Technical Proofs

EC.1. Proof of Theorems 1 and 2

Proof of Theorem 1. We first characterize the diffusion limit of $\widetilde{Y}^n = \{\widetilde{Y}^n(t) : t \geq 0\}$ given by (17). Condition (3) implies that, for any $T > 0$,

$$\sup_{0\leq t\leq T}\left|\frac{(\lambda_n F_c^n(\omega^n) - s_n\mu)t}{\sqrt{\lambda_n}} - \beta t\right| = \left|\frac{(\lambda_n F_c^n(\omega^n) - s_n\mu)}{\sqrt{\lambda_n}} - \beta\right| T \to 0. \tag{EC.1}$$

It then follows from (1), (3), (EC.1), Lemmas EC.2 and EC.4 that $\widetilde{Y}^n \Rightarrow \widetilde{Y}$, where $\widetilde{Y} = \{\widetilde{Y}(t) : t \geq 0\}$ with

$$\widetilde{Y}(t) = \rho\beta t + \widetilde{\Lambda}(t) - \sqrt{\rho - 1}\mathcal{B}_A(t) - \sqrt{\rho}\mathcal{B}(t).$$

By Lemma EC.5 in Section EC.3, any subsequence of $\{\widetilde{V}^n\}_{n\in\mathbb{N}}$ has a further convergent subsequence, written as $\{\widetilde{V}^{n_k}\}_{k\in\mathbb{N}}$ such that

$$\widetilde{V}^{n_k} \Rightarrow V_\star \quad \text{as } k \to \infty, \tag{EC.2}$$

for a limit $\{V_\star(t) : t \geq 0\}$. The objective is to characterize the limit as the solution to (19). To this end, write the second term on the right-hand side of (16) as

$$\begin{aligned}
&\int_0^t \sqrt{\lambda_n}\Big(F^n(\omega^n + \frac{\widetilde{V}^n(x-)}{\sqrt{\lambda_n}}) - F^n(\omega^n)\Big)\mathsf{d}\bar{\Lambda}^n(x) \qquad\qquad \text{(EC.3)}\\
&\quad = \int_0^t f_\omega(\widetilde{V}^n(x-))\mathsf{d}\bar{\Lambda}^n(x) + \int_0^t \Big(\sqrt{\lambda_n}\big(F^n(\omega^n + \frac{\widetilde{V}^n(x-)}{\sqrt{\lambda_n}}) - F^n(\omega^n)\big) - f_\omega(\widetilde{V}^n(x-))\Big)\mathsf{d}\bar{\Lambda}^n(x).
\end{aligned}$$

According to Lemma 8.3 of Dai and Dai (1999) and (EC.2), the first term in (EC.3) converges to $\int_0^t f_\omega(V_\star(x))\mathsf{d}x$ along the subsequence $\{n_k\}_{k\in\mathbb{N}}$ as $k \to \infty$. By Condition (1) on the arrival process, for any $\varepsilon > 0$ there exists an N_1 such that when $n \geq N_1$,

$$\mathbb{P}\left(\sup_{0\leq t\leq T}|\bar{\Lambda}^n(t)| \geq 2T\right) \leq \frac{\varepsilon}{2}.$$

By the tightness proved in Lemma EC.5 of Section EC.3, for the above ε, there also exist $M > 0$ and N_2 such that for all $n \geq N_2$,

$$\mathbb{P}\left(\sup_{0\leq t\leq T}|\widetilde{V}^n(t)| \geq M\right) \leq \frac{\varepsilon}{2}.$$

With the help of Lemma 4.1 of Dai (1995), by Condition (4), $\sqrt{\lambda_n}\Big(F^n(\omega^n + \frac{x}{\sqrt{\lambda_n}}) - F^n(\omega^n)\Big)$ converges to $f_\omega(x)$ uniformly on compact sets. Thus, for any given $\delta > 0$, we can find an N_3 such that for $n \geq N_3$ and $x \in [-M, M]$,

$$\left|\sqrt{\lambda_n}\Big(F^n(\omega^n + \frac{x}{\sqrt{\lambda_n}}) - F^n(\omega^n)\Big) - f_\omega(x)\right| \leq \frac{\delta}{2T}.$$

So we can conclude that for all $n \geq \max(N_1, N_2, N_3)$,

$$\begin{aligned}&\mathbb{P}\left(\sup_{0\le t\le T}\left|\int_0^t \left(\sqrt{\lambda_n}\big(F^n(\omega^n+\frac{\widetilde{V}^n(x-)}{\sqrt{\lambda_n}})-F^n(\omega^n)\big)-f_\omega(\widetilde{V}^n(x-))\right)\mathsf{d}\bar{\Lambda}^n(x)\right|\geq\delta\right)\\&\leq\mathbb{P}\left(\sup_{0\le t\le T}|\widetilde{V}^n(t)|\geq M\right)+\mathbb{P}\left(\sup_{0\le t\le T}|\bar{\Lambda}^n(t)|\geq 2T\right)\leq\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon.\end{aligned}$$

This proves that the last term in (EC.3) converges to 0. As a result, (EC.3) converges to $\int_0^t f_\omega(V_\star(x))\mathsf{d}x$. It follows from (16), Assumptions (18) and the above proved convergence of (EC.3) that the limit process $\{V_\star(t) : t \geq 0\}$ is a solution to (19).

In view of Theorem 5.15 on page 341 of Karatzas and Shreve (1991), we know that when $f_\omega(\cdot)$ is locally integrable, the solution of (19) is unique in the sense of probability law. Hence, from Condition (4), we conclude weak convergence of $\{\widetilde{V}^n\}_{n\in\mathbb{N}}$ and that the corresponding limit satisfies (19). □

Proof of Theorem 2. From Theorem 2.8 in Whitt (1980), proving the convergence on (ω, ∞) is equivalent to proving the convergence on $[\omega + \delta, \infty)$, for any $\delta > 0$, which we now proceed to prove.

By (7), when $t \geq V^n(0)$, the queue length process can be written as

$$\begin{aligned}Q^n(t)&=\sum_{i=\Lambda^n(\kappa^n(t))+1}^{\Lambda^n(t-\omega^n)}\mathbf{1}_{\{u_i^n+\tau_i^n>t\}}+\sum_{i=\Lambda^n(t-\omega^n)+1}^{\Lambda^n(t)}\mathbf{1}_{\{u_i^n+\tau_i^n>t\}}\\&=\sum_{i=\Lambda^n(\kappa^n(t))+1}^{\Lambda^n(t-\omega^n)}\mathbf{1}_{\{u_i^n+\tau_i^n>t\}}+\sum_{i=\Lambda^n(t-\omega^n)+1}^{\Lambda^n(t)}\left(\mathbf{1}_{\{u_i^n+\tau_i^n>t\}}-F_c^n(t-\tau_i^n)\right)\\&\quad+\int_{t-\omega^n}^t F_c^n(t-x)\mathsf{d}\left(\Lambda^n(x)-\lambda_n x\right)+\lambda_n\int_{t-\omega^n}^t F_c^n(t-x)\mathsf{d}x.\end{aligned}$$

Applying the diffusion scaling (20),

$$\widetilde{Q}^n(t)=\widetilde{M}^n(t)+\frac{1}{\sqrt{\lambda_n}}\sum_{i=\Lambda^n(\kappa^n(t))+1}^{\Lambda^n(t-\omega^n)}\mathbf{1}_{\{u_i^n+\tau_i^n>t\}},\tag{EC.4}$$

where

$$\widetilde{M}^n(t)=\frac{1}{\sqrt{\lambda_n}}\sum_{i=\Lambda^n(t-\omega^n)+1}^{\Lambda^n(t)}\left(\mathbf{1}_{\{u_i^n+\tau_i^n>t\}}-F_c^n(t-\tau_i^n)\right)+\int_{t-\omega^n}^t F_c^n(t-x)\mathsf{d}\tilde{\Lambda}^n(x).\tag{EC.5}$$

Following the idea given by Liu and Whitt (2014), the process $\widetilde{M}^n(\cdot)$ can be viewed as the diffusion-scaled queue length process of an infinite-server queue, with service times $u_i^n \wedge \omega^n$. With a modification to the proof for Theorem 3.1 (more specifically, Lemma 5.3) in Krichagina and Puhalskii (1997), we obtain that the first term in (EC.5) weakly converges to the process $\{\mathcal{G}(t) : t \geq \omega + \delta\}$. See Appendix EC.6 for details on the modifications. For the second term, integrating-by-part, we have

$$\int_{t-\omega^n}^t F_c^n(t-x)\mathsf{d}\tilde{\Lambda}^n(x)=\tilde{\Lambda}^n(t)-F_c^n(\omega^n)\tilde{\Lambda}^n(t-\omega^n)-\int_{t-\omega^n}^t\tilde{\Lambda}^n(x)\mathsf{d}F_c^n(t-x).\tag{EC.6}$$

Using the Skorohod representation theorem, we embed all the random objects in a common probability space. We maintain the original notation for the mapped random objects. On the new probability space, we have

$$\sup_{0\le t\le T} |\tilde{\Lambda}^n(t) - \tilde{\Lambda}(t)| \to 0, \quad \text{as } n\to\infty, \tag{EC.7}$$

on each sample path. Note that

$$\begin{aligned}&\int_{t-\omega^n}^{t} \tilde{\Lambda}^n(t)\mathsf{d}F_c^n(t-x) - \int_{t-\omega}^{t} \tilde{\Lambda}(x)\mathsf{d}F_c(t-x)\\ &= \int_{t-\omega^n}^{t} \left(\tilde{\Lambda}^n(t) - \tilde{\Lambda}(x)\right)\mathsf{d}F_c^n(t-x) + \int_{t-\omega^n}^{t-\omega} \tilde{\Lambda}(x)\mathsf{d}F_c^n(t-x)\\ &\quad + \int_{t-\omega}^{t} \tilde{\Lambda}(x)\mathsf{d}\left(F_c^n(t-x) - F_c(t-x)\right).\end{aligned}$$

The first two terms on the right-hand side converge to zero in probability, following (EC.7) and $|F^n(\omega^n) - F^n(\omega)| \le |F^n(\omega) - F(\omega)| + |F^n(\omega^n) - F(\omega)| \to 0$. Since $\tilde{\Lambda}$ is a Brownian motion, thus for any fixed $T > 0$,

$$\lim_{\Gamma\to\infty} \mathbb{P}\left(\sup_{0\le t\le T} |\tilde{\Lambda}(t)| \ge \Gamma\right) = 0.$$

This and $\{F^n\}_{n\in\mathbb{N}}$ converges to F in total variation on $[0,\omega]$ imply that the last term also converges to zero in probability. Combining the above convergence with (EC.6), we conclude that the second term in (EC.5) weakly converges to $\{\int_{t-\omega}^{t} F_c(t-x)\mathsf{d}\tilde{\Lambda}(x) : t \ge \omega+\delta\}$. So we have

$$\widetilde{M}^n(t) \Rightarrow \mathcal{G}(t) + \int_{t-\omega}^{t} F_c(t-x)\mathsf{d}\tilde{\Lambda}(x) \quad \text{on} \quad [\omega+\delta, \infty). \tag{EC.8}$$

