ONLINE APPENDIX
to
Many-Server Gaussian Limits for Overloaded Non-Markovian Queues with Customer Abandonment

A. Korhan Aras
SAS Institute, Cary, NC. akaras@ncsu.edu

Xinyun Chen
School of Economics and Management, Wuhan University, Wuhan, China
xinyun.chen@whu.edu.cn

Yunan Liu
Department of Industrial and Systems Engineering, North Carolina State University
yliu48@ncsu.edu

February 25, 2018

This appendix supplements the main paper [1] by providing additional materials. In §1 we review the Gronwall's inequality. In §2 we provide simulation results to confirm the effectiveness of the FCLT-based performance formulas. In §3 we extend our results to queues having positive initial queue content (with existing initial customers waiting in line at time 0).

1 Gronwall's Inequality

Gronwall's inequality is used in the proofs of the main results (See §5). We review Gronwall's inequality below.

Lemma 1.1 (Gronwall's Inequality)**.** *Consider measurable functions $x, h \geq 0 : [0,T] \to [0,\infty)$ and a locally-finite nonnegative measure μ on $[0,T]$. If*

$$x(t) \leq h(t) + \int_0^t x(u)\,\mu(u)du \quad \textit{with} \quad \int_0^T h(u)\mu(u)du < \infty,$$

then

$$x(t) \leq h(t) + \int_0^t h(u)e^{\left(\int_u^t \mu(r)\,dr\right)}\mu(u)du. \tag{1.1}$$

See [9] for a reference. Also see [5].

2 Numerical Examples

In this section, we provide numerical examples to demonstrate the effectiveness of the engineering formulas (based on the variance and covariance formulae) given in §4.3 of [1]. We provide simulation comparisons for the steady-state performance of the $G/GI/n+GI$ model.

Specifically, we consider the $H_2(\lambda^{-1}, c_\lambda^2)/GI/n+GI$ model having H_2 interarrival times, $n = 100$ servers, phase-type (PH) service times and PH abandonment times. Our PH distributions include H_2, M and E_2 distributions, representing high variability (with SCV 4), medium variability (with SCV 1) and low variability (with SCV 0.5). In Tables 1 and 2, we report simulation estimations of the means and variances of the steady-state waiting time W and queue length Q.

We observe that the service-time and abandonment-time distributions do not have an impact on the steady-state mean values. However, they do have a significant impact on the steady-state variances (and distributions) of W and Q. Our results show that our FCLT-based performance formulas provide accurate approximations for the desired performance.

Table 1: $H_2(\lambda^{-1}, c_\lambda^2)/GI/100 + H_2(\theta^{-1}, c_a^2)$ with $(\lambda, \rho, \theta, c_\lambda^2, c_a^2) = (120, 1.2, 0.5, 4, 4)$

	E_2 service ($c_s^2 = 0.5$)		M service ($c_s^2 = 1$)		H_2 service ($c_s^2 = 4$)	
Perf.	Sim	Num	Sim	Num	Sim	Num
$\mathbb{E}[W]$	0.237	0.240	0.239	0.237	0.240	0.240
rel. err.	±1.4E-3	1.4%	±1.5E-3	1%	±2E-3	0.8%
Var(W)	0.022	0.022	0.024	0.0245	0.0286	0.0288
rel. err.	±7.5E-4	0.9%	±8.3E-4	2.1%	±1.1E-3	0.7%
$\mathbb{E}[Q]$	26.21	25.84	26.33	26.12	26.40	26.41
rel. err.	±0.168	1.4%	±0.177	0.7%	±0.215	0.0%
Var(Q)	302.78	305.02	320.97	316.83	360.78	343.67
rel. err.	±10.31	4.6%	±12.12	4.7%	±13.65	4.7%

3 The $GI/GI/n+GI$ Queue with Positive Initial Queue Content

In the main paper [1], we have developed the heavy-traffic fluid and diffusion limits for the $GI/GI/n+GI$ queue having special initial conditions. Specifically, in [1], we assume that the queue is initially empty, i.e., $Q_n(0) = W_n(0) = V_n(0) = 0$, and all servers are initially busy with equilibrium service times. We advocate that was adequate because our focus there is to develop steady state performance as $t \to \infty$ (where initial condition becomes asymptotically negligible).

Table 2: $H_2(\lambda^{-1}, c_\lambda^2)/GI/100 + E_2(\theta^{-1})$ with $(\lambda, \rho, \theta, c_\lambda^2) = (120, 1.2, 0.5, 4)$

	E_2 service ($c_s^2 = 0.5$)		M service ($c_s^2 = 1$)		H_2 service ($c_s^2 = 4$)	
Perf.	Sim	Num	Sim	Num	Sim	Num
$\mathbb{E}[W]$	0.705	0.732	0.7046	0.732	0.700	0.732
rel. err.	±3.4E-3	3.7%	±3.7E-3	3.9%	±4.2E-3	4.6%
Var(W)	0.051	0.047	0.0576	0.0532	0.0712	0.0659
rel. err.	±4.5E-3	7.8%	±4.9E-3	7.6%	±5.7E-3	7.5%
$\mathbb{E}[Q]$	79.84	82.32	79.65	82.32	78.83	82.32
rel. err.	±0.427	3.0%	±0.454	3.3%	±0.492	4.4%
Var(Q)	818.86	785.86	883.42	847.02	1070.8	976.20
rel. err.	±67.60	4.0%	±71.62	4.1%	±77.01	8.8%

In this appendix, we give heavy-traffic fluid and diffusion limits for the overloaded $G_t/GI/n+GI$ model having (i) a time-varying arrival rate $\lambda(t)$ and (ii) a positive initial queue content (e.g., $Q_n(0) > 0$, $W_n(0) > 0$). Our focus here is to study the dynamics of initial queue content and its impact on the transient system performance. In §3.1 we give preliminary results for key system performance processes. In §3.2, we develop the heavy-traffic fluid and diffusion limits for the $G_t/GI/n + GI$ queue with positive initial queue content. We provide the proofs in §4.

3.1 Prelimit Processes for Models with Positive Initial Queue Content

In [2], the authors developed the heavy-traffic limits for the $G_t/GI^o, GI^\nu/\infty$ infinite-server queue. Hereby, our performance representations for the $G_t/GI/n + GI$ queue with positive initial queue content build on results in [2]. The key idea is that we model the initial queue content (i.e., number of customers waiting in line) as a "trancated" infinite-server queue with "service times" that are customers' patience times. Paralleling [1], we first discuss the enter-service process.

Enter-service process. First, we give an expression for the $E_n(t)$, the total number of customers who enter service from queue in an OL interval by t. Let $E_n^o(t)$ be the number of initial customers, who were waiting in line at time 0, have entered service in the interval $[0, t]$, and let $E_n^\nu(t)$ be the number of customers who has arrived after time 0 and entered service in the interval $[0, t]$. Then

$$E_n(t) = E_n^o(t) + E_n^\nu(t) \tag{3.1}$$

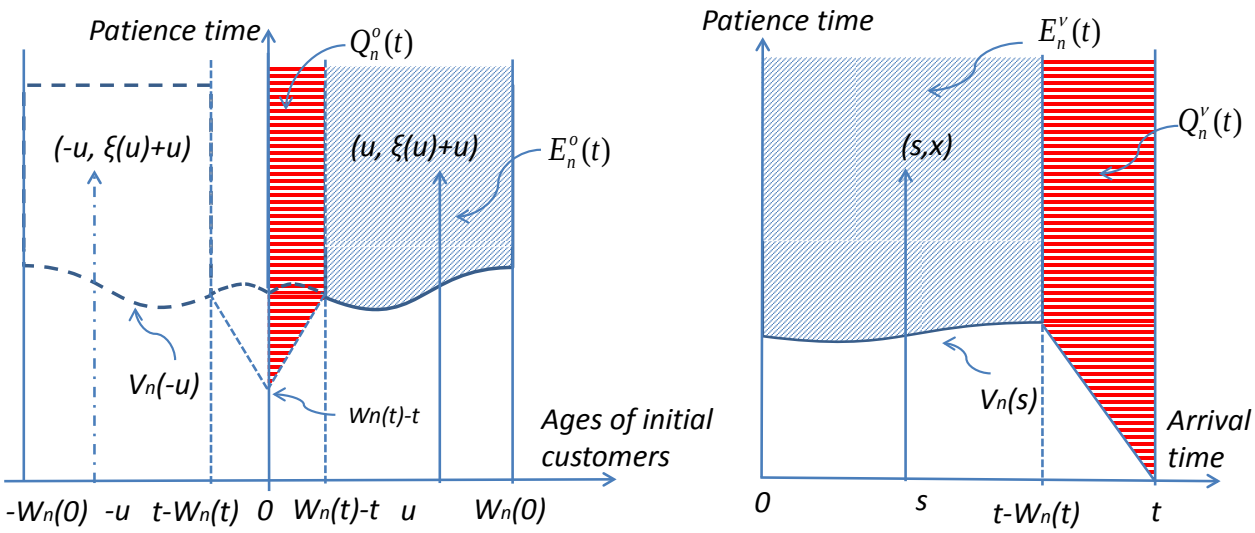

Figure 1: Graphic demonstration of $E_n(t)$.

where $E_n^v(t)$ is discussed in (3.3) in the main paper [1], which can be decomposed into 3 asymptotically independent terms given in (3.4)–(3.6) in [1], and

$$E_n^o(t) \equiv \sum_{i=Q_n(0,(W_n(t)-t)^+)+1}^{Q_n(0,W_n(0))} \mathbf{1}(\zeta_i(\eta_{n,i}) > V_n(-\eta_{n,i}) - \eta_{n,i}) \tag{3.2}$$

$$= n \int_{(W_n(t)-t)^+}^{W_n(0)} \int_0^1 \mathbf{1}(y > H_u(V_n(-u) - u))\, d\bar{U}_n^o(\bar{Q}_n(0,u), y), \tag{3.3}$$

where U_n^o is an independent sequential empirical processes defined as in §3.1. The random variables $0 \le \eta_{n,1} \le \eta_{n,2} \le \cdots$ are the ordered ages (elapsed waiting times) of the initial customers in line, and $\zeta_i(\eta_{n,i})$ is the remaining patience time of customer i who has elapsed waiting time $\eta_{n,i}$.

Remark 3.1 (Understanding the physics of representations (3.3))**.** *We now carefully explain why Equations* (3.2)–(3.3) *hold. Note if there are old customers waiting in line, it must hold that* $W_n(t) > t$*. To understand* (3.2)*, we note that all initial customers with ages greater than* $W_n(t) - t$ *at time 0 have already entered service by time t provided that such a customer does not abandon; hence the upper and lower limits of the sum in* (3.2)*. In order for such a customer not to abandon, we require its "full" patience time* $\zeta_i(\eta_{n,i}) + \eta_{n,i}$ *(sum of the remaining patience time and the age) to be greater than* $V_n(-\eta_{n,i})$*, because this initial customer with age* $\eta_{n,i}$*, satisfying* $0 < W_n(t) - t < \eta_{n,i} \le W_n(0)$*, can intuitively be treated as an arrival at the negative arrival time* $-\eta_{n,i}$*, satisfying* $-W_n(0) \le -\eta_{n,i} < t - W_n(t) < 0$*. See the blue shaded area in the left-hand figure of Figure 1 for an illustration.*

Paralleling the decomposition for $E_n^\nu(t)$ in (3.3) of [1], we have

$$E_n^o(t) = E_{n,1}^o(t) + E_{n,2}^o(t) + E_{n,3}^o(t), \tag{3.4}$$

where

$$E_{n,1}^o(t) \equiv \sqrt{n} \int_{(W_n(t)-t)^+}^{W_n(0)} H_u^c(V_n(-u) - u)\, d\hat{Q}_n(0,u) \tag{3.5}$$

$$E_{n,2}^o(t) \equiv \sqrt{n} \int_{(W_n(t)-t)^+}^{W_n(0)} \int_0^1 \mathbf{1}(y > H_u(V_n(-u) - u))\, d\hat{U}_n^o(\bar{Q}_n(0,u), y) \tag{3.6}$$

$$E_{n,3}^o(t) \equiv n \int_{(W_n(t)-t)^+}^{W_n(0)} H_u^c(V_n(-u) - u)\, dQ_n(0,u). \tag{3.7}$$

Queue-length process. The number of customer waiting in line at time t can be represented as

$$Q_n(t) = Q_n^o(t) + Q_n^\nu(t) \tag{3.8}$$

where $Q_n^\nu(t)$ is the number of customers waiting in line at time t who arrived after time 0 (already discussed in (3.7) in [1], $Q_n^o(t)$ is the number of customers waiting in line at time t who were present at time 0:

$$Q_n^o(t) = \sum_{i=1}^{Q_n(0,(W_n(t)-t)^+)} \mathbf{1}(\zeta_i(\eta_{n,i}) > t) = n \int_0^{(W_n(t)-t)^+} \int_0^1 \mathbf{1}(y > H_u(t))\, d\bar{U}_n^o(\bar{Q}_n(0,u), y), \tag{3.9}$$

Similar to $Q_n^\nu(t)$, (3.9) can be represented as sum of three terms, i.e,

$$Q_n^o(t) \equiv Q_{n,1}^o(t) + Q_{n,2}^o(t) + Q_{n,3}^o(t)$$

where

$$Q_{n,1}^o(t) \equiv \sqrt{n} \int_0^{(W_n(t)-t)^+} H_u^c(t)\, d\hat{Q}_n(0,u) \tag{3.10}$$

$$Q_{n,2}^o(t) \equiv \sqrt{n} \int_0^{(W_n(t)-t)^+} \int_0^1 \mathbf{1}(y > H_u(t))\, d\hat{U}_n^o(\bar{Q}_n(0,u), y) \tag{3.11}$$

$$Q_{n,3}^o(t) \equiv n \int_0^{(W_n(t)-t)^+} H_u^c(t)\, dQ(0,u). \tag{3.12}$$

Remark 3.2 (Understanding the queue-length decomposition)**.** *We now provide insights into the decompositions of Q_n^o. The first term* (3.10) *captures the randomness of the number of initial customers $\hat{Q}_n(0,\cdot)$ and its age distribution, the second term* (3.11) *captures the randomness of the remaining service times of these old customers given that $Q_n(0,y) \approx nQ(0,y)$, and the last term* (3.12) *involves the randomness in W_n.*

3.2 Main Results

In this section, we give many-server heavy-traffic FWLLN and FCLT results for the overloaded $G_t/GI/n+GI$ queue with a positive initial queue content, namely, there may be customers initially waiting in line at time 0. To establish convergence we need appropriate assumptions on initial conditions. Therefore, we require the following conditions to hold. The proofs are given in §4 of the appendix.

