1/4}\int_0^t \left(\widetilde{U}_1^n(s) - \theta\widetilde{U}_2^n(s)\right)ds + \widetilde{M}_A^n(t) - \widetilde{M}_S^n(t) - \widetilde{M}_R^n(t). \tag{33}$$

The proof of Theorem 7 builds on the following three supporting propositions, whose proofs appear in Section 6. Throughout, we assume that (Ib) holds.

Proposition 6. *As $n \to \infty$,*

a. $n^{1/4}\left(\int_0^\cdot Z_0^n(s)ds, \int_0^\cdot Q_0^n(s)ds\right) \Rightarrow (0\eta, 0\eta)$ in D^2 *and* $T_0^n \Rightarrow 0$ *in* $\mathbb{R}$;

b. $(\widetilde{M}_A^n, \widetilde{M}_S^n, \widetilde{M}_R^n) \Rightarrow (0\eta, 0\eta, 0\eta)$ *in* D^3;

c. $n^{1/4}\int_0^\cdot \widetilde{U}_1^n(s)ds \Rightarrow 0\eta$ *in* D; *and*

d. $n^{1/4}\widetilde{U}_2^n \Rightarrow 0\eta$, *so that* $n^{1/4}\int_0^\cdot \widetilde{U}_2^n(s)ds \Rightarrow 0\eta$ *in* D.

Proposition 7. $\{\widetilde{Q}^n : n \geq 1\}$ *is C-tight in* D.

Proposition 8. *As* $n \to \infty$

$$\widetilde{V}^n(\cdot) + \frac{\theta^2}{2}\int_0^\cdot (\widetilde{Q}^n(s))^2 ds \Rightarrow 0\eta \text{ in } D. \tag{34}$$

For a given $\phi \in D_0$, we say that $(y, \psi) \in D^2$ is a solution to the Skorokhod problem if

$$y = \phi + \psi;$$
$$\int_0^t y(s)d\psi(s) = 0, \text{ for all } t \geq 0;$$
$$y \geq 0, \psi(0) = 0 \text{ and } \psi \text{ is nondecreasing.} \tag{35}$$

It is well known (e.g., see Chen and Yao [2, theorem 6.1]) that the Skorokhod problem in (35) admits a unique solution (y, ψ), and that $h : D_0 \to D^2$, mapping the input ϕ to that solution, namely, the map defined via

$$h(\phi) := (y, \psi), \tag{36}$$

is (Lipschitz) continuous in the J_1 topology; see theorems 13.4.1 and 13.5.1 in Whitt [21]. (Continuity of h is proved only in the uniform topology in Chen and Yao [2].) Further, if ϕ is continuous, then so is $h(\phi)$.

Proof of Theorem 7. Due to Proposition 7, any subsequence of $\{\widetilde{Q}^n : n \geq 1\}$ has a further weakly converging subsequence in D. Let $\{\widetilde{Q}^k : k \geq 1\}$ denote such a converging subsequence, and let Q denote its weak limit. Let $\Phi^k \in D$

and $\Phi \in C$ be defined via

$$\Phi^k(t) = \widetilde{X}^k(t) + k^{1/4} \int_0^t \widetilde{Z}^k(s)ds, \tag{37}$$

$$\Phi(t) = X_0 - \beta t - \frac{\theta^2}{2} \int_0^t Q^2(s)ds. \tag{38}$$

By (33), Propositions 6 and 8, and the continuous mapping theorem, it holds that

$$\Phi^k - \widetilde{X}^k(0) + \beta^k \eta + \frac{\theta^2}{2} \int_0^{\cdot} (\widetilde{Q}^k(s))^2 ds \Rightarrow 0\eta \quad \text{in } D \quad \text{as } k \to \infty.$$

The convergence $\widetilde{Q}^k \Rightarrow Q$ in D and the continuous mapping theorem together give

$$(\widetilde{Q}^k, \Phi^k) \Rightarrow (Q, \Phi) \text{ in } D^2 \text{ as } k \to \infty.$$

We need the following lemma, the proof of which appears at the end of this section. Recall h from (36).

Lemma 1. $(\Phi^k, \widetilde{X}^k, \widetilde{X}^k - \Phi^k) \Rightarrow (\Phi, h(\Phi))$ *as* $k \to \infty$ *in* D^3.

Denote $(X, \Psi) := h(\Phi)$, so that $X = \Phi + \Psi$ and $X \geq 0$ w.p.1. Because h maps C_0 to C^2 and $\Phi \in C_0$, we have $(X, \Psi) \in C^2$. The convergence $X^k \Rightarrow X$ and the continuous mapping theorem imply that

$$\widetilde{Q}^k = \widetilde{X}^k \vee 0 \Rightarrow X \vee 0, \quad \text{in } D \quad \text{as } n \to \infty,$$

and thus $Q = X \vee 0 = X$, w.p.1. In particular, (38) simplifies to

$$\Phi(t) = X_0 - \beta t - \frac{\theta^2}{2} \int_0^t X^2(s)ds. \tag{39}$$

It follows from (35) and the fact that $(X, \Psi) = h(\Phi)$ that Ψ is a nondecreasing process with $\Psi(0) = 0$, such that

$$X = \Phi + \Psi \quad \text{and} \quad \int_0^{\cdot} 1\{X(s) > 0\} d\Psi(s) = 0\eta(\cdot).$$

Hence, conditional on $\{X_0 = x_0\}$, for $x_0 \geq 0$, and using (39), $(y, \psi) := (X, \Psi)$ satisfies the following:

$$\begin{aligned} &y(t) = x_0 - \beta t - \frac{\theta^2}{2} \int_0^t y^2(s)ds + \psi, \\ &\int_0^t 1\{y > 0\} d\psi = 0, \\ &(y, \psi) \in C^2,\ y \geq 0,\ \psi(0) = 0, \text{ and } \psi \text{ is a nondecreasing process.} \end{aligned} \tag{40}$$

The next lemma is proved at the end of this section.

Lemma 2. *There exists a unique solution* $(y, \psi) = (x_F, 0\eta)$ *to* (40) *for any input* $x_0 \geq 0$ *and* $\beta \leq 0$, *where* x_F *is the unique solution to* (9). *Further, the function* $g : \mathbb{R}_+ \to C^2$, *mapping* x_0 *to* (y, ψ), *is continuous.*

It follows that, conditional on $\{X_0 = x_0\}$, $X = x_F$ w.p.1, so that $(X, Q, Z) = (x_F, x_F \vee 0, 0\eta)$ w.p.1. The uniqueness of the limit implies the stated weak convergence. □

Proof of Lemma 1. For fixed $\tau > 0$ and $\epsilon > 0$, and for each $k \geq 1$ such that $k^{-3/4} < \epsilon$, define the event

$$\Xi^k \equiv \Xi^k(\epsilon, \tau) := \left\{ -\epsilon < \widetilde{X}^k(0),\ \inf_{t \leq \tau} \widetilde{X}^k(t) \wedge 0 < -3\epsilon \right\}.$$

We first show that Ξ^k is an asymptotically null event in the sense that $P(\Xi^k) \to 0$ as $n \to \infty$. To this end, let t_1^k be such that $\widetilde{X}^k(t_1^k) < -3\epsilon$. For

$$t_2^k := \sup \left\{ t < t_1^k : \widetilde{X}^k(t) \geq -\epsilon \right\},$$

it holds that $\widetilde{X}^k(t_2^k-) \geq -\epsilon$. As X^k is a pure jump process with jumps of size 1 and –1 w.p.1,

$$\widetilde{X}^k(t_2^k) \geq \widetilde{X}^k(t_2^k-) - k^{-3/4} > -2\epsilon.$$

Let $\phi := -\beta - \theta^2(Q)^2/2 \in D$ so that $\Phi(t) = \Phi(0) + \int_0^t \phi(s)ds$ for $t \geq 0$:

$$\begin{aligned} -\epsilon > \widetilde{X}^k(t_1^k) - \widetilde{X}^k(t_2^k) &= \Phi^k(t_1^k) - \Phi^k(t_2^k) - k^{1/4}\int_{t_2^k}^{t_1^k} \widetilde{X}^k(s) \wedge 0 ds \\ &\geq \Phi(t_1^k) - \Phi(t_2^k) - 2\|\Phi^k - \Phi\|_\tau + k^{1/4}(t_1^k - t_2^k)\epsilon \\ &\geq -(t_1^k - t_2^k)\|\phi\|_\tau - 2\|\Phi^k - \Phi\|_\tau + k^{1/4}(t_1^k - t_2^k)\epsilon. \end{aligned}$$

This strict inequality can hold if either $\|\phi\|_\tau \geq k^{1/4}$ or $\|\Phi^k - \Phi\|_\tau \geq \epsilon/2$, implying that

$$\Xi^k \subseteq \{\|\Phi^k - \Phi\|_\tau \geq \epsilon/2\} \cup \{\|\phi\|_\tau \geq k^{1/4}\}.$$

As both events on the right-hand side are asymptotically null under the probability measure P, we conclude that $P(\Xi^k) \to 0$ as $n \to \infty$.

Next, $X_0 \geq 0$ implies that $P(\widetilde{X}^k(0) > -\epsilon) \to 1$. Together with the fact that $P(\Xi^k) \to 0$, we have

$$P\left(\inf_{t\leq\tau} \widetilde{X}^k(t) \wedge 0 < -3\epsilon\right) \to 0, \quad \text{for all } \epsilon > 0 \quad \text{and } \tau > 0,$$

and thus

$$\widetilde{X}^k \wedge 0 \Rightarrow 0\eta \quad \text{in } D \quad \text{as } n \to \infty. \tag{41}$$

It is easy to check that $\Phi^k - \widetilde{X}^k \wedge 0 \in D_0$ and that

$$\left(\widetilde{X}^k \vee 0, k^{1/4}\int_0^\cdot \widetilde{X}^k(s) \wedge 0 ds\right) = h(\Phi^k - \widetilde{X}^k \wedge 0).$$

Now, due to (41)

$$\Phi^k - \widetilde{X}^k \wedge 0 \Rightarrow \Phi \quad \text{in } D \quad \text{as } n \to \infty,$$

and so

$$(\widetilde{X}^k \vee 0, \widetilde{X}^k - \Phi^k) \Rightarrow h(\Phi) \quad \text{in } D^2 \quad \text{as } n \to \infty.$$

Thus

$$(\Phi^k, \widetilde{X}^k \vee 0, \widetilde{X}^k - \Phi^k) \Rightarrow (\Phi, h(\Phi)) \quad \text{in } D^3 \quad \text{as } n \to \infty.$$

Writing $\widetilde{X}^k = \widetilde{X}^k \wedge 0 + \widetilde{X}^k \vee 0$ and employing (41) gives the stated limit. □

Proof of Lemma 2. First, it follows from the standard theory of ordinary differential equation that (9) has a unique solution. (It is easy to check that x_F in (10) and (11) satisfies (9) when $\beta < 0$ and $\beta = 0$, respectively.) Then, $(x_F, 0\eta)$ trivially satisfies (40), and it remains to show that it is the unique element in C^2 to have this property.

To this end, let $(y_1, \psi_1) \in C^2$ be a solution to (40). The fact that $\psi_1 \geq 0$ implies that

$$\begin{aligned} y_1(t) - x_F(t) &= -\frac{\theta^2}{2}\int_0^t (y_1^2(s) - x_F^2(s))ds + \psi_1(t) \\ &\geq -\frac{\theta^2}{2}\int_0^t (y_1(s) + x_F(s))(y_1(s) - x_F(s))^+ ds, \end{aligned}$$

and

$$(y_1(t)-x_F(t))^- \le (y_1(t)-x_F(t))^+ + \frac{\theta^2}{2}\int_0^t (y_1(s)+x_F(s))(y_1(s)-x_F(s))^+ ds, \quad \text{for all } t \ge 0.$$

By Gronwall's inequality, for each t, there is a $c_t \ge 0$ such that

$$(y_1(t)-x_F(t))^- \le c_t (y_1(t)-x_F(t))^+. \tag{42}$$

As $a^- > 0$ implies that $a^+ = 0$ for $a \in \mathbb{R}$, (42) implies that $(y_1 - x_F)^- = 0$, so that $y_1 \ge x_F$. Therefore, if either $\beta < 0$ or $x_0 > 0$, we have $y_1(t) \ge x_F(t) > 0$ for all $t > 0$. By (40), we immediately have $\psi_1 = 0\eta$, and thus y_1 solves (9) and must equal x_F.

Next, consider the case $\beta = 0$ and $x_0 = 0$. For t such that $y_1(t) > 0$, we have

$$dy_1(t)/dt = -\theta^2 y_1(t)^2/2 < 0, \quad \text{for all } t \ge 0 \text{ such that } y_1(t) > 0.$$

Together with $y_1(0) = 0$ and $y_1 \in C$, we have $y_1 = 0\eta$, so that $\psi_1 = -y_1 = 0\eta$.

Finally, it follows (10) (or (11)) and $\psi = 0\eta$ that the map $x_0 \mapsto (y, \psi)$ is continuous, completing the proof of Lemma 2. □

5.5. Proof of Theorems 3 and 5

We will need the following two supporting propositions, the proof of which are given in Section 6.5. We omit the proof of Theorem 3 since its assertion (i) follows immediately from Theorem 5, and its assertion (ii) follows from Proposition 10 and Theorem 6 by similar arguments used to show that Theorem 5 follows from Proposition 9 and Theorem 7.

Proposition 9. *For any $\beta \in \mathbb{R}$, $\{\widetilde{X}^n(\infty) : n \ge 1\}$ is tight in $\mathbb{R}$. Further, $\widetilde{X}^n(\infty) \wedge 0 \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$.*

Proposition 10. *If $\beta > 0$, then $\{\widehat{X}^n(\infty) : n \ge 1\}$ is tight in $\mathbb{R}$.*

Proof of Theorem 5. For each $n \ge 1$, we consider a stationary version of the processes X^n and L^n by taking

$$X^n(0) \overset{\mathrm{d}}{=} X^n(\infty) \quad \text{and} \quad L^n(0) \overset{\mathrm{d}}{=} L^n(\infty). \tag{43}$$

Due to Proposition 9, each subsequence of $\{\widetilde{X}^n(\infty) : n \ge 1\}$ has a further weakly converging subsequence; let $\{\widetilde{X}^k(\infty) : k \ge 1\}$ be such a converging subsequence, and let X_0 be its weak limit. Then by our choice of the initial distribution, it holds that $\widetilde{X}^k(0) \Rightarrow X_0$, and the stated convergence in Proposition 9 implies that $X_0 \ge 0$ w.p.1. Moreover, by Proposition 4 it holds that $\widetilde{L}^k(0) \Rightarrow 0$ as $k \to \infty$.

Now, conditional on the event $\{X_0 = x_0\}$, for $x_0 \in \mathbb{R}_+$, we have $\widetilde{X}^n \Rightarrow X_F^0$ in D as $n \to \infty$ by virtue of Theorem 7, where X_F^0 is the unique solution to the IVP (9) with initial condition $X_F^0(0) := X_0 = x_0$. Moreover, the stationarity of the prelimit $\{\widetilde{X}^n : n \ge 1\}$ implies that the limit X_F^0 is strictly stationary as well, so that $X_F^0(t) \overset{\mathrm{d}}{=} X_0$ for all $t \ge 0$.