Now we consider the second term on the left-hand side in (EC.4). For any $M > 0$ and $n\in\mathbb{N}$, define the event $\Omega_M^n = \{\sup_{\omega+\delta\le t\le T} |\sqrt{\lambda_n}\big(t - \omega^n - \kappa^n(t)\big)| \le M\}$. It is clear that, on the event Ω_M^n, we have

$$\sum_{i=\Lambda^n(\kappa^n(t))+1}^{\Lambda^n(t-\omega^n)} \mathbf{1}_{\{u_i^n > \omega^n + \frac{M}{\sqrt{\lambda_n}}\}} \le \sum_{i=\Lambda^n(\kappa^n(t))+1}^{\Lambda^n(t-\omega^n)} \mathbf{1}_{\{u_i^n + \tau_i^n > t\}} \le \sum_{i=\Lambda^n(\kappa^n(t))+1}^{\Lambda^n(t-\omega^n)} \mathbf{1}_{\{u_i^n > \omega^n - \frac{M}{\sqrt{\lambda_n}}\}}.$$

Introduce

$$\begin{aligned}\widetilde{G}_{M-}^n(t) = \frac{1}{\sqrt{\lambda_n}}\Big[&\sum_{i=\Lambda^n(\kappa^n(t))+1}^{\Lambda^n(t-\omega^n)} \Big(\mathbf{1}_{\{u_i^n > \omega^n - \frac{M}{\sqrt{\lambda_n}}\}} - F_c^n\Big(\omega^n - \frac{M}{\sqrt{\lambda_n}}\Big)\Big)\\ &+ F_c^n\Big(\omega^n - \frac{M}{\sqrt{\lambda_n}}\Big)\Big(\Lambda^n(t-\omega^n) - \Lambda^n(\kappa^n(t)) - \lambda_n(t-\omega^n-\kappa^n(t))\Big)\\ &+ \Big(F_c^n\Big(\omega^n - \frac{M}{\sqrt{\lambda_n}}\Big) - F_c^n(\omega^n)\Big)\lambda_n\Big(t - \omega^n - \kappa^n(t)\Big)\Big],\end{aligned}$$

$$
\begin{aligned}
\widetilde{G}^n_{M+}(t) = \frac{1}{\sqrt{\lambda_n}}\Big[& \sum_{i=\Lambda^n(\kappa^n(t))+1}^{\Lambda^n(t-\omega^n)} \Big(\mathbf{1}_{\{u_i^n > \omega^n + \frac{M}{\sqrt{\lambda_n}}\}} - F_c^n(\omega^n + \frac{M}{\sqrt{\lambda_n}})\Big) \\
& + F_c^n\Big(\omega^n + \frac{M}{\sqrt{\lambda_n}}\Big)\Big(\Lambda^n(t-\omega^n) - \Lambda^n(\kappa^n(t)) - \lambda_n(t-\omega^n-\kappa^n(t))\Big) \\
& + \Big(F_c^n\Big(\omega^n + \frac{M}{\sqrt{\lambda_n}}\Big) - F_c^n(\omega^n)\Big)\lambda_n\Big(t-\omega^n-\kappa^n(t)\Big)\Big].
\end{aligned}
$$

Then

$$
\begin{aligned}
\frac{1}{\sqrt{\lambda_n}} \sum_{i=\Lambda^n(\kappa^n(t))+1}^{\Lambda^n(t-\omega^n)} \mathbf{1}_{\{u_i^n > \omega^n - \frac{M}{\sqrt{\lambda_n}}\}} &= \widetilde{G}^n_{M-}(t) + F_c^n(\omega^n)\sqrt{\lambda_n}(t-\omega^n-\kappa^n(t)), \\
\frac{1}{\sqrt{\lambda_n}} \sum_{i=\Lambda^n(\kappa^n(t))+1}^{\Lambda^n(t+\omega^n)} \mathbf{1}_{\{u_i^n > \omega^n + \frac{M}{\sqrt{\lambda_n}}\}} &= \widetilde{G}^n_{M+}(t) + F_c^n(\omega^n)\sqrt{\lambda_n}(t-\omega^n-\kappa^n(t)).
\end{aligned}
$$

Due to being on the event Ω^n_M,

$$
\begin{aligned}
&\sup_{\omega+\delta\le t\le T}\Big|\widetilde{Q}^n(t) - \widetilde{M}^n(t) - F_c^n(\omega^n)\sqrt{\lambda_n}(t-\omega^n-\kappa^n(t))\Big| \\
&\quad \le \sup_{\omega+\delta\le t\le T}|\widetilde{G}^n_{M-}(t)| + \sup_{\omega+\delta\le t\le T}|\widetilde{G}^n_{M+}(t)|.
\end{aligned}
$$

As a result, for any $\eta > 0$,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{\omega+\delta\le t\le T}\Big|\widetilde{Q}^n(t) - \widetilde{M}^n(t) - F_c^n(\omega^n)\sqrt{\lambda_n}(t-\omega^n-\kappa^n(t))\Big| \ge \eta\right) \\
&\quad \le \mathbb{P}((\Omega^n_M)^c) + \mathbb{P}\left(\sup_{\omega+\delta\le t\le T}|\tilde{G}^n_{M-}(t)| + \sup_{\omega+\delta\le t\le T}|\tilde{G}^n_{M+}(t)| \ge \eta\right).
\end{aligned}
\tag{EC.9}
$$

Note that the definition of $\kappa^n(\cdot)$ in (6) and Proposition 1 imply that, as $n\to\infty$,

$$
\sup_{V^n(0)\le t\le T}\Big|\kappa^n(t) - t + \omega\Big| \Rightarrow 0.
$$

By the initial condition (18), the probability that $V^n(0) > \omega+\delta$ is vanishing with $n\to\infty$. As a result, we have

$$
\sup_{\omega+\delta\le t\le T}\Big|\kappa^n(t) - t + \omega\Big| \Rightarrow 0.
$$

Then it is clear, for any fixed $M > 0$,

$$
\sup_{\omega+\delta\le t\le T}|\tilde{G}^n_{M-}(t)| + \sup_{\omega+\delta\le t\le T}|\tilde{G}^n_{M+}(t)| \Rightarrow 0. \tag{EC.10}
$$

By Theorem 1, $\sup_{0\le t\le T}\sqrt{\lambda_n}|V^n(t) - V^n(t-)| \Rightarrow 0$, as $n\to\infty$. From the definition (6), we know that for $\omega+\delta\le t\le T$, $t\le \kappa^n(t) + V^n(\kappa^n(t)) \le t + \sup_{\omega+\delta\le t\le T}|V^n(\kappa^n(t)) - V^n(\kappa^n(t)-)|$. This, together with (11), implies that, as $n\to\infty$,

$$
\begin{aligned}
&\sup_{\omega+\delta\le t\le T}\Big|\sqrt{\lambda_n}\big(t-\kappa^n(t) - V^n(\kappa^n(t))\big)\Big| \\
&\quad = \sup_{\omega+\delta\le t\le T}\Big|\sqrt{\lambda_n}\big(t-\kappa^n(t)-\omega^n\big) - \widetilde{V}^n(\kappa^n(t))\Big| \Rightarrow 0.
\end{aligned}
\tag{EC.11}
$$

The first implication of (EC.11) is that

$$\lim_{M\to\infty}\lim_{n\to\infty}\mathbb{P}\left((\Omega_M^n)^c\right)=0.$$

Combining this with (EC.10) and (EC.9), we know that $\{\widetilde{Q}^n(t)\}$ and $\{\widetilde{M}^n(t)+F_c^n(\omega^n)\sqrt{\lambda_n}(t-\omega^n-\kappa^n(t))\}$ have the same weak limit. Since (EC.11) also implies that $\sqrt{\lambda_n}\big(t-\kappa^n(t)-\omega^n\big)\Rightarrow\widetilde{V}^n(t-\omega)$, the result of the theorem follows from (EC.8). □

EC.2. Proofs of Propositions 1–3

In this section, we provide the proofs for Propositions 1–3.

Proof of Proposition 1. It suffices to show, in view of the convergence $\omega^n\to\omega$, as $n\to\infty$, that for any $T>0$ and $\delta\in(0,\ \omega/2)$,

$$\mathbb{P}\Big(\sup_{0\le t\le T}|V^n(t)-\omega^n|\ge\delta\Big)\to 0 \text{ as } n\to\infty. \tag{EC.12}$$

Define $\bar{V}^n(t)=V^n(t)-\omega^n$, $\eta_1^n=\inf\{t\ge 0:\bar{V}^n(t)\ge\delta\}$ and $\eta_2^n=\inf\{t\ge 0:\bar{V}^n(t)\le-\delta\}$. Let $\bar{\Omega}_1^n(\delta,T)=\{\eta_1^n\le\eta_2^n,\eta_1^n\le T\}$, $\bar{\Omega}_2^n(\delta,T)=\{\eta_1^n>\eta_2^n,\eta_2^n\le T\}$, and $\bar{\Omega}_0^n(\delta)=\{\bar{V}^n(0)\le\delta/4\}$. In view of (18), to get (EC.12), it is sufficient to prove that the probabilities of the events $\bar{\Omega}_1^n(\delta,T)\cap\bar{\Omega}_0^n(\delta)$ and $\bar{\Omega}_2^n(\delta,T)\cap\bar{\Omega}_0^n(\delta)$ vanish as n converges to infinity. We will only consider the event $\bar{\Omega}_1^n(\delta,T)\cap\bar{\Omega}_0^n(\delta)$, since the analysis of $\bar{\Omega}_2^n(\delta,T)\cap\bar{\Omega}_0^n(\delta)$ is similar. On the set $\bar{\Omega}_1^n(\delta,T)\cap\bar{\Omega}_0^n(\delta)$, define $\eta_{12}^n=\sup\{0\le t\le\eta_1^n:\bar{V}^n(t)\le\delta/3\}\vee 0$. By the definitions of η_1^n and η_{12}^n, we clearly have that

$$\bar{V}^n(\eta_1^n)\ge\delta,\quad\text{and}\quad\bar{V}^n(\eta_{12}^n-)\le\frac{\delta}{3}.$$

In view of $F^n(\omega^n+x)\ge F^n(\omega^n)$ for any $x\ge 0$, by (16)–(17), we have that

$$\begin{aligned}
&\bar{V}^n(\eta_1^n)-\bar{V}^n(\eta_{12}^n-)\\
&\le\frac{\lambda_n}{s_n\mu}\Bigg[F_c^n(\omega^n)\frac{\widetilde{\Lambda}^n(\eta_1^n)-\widetilde{\Lambda}^n(\eta_{12}^n-)}{\sqrt{\lambda_n}}-\frac{1}{\lambda_n}\sum_{i=\Lambda(\eta_{12}^n)}^{\Lambda^n(\eta_1^n)}\big(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n)\big)\\
&\quad+\frac{(\lambda_nF_c^n(\omega)-s_n\mu)(\eta_1^n-\eta_{12}^n)}{\lambda_n}-\frac{\widetilde{B}^n(\eta_1^n+V^n(\eta_1^n))-\widetilde{B}^n(\eta_{12}^n+V^n(\eta_{12}^n-))}{\sqrt{\lambda_n}}\Bigg].
\end{aligned}\tag{EC.13}$$

By (1)–(3), we know that, as $n\to\infty$,

$$\mathbb{P}\Big(\Big|\frac{\lambda_nF_c^n(\omega)}{s_n\mu}\cdot\frac{\widetilde{\Lambda}^n(\eta_1^n)-\widetilde{\Lambda}^n(\eta_{12}^n-)}{\sqrt{\lambda_n}}\Big|>\frac{\delta}{6}\Big)\to 0, \tag{EC.14}$$

$$\mathbb{P}\Big(\Big|\frac{1}{s_n\mu}(\lambda_nF_c^n(\omega^n)-s_n\mu)(\eta_1^n-\eta_{12}^n)\Big|>\frac{\delta}{6}\Big)\to 0. \tag{EC.15}$$