Assumption 1 (FCLT for initial ages in queue and initial HOL waiting time)**.** *There are customers waiting in line at time 0 in the nth system with positive elapsed waiting times, that is,* $B_n(0) = n$, $Q_n(0) > 0$ *and* $W_n(0) > 0$. *Moreover, the sequences* $\{\hat{Q}_n(0,\cdot)\}$ *and* $\{\hat{W}_n(0)\}$ *satisfy a joint FCLT, i.e.,*

$$(\hat{Q}_n(0,\cdot), \bar{Q}_n(0,\cdot), \hat{W}_n(0), W_n(0)) \Rightarrow (\hat{Q}(0,\cdot), Q(0,\cdot), \hat{W}(0), w(0))$$

in $\mathcal{D}^2 \times \mathbb{R}^2$ *as* $n \to \infty$ *where the two limits* $\hat{Q}(0,\cdot)$ *and* $\hat{W}(0)$ *are independent, and*

$$Q(0,y) = \int_0^{w(0)\wedge y} q(0,x)\,dx \quad \textit{and} \quad w(0) \geq 0.$$

Remark 3.3 (Understanding Assumption 1)**.** Assumption 1 says that given $W_n(0) \approx w(0)$ and ignoring the fluctuation $\hat{W}_n(0)$ when n is large, the limit $\hat{Q}(0,y)$ estimates the stochastic fluctuations of the initial age process $Q_n(0,y)$. Therefore, it is not restrictive to assume $\hat{Q}_n(0,y)$, $0 \leq y \leq W_n(0)$, is asymptotically independent with $\hat{W}_n(0)$ (which estimates the stochastic fluctuation of the upper bound $W_n(0)$ for the ages of initial customers).

Theorem 3.1 (FWLLN for $G_t/GI/n+GI$ with positive initial queue content)**.** *Suppose Assumption 1 and all assumptions in Theorem 4.1 of [1] hold. Then, as* $n \to \infty$,

$$(\bar{N}_n, \bar{Q}_n(0,\cdot), W_n(0), \bar{W}_n, \bar{V}_n, \bar{D}_n, \bar{E}_n, \bar{B}_n, \bar{Q}_n, \bar{X}_n, \bar{A}_n) \Rightarrow (\Lambda, Q(0,\cdot), w(0), w, v, D, E, \mathbf{1}, Q, X, A)$$

in $\mathcal{D}^9([0,T];\mathbb{R}) \times \mathcal{D}([0,\infty);\mathbb{R}) \times \mathbb{R}$ *where*

$$\Lambda(t) = \int_0^t \lambda(u)\,du \quad \textit{and} \quad Q(0,y) = \int_0^y q(0,u)\,du, \quad t \geq 0, \quad 0 \leq y \leq w(0).$$

The deterministic limits w *and* v *satisfy*

$$w(t) = w(0) + \int_0^t \left(1 - \frac{\mu}{q(u,w(u))}\right) du, \quad v(t) = w(t + v(t)), \quad t \geq 0. \tag{3.13}$$

where the function $q(\cdot,\cdot)$ *is defined in Theorem 4.1 in [1].*

Moreover, for $t \geq 0$, $D(t) = E(t) = \mu t$,

$$Q(t) = \begin{cases} \int_0^{w(t)-t} H_u^c(t)\, dQ(0,u), & \text{for} \quad t < t^* \\ \int_{t-w(t)}^t F^c(t-s)\lambda(s)\, ds, & \text{for} \quad t \geq t^* \end{cases},$$

where t^* *is the boundary point for which the conditions* (i) $t^* = w(t^*)$, (ii) $w(t) > t$ *for* $t < t^*$, *and* (iii) $w(t) < t$ *for* $t > t^*$ *are satisfied. Moreover,* $X(t) = Q(t) + 1$ *and* $A(t) = \Lambda(t) - E(t) - X(t) + X(0)$.

Theorem 3.2 (FCLT for $G_t/GI/n+GI$ with positive initial queue content)**.** *Suppose Assumption 1 and all assumptions in Theorem 4.1 of [1] hold. Then, as* $n \to \infty$,

$$\begin{aligned} &(\hat{N}_n, \hat{Q}_n(0,\cdot), \hat{W}_n(0), \hat{W}_n, \hat{V}_n, \hat{D}_n, \hat{E}_n, \hat{B}_n, \hat{Q}_n, \hat{X}_n, \hat{A}_n) \\ &\quad \Rightarrow (\hat{N}, \hat{Q}(0,\cdot), \hat{W}(0), \hat{W}, \hat{V}, \hat{E}, \hat{E}, \mathbf{0}, \hat{Q}, \hat{Q}, \hat{A}) \quad \text{in} \quad \mathcal{D}^9([0,T];\mathbb{R}) \times \mathcal{D}([0,\infty);\mathbb{R}) \times \mathbb{R} \end{aligned}$$

where $\hat{N}$, $\hat{Q}(0,\cdot)$ *and* $\hat{W}(0)$ *are given in Assumption 1, respectively. Moreover,* $\hat{A}(t) = \hat{N}(t) - \hat{Q}(t) - \hat{E}(t) + \hat{Q}(0)$, *and* $\hat{D}(t) = \hat{E}(t)$ *for all* $t \geq 0$.

The limit enter-service process $(\hat{E}(t) : t \geq 0)$ *is a zero-mean Gaussian process given in Theorem 4.2 of [1].*

The limit of head-of-line waiting time $(\hat{W}(t) : t \geq 0)$ *uniquely solves the piecewise stochastic differential equation (PSIE)*

$$\begin{aligned} \hat{W}(t) = &-\frac{1}{\lambda^*(t-w(t))\tilde{F}^c_{t-w(t)}(w(t))} \int_0^t \tilde{f}_s(w(s))\lambda^*(s-w(s))\hat{W}(s)\, ds \\ &+ \frac{1}{\lambda^*(t-w(t))\tilde{F}^c_{t-w(t)}(w(t))}\, \hat{G}(t) + \frac{\lambda^*(-w(0))}{\lambda^*(t-w(t))\tilde{F}^c_{t-w(t)}(w(t))}\, \hat{W}(0), \end{aligned}$$

where w *is as in* (3.13), $\widetilde{f}_s(x) \equiv (\partial/\partial x)\widetilde{F}_s(x)$,

$$\begin{aligned} \widetilde{F}^c_s(x) &= \begin{cases} \frac{F^c(x)}{F^c(-s)} & \text{if} \quad s \leq 0 \\ F^c(x) & \text{if} \quad s > 0 \end{cases}, \\ \hat{G}(t) &\equiv \begin{cases} \int_{w(t)-t}^{w(0)} \frac{F^c(v(-s))}{F^c(s)}\, d\hat{Q}(0,s) + \mathcal{B}^o\left(T^o(t)\right) - \hat{E}(t) & \text{for} \quad t < t* \\ Z^o + \int_0^t F^c(w(s))\, d\hat{N}(s-w(s)) + \mathcal{B}^\nu\left(T^\nu(t)\right) - \hat{E}(t) & \text{for} \quad t \geq t^* \end{cases}, \\ T^o(t) &\equiv \int_{w(t)-t}^{w(0)} \frac{F^c(v(-u))}{F^c(u)} \left(1 - \frac{F^c(v(-u))}{F^c(u)}\right) q(0,u)\, du, \quad t \geq 0, \\ T^\nu(t) &\equiv \int_0^t F^c(v(u))F(v(u))\lambda(u)\, du, \quad t \geq 0, \\ Z^o &\equiv \int_0^{w(0)} \frac{F^c(v(-u))}{F^c(u)}\, d\hat{Q}(0,u) + \mathcal{B}^o\left(T^o(0)\right) \end{aligned} \tag{3.14}$$

with $(\mathcal{B}^\nu(t) : t \geq 0)$ *and* $(\mathcal{B}^o(t) : t \geq 0)$ *being two independent standard Brownian motions. The deterministic function* v *in* (3.14) *is characterized by* (3.13). *The boundary point* t^* *satisfies the following conditions: (i)* $t^* = w(t^*)$, *(ii)* $w(t) > t$ *for* $t < t^*$, *and (iii)* $w(t) < t$ *for* $t > t^*$. *The integrals in* (3.14) *are interpreted as the form after integration by parts.*

The limit virtual waiting time process $\hat{V}$ *uniquely solves*

$$\hat{V}(t) = \frac{\hat{W}(t + v(t))}{1 - \dot{w}(t + v(t))}, \quad t \geq 0$$

where $\dot{w}$ *is the derivative of* w, *and* v *is as in* (3.13).

The limit queue-length process $\hat{Q}$ *is the sum of three independent terms, specifically,*

$$\hat{Q}(t) \equiv \hat{Q}_1(t) + \hat{Q}_2(t) + \hat{Q}_3(t),$$

with

$$\begin{aligned}
\hat{Q}_1(t) &\equiv \int_{(t-w(t))^+}^{t} F^c(t-s)\, d\hat{N}(s) + \int_0^{(w(t)-t)^+} H_s^c(t)\, d\hat{Q}(0,s), \quad t \geq 0, \\
\hat{Q}_2(t) &\equiv \int_{t-w(t)}^{t} \int_0^1 \mathbf{1}(x > F(t-s))\, d\hat{U}^\nu(\Lambda(s), y) \\
&\quad + \int_0^{(w(t)-t)^+} \int_0^1 \mathbf{1}(x > H_s^c(t))\, d\hat{U}^o(Q(0,s), y), \quad t \geq 0, \\
\hat{Q}_3(t) &\equiv q(t, w(t))\hat{W}(t), \quad t \geq 0
\end{aligned}$$

where $\hat{U}^\nu$ *and* $\hat{U}^o$ *are two independent standard Kiefer processes, and*

$$H_s^c(t) \equiv \frac{F^c(t+s)}{F^c(s)}, \quad s, t \geq 0$$

4 Proofs of Theorems 3.1 and 3.2

In §4.1, we first provide a proof for FWLLN by using compactness approach. The key step in our proof of FWLLN for all processes is to establish convergence for the sequence of head-of-line waiting times $\{W_n : n \geq 1\}$. Given the convergence of $\{W_n : n \geq 1\}$, convergence of the other sequences follows by continuous mapping theorem. The limit processes obtained in §4.1 are used as centering terms in §4.2 to define the CLT-scaled processes. Then, in the same section, we prove a FCLT for these processes using a different approach. The key step is to use Gronwall's inequality to first prove stochastic boundedness of and then convergence for the sequence of scaled processes $\{\hat{W}_n : n \geq 1\}$ in $\mathcal{D}([0,T];\mathbb{R})$.

4.1 Proof of Theorem 3.1

We first establish a FWLLN for $\{W_n : n \geq 1\}$ by using compactness approach, i.e., (i) we show the sequence $\{W_n : n \geq 1\}$ is tight in $\mathcal{D}([0,T];\mathbb{R})$ which implies every subsequence has a convergent subsequence (see §4.1.1); (ii) every convergent subsequence converges to the same limit which uniquely solves the ODE in Theorem 3 of [6] (see §4.1.2). Finally, in §4.1.3, we establish convergence for the other processes and characterize their limits.