To show that $X_0 = x^*$, w.p.1., recall that any solution to the ODE in (9) converges monotonically to x^* as $t \to \infty$. Hence, on the event $E_0 := \{X_0 \ne x^*\}$, it holds that

$$|X_F^0(t) - x^*| < |X_0 - x^*| \quad \text{for all } t > 0,$$

in contradiction to the stationarity of X_0, so that E_0 is a P-null event. Thus, the limit of all weakly converging subsequences of $\{\widetilde{X}^n(0) : n \ge 1\}$ is x^*, implying that $\widetilde{X}^n(0) \Rightarrow x^*$ as $n \to \infty$. The result follows from our choice of the initial conditions in (43). □

5.6. Proof of Proposition 1

To prove Proposition 1, we need the following comparison lemma, whose proof appears at the end of this section. Consider two $M/M_{pc}/n + M_{pc}$ systems, denoted by $\mathcal{P}_1$ and $\mathcal{P}_2$, both having service rate $\mu = 1$. Let the arrival rates λ_i in $\mathcal{P}_i$, $i = 1, 2$, satisfy $\lambda_1 \ge \lambda_2$, the abandonment rate θ_1 of $\mathcal{P}_1$ satisfy $0 \le \theta_1 < 1$, and the abandonment rate θ_2 of $\mathcal{P}_2$ satisfy

$$\theta_2 \ge \theta_1/(1-\theta_1) \ge \theta_1.$$

Note that we allow for $\theta_1 = 0$, in which system $\mathcal{P}_1$ reduces to an $M/M/n$ system. (In this case, we assume that all the customers that are initially in the system have exponentially distributed remaining service times, each with mean 1.)

Let X_i denote the number-in-system process in $\mathcal{P}_i$, $i = 1, 2$. If either $\theta_i > 0$ or $\lambda_i < n$, Theorem 1 implies that X_i has a stationary distribution, which we denote by $X_i(\infty)$.

Lemma 3. *If $\theta_1 > 0$ or $\lambda_1 < n$ so that $X_1(\infty)$ and $X_2(\infty)$ exist, then $X_1(\infty) \geq_{st.} X_2(\infty)$.*

Proof of Proposition 1. We write $\widehat{X}_C(t;\beta)$ to make explicit the dependence of the distribution of the process $\widehat{X}_C$ on the value of β, as well as of its stationary distribution (when $t := \infty$). Fix $\epsilon > 0$, and consider a sequence $\{\beta_\epsilon^n : n \geq 1\} \subset \mathbb{R}_+$ satisfying $\beta_\epsilon^n \geq \beta^n$ and $\beta_\epsilon^n \to \epsilon$ as $n \to \infty$. Let a sequence of $M/M_{pc}/n + M_{pc}$ systems be labeled by n, with arrival rate $\lambda_\epsilon^n := n - \beta_\epsilon^n$, service rate 1, and patience rate $\theta/(1-\theta)$. Denote by $X_\epsilon^n(\infty)$ the stationary distribution of the number-in-system process of the nth system. Lemma 3 and the existence of $X_\epsilon^n(\infty)$ imply that $X_\epsilon^n(\infty) \leq_{st.} X^n(\infty)$, so that, for any $M > 0$,

$$P(\widehat{X}^n(\infty) > M) \geq P(\widehat{X}_\epsilon^n(\infty) > M).$$

On the other hand, Theorem 3 gives

$$P(\widehat{X}_\epsilon^n(\infty) > M) \to P(\widehat{X}_C(\infty;\epsilon) > M) \quad \text{as } n \to \infty,$$

so that

$$\liminf_{n\to\infty} P(\widehat{X}^n(\infty) > M) \geq P(\widehat{X}_C(\infty;\epsilon) > M), \quad \text{for all } M > 0.$$

Finally, (5)–(7) give

$$\lim_{\epsilon\to 0^+} P(\widehat{X}_C(\infty;\epsilon) > M) = 1.$$

Therefore, $P(\widehat{X}^n(\infty) > M) \to 1$ as $n \to \infty$ for any $M > 0$, implying the result. □

It remains to prove Lemma 3.

Proof of Lemma 3. We assume that the arrival process to $\mathcal{P}_1$ is the superposition of two independent Poisson streams, with stream 1 having rate λ_2 and stream 2 having rate $\lambda_1 - \lambda_2$. We consider a coupling of $\mathcal{P}_1$ and $\mathcal{P}_2$ such that (i) both systems start empty; (ii) stream 1 arrivals to $\mathcal{P}_1$ and all arrivals of $\mathcal{P}_2$ follows the same Poisson process; and (iii) any stream 1 arrival to $\mathcal{P}_1$ and the corresponding arrival of $\mathcal{P}_2$ have the same service time. Using this coupling, we will show that the sojourn time of every stream 1 arrival to $\mathcal{P}_1$ is at least as long as that of the same arrival in $\mathcal{P}_2$. We label the stream 1 arrivals to $\mathcal{P}_1$, that also constitute the arrivals to $\mathcal{P}_2$, by $1, 2, \cdots$, and denote by S_i the service time of customer i. We denote the coupled number-in-system processes in $\mathcal{P}_1$ and $\mathcal{P}_2$ by $\check{X}_1$ and $\check{X}_2$, respectively.

The proof proceeds by induction. First, customer 1 in $\mathcal{P}_2$ enters service immediately upon arrival, so that the customer's sojourn time is S_1. The same customer in $\mathcal{P}_1$ may (i) enter service immediately, and experience the same sojourn time S_1; (ii) enter service after waiting in queue, so that the customer's sojourn time is greater than S_1; or (iii) abandon the system, after waiting for $\theta_1^{-1}S_1 > S_1$ units of time. In all three scenarios, the sojourn time of customer 1 in $\mathcal{P}_1$ is at least as large as in $\mathcal{P}_2$.

Take the induction hypothesis that the first j customers have equal or shorter sojourn times in $\mathcal{P}_2$ than in $\mathcal{P}_1$. There are three cases to consider in order to show that the same is true for the $(j+1)$ st customer.

Case* 1: *Customer $j+1$ abandons $\mathcal{P}_1$. In this case, that customer's sojourn time in $\mathcal{P}_1$ is equal to $\theta_1^{-1}S_{j+1}$. On the other hand, the sojourn time of customer $j+1$ in $\mathcal{P}_2$ is bounded from above by the service requirement S_{j+1} plus the patience time $\theta_2^{-1}S_{j+1}$. Because $(1+\theta_2^{-1})S_{j+1} \leq \theta_1^{-1}S_{j+1}$, the ordering of the sojourn times for the first j customers in the two systems remains to hold for the $(j+1)$ st customer.

Case* 2: *Customer $j+1$ is served in $\mathcal{P}_1$ ***but abandons*** $\mathcal{P}_2$. For $1 \leq k \leq j+1$, denote by D_k^1 and D_k^2 the time when the kth customer leaves system $\mathcal{P}_1$ and system $\mathcal{P}_2$, respectively. By the induction hypothesis, we have $D_k^1 \geq D_k^2$ for $k = 1, 2, \cdots, j$. Denote by F_{j+1}^1 the time when customer $j+1$ enters service in $\mathcal{P}_1$. Clearly, the first j customers are not in queue at this time, namely, each of them is either in service or has left $\mathcal{P}_1$ (either via abandonment or

service completion). In particular, if customer $k \le j$ is still in system $\mathcal{P}_1$, this customer must be in service. For a system with n servers, we then have

$$\sum_{i=1}^{j+1} 1\{D_i^1 \le F_{j+1}^1\} \le n, \quad \text{and } D_{j+1}^1 = F_{j+1}^1 + S_{j+1} > F_{j+1}^1,$$

so that

$$\sum_{i=1}^{j} 1\{D_i^1 \le F_{j+1}^1\} \le n-1.$$

Using $D_k^1 \ge D_k^2$ for $k = 1, 2, \cdots, j$, we have

$$\sum_{i=1}^{j} 1\{D_i^2 \le F_{j+1}^1\} \le \sum_{i=1}^{j} 1\{D_i^1 \le F_{j+1}^1\} \le n-1,$$

so that there are no more than $n-1$ of the first j stream 1 customers in $\mathcal{P}_2$ at time F_{j+1}^1. As system $\mathcal{P}_2$ has n servers and customer $j+1$ abandons system $\mathcal{P}_2$, customer $j+1$ must have abandoned system $\mathcal{P}_2$ by time F_{j+1}^1. Therefore, customer $j+1$ has equal or shorter sojourn time in system $\mathcal{P}_2$ than system $\mathcal{P}_1$.

Case 3: Customer $j+1$ is served in both systems $\mathcal{P}_1$ ***and*** $\mathcal{P}_2$. As in Case 2, there are no more than $n-1$ customers of label $1, 2, \cdots, j$ present in system $\mathcal{P}_2$ at time F_{j+1}^1. In this case, customer $j+1$ is served in system $\mathcal{P}_2$, so this customer must have entered service by time F_{j+1}^1, implying that the customer's delay in queue in $\mathcal{P}_2$ is no longer than the customer's delay in queue in $\mathcal{P}_1$. Because the service time of this customer is the same in both coupled systems, the ordering of the customer's sojourn times in both systems remains as in the previous two cases.

In either of these three cases, the ordering of the sojourn times of the stream 1 customers imply that $\check{X}_1(t) \ge \check{X}_2(t)$ w.p.1, so that $X_1(t) \ge_{st.} X_2(t)$, for all $t \ge 0$. Because $X_i(t) \Rightarrow X_i(\infty)$ as $t \to \infty$, for $i = 1, 2$, by Theorem 1 (independently of the initial condition), the result follows from the fact that stochastic order is maintained under weak convergence (Kamae et al. [8, proposition 3]). □

5.7. Proof of Proposition 4

Let w_v^n be the offered waiting-time process in the nth system, namely, $w_v^n(t)$ is the time that an infinite-patient customer (that does not abandon) would have to wait if the customer arrives at time t. Similar to the proofs of Theorem 1 and Proposition 3, $(X^n, Q^n, Z^n, L^n, w_v^n)$ has a unique joint stationary distribution. Let $Z^n(\infty)$ and $w_v^n(\infty)$ follow the marginal stationary distribution of Z^n and w_v^n, respectively. Notice that $Z^n(\infty) = X^n(\infty) \wedge n$ and let

$$\widehat{Z}^n(\infty) := n^{-1/2}(\widehat{Z}^n(\infty) - n) = \widehat{X}^n(\infty) \wedge 0.$$

Consider a generic customer with service requirement S and patience time $T = \theta^{-1}S$ arriving at the system in steady state. The offered waiting time of such a customer is distributed like $w_v^n(\infty)$ due to PASTA (Possion arrivals see time average), and is independent of S and T.

To prove Proposition 4, we need the following lemma, the proof of which appears in Section 6.5.

Lemma 4. *For any* $\beta \in \mathbb{R}$,

$$E[Z^n(\infty)]/\lambda^n + \theta^2 E[(w_v^n(\infty) \wedge T)^2]/2 = 1, \quad \textit{for all } n \ge 1, \quad \textit{and } \limsup_{n\to\infty} E[\widehat{Z}^n(\infty)] > -\infty. \tag{44}$$

Proof of Proposition 4. We will show that

$$E[L^n(\infty)] \le \lambda^n(1+\theta^2)E[(w_v^n(\infty) \wedge T)^2]/2, \quad \text{for } \beta \in \mathbb{R}. \tag{45}$$

and give separate estimation of the right-hand side to prove assertions (a) and (b).

To prove (45), we consider the generalization of Little's law, known as $H = \lambda G$; for example, see Wolff [22, chapter 5]. Assume that the system is initialized in steady state, and let E_j^n, v_j^n, and T_j^n be, respectively, the arrival time, offered wait, and the patience of the jth arrival. Also let

$$g_j^n(t) := (t - E_j^n)1\{t \in [E_j^n, E_j^n + (v_j^n \wedge T_j^n)]\}$$
$$+ (\theta v_j^n - (t - E_j^n - v_j^n))1\{t \in [E_j^n + v_j^n, E_j^n + (1+\theta)v_j^n], v_j^n \le T_j^n\},$$

We claim that

$$L^n(t) = \sum_{j=0}^{\infty} g_j^n(t), \text{ for all } t \geq 0. \tag{46}$$

To see this, recall that $L^n(t)$ is the sum of the elapsed waiting time for all customers that are in the queue, plus the remaining phase 1 service time for all customers in service. Now, customer j is in the queue at time t if j is an element of the set $\{j : E_j^n \leq t \leq E_j^n + (v_j^n \wedge T_j^n)\}$, and the elapsed waiting time of that customer is $t - E_j^n$. On the other hand, customer j is in phase 1 of service if j is an element of the set $\{j : T_j^n \geq v_j^n, E_j^n + v_j^n \leq t \leq E_j^n + (1+\theta)v_j^n\}$, and the remaining phase 1 service time for that customer is $\theta v_j^n - (t - E_j^n - v_j^n)$. Hence, we obtain (46).

Let

$$G_j^n := \int_0^{\infty} g_j^n(t)dt = (v_j^n \wedge T_j^n)^2/2 + 1\{T_j^n \geq v_j^n\}(\theta v_j^n)^2/2.$$

Because the system is considered to be in steady state, G_j^n is, for each $j \geq 1$, distributed like

$$G^n := (w_v^n(\infty) \wedge T)^2/2 + 1\{w_v^n(\infty) \leq T\}(\theta w_v^n(\infty))^2/2,$$

where $w_v^n(\infty)$ is the stationary offered wait defined in Section 6.5, and T is an exponentially distributed random variable with rate θ that is independent of $w_v^n(\infty)$.

It follows from the following inequality,

$$G^n = (w_v^n(\infty) \wedge T)^2/2 + 1\{w_v^n(\infty) \leq T\}(\theta w_v^n(\infty))^2/2 \leq \frac{1+\theta^2}{2}(w_v^n(\infty) \wedge T)^2, \tag{47}$$

and the trivial inequality $w_v^n(\infty) \wedge T \leq T$, that $E[G^n] < \infty$. It is also easy to check that (198) in chapter 5 of Wolff [22] holds, so that, by theorem 5 in this reference,

$$E[L^n(\infty)] = \lambda^n E[G^n],$$

which, together with the inequality in (47), gives (45).

To prove assertion (a), it is sufficient to prove that, if $\beta > 0$, then

$$E[(w_v^n(\infty) \wedge T)^2] = O(n^{-1}). \tag{48}$$

Consider a sequence of $M/M/n$ (Erlang-C) systems, each with service rate 1, and with arrival rate λ^n to the nth system. Notice that the $M/M/n$ system can be regarded as an $M/M_{pc}/n + M_{pc}$ system with no abandonment, so that we can apply the coupling in Lemma 3 between the two $M/M_{pc}/n + M_{pc}$ systems (one with abandonment rate that is equal to 0, and the other with rate θ).

Let the two coupled systems be initially empty and consider a customer that arrives at both systems. Inspecting the three cases in the proof of Lemma 3, we claim that the customer experiences longer delay in the Erlang-C system. First, Case 1 is irrelevant, because there is no abandonment in the Erlang-C system. The proof of Case 2 in Lemma 3 shows that the patience of the customer in the $M/M_{pc}/n + M_{pc}$ is shorter than the waiting time of that customer in the Erlang-C system. In particular, the delay in queue of the customer is shorter in the $M/M_{pc}/n + M_{pc}$ system than in the Erlang-C system. Finally, the proof of Case 3 in the proof of Lemma 3 shows again that the waiting time of the customer in $M/M_{pc} + n/M_{pc}$ system is shorter than in the Erlang-C system. Therefore, the waiting time of any customer is smaller in the $M/M_{pc}/n + M_{pc}$ system than in the coupled Erlang-C system. As $\beta > 0$ implies that $\beta^n > 0$ for sufficiently large n, there exists $N_0 \in \mathbb{Z}_+$ such that the Erlang-C system is stable for all $n > N_0$. In particular, for $n \geq N_0$, the stationary waiting time of the $M/M_{pc}/n + M_{pc}$ system is stochastically dominated from above by the stationary waiting time of the Erlang-C system.