Using Lemma EC.1, we have

$$\mathbb{P}\Big(\Big|\frac{1}{s_n\mu}\sum_{i=\Lambda(\eta_{12}^n)}^{\Lambda^n(\eta_1^n)}\big(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n)\big)\Big|>\frac{\delta}{6}\Big)\to 0. \tag{EC.16}$$

To get that the last term in (EC.13) also vanishes, let $S^n(t)$ denote the number of customers in service at time t, and $D^n(t)$ the number of departures through service completion by time t. We can relate these two processes with $B^n(t)$ by

$$B^n(t) = D^n(t) + S^n(t) - S^n(0), \tag{EC.17}$$

which implies that

$$\begin{aligned} B^n(\eta_1^n + V^n(\eta_1^n)) - B^n(\eta_{12}^n + V^n(\eta_{12}^n-)) &= D^n(\eta_1^n + V^n(\eta_1^n)) - D^n(\eta_{12}^n + V^n(\eta_{12}^n-)) \\ &\quad + S^n(\eta_1^n + V^n(\eta_1^n)) - S^n(\eta_{12}^n + V^n(\eta_{12}^n-)). \end{aligned}$$

As $V^n(\cdot)$ is always positive on $[\eta_{12}^n + V^n(\eta_{12}^n-), \eta_1^n + V^n(\eta_1^n))$, all the servers are busy; hence $S^n(\eta_1^n + V^n(\eta_1^n)) = S^n(\eta_{12}^n + V^n(\eta_{12}^n-)) = s_n$. As a result, noticing that the service time is exponential with rate μ,

$$\begin{aligned} B^n(\eta_1^n + V^n(\eta_1^n)) - B^n(\eta_{12}^n + V^n(\eta_{12}^n-)) &= D^n(\eta_1^n + V^n(\eta_1^n)) - D^n(\eta_{12}^n + V^n(\eta_{12}^n-)) \\ &= \mathcal{S}\left(s_n(\eta_1^n + V^n(\eta_1^n))\right) - \mathcal{S}\left(s_n(\eta_{12}^n + V^n(\eta_{12}^n-))\right), \end{aligned}$$

where $\{\mathcal{S}(t) : t \geq 0\}$ is a Poisson process with rate μ. Hence, we have that, as $n \to \infty$,

$$\mathbb{P}\Big(\frac{\lambda_n}{s_n\mu}\Big|\frac{\widetilde{B}^n(\eta_1^n + V^n(\eta_1^n)) - \widetilde{B}^n(\eta_{12}^n + V^n(\eta_{12}^n-))}{\sqrt{\lambda_n}}\Big| > \frac{\delta}{6}\Big) \to 0. \tag{EC.18}$$

Combining (EC.13)–(EC.16) and (EC.18), the probability of the event $\bar{\Omega}_1^n(\delta, T) \cap \bar{\Omega}_0^n(\delta)$ will vanish as $n \to \infty$. □

Proof of Proposition 2. First consider the stationary distribution of the diffusion limit for the virtual waiting time process. Introduce $g(x) = \rho(f_\omega(x) - \beta)$. Note that, in view of (26), $\lim_{x\to\infty} g(x) > 0$ and $\lim_{x\to-\infty} g(x) < 0$. Now let $\mathcal{X} = \{\mathcal{X}(t) : t \geq 0\}$ be the solution to the following stochastic differential equation:

$$\mathsf{d}\mathcal{X}(t) = -g(\mathcal{X}(t))\mathsf{d}t + \sigma \mathsf{d}\mathcal{W}(t), \quad t \geq 0.$$

It is enough to prove that the stationary distribution of $\mathcal{X}$ has the density

$$\pi(y) = C\exp\left(-\frac{2}{\sigma^2}\int_0^y g(x)\mathsf{d}x\right), \tag{EC.19}$$

where C is a normalizing constant. Noting that the generator of $\mathcal{X}$ is

$$\mathcal{A} = \frac{\sigma^2}{2}\frac{\mathsf{d}^2}{\mathsf{d}x^2} - g(x)\frac{\mathsf{d}}{\mathsf{d}x},$$

it is enough to prove that the function π in (EC.19) satisfies

$$\int_{\mathbb{R}} \mathcal{A}f(x)\pi(x)\mathsf{d}x = 0, \tag{EC.20}$$

for all $f(\cdot)$ in the class of bounded, twice continuously differentiable functions (see Ethier and Kurtz (1986), page 248). However, with $\lim_{x\to\infty} g(x) > 0$, it can be easily verified that, with π given in (EC.19), we have

$$\int_{\mathbb{R}_+} \mathcal{A}f(x)\pi(x)\mathsf{d}x = \frac{C\sigma^2}{2}\int_{\mathbb{R}_+} \mathsf{d}\Big[\exp\Big(-\frac{2}{\sigma^2}\int_0^y g(x)\mathsf{d}x\Big)f'(y)\Big] = -\frac{C\sigma^2}{2}f'(0). \quad \text{(EC.21)}$$

Similarly,

$$\int_{\mathbb{R}_-} \mathcal{A}f(x)\pi(x)\mathsf{d}x = \frac{C\sigma^2}{2}\int_{\mathbb{R}_-} \mathsf{d}\Big[\exp\Big(-\frac{2}{\sigma^2}\int_0^y g(x)\mathsf{d}x\Big)f'(y)\Big] = \frac{C\sigma^2}{2}f'(0). \quad \text{(EC.22)}$$

We now conclude (EC.20) by summing up (EC.21) and (EC.22). This implies (EC.19), and hence (27).

For the stationary distribution of the diffusion limit of the queue length, note that, for $t>\omega$, $\mathcal{G}(t)$ is normally distributed with zero mean and variance $\int_0^\omega F(x)F_c(x)\mathsf{d}x$ due to (23). Similarly, $\int_{t-\omega}^t F_c(t-x)\mathsf{d}\tilde{\Lambda}(x)$ follows a zero-mean normal distribution with variance $\theta^2\int_0^\omega (F_c(x))^2\mathsf{d}x$. Hence the second result is implied by Theorem 2. This completes the proof. □

Proof of Proposition 3. We will prove the statement by contradiction, via considering two cases:

$$\text{(i)}\ \limsup_{\lambda\to\infty}\sqrt{\lambda}(\tau_* - \tau_*^\lambda) > 0 \quad \text{and} \quad \text{(ii)}\ \liminf_{\lambda\to\infty}\sqrt{\lambda}(\tau_* - \tau_*^\lambda) < 0.$$

To that end, we first note that for any nondecreasing function $f(\cdot)$, the function defined by

$$\frac{\int_0^\infty \exp\left(-\frac{\rho}{2\sigma^2}\int_0^y [f(x)-\beta]\mathsf{d}x\right)\mathsf{d}y}{\int_{-\infty}^\infty \exp\left(-\frac{\rho}{2\sigma^2}\int_0^y [f(x)-\beta]\mathsf{d}x\right)\mathsf{d}y} \quad \text{(EC.23)}$$

is continuous and strictly increasing in β. In the remainder of the proof, we first propose a feasible solution, and then compare it with any optimal solution that satisfies either of the above two cases, to get a contradiction to the optimality of an optimal solution.

A feasible solution. This solution is constructed as follows. Suppose an announcement is made exactly at time τ_*. Then by the definition of τ_*, $\omega_{\tau_*}=\tau_*$. Hence, the first constraint on the fraction of abandonment in (44) holds. Thus we only consider the second constraint on waiting time in (44), which further becomes $\mathbb{P}(W^\lambda(\infty) > \tau_*) \le \alpha_2$. Let $\widehat{s}_*^\lambda$ be its optimal solution. (We append the superscript λ to emphasize the dependency on the arrival rate.) It follows from (29) and (32) that the optimal number of servers for announcement time τ_* is given by (see also problem (38), and (39)–(40))

$$\widehat{s}_*^\lambda = \frac{\lambda}{\mu}H_c(\tau_*|\tau_*) - \frac{\beta_*}{\mu}\sqrt{\lambda} + o(\sqrt{\lambda}), \quad \text{(EC.24)}$$

where β_* solves (46). Then by the continuity and monotonicity of the function given by (EC.23) with $f(\cdot)=h_{\tau_*}(\cdot)$,

$$H_c(\tau_*|\tau_*)\cdot\frac{\int_0^\infty \exp\left(-\frac{\rho}{2\sigma^2}\int_0^y [h_{\tau_*}(x)-\beta_*]\mathsf{d}x\right)\mathsf{d}y}{\int_{-\infty}^\infty \exp\left(-\frac{\rho}{2\sigma^2}\int_0^y [h_{\tau_*}(x)-\beta_*]\mathsf{d}x\right)\mathsf{d}y} = \alpha_2. \quad \text{(EC.25)}$$

This is equivalent to

$$
\begin{aligned}
(1-\alpha_1-\alpha_2)\cdot \int_0^\infty \exp\left(-\frac{\rho}{2\sigma^2}\int_0^y [h_{\tau_*}(x)-\beta_*]\mathsf{d}x\right)\mathsf{d}y \\
=\alpha_2\int_{-\infty}^0 \exp\left(-\frac{\rho}{2\sigma^2}\int_0^y [h_{\tau_*}(x)-\beta_*]\mathsf{d}x\right)\mathsf{d}y. \qquad \text{(EC.26)}
\end{aligned}
$$

By the definition of $f_{\tau_*}(\cdot)$,

$$
h_{\tau_*}(x)=\begin{cases} e^{-h_0\tau_*}h_1x, & \text{if } x\ge 0,\\ e^{-h_0\tau_*}h_0x, & \text{if } x<0.\end{cases} \qquad \text{(EC.27)}
$$

Obviously, $(\widehat{s}_*^\lambda,\tau_*)$ is a feasible solution to our original problem (44). (Indeed it is the staffing level (45).)