4.1.1 Tightness of the sequence $\{W_n\}$

We prove tightness in two steps: First we show that $\{W_n : n \geq 1\}$ is stochastically bounded in $\mathcal{D}([0,T];\mathbb{R})$ and then show that the following criterion involving modulus of continuity is satisfied: For each $T > 0$ and $\epsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathbb{P}(w(W_n, \delta, T) > \epsilon) = 0 \tag{4.1}$$

where $w(W_n, \delta, T)$ is the modulus of continuity of W_n, i.e., $\sup\{w(W_n, [t_1, t_2] : 0 \leq t_1 < t_2 \leq (t_1 + \delta) \wedge T)\}$ with $w(W_n, A) \equiv \sup\{W_n(s_1) - W_n(s_2) : s_1, s_2 \in A\}$.

Stochastic boundedness. Since we consider the system in the interval $[0,T]$, we immediately see that the HOL waiting time for new customers satisfies $0 \leq W_n(t) \leq T$ for all $n \geq 1$, $t \in [0,T]$. For initial customers, on the other hand, we make use of the assumed convergence $W_n(0) \Rightarrow w(0) < \infty$ in $\mathbb{R}$. In particular, we can bound from above HOL waiting time of an initial customer by $W_n(0)+T$ for all $n \geq 1$ thanks to FCFS service discipline. The upper bound $W_n(0)+T$ is stochastically bounded due to the assumed convergence. Therefore, the sequence of HOL waiting time for initial customers is stochastically bounded. Hence we conclude that $\{W_n : n \geq 1\}$ is stochastically bounded.

Modulus of continuity. $W_n(t+\delta) - W_n(t) \leq \delta$ for $\delta > 0$ and $t \geq 0$ holds because the HOL waiting time can increase at most at rate 1. Therefore, it remains to find a bound on $W_n(t) - W_n(t+\delta)$ to conclude that the criterion involving modulus of continuity of $W_n(t)$ is satisfied.

To this end, let us define the δ-increment of $\bar{E}_{n,3}(t)$,

$$\begin{aligned}\bar{E}_{n,3}(t,\delta) &\equiv \bar{E}^{\nu}_{n,3}(t,\delta) + \bar{E}^{o}_{n,3}(t,\delta) \equiv \bar{E}^{\nu}_{n,3}(t+\delta) - \bar{E}^{\nu}_{n,3}(t) + \bar{E}^{o}_{n,3}(t+\delta) - \bar{E}^{o}_{n,3}(t) \\ &= \int_{(t-W_n(t))^+}^{(t+\delta-W_n(t+\delta))^+} F^c(V_n(s))\lambda(s)\,ds + \int_{(W_n(t)-t)^+}^{(W_n(t+\delta)-t-\delta)^+} H^c_u(V_n(-u)-u)q(0,u)\,du. \end{aligned} \tag{4.2}$$

Because F^c is continuous and $F^c(x) > 0$ for all $x \geq 0$, it holds that $\inf_{x\in[0,T]}\{F^c(x)\} = c_1 > 0$ for any $T > 0$. Similarly, $\inf_{x\in[0,T]}\{H^c_x(V_n(-x)-x)\} = c_2 > 0$ for any $T > 0$. Without loss of generality, we assume that $\lambda(x) > 0$ for all $x \in [t - W_n(t), t + \delta - W_n(t+\delta)]$ and $q(0,x) > 0$ for all

$x \in [W_n(t) - t, W_n(t+\delta) - t - \delta]$ so that both integrands in (4.2) are bounded below by a constant $c > 0$. Then replacing both integrands with the constant c yields a lower bound on $\bar{E}_{n,3}(t,\delta)$. In particular, we have

$$(W_n(t) - W_n(t+\delta) + \delta)\mathbf{1}_{\{t > W_n(t)\}} + (W_n(t+\delta) - W_n(t) - \delta)\mathbf{1}_{\{t < W_n(t)\}} \leq \frac{\bar{E}_{n,3}(t,\delta)}{c}$$

and by the convergence $\bar{D}_n \Rightarrow D$ in $\mathcal{D}([0,T];\mathbb{R})$ and that $E(t) = D(t)$, we have

$$\lim_{n\to\infty} \{W_n(t) - W_n(t+\delta)\} \leq \frac{D(t,\delta)}{c}$$

which implies the modulus of continuity condition (4.1). Hence, $\{W_n : n \geq 1\}$ is tight in $\mathcal{D}([0,T];\mathbb{R})$. More specifically, $\{W_n : n \geq 1\}$ is C-tight because (4.1) is sufficient for C-tightness together with stochastic boundedness.

4.1.2 Characterizing the limit of the sequence $\{W_n\}$.

Due to C-tightness, we know that (i) every subsequence of $\{W_n : n \geq 1\}$ has a convergent subsequence. Let $\{W_n^* : n \geq 1\}$ be such a convergent subsequence with the limit w^*, i.e., $W_n^* \Rightarrow w^*$ in $\mathcal{D}([0,T];\mathbb{R})$. The convergence of the subsequence $\{W_n^* : n \geq 1\}$ implies that there exists a corresponding subsequence $\{V_n^* : n \geq 1\}$ that converges to some v^* satisfying the fluid equations

$$v^*(t) \;=\; w^*(t + v^*(t)) \quad \text{and} \quad v^*(t - w^*(t)) = w^*(t), \quad t \geq 0. \tag{4.3}$$

Next we derive an ordinary differential equation for w^* using the LLN-scaled enter-service process $\bar{E}_n(t)$. First, from the FCLT in Theorem 2 of [12], we deduce that $\hat{D} \Rightarrow D$ in $\mathcal{D}([0,T];\mathbb{R})$ which, together with the assumption that the fluid model is OL throughout entire time interval $[0,T]$, implies $\sup_{t\in[0,T]}\{|E_n(t) - D_n(t)|\} = o(\sqrt{n})$ and

$$\bar{E}_n(t) \Rightarrow E(t) = \int_0^t b(u,0)\,du = \mu t \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty. \tag{4.4}$$

On the other hand, from (3.4) and that $(W_n^*, V_n^*) \Rightarrow (w^*, v^*)$ in $\mathcal{D}^2([0,T];\mathbb{R})$, the sequence $\{\bar{E}_n : n \geq 1\}$ along the subsequence associated with $\{W_n^*\}$ and $\{V_n^*\}$ converges to a limit E^* satisfying

$$\begin{aligned} \bar{E}_n^*(t) \;&\Rightarrow\; E^*(t) = E_3^{o,*}(t) + E_3^{\nu,*}(t) \\ &\equiv\; \int_0^{(t-w^*(t))^+} F^c(v^*(s))\lambda(s)\,ds + \int_{(w^*(t)-t)^+}^{w^*(0)} H_u^c(v^*(-u) - u)q(0,u)\,du \end{aligned} \tag{4.5}$$

with

$$\bar{E}_{n,1}^*(t) \equiv \bar{E}_{n,1}^{o,*}(t) + \bar{E}_{n,1}^{\nu,*}(t) \Rightarrow (0e)(t) \quad \text{and} \quad \bar{E}_{n,2}^*(t) \equiv \bar{E}_{n,2}^{o,*}(t) + \bar{E}_{n,2}^{\nu,*}(t) \Rightarrow (0e)(t)$$

in $\mathcal{D}([0,T];\mathbb{R})$ as $n \to \infty$. From [10], we know that LLN-scaled versions of (3.5) and (3.6) vanish in the limit if W_n and V_n are replaced by the deterministic fixed (independent of n) functions w and v, respectively. Therefore, we conclude that $\{\bar{E}^*_{n,1}\}$ and $\{\bar{E}^*_{n,2}\}$ vanish because the limits w^* and v^* are deterministic.

We now derive an ODE for the limit w^*. Equating (4.4) and (4.5), and taking the derivative of both sides yield

$$\begin{aligned} b(t,0) &= (1-\dot{w}^*(t))F^c(v^*(t-w^*(t))\lambda(t-w^*(t))\mathbf{1}(t \geq w^*(t)) \\ &\quad + (1-\dot{w}^*(t))H^c_{w^*(t)-t}\left(v^*(t-w^*(t)) - w^*(t) + t\right) q(0, w^*(t)-t)\mathbf{1}(t < w^*(t)) \\ &= (1-\dot{w}^*(t))\left(F^c(w^*(t))\lambda(t-w^*(t))\mathbf{1}(t \geq w^*(t)) + H^c_{w^*(t)-t}(t)\, q(0, w^*(t)-t)\mathbf{1}(t < w^*(t))\right), \end{aligned} \tag{4.6}$$

where the second equality holds by (4.3). Equation (4.6) implies that

$$\begin{aligned} \dot{w}^*(t) &= 1 - \frac{b(t,0)}{F^c(w^*(t))\lambda(t-w^*(t))\mathbf{1}(t \geq w^*(t)) + H^c_{w^*(t)-t}(t)\, q(0, w^*(t)-t)\mathbf{1}(t < w^*(t))} \\ &= 1 - \frac{b(t,0)}{q(t, w^*(t))}, \end{aligned}$$

which coincides with the ODE (3.13) which has a unique solution. This implies that any convergent subsequence $\{W^*_n\}$ must converge to the same limit and hence full convergence of $\{W_n : n \geq 1\}$.

4.1.3 FWLLN of the other processes.

First, we prove full convergence of $\{V_n : n \geq 1\}$. In particular, for $t \geq 0$,

$$\begin{aligned} |V_n(t - W_n(t)) - v(t - w(t))| &\leq |V_n(t - W_n(t)) - V_n(t - w(t))| + |V_n(t - w(t)) - v(t - w(t))| \\ &= |W_n(t) - w(t) + O(1/n)| + |w(t) + O(1/n) - w(t)| \\ &\leq |W_n(t) - w(t)| + O(1/n) \end{aligned} \tag{4.7}$$

where the equality follows from (2.3) and (2.4) in [1]. This implies convergence of $V_n \Rightarrow v$ in $\mathcal{D}([0,T];\mathbb{R})$ thanks to $W_n \Rightarrow w$ in $\mathcal{D}([0,T];\mathbb{R})$.

We further obtain (4.8) by applying change of variable to (4.7) with $u_n \equiv t - W_n(t)$ and $u \equiv t - w(t)$, i.e., for a constant $\gamma > 0$,

$$\|V_n - v\| \leq \frac{\|W_n - w\|}{\gamma} + O(1/n) = O(1/n) \tag{4.8}$$

where the equality holds because $u_n = u + o(1)$. We will make use of (4.8) establishing an FCLT limit for $\{\hat{V}_n : n \geq 1\}$ in §4.2.

We readily have the limit for the enter-service process $\{\bar{E}_n : n \geq 1\}$ along the convergent subsequence in (4.5). Full convergence of enter-service follows from full convergence of $\{W_n\}$ and $\{V_n\}$ established above. Therefore, we can now drop the superscripts in (4.5). The limit of the sequence of departure processes must coincide with the limit of the sequence of enter-service processes because, in the limit, the system will be overloaded over the entire interval $[0, T]$.

The limit of the sequences of processes (3.8) and (3.9) in [1] can be obtained the same way it is done in [8]. which make use of Theorem 3.1. of [10] and then apply continuous mapping theorem given $W_n \Rightarrow w$. From (6.17) of [8], we immediately write

$$\bar{Q}^{\nu}_{n,i} \Rightarrow 0e \quad \text{for} \quad i = 1, 2; \quad \bar{Q}^{\nu}_{n,3} \Rightarrow Q^{\nu}_3(t) \equiv \int_{(t-w(t))^+}^{t} F^c(t-s)\lambda(s)\, ds$$

in $\mathcal{D}([0, T]; \mathbb{R})$ as $n \to \infty$.

The limit of the sequences of processes (3.10)–(3.12) can be obtained in a similar. Although not treated in [8], we can use similar arguments because they have similar mathematical forms as (3.8)–(3.9) in [1]. Therefore, we obtain

$$\bar{Q}^{\nu}_{n,i} \Rightarrow 0e \quad \text{for} \quad i = 1, 2; \quad \bar{Q}^{\nu}_{n,3} \Rightarrow Q^{\nu}_3(t) \equiv \int_{0}^{(w(t)-t)^+} H^c_u(t)\, dQ(0, u)$$

in $\mathcal{D}([0, T]; \mathbb{R})$ as $n \to \infty$ where $Q(0, \cdot)$ is the limit of the sequence $\{\bar{Q}_n(0, \cdot)\}$ implied by the FCLT $\hat{Q}_n(0, \cdot) \Rightarrow \hat{Q}(0, \cdot)$ in Assumption 1. Hence convergence of (3.8) follows from continuous mapping theorem with addition.

4.2 Proof of Theorem 3.2.

In §4.2.1, we first establish an FCLT for the sequences $\{\hat{W}_n : n \geq 1\}$ and $\{\hat{V}_n : n \geq 1\}$. Then, in §4.2.2, we establish FCLT for the other processes given the FCLT in §4.2.1.