Let $w_U^n(\infty)$ denote the stationary waiting time in the Erlang-C system, and note that the stationary waiting time of a generic customer in the $M/M_{pc}/n + M_{pc}$ is distributed like $w_v^n(\infty) \wedge T$. Then, the stochastic ordering $w_v^n(\infty) \wedge T \leq_{st.} w_U^n(\infty)$ just argued implies that

$$E[(w_v^n(\infty) \wedge T)^2] \leq E[(w_U^n(\infty))^2]. \tag{49}$$

Now, the waiting time of an arriving customer to the Erlang-C system that finds $q-1$ customers in queue, $q \geq 1$, is distributed like the sum of q independent exponential variables with mean n^{-1}. Letting $\{\gamma_i^n\}$ be a sequence of i.i.d. exponential random variables with mean n^{-1} and $Q_U^n(\infty)$ be the stationary distribution of the queue length

process in the Erlang-C system, it holds that $w_U^n(\infty) \stackrel{\mathrm{d}}{=} \sum_{i=1}^{Q_U^n(\infty)} \gamma_i^n$, due to PASTA, so that

$$E[(w_U^n(\infty))^2] = E\left[\left(\sum_{i=1}^{Q_U^n(\infty)} \gamma_i^n\right)^2\right] = n^{-2}E[(Q_U^n(\infty))^2] + n^{-2}E[Q_U^n(\infty)].$$

Because $\{Q_U^n(\infty) : n \geq 1\}$ is a sequence of stationary queues of $M/M/n$ systems staffed according to (1), we can apply the (explicit) limits for the first and second moments of the diffusion-scaled process in Halfin and Whitt [7, corollary 1], to conclude that $E[(w_U^n(\infty))^2] = O(n^{-1})$. Hence, (48), (45), and thus assertion (a), follow from (49). Finally, by Lemma 4,

$$E[(w_v^n(\infty) \wedge T)^2] = -2\theta^{-2}(\lambda^n)^{-1}n^{1/2}E[\widehat{Z}^n(\infty) + \beta^n] = O(n^{-1/2}),$$

(45), and thus assertion (b), immediately follow. □

6. Proofs of Supporting Results for Process Limits

In this section we prove Propositions 5–10 and Lemma 4. The proofs of Proposition 6–8 appear in Sections 6.1–6.3, respectively. A few supporting lemmas that are used in the proofs of the propositions are given in Appendix A. The proof of Proposition 5 is given in Section 6.4, as it requires arguments that are developed in Sections 6.1–6.3. The proofs of Propositions 9 and 10 and Lemma 4 appear in Section 6.5.

6.1. Proof of Proposition 6

We refer to the customers that are in the system at time 0 as the *initial customers*.

Proof of Assertion (a). For $i = 1, 2, \cdots, Z^n(0)$ and $j = 1, 2, \cdots, Q^n(0)$, let $g_i^n(t)$ be the elapsed phase 1 service time of the ith initial customer in service, and $h_j^n(t)$ be the elapsed phase 1 service time of the jth initial customer in the queue, at time t. Then

$$\int_0^t Z_0^n(s)ds = \sum_{i=1}^{Z^n(0)} g_i^n(t) + \sum_{i=1}^{Q^n(0)-Q_0^n(t)} h_i^n(t).$$

Notice that $g_i^n(t) \leq r_i^n(0)$, for any $i = 1, 2, \cdots, Z^n(0)$ and $t \geq 0$ and that

$$h_i^n(t) \leq \theta\ell_i^n(0) + \theta(T_0^n \wedge t)$$

for an initial customer i who has left the queue by time t. Hence,

$$\begin{aligned}\int_0^t Z_0^n(s)ds &\leq \sum_{i=1}^{Z^n(0)} r_i^n(0) + \theta\sum_{j=1}^{Q^n(0)} \ell_i^n(0) + \theta(T_0^n \wedge t)(Q^n(0) - Q_0^n(t)) \\ &\leq (1+\theta)L^n(0) + \theta(T_0^n \wedge t)(Q^n(0) - Q_0^n(t)), \text{ for all } t \geq 0. \end{aligned} \tag{50}$$

To bound $\int_0^t Q_0^n(s)ds$, notice that each initial customer in the queue waits for at most $T_0^n \wedge t$ during $[0, t]$, so that

$$\int_0^t Q_0^n(t)dt \leq (T_0^n \wedge t)Q^n(0), \text{ for all } t \geq 0. \tag{51}$$

Now, since $\widetilde{Z}_0^n$ and $\widetilde{Q}_0^n$ are nonnegative processes, (50) and (51) give

$$\begin{aligned} 0 \leq n^{1/4}\int_0^\infty \widetilde{Z}_0^n(s)ds &\leq (1+\theta)\widetilde{L}^n(0) + \theta T_0^n\widetilde{Q}^n(0), \\ 0 \leq n^{1/4}\int_0^\infty \widetilde{Q}_0^n(s)ds &\leq T_0^n\widetilde{Q}^n(0). \end{aligned}$$

Due to (Ib), it suffices to prove that $T_0^n \Rightarrow 0$ in $\mathbb{R}$, as $n \to \infty$; in particular, we need only consider the event $\{T_0^n > 0 \text{ for all } n \text{ large enough}\}$. Let

$$T_1^n := 4n^{-1}(1+\theta)(L^n(0) + 1),$$

and note that $T_1^n \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$. For each $n \geq 1$, define the event

$$\Upsilon^n := \{T_1^n < T_0^n \text{ and } Z_2^n(t_0) < n/2 \text{ for some } t_0 \in [T_1^n, T_0^n]\}.$$

We will show that $P(\Upsilon^n) \to 0$ as $n \to \infty$. To this end, note that, because $Z_1^n(s) = Z_0^n(s)$ for $s \le T_0^n$ (because $Z_0^n(s)$ is the number of initial customers that are in phase 1 at time s) and $Z_0^n + Q_0^n$ is nonincreasing, it holds on the event Υ that, for all $s \in [0, T_1^n]$,

$$Z_0^n(s) + Q_0^n(s) \ge Z_0^n(t_0) + Q_0^n(t_0) \ge Z_0^n(t_0) = Z_1^n(t_0) > n/2.$$

The last inequality and (50) give the bounds

$$\begin{aligned} \frac{nT_1^n}{2} < \int_0^{T_1^n} (Z_0^n(s) + Q_0^n(s))ds &\le (1+\theta)(L^n(0) + 2T_1^n Q^n(0)) \\ &< \frac{nT_1^n}{4} + 2(1+\theta)T_1^n Q^n(0), \end{aligned} \tag{52}$$

where the equality in (52) follows from the definition of T_1^n. Notice that (52) cannot hold when $Q^n(0) \le (1+\theta)^{-1} n/8$, and so, together with (Ib),

$$P(\Upsilon^n) \le P(Q^n(0) > (1+\theta)^{-1} n/8) \to 0 \quad \text{as } n \to \infty.$$

Thus, we need only consider sample paths on the complementary event Υ^c. On this event, either $T_0^n \le T_1^n$ or, if $T_0^n > T_1^n$, then there are at least $n/2$ customers in phase 2 service over the interval $[T_1^n, T_0^n]$, in which case the total service rate is at least $n/2$. In either case, for a sequence of i.i.d. exponentially distributed random variables $\{\mathcal{E}_k^n : k \ge 1\}$, each having rate $n/2$, it holds that

$$T_0^n \le_{st.} T_1^n + \sum_{k=1}^{Q^n(0)} \mathcal{E}_k^n,$$

where the latter sum is defined to be equal to 0 on $\{Q^n(0) = 0\}$. It follows that $T_0^n \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$, implying the result. □

We need the following lemma, whose proof appears in Appendix A, for the proofs of assertions (b) and (c).

Lemma 5. *If* (Ib) *holds, then* $\{\widetilde{Q}^n : n \ge 1\}$ *is stochastically bounded.*

Proof of Assertion (b). Consider the predictable quadratic variation of the (local) martingales $(\widetilde{M}_A^n, \widetilde{M}_S^n, \widetilde{M}_R^n)$. As $n \to \infty$, the following limits hold in D:

$$\begin{aligned} \langle \widetilde{M}_A^n \rangle(\cdot) &= n^{-3/2}(n^{1/4}\lambda^n)\eta(\cdot) \to 0\eta, \\ 0 \le \langle \widetilde{M}_S^n \rangle(\cdot) &= n^{-3/2} \int_0^{n^{1/4}\cdot} Z^n(s)ds \le n^{-1/4}\eta \to 0\eta, \\ 0 \le \langle \widetilde{M}_R^n \rangle(\cdot) &= n^{-3/4} \int_0^{\cdot} \widetilde{Q}^n(s)ds \Rightarrow 0\eta, \end{aligned}$$

where the last weak convergence follows from Lemma 5. Hence, $(\widetilde{M}_A^n, \widetilde{M}_S^n, \widetilde{M}_R^n) \Rightarrow (0\eta, 0\eta, 0\eta)$ in D^3, as $n \to \infty$ by, for example, theorem 8.1 in Pang et al. [12]. □

To prove assertions (c) and (d), we need the following lemma, whose proof appears in Appendix A.

Lemma 6. $\{\widetilde{w}^n : n \ge 1\}$ *is stochastically bounded in D.*

Proof of Assertil (c). We will show that $\{n^{1/2} \int_0^{\cdot} \widetilde{U}_1^n(s)ds : n \ge 1\}$ is stochastically bounded in D, from which the assertion follows immediately. To this end, note that similar arguments to those in (24) give

$$\begin{aligned} n^{1/2} \int_0^t \widetilde{U}_1^n(s)ds &= n^{1/2} \int_{\tilde{T}_0^n \wedge t}^t (n^{-1/2}\theta\widetilde{w}^n(s_1-)) \wedge (t - s_1) d\widetilde{M}_S^n(s_1) \\ &= \int_{\tilde{T}_0^n \wedge t}^t \theta\widetilde{w}^n(s_1-) d\widetilde{M}_S^n(s_1) \\ &\quad - n^{1/2} \int_{\tilde{T}_0^n \wedge t}^t \left(n^{-1/2}\theta\widetilde{w}^n(s_1-) - t + s_1\right)^+ d\widetilde{M}_S^n(s_1). \end{aligned} \tag{53}$$

Because $\widetilde{M}_S^n$ is an $\mathcal{F}_t^n$-martingale and $\widetilde{w}^n(s_1-)$ is a predictable process, $\int_0^{\cdot} \widetilde{w}^n(s_1-)d\widetilde{M}_S^n(s_1)$ is also an $\mathcal{F}_t^n$-martingale, with corresponding predictable quadratic variation process,

$$\left\langle \int_0^t \theta\widetilde{w}^n(s_1-)d\widetilde{M}_S^n(s_1) \right\rangle = \int_0^t (\theta\widetilde{w}^n(s_1-))^2 d\langle \widetilde{M}_S^n\rangle(s_1) \le (\|\theta\widetilde{w}^n\|_t)^2 \langle \widetilde{M}_S^n\rangle(t).$$

If follows from Lemma 6 and the proof of assertion (b) that

$$\left\langle \int_0^{\cdot} \theta\widetilde{w}^n(s_1-)d\widetilde{M}_S^n(s_1) \right\rangle \Rightarrow 0\eta \quad \text{in } D \text{ as } n \to \infty,$$

implying that

$$\int_0^{\cdot} \widetilde{w}^n(s_1-)d\widetilde{M}_S^n(s_1) \Rightarrow 0\eta \quad \text{in } D \text{ as } n \to \infty,$$

due to the martingale FCLT (e.g., theorem 8.1 in Pang et al. [12]). Hence, for any $t \ge 0$,

$$\epsilon^n(t) := \sup_{s\in[0,t]} \left| \int_{\widetilde{T}^n \wedge s}^s \widetilde{w}^n(s_1-)d\widetilde{M}_S^n(s_1) \right| \Rightarrow 0 \quad \text{in } \mathbb{R} \quad \text{as } n \to \infty. \tag{54}$$

To treat the second integral in the right-hand side of (53), we first observe that, for $s_1 \in [0,t]$,

$$\begin{aligned}(n^{-1/2}\theta\widetilde{w}^n(s_1-) - t + s_1)^+ &\le n^{-1/2}\theta\|\widetilde{w}^n\|_t 1\{s_1 + n^{-1/2}\theta\widetilde{w}^n(s_1) \ge t\} \\ &\le n^{-1/2}\theta\|\widetilde{w}^n\|_t 1\{s_1 \ge t - n^{-1/2}\theta\|\widetilde{w}^n\|_t\},\end{aligned}$$

so that

$$\left| n^{1/2} \int_{\widetilde{T}_0^n \wedge t}^t \left(n^{-1/2}\theta\widetilde{w}^n(s_1-) - t + s_1\right)^+ d\widetilde{M}_S^n(s_1) \right| \le \|\theta\widetilde{w}^n\|_t \int_{t-n^{-1/2}\theta\|\widetilde{w}^n\|_t}^t d|\widetilde{M}_S^n(s_1)|. \tag{55}$$

Furthermore (recalling that $\widetilde{Z}_2^n$ is centered about n),

$$\widetilde{M}_S^n(t) + n^{1/4}\int_0^t (n^{1/4} + \widetilde{Z}_2^n(s))ds = n^{-3/4}D^n(n^{1/4}t), \tag{56}$$

is a nondecreasing pure jump process and $-n^{1/4} \le \widetilde{Z}_2^n \le 0$, we have

$$\begin{aligned}\int_s^t d|\widetilde{M}_S^n(s_1)| &\le \int_s^t (n^{-3/4}dD^n(n^{1/4}t) + |n^{1/2} + n^{1/4}\widetilde{Z}_2^n(s_1)|ds_1) \\ &\le n^{-3/4}D^n(n^{1/4}t) - n^{-3/4}D^n(n^{1/4}s) + \sqrt{n}(t-s) \\ &= \widetilde{M}_S^n(t) - \widetilde{M}_S^n(s) + 2\sqrt{n}(t-s) + n^{1/4}\int_s^t \widetilde{Z}_2^n(s_1)ds_1 \\ &\le 2\|\widetilde{M}_S^n\|_t + 2\sqrt{n}(t-s),\end{aligned} \tag{57}$$

for all $0 \le s \le t$. Plugging (57) in (55) gives

$$\left| n^{1/2} \int_{\widetilde{T}^n \wedge t}^t \left(n^{-1/2}\theta\widetilde{w}^n(s_1-) - t + s_1\right)^+ d\widetilde{M}_S^n(s_1) \right| \le 2\|\theta\widetilde{w}^n\|_t(\|\widetilde{M}_S^n\|_t + \|\theta\widetilde{w}^n\|_t). \tag{58}$$

It follows from Lemma 6 and assertion (b) that the right-hand side of (58) is tight in $\mathbb{R}$. Next, plugging (54) and (58) in (53) gives

$$\left\| n^{1/2}\int_0^{\cdot} \widetilde{U}_1^n(s)ds \right\|_t \le \epsilon^n(t) + 2\|\theta\widetilde{w}^n\|_t(\|\widetilde{M}_S^n\|_t + \|\theta\widetilde{w}^n\|_t) \quad \text{for all } t \ge 0.$$

Therefore, $\left\{ n^{1/2}\int_0^{\cdot} \widetilde{U}_1^n(s)ds : n \ge 1 \right\}$ is tight in D for any $t \ge 0$. □

To prove assertion (d), we need the following lemma, the proof of which appears in Section A of the appendix. Recall $F^n(s,t)$ from (25).

Lemma 7. *For each $t>0$, $\{F^n(s,t): s\in[0,t]\}$ is a martingale, and $\{e^{\theta t}\sup_{s\in[0,t]}|F^n(s,t)| : t\ge 0\}$ is a submartingale, both with respect to their augumented natural filtration.*

Proof of Assertion (d). For $\tau>0$, $K>0$, and $F^n(s,t)$ in (25), let

$$\mathcal{M}^n(\tau,K;s,t) := \sup_{\substack{s,t\in[0,n^{1/4}\tau],\\ t-s\in[0,n^{-1/4}K]}} |F^n(s,t)|,$$

and observe that

$$n^{1/4}\|\widetilde{U}_2^n\|_\tau \le 2n^{-1/2}\mathcal{M}^n(\tau,\|\widetilde{w}^n\|_\tau;s,t).$$

Thus, the proof of the assertion will follow if we show that, for any $\epsilon>0$,

$$P(n^{-1/2}\mathcal{M}^n(\tau,\|\widetilde{w}^n\|_\tau;s,t)>\epsilon)\to 0 \quad \text{as } n\to\infty,$$

which is what we prove next.