Case (i) There is a subsequence along which the limit will be positive. To simplify the notation, we still use λ to index the subsequence, i.e. $\lim_{\lambda\to\infty}\sqrt{\lambda}(\tau_*-\tau_*^\lambda)>0$. We will first focus on the subcase that

$$
0<\lim_{\lambda\to\infty}\sqrt{\lambda}(\tau_*-\tau_*^\lambda)<\infty. \qquad \text{(EC.28)}
$$

Note that in this case $\omega_{\tau_*^\lambda}>\tau_*^\lambda$ (because $H(\tau_*^\lambda|\tau_*^\lambda)<\alpha_1$), then the constraint on waiting time in (44) becomes $\mathbb{P}\left(W^\lambda(\infty)>\tau_*^\lambda\right)$. Similar to (EC.24)–(EC.25) (noticing that the constraint $\mathbb{P}(\mathrm{Ab})$ can be achieved from the first order), the optimal number of servers is

$$
s_*^\lambda=\frac{\lambda}{\mu}H_c(\tau_*^\lambda|\tau_*^\lambda)-\frac{\beta_*^\lambda}{\mu}\sqrt{\lambda}+o(\sqrt{\lambda}), \qquad \text{(EC.29)}
$$

where β_*^λ solves

$$
\begin{aligned}
&\max_\beta \quad \beta\\
&\text{s.t.} \quad H_c(\tau_*^\lambda|\tau_*^\lambda)\cdot\frac{\int_0^\infty \exp\left(-\frac{\rho}{2\sigma^2}\int_0^y[f_{\tau_*^\lambda}(x)-\beta]\mathsf{d}x\right)\mathsf{d}y}{\int_{-\infty}^\infty \exp\left(-\frac{\rho}{2\sigma^2}\int_0^y[f_{\tau_*^\lambda}(x)-\beta]\mathsf{d}x\right)\mathsf{d}y}\le\alpha_2
\end{aligned} \qquad \text{(EC.30)}
$$

with $f_{\tau_*^\lambda}(x)=\sqrt{\lambda}[H(\tau_*^\lambda+\frac{x}{\sqrt{\lambda}}|\tau_*^\lambda)-H(\tau_*^\lambda|\tau_*^\lambda)]$. From the definitions of $f_{\tau_*^\lambda}(x)$ and H,

$$
f_{\tau_*^\lambda}(x)=\begin{cases} e^{-h_0\tau_*^\lambda}\sqrt{\lambda}\Big(1-\exp(-h_1\frac{x}{\sqrt{\lambda}})\Big), & \text{if } x\ge 0,\\ e^{-h_0\tau_*^\lambda}\sqrt{\lambda}\Big(1-\exp(-h_0\frac{x}{\sqrt{\lambda}})\Big), & \text{if } -\tau_*^\lambda\sqrt{\lambda}\le x<0.\end{cases} \qquad \text{(EC.31)}
$$

This, by the continuity and monotonicity of the function given by (EC.23), similar to (EC.26), implies that

$$
\begin{aligned}
\Big(H_c(\tau_*^\lambda|\tau_*^\lambda)-\alpha_2\Big)\cdot\int_0^\infty \exp\Big(-\frac{\rho}{2\sigma^2}\int_0^y[f_{\tau_*^\lambda}(x)-\beta_*^\lambda]\mathsf{d}x\Big)\mathsf{d}y\\
=\alpha_2\cdot\int_{-\infty}^0 \exp\Big(-\frac{\rho}{2\sigma^2}\int_0^y[f_{\tau_*^\lambda}(x)-\beta_*^\lambda]\mathsf{d}x\Big)\mathsf{d}y. \qquad \text{(EC.32)}
\end{aligned}
$$

In view of (EC.28) and (EC.31),

$$
\lim_{\lambda\to\infty}\tau_*^\lambda=\tau_* \text{ and } \lim_{\lambda\to\infty}f_{\tau_*^\lambda}(x)=h_{\tau_*}(x)=\begin{cases} e^{-h_0\tau_*}h_1x, & \text{if } x\ge 0,\\ e^{-h_0\tau_*}h_0x, & \text{if } x<0.\end{cases} \qquad \text{(EC.33)}
$$

This together with (EC.32) implies that

$$\begin{aligned}&(1-\alpha_1-\alpha_2)\cdot\int_0^{\infty}\exp\left(-\frac{\rho}{2\sigma^2}\int_0^y[h_{\tau_*}(x)-\liminf_{\lambda\to\infty}\beta_*^\lambda]\mathsf{d}x\right)\mathsf{d}y\\&=\alpha_2\int_{-\infty}^{0}\exp\left(-\frac{\rho}{2\sigma^2}\int_0^y[h_{\tau_*}(x)-\liminf_{\lambda\to\infty}\beta_*^\lambda]\mathsf{d}x\right)\mathsf{d}y.\end{aligned}\tag{EC.34}$$

Using the continuity and monotonicity of the function given by (EC.23), therefore, we have

$$\beta_*=\liminf_{\lambda\to\infty}\beta_*^\lambda.\tag{EC.35}$$

Similarly, taking the upper limit on both sides of (EC.32), we have

$$\beta_*=\limsup_{\lambda\to\infty}\beta_*^\lambda.\tag{EC.36}$$

On the other hand, by (EC.28)–(EC.29) and the definition of H, we have

$$0<\lim_{\lambda\to\infty}\sqrt{\lambda}\Big(H_c(\tau_*^\lambda|\tau_*^\lambda)-H_c(\tau_*|\tau_*)\Big)=h_0e^{-h_0\tau_*}\times\lim_{\lambda\to\infty}\sqrt{\lambda}(\tau_*-\tau_*^\lambda)<\infty.\tag{EC.37}$$

It then follows from (EC.24), (EC.29) and (EC.35)–(EC.37) that $\lim_{\lambda\to\infty}(s_*^\lambda-\widehat{s}_*^\lambda)/\sqrt{\lambda}>0$, which is a contradiction with the optimality of s_*^λ.

Now consider the subcase

$$\lim_{\lambda\to\infty}\sqrt{\lambda}(\tau_*-\tau_*^\lambda)=\infty.$$

The above argument still works if we replace τ_* by $\tau_*^\lambda+\frac{M}{\sqrt{\lambda}}$ for any $M>0$. Then we again obtain a contradiction with the optimality of s_*^λ. Hence the proof of *Case* (i) is complete.

Case (ii) Similar to case (i) we assume $\lim_{\lambda\to\infty}\sqrt{\lambda}(\tau_*-\tau_*^\lambda)<0$. Now $\omega_{\tau_*^\lambda}=\tau_*$ as τ_*^λ is larger than τ_*. Note that the constraint on waiting time in (44) becomes $\mathbb{P}\left(W^\lambda(\infty)>\omega_{\tau_*^\lambda}\right)$. So the optimal number of servers is

$$s_*^\lambda=\frac{\lambda}{\mu}H_c(\omega_{\tau_*^\lambda}|\tau_*^\lambda)-\frac{\widehat{\beta}_*^\lambda}{\mu}\sqrt{\lambda}+o(\sqrt{\lambda}),\tag{EC.38}$$

where $\widehat{\beta}_*^\lambda$ is the optimal solution to problem given by (EC.30) with replacing $f_{\tau_*^\lambda}(\cdot)$ by $f_{\omega_{\tau_*^\lambda},\tau_*^\lambda}(\cdot)$, where $f_{\omega_{\tau_*^\lambda},\tau_*^\lambda}(x)=\sqrt{\lambda}[H(\omega_{\tau_*^\lambda}+\frac{x}{\sqrt{\lambda}}|\tau_*^\lambda)-H(\omega_{\tau_*^\lambda}|\tau_*^\lambda)]$. Note that $H_c(\omega_{\tau_*^\lambda}|\omega_{\tau_*^\lambda})=H_c(\omega_{\tau_*^\lambda}|\tau_*^\lambda)$, so the difference between $\widehat{s}_*^\lambda$ (see (EC.24)) and s_*^λ lies in the difference between β_* and $\widehat{\beta}_*^\lambda$. If

$$\liminf_{\lambda\to\infty}(\beta_*-\widehat{\beta}_*^\lambda)>0,\tag{EC.39}$$

then $\lim_{\lambda\to\infty}(s_*^\lambda-\widehat{s}_*^\lambda)/\sqrt{\lambda}>0$, which is again a contradiction with the optimality of s_*^λ. Thus to complete the proof for *Case* (ii), it is sufficient to show (EC.39).

To prove (EC.39), again by the continuity and monotonicity of the function given by (EC.23), similar to (EC.26), we have

$$(1-\alpha_1-\alpha_2)\cdot\int_0^{\infty}\exp\Big(-\frac{\rho}{2\sigma^2}\int_0^y[f_{\omega_{\tau_*^\lambda},\tau_*^\lambda}(x)-\widehat{\beta}_*^\lambda]\mathsf{d}x\Big)\mathsf{d}y$$

$$=\alpha_2\int_{-\infty}^{0}\exp\Big(-\frac{\rho}{2\sigma^2}\int_0^y[f_{\omega_{\tau_*^\lambda},\tau_*^\lambda}(x)-\widehat{\beta}_*^\lambda]\mathsf{d}x\Big)\mathsf{d}y. \tag{EC.40}$$

From the definitions of $f_{\omega_{\tau_*^\lambda},\tau_*^\lambda}(x)$ and $H(x|\tau)$, and $\omega_{\tau_*^\lambda}=\tau_*$,

$$f_{\omega_{\tau_*^\lambda},\tau_*^\lambda}(x)=\begin{cases}e^{-h_0\tau_*}\sqrt{\lambda}\Big(1-\exp(-h_0\frac{x}{\sqrt{\lambda}})\Big), & \text{if } -\sqrt{\lambda}\tau_*\le x\le\sqrt{\lambda}(\tau_*^\lambda-\tau_*),\\ e^{-h_0\tau_*}\sqrt{\lambda}\Big(1-\exp(-h_1\frac{(x-\sqrt{\lambda}(\tau_*^\lambda-\tau_*))}{\sqrt{\lambda}}-h_0\frac{\sqrt{\lambda}(\tau_*^\lambda-\tau_*))}{\sqrt{\lambda}})\Big), & \text{if } x>\sqrt{\lambda}(\tau_*^\lambda-\tau_*).\end{cases}$$

For notation simplicity, assume that $\lim_{\lambda\to\infty}\sqrt{\lambda}(\tau_*^\lambda-\tau_*)$ exists and denote it by $\tilde{\tau}$. Then the above equation yields that

$$\lim_{\lambda\to\infty}f_{\omega_{\tau_*^\lambda},\tau_*^\lambda}(x)=\begin{cases}e^{-h_0\tau_*}h_0x, & \text{if } x\le\tilde{\tau},\\ e^{-h_0\tau_*}(h_1(x-\tilde{\tau})+h_0\tilde{\tau}), & \text{if } x>\tilde{\tau}.\end{cases} \tag{EC.41}$$

Combining (EC.40) and (EC.41) yields that

$$\begin{aligned}&(1-\alpha_1-\alpha_2)\cdot\Big[\int_0^{\tilde{\tau}}\exp\Big(-\frac{\rho}{2\sigma^2}\int_0^y[e^{-h_0\tau_*}h_0x-\limsup_{\lambda\to\infty}\widehat{\beta}_*^\lambda]\mathsf{d}x\Big)\mathsf{d}y\\ &+\int_{\tilde{\tau}}^{\infty}\exp\Big(-\frac{\rho}{2\sigma^2}\int_0^{\tilde{\tau}}[e^{-h_0\tau_*}h_0x-\limsup_{\lambda\to\infty}\widehat{\beta}_*^\lambda]\mathsf{d}x-\int_{\tilde{\tau}}^{y}[e^{-h_0\tau_*}(h_1(x-\tilde{\tau})+h_0\tilde{\tau})-\limsup_{\lambda\to\infty}\widehat{\beta}_*^\lambda]\mathsf{d}x\Big)\mathsf{d}y\Big]\\ &=\alpha_2\int_{-\infty}^{0}\exp\Big(-\frac{\rho}{2\sigma^2}\int_0^y[e^{-h_0\tau_*}h_0x-\limsup_{\lambda\to\infty}\widehat{\beta}_*^\lambda]\mathsf{d}x\Big)\mathsf{d}y.\end{aligned}$$

Notice that by (EC.27) and (EC.41), $\lim_{\lambda\to\infty}f_{\omega_{\tau_*^\lambda},\tau_*^\lambda}(x)=h_{\tau^*}(x)$ for $x\le 0$, and $\lim_{\lambda\to\infty}f_{\omega_{\tau_*^\lambda},\tau_*^\lambda}(x)<h_{\tau^*}(x)$ for $x>0$. Thus using the fact $1-\alpha_1-\alpha_2>0$ and $h_0<h_1$, and the definition of β_* (see (EC.26)), we have that $\limsup_{\lambda\to\infty}\widehat{\beta}_*^\lambda<\beta_*$ by the monotonicity of the function given by (EC.23), which is equivalent to (EC.39).