4.2.1 FCLT for $\hat{W}_n$ and $\hat{V}_n$

To prove the FCLTs for $\hat{W}_n$ and $\hat{V}_n$, one could do the compactness approach ((i) C-tightness and (ii) characterization of the limit of convergent subsequence). This could be done by mimicking the approach here in §4.1 and the approach in [8]. However, we hereby adopt a new approach: we establish the convergence $\hat{W}_n \Rightarrow \hat{W}$ and $\hat{V}_n \Rightarrow \hat{V}$ using the continuous mapping theorem and Gronwall's inequality. We show that the limit process $\hat{W}$ uniquely solves a stochastic differential equation (SDE). Our SDE here generalizes the SDE given in (6.64) of [8] in two ways. First, our generalization from M service to GI will replace the Brownian motion $\mathcal{B}_s$ there by a general

Gaussian process. Second, our general initial condition here will generate a piecewise structure both in the drift term and in the volatility term.

By definition of $E_n(t)$, the number of customers entered service by time t

$$\begin{aligned} E_n(t) &\equiv E_n^\nu(t) + E_n^o(t) \\ &\equiv \sum_{i=1}^{N_n((t-W_n(t))^+)} \mathbf{1}(\gamma_i^n > V_n(\tau_i^n -)) + \sum_{i=Q_n(0,(W_n(t)-t)^+)+1}^{Q_n(0,W_n(0))} \mathbf{1}(\gamma_{-i}^n > V_n(-\eta_{-i}^n) - \eta_{-i}^n) \\ &= n \int_0^{(t-W_n(t))^+} \int_0^1 \mathbf{1}(y > F(V_n(s)))\, d\bar{U}_n^\nu(\bar{N}_n(s), y) \qquad (4.9) \\ &\quad + n \int_{(W_n(t)-t)^+}^{W_n(0)} \int_0^1 \mathbf{1}(y > H_u(V_n(-u) - u))\, d\bar{U}_n^o(\bar{Q}_n(0,u), y). \qquad (4.10) \end{aligned}$$

We use the representation in [10] to rewrite (4.9) and (4.10) as

$$\begin{aligned} & n \int_0^{(t-W_n(t))^+} \int_0^1 \mathbf{1}(y > F(V_n(s)))\, d\bar{U}_n^\nu(\bar{N}_n(s), y) \\ &= \sqrt{n} \int_0^{(t-W_n(t))^+} F^c(V_n(s))\, d\hat{N}_n(s) + n \int_0^{(t-W_n(t))^+} F^c(V_n(s))\lambda(s)\, ds + \sum_{i=1}^{N_n((t-W_n(t))^+)} \left(\mathbf{1}(\gamma_i^n > w_i^n) - F^c(w_i^n)\right), \end{aligned}$$

and

$$\begin{aligned} & n \int_{(W_n(t)-t)^+}^{W_n(0)} \int_0^1 \mathbf{1}(y > H_u(V_n(-u) - u))\, d\bar{U}_n^o(\bar{Q}_n(0,u), y) \\ &= \sqrt{n} \int_{(W_n(t)-t)^+}^{W_n(0)} H_u^c(V_n(-u) - u)\, d\hat{Q}_n(0,u) + n \int_{(W_n(t)-t)^+}^{W_n(0)} H_u^c(V_n(-u) - u) q(0,u)\, du \\ &\quad + \sum_{i=Q_n(0,(t-W_n(t))^+)+1}^{Q_n(0,W_n(0))} \left(\mathbf{1}(\gamma_{-i}^n > w_{-i}^n) - \frac{F^c(w_{-i}^n + \eta_{-i}^n)}{F^c(\eta_{-i}^n)}\right) \qquad (4.11) \end{aligned}$$

We now introduce new notation to simplify the expressions in (4.11). Define

$$\Lambda_n(s) \equiv \begin{cases} Q_n(0,-s) & \text{for } s \le 0 \\ N_n(s) & \text{for } s > 0 \end{cases}, \quad \Lambda^*(s) \equiv \int_0^s \lambda^*(u)\, du \quad \text{and} \quad \widetilde{F}_s^c(x) \equiv \begin{cases} \frac{F^c(x)}{F^c(-s)} & \text{for } s \le 0 \\ F^c(x) & \text{for } s > 0 \end{cases} \qquad (4.12)$$

where

$$\lambda^*(s) \equiv \begin{cases} q(0,-s) & \text{if } s \le 0 \\ \lambda(s) & \text{if } s > 0 \end{cases}.$$

We now decompose $E_n(t)$ into sum of three terms, i.e.,

$$E_n(t) = E_{n,1}(t) + E_{n,2}(t) + E_{n,3}(t)$$

where

$$E_{n,1}(t) = \sqrt{n} \int_{-W_n(0)}^{t-W_n(t)} \widetilde{F}_s^c(V_n(s))\, d\hat{\Lambda}_n(s), \tag{4.13}$$

$$E_{n,2}(t) = \sum_{i=Q_n(0,(W_n(t)-t)^+)+1}^{Q_n(0,W_n(0))} \left(\mathbf{1}(\gamma_{-i}^n > w_{-i}^n) - \frac{F^c(w_{-i}^n + \eta_{-i}^n)}{F^c(\eta_{-i}^n)} \right) \tag{4.14}$$

$$+ \sum_{i=1}^{N_n((t-W_n(t))^+)} \left(\mathbf{1}(\gamma_i^n > w_i^n) - F^c(w_i^n) \right), \tag{4.15}$$

$$E_{n,3}(t) = n \int_{-W_n(0)}^{t-W_n(t)} \widetilde{F}_s^c(V_n(s)) \lambda^*(s)\, ds. \tag{4.16}$$

From (4.5) and the discussion in §4.1.3, we deduce that

$$\bar{E}_{n,1}(t) \Rightarrow (0e)(t),\ \bar{E}_{n,2}(t) \Rightarrow (0e)(t) \quad \text{in} \quad \mathcal{D}([0,T];\mathbb{R}),$$
$$\bar{E}_{n,3}(t) \Rightarrow E_3(t) \equiv \int_{-w(0)}^{t-w(t)} \widetilde{F}_s^c(v(s))\, d\Lambda^*(s) \quad \text{in} \quad \mathcal{D}([0,T];\mathbb{R}).$$

Then, by definition, it follows that

$$\hat{E}_n(t) = \frac{1}{\sqrt{n}} E_{n,1}(t) + \frac{1}{\sqrt{n}} E_{n,2}(t) + \frac{1}{\sqrt{n}} (E_{n,3}(t) - nE_3(t)). \tag{4.17}$$

Overview of proof. The key step of our proof is establishing convergence and characterizing the limit of scaled head-of-line waiting-time processes $\{\hat{W}_n(t) : n \geq 1\}$. To do so, we first obtain an SDE for the scaled prelimit head-of-line waiting-time process $\hat{W}_n(t)$ (see (4.39)). Then using Gronwall's inequality (Lemma 1.1), we show that the sequence $\{\hat{W}_n(t) : n \geq 1\}$ is stochastically bounded in $\mathcal{D}([0,T];\mathbb{R})$. Next, we show that $\hat{W}_n$ converges uniformly to a limit process $\hat{W}$ as $n \to \infty$ where the limit $\hat{W}$ is given in (4.43). Given convergence $\hat{W}_n \Rightarrow \hat{W}$, we establish FCLT for the other processes and characterize their limits in §4.2.2.

The following lemmas will help deriving an SDE for the process $\hat{W}_n$ for all $n \geq 1$ and characterize the limit. In Lemma 4.1 and Lemma 4.2, we respectively establish convergence of the first and the second component on the right-hand side of the equality in (4.17).

Lemma 4.1. *For $T > 0$, if*

$$(\hat{N}_n, \hat{Q}_n(0,\cdot)) \Rightarrow (\hat{N}, \hat{Q}(0,\cdot)) \quad \textit{in} \quad \mathcal{D}([0,T];\mathbb{R}) \times \mathcal{D}([0,\infty);\mathbb{R})$$

as $n \to \infty$, then

$$\frac{1}{\sqrt{n}}\, E_{n,1}(t) \Rightarrow \int_0^t \widetilde{F}_{s-w(s)}^c(w(s))\, d\hat{\Lambda}(s - w(s)) \quad \textit{in} \quad \mathcal{D}([0,T];\mathbb{R}). \tag{4.18}$$

Proof. We consider the modified processes $n^{-1/2}\tilde{E}_{n,1}(t)$ given below. We first prove convergence for the sequence $\{n^{-1/2}\tilde{E}_{n,1}(t) : n \geq 1\}$ and then show that the difference between the modified sequences $\{n^{-1/2}\tilde{E}_{n,1} : n \geq 1\}$ and the desired sequence $\{n^{-1/2}E_{n,1}(t) : n \geq 1\}$ is asymptotically negligible which proves the desired convergence of $n^{-1/2}\tilde{E}_{n,1}$.

Now consider for $t \geq 0$ the processes

$$\begin{aligned}
\frac{1}{\sqrt{n}}\tilde{E}_{n,1}(t) &= \int_{-w(0)}^{t-w(t)} \widetilde{F}^c_s(v(s))\, d\hat{\Lambda}_n(s) \\
&= \widetilde{F}^c_{t-w(t)}(v(t-w(t)))\hat{\Lambda}_n(t-w(t)) - \widetilde{F}^c_{-w(0)}(v(-w(0)))\hat{\Lambda}_n(-w(0)) \\
&\quad - \int_{-w(0)}^{t-w(t)} \hat{\Lambda}_n(s-)\, d\widetilde{F}^c_s(v(s)) \\
&= \widetilde{F}^c_{t-w(t)}(w(t))\hat{\Lambda}_n(t-w(t)) - \hat{\Lambda}_n(-w(0)) - \int_0^t \hat{\Lambda}_n(s-w(s))\, d\widetilde{F}^c_{s-w(s)}(w(s)). \quad (4.19)
\end{aligned}$$

The second equality holds since $\hat{\Lambda}_n(s)$ is of bounded variation and therefore the integral can be represented as the form after integration by parts. The last equality follows from the fluid equation $v(t-w(t)) = w(t)$, for $t \geq 0$, and that $\widetilde{F}^c_{-w(0)}(w(0)) = 1$ by definition. Next we define a mapping $\psi : \mathcal{D}([0,T];\mathbb{R}) \to \mathcal{D}([0,T];\mathbb{R})$ such that for $z \in \mathcal{D}([0,T];\mathbb{R})$,

$$\psi(z)(t) \equiv \widetilde{F}^c_{t-w(t)}(w(t))z(t) - z(0) - \int_0^t z(s)\, d\widetilde{F}^c_{s-w(s)}(w(s)), \quad 0 \leq t \leq T.$$

We now prove that the mapping ψ is continuous in $\mathcal{D}([0,T];\mathbb{R})$. Let $\{x_n\}$ be a sequence in $\mathcal{D}([0,T];\mathbb{R})$ such that $\|x_n - x\|_T \to 0$. Then

$$\begin{aligned}
&|\psi(x_n)(t) - \psi(x)(t)| \\
&= \Big| \widetilde{F}^c_{t-w(t)}(w(t))x_n(t) - x_n(0) - \int_0^t x_n(s)\, d\widetilde{F}^c_{s-w(s)}(w(s)) \\
&\quad - \widetilde{F}^c_{t-w(t)}(w(t))x(t) + x(0) + \int_0^t x(s)\, d\widetilde{F}^c_{s-w(s)}(w(s)) \Big| \\
&\leq \widetilde{F}^c_{t-w(t)}(w(t))|x_n(t) - x(t)| + |x_n(0) - x(0)| + \|x_n - x\|_T \left| \int_0^t d\widetilde{F}^c_{s-w(s)}(w(s)) \right| \leq 4\,\|x_n - x\|.
\end{aligned}$$

Hence the mapping ψ is continuous. In general, proving convergence in the uniform topology does not necessarily imply J_1 convergence because there may be measurability issues (see e.g. [13, 3]). However, we will be interested in the case where the limit x is continuous, i.e., $x \in \mathcal{C}([0,T];\mathbb{R})$. Therefore, we will not have any measurability issues and obtain the desired convergence in $\mathcal{D}([0,T];\mathbb{R})$ with respect to Skorokhod's J_1 metric.