Fix $\epsilon>0$. Because $\{\widetilde{w}^n : n\ge 1\}$ is stochastically bounded in D by Lemma 6, we can find a $K:=K(\epsilon)>2$, such that $P(\|\widetilde{w}^n\|_\tau>K)<\epsilon$ for any n. Notice that the value of $F^n(s,t)$ only depends on arrival times at $(s,t]$ and the patience times of those arrivals, implying that F^n is time-invariant in its two parameters, in the sense that $F^n(s,t)$ and $F^n(s+r,t+r)$ have the same law for any $r\ge 0$.

Let J^n be the smallest integer satisfying $J^n+1\ge n^{1/2}\tau/K$, and let

$$I_j^n := [n^{-1/4}K(j-1), n^{-1/4}K(j+1)]\cap[0,n^{1/4}\tau], \quad \text{for } j=1,2,\cdots,J^n.$$

Observe that, for $j=1,2,\cdots,J^n-1$, the length of each I_j^n is $2n^{-1/4}K$ and the length of $I_j^n\cap I_{j+1}^n$ is $n^{-1/4}K$. It holds that, for any $s,t\in[0,n^{1/4}\tau]$ for which $[s,t]\subseteq[0,n^{1/4}\tau]$ and $t-s<n^{-1/4}K$, the interval $[s,t]$ is contained in at least one of the intervals $\{I_j^n : j=1,2,\cdots,J^n\}$. Therefore,

$$\begin{aligned} P(n^{-1/2}\mathcal{M}^n(\tau,K;s,t)>\epsilon) &\le \sum_{j=1}^{J^n} P\left(n^{-1/2}\sup_{s,t\in I_j^n, s\le t}|F^n(s,t)|>\epsilon\right)\\ &\le J^n P\left(n^{-1/2}\sup_{s,t\in I_1^n, s\le t}|F^n(s,t)|>\epsilon\right), \end{aligned} \tag{59}$$

where the second inequality is due to the aforementioned time-invariance property of F^n, which implies that all the probabilities in the sum, except possibly the last one (which may be smaller than the rest), are equal.

Employing Lemma 7, we have

$$\begin{aligned} &P\left(n^{-1/2}\sup_{0\le s\le t\le 2n^{-1/4}K}|F^n(s,t)|>\epsilon\right)\\ &\le P\left(\sup_{t\in[0,\,2n^{-1/4}K]} e^{\theta t}\sup_{s\in[0,\,t]}|F^n(s,t)|>n^{1/2}\epsilon\right)\\ &\le \epsilon^{-6}n^{-3}E\left[\left(e^{2\theta n^{-1/4}K}\sup_{s\in[0,2n^{-1/4}K]}|F^n(s,2n^{-1/4}K)|\right)^6\right]\\ &\le \epsilon^{-6}e^{12\theta n^{-1/4}K}(6/5)^6 E\left[(n^{-1/2}F^n(2n^{-1/4}K,2n^{-1/4}K))^6\right], \end{aligned} \tag{60}$$

where the second and the last inequalities follow from Doob's L^p-maximal inequality (e.g., Revuz and Yor [17, theorem 1.7]) for $p=6$, for the (sub)martingales in Lemma 7.

It remains to compute $E[(n^{-1/2}F^n(2n^{-1/4}K,2n^{-1/4}K))^6]$ to bound the right-hand side of (59). Note that, conditional on $A^n(t)$, the vector of arrival times $(E_1^n,E_2^n,\ldots,E_{A^n(t)}^n)$ is distributed as the vector of ordered statistic of $A^n(t)$

uniform random variables on $[0,t]$. Therefore,

$$\int_0^t 1\{E^n_{A^n(s)} + T^n_{A^n(s)} \geq t\} dA^n(s)$$

is, conditional on $A^n(t)$, distributed like $\sum_{k=1}^{A^n(t)} B_k(t)$, where, for each $t \geq 0$, $\{B_k(t) : k \geq 1\}$ is a sequence of i.i.d. Bernoulli random variables, each distributed like $B_t := 1\{U + T \geq t\}$, where U is uniform on $[0,t]$, and T is exponentially distributed with rate θ that is independent of U. Thus, $E[B_t] = (\theta t)^{-1}(1 - e^{-\theta t})$, and

$$F^n(t,t) \stackrel{\mathrm{d}}{=} \sum_{k=1}^{A^n(t)} B_k(t) - E[B_t]\lambda^n t.$$

Let $\bar{b}_n$ denote $E[B_t]$ for $t = n^{-1/4}K$;

$$\bar{b}_n := E[B_{n^{-1/4}K}] = (n^{-1/4}\theta K)^{-1}(1 - e^{-n^{-1/4}\theta K}).$$

Let φ_n denote the moment generating function of $n^{-1/2}F^n(n^{-1/4}K, n^{-1/4}K)$. Using the identity $E[a^{A^n(t)}] = \exp((a-1)\lambda^n t)$ for each $a > 0$,

$$\begin{aligned}
\varphi_n(s) &:= E[\exp(sn^{-1/2}F^n(n^{-1/4}K, n^{-1/4}K))] \\
&= E[(E[e^{n^{-1/2}sB_1}])^{A^n(n^{-1/4}K)}]\exp(-s\lambda^n n^{-3/4}K\bar{b}_n) \\
&= E[(\bar{b}_n e^{n^{-1/2}s} + 1 - \bar{b}_n)^{A^n(n^{-1/4}K)}]\exp(-s\lambda^n n^{-3/4}K\bar{b}_n) \\
&= \exp(n^{-1/4}\lambda^n K\bar{b}_n(e^{n^{-1/2}s} - 1))\exp(-s\lambda^n n^{-3/4}K\bar{b}_n) \\
&= \exp(\gamma^n(e^{n^{-1/2}s} - 1 - n^{-1/2}s)), \quad \text{for all } s \geq 0,
\end{aligned}$$

where

$$\gamma^n := n^{-1/4}\lambda^n K\bar{b}^n = \theta^{-1}\lambda^n(1 - e^{-n^{-1/4}\theta K}) = O(n^{3/4}).$$

We claim that $\varphi_n^{(k)}(0) = O(n^{-k/8})$ for all $k \in \mathbb{Z}_+$, where $\varphi_n^{(k)}(s)$ denotes the kth derivative of φ_n taking value at s. We let

$$g_n(s) := \gamma^n(e^{n^{-1/2}s} - 1 - n^{-1/2}s), \quad \text{for } s \geq 0,$$

so that $\varphi_n = \exp(g_n)$, and note that $g_n(0) = g_n'(0) = 0$ and $g_n^{(k)}(0) = O(n^{3/4-k/2})$ for $k \geq 2$.

We prove this latter claim by induction. First, for $k = 1$, we have

$$\varphi_n'(0) = \varphi_n(0)g_n'(0) = 0 = O(n^{-1/8}).$$

Next, take the induction hypothesis that $\varphi_n^{(m)}(0) = O(n^{-m/8})$ for all $m \leq k$, and consider the $(k+1)$ st derivative:

$$\begin{aligned}
\varphi_n^{(k+1)}(0) &= (\varphi_n g_n')^{(k)}(0) \\
&= \sum_{j=0}^{k}\binom{k}{j}\varphi_n^{(k-j)}(0)g_n^{(j+1)}(0) \\
&= \sum_{j=1}^{k}\binom{k}{j}\varphi_n^{(k-j)}(0)n^{1/4-j/2} \\
&= \sum_{j=1}^{k} O(n^{-1/8(k-j)})n^{1/4-j/2} \\
&= O(n^{-1/8(k-1)})n^{-1/4} \\
&= O(n^{-1/8(k+1)}).
\end{aligned}$$

This proves our claim that $\varphi_n^{(k)}(0) = O(n^{-k/8})$ for all $k \in \mathbb{Z}_+$. In particular, taking $k = 6$ gives

$$E[(n^{-1/2}F^n(n^{-1/4}K, n^{-1/4}K))^6] = O(n^{-3/4}).$$

Using this fact in (59), and then in the upper bound in (60), we obtain

$$P(n^{-1/2}\mathcal{M}^n(\tau, K; s, t) > \epsilon) = O(n^{-1/4}), \quad \text{for all } \epsilon > 0. \quad \square$$

6.2. Proofs of Proposition 7

To prove Proposition 7, we need the following lemma, the proof of which appears in Section A of the appendix, together with the proofs of the rest of the supporting lemmas of this section.

Lemma 8. *If* (Ib) *holds, then* $\{\widetilde{X}^n : n \geq 1\}$, $\{\widetilde{Z}^n : n \geq 1\}$, $\{\widetilde{Z}_1^n - \widetilde{Z}_0^n : n \geq 1\}$, *and* $\{\widetilde{Z}_2^n + \widetilde{Z}_0^n : n \geq 1\}$ *are stochastically bounded in* D.

Proof of Proposition 7. We first show that, as $n \to \infty$,

$$\widetilde{V}^n(\cdot) - \frac{1}{2}\int_0^{\cdot} \theta^2(\widetilde{w}^n)^2(s)ds + \int_0^{\cdot} \theta(\widetilde{Z}_1^n(s) - \widetilde{Z}_0^n(s))\widetilde{w}^n(s)ds \Rightarrow 0\eta \text{ in } D. \tag{61}$$

Using the definition of V^n in (29), we have

$$\begin{aligned}
\widetilde{V}^n(t) &= \int_{\tilde{T}_0^n \wedge t}^t [(\theta\widetilde{w}^n(s)) \wedge (n^{1/2}t - n^{1/2}s)](\widetilde{Z}_2^n(s) + n^{1/4})ds \\
&\quad - \int_0^t n^{-1/2}\lambda^n(1 - e^{-n^{-1/4}\theta\widetilde{w}^n(s)})ds \\
&= \int_0^t \theta\widetilde{w}^n(s)\widetilde{Z}_2^n(s)ds + \frac{1}{2}\int_0^t (\theta\widetilde{w}^n(s))^2 ds - \int_0^{\tilde{T}_0^n \wedge t} \theta\widetilde{w}^n(s)(\widetilde{Z}_2^n(s) + n^{1/4})ds \\
&\quad - \int_{\tilde{T}_0^n \wedge t}^t (n^{-1/2}\theta\widetilde{w}^n(s) - t + s)^+(n^{3/4} + n^{1/2}\widetilde{Z}_2^n(s))ds \\
&\quad + \int_0^t \left(n^{1/4}\theta\widetilde{w}^n(s) - \frac{1}{2}(\theta\widetilde{w}^n(s))^2 - n^{-1/2}\lambda^n(1 - e^{-n^{-1/4}\theta\widetilde{w}^n(s)})\right)ds.
\end{aligned} \tag{62}$$

Noting that $0 \leq \widetilde{Z}_2^n + n^{1/4} \leq n^{1/4}$ and that $n^{1/4}\widetilde{T}_0^n = T_0^n \to 0$ in $\mathbb{R}$ as $n \to \infty$. By Proposition 6(a), we have that, for all $t > 0$,

$$\left\| \int_0^{\tilde{T}_0^n \wedge \cdot} \theta\widetilde{w}^n(s)(\widetilde{Z}_2^n(s) + n^{1/4})ds \right\|_t \leq n^{1/4}\widetilde{T}_0^n \|\widetilde{w}^n\|_t \Rightarrow 0, \quad \text{as } n \to \infty. \tag{63}$$

Next, using the fact that

$$(n^{-1/2}\theta\widetilde{w}^n(s_1) - t + s_1)^+ \leq n^{-1/2}\theta\|\widetilde{w}^n\|_t 1\{s_1 \geq t - n^{-1/2}\theta\|\widetilde{w}^n\|_t\}, \text{ for } s_1 \in [0, t],$$

we have

$$\begin{aligned}
0 &\leq \int_{\tilde{T}_0^n \wedge t}^t \left(n^{3/4} + n^{1/2}\widetilde{Z}_2^n(s_1)\right)(n^{-1/2}\theta\widetilde{w}^n(s_1) - t + s_1)^+ ds_1 \\
&\leq \int_0^t n^{1/4}\theta\|\widetilde{w}^n\|_t 1\{s_1 \geq t - n^{-1/2}\theta\|\widetilde{w}^n\|_t\}ds_1 \\
&= n^{1/4}\theta\|\widetilde{w}^n\|_t(t - (t - n^{-1/2}\theta\|\widetilde{w}^n\|_t)^+) \\
&\leq n^{-1/4}\theta^2(\|\widetilde{w}^n\|_t)^2 \Rightarrow 0\eta \text{ in } D, \text{ as } n \to \infty,
\end{aligned} \tag{64}$$

where the equality follows from

$$\int_a^b 1\{s \geq c\}ds = b \vee c - a \vee c \quad \text{for all } a, b, c \in \mathbb{R}.$$

Define the functions

$$f_1(x) := \begin{cases} (e^{-x} - 1 + x)/x & \text{if } x \neq 0 \\ 0 & \text{if } x = 0, \end{cases}$$

$$f_2(x) := \begin{cases} \left(e^{-x} - 1 + x - \frac{1}{2}x^2\right)/x^2 & \text{if } x \neq 0 \\ 0 & \text{if } x = 0, \end{cases} \tag{65}$$

and note that both f_1 and f_2 are continuous at $\mathbb{R}$. It follows from Lemma 6 that $n^{-1/4}\widetilde{w}^n \Rightarrow 0\eta$ in D as $n \to \infty$, and so $f_i(n^{-1/4}\widetilde{w}^n) \Rightarrow 0\eta$ in D as $n \to \infty$, for $i = 1, 2$, by virtue of the continuous mapping theorem. Writing $\lambda^n = n - \beta^n\sqrt{n}$, we have

$$\begin{aligned} &n^{1/4}\theta\widetilde{w}^n - 1/2(\theta\widetilde{w}^n)^2 - n^{-1/2}\lambda^n(1 - e^{-n^{-1/4}\theta\widetilde{w}^n}) \\ &= (\theta\widetilde{w}^n)^2 f_2(\theta n^{-1/4}\widetilde{w}^n) - n^{-1/4}\beta^n\theta\widetilde{w}^n(1 + f_1(n^{-1/4}\theta\widetilde{w}^n)) \Rightarrow 0\eta \text{ in } D \text{ as } n \to \infty. \end{aligned} \tag{66}$$

Using the weak limits established in (63), (64), and (66) in (62), gives

$$\widetilde{V}^n(\cdot) - \frac{1}{2}\int_0^{\cdot} \theta^2(\widetilde{w}^n)^2(s)ds - \int_0^{\cdot} \theta\widetilde{Z}_2^n(s)\widetilde{w}^n(s)ds \Rightarrow 0\eta \quad \text{in } D \quad \text{as } n \to \infty. \tag{67}$$

Now, it follows from (21) and $\widetilde{Z}^n = \widetilde{Z}_1^n + \widetilde{Z}_2^n$ that $\widetilde{Z}^n\widetilde{w}^n = 0\eta$, so that

$$\widetilde{Z}_1^n\widetilde{w}^n = -\widetilde{Z}_2^n\widetilde{w}_2^n. \tag{68}$$

Finally, by Proposition 6(a) and Lemma 6,

$$\left\| \int_0^{\cdot} \widetilde{Z}_0^n(s)\widetilde{w}^n(s)ds \right\|_t \leq \|\widetilde{w}^n\|_t \int_0^t \widetilde{Z}_0^n(s)ds \Rightarrow 0\eta \quad \text{in } D.$$

This, together with (67) and (68), gives (61).