In summary, Cases (i) and (ii) do not hold. Thus, we have $\lim_{\lambda\to\infty}\sqrt{\lambda}(\tau_*-\tau_*^\lambda)=0$. Also from this proof, we see that (45) is the minimal number of servers. Hence, the proof of the proposition is complete. □

EC.3. Several Auxiliary Lemmas

In the following, we establish several technical lemmas which support the proofs of Proposition 1, Theorems 1 and 2.

Define the associated filtration with the nth system by $\{\mathcal{F}_k^n;k\ge 0\}$ by

$$\mathcal{F}_k^n=\sigma\{\tau_{\ell+1}^n,v_\ell^n,u_\ell^n;\ell\le k\}. \tag{EC.42}$$

Then we have

Lemma EC.1. *$\{\sum_{i=1}^{[\lambda_n t]}(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n)):t\ge 0\}$ is a Martingale with respect to the filtration $\{\mathcal{F}_{\lfloor\lambda_n t\rfloor}^n;t\ge 0\}$. Furthermore,*

$$\frac{1}{\lambda_n}\sum_{i=1}^{[\lambda_n t]}(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n))\Rightarrow 0.$$

Proof. First note that for each i, u_{i+1}^n is independent of $\mathcal{F}_i^n$. By Lemma 3.1 in Dai and He (2010), ω_i^n and ω_j^n are $\mathcal{F}_{i-1}^n$-measurable for $j<i$. Also u_j^n is measurable w.r.t. $\mathcal{F}_{i-1}^n$. Since the conditional probability

$$\mathbb{E}[(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n))|\mathcal{F}_{i-1}^n]=0, \tag{EC.43}$$

so $\{\sum_{i=1}^{[\lambda_n t]}(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n)):t\ge 0\}$ is a Martingale with respect to the filtration $\{\mathcal{F}_{\lfloor\lambda_n t\rfloor}^n;t\ge 0\}$. Therefore, we have the first part of the lemma.

Now we prove the second part of the lemma. Clearly, $\{\frac{1}{\lambda_n}\sum_{i=1}^{[\lambda_n t]}(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n)):t\ge 0\}$ is also a Martingale. Its quadratic variation is given by

$$\frac{1}{\lambda_n^2}\sum_{i=1}^{[\lambda_n t]}(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n))^2\le\frac{1}{\lambda_n^2}[\lambda_n t]\to 0.$$

Therefore, the second part is proved. □

Lemma EC.2. *Under the same assumptions as Theorem* 1*, as* $n\to\infty$

$$\widetilde{B}^n\Rightarrow\frac{1}{\sqrt{\rho}}\mathcal{B}, \tag{EC.44}$$

where $\mathcal{B}=\{\mathcal{B}(t):t\ge 0\}$ *is a standard Brownian motion.*

Proof. In view of (EC.17), we first look at the departure process $\{D^n(t):t\ge 0\}$. We introduce the following two diffusion scalings:

$$\widetilde{D}^n(t)=\frac{D^n(t)-s_n\mu t}{\sqrt{\lambda_n}},\ \widetilde{S}^n(t)=\frac{S^n(t)-s_n}{\sqrt{\lambda_n}}.$$

Then

$$\widetilde{B}^n(t)=\widetilde{D}^n(t)+\widetilde{S}^n(t)-\widetilde{S}^n(0). \tag{EC.45}$$

Let $X^n(t)$ denote the total number of customers at time t in the nth system. Then the departure process $D^n(t)$ can be represented as $\mathcal{S}(\int_0^t(X^n(x)\wedge s_n)\mathrm{d}x)$, where $\{\mathcal{S}(t):t\ge 0\}$ is a Poisson process with rate μ. By (10),

$$\sup_{0\le t\le T}\frac{(s_n-X^n(t))^+}{\lambda_n}\Rightarrow 0. \tag{EC.46}$$

This, together with (2), implies

$$\widetilde{D}^n\Rightarrow\frac{1}{\sqrt{\rho}}\mathcal{B}. \tag{EC.47}$$

By the initial condition (18) and (EC.46), we have that the last two terms in (EC.45) will converge to zero. Hence, (EC.44) directly follows from (EC.47). □

Lemma EC.3. *Under the same assumptions as Theorem* 1*, the sequence of stochastic processes* $\{\widetilde{V}^n\}_{n\in\mathbb{N}}$ *is stochastically bounded.*

Proof. It suffices to show that, for any $T>0$ and $\varepsilon>0$, the following holds for all large enough n and M:

$$\mathbb{P}\left\{\sup_{0\le t\le T}|\widetilde{V}^n(t)|\ge M\right\}\le 4\varepsilon.$$

To this end, define

$$\varsigma_1^n=\inf\{t\ge 0:\widetilde{V}^n(t)\ge M\},\quad \varsigma_2^n=\inf\{t\ge 0:\widetilde{V}^n(t)\le -M\},$$
$$\Omega_1^n(M,T)=\{\varsigma_1^n\le\varsigma_2^n,\varsigma_1^n\le T\},\quad \Omega_2^n(M,T)=\{\varsigma_1^n>\varsigma_2^n,\varsigma_2^n\le T\}.$$

Hence we only need to show that, for all large enough n and M,

$$\mathbb{P}\left(\Omega_1^n(M,T)\right)\le 2\varepsilon\quad\text{and}\quad\mathbb{P}\left(\Omega_2^n(M,T)\right)\le 2\varepsilon. \tag{EC.48}$$

We will first consider the event $\Omega_1^n(M,T)$. By the definition of ς_1^n, we must have that $\widetilde{V}^n(\varsigma_1^n)\ge\widetilde{V}^n(\varsigma_1^n-)$. In other words, if $\widetilde{V}^n$ has a jump at ς_1^n, then it must be an upward jump. Since $\widetilde{V}^n(t)\in[-M,\ M]$ on $[0,\varsigma_1^n]$, for any $t\in(0,\varsigma_1^n]$ and small positive $\delta\in(0,t)$, by (16),

$$\begin{aligned}\widetilde{V}^n(t)-\widetilde{V}^n(t-\delta)=&-\frac{\lambda_n}{s_n\mu}\int_{t-\delta}^{t}\sqrt{\lambda_n}\Big(F^n(\omega^n+\frac{\widetilde{V}^n(x-)}{\sqrt{\lambda_n}})-F^n(\omega^n)\Big)\mathsf{d}\bar{\Lambda}^n(x)\\&+\widetilde{Y}^n(\varsigma_1^n)-\widetilde{Y}^n(\varsigma_1^n-\delta).\end{aligned}\tag{EC.49}$$

Since $\widetilde{V}^n(0)$ is stochastically bounded, we can choose M large enough such that

$$\mathbb{P}\left(\Omega_0^n(M)\right)=\mathbb{P}\left(\widetilde{V}^n(0)\le\frac{M}{4}\right)\ge 1-\varepsilon,$$

where $\Omega_0^n(M)$ is defined in the proof of Proposition 1. Define $\varsigma_{12}^n=\sup\{0\le t\le\varsigma_1^n:\widetilde{V}^n(t)\le M/2\}\vee 0$. We know that on the event $\Omega_0^n(M)$, $\varsigma_{12}^n>0$. By the definition of ς_1^n and ς_{12}^n, we clearly have that

$$\widetilde{V}^n(\varsigma_1^n)\ge M,\quad\text{and}\quad\widetilde{V}^n(\varsigma_{12}^n-)\le\frac{M}{2}. \tag{EC.50}$$

Note that the process $\widetilde{V}^n(\cdot)$ is larger than $M/2$ (thus larger than 0) on the interval $[\varsigma_{12}^n,\ \varsigma_1^n]$. By (EC.49) and the fact that $F^n(\omega^n+x)\ge F^n(\omega^n)$ for any $x\ge 0$,

$$\widetilde{V}^n(\varsigma_1^n)-\widetilde{V}^n(\varsigma_{12}^n-)\le\widetilde{Y}^n(\varsigma_1^n)-\widetilde{Y}^n(\varsigma_{12}^n-). \tag{EC.51}$$

By (EC.50) and (EC.51),

$$\mathbb{P}\Big(\Omega_0^n(M)\cap\Omega_1^n(M,T)\Big)\le\mathbb{P}\Big(\sup_{t\in[0,T]}\Big|\widetilde{Y}^n(t)\Big|\ge\frac{M}{4}\Big). \tag{EC.52}$$

We now prove the stochastic boundedness of $\widetilde{Y}^n$. Recall the definition of $\widetilde{Y}^n$ in (17). The first and the third term on the right side of (17) is stochastically bounded by Conditions (1), (2) and (3). The last two terms are stochastically bounded by Lemma EC.2. It now remains to show the stochastic boundedness of the third term, which can be written as $\frac{1}{\sqrt{\lambda_n}}\sum_{i=1}^{\Lambda^n(t)}(1_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n))$. According to Condition (1), it is enough to show the stochastic boundedness of

$\frac{1}{\sqrt{\lambda_n}}\sum_{i=1}^{[\lambda_n t]}(1_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n))$. From Doob's inequality for martingale (see Lemma EC.1), for any $M\ge 0$,

$$\mathbb{P}\Big(\sup_{0\le t\le T}\Big|\frac{1}{\sqrt{\lambda_n}}\sum_{i=1}^{[\lambda_n t]}(\mathbf{1}_{\{\{u_i^n\le\omega_i^n\}\}}-F^n(\omega_i^n))\Big|\ge M\Big)\le\frac{1}{M^2}\mathbb{E}\Big[\Big|\frac{1}{\sqrt{\lambda_n}}\sum_{i=1}^{[\lambda_n T]}(\mathbf{1}_{\{\{u_i^n\le\omega_i^n\}\}}-F^n(\omega_i^n))\Big|\Big]^2$$
$$=\frac{1}{M^2\lambda_n}\sum_{i=1}^{[\lambda_n T]}\mathbb{E}(\mathbf{1}_{\{\{u_i^n\le\omega_i^n\}\}}-F^n(\omega_i^n))^2\le M^{-2}T.$$

Using the stochastic boundedness of $\{\widetilde{Y}^n, n\ge 1\}$, we can choose M large enough such that the probability on the right-hand side of (EC.52) is less than ε. So we have that $\mathbb{P}(\Omega_1^n(M,T))\le 2\varepsilon$ for large enough M. A symmetric argument shows that $\mathbb{P}(\Omega_2^n(M,T))\le 2\varepsilon$ for large enough M. So we have proved stochastic boundedness. □

Lemma EC.4. *Under the same assumptions as Theorem* 1, *as* $n\to\infty$

$$\widetilde{H}^n(\cdot)\Rightarrow(1/\rho)\sqrt{\rho-1}\mathcal{B}_A(\cdot). \tag{EC.53}$$

Here $\widetilde{H}^n(\cdot)$ *is given by* (15) *and* $\mathcal{B}_A=\{\mathcal{B}_A(t):t\ge 0\}$ *is a standard Brownian motion which is independent of* $\{\mathcal{B}(t):t\ge 0\}$.