Convergence of the modified process in (4.19) follows by continuous mapping theorem with composition. In particular, let $Z_n(\cdot) \equiv \hat{\Lambda}_n(\cdot - W_n(\cdot))$. Then $Z_n : \mathcal{D}([0,T];\mathbb{R}) \to \mathcal{D}([0,T];\mathbb{R})$ and

$Z_n \Rightarrow Z$ in $\mathcal{D}([0,T];\mathbb{R})$ where $Z(\cdot) \equiv \hat{\Lambda}(\cdot - w(\cdot))$. Convergence of $\{Z_n\}$ follows from continuous mapping theorem with composition. In particular, we apply Theorem 13.2.2 of [13] with sequences $t - W_n(t)$ and $\hat{\Lambda}_n$ converging to the continuous limits $t - w(t)$ and $\hat{\Lambda}(t)$, respectively. Then we have $n^{-1/2}\tilde{E}_{n,1}(\cdot) = \psi(Z_n)(\cdot) \Rightarrow \psi(Z)(\cdot)$ in $\mathcal{D}([0,T];\mathbb{R})$. We denote the limit $\psi(Z)$ as

$$\int_0^t \widetilde{F}^c_{s-w(s)}(w(s))\, d\hat{\Lambda}(s - w(s)) \equiv \widetilde{F}^c_{t-w(t)}(w(t))\hat{\Lambda}(t - w(t)) - \hat{\Lambda}(-w(0)) - \int_0^t \hat{\Lambda}(s - w(s))\, d\widetilde{F}^c_{s-w(s)}(w(s))$$

for $0 \le t \le T$ where $\hat{\Lambda}(s) \equiv \hat{N}(s)$ for $s > 0$ and $\hat{\Lambda}(s) \equiv \hat{Q}(0,s)$ for $s \le 0$.

Finally, we show that the difference between the processes $n^{-1/2}E_{n,1}(t)$ and $n^{-1/2}\tilde{E}_{n,1}(t)$ is asymptotically negligible. In particular,

$$\begin{aligned}
\frac{1}{\sqrt{n}}|E_{n,1}(t) - \tilde{E}_{n,1}(t)| &= \frac{1}{\sqrt{n}}\left|\int_{-W_n(0)}^{t-W_n(t)} \widetilde{F}^c_s(V_n(s))\, d\hat{\Lambda}_n(s) - \int_{-w(0)}^{t-w(t)} \widetilde{F}^c_s(v(s))\, d\hat{\Lambda}_n(s)\right| \\
&= \frac{1}{\sqrt{n}}\left|\int_{-W_n(0)}^{-w(0)} \widetilde{F}^c_s(V_n(s))\, d\hat{\Lambda}_n(s) + \int_{t-W_n(t)}^{t-w(t)} \widetilde{F}^c_s(v(s))\, d\hat{\Lambda}_n(s)\right| \\
&\le \frac{1}{\sqrt{n}}\left|\hat{\Lambda}_n(-w(0)) - \hat{\Lambda}_n(-W_n(0))\right| + \frac{1}{\sqrt{n}}\left|\hat{\Lambda}_n(t - w(t)) - \hat{\Lambda}_n(t - W_n(t))\right|
\end{aligned}$$

which converges to 0 due to $W_n \Rightarrow w$ and $\hat{\Lambda}_n(\cdot - W_n(\cdot)) \Rightarrow \hat{\Lambda}(\cdot - w(\cdot))$ in $\mathcal{D}([0,T];\mathbb{R})$. Hence this completes the proof.

We next present a convergence result for the second component in (4.17).

Lemma 4.2. *For $T > 0$, if*

$$(\hat{N}_n, \hat{Q}_n(0,\cdot)) \Rightarrow (\hat{N}, \hat{Q}(0,\cdot)) \quad \textit{in} \quad \mathcal{D}([0,T];\mathbb{R}) \times \mathcal{D}([0,\infty);\mathbb{R})$$

as $n \to \infty$, then

$$\frac{1}{\sqrt{n}}\, E_{n,2}(t) \Rightarrow \mathcal{B}^o(T^o(t)) + \mathcal{B}^\nu(T^\nu(t)) \quad \textit{in} \quad \mathcal{D}([0,T];\mathbb{R}) \quad \textit{as} \quad n \to \infty \tag{4.20}$$

where $\mathcal{B}^o$ and $\mathcal{B}^\nu$ are independent standard Brownian motions with time transforms,

$$\begin{aligned}
T^o(t) &= \int_{(w(t)-t)^+}^{w(0)} \widetilde{F}^c_u(v(-u))\widetilde{F}_u(v(-u))q(0,u)\, du, \\
T^\nu(t) &= \int_0^t F^c(v(u))F(v(u))\lambda(u)\, du.
\end{aligned}$$

Proof. We prove that the first component of the process $n^{-1/2}E_{n,2}(t)$ weakly converges in $\mathcal{D}([0,T];\mathbb{R})$ to a Brownian motion. We apply continuous mapping to conclude that the sum of the two processes converges to the sum of the limits.

Proof of convergence of the first component. Our ultimate goal is to use martingale FCLT to prove convergence for (4.14) which is the CLT-scaled total number of initial customers entered service. First we define a sequence of discrete-time processes (see (4.21)) and argue that it is a sequence of martingales adapted to a specific filtration $\mathscr{H}_k^n$ as defined below. Next, we define continuous-time martingales using the discrete-time martingales in (4.21). Then we invoke Theorem 7.1.4. on p.339 in [5] to establish convergence and characterize the limit.

Consider the discrete-time processes

$$\hat{H}_k^n \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{k} \left(\mathbf{1}(\gamma_{-i}^n > w_{-i}^n) - \frac{F^c(w_{-i}^n + \eta_{-i}^n)}{F^c(\eta_{-i}^n)} \right) \quad \text{for} \quad k = 1, 2, \dots \tag{4.21}$$

where the sequence $\{\gamma_{-i} : i \geq 1\}$ depends on elapsed times of customers in queue. This process is different than (4.14) in upper bound of the summation. Recall that, for given ages $\{\eta_{-1}^n, \eta_{-2}^n, \dots\}$, $\{\gamma_{-i} : i \geq 1\}$ have the complementary cdf

$$\mathbb{P}(\gamma_{-i} > s \,|\, \eta_{-i}^n = x_{-i}^n) = \frac{F^c(s + x_{-i}^n)}{F^c(x_{-i}^n)}$$

for $x_{-i}^n \in \mathbb{R}$. Also consider the filtration $\mathscr{H}_k^n \equiv \sigma\{\gamma_{-i}, w_{-i-1}^n, \eta_{-i-1}^n : 1 \leq i \leq k\}$. Then $\mathbb{E}[|\hat{H}_k^n|] \leq k/\sqrt{n}$ and

$$\mathbb{E}[H_k^n - H_{k-1}^n | \mathscr{H}_{k-1}^n] = \frac{1}{\sqrt{n}} \left(\mathbb{E}[\mathbf{1}(\gamma_{-k}^n > w_{-k}^n) | \mathscr{H}_{k-1}^n] - \frac{F^c(w_{-k}^n + \eta_{-k}^n)}{F^c(\eta_{-k}^n)} \right) = 0.$$

which implies that the process $\{(\hat{H}_k^n, \mathscr{H}_k^n) : k \geq 1\}$ is a discrete-time martingale for each $n \geq 1$.

Our next step is to replace k with $\lfloor ny \rfloor$ and extend the above result to continuous-time setting. To do so we invoke Lemma 4.2 of [4] (also see Theorem 2.26 of [11]). By a direct application of that lemma, we deduce that the continuous-time process $(\hat{H}^n(y), \mathscr{H}^n(y) : y \geq 0) \equiv (\hat{H}_{\lfloor ny \rfloor}^n, \mathscr{H}_{\lfloor ny \rfloor}^n : y \geq 0)$ is a martingale with quadratic variation

$$\langle \hat{H}^n \rangle(y) = \frac{1}{n} \sum_{i=1}^{\lfloor ny \rfloor} \left(\mathbf{1}(\gamma_{-i}^n > w_{-i}^n) - \frac{F^c(w_{-i}^n + \eta_{-i}^n)}{F^c(\eta_{-i}^n)} \right)^2 \equiv \frac{1}{n} \sum_{i=1}^{\lfloor ny \rfloor} \left(\mathbf{1}(\gamma_{-i}^n > w_{-i}^n) - H_{\eta_{-i}^n}^c(w_{-i}^n) \right)^2. \tag{4.22}$$

We next show that the sequence of martingales $(\hat{H}^n(y), \mathscr{H}^n(y) : y \geq 0)$ satisfies the conditions of Theorem 7.1.4. of [5]. It particular, it is required that (i) jumps of the processes $\hat{H}^n(y)$ are asymptotically negligible and (ii) quadratic variation of the processes converges in probability to a limit characterized in Theorem 7.1.1. of [5].

(*i*) *Negligibility of jumps.* We show that condition (*a*) of Theorem 7.1.4. holds. For each $T > 0$, we have $\sup_{0\le t\le T} |\hat{H}^n(t) - \hat{H}^n(t-)| \le 1/\sqrt{n}$ and hence

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{0\le t\le T} |\hat{H}^n(t) - \hat{H}^n(t-)|\right] = 0.$$

which is the desired condition.

(*ii*) *Convergence of quadratic variations.* We prove that the quadratic variation processes given in (4.22) converges in L^2 sense as $n \to \infty$. In particular,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{\lfloor ny\rfloor}\left(\mathbf{1}(\gamma^n_{-i} > w^n_{-i}) - H^c_{\eta^n_{-i}}(w^n_{-i})\right)^2 - \int_0^{Q^{-1}(0,y)} H^c_u(v(-u)-u)H_u(v(-u)-u)\,dQ(0,u)\right)^2\right] \\
&\le 2\,\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{\lfloor ny\rfloor}\left[\left(\mathbf{1}(\gamma^n_{-i} > w^n_{-i}) - H^c_{\eta^n_{-i}}(w^n_{-i})\right)^2 - H^c_{\eta^n_{-i}}(w^n_{-i})H_{\eta^n_{-i}}(w^n_{-i})\right]\right)^2\right] \\
&\quad + 4\,\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{\lfloor ny\rfloor}\left[H^c_{\eta^n_{-i}}(w^n_{-i})H_{\eta^n_{-i}}(w^n_{-i}) - H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})H_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})\right]\right)^2\right] \\
&\quad + 4\,\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{\lfloor ny\rfloor} H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})H_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})\right.\right. \\
&\qquad\qquad \left.\left. - \int_0^{Q^{-1}(0,y)} H^c_u(v(-u)-u)H_u(v(-u)-u)\,dQ(0,u)\right)^2\right] \\
&\le \frac{2}{n^2}\sum_{i=1}^{\lfloor ny\rfloor}\mathbb{E}\left[\left(\mathbf{1}(\gamma^n_{-i} > w^n_{-i}) - H^c_{\eta^n_{-i}}(w^n_{-i})\right)^2\left(H_{\eta^n_{-i}}(w^n_{-i}) - H^c_{\eta^n_{-i}}(w^n_{-i})\right)^2\right] \\
&\quad + \frac{2}{n^2}\mathbb{E}\sum_{i\ne j}\left[\left(\mathbf{1}(\gamma^n_{-i} > w^n_{-i}) - H^c_{\eta^n_{-i}}(w^n_{-i})\right)\left(\mathbf{1}(\gamma^n_{-j} > w^n_{-j}) - H^c_{\eta^n_{-j}}(w^n_{-j})\right)\right. \\
&\qquad\qquad \left.\left(H_{\eta^n_{-i}}(w^n_{-i}) - H^c_{\eta^n_{-i}}(w^n_{-i})\right)\left(H_{\eta^n_{-j}}(w^n_{-j}) - H^c_{\eta^n_{-j}}(w^n_{-j})\right)\right] \\
&\quad + 4\,\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{\lfloor ny\rfloor}\left[H^c_{\eta^n_{-i}}(w^n_{-i})H_{\eta^n_{-i}}(w^n_{-i}) - H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})H_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})\right]\right)^2\right] \qquad (4.23) \\
&\quad + 4\,\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{\lfloor ny\rfloor} H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})H_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})\right.\right. \\
&\qquad\qquad \left.\left. - \int_0^{Q^{-1}(0,y)} H^c_u(v(-u)-u)H_u(v(-u)-u)\,dQ(0,u)\right)^2\right] \qquad (4.24)
\end{aligned}$$

The first sum vanishes as $n \to \infty$ because the summands are bounded by 1 and hence the first term is bounded by $2\lfloor nt\rfloor/n^2 \le 2t/n \to 0$ as $n \to \infty$. Summands of the second sum are independent

conditioned on the sequences $\{\eta^n_{-i}\}$ and $\{w^n_{-i}\}$. Therefore, the conditional expectation of each summand is zero. Hence, the second term is equal to 0.