Next, for $x \in D$, $\tau > 0$, and $\delta > 0$, consider the modulus of continuity

$$v_\tau(x, \delta) := \sup_{t-s \leq \delta} \{|x(s) - x(t)| : 0 \leq s < t \leq \tau\}.$$

Given the assumed convergence of the sequence of initial conditions $\{\widetilde{Q}^n(0) : n \geq 1\}$ the statement of the proposition will follow from Billingsley [1, theorem 15.5] once we show that

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P(v_\tau(\widetilde{Q}^n, \delta) \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0. \tag{69}$$

To estimate $v_\tau(\widetilde{Q}^n, \delta)$, note that, due to Proposition 6 and (61), we can write (33) as follows:

$$\begin{aligned} \widetilde{X}^n(\cdot) = \widetilde{X}^n(0) - \beta^n\eta - n^{1/4}\int_0^{\cdot} \widetilde{Z}^n(s) + \frac{1}{2}\int_0^{\cdot} \theta^2(\widetilde{w}^n)^2(s)ds \\ - \int_0^{\cdot} \theta(\widetilde{Z}_1^n(s) - \widetilde{Z}_0^n(s))\widetilde{w}^n(s)ds + \varepsilon^n(t), \end{aligned} \tag{70}$$

for some $\varepsilon^n \in D$ satisfying $\varepsilon^n = o_P(1)$. Let

$$\xi^n(t) := -\beta^n t + \frac{1}{2}\int_0^t \theta^2(\widetilde{w}^n)^2(s)ds - \int_0^t \theta(\widetilde{Z}_1^n(s) - \widetilde{Z}_0^n(s))\widetilde{w}^n(s)ds, \quad t \geq 0.$$

Because $\widetilde{Z}^n = \widetilde{X}^n \wedge 0$, we have

$$\widetilde{X}^n(t) = \xi^n(t) - n^{1/4}\int_0^t \widetilde{X}^n(s) \wedge 0ds + \varepsilon^n(t), \quad t \geq 0. \tag{71}$$

Fix $0 \le s \le t \le \tau$. Conditional on the event $\mathcal{E}_+^n := \{\inf_{u\in[s,t)} \widetilde{Q}^n(u) > 0\}$, we have that $\widetilde{X}^n(u) = \widetilde{Q}^n(u) > 0$ for all $u \in [s,t)$, in which case (71) implies that

$$|\widetilde{Q}^n(t) - \widetilde{Q}^n(s)| \le |\xi^n(t) - \xi^n(s)| + \|\varepsilon^n\|_\tau.$$

Next consider the event $\mathcal{E}_0^n := \{\inf_{u\in[s,t)} \widetilde{Q}^n(u) = 0\}$. Take

$$s_0 := \inf\{u \in [s,t) : \widetilde{Q}^n(u) = 0\} \quad \text{and} \quad t_0 := \sup\{u \in [s,t) : \widetilde{Q}^n(u) = 0\},$$

and note that $\widetilde{Q}^n$ is a pure jump process, so that $s_0 < t_0$ w.p.1. Then $\widetilde{Q}^n(s_0) = \widetilde{Q}^n(t_0-) = 0$, and $\widetilde{X}^n(u) = \widetilde{Q}^n(u) > 0$ for all $u \in [s,s_0) \cup [t_0,t)$. Thus, on $\mathcal{E}_0^n$,

$$\begin{aligned}|\widetilde{Q}^n(t) - \widetilde{Q}^n(s)| &\le |\widetilde{Q}^n(s_0) - \widetilde{Q}^n(s)| + |\widetilde{Q}^n(t) - \widetilde{Q}^n(t_0-)| \\ &\le |\xi^n(s_0) - \xi^n(s)| + |\xi^n(t) - \xi^n(t_0)| + 2\|\varepsilon^n\|_\tau.\end{aligned}$$

Overall, we see that

$$|\widetilde{Q}^n(t) - \widetilde{Q}^n(s)| \le 2 \sup_{s_1,t_1\in[s,t]} |\xi^n(s_1) - \xi^n(t_1)| + 2\|\varepsilon^n\|_\tau,$$

and thus

$$v_\tau(\widetilde{Q}^n, \delta) \le 2v_\tau(\xi^n, \delta) + 2\|\varepsilon^n\|_\tau. \tag{72}$$

Now,

$$|\xi^n(t) - \xi^n(s)| \le (t-s)\left(-\beta^n + \frac{1}{2}\theta^2(\|\widetilde{w}^n\|_\tau)^2 + \theta\|\widetilde{Z}_1^n - \widetilde{Z}_0^n\|_\tau\|\widetilde{w}^n\|_\tau\right),$$

and so, Lemmas 6 and 8 imply that, for any $\epsilon > 0$, there is an $M := M(\epsilon) > 0$ for which

$$\limsup_{n\to\infty} P(|\xi^n(t) - \xi^n(s)| \ge M(t-s)) \le \epsilon.$$

Thus,

$$\limsup_{n\to\infty} P(v_\tau(\xi^n, \delta) \ge M\delta) \le \epsilon, \text{ for all } \delta \in [0,\tau), \tag{73}$$

implying that

$$\lim_{\delta\to 0} \limsup_{n\to\infty} P(v_\tau(\xi^n, \delta) \ge \epsilon') = 0, \quad \text{for all } \epsilon' > 0.$$

This, together with (72) and the fact that $\varepsilon^n = o_P(1)$, gives (69), proving the statement of the proposition. □

6.3. Proof of Proposition 8

We start by proving that

$$\widetilde{Q}^n - \widetilde{Q}_0^n - \widetilde{w}^n \Rightarrow 0\eta \quad \text{in } D. \tag{74}$$

To this end, consider the LOF-scaled version of (27),

$$\widetilde{Q}^n - \widetilde{Q}_0^n = \theta^{-1}n^{-3/4}\lambda^n(1 - e^{-\theta n^{-1/4}\widetilde{w}^n}) + \widetilde{U}_2^n. \tag{75}$$

and the (continuous) function f_1 in (65). It follows from the proof of Proposition 7 (the arguments below (66)) that $f_1(n^{-1/4}\theta\widetilde{w}^n) \Rightarrow 0$ in D as $n \to \infty$, so that

$$n^{-3/4}\lambda^n(1 - e^{-n^{-1/4}\theta\widetilde{w}^n}) = n^{-1}\lambda^n\theta\widetilde{w}^n\left(1 - f_1(n^{-1/4}\theta\widetilde{w}^n)\right) = n^{-1}\lambda^n\theta\widetilde{w}^n + o_P(1).$$

Using the latter equality, $\lambda^n/n \to 1$, and Proposition 6(c) in (75), gives (74).

We next prove that

$$\int_0^{\cdot} \left|\widetilde{Z}_1^n(s) - \widetilde{Z}_0^n(s) - \theta\widetilde{w}^n(s)\right| ds \Rightarrow 0\eta \text{ in } D, \text{ as } n \to \infty. \tag{76}$$

Consider the LOF-scaled version of (19):

$$\widetilde{Z}_1^n(t) - \widetilde{Z}_0^n(t) = n^{-3/4} \int_{\widetilde{T}_0^n \wedge t}^t 1\{n^{-1/2}\theta\widetilde{w}^n(s-) + s > t\} dD^n(n^{1/4}s), \quad t \geq 0. \tag{77}$$

Fix a constant $\tau > 0$ and let

$$\Delta^n := \sup_{\substack{t\in[0,\tau], s\in[\widetilde{T}_0^n\wedge t, t],\\ t-s \leq n^{-1/2}\theta\|\widetilde{w}^n\|_\tau}} |\widetilde{w}^n(s-) - \widetilde{w}^n(t)|.$$

Using (74), the fact that $\widetilde{Q}_0^n(t) = 0$ for all $t \geq \widetilde{T}_0^n$, and noting that the jumps of $\widetilde{Q}^n$ are of size $\pm n^{-1/4}$ w.p.1, so that $\sup_{s\in[0,\tau]}|\widetilde{Q}^n(s) - \widetilde{Q}^n(s-)| \to 0$ as $n \to \infty$ w.p.1,

$$\Delta^n = \sup_{\substack{t\in[0,\tau], s\in[\widetilde{T}_0^n\wedge t, t],\\ t-s \leq n^{-1/2}\theta\|\widetilde{w}^n\|_\tau}} |\widetilde{Q}^n(s-) - \widetilde{Q}^n(t)| + \delta^n \leq \sup_{\substack{0\leq s<t\leq\tau,\\ t-s\leq n^{-1/2}\theta\|\widetilde{w}^n\|_\tau}} |\widetilde{Q}^n(t) - \widetilde{Q}^n(s)| + \delta^n,$$

where $\delta^n \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$. Thus, Lemma 6 and the C-tightness of $\{\widetilde{Q}^n : n \geq 1\}$ in Proposition 7 imply that $\Delta^n \Rightarrow 0$ in $\mathbb{R}$, as $n \to \infty$. Then, for $s \in [\widetilde{T}_0^n \wedge t, t]$,

$$1\{n^{-1/2}\theta(\widetilde{w}^n(t) - \Delta^n) + s > t\} \leq 1\{n^{-1/2}\theta\widetilde{w}^n(s-) + s > t\} \leq 1\{n^{-1/2}\theta(\widetilde{w}^n(t) + \Delta^n) + s > t\}.$$

For $\widetilde{T}_\Delta^n := \widetilde{T}_0^n + n^{-1/2}\theta(\|\widetilde{w}^n\|_\tau + \Delta^n)$ and $t \in [0,\tau]$, let $\Upsilon_t^n := \{T_\Delta^n < t\}$, and note that $T_0^n \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$ by Proposition 6(a), $w^n = O_P(1)$ by Lemma 6, and $\Delta^n \Rightarrow 0$ as shown previously, imply together that, for all $t \in (0,\tau]$,

$$\widetilde{T}_\Delta^n \Rightarrow 0 \quad \text{in } \mathbb{R} \quad \text{as } n \to \infty, \quad \text{so that} \quad P(\Upsilon_t^n) \to 1 \quad \text{as } n \to \infty.$$

Now, on the event Υ_t,

$$t - n^{-1/2}\theta(\widetilde{w}^n(t) \pm \Delta^n) \geq \widetilde{T}_0^n = \widetilde{T}_0^n \wedge t,$$

and it follows from (77) and the equality

$$\int_b^a 1\{s > c\} dF(s) = F(a\vee c) - F(b\vee c),$$

that

$$\begin{aligned}\widetilde{Z}_1^n(t) - \widetilde{Z}_0^n(t) &\geq n^{-3/4} \int_{\widetilde{T}_0^n\wedge t}^t 1\{s > t - n^{-1/2}\theta(\widetilde{w}^n(t) - \Delta^n)\} dD^n(n^{1/4}t)\\ &= n^{-3/4}D^n(n^{1/4}t) - n^{-3/4}D^n(n^{1/4}t - n^{-1/4}\theta(\widetilde{w}^n(t) - \Delta^n)).\end{aligned} \tag{78}$$

Similarly,

$$\widetilde{Z}_1^n(t) - \widetilde{Z}_0^n(t) \leq n^{-3/4}D^n(n^{1/4}t) - n^{-3/4}D^n(n^{1/4}t - n^{-1/4}\theta(\widetilde{w}^n(t) + \Delta^n)). \tag{79}$$

For any $0 \leq s_1 \leq t_1 \leq \tau$, (56) gives

$$\begin{aligned}&n^{-3/4}|D^n(n^{1/4}t_1) - D^n(n^{1/4}s_1) - n^{5/4}(t_1 - s_1)|\\ &= \left|\int_{s_1}^{t_1} n^{1/4}\widetilde{Z}_2^n(s)ds + \widetilde{M}_S^n(t_1) - \widetilde{M}_S^n(s_1)\right|\\ &\leq n^{1/4}\int_{s_1}^{t_1} |\widetilde{Z}_2^n(s)|ds + 2\|\widetilde{M}_S^n\|_\tau\\ &\leq n^{1/4}\int_0^\tau \widetilde{Z}_0^n(s)ds + n^{1/4}\int_{s_1}^{t_1} |\widetilde{Z}_2^n(s) + \widetilde{Z}_0^n(s)|ds + 2\|\widetilde{M}_S^n\|_\tau\\ &\leq n^{1/4}\|\widetilde{Z}_2^n + \widetilde{Z}_0^n\|_\tau(t_1 - s_1) + n^{1/4}\int_0^\tau \widetilde{Z}_0^n(s)ds + 2\|\widetilde{M}_S^n\|_\tau.\end{aligned}$$

Plugging $t_1 = t$, and the values $t - n^{-1/2}\theta(\widetilde{w}^n(t) + \Delta^n)$, as well as $t - n^{-1/2}\theta(\widetilde{w}^n(t) - \Delta^n)$ instead of s_1, shows that, for all $t \in [0, \tau]$,

$$|n^{-3/4}D^n(n^{1/4}t) - n^{-3/4}D^n(n^{1/4}t - n^{-1/4}\theta(\widetilde{w}^n(t) \pm \Delta^n)) - \theta\widetilde{w}^n(t)|$$
$$\le \theta\Delta^n + n^{-1/4}\theta\|\widetilde{Z}_2^n + \widetilde{Z}_0^n\|_\tau(\|\widetilde{w}^n\|_\tau + \Delta^n) + n^{1/4}\theta\int_0^\tau \widetilde{Z}_0^n(s)ds + 2\|\widetilde{M}_S^n\|_\tau =: \delta_\tau^n. \tag{80}$$

It follows from assertions (a) and (b) of Proposition 6, Lemma 8, and the fact that $\Delta^n \Rightarrow 0$ in $\mathbb{R}$, that $\delta_\tau^n \Rightarrow 0$ in $\mathbb{R}$. Further, by (78) and (79),

$$|\widetilde{Z}_1^n(t) - \widetilde{Z}_0^n(t) - \theta\widetilde{w}^n(t)| \le \delta_\tau^n \quad \text{for all } t \in [0, \tau],$$

so that

$$\int_{T_\Delta^n \wedge \tau}^\tau |\widetilde{Z}_1^n(s) - \widetilde{Z}_0^n(s) - \theta\widetilde{w}^n(s)|ds \Rightarrow 0 \quad \text{in } \mathbb{R}. \tag{81}$$

Finally, notice that

$$\int_0^{T_\Delta^n} |\widetilde{Z}_1^n(s) - \widetilde{Z}_0^n(s) - \theta\widetilde{w}^n(s)|ds \le T_\Delta^n(\|\widetilde{Z}_1^n - \widetilde{Z}_0^n\|_{T_\Delta^n} + \|\theta\widetilde{w}^n\|_{T_\Delta^n}) \Rightarrow 0 \quad \text{in } \mathbb{R},$$

where the equality (order of magnitude) follows from the stochastic boundedness of $\{\widetilde{Z}_1^n - \widetilde{Z}_0^n : n \ge 1\}$ and $\{\widetilde{w}^n : n \ge 1\}$ in D, established in Lemmas 8 and 6, respectively. Together with (81), this shows that

$$\int_0^\tau |\widetilde{Z}_1^n(s) - \widetilde{Z}_0^n(s) - \theta\widetilde{w}^n(s)|ds \Rightarrow 0 \quad \text{in } \mathbb{R}, \quad \text{for all } \tau > 0.$$

The uniform convergence over compact intervals in (76) follows from to the monotonicity in τ of the integral; see Dai [3, lemma 4.1].