Proof. We first prove a convergence result for the sequence of processes given by $\{\frac{1}{\sqrt{\lambda_n}}\sum_{i=1}^{[\lambda_n t]}(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n)):t\ge 0\}$. By Lemma EC.1, the quadratic variation of martingale $\{\frac{1}{\sqrt{\lambda_n}}\sum_{i=1}^{[\lambda_n t]}(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n)):t\ge 0\}$ is

$$\frac{1}{\lambda_n}\sum_{i=1}^{[\lambda_n t]}(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n))^2.$$

We calculate it in the following:

$$\begin{aligned}
&\mathbb{E}\Big[\frac{1}{\lambda_n}\sum_{i=1}^{[\lambda_n t]}\big((\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n))^2-F^n(\omega_i^n)F_c^n(\omega_i^n)\big)\Big]^2\\
&=\mathbb{E}\Big[\frac{1}{\lambda_n}\sum_{i=1}^{[\lambda_n t]}\big((\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n))(F_c^n(\omega_i^n)-F^n(\omega_i^n))\big)\Big]^2\\
&=\frac{2}{(\lambda_n)^2}\sum_{1\le j<i\le[\lambda_n t]}\mathbb{E}\Big[(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n))(F_c^n(\omega_i^n)-F^n(\omega_i^n))\\
&\qquad\cdot(\mathbf{1}_{\{u_j^n\le\omega_j^n\}}-F^n(\omega_j^n))(F_c^n(\omega_j^n)-F^n(\omega_j^n))\Big]\\
&\quad+\frac{1}{(\lambda_n)^2}\sum_{i=1}^{[\lambda_n t]}\mathbb{E}\big((\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n))(F_c^n(\omega_i^n)-F^n(\omega_i^n))\big)^2.
\end{aligned}\tag{EC.54}$$

Then by conditioning on $\mathcal{F}_{i-1}^n$, we have

$$\begin{aligned}
&\mathbb{E}\left[(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n))(F_c^n(\omega_i^n)-F^n(\omega_i^n))(\mathbf{1}_{\{u_j^n\le\omega_j^n\}}-F^n(\omega_j^n))(F_c^n(\omega_j^n)-F^n(\omega_j^n))\right]\\
&=\mathbb{E}\left[\mathbb{E}[(\mathbf{1}_{\{u_i^n\le\omega_i^n\}}-F^n(\omega_i^n))|\mathcal{F}_{i-1}^n](F_c^n(\omega_i^n)-F^n(\omega_i^n))(\mathbf{1}_{\{u_j^n\le\omega_j^n\}}-F^n(\omega_j^n))(F_c^n(\omega_j^n)-F^n(\omega_j^n))\right].
\end{aligned}$$

By (EC.43) the first term on the right-hand side of (EC.54) is 0. Note that the second term on the right-hand side of (EC.54) converges to 0 as $n \to \infty$, so the expectation on the left-hand side of (EC.54) must converge to 0. As a result,

$$\frac{1}{\lambda_n}\sum_{i=1}^{[\lambda_n t]}\left((\mathbf{1}_{\{u_i^n \le \omega_i^n\}} - F^n(\omega_i^n))^2 - F^n(\omega_i^n)F_c^n(\omega_i^n)\right) \Rightarrow 0.$$

On the other hand, on the event given by $\{\Lambda^n(t+1) \ge \lambda_n t\}$

$$\begin{aligned}
&\frac{1}{\lambda_n}\sum_{i=1}^{[\lambda_n t]}(F^n(\omega_i^n)F_c^n(\omega_i^n) - F^n(\omega^n)F_c^n(\omega^n)) \\
&= \frac{1}{\lambda_n}\sum_{i=1}^{[\lambda_n t]}\Big([F^n(\omega_i^n) - F^n(\omega^n)]\,F_c^n(\omega_i^n) + F^n(\omega^n)\,[F_c^n(\omega_i^n) - F_c^n(\omega^n)]\Big) \\
&\le \sup_{0\le s\le t+1} 2\,|F^n(V^n(s)) - F^n(\omega^n)|\,t \\
&= \frac{2}{\sqrt{\lambda_n}}\sup_{0\le s\le t+1}\left|\sqrt{\lambda_n}\left(F^n(\omega^n + \frac{1}{\sqrt{\lambda_n}}\widetilde{V}^n(s)) - F^n(\omega^n)\right)\right| t,
\end{aligned}$$

which vanishes to 0 following from Condition (4) and the stochastic boundedness of $\{\widetilde{V}^n\}_{n\in\mathbb{N}}$ given by Lemma EC.3. Note that (2) and (3) imply that for any $T > 0$, as $n \to \infty$,

$$F^n(\omega^n) \to 1 - \frac{1}{\rho} \quad \text{and} \quad \mathbb{P}\Big(\inf_{0\le t\le T}(\Lambda^n(t+1) - \lambda_n t) \ge 0\Big) \to 1.$$

As a result,

$$\frac{1}{\lambda_n}\sum_{i=1}^{[\lambda_n t]}(\mathbf{1}_{\{u_i^n \le \omega_i^n\}} - F^n(\omega_i^n))^2 \Rightarrow \lim_{n\to\infty} F^n(\omega^n)F_c^n(\omega^n)t = \frac{\rho-1}{\rho^2}t.$$

Then from the martingale convergence theorem (Theorem 8.1 (ii) of Pang et al. (2007)), we know that the sequence of the processes given by $\{\frac{1}{\sqrt{\lambda_n}}\sum_{i=1}^{[\lambda_n t]}(\mathbf{1}_{\{u_i^n \le \omega_i^n\}} - F^n(\omega_i^n)) : t \ge 0\}$ weakly converges to the process $\frac{\sqrt{\rho-1}}{\rho}\mathcal{B}_A$. The result of this lemma then follows from the random-time-change theorem. □

Lemma EC.5. *Under the same assumptions as Theorem* 1*, the sequence of stochastic processes* $\{\widetilde{V}^n\}_{n\in\mathbb{N}}$ *is tight.*

Proof. In view of Lemma EC.3, it suffices to study the modulus of continuity for $\{\widetilde{V}^n\}_{n\in\mathbb{N}}$. By (1), for any $\varepsilon > 0$,

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}\Big(\sup_{\substack{s,t\in[0,T]\\|s-t|<\delta}}|\widetilde{\Lambda}^n(s) - \widetilde{\Lambda}^n(t)| > \varepsilon\Big) = 0. \tag{EC.55}$$

By Conditions (1)–(3), and Lemma EC.4, for any $\varepsilon > 0$,

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}\Big(\sup_{\substack{s,t\in[0,T]\\|s-t|<\delta}}|\widetilde{Y}^n(s) - \widetilde{Y}^n(t)| > \varepsilon\Big) = 0. \tag{EC.56}$$

Let $\Omega_c^n(M,T)$ be the complement of $\Omega_1^n(M,T) \cup \Omega_2^n(M,T)$ which are given in the proof of Lemma EC.3. Then

$$\lim_{M\to\infty}\liminf_{n\to\infty}\mathbb{P}(\Omega_c^n(M,T)) = 1. \tag{EC.57}$$

On the event $\Omega_c^n(M,T)$, it follows from (EC.49) that

$$\left|\widetilde{V}^n(t) - \widetilde{V}^n(t-\delta)\right| \le C_M^n \cdot \left(\bar{\Lambda}^n(t) - \bar{\Lambda}^n(t-\delta)\right) + \left|\widetilde{Y}^n(t) - \widetilde{Y}^n(t-\delta)\right|,$$

where $C_M^n = \max\left\{\sqrt{\lambda_n}(F^n(\omega^n + \frac{M}{\sqrt{\lambda_n}}) - F^n(\omega^n)), \sqrt{\lambda_n}(F^n(\omega^n) - F^n(\omega^n - \frac{M}{\sqrt{\lambda_n}}))\right\}$. By Condition (4), C_M^n is bounded by a finite number C_M which may depend on M. So for any $M > 0$,

$$\mathbb{P}\Big(\sup_{\substack{s,t\in[0,T]\\|s-t|<\delta}} |\widetilde{V}^n(s) - \widetilde{V}^n(t)| > \varepsilon\Big) \le (1 - \mathbb{P}(\Omega_c^n(M,T))) + \mathbb{P}\Big(\sup_{\substack{s,t\in[0,T]\\|s-t|<\delta}} |\bar{\Lambda}^n(s) - \bar{\Lambda}^n(t)| > \frac{\varepsilon}{2C_M}\Big)$$
$$+\mathbb{P}\Big(\sup_{\substack{s,t\in[0,T]\\|s-t|<\delta}} |\widetilde{Y}^n(s) - \widetilde{Y}^n(t)| > \frac{\varepsilon}{2}\Big).$$

By first letting n go to infinite, then δ to zero and finally M go to infinite, we can show that

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}\Big(\sup_{\substack{s,t\in[0,T]\\|s-t|<\delta}} |\widetilde{V}^n(s) - \widetilde{V}^n(t)| > \varepsilon\Big) = 0.$$

This shows that the modulus of continuity for $\{\widetilde{V}^n\}_{n\in\mathbb{N}}$ will vanish as $n \to \infty$. Hence we have the lemma. □

EC.4. Discussion on the Sequence of $\{\omega^n\}$

As discussed after condition (4) that both β and f_ω depend on the sequence $\{\omega^n\}_{n\in\mathbb{N}}$, one may wonder whether different sequences of $\{\omega^n\}_{n\in\mathbb{N}}$ satisfying conditions (3)–(4) and assumption (18) will give us inconsistent results. (Inconsistence means that arguments based on different sequences of $\{\omega^n\}_{n\in\mathbb{N}}$ may give contradictions.) In this section, we argue that this inconsistence is impossible.

We are given two sequences $\{\omega_{(1)}^n\}_{n\in\mathbb{N}}$ and $\{\omega_{(2)}^n\}_{n\in\mathbb{N}}$ with $\lim_{n\to\infty}\omega_{(1)}^n = \lim_{n\to\infty}\omega_{(2)}^n = \omega$ such that $i = 1, 2$,

$$\frac{\lambda_n F_c^n(\omega_{(i)}^n) - s_n\mu}{\sqrt{\lambda_n}} \to \beta^{(i)}, \tag{EC.58}$$

$$\sqrt{\lambda_n}\Big[F^n(\omega_{(i)}^n + \frac{x}{\sqrt{\lambda_n}}) - F^n(\omega_{(i)}^n)\Big] \to f_\omega^{(i)}(x), \tag{EC.59}$$

$$\sqrt{\lambda_n}\Big(V^n(0) - \omega_{(i)}^n\Big) \Rightarrow \widetilde{V}_0^{(i)}. \tag{EC.60}$$

It directly follows from (EC.60) that

$$\lim_{n\to\infty}\sqrt{\lambda_n}(\omega_{(2)}^n - \omega_{(1)}^n) = x_0 \in \mathbb{R}. \tag{EC.61}$$

We first look at the case of the virtual waiting time. Define $\widetilde{V}_1^n(t) = \sqrt{\lambda_n}(V^n(t) - \omega_{(1)}^n)$ and $\widetilde{V}_2^n(t) = \sqrt{\lambda_n}(V^n(t) - \omega_{(2)}^n)$. Then

$$\widetilde{V}_2^n(t) = \widetilde{V}_1^n(t) + \sqrt{\lambda^n}(\omega_{(1)}^n - \omega_{(2)}^n). \tag{EC.62}$$

It directly follows (EC.61)–(EC.62) that

$$\text{if } \widetilde{V}_i^n \Rightarrow \widetilde{V}^{(i)}, \text{ then } \widetilde{V}_{3-i}^n \Rightarrow \widetilde{V}^{(3-i)} \text{ with } \widetilde{V}^{(2)} = \widetilde{V}^{(1)} - x_0,\ \ i = 1, 2. \tag{EC.63}$$

To prove the results for two sequences $\{\omega^n_{(1)}\}_{n\in\mathbb{N}}$ and $\{\omega^n_{(2)}\}_{n\in\mathbb{N}}$ to be consistent, it is sufficient to show that the diffusion approximations for $\omega^n_{(1)}$ and $\omega^n_{(2)}$ given by Theorem 1 also satisfy (EC.63). To this end, note that (EC.61) together with (EC.58)–(EC.59) implies that

$$\beta^{(2)} = -f^{(1)}_\omega(x_0) + \beta^{(1)}, \quad f^{(2)}_\omega(x) = f^{(1)}_\omega(x_0 + x) - f^{(1)}_\omega(x_0). \tag{EC.64}$$

By (EC.58)–(EC.60) and Theorem 1, we have, as $n \to \infty$

$$\widetilde{V}^n_i \Rightarrow \widetilde{V}^{(i)}, \tag{EC.65}$$

where

$$\widetilde{V}^{(i)}(t) = \widetilde{V}^{(i)}(0) - \rho \int_0^t \left[f^{(i)}_\omega(\widetilde{V}^{(i)}(x)) - \beta^{(i)} \right] \mathrm{d}x + \left[\widetilde{\Lambda}(t) - \sqrt{\rho}\mathcal{B}(t) - \sqrt{\rho - 1}\mathcal{B}_A(t) \right]. \tag{EC.66}$$

In view of (EC.64), it follows from (EC.66) that

$$\widetilde{V}^{(2)} = \widetilde{V}^{(1)} - x_0. \tag{EC.67}$$

Therefore, different choices of $\{\omega^n\}_{n\in\mathbb{N}}$ do not give us inconsistent results based on Theorem 1 when conditions (1)–(4) and assumption (18) hold!