To prove convergence of (4.23), we first rewrite the summands of (4.23) as

$$\begin{aligned}
&H^c_{\eta^n_{-i}}(w^n_{-i})(1-H^c_{\eta^n_{-i}}(w^n_{-i}))-H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})(1-H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i}))\\
&=H^c_{\eta^n_{-i}}(w^n_{-i})-H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})-\left(H^c_{\eta^n_{-i}}(w^n_{-i})^2-H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})^2\right) \qquad (4.25)
\end{aligned}$$

Next we make use of fluid scale convergence of virtual waiting time process $V_n(t)$, i.e., $V_n \Rightarrow v$ in $\mathcal{D}$, and continuity of the function $\cdot \mapsto H_{\eta^n_{-i}}(\cdot)$ to show that (4.23) converges to 0. In particular, for all $i \geq 1$,

$$H^c_{\eta^n_{-i}}(w^n_{-i}) = H^c_{\eta^n_{-i}}(V_n(-\eta^n_{-i})-\eta^n_{-i}) = H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i}+o(1)) \quad \text{for} \quad n \geq 1$$

due to the fact that $V_n \Rightarrow v$ in $\mathcal{D}$. Combined with (4.25), this implies that summands in (4.23) can be bounded above by

$$\begin{aligned}
&|H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i}+o(1))-H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})|+|H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i}+o(1))^2-H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})^2|\\
&=\left|\frac{F^c(v(-\eta^n_{-i})+o(1))-F^c(v(-\eta^n_{-i}))}{F^c(\eta^n_{-i})}\right|+\left|\frac{F^c(v(-\eta^n_{-i})+o(1))^2-F^c(v(-\eta^n_{-i}))^2}{F^c(\eta^n_{-i})^2}\right|\\
&\leq \frac{K_1\,|o(1)|}{F^c(\eta^n_{-i})}+\frac{K_2\,|o(1)|}{F^c(\eta^n_{-i})^2} \leq K|o(1)|
\end{aligned}$$

where a candidate $K = 2\max\{K_1/m, K_2/m\}$ and $m = \min\{F^c(\eta^n_{-i})^2 : i \geq 1\} > 0$. The equality holds by definition and the inequality holds by differentiability (Lipschitz continuity) of the service-time cdf F. This implies that the squared sum inside expectation in (4.23) is bounded above by $(K|o(1)|\lfloor ny\rfloor/n)^2 \leq (Ky)^2|o(1)| = o(1)$ for all $y \geq 0$. Hence convergence of (4.23) to 0 by dominated convergence theorem.

The summation in (4.24) can be alternatively represented as

$$\begin{aligned}
\frac{1}{n}\sum_{i=0}^{\lfloor ny\rfloor} H^c_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i})H_{\eta^n_{-i}}(v(-\eta^n_{-i})-\eta^n_{-i}) &= \int_0^{\bar{Q}_n^{-1}(0,y)} H^c_u(v(-u)-u)H_u(v(-u)-u)\,d\bar{Q}_n(0,u)\\
&\Rightarrow \int_0^{Q^{-1}(0,y)} H^c_u(v(-u)-u)H_u(v(-u)-u)\,dQ(0,u) \qquad (4.26)
\end{aligned}$$

where the convergence in (4.26) follows from [7]. Having established the convergence in (4.26), convergence in mean square is obtained by first applying continuous mapping theorem with the function $f(x) = x^2$ and then applying dominated convergence theorem by using the fact that both

the summation and the limit integral in (4.26) are bounded by y. Hence (4.24) converges to 0. That completes the proof of convergence of the quadratic variation (4.22).

Having proved conditions (i) and (ii) are indeed satisfied, by Theorem 7.1.4 of [5], we deduce that $\hat{H}^n \Rightarrow \hat{H}$ in $\mathcal{D}$ where $\hat{H}$ is a Gaussian process with independent increments and sample paths in $\mathcal{C}$. Moreover, as implied by the proof of Theorem 7.1.1. of [5], the limit $\hat{H}$ is indeed a time-changed Brownian motion where time-change is the limit of the quadratic variation, i.e.,

$$\hat{H}(y) = \mathcal{B}(\langle \hat{H} \rangle(y)) = \mathcal{B}\left(\int_0^{Q^{-1}(0,y)} H_u^c(v(-u)-u) H_u(v(-u)-u) q(0,u)\, du \right), \quad y \geq 0.$$

where $\mathcal{B}$ is the standard Brownian motion.

Finally, to prove the convergence of (4.14) we consider the compositions $\hat{H}^n(\bar{Q}_n(0, W_n(0)))$ and $\hat{H}^n(\bar{Q}_n(0, (W_n(t) - t)^+))$. By continuous mapping theorem, we have $\hat{H}^n(\bar{Q}_n(0, W_n(0))) \Rightarrow \hat{H}(Q(0, w(0)))$ in $\mathcal{D}$ and $\hat{H}^n(\bar{Q}_n(0, (W_n(t)-t)^+)) \Rightarrow \hat{H}(Q(0, (w(t)-t)^+))$ in $\mathcal{D}$. The limit is obtained by writing (4.14) as $\hat{H}^n(\bar{Q}_n(0, W_n(0))) - \hat{H}^n(\bar{Q}_n(0, (W_n(t)-t)^+))$ and applying continuous mapping theorem once again. Hence

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=Q_n(0,(W_n(t)-t)^+)+1}^{Q_n(0,W_n(0))} \left(\mathbf{1}(\gamma_{-i}^n > w_{-i}^n) - \frac{F^c(w_{-i}^n) + \eta_{-i}^n}{F^c(\eta_{-i}^n)} \right) \\ &\qquad \Rightarrow \mathcal{B}^o \left(\int_{(w(t)-t)^+}^{w(0)} H_u^c(v(-u)-u) H_u(v(-u)-u) q(0,u)\, du \right). \end{aligned} \tag{4.27}$$

due to stationary increments of the standard Brownian motion.

Application of continuous mapping theorem. The fact that the processes $\hat{H}^n(t)$ and $\hat{L}^n(t)$ are independent for all $n \geq 1$ and that respective filtrations $\mathscr{F}_k^n$ and $\mathscr{G}_k^n$ are orthogonal for all $n \geq 1$ implies that

$$(\hat{H}^n, \hat{L}^n, \hat{H}^n + \hat{L}^n, \bar{N}_n, \bar{Q}_n(0,\cdot), W_n) \Rightarrow (\hat{H}, \hat{L}, \hat{H} + \hat{L}, \Lambda, Q(0,\cdot), w) \quad \text{in} \quad \mathcal{D}^6([0,T];\mathbb{R}) \tag{4.28}$$

where joint convergence of last three processes directly follows because the limits are deterministic. Having established (4.28), the desired result follows from continuous mapping theorem. ∎

We now consider the third term in (4.17) without the scaling factor $1/\sqrt{n}$, i.e.,

$$
\begin{aligned}
&E_{n,3}(t) - nE_3(t) \\
&= n\int_{-W_n(0)}^{t-W_n(t)} \widetilde{F}_s^c(V_n(s))\lambda^*(s)\,ds - n\int_{-w(0)}^{t-w(t)} \widetilde{F}_s^c(v(s))\lambda^*(s)\,ds \\
&= n\int_{-W_n(0)}^{t-W_n(t)} \widetilde{F}_s^c(V_n(s))\lambda^*(s)\,ds - n\int_{-w(0)}^{t-w(t)} \widetilde{F}_s^c(V_n(s))\lambda^*(s)\,ds + n\int_{-w(0)}^{t-w(t)} \widetilde{F}_s^c(V_n(s))\lambda^*(s)\,ds \\
&\quad - n\int_{-w(0)}^{t-w(t)} \widetilde{F}_s^c(v(s))\lambda^*(s)\,ds \\
&= n\int_{-W_n(0)}^{-w(0)} \widetilde{F}_s^c(V_n(s))\lambda^*(s)\,ds + n\int_{t-W_n(t)}^{t-w(t)} \widetilde{F}_s^c(V_n(s))\lambda^*(s)\,ds \\
&\quad + n\int_{-w(0)}^{t-w(t)} [\widetilde{F}_s^c(V_n(s)) - \widetilde{F}_s^c(v(s))]\lambda^*(s)\,ds \\
&= \sqrt{n}\widetilde{F}_{-w(0)}^c(\theta_{1,n})\lambda^*(-w(0))\hat{W}_n(0) + \sqrt{n}\widetilde{F}_{t-w(t)}^c(\theta_{2,n}(t))\lambda^*(t-w(t))\hat{W}_n(t) \\
&\quad - \sqrt{n}\int_{-w(0)}^{t-w(t)} \widetilde{f}_s(\theta_{3,n}(s))\hat{V}_n(s)\lambda^*(s)\,ds
\end{aligned} \tag{4.29}
$$

where $\widetilde{f}_s(u) \equiv \frac{\partial}{\partial u}\,\widetilde{F}_s(u)$,

$$V_n(-W_n(0)) \wedge V_n(-w(0)) \le \theta_{1,n} \le V_n(-W_n(0)) \vee V_n(-w(0)), \tag{4.30}$$

$$V_n(t-W_n(t)) \wedge V_n(t-w(t)) \le \theta_{2,n}(t) \le V_n(t-W_n(t)) \vee V_n(t-w(t)), \tag{4.31}$$

$$V_n(t) \wedge v(t) \le \theta_{3,n}(t) \le V_n(t) \vee v(t). \tag{4.32}$$

Having treated all the terms in (4.17), we next derive an SDE for the prelimit processes $W_n(t)$ by using two representations of the process $E_n(t)$. Then we will obtain an integral equation for $\hat{W}_n(t)$ for each $n \ge 1$ and use Gronwall's inequality to prove that $\{\hat{W}_n(t) : n \ge 1\}$ is stochastically bounded in $\mathcal{D}([0,T];\mathbb{R})$. Gronwall's inequality will also be used for establishing the convergence $\hat{W}_n \Rightarrow \hat{W}$ in $\mathcal{D}([0,T];\mathbb{R})$.

From Lemma 4.1, Lemma 4.2 and (4.29), we have

$$
\begin{aligned}
&E_n(t) = E_{n,1}(t) + E_{n,2}(t) + E_{n,3}(t) \\
&= \sqrt{n}\int_0^t \widetilde{F}_{s-w(s)}^c(w(s))\,d\hat{\Lambda}(s-w(s)) + \sqrt{n}\mathcal{B}^o\left(\int_{(w(t)-t)^+}^{w(0)} \widetilde{F}_u^c(v(-u))\widetilde{F}_u(v(-u))q(0,u)\,du\right) \\
&+ \sqrt{n}\mathcal{B}^\nu\left(\int_0^t F^c(v(u))F(v(u))\lambda(u)\,du\right) + \sqrt{n}\widetilde{F}_{-w(0)}^c(\theta_{1,n})\lambda^*(-w(0))\hat{W}_n(0) \\
&+ \sqrt{n}\widetilde{F}_{t-w(t)}^c(\theta_{2,n}(t))\lambda^*(t-w(t))\hat{W}_n(t) - \sqrt{n}\int_{-w(0)}^{t-w(t)} \widetilde{f}_s(\theta_{3,n}(s))\hat{V}_n(s)\lambda^*(s)\,ds \\
&+ nE(t) + o(\sqrt{n})
\end{aligned} \tag{4.33}
$$

We want to rewrite the last integral term in (4.33) as a function of W_n instead of V_n so that we will be able to apply Gronwall's inequality to prove stochastic boundedness of the sequence $\{\hat{W}_n : n \geq 1\}$ in $\mathcal{D}([0,T];\mathbb{R})$. To rewrite the integral term as a function of W_n, we will apply change of variable. However, first, we present some results on the relation between $\hat{W}_n(t)$ and $\hat{V}_n(t)$ that will be useful as we will be applying change of variable.

Let $\Delta V_n(t) \equiv V_n(t) - v(t)$ and $\Delta W_n(t) \equiv W_n(t) - w(t)$. We write

$$\begin{aligned}\Delta V_n(t) &= \Delta W_n(t + V_n(t) + O(1/n)) + w(t + V_n(t)) - w(t + v(t)) + O(1/n) \\ &= \Delta W_n(t + V_n(t) + O(1/n)) + \dot{w}(t + v(t))\Delta V_n(t) + o(\Delta V_n(t)) + O(1/n)\end{aligned}$$

which implies

$$\Delta V_n(t) = \frac{\Delta W_n(t + V_n(t) + O(1/n))}{1 - \dot{w}(t + v(t))} + o(\Delta V_n(t)) + O(1/n). \tag{4.34}$$

Combining (4.34) with (4.8), we deduce that $o(\Delta V_n(t)) = o(1/n)$ since $\Delta V_n(t)$ is of $O(1/n)$. Hence we have

$$\sup_{0 \leq t \leq T} \left| \hat{V}_n(t) - \frac{\hat{W}_n(t + v(t))}{1 - \dot{w}(t + v(t))} \right| = \sqrt{n}\, o(1/n) = o(1/\sqrt{n}) \tag{4.35}$$

which concludes discussion of the relation between $\hat{V}_n(t)$ and $\hat{W}_n(t)$.