Now,

$$\left|\int_0^t \widetilde{w}^n(s)(\theta\widetilde{w}^n(s) - \widetilde{Z}_1^n(s) + \widetilde{Z}_0^n(s))ds\right| \le \|\widetilde{w}^n\|_t \int_0^t \left|\theta\widetilde{w}^n(s) + \widetilde{Z}_0^n(s) - \widetilde{Z}_1^n(s)\right|ds, \tag{82}$$

for all $t \ge 0$. It follows from (76) and the fact that $\widetilde{w}^n = O_P(1)$, that the right-hand side of (82) is stochastically bounded in $\mathbb{R}$ for each $t \ge 0$, and because it is also nondecreasing in t,

$$\int_0^\cdot \widetilde{w}^n(s)(\theta\widetilde{w}^n(s) - \widetilde{Z}_1^n(s) + \widetilde{Z}_0^n(s))ds = o_P(1),$$

so that

$$\int_0^\cdot (\widetilde{Z}_1^n(s) - \widetilde{Z}_0^n(s))\widetilde{w}^n(s)ds = \int_0^\cdot \theta(\widetilde{w}^n(s))^2 ds + o_P(1). \tag{83}$$

On the other hand, for all $t \ge 0$,

$$\left|\int_0^t ((\widetilde{w}^n(s))^2 - (\widetilde{Q}^n(s))^2)ds\right|$$
$$= \left|\int_0^t (\widetilde{w}^n(s) + \widetilde{Q}^n(s))(\widetilde{w}^n(s) - \widetilde{Q}^n(s))ds\right|$$
$$\le (\|\widetilde{w}^n\|_t + \|\widetilde{Q}^n\|_t)\int_0^t (|\widetilde{w}^n(s) + \widetilde{Q}_0^n(s) - \widetilde{Q}^n(s)| + \widetilde{Q}_0^n(s))ds. \tag{84}$$

By Proposition 6(a), (74), and the facts that $\widetilde{w}^n = O_P(1)$ and $\widetilde{Q}^n = O_P(1)$, the right-hand side of (84) weakly converges to 0 in $\mathbb{R}$ as $n \to \infty$, for any $t \ge 0$. Notice that the right-hand side of (84) is nondecreasing in t, we obtain

$$\int_0^\cdot ((\widetilde{w}^n(s))^2 - (\widetilde{Q}^n(s))^2)ds \Rightarrow 0\eta \quad \text{in } D \quad \text{as } n \to \infty,$$

so that

$$\int_0^{\cdot} (\widetilde{w}^n(s))^2 ds = \int_0^{\cdot} (\widetilde{Q}^n(s))^2 ds + o_P(1). \tag{85}$$

The statement of the proposition follows by employing (85) in (83), and then in (61). □

6.4. Proof of Proposition 5

We now prove Proposition 5, building on some of the previous arguments. Of course, Condition (Ia) is stronger than Condition (Ib), and we can therefore use Propositions 6–8 in the current proof.

Proof of Assertion (a). The inequalities in (50) and (51) give

$$\int_0^{\infty} \widehat{Z}_0^n(s) ds \le \widehat{L}^n(0) + T_0^n \widehat{Q}^n(0) \quad \text{and} \quad \int_0^{\infty} \widehat{Q}_0^n(s) ds \le T_0^n \widehat{Q}^n(0). \tag{86}$$

The weak limit $T_0^n \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$ in Proposition 6(a) implies the assertion. □

Proof of Assertion (b). Notice that

$$\widehat{U}_1^n(t) = n^{1/4} \widetilde{U}_1^n(n^{-1/4}t), \ \widehat{U}_2^n(t) = n^{1/4} \widetilde{U}_2^n(n^{-1/4}t) \text{ and } \widehat{V}^n(t) = n^{1/4} \widetilde{V}^n(n^{-1/4}t), \quad t \ge 0.$$

Proposition 6(d) implies that $\widehat{U}_2^n \Rightarrow 0\eta$ in D, and thus $\int_0^{\cdot} \widehat{U}_2^n(s) ds \Rightarrow 0\eta$ in D, as $n \to \infty$.

To prove

$$\int_0^{\cdot} \widehat{U}_1^n(s) ds = n^{1/2} \int_0^{n^{-1/4}\cdot} \widetilde{U}_1^n(s) ds \Rightarrow 0\eta \quad \text{in } D \quad \text{as } n \to \infty. \tag{87}$$

Using similar arguments as in the proof of Proposition 6(a), one can check that, under (Ia), $n^{1/4} T_0^n \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$. Inspecting the proof of Proposition 6(c) (see, in particular, (53), (54), and (58)), it is sufficient to prove that $\|\widetilde{w}^n\|_{n^{-1/4}\tau} \Rightarrow 0$ in $\mathbb{R}$ for all $\tau \ge 0$. Notice that $\widetilde{w}^n(\widetilde{T}_0^n) \le n^{1/4} T_0^n$ and $\widetilde{Q}_0^n(s) = 0$ for $s \ge \widetilde{T}_0^n$. Then, for $\tau \ge 0$,

$$\begin{aligned} \|\widetilde{w}^n\|_{n^{-1/4}\tau} &\le n^{1/4} T_0^n + \sup\{\widetilde{w}^n(s) : s \in [\widetilde{T}_0^n \wedge (n^{-1/4}\tau), n^{-1/4}\tau]\} \\ &\le n^{1/4} T_0^n + \|\widetilde{Q}^n\|_{n^{-1/4}\tau} + \sup\{|\widetilde{w}^n(s) - \widetilde{Q}^n(s) - \widetilde{Q}_0^n(s)| : s \in [\widetilde{T}_0^n, n^{-1/4}\tau]\} \\ &\le n^{1/4} T_0^n + \|\widetilde{Q}^n\|_{n^{-1/4}\tau} + \|\widetilde{w}^n - \widetilde{Q}^n - \widetilde{Q}_0^n\|_{n^{-1/4}\tau}. \end{aligned} \tag{88}$$

Now,

$$\begin{aligned} \|\widetilde{Q}^n\|_{n^{-1/4}\tau} &\le \|\widetilde{Q}^n(t) - \widetilde{Q}^n(0)\|_{n^{-1/4}\tau} + \|\widetilde{Q}^n(0)\|_{n^{-1/4}\tau} \\ &\le \sup_{s,t \in [0, n^{-1/4}\tau]} |\widetilde{Q}^n(t) - \widetilde{Q}^n(s)| + \|\widetilde{Q}^n(0)\|_{n^{-1/4}\tau} \Rightarrow 0 \quad \text{in } \mathbb{R} \quad \text{as } n \to \infty, \end{aligned}$$

where the convergence follows from Proposition 7 and (Ia). Further, $\|\widetilde{w}^n - \widetilde{Q}^n - \widetilde{Q}_0^n\|_{n^{-1/4}\tau} \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$ by (74). Because $n^{1/4} T_0^n \Rightarrow 0$ in $\mathbb{R}$, as was mentioned earlier, $\|\widetilde{w}^n\|_{n^{-1/4}\tau} \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$, for $\tau > 0$ by (88).

The proof that $\widehat{V}^n \Rightarrow 0\eta$ in D builds on arguments in the proof of Proposition 7, by replacing t in the proof (61) with $n^{-1/4}t$. Because $n^{1/2}\widetilde{T}_0^n = n^{1/4} T_0^n \Rightarrow 0$ and $\|\widetilde{w}^n\|_{n^{-1/4}\tau} \Rightarrow 0$ in $\mathbb{R}$ for all $\tau \ge 0$, the left-hand side of (63), (64), and (66), regarded as processes of t, are all $o_P(n^{-1/4})$. Using this in (62) gives that

$$\widetilde{V}^n(n^{-1/4}\cdot) - \frac{1}{2}\int_0^{n^{-1/4}\cdot} \theta^2 (\widetilde{w}^n)^2(s) ds + \int_0^{n^{-1/4}\cdot} \theta(\widetilde{Z}_1^n(s) - \widetilde{Z}_0^n(s)) \widetilde{w}^n(s) ds = o_P(n^{-1/4}).$$

The stochastic boundedness of $\{\widetilde{Z}_1^n - \widetilde{Z}_0^n : n \ge 1\}$ (Lemma 8), and the fact that $\|\widetilde{w}\|_{n^{-1/4}\tau} \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$, imply that

$$\widehat{V}^n = n^{1/4} \widetilde{V}^n(n^{-1/4}\cdot) \Rightarrow 0\eta, \quad \text{in } D \quad \text{as } n \to \infty. \quad \square$$

Proof of Assertion (c). By the Poisson FCLT (e.g., theorem 4.2 in Pang et al. [12]),

$$\left(\frac{A(nt)-nt}{\sqrt{n}}, \frac{S(nt)-nt}{\sqrt{n}}, \frac{S(nt)-nt}{\sqrt{n}}\right) \Rightarrow (B_1, B_2, 0\eta), \quad \text{in } D^3 \quad \text{as } n \to \infty, \tag{89}$$

for two independent standard Brownian motions (B_1, B_2). Notice that $\widehat{M}_A^n$, $\widehat{M}_S^n$, and $\widehat{M}_S^n$ are the compositions of the scaled compensated Poisson processes in (89) with the time changes

$$\Phi_A^n : t \longmapsto n^{-1}\lambda^n t, \quad \Phi_S^n : t \longmapsto n^{-1}\int_0^t Z_2^n(s)ds, \quad \text{and } \Phi_R^n : t \longmapsto n^{-1}\int_0^t Q^n(s)ds,$$

respectively. By (15), (86), and the stochastic boundedness of $\{\widetilde{Q}^n : n \geq 1\}$ and $\{\widetilde{Z}_2^n + \widetilde{Z}_0^n : n \geq 1\}$ in D, established in Lemmas 5 and 8, respectively,

$$n^{-1}\lambda^n t = t + o(1), \quad n^{-1}\int_0^{\cdot} Q^n(s)ds = \int_0^{n^{-1/4}\cdot} \widetilde{Q}(s)ds = o_P(1),$$

$$n^{-1}\int_0^{\cdot} Z_2^n(s)ds = \eta + \int_0^{n^{-1/4}\cdot} (\widetilde{Z}_2^n(s) + \widetilde{Z}_0^n(s))ds - n^{-1/2}\int_0^{\cdot} \widehat{Z}_0^n(s)ds = \eta + o_P(1),$$

implying that

$$(\Phi_A^n, \Phi_S^n, \Phi_R^n) \Rightarrow (\eta, \eta, 0\eta), \quad \text{in } D^3 \quad \text{as } n \to \infty,$$

and the joint convergence in assertion (c) follows from the continuity of the composition map, for example, theorem 13.2.1 in Whitt [21]. □

6.5. Proofs of Lemma 4 and Propositions 9 and 10

An essential step in the proofs in this section is the following stochastic-order lower bound for X^n. For $n \geq 1$, consider an $M/M/n+M$ (Erlang-A) system, having independent service and patience times, with arrival rate λ^n, service rate 1, and patience rate θ. Let X_A^n, Q_A^n, Z_A^n denote the queueing processes in this Erlang-A system, analogously to the corresponding processes X_n, Q^n, L^n, and Z^n in the $M/M_{pc}/n+M_{pc}$.

Lemma 9. $X_A^n(\infty) \leq_{st.} X^n(\infty)$.

Proof. We prove the lemma by coupling $M/M_{pc}/n+M_{pc}$ and the presented Erlang-A system, and showing that the inequality in the statement holds w.p.1 for the coupled systems. In particular, we give the $M/M_{pc}/n+M_{pc}$ system and the $M/M/n+M$ system the same arrival stream and initial condition. Let E_k^n denote by the arrival epoch of the kth customer to the nth system. Exploiting the PASTA (Poisson arrivals see time averages) property, and using induction, it is sufficient to prove that, if $X^n(E_k^n) \geq X_A^n(E_k^n)$, then $X^n(E_{k+1}^n) \geq X_A^n(E_{k+1}^n)$, for all $k \geq 1$, where the inequalities hold w.p.1 for the coupled systems, from which the stochastic ordering in the statement follows.

Hence, we initialize both systems with the same number of customers, so that $X^n(0) = X_A^n(0)$, and take the induction hypothesis that $X^n(E_k^n) \geq X_A^n(E_k^n)$. Consider the dynamics of the $M/M_{pc}/n+M_{pc}$ when all arrivals are "turned off" after the kth arrival, and let $(X', Q', Z_1', \{\ell'\}, \{r'\})$ denote the corresponding Markov process. Let X_A' be the corresponding pure-death process for the Erlang-A system with arrivals turned off after the kth arrival. Note that the death rate of this process at state $m \geq 1$ is

$$d_A(m) := \theta(m-n)\vee 0 + m \wedge n,$$

and that $X'(E_k^n + \cdot)$ is a pure jump process with $X'(E_k^n)$ jumps until it reaches state 0. For $j = 1, \ldots, X'(E_k^n)$, let N_j denote the jth jump time of $X'(E_k^n + \cdot)$, so that N_j is the jth customer that leaves the system after E_k. Due to the memoryless property of the exponential distribution, at $t \geq E_k^n$,

i. The number of customers in queue is $(X'(t) - n)\vee 0$, each having a remaining patience time that is exponentially distributed with rate θ, and is independent of everything else.

ii. The number of customers in phase 2 service is $X'(t) \wedge n - \sum_{i=1}^{Z_1'(t)} 1\{t \leq E_k^n + r_i'(E_k^n)\}$, with each of those customers having a remaining service time that is exponentially distributed with rate 1, independently of everything else.

Then $N_{j+1} - N_j$ is, conditional on $(X'(N_j), Z'_1(N_j))$, distributed as the interarrival time in a nonhomogeneous Poisson process with intensity function

$$d_j(t) := \theta(X'(N_j) - n) \vee 0 + X'(N_j) \wedge n - \sum_{i=1}^{Z'_1(N_j)} 1\{t \le N_j + r'_i(N_j)\}.$$

Clearly, $d_j(t) \le d_A(X'(N_j))$ for all $t \ge 0$ and $j = 1, 2, \cdots X'(E_k^n)$, implying that, for $j \le X'_A(E_k^n)$, the sojourn time of the process $X'_A(E_k^n + \cdot)$ in state j is dominated by the corresponding sojourn time of $X'(E_k^n + \cdot)$. Using the induction hypothesis $X'_A(E_k^n) \le X'(E_k^n)$, we conclude that $X'_A(E_k^n + t) \le X'(E_k^n + t)$ for all $t \ge 0$. Finally, because $E_{k+1}^n - E_k^n$ is independent of $(X'_A(E_k^n + \cdot), X'(E_k^n + \cdot))$, we have that $X'_A(E_{k+1}^n) \le X'(E_{k+1}^n)$, implying that $X_A(E_{k+1}^n) \le X(E_{k+1}^n)$ for the two coupled systems. □

Proof of Lemma 4. To prove the equality in (44), consider a generic customer arriving at the system in steady state, and let v^n be the offered waiting time of the customer, namely, the waiting time of the customer if the customer never abandons. By PASTA, $v^n \overset{\mathrm{d}}{=} w_v^n(\infty)$. When the customer has service time S and patience time $T = \theta^{-1}S$, the waiting time of the customer is $T \wedge v^n$. Note that T and v^n are independent, the customer enters service if and only if $T \ge v^n$, and the customer contributes $1\{T \ge v^n\}S$ of work to the workload. In particular, the Poisson arrivals contribute to the mean workload of the system

$$\lambda^n E[1\{T \ge v^n\}S] = \theta\lambda^n E[1\{T \ge v^n\}T].$$

On the other hand, each working server reduces the workload at a constant rate 1, so that the pool of servers reduces the workload by $Z^n(\infty)$ per unit time in steady state. Because the mean workload is constant in steady state, we have

$$E[Z^n(\infty)] = \theta\lambda^n E[1\{T \ge v^n\}T]$$

One can check that, for any $x \in \mathbb{R}_+$,

$$E[(x \wedge T)^2] = 2\theta^{-2}(1 - e^{-\theta x}(1 + \theta x)) \quad \text{and} \quad \theta E[1\{T \ge x\}T] = e^{-\theta x}(1 + \theta x),$$

so that

$$\theta^2 E[(x \wedge T)^2]/2 + \theta E[1\{T \ge x\}T] = 1. \tag{90}$$

Exploiting the independence of T and v^n, taking $x = v^n$ in (90), and using $E[T \wedge v^n] = E[T \wedge w_v^n(\infty)]$, give the equality in (44).