Similarly, from Theorem 2, we can prove that different choices of $\{\omega^n\}_{n\in\mathbb{N}}$ also do not give us inconsistent results for the diffusion approximations for the queue length process.

EC.5. Discussion on the Initial State

We discuss Assumption (18) on the initial state. Usually, the initial state is given by the queue length and patience times; see Liu and Whitt (2014), Mandelbaum and Momčilović (2012), and Reed and Tezcan (2012). In the following, we hence provide a sufficient condition for (18) in terms of queue length and patience times. Analysis for what general initial conditions imply (18) is left for future research. To the best of our knowledge, the first work focusing on the initial state is Aras et al. (2017), which studied the impact of initial content (e.g., initial age process) on the system performances.

Lemma EC.6. *Denote by $Q^n(0)$ the number of customers who are initially in queue. Assume those initial customers in queue are infinitely patient. If $\frac{Q^n(0) - s_n\mu\omega^n}{\sqrt{\lambda_n}} \Rightarrow \widetilde{Q}_0$ for a random variable $\widetilde{Q}_0$ and $\omega^n \to \omega$, then as $n \to \infty$,*

$$\widetilde{V}^n(0) \Rightarrow \widetilde{V}_0, \tag{EC.68}$$

where $\widetilde{V}_0 = \rho(\widetilde{Q}_0 - \sqrt{\frac{\omega}{\rho}}\mathcal{N})$ with $\mathcal{N}$ being a standard normal random variable independent of $\widetilde{Q}_0$.

Proof. Recall that $D^n(t)$ is the number of departures through service completion by time t. As all customers initially in queue will eventually receive service, the virtual waiting time $V^n(0)$ satisfies

$$D^n(V^n(0)-) \le Q^n(0) < D^n(V^n(0)).$$

As D^n is a Poisson process with rate $s_n\mu$ on $[0, V^n(0)]$, with probability one, there is only one departure at time $V^n(0)$. This gives

$$D^n(V^n(0)) = Q^n(0) + 1. \tag{EC.69}$$

Dividing both sides by λ_n, then one can see that

$$V^n(0) \Rightarrow \omega, \quad \text{as} \quad n \to \infty.$$

From (EC.69), we get

$$\widetilde{V}^n(0) = \frac{\lambda_n}{s_n\mu}\left(\frac{Q^n(0) - s_n\mu\omega^n}{\sqrt{\lambda_n}} - \frac{D^n(V^n(0)) - s_n\mu V^n(0)}{\sqrt{\lambda_n}} + \frac{1}{\sqrt{\lambda_n}}\right).$$

With $V^n(0) \Rightarrow \omega$, we have $\frac{D^n(V^n(0)) - s_n\mu V^n(0)}{\sqrt{\lambda_n}} \Rightarrow \sqrt{\frac{\omega}{\rho}}\mathcal{N}$; here $\mathcal{N}$ follows the standard normal distribution. Together with the assumption that $\frac{Q^n(0) - s_n\mu\omega^n}{\sqrt{\lambda_n}} \Rightarrow \widetilde{Q}_0$, we get the convergence of $\widetilde{V}^n(0)$. This completes the proof. □

EC.6. Proving the Convergence of the First Term in (EC.5)

To analyze the first term on the right-hand side of (EC.5), we need a modification of the proof of Lemma 5.3 in Krichagina and Puhalskii (1997). Such a modification is needed because we allow the distribution $F^n(\cdot)$ to vary with n while Lemma 5.3 in Krichagina and Puhalskii (1997) only deals with a fixed $F(\cdot)$ (i.e., $F^n(\cdot) \equiv F(\cdot)$ for all n). We now demonstrate how to modify their proof to allow the distribution $F^n(\cdot)$ to vary with n. For a function $g(\cdot)$, let $g(x-)$ denote its left-hand limit at x.

To make the connection easy, we adopt the same notation as theirs without conflicting with the notation already used in the above. We denote the first term on the right-hand side of our (EC.5) by

$$M_2^n(t) := \frac{1}{\sqrt{\lambda_n}} \sum_{i=\Lambda^n(t-\omega^n)+1}^{\Lambda^n(t)} \left(\mathbf{1}_{\{u_i^n + \tau_i^n > t\}} - F_c^n(t - \tau_i^n)\right).$$

Define $v_i^n = u_i^n \wedge \omega^n$ and let $F_\wedge^n(\cdot)$ be the distribution of v_1^n. Let u_1 be a random variable with distribution $F(\cdot)$ and $F_\wedge(\cdot)$ be the distribution of $u_1 \wedge \omega$. Define U^n as in their (2.23) but change n to λ_n. Change V^n in their (3.24) to $V^n(t,x) = U^n(\bar{\Lambda}^n(t), F_\wedge^n(x)), t \geq 0, x \geq 0$, and $L^n(t,x)$ in their (3.18) to

$$L^n(t,x) = \frac{1}{\sqrt{\lambda_n}} \sum_{i=1}^{\Lambda^n(t)} \left[\mathbf{1}_{\{v_i^n \leq x\}} - \int_0^{x \wedge v_i^n} \frac{\mathsf{d}F_\wedge^n(y)}{1 - F_\wedge^n(y-)}\right].$$

Then following the same argument leading to their (3.19), we have

$$V^n(t,x) = -\int_0^x \frac{V^n(t,y-)}{1 - F_\wedge^n(y-)} \mathsf{d}F_\wedge^n(y) + L^n(t,x), \quad t \geq 0, x \geq 0.$$

Also, following the same argument leading to their (3.20), we have

$$M_2^n(t) = G^n(t) - H^n(t),$$

where

$$G^n(t)=\int_0^t \frac{V^n(t-x,x-)}{1-F^n_\wedge(x-)}\mathsf{d}F^n_\wedge(x),$$
$$H^n(t)=\frac{1}{\sqrt{\lambda_n}}\sum_{i=1}^{\Lambda^n(t)}\Big(\mathbf{1}_{\{v_i^n\le t-\tau_i^n\}}-\int_0^{v_i^n\wedge(t-\tau_i^n)^+}\frac{\mathsf{d}F^n_\wedge(u)}{1-F^n_\wedge(u-)}\Big).$$

In Krichagina and Puhalskii (1997), the proof of Lemma 5.3, which shows the convergence of $\{M_2^n(t)\}$, needs (i) the tightness of $\{M_2^n\}_{n\in\mathbb{N}}$ with $M_2^n=\{M_2^n(t):t\ge 0\}$ (Lemma 3.8) and (ii) finite dimensional distribution convergence of $\{M_2^n\}_{n\in\mathbb{N}}$. In the following we show (i) and (ii) separately.

(i) Tightness. Lemma 3.8 is essentially a summary of Lemmas 3.4 (tightness of $\{G^n\}_{n\in\mathbb{N}}$ with $G^n=\{G^n(t):t\ge 0\}$) and 3.7 (tightness of $\{H^n\}_{n\in\mathbb{N}}$ with $H^n=\{H^n(t):t\ge 0\}$), both of which require Lemmas 3.1 and 3.2 and an additional analysis. We can directly use their Lemma 3.1 because it is for uniform distribution and does not involve general distributions. In our model, $\bar{\Lambda}^n$ plays the role of a^n in their Lemma 3.2. Our assumption (1) enables us to use their Lemma 3.2.

We now show that the conclusion of their Lemma 3.4 (tightness of $\{G^n\}_{n\in\mathbb{N}}$) holds for our model. Let $T_0=\sup_n\omega^n+1$. Let $T_0=\sup_n\omega^n+1$, and define piecewise-linear function $\psi^n(\cdot)$ such that its graph linearly connects $(0,0)$, (ω_1^n,ω_1^n), (ω,ω^n) and (T_0,T_0) (that is, $\psi^n(0)=0$, $\psi^n(\omega_1^n)=\omega_1^n$, $\psi^n(\omega)=\omega^n$ and $\psi^n(T_0)=T_0$), where

$$\omega_1^n=\begin{cases}\omega^n-d_n, \text{ if } \omega^n\le\omega,\\ \omega-d_n, \quad \text{if } \omega^n>\omega\end{cases}\quad \text{with}\quad \begin{cases}F^n_\wedge(\omega^n-)-F^n_\wedge(\omega_1^n)<\frac{1}{\lambda_n^{1/4}}, \text{ if } \omega^n\le\omega,\\ F_\wedge(\omega-)-F_\wedge(\omega_1^n)<\frac{1}{\lambda_n^{1/4}}, \quad \text{if } \omega^n>\omega,\end{cases}$$

and $d_n\in(0,1/\lambda_n^{1/4})$. Assumptions (3)–(4) and assumption $F(\omega-)=1-1/\rho$ imply that $F^n_\wedge(\omega^n-)-F_\wedge(\omega-)\to 0$ as $n\to\infty$. Note that $\psi^n(\omega)=\omega^n$, then $\Delta F^n_\wedge(\psi^n(\omega))\to\Delta F_\wedge(\omega)$. (For a function f, $\Delta f(t):=f(t)-f(t-)$.) Then, $\sup_{0\le t\le T_0}|\psi^n(t)-t|\le|\omega^n-\omega|\to 0$ and $\sup_{0\le t\le T_0}|F^n_\wedge(\psi^n(t))-F_\wedge(t)|\le\frac{1}{\lambda_n^{1/4}}+|F^n_\wedge(\omega^n-)-F_\wedge(\omega-)|+\sup_{0\le t\le\omega}|F^n(t)-F(t)|\to 0$ as $n\to\infty$. One can also verify that the probability distribution $F^n_\wedge(\psi^n(\cdot))$ converges to $F_\wedge(\cdot)$ in total variation. Indeed, for any measurable set A,

$$\begin{aligned}|F^n_\wedge(\psi^n(A))-F_\wedge(A)|\le&|F^n_\wedge(\psi^n(A,x\le\omega_d^n))-F_\wedge(A,x\le\omega_d^n)|\\ &+|F^n_\wedge(\psi^n(A,\omega_d^n<x<\omega))-F_\wedge(A,\omega_d^n<x<\omega)|\\ &+|F^n_\wedge(\psi^n(A,x\ge\omega))-F_\wedge(A,x\ge\omega)|\\ \le&|F^n_\wedge(A,x\le\omega_d^n)-F_\wedge(A,x\le\omega_d^n)|\\ &+\frac{1}{\lambda_n^{1/4}}+|F^n_\wedge(\omega^n-)-F_\wedge(\omega-)|+|F^n_\wedge(\omega_d^n)-F_\wedge(\omega_d^n)|\\ &+|\Delta F^n_\wedge(\psi^n(\omega))-\Delta F_\wedge(\omega)|.\end{aligned}$$

Then the convergence in total variation of $F^n_\wedge(\psi^n(\cdot))$ to $F_\wedge$ follows from the convergence in total variation of $F^n_\wedge$ to $F_\wedge$ on $[0,\omega)$ (from the condition (21)), $F^n(\omega^n)\to F(\omega)$ and the fact that $|\Delta F^n_\wedge(\psi^n(\omega))-\Delta F_\wedge(\omega)|\to 0$.