We next move on to apply change of variable in the last integral term of (4.33). By using (4.35) and the fluid equations $w(t) = v(t - w(t))$ and $v(t) = w(t + v(t))$, we write

$$\begin{aligned}&\sqrt{n} \int_{-w(0)}^{t-w(t)} \widetilde{f}_s(\theta_{3,n}(s)) \hat{V}_n(s)\, d\Lambda^*(s) \\ &= \sqrt{n} \int_{-w(0)}^{t-w(t)} \widetilde{f}_s(\theta_{3,n}(s)) \left(\frac{\hat{W}_n(s + v(s))}{1 - \dot{w}(s + v(s))} + o(1/\sqrt{n}) \right) d\Lambda^*(s) \\ &= \sqrt{n} \int_0^t \widetilde{f}_{s-w(s)}(\theta'_{3,n}(s)) \left(\frac{\hat{W}_n(s)}{1 - \dot{w}(s)} + o(1/\sqrt{n}) \right) (1 - \dot{w}(s)) \lambda^*(s - w(s))\, ds \\ &= \sqrt{n} \int_0^t \widetilde{f}_{s-w(s)}(\theta'_{3,n}(s))\, \hat{W}_n(s)\, \lambda^*(s - w(s))\, ds + o(1)\end{aligned} \tag{4.36}$$

for some sequence $\{\theta'_{3,n}(t) : n \geq 1\}$ satisfying

$$V_n(t - w(t)) \wedge v(t - w(t)) \leq \theta'_{3,n}(t) \leq V_n(t - w(t)) \vee v(t - w(t)) \tag{4.37}$$

where $\dot{w}(t) \equiv (d/dt)w(t)$. The second inequality in (4.36) holds since $\|\lambda^*\|_t < \infty$ for any $0 < t < \infty$ and $\widetilde{f}$ is a probability density function, and hence, the integral over any subset of its domain is at most 1.

On the other hand, we write

$$E_n(t) = nE(t) + \sqrt{n}\hat{E}(t) + o(\sqrt{n}). \tag{4.38}$$

by the implied convergence $\hat{E}_n \Rightarrow \hat{E}$ in $\mathcal{D}([0,T];\mathbb{R})$ due to Theorem 2 of [12]. In particular, the service completion process at each of n servers is asymptotically identical to an equilibrium renewal process since we assume that the system is asymptotically overloaded. We know from the FCLT for equilibrium process in Theorem 2 of [12] that $\hat{D}_n \Rightarrow \hat{D}$ in $\mathcal{D}([0,T];\mathbb{R})$. Consequently, this implies that $\sup_{0\le t\le T} |\hat{E}(t) - \hat{D}(t)| = o(\sqrt{n})$. Then plugging (4.36) in (4.33) and equating (4.33) with (4.38) yields

$$\begin{aligned}\hat{W}_n(t) = &-\frac{1}{g_n(t)} \int_0^t \widetilde{f}_{s-w(s)}(\theta'_{3,n}(s))\, \hat{W}_n(s)\, \lambda^*(s-w(s))\, ds + \frac{1}{g_n(t)}\, \hat{G}(t) \\ &+ \frac{\widetilde{F}^c_{-w(0)}(\theta_{1,n})\lambda^*(-w(0))}{g_n(t)}\, \hat{W}_n(0) + o(1),\end{aligned} \tag{4.39}$$

where $g_n(t) \equiv -\widetilde{F}^c_{t-w(t)}(\theta_{2,n}(t))\lambda^*(t-w(t))$ and

$$\begin{aligned}\hat{G}(t) \equiv &\int_0^t \widetilde{F}^c_{s-w(s)}(w(s))\, d\hat{\Lambda}(s-w(s)) + \mathcal{B}^o\left(\int_{(w(t)-t)^+}^{w(0)} \widetilde{F}^c_u(v(-u))\widetilde{F}_u(v(-u))q(0,u)\, du\right) \\ &+ \mathcal{B}^\nu\left(\int_0^t F^c(v(u))F(v(u))\lambda(u)\, du\right) - \hat{E}(t).\end{aligned}$$

Then, equation (4.39) implies that

$$\begin{aligned}\left|\hat{W}_n(t)\right| \le &\frac{1}{g_n(t)} \int_0^t \widetilde{f}_{s-w(s)}(\theta'_{3,n}(s))\, |\hat{W}_n(s)|\, \lambda^*(s-w(s))\, ds + \frac{1}{g_n(t)}\, |\hat{G}(t)| \\ &+ \frac{\widetilde{F}^c_{-w(0)}(\theta_{1,n})\lambda^*(-w(0))}{g_n(t)}\, |\hat{W}_n(0)| + o(1).\end{aligned} \tag{4.40}$$

For fixed n, we apply Gronwall's inequality in §1 to (4.40) with

$$\begin{aligned}h_n(t) &= \frac{1}{g_n(t)}|\hat{G}(t)| + \frac{\tilde{F}^c_{-w(0)}(\theta_{1,n})\lambda^*(-w(0))}{g_n(t)}|\hat{W}_n(0)| + o(1), \\ \mu_n(t) &= \frac{1}{g_n(t)}\, \widetilde{f}_{t-w(t)}(\theta'_{3,n}(t))\lambda^*(t-w(t)).\end{aligned}$$

Then we have

$$\begin{aligned}\left|\hat{W}_n(t)\right| &\le h_n(t) + \int_0^t h_n(u) \exp\left(\|\lambda^*/g_n\|_{[u,t]}\right)\, d\mu_n(u) \\ &\le h_n(t) + \exp\left(\|\lambda^*/g_n\|_t\right) \int_0^t h_n(u)\mu_n(u)\, du.\end{aligned} \tag{4.41}$$

The exponential term in (4.41) is bounded because λ^* is bounded over any closed finite interval with $\lambda^*(t) > 0$ for all $t \geq 0$. Furthermore, since $\widetilde{F}^c(t) > 0$ for all $t \geq 0$, $\|g_n\|_t > 0$ for all $t \geq 0$. This implies that the norm in (4.41) bounded by a constant different than 0 over all bounded intervals. Hence the second term in (4.41) is bounded since the integral on the right-hand side is bounded.

The fact that the sequences $\{\hat{G}_n : n \geq 1\}$ and $\{\hat{W}_n(0) : n \geq 1\}$ are stochastically bounded in $\mathcal{D}([0,T];\mathbb{R})$ and in $\mathbb{R}$, respectively, and that the second term in (4.41) is finite for all $n \geq 1$. Together with the discussion in the above paragraph, this implies that the sequence $\{\hat{W}_n : n \geq 1\}$ is stochastically bounded in $\mathcal{D}([0,T];\mathbb{R})$.

Stochastic boundedness of $\{\hat{W}_n : n \geq 1\}$ plays a key role because we want the error caused by replacing $\theta_{1,n}$ by $w(0)$, $\theta_{2,n}(t)$ by $w(t)$, $\theta_{3,n}(t)$ by $v(t)$ and $\theta'_{3,n}(t)$ by $w(t)$ to be asymptotically negligible. More specifically, from (4.30)-(4.32), (4.37) and $W_n \Rightarrow w$, $V_n \Rightarrow v$ in $\mathcal{D}([0,T];\mathbb{R})$, we deduce that $\theta_{1,n} = w(0) + o(1)$, $\theta_{2,n}(t) = v(t - w(t)) + o(1) = w(t) + o(1)$, $\theta_{3,n}(t) = v(t) + o(1)$ and $\theta'_{3,n}(t) = v(t - w(t)) + o(1) = w(t) + o(1)$. When $\theta_{1,n}, \theta_{2,n}(t), \theta_{3,n}(t), \theta'_{3,n}(t)$ are to be replaced by respective expressions on the right-hand side of the equalities, we end up with extra terms where $o(1)$ terms are multiplied by $\hat{W}_n(\cdot)$. To guarantee that such terms are not too irregular, we need stochastic boundedness of $\{\hat{W}_n : n \geq 1\}$. Once stochastic boundedness is proved, such extra terms can be treated as $o(1)$.

In light of the above discussion, (4.39) can be rewritten as

$$\begin{aligned}\hat{W}_n(t) = {} & \frac{1}{\widetilde{F}^c_{t-w(t)}(w(t))\lambda^*(t-w(t))} \int_0^t \widetilde{f}_{s-w(s)}(w(s))\, \hat{W}_n(s)\, \lambda^*(s-w(s))\, ds \\ & - \frac{1}{\widetilde{F}^c_{t-w(t)}(w(t))\lambda^*(t-w(t))}\, \hat{G}(t) - \frac{\widetilde{F}^c_{-w(0)}(w(0))\lambda^*(-w(0))}{\widetilde{F}^c_{t-w(t)}(w(t))\lambda^*(t-w(t))}\, \hat{W}_n(0) + o(1). \qquad (4.42)\end{aligned}$$

In order to show that $\hat{W}_n \Rightarrow \hat{W}$, where $\hat{W}$ satisfies

$$\begin{aligned}\hat{W}(t) = {} & \frac{1}{\widetilde{F}^c_{t-w(t)}(w(t))\lambda^*(t-w(t))} \int_0^t \widetilde{f}_{s-w(s)}(w(s))\, \hat{W}(s)\, \lambda^*(s-w(s))\, ds \\ & - \frac{1}{\widetilde{F}^c_{t-w(t)}(w(t))\lambda^*(t-w(t))}\, \hat{G}(t) - \frac{\widetilde{F}^c_{-w(0)}(w(0))\lambda^*(-w(0))}{\widetilde{F}^c_{t-w(t)}(w(t))\lambda^*(t-w(t))}\, \hat{W}(0), \qquad (4.43)\end{aligned}$$

it suffices to show that $\|\hat{W}_n - \hat{W}\|_T \to 0$ for $\hat{W}_n$ and $\hat{W}$ satisfying (4.42) and (4.43). Equations

(4.42) and (4.43) imply that

$$\begin{aligned}\left|\hat{W}_n(t)-\hat{W}(t)\right| &\le \frac{1}{\widetilde{F}^c_{t-w(t)}(w(t))\lambda^*(t-w(t))}\int_0^t \widetilde{f}_{s-w(s)}(w(s))\,|\hat{W}_n(s)-\hat{W}(s)|\,\lambda^*(s-w(s))\,ds \\ &\quad + \frac{\widetilde{F}^c_{-w(0)}(w(0))\lambda^*(-w(0))}{\widetilde{F}^c_{t-w(t)}(w(t))\lambda^*(t-w(t))}\,|\hat{W}_n(0)-\hat{W}(0)| + o(1) \\ &\equiv \frac{1}{\widetilde{g}_n(t)}\int_0^t |\hat{W}_n(s)-\hat{W}(s)|\,\widetilde{\mu}(s)\,ds + \widetilde{h}_n(t).\end{aligned}$$

Once again a direct application of Gronwall's inequality similar to (4.41) yields

$$\sup_{0\le t\le T}\left|\hat{W}_n(t)-\hat{W}(t)\right| \le \widetilde{h}_n(t) + \exp\left(\|\lambda^*/g_n\|_t\right)\int_0^t \widetilde{h}_n(u)\widetilde{\mu}_n(u)\,du. \tag{4.44}$$

for all $n \ge 1$. This concludes the convergence $\|\hat{W}_n - \hat{W}\|_T \to 0$ since the exponential term is finite, $\widetilde{h}_n \Rightarrow 0$ in $\mathbb{R}$, thanks to $\hat{W}_n(0) \Rightarrow \hat{W}(0)$ in $\mathbb{R}$, and the integral is continuous in the sense of Theorem 11.5.1. of [13].

4.2.2 FCLT for Other Processes

If $b(t,0) > b^{\downarrow} > 0$ for all t, then we know from [6] that the fluid HWT $w(t)$, with $w(0) > 0$, will eventually cross over the 45 degree line. See Figure 5 there. Let $t^* \equiv \inf\{t \ge 0 : w(t) \le t\}$ be the time of the crossover. Therefore, we have (i) $w(t) > t$ for $t \in \mathcal{I}^o \equiv [0, t^*)$, (ii) $w(t^*) = t^*$ and (iii) $w(t) < t$ for $t \in \mathcal{I}^\nu \equiv (t^*, \infty)$. We remark that all old contents (i.e., customers that were in the system at time 0) will either abandon or enter service for $t > t^*$, asymptotically when n is large.