For the inequality in (44), observe that, by Lemma 9,

$$Z^n(\infty) = X^n(\infty) \wedge n \ge_{st.} X_A^n(\infty) \wedge n = Z_A^n(\infty),$$

where $Z_A^n(\infty)$ is the stationary distribution of the Erlang-A system. To estimate $E[Z_A^n(\infty)]$, let $P(Ab_A^n)$ denote the long-run fraction of customers that abandon the system, so that $E[Z_A^n(\infty)] = \lambda^n(1 - P(Ab_A^n))$. By Garnett et al. [6, theorem 4], $P(Ab_A^n) = O(n^{-1/2})$. Therefore, using $\lambda^n/n \to 1$ as $n \to \infty$,

$$1 - n^{-1}E[Z_A^n(\infty)] = O(n^{-1/2}),$$

implying the inequality in (44). □

Proof of Proposition 9. Let

$$\widetilde{Z}^n(\infty) := n^{-3/4}(Z^n(\infty) - n) = n^{-1/4}\widehat{Z}^n(\infty) \le 0.$$

By Lemma 4, $E[\widetilde{Z}^n(\infty)] \to 0$ as $n \to \infty$. By Markov's inequality,

$$P(-\widetilde{Z}^n(\infty) > \epsilon) \le -\epsilon^{-1}E[\widetilde{Z}^n(\infty)], \quad \text{for all } \epsilon > 0,$$

implying that $\widetilde{X}^n(\infty) \wedge 0 = \widetilde{Z}^n(\infty) \Rightarrow 0$ in $\mathbb{R}$.

To prove the tightness of $\{\widetilde{X}^n(\infty) : n \ge 1\}$, it remains to show that $\{\widetilde{Q}^n(\infty) : n \ge 1\}$ is tight, where

$$\widetilde{Q}^n(\infty) := n^{-3/4}Q^n(\infty) = \widetilde{X}^n(\infty) - \widetilde{Z}^n(\infty).$$

Again notice that the stationary waiting time of the arrivals is distributed as $w_v^n(\infty) \wedge T$. By Little's law,

$$E[Q^n(\infty)] = \lambda^n E[w_v^n(\infty) \wedge T] \le \lambda^n (E[(w_v^n(\infty) \wedge T)^2])^{1/2}.$$

By Lemma 4,

$$E[Q^n(\infty)] \le \theta^{-1} n^{1/4} (\lambda^n)^{1/2} (-\beta^n - E[\widehat{Z}^n(\infty)])^{1/2} = O(n^{3/4}).$$

By Markov's inequality, we have that, for any $M > 0$,

$$\limsup_{n\to\infty} P(n^{-3/4} Q^n(\infty) \ge M) \le M^{-1} \limsup_{n\to\infty} n^{-3/4} E[Q^n(\infty)] < \infty.$$

Therefore, $\{\widetilde{Q}^n(\infty) : n \in \mathbb{Z}_+\}$ is tight in $\mathbb{R}$, and so is $\{\widetilde{X}^n(\infty) : n \in \mathbb{Z}_+\}$. □

Proof of Proposition 10. First, $\beta^n \to \beta > 0$ as $n \to \infty$ implies that there exists N, such that $\beta^n > 0$ for $n \ge N$ so that $X_U^n(\infty)$ exists. As in the proof of Proposition 4, we consider a coupling of the $M/M_{pc}/n + M_{pc}$ system with an Erlang-C system having the same arrival process and service rate 1. In turn, the Erlang-C system can be considered to be an $M/M_{pc}/n + M_{pc}$ system with patience that is exponentially distributed with rate 0, so that the coupling in Lemma 3 can be applied. Denote by (X_U^n, Q_U^n, Z_U^n) the number-in-system process, the queue length process, and the number-in-service process in the nth Erlang-C system. For $n \ge N$, Lemma 3 implies that $X_U^n(\infty) \ge_{st.} X^n(\infty)$. In particular,

$$E[\widehat{X}^n(\infty)] \le E[\widehat{X}_U^n(\infty)] \to E[\widehat{X}_C(\infty)] < \infty \quad \text{as } n \to \infty,$$

where the convergence follows from Halfin and Whitt [7, theorem 1] and the last inequality follows from corollary 1 in this reference.

On the other hand, Lemma 4 gives

$$\liminf_{n\to\infty} E[\widehat{X}^n(\infty)] \ge \liminf_{t\to\infty} E[\widehat{Z}^n(\infty)] > -\infty.$$

Therefore,

$$\limsup_{n\to\infty} E[|\widehat{X}^n(\infty)|] < \infty,$$

implying the statement of the proposition. □

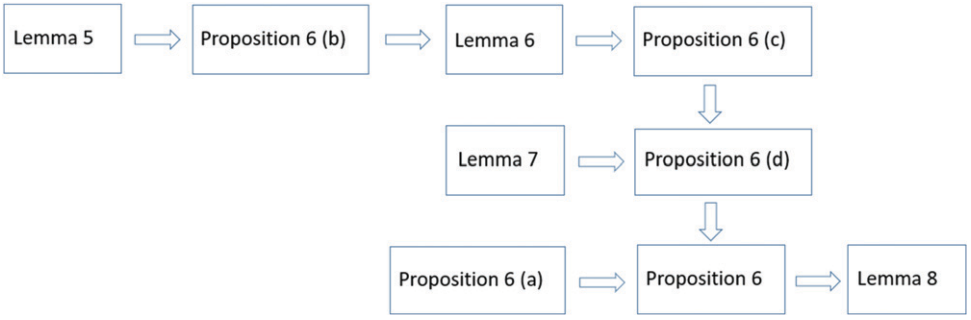
Acknowledgments

The work on this paper began while the first author was completing his PhD in the Industrial Engineering and Management Science Department at Northwestern University.

Appendix A. Remaining Proofs of Lemmas in Section 6

In this section we prove Lemmas 5–8. The flowchart in Figure A.1 depicts how the proofs in this section depend on each other, as well as on other results that were established in Section 6.

Figure A.1. (Color online) Flowchart detailing how the proofs in this section are related to each other and to proofs of other results.



Proof of Lemma 5. We again use a coupling of the $M/M_{pc}/n+M_{pc}$ system with another queueing system, which we denote by $\mathcal{U}^n$, using the same notation as in the proof of Proposition 10 for the corresponding process (X_U^n, Q_U^n, Z_U^n). We take system $\mathcal{U}^n$ is a degenerated $M/M_{pc}/n+M_{pc}$ system with arrival rate λ^n and service rate 1, in which customers have infinite patience. For the coupling, we initialize system $\mathcal{U}^n$ and the $M/M_{pc}/n+M_{pc}$ system as follows: first, we take $X_U^n(0)=X^n(0)$; second, any initial customer in queue has the same service time in both systems; third, any initial customer in service system $\mathcal{U}$ has the same remaining service time in the $M/M_{pc}/n+M_{pc}$ system. Note that system $\mathcal{U}^n$ is not an Erlang-C system per se because some of the initial customers in service may be in their phase 1. (There is no phase-1 service for any of the customers that arrive after time 0 in this system.

Using the same arguments as in the proof of Lemma 3, we can construct a coupling between X^n and X_U^n such that $X^n(t)\le X_U^n(t)$, and thus $Q^n(t)\le Q_U^n(t)$, w.p.1 for all $t\ge 0$. Let $\widetilde{Q}_U^n(t):=n^{-3/4}Q^n(n^{1/4}t)$. It is sufficient to prove that $\{\widetilde{Q}_U^n : n\ge 1\}$ is stochastically bounded in D.

Let A and S be two unit-rate Poisson processes. Let Z_{U0}^n and Z_U^n be the processes that characterize the number of customers in phase 1 service and phase 2 service, respectively. (Recall that arrivals have only phase 2 service, but initial customers may have phase 1 service). Let

$$\widetilde{X}_U^n(t):=n^{-3/4}(X_U^n(n^{1/4}t)-n),\quad \widetilde{Z}_U^n(t):=n^{-3/4}(Z_U^n(n^{1/4}t)-n),\widetilde{Z}_{U0}^n(t):=n^{-3/4}\widetilde{Z}_{U0}^n(n^{1/4}t).$$

Following similar arguments as in Section 5.2, $\widetilde{X}_U^n$ admits the following martingale representation

$$\widetilde{X}_U^n(t)=\widetilde{X}_U^n(0)-\beta^n t-n^{1/4}\int_0^t \widetilde{Z}_U^n(s)ds+\widetilde{M}_{UA}^n(t)-\widetilde{M}_{US}^n(t),\quad \text{for all } t\ge 0, \tag{A.1}$$

where

$$\widetilde{M}_{UA}^n(t)=n^{-3/4}(A(n^{1/4}\lambda^n t)-n^{1/4}\lambda^n t),\quad \text{and}$$

$$\widetilde{M}_{US}^n(t)=n^{-3/4}\left(S\left(\int_0^{n^{1/4}t} Z_U^n(s)ds\right)-\int_0^{n^{1/4}t} Z_U^n(s)ds\right),\quad \text{for } t\ge 0.$$

It follows from the Poisson FCLT (e.g., theorem 4.2 in Pang et al. [12]) that

$$n^{-5/8}(A(n^{5/4}\cdot)-n^{5/4}\eta(\cdot))\Rightarrow B(\cdot)\quad \text{and } n^{-5/8}(S(n^{5/4}\cdot)-n^{5/4}\eta(\cdot))\Rightarrow B(\cdot),\quad \text{for all } t\ge 0,$$

for a standard Brownian motion B, so that

$$\|n^{-3/4}(A-\eta)\|_{n^{5/4}t}\Rightarrow 0\quad \text{and } \|n^{-3/4}(S-\eta)\|_{n^{5/4}t}\Rightarrow 0,\quad \text{for all } t\ge 0.$$

Therefore,

$$n^{1/4}\lambda^n t=O(n^{5/4})t\quad \text{and}\quad \int_0^{n^{1/4}t} Z_U^n(s)ds\le n^{5/4}t$$

imply that

$$(\widetilde{M}_{UA}^n,\widetilde{M}_{US}^n)\Rightarrow(0\eta,0\eta)\quad \text{in } D^2,\quad \text{as } n\to\infty.$$

Consider the process

$$\xi^n(t):=\widetilde{M}_{UA}^n(t)-\widetilde{M}_{US}^n(t)+n^{1/4}\int_0^t \widetilde{Z}_{U0}^n(s)ds,\quad t\ge 0.$$

Using similar arguments as in the proof of Proposition 6(a), one can show that

$$n^{1/4}\int_0^{\cdot}\widetilde{Z}_{U0}^n(s)ds=o_P(1),$$

so that $\xi^n=o_P(1)$. Using the equality $\widetilde{Z}_U^n=\widetilde{X}_U^n\wedge 0-Z_{U0}^n$, (A.1) becomes

$$\widetilde{X}_U^n(t)=\widetilde{X}_U^n(s)-\beta^n(t-s)-n^{1/4}\int_s^t \widetilde{X}_U^n(s_1)\wedge 0ds_1+\xi^n(t)-\xi^n(s),\quad t\ge s\ge 0. \tag{A.2}$$

Take $s:=\sup\{u\in[0,t]: X_U^n(u)<0\}$, where, for $\emptyset$ denoting the empty set, $\sup\emptyset:=0$. Then either $s=t$ or $X_U^n\ge 0$ on $[s,t)$, implying that $n^{1/4}\int_s^t \widetilde{X}_U^n(s)\wedge 0ds=0$. Moreover, either $s=0$ or $X_U^n(s-)\le 0$, where in the latter case $X_U^n(s)\le 1$ because X_U^n only has either positive or negative jumps of size 1. In particular, $X_U^n(s)\le X_U^n(0)\vee 0+1$. Therefore,

$$\widetilde{X}_U^n(t)\le \widetilde{X}_U^n(0)\vee 0+n^{-3/4}-(\beta^n\wedge 0)t+2\|\xi^n\|_t. \tag{A.3}$$

Notice that the right-hand side is strictly positive and nondecreasing in t. Using $\widetilde{Q}_U^n = \widetilde{X}_U^n \vee 0$,

$$\|\widetilde{Q}_U^n\|_t \le \widetilde{X}_U^n(0) \vee 0 + n^{-3/4} - (\beta^n \wedge 0)t + 2\|\xi^n\|_t.$$

As $n \to \infty$, $\widetilde{X}_U^n(0) \Rightarrow X_0$ in $\mathbb{R}$ and $\beta^n \to \beta \in \mathbb{R}$, so that the right-hand side is stochastically bounded in $\mathbb{R}$ for any $t \ge 0$, so that $\{\widetilde{Q}^n : n \ge 1\}$ is stochastically bounded in D, as stated. □

Proof of Lemma 6. We first observe that, for any $t > 0$,

$$Q^n(t) \ge A^n(t) - A^n(t - w^n(t)) - R^n(t) + R^n(t - w^n(t)). \tag{A.4}$$

To see this, note that the head-of-line customer arrived at time $t - w^n(t)$. Thus, any waiting customer at time t must either be an initial customer, or a customer that arrived to the system during $[t - w^n(t), t)$. On the other hand, the number of those customers that arrived during $[t - w^n(t), t]$ and abandoned by time t is clearly no larger than the total number of abandonments during $[t - w^n(t), t]$. Thus, we get (A.4).

Notice that

$$\begin{aligned}
&n^{-3/4}A^n(n^{1/4}t) = \widetilde{M}_A^n(t) + n^{-1/2}\lambda^n t,\\
&n^{-3/4}A^n(n^{1/4}t - w^n(n^{1/4}t)) = \widetilde{M}_A^n(t - n^{-1/2}\widetilde{w}^n(t)) + n^{-1/2}\lambda^n(t - n^{-1/2}\widetilde{w}^n(t))\\
&n^{-3/4}R^n(n^{1/4}t) = \widetilde{M}_R^n(t) + \theta\int_0^t n^{1/4}\widetilde{Q}^n(s)ds\\
&n^{-3/4}R^n(n^{1/4}t - w^n(n^{1/4}t)) = \widetilde{M}_R^n(t - n^{-1/2}\widetilde{w}^n(t)) + \theta\int_0^{t - n^{-1/2}\widetilde{w}^n(t)} n^{1/4}\widetilde{Q}^n(s)ds,
\end{aligned}$$

Plugging these equalities in (A.4) and using the LOF scaling gives

$$\begin{aligned}
\widetilde{Q}^n(t) &\ge n^{-3/4}(A^n(n^{1/4}t) - A^n(n^{1/4}t - w^n(n^{1/4}t))\\
&\quad - R^n(n^{1/4}t) + R^n(n^{1/4}t - w^n(n^{1/4}t)))\\
&= n^{-1}\lambda^n\widetilde{w}^n(t) + \widetilde{M}_A^n(t) - \widetilde{M}_A^n(t - n^{-1/2}\widetilde{w}^n(t))\\
&\quad - \theta\int_{t - n^{-1/2}\widetilde{w}^n(t)}^t n^{1/4}\widetilde{Q}^n(s)ds - \widetilde{M}_R^n(t) + \widetilde{M}_R^n(t - n^{-1/2}\widetilde{w}^n(t))\\
&\ge \widetilde{w}^n(t)(n^{-1}\lambda^n - n^{-1/4}\theta\|\widetilde{Q}^n\|_t) - 2\|\widetilde{M}_A^n\|_t - 2\|\widetilde{M}_R^n|_t.
\end{aligned}$$

The statement follows from the facts that the processes $\widetilde{Q}^n$, $\widetilde{M}_A^n$, and $\widetilde{M}_R^n$ are all $o_P(1)$, $n^{-1/4}\|\widetilde{Q}^n\|_t \Rightarrow 0$ in $\mathbb{R}$, and $\lambda^n/n \to 1$. □

Proof of Lemma 7. Fix $n \ge 1$. That $F^n(s,t)$ is integrable follows from $-\theta^{-1}\lambda^n \le F^n(s,t) \le A^n(s)$. Let $\{\mathcal{G}_{s,t}^n : s \ge 0\}$ be the natural filtration generated by $F^n(s,t)$, augmented by including all P-null sets. Note that, for $0 \le s_1 < s_2 \le t$,

$$F^n(s_2,t) - F^n(s_1,t) = \int_{s_1}^{s_2} 1\{E_{A^n(s)}^n + T_{A^n(s)}^n > t\}dA^n(s) - \theta^{-1}\lambda^n(e^{\theta(s_2 - t)} - e^{\theta(s_1 - t)}),$$

and that the right-hand side is independent of $\mathcal{G}_{s_1,t}^n$. Hence,

$$\begin{aligned}
&E\left[\int_{s_1}^{s_2} 1\{E_{A^n(s)}^n + T_{A^n(s)}^n > t\}dA^n(s)\,\middle|\,\mathcal{G}_{s_1,t}^n\right]\\
&= E\left[\int_{s_1}^{s_2} 1\{E_{A^n(s)}^n + T_{A^n(s)}^n > t\}dA^n(s)\right]\\
&= E\left[\int_{s_1}^{s_2} E[1\{E_{A^n(s)}^n + T_{A^n(s)}^n > t\}|E_{A^n(s)}^n]dA^n(s)\right].
\end{aligned}$$

Because $T_{A^n(s)}^n$—the patience of the last customer to arrive before time s—is exponentially distributed and is independent of the arrival time $E_{A^n(s)}^n$,

$$E[1\{E_{A^n(s)}^n + T_{A^n(s)}^n > t\}|E_{A^n(s)}^n] = \exp(-\theta(E_{A^n(s)}^n - t)).$$

Therefore,

$$E\left[\int_{s_1}^{s_2} E[1\{E^n_{A^n(s)} + T^n_{A^n(s)} > t\}|E^n_{A^n(s)}]dA^n(s)\right] = E\left[\int_{s_1}^{s_2} \exp\left(-\theta(E^n_{A^n(s)} - t)\right)dA^n(s)\right].$$

Finally, A^n is a simple counting process, and $E^n_{A^n(s)} = s$ when $dA^n(s) = 1$, so that

$$\begin{aligned} E\left[\int_{s_1}^{s_2} \exp\left(-\theta(E^n_{A^n(s)} - t)\right)dA^n(s)\right] &= E\left[\int_{s_1}^{s_2} e^{-\theta(s-t)}dA^n(s)\right] \\ &= \theta^{-1}\lambda^n(e^{\theta(s_2-t)} - e^{\theta(s_1-t)}). \end{aligned}$$

Thus,

$$E[F^n(s_2,t) - F^n(s_1,t)|\mathcal{G}_{s_1,t}] = 0, \quad \text{for } 0 \le s_1 \le s_2 \le t,$$

implying that $\{F^n(s,t) : s \in [0,t]\}$ is a martingale.