Let $\mathbf{D}([0,\infty),\mathbf{D}[0,\infty))$ denote the space of all $\mathbf{D}[0,\infty)$-valued right continuous functions with left limits defined on $[0,\infty)$; see Talreja and Whitt (2009) for detailed analysis on this space. Introduce a sequence of mappings $\Psi^n:\mathbf{D}([0,\infty),\mathbf{D}[0,\infty))\to\mathbf{D}[0,\infty)$ and $\Psi:\mathbf{D}([0,\infty),\mathbf{D}[0,\infty))\to\mathbf{D}[0,\infty)$ by

$$\begin{aligned}\Psi^n(z)(t)&=\int_0^t\frac{z(\psi^n(t)-\psi^n(x),F^n_\wedge(\psi^n(x)-))}{1-F^n_\wedge(\psi^n(x)-)}\mathsf{d}F^n_\wedge(\psi^n(x)),\quad t\ge 0,\\ \Psi(z)(t)&=\int_0^t\frac{z(t-x,F_\wedge(x-))}{1-F_\wedge(x-)}\mathsf{d}F_\wedge(x),\quad t\ge 0,\end{aligned}$$

for any $z\in\mathbf{D}([0,\infty),\mathbf{D}[0,\infty))$. Then for any $T\ge 0$,

$$\begin{aligned}&\sup_{0\le t\le T}|\Psi^n(z^n)(t)-\Psi(z)(t)|\\ &\le\sup_{0\le t\le T}\left|\int_0^t\frac{z^n(\psi^n(t)-\psi^n(x),F^n_\wedge(\psi^n(x)-))}{1-F^n_\wedge(\psi^n(x)-)}\mathsf{d}\left(F^n_\wedge(\psi^n(x))-F_\wedge(x)\right)\right|\\ &\quad+\sup_{0\le t\le T}\left|\int_0^t\left(\frac{z^n(\psi^n(t)-\psi^n(x),F^n_\wedge(\psi^n(x)-))}{1-F^n_\wedge(\psi^n(x)-)}-\frac{z(t-x,F_\wedge(x-))}{1-F_\wedge(x-)}\right)\mathsf{d}F_\wedge(x)\right|.\end{aligned}$$

The first term on the right-hand side in the above converges to zero because $F^n_\wedge(\psi^n(\cdot))$ converges to $F_\wedge(\cdot)$ in total variation. If z^n converges to z which is continuous in both variables, then the second term converges to zero because

$$\sup_{0\le t\le T}\left|\psi^n(t)-t\right|\to 0\text{ and }\sup_{0\le t\le T}\left|F^n_\wedge(\psi^n(t)-)-F_\wedge(t-)\right|\to 0\text{ as }n\to\infty.$$

Note that

$$G^n(\psi^n(t))=\int_0^{\psi^n(t)}\frac{V^n(\psi^n(t)-x,x-)}{1-F^n_\wedge(x-)}\mathsf{d}F^n_\wedge(x)=\int_0^t\frac{V^n(\psi^n(t)-\psi^n(x),\psi^n(x)-)}{1-F^n_\wedge(\psi^n(x)-)}\mathsf{d}F^n_\wedge(\psi^n(x))$$

can be represented as $G^n(\psi^n(t))=\Psi^n(U^n(\bar\Lambda^n(\cdot),\cdot))(t)$. From the fact that $U^n(\bar\Lambda^n(\cdot),\cdot)$ converges to a function which is continuous in both variables, then similar to the argument below (3.26) in Krichagina and Puhalskii (1997) we know that $\{G^n\}_{n\in\mathbb{N}}$ is C-tight, i.e., the conclusion in their Lemma 3.4 holds for our model.

We next show that the conclusion of their Lemma 3.7 (tightness of $\{H^n\}_{n\in\mathbb{N}}$) holds for our model. Define H^n_k the same as their (3.28) with changing their $F(\cdot)$ to $F^n_\wedge(\cdot)$. The conclusion of their Lemma 3.5, which shows that H^n_k is a square-integrable martingale, still holds because it is for a specific system and no limit is involved. Based on Lemma 3.5, the arguments for proving their Lemma 3.7 still work after we change their $F(\cdot)$ to $F^n_\wedge(\cdot)$ (i.e., for our model). This is because the arguments for their (3.60) and thereafter still hold for $F^n_\wedge(\cdot)$, except that we need to use the weak law of large numbers for the triangular array $\frac{1}{\lambda_n}\sum_{i=1}^{[\lambda_n t]}\int_0^{v_i^n}\frac{\mathsf{d}F^n_\wedge(u)}{1-F^n_\wedge(u-)}$. The weak law of large numbers for triangular array can be applied because we can bound its second moment:

$$\mathbb{E}\left[\Big(\int_0^{v_i^n}\frac{\mathsf{d}F^n_\wedge(x)}{1-F^n_\wedge(x-)}\Big)^2\right]=\int_0^\infty\Big(\int_0^x\frac{\mathsf{d}F^n_\wedge(x_1)}{1-F^n_\wedge(x_1-)}\Big)^2\mathsf{d}F^n_\wedge(x)$$

$$
\begin{aligned}
&= \int_0^\infty \int_0^\infty \int_0^\infty \mathbf{1}_{\{x_1 \vee x_2 \le x\}} \frac{\mathsf{d}F_\wedge^n(x_1)}{1-F_\wedge^n(x_1-)} \frac{\mathsf{d}F_\wedge^n(x_2)}{1-F_\wedge^n(x_2-)} \mathsf{d}F_\wedge^n(x) \\
&= \int_0^\infty \int_0^\infty \int_0^\infty \mathbf{1}_{\{x_1 \vee x_2 \le x\}} \mathsf{d}F_\wedge^n(x) \frac{\mathsf{d}F_\wedge^n(x_1)}{1-F_\wedge^n(x_1-)} \frac{\mathsf{d}F_\wedge^n(x_2)}{1-F_\wedge^n(x_2-)} \\
&= \int_0^\infty \int_0^\infty \frac{1-F_\wedge^n((x_1 \vee x_2)-)}{(1-F_\wedge^n(x_1-))\,(1-F_\wedge^n(x_2-))} \mathsf{d}F_\wedge^n(x_1)\mathsf{d}F_\wedge^n(x_2) \\
&\le 2\int_0^\infty \int_0^\infty \mathbf{1}_{\{x_1 \ge x_2\}} \frac{1-F_\wedge^n((x_1 \vee x_2)-)}{(1-F_\wedge^n(x_1-))\,(1-F_\wedge^n(x_2-))} \mathsf{d}F_\wedge^n(x_1)\mathsf{d}F_\wedge^n(x_2) \\
&= 2\int_0^\infty \int_0^\infty \mathbf{1}_{\{x_1 \ge x_2\}} \frac{1}{1-F_\wedge^n(x_2-)} \mathsf{d}F_\wedge^n(x_1)\mathsf{d}F_\wedge^n(x_2) \\
&= 2\int_0^\infty \frac{1-F_\wedge^n(x_2-)}{1-F_\wedge^n(x_2-)} \mathsf{d}F_\wedge^n(x_2) = 2.
\end{aligned}
$$

(ii) Finite dimensional distribution convergence. Introduce

$$
M_{2k}^n(t) = \sum_{i=1}^k \Box U^n((\bar{\Lambda}^n(s_{i-1}^k), 0), (\bar{\Lambda}^n(s_i^k), F_\wedge^n(t-s_i^k))),
$$

with the operator $\Box$ defined in the same way as the equation below their (2.19). Essentially, we change their $F(\cdot)$ in their (5.18) to $F_\wedge^n(\cdot)$. From the definitions, $F_\wedge^n(x) = F_\wedge(x) = 1$ if $x \ge \omega^n \vee \omega$. As a result, $F_\wedge^n(t) \to F_\wedge(t)$ for $t > \omega$. From (21) and $\omega^n \to \omega$, $F_\wedge^n(t) \to F_\wedge(t)$ for $t < \omega$. As a result, for all $t \ne \omega$, $F_\wedge^n(t) \to F_\wedge(t)$. In the proof of (a) and (b) on their page 270, it is enough to consider $\{s_i^k\}$ such that $t - s_i^k \ne \omega$ for all $i \ge 1$ and k. Then the convergence of their (5.17) holds by also using the convergence of $F_\wedge^n(t-s_i^k)$ to $F_\wedge(t-s_i^k)$ and $\bar{\Lambda}^n(s_{i-1}^k)$ to $\bar{\Lambda}(s_{i-1}^k)$. As a result, their (a) on page 270 still holds. For (b) on their page 270, their (5.20) and the argument on the top of page 273 still works. Thus, we have the convergence of finite dimensional distribution.

Combining (i) and (ii), we have shown that their Lemma 5.3 also holds for our model.

References

Aras, A. K., Y. Liu, and W. Whitt (2017). Heavy-traffic limit for the initial content process. *Stoch. Syst. 7*(1), 95–142.

Dai, J. G. (1995). On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab. 5*(1), 49–77.

Dai, J. G. and W. Dai (1999). A heavy traffic limit theorem for a class of open queueing networks with finite buffers. *Queueing Syst. 32*(1-3), 5–40.

Dai, J. G. and S. He (2010). Customer abandonment in many-server queues. *Math. Oper. Res. 35*(2), 347–362.

Ethier, S. N. and T. G. Kurtz (1986). *Markov processes.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc.

Karatzas, I. and S. E. Shreve (1991). *Brownian motion and stochastic calculus*, Volume 113 of *Graduate Texts in Mathematics.* New York: Springer-Verlag.

Krichagina, E. V. and A. A. Puhalskii (1997). A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Syst. 25*(1-4), 235–280.

Liu, Y. and W. Whitt (2014). Many-server heavy-traffic limit for queues with time-varying parameters. *Ann. Appl. Probab. 24*(1), 378–421.

Mandelbaum, A. and P. Momčilović (2012). Queues with many servers and impatient customers. *Math. Oper. Res. 37*(1), 41–65.

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

Reed, J. E. and T. Tezcan (2012). Hazard rate scaling of the abandonment distribution for the $GI/M/n+GI$ queue in heavy traffic. *Oper. Res. 60*(4), 981–995.

Talreja, R. and W. Whitt (2009). Heavy-traffic limits for waiting times in many-server queues with abandonment. *Ann. Appl. Probab. 19*(6), 2137–2175.

Whitt, W. (1980). Some useful functions for functional limit theorems. *Math. Oper. Res. 5*(1), 67–85.