FCLT limits for queue length $\hat{Q}_n$. We next prove the FCLTs for the queue lengths using the FCLT of $\hat{W}_n$ and continuous mapping theorem. Let $Q_n(t)$ be the number of customers entered service in the interval $[0,t]$, let $Q^o_n(t)$ be the number of old customers (who were initial in queue) entered service in the interval $[0,t]$, and let $Q^\nu_n(t)$ be the number of new customers (arriving after time 0) entered service in the interval $[0,t]$. Then we have Letting $n \to \infty$, we have

$$\begin{aligned}\hat{Q}_{n,1}(t) \equiv \frac{1}{\sqrt{n}}\left(Q^\nu_{n,1}(t) + Q^o_{n,1}(t)\right) &\Rightarrow \hat{Q}_1(t) \equiv \int_{(t-w(t))^+}^t F^c(t-s)\,d\hat{N}(s) + \int_0^{(w(t)-t)^+} H^c_u(t)\,d\hat{Q}(0,u) \\ &= \int_{t\vee t^*}^{\Gamma^{-1}(t)} F^c(t-u+w(u))\,d\hat{N}(u-w(u)) - \int_{t\wedge t^*}^{t^*} H^c_{w(u)-u}(t)\,d\hat{Q}(0,w(u)-u) \\ &= \int_t^{\Gamma^{-1}(t)} J_\lambda(t,u)\,d\tilde{Y}(u),\end{aligned} \tag{4.45}$$

where the convergence holds by the FWLLN and FCLT of W_n, along with the continuous mapping theorem. Here $\Gamma(t) \equiv t - w(t)$, $J_\lambda(t,u) \equiv -H^c_{w(u)-u}(t)\mathbf{1}_{\{u\le t^*\}} + F^c(t-u+w(u))\mathbf{1}_{\{u>t^*\}}$. Next, we

have,

$$
\begin{aligned}
&\hat{Q}_{n,2}(t) \equiv \frac{1}{\sqrt{n}}\left(Q^{\nu}_{n,2}(t) + Q^{o}_{n,2}(t)\right) \Rightarrow \hat{Q}_2(t) \\
&\equiv \int_{(t-w(t))^+}^{t} \int_0^{\infty} \mathbf{1}(x > F(t-s))\, d\hat{U}^{\nu}(\Lambda(s), y) + \int_0^{(w(t)-t)^+} \int_0^1 \mathbf{1}(y > H_u(t))\, d\hat{U}^{o}(\bar{Q}(0,u), y), \\
&\stackrel{\mathrm{d}}{=} -\int_{(t-w(t))^+}^{t} \sqrt{F(t-s)F^c(t-s)}\, d\tilde{\mathcal{B}}^{\nu}(\Lambda(s)) - \int_0^{(w(t)-t)^+} \sqrt{H_u(t)H^c_u(t)}\, d\tilde{\mathcal{B}}^{o}(Q(0,u)) \\
&= -\int_{t\vee t^*}^{\Gamma^{-1}(t)} \sqrt{F(t-u+w(u))F^c(t-u+w(u))}\, d\tilde{\mathcal{B}}^{\nu}(\Lambda(u-w(u))) \\
&\quad - \int_{t^*}^{t\wedge t^*} \sqrt{H_{w(u)-u}(t)H^c_{w(u)-u}(t)}\, d\tilde{\mathcal{B}}^{o}(Q(0,w(u)-u)) \\
&= -\int_{t\vee t^*}^{\Gamma^{-1}(t)} \sqrt{F(t-u+w(u))F^c(t-u+w(u))\lambda(u-w(u))(1-\dot{w}(u))}\, d\mathcal{B}^{\nu}(u) \\
&\quad - \int_{t\wedge t^*}^{t^*} \sqrt{H_{w(u)-u}(t)H^c_{w(u)-u}(t)q(0,w(u)-u)(1-\dot{w}(u))}\, d\mathcal{B}^{o}(u) \\
&= \int_t^{\Gamma^{-1}(t)} J_a(t,u)\, d\mathcal{B}(u),
\end{aligned} \tag{4.46}
$$

where

$$
\begin{aligned}
J_a(t,u) \quad \equiv \quad & -\sqrt{H_{w(u)-u}(t)H^c_{w(u)-u}(t)q(0,w(u)-u)(1-\dot{w}(u))}\mathbf{1}_{\{u\le t^*\}} \\
& -\sqrt{F(t-u+w(u))F^c(t-u+w(u))\lambda(u-w(u))(1-\dot{w}(u))}\mathbf{1}_{\{u>t^*\}}.
\end{aligned}
$$

Next, let

$$
Q(t) \equiv \int_{(t-W_n(t))^+}^{t} F^c(t-s)\lambda(s)\, ds + \int_0^{(W_n(t)-t)^+} H^c_u(t)\, dQ(0,u),
$$

we have

$$
\begin{aligned}
\hat{Q}_{n,3}(t) &\equiv \frac{1}{\sqrt{n}}\left(Q^{\nu}_{n,3}(t) + Q^{o}_{n,3}(t) - n\,Q(t)\right) \\
&= \sqrt{n}\int_{(t-W_n(t))^+}^{t-w(t)} F^c(t-s)\lambda(s)\, ds + \sqrt{n}\int_{w(t)-t}^{(W_n(t)-t)^+} H^c_u(t)\, dQ(0,u) \\
&= \hat{W}_n(t)F^c(w(t))\lambda(t-w(t))\mathbf{1}_{\{W_n(t)\le t\}} + \hat{W}_n(t)H^c_{w(t)-t}(t)q(0,w(t)-t)\mathbf{1}_{\{W_n(t)>t\}} \\
&\Rightarrow \hat{W}(t)F^c(w(t))\lambda(t-w(t))\mathbf{1}_{\{w(t)\le t\}} + \hat{W}(t)H^c_{w(t)-t}(t)q(0,w(t)-t)\mathbf{1}_{\{w(t)>t\}} \\
&= \hat{W}(t)q(t,w(t)) \equiv \hat{Q}_3(t).
\end{aligned} \tag{4.47}
$$

Although $\hat{Q}_3$ involves $\hat{W}$, which involves $\hat{N}$, $\hat{Q}(0,\cdot)$, $\mathcal{B}^{\nu}_a$ and $\mathcal{B}^{o}_a$, $\hat{Q}_3$ and $\hat{Q}_2$ are independent because $\mathcal{B}^{\nu}_a$ and $\mathcal{B}^{o}_a$ have independent increments and the intervals of integrals do not overlap. However, $\hat{Q}_3$ and $\hat{Q}_1$ may not be independent (although the intervals of integrals do not overlap), because

the process $\hat{Q}(0,\cdot)$ may not have independent increments. After all, we have the joint convergence: Combining (4.45), (4.46) and (4.47), we have

$$\hat{Q}_n(t) \Rightarrow \hat{Q}(t) \equiv \hat{Q}_0(t) + \hat{Q}_1(t) + \hat{Q}_2(t) + \hat{Q}_3(t) \equiv \hat{Q}_0(t) + \hat{Q}_1^{\nu,*}(t) + \hat{Q}_1^{o,*}(t) + \hat{Q}_2^{\nu,*}(t) + \hat{Q}_2^{o,*}(t) + \hat{Q}_3(t),$$

where the six terms $\hat{Q}_0(t)$, $\hat{Q}_1^{\nu,*}(t)$, $\hat{Q}_1^{o,*}(t)$, $\hat{Q}_2^{\nu,*}(t)$, $\hat{Q}_2^{o,*}(t)$ and $\hat{Q}_3(t)$ are independent with

$$\begin{aligned}
\hat{Q}_0(t) &\equiv \hat{W}(0)H(t,0)q(t,w(t)) \\
\hat{Q}_1(t) &\equiv \int_0^t q(t,w(t))\frac{H(t,u)K_\lambda(u)}{q(u,w(u))}\,d\tilde{Y}(u) + \int_t^{\Gamma^{-1}(t)} J_\lambda(t,u)\,d\tilde{Y}(u) \\
&= \int_0^{\Gamma^{-1}(t)} L_\lambda(t,u)\,d\tilde{Y}(u) \\
&\equiv \hat{Q}_1^{\nu,*}(t) + \hat{Q}_1^{o,*}(t) \equiv \int_{t^*}^{\Gamma^{-1}(t)} L_\lambda^\nu(t,u)\,d\hat{N}(u-w(u)) + \int_0^{t^*} L_\lambda^o(t,u)\,d\hat{Q}(0,w(u)-u) \\
&\quad \text{where} \quad L_\lambda(t,u) \equiv q(t,w(t))\frac{H(t,u)K_\lambda(u)}{q(u,w(u))}\,\mathbf{1}_{\{u\le t\}} + J_\lambda(t,u)\mathbf{1}_{\{u>t\}}, \\
&\qquad\qquad L_\lambda^\nu(t,u) \equiv \frac{q(t,w(t))H(t,u)F^c(w(u))}{q(u,w(u))}\,\mathbf{1}_{\{u\le t\}} + F^c(t-u+w(u))\mathbf{1}_{\{u>t\}}, \\
&\qquad\qquad L_\lambda^o(t,u) \equiv \frac{q(t,w(t))H(t,u)H_{w(u)-u}(u)}{q(u,w(u))}\,\mathbf{1}_{\{u\le t\}} - H^c_{w(u)-u}(t)\mathbf{1}_{\{u>t\}}, \\
\hat{Q}_2(t) &\equiv \int_0^t q(t,w(t))\frac{H(t,u)K_a(u)}{q(u,w(u))}\,d\mathcal{B}_a(u) + \int_t^{\Gamma^{-1}(t)} J_a(t,u)\,d\mathcal{B}_a(u) \\
&= \int_0^{\Gamma^{-1}(t)} L_a(t,u)\,d\mathcal{B}_a(u) \\
&\equiv \hat{Q}_2^{\nu,*}(t) + \hat{Q}_2^{o,*}(t) \equiv \int_{t^*}^{\Gamma^{-1}(t)} L_a^\nu(t,u)\,d\mathcal{B}_a^\nu(u) + \int_0^{t^*} L_a^o(t,u)\,d\mathcal{B}_a^o(u) \\
&\quad \text{where} \quad L_a(t,u) \equiv q(t,w(t))\frac{H(t,u)K_a(u)}{q(u,w(u))}\,\mathbf{1}_{\{u\le t\}} + J_a(t,u)\mathbf{1}_{\{u>t\}}, \\
&\qquad\qquad L_a^\nu(t,u) \equiv -\frac{q(t,w(t))H(t,u)\sqrt{F(w(u))b(u,0)}}{q(u,w(u))}\,\mathbf{1}_{\{u\le t\}} \\
&\qquad\qquad\qquad - \sqrt{F(t-u+w(u))F^c(t-u+w(u))\lambda(u-w(u))(1-\dot{w}(u))}\mathbf{1}_{\{u>t\}}, \\
&\qquad\qquad L_a^o(t,u) \equiv -\frac{q(t,w(t))H(t,u)\sqrt{H_{w(u)-u}(u)b(u,0)}}{q(u,w(u))}\,\mathbf{1}_{\{u\le t\}} \\
&\qquad\qquad\qquad - \sqrt{H_{w(u)-u}(t)H^c_{w(u)-u}(t)q(0,w(u)-u)(1-\dot{w}(u))}\mathbf{1}_{\{u>t\}}, \\
\hat{Q}_3(t) &\equiv -q(t,w(t))\int_0^t \frac{H(t,u)}{q(u,w(u))}d\hat{E}(u).
\end{aligned}$$

Remark 4.1 (Separation of variability for $\hat{Q}(t)$)**.** We now explain the meaning of the six terms: (i) $\hat{Q}_0(t)$ captures the randomness of the initial HWT (as a function of $\hat{W}(0)$; (ii) $\hat{Q}_1^{\nu,*}(t)$ captures the randomness of the new arrivals after time 0 (as a function of $\hat{N}$); (iii) $\hat{Q}_1^{o,*}(t)$ captures the

randomness of the ages of customers in queue at time 0 (as a function of $\hat{Q}(0,\cdot)$; (iv) $\hat{Q}_2^{\nu,*}(t)$ captures the randomness of the remaining patience times of customers in queue at time 0; (v) $\hat{Q}_2^{o,*}(t)$ captures the randomness of the patience times of new arrivals after time 0; (vi) $\hat{Q}_3(t)$ captures the randomness of the service times of all customers (as a function of $\hat{E}$).

References

[1] A. K. Aras, X. Chen, and Y. Liu. Many-server gaussian limits for overloaded non-markovian queues with customer abandonment. *Queueing Systems*, 2017.

[2] A. K. Aras, Y. Liu, and W. Whitt. Heavy-traffic limit for the initial content process. *Stochastic Systems*, 1:95–142, 2017.

[3] P. Billingsley. *Convergence of Probability Measures*. Wiley-Interscience, second edition, 1999.

[4] J. G. Dai and S. He. Customer abandonment in many-server queues. *Mathematics of Operations Research*, 35(2):347–362, May 2010.

[5] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, 1986.

[6] Y. Liu and W. Whitt. The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems*, 71:405–444, 2012.

[7] Y. Liu and W. Whitt. A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. *Operations Research Letters*, 40:307–312, 2012.

[8] Y. Liu and W. Whitt. Many-server heavy-traffic limits for queues with time-varing parameters. *The Annals of Applied Probability*, 24:378–421, 2014.

[9] A. Mandelbaum, W. A. Massey, and Reiman. Strong approximations for markovian service networks. *Queueing Systems*, 30:149–201, 1998.

[10] G. Pang and W. Whitt. Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems*, 65:325–364, 2010.

[11] P. Protter. *Stochastic Integration and Differential Equations*. Springer, 2005.

[12] W. Whitt. Queues with superposition arrival process in heavy traffic. *Stochastic Processes and their Applications*, 21:81 – 91, 1985.

[13] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Ther Application to Queues*. Springer, 2002.