To prove that $\{e^{\theta t}\sup_{s\in[0,t]}|F^n(s,t)| : t \ge 0\}$ is a submartingale, let $\{\mathcal{G}^n_t : t \ge 0\}$ be the right-continuous filtration generated by

$$\left(A^n(s), 1\{T^n_k + E^n_k < t\} : s \le t, k \le A^n(t)\right),$$

and augmented by including all P-null sets. It is easy to check that $F^n(s,s+t) \in \mathcal{G}^n_{s+t}$ and $\sup_{s\in[0,t]}|F(s,t)| \in \mathcal{G}^n_t$. We will show that $\sup_{s\in[0,t]}|F(s,t)|$ is a $\mathcal{G}^n$-submartingale, and therefore also a submartingale with respected to the (augmented) natural filtration.

To this end, fix $0 \le t_1 \le t_2$. for $s_1 \in [0,t_1]$ such that $E^n_{A^n(s_1)} + T^n_{A^n(s_1)} > t_1$. Due to the memoryless property of $T^n_{A^n(s_1)}$, $E^n_{A^n(s_1)} + T^n_{A^n(s_1)} - t_1$ is also an exponential random variable with rate θ, so that

$$P\left(E^n_{A^n(s_1)} + T^n_{A^n(s_1)} > t_2 | E^n_{A^n(s_1)} + T^n_{A^n(s_1)} > t_1\right) = e^{\theta(t_1-t_2)}.$$

Trivially,

$$E^n_{A^n(s_1)} + T^n_{A^n(s_1)} \le t_2 \quad \text{if} \quad E^n_{A^n(s_1)} + T^n_{A^n(s_1)} \le t_1.$$

Now, $E^n_{A^n(s_1)} + T^n_{A^n(s_1)} > t_2$ is, conditional on the event $\{E^n_{A^n(s_1)} + T^n_{A^n(s_1)} > t_1\}$, independent of $\mathcal{G}^n_{t_1}$. Finally, $1\{E^n_{A^n(s_1)} + T^n_{A^n(s_1)} > t_1\} \in \mathcal{G}^n_{t_1}$, implying that

$$E[1\{E^n_{A^n(s_1)} + T^n_{A^n(s_1)} > t_2\}|\mathcal{G}^n_{t_1}] = e^{\theta(t_1-t_2)}1\{E^n_{A^n(s_1)} + T^n_{A^n(s_1)} > t_1\}.$$

Integrating both sides of the equality with respect to s_1 over $[0,s]$, and using the equality

$$e^{\theta(t_1-t_2)}\theta^{-1}\lambda^n(e^{-\theta(t_1-s)} - e^{-\theta t_1}) = \theta^{-1}\lambda^n(e^{-\theta(t_2-s)} - e^{-\theta t_2}),$$

gives

$$E[F^n(s,t_2)|\mathcal{G}^n_{t_1}] = e^{\theta(t_1-t_2)}F^n(s,t_1), \quad \text{for all } 0 \le s \le t_1 \le t_2.$$

In particular, $\{e^{\theta(s+t)}F^n(s,s+t) : t \ge 0\}$ is a $\{\mathcal{G}^n_{s+t} : t \ge 0\}$-martingale.

Now, let $0 \le t_1 \le t_2$ and an arbitrary random time $S \le t_1$ such that $S \in \mathcal{G}^n_{t_1}$. We have

$$E\left[e^{\theta t_2}\sup_{s\in[0,t_2]}|F^n(s,t_2)|\middle|\mathcal{G}^n_{t_1}\right] \ge E[|e^{\theta t_2}F^n(S,t_2)||\mathcal{G}^n_{t_1}].$$

It follows from the facts that $S \in \mathcal{G}^n_{t_1}$, and that $\{e^{\theta(s+t)}F^n(s,s+t) : t \ge 0\}$ is a $\{\mathcal{G}^n_{s+t} : t \ge 0\}$-martingale, that

$$E[e^{\theta t_2}F^n(S,t_2)|\mathcal{G}^n_{t_1}] = e^{\theta t_1}F^n(S,t_1).$$

By Jensen's inequality

$$E[|e^{\theta t_2}F^n(S,t_2)||\mathcal{G}^n_{t_1}] \ge e^{\theta t_1}|F^n(S,t_1)|,$$

so that

$$E\left[e^{\theta t_2}\sup_{s\in[0,t_2]}|F^n(s,t_2)|\middle|\mathcal{G}^n_{t_1}\right] \ge e^{\theta t_1}|F^n(S,t_1)|. \tag{A.5}$$

Finally, $|F^n(s,t_1)| \in \mathcal{G}^n_{t_1}$ for any $s \in [0,t_1]$. Because $F^n(\cdot \wedge t_1, t_1)$ has right-continuous paths, for each $\epsilon > 0$, we can choose $S_\epsilon \in \mathcal{G}^n_{t_1}$ such that

$$\sup_{s\in[0,t_1]}|F^n(s,t_1)| \le |F^n(S_\epsilon,t_1)| + \epsilon, \text{w.p.1.}$$

Taking $S = S_\epsilon$ in (A.5) gives

$$E\left[e^{\theta t_2}\sup_{s\in[0,t_2]}|F^n(s,t_2)|\,\middle|\,\mathcal{G}^n_{t_1}\right] \geq e^{\theta t_1}\sup_{s\in[0,t_1]}|F^n(s,t_1)| - \epsilon, \text{ for all } t_1 \geq 0 \quad \text{and } \epsilon > 0.$$

The proof follows upon taking $\epsilon \to 0$. □

Proof of Lemma 8. We first prove that $\{\widetilde{U}^n_1 : n \geq 1\}$ is stochastically bounded in D. Consider the LOF-scaled process in (22),

$$\widetilde{U}^n_1(t) = \int_{\tilde{T}^n_0 \wedge t}^t 1\{n^{-1/2}\theta\widetilde{w}^n(s-) + s > t\}d\widetilde{M}^n_S(s).$$

Notice that the integrand is nonnegative and satisfies

$$1\{n^{-1/2}\theta\widetilde{w}^n(s-) + s > t\} \leq 1\{s \geq t - n^{-1/2}\theta\|\widetilde{w}^n\|_t\}, \text{ for all } 0 \leq s \leq t. \tag{A.6}$$

Thus,

$$|\widetilde{U}^n_1(t)| \leq \int_0^t 1\{s \geq t - n^{-1/2}\theta\|\widetilde{w}^n\|_t\}d|\widetilde{M}^n_S(t)| \leq \int_{t-n^{-1/2}\theta\|\widetilde{w}^n\|_t}^t d|\widetilde{M}^n_S(t)| \leq 2\|\widetilde{M}^n_S\|_t + 2\theta\|\widetilde{w}^n\|_t,$$

where the last inequality follows from (57). By Proposition 6(b) and Lemma 6, the right-hand side is stochastically bounded in D, implying that $\{\widetilde{U}^n_1 : n \geq 1\}$ is stochastically bounded in D as well.

To prove that $\{\widetilde{Z}^n_1 - \widetilde{Z}^n_0 : n \geq 1\}$ is stochastically bounded in D, consider the LOF-scaled process in (23):

$$\widetilde{Z}^n_1(t) - \widetilde{Z}^n_0(t) = \int_{\tilde{T}^n_0 \wedge t}^t 1\{n^{-1/2}\theta\widetilde{w}^n(s-) + s > t\}(n^{1/2} + n^{1/4}\widetilde{Z}^n_2(s))ds + \widetilde{U}^n_1(t).$$

Again, using $\widetilde{Z}^n_2 \leq 0$ w.p.1 and (A.6),

$$|\widetilde{Z}^n_1(t) - \widetilde{Z}^n_0(t)| \leq n^{1/2}\int_0^t 1\{s \geq t - n^{-1/2}\theta\|\widetilde{w}^n\|_t\}ds + |\widetilde{U}^n_1(t)| \leq \theta\|\widetilde{w}^n\|_t + \|\widetilde{U}_1\|_t.$$

Since the right-hand side is nondecreasing in t, Lemma 6 and the stochastic boundedness of $\{\widetilde{U}^n : n \geq 1\}$ in D imply that $\{\widetilde{Z}^n_1 - \widetilde{Z}^n_0 : n \geq 1\}$ is also stochastically bounded in D.

To prove that $\{\widetilde{X}^n : n \geq 1\}$ is stochastically bounded in D, we use the same arguments as in the proof of Proposition 7 to obtain (71), and in particular,

$$\widetilde{X}^n \geq \xi^n + \varepsilon^n,$$

where $\varepsilon^n = o_P(1)$. It follows from the stochastic boundedness of $\{\widetilde{w}^n : n \geq 1\}$ and $\{\widetilde{Z}^n_1 - \widetilde{Z}^n_0 : n \geq 1\}$ in D that ξ^n is also $O_P(1)$. Therefore, $\xi^n + \varepsilon^n \leq \widetilde{X}^n \leq \widetilde{Q}^n$ implies that $\{\widetilde{X}^n\}$ is stochastically bounded in D, and thus $\widetilde{Z}^n = \widetilde{X}^n \wedge 0$ implies that $\{\widetilde{Z}^n : n \geq 1\}$ is stochastically bounded in D. Finally, $\widetilde{Z}^n_2 + \widetilde{Z}^n_0 = \widetilde{Z}^n - (\widetilde{Z}^n_1 - \widetilde{Z}^n_0)$ implies that $\{\widetilde{Z}^n_2 + \widetilde{Z}^n_0 : n \geq 1\}$ is stochastically bounded in D. □

Appendix B. Proof of Proposition 2

Recall system $\mathcal{U}^n$ from the proof of Lemma 5, and notice that this system becomes an Erlang-C system if all the customers that are initially in the system have zero remaining phase 1 service time. As a result, all arguments regarding X^n_U in the proof of Lemma 5 hold for X^n_C as well. Therefore, (A.2) (taking $s = 0$) and (A.3) together give

$$\widetilde{X}^n_C(t) \geq \widetilde{X}^n_C(0) - \beta^n t + \xi^n(t), \quad \text{and } \widetilde{X}^n_C(t) \leq \widetilde{X}^n_C(0) \vee 0 + n^{-3/4} - (\beta^n \wedge 0)t + 2\|\xi^n\|_t, \quad t \geq 0,$$

for some process $\xi^n \in D$, which is $o_P(1)$. Using $\beta^n \to \beta \leq 0$ and $\widetilde{X}^n_C(0) \Rightarrow x_0 \geq 0$ in $\mathbb{R}$, we have $\widetilde{X}^n_C = x_C + o_P(1)$, implying that $\widetilde{X}^n_C \Rightarrow x_C$ in D.

References

[1] Billingsley P (1968) *Convergence of Probability Measures* (Wiley, Hoboken, NJ).
[2] Chen H, Yao DD (2013) *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*, vol. 46 (Springer Science & Business Media, Berlin).
[3] Dai JG (1995) On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* 5(1):49–77.
[4] De Vries J, Roy D, De Koster R (2017) Worth the wait? How waiting influences customer behavior and their inclination to return. *J. Oper. Management* 63(1):59–78.

[5] Gamarnik D, Goldberg DA (2013) Steady-state $GI/G/n$ queue in the Halfin–Whitt regime. *Ann. Appl. Probab.* 23(6):2382–2419.
[6] Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.
[7] Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.
[8] Kamae T, Krengel U, O'Brien GL (1977) Stochastic inequalities on partially ordered spaces. *Ann. Probab.* 5(6):899–912.
[9] Kang W, Ramanan K (2010) Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.* 20(6):2204–2260.
[10] Kang W, Ramanan K (2012) Asymptotic approximations for stationary distributions of many-server queues with abandonment. *Ann. Appl. Probab.* 22(2):477–521.
[11] Moyal P, Perry O (2022) Many-server limits for service systems with dependent service and patience times. *Queueing Systems* 100(3): 337–339.
[12] Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys* 4:193–267.
[13] Puha AL, Ward AR (2019) Scheduling an overloaded multiclass many-server queue with impatient customers. *Operations Research & Management Science in the Age of Analytics* (INFORMS, Catonsville, MD), 189–217.
[14] Puhalskii AA, Reiman MI (2000) The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* 32(2):564–595.
[15] Reed J (2009) The $G/GI/N$ queue in the Halfin–Whitt regime. *Ann. Appl. Probab.* 19(6):2211–2269.
[16] Reich M (2012) The offered-load process: Modeling, inference and applications. PhD thesis, Technion-Israel Institute of Technology, Haifa, Israel.
[17] Revuz D, Yor M (2013) *Continuous Martingales and Brownian Motion*, vol. 293 (Springer Science & Business Media, Berlin).
[18] Sigman K, Wolff RW (1993) A review of regenerative processes. *SIAM Rev.* 35(2):269–288.
[19] van Leeuwaarden JSH, Mathijsen BWJ, Zwart B (2019) Economies-of-scale in many-server queueing systems: Tutorial and partial review of the QED Halfin–Whitt heavy-traffic regime. *SIAM Rev.* 61(3):403–440.
[20] Whitt W (1992) Understanding the efficiency of multi-server service systems. *Management Sci.* 38(5):708–723.
[21] Whitt W (2002) *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues* (Springer, Berlin).
[22] Wolff RW (1989) *Stochastic Modeling and the Theory of Queues*, vol. 14 (Prentice Hall, Englewood Cliffs, NJ).
[23] Wu C, Bassamboo A, Perry O (2018) Service system with dependent service and patience times. *Management Sci.* 65(3):1151–1172.
[24] Wu CA, Bassamboo A, Perry O (2021) When service times depend on customers' delays: A solution to two empirical challenges. *Oper. Res.* ePub ahead of print December 1, https://doi.org/10.1287/opre2021.2179.