HEAVY-TRAFFIC LIMITS FOR MANY-SERVER QUEUES WITH SERVICE INTERRUPTIONS

by

Guodong Pang and Ward Whitt

IEOR Department
Columbia University
{gp2224, ww2240}@columbia.edu

Abstract

We establish many-server heavy-traffic limits for $G/M/n + M$ queueing models, allowing customer abandonment (the $+M$), subject to exogenous regenerative service interruptions. With unscaled service interruption times, we obtain a FWLLN for the queue-length process, where the limit is an ordinary differential equation in a two-state random environment. With asymptotically negligible service interruptions, we obtain a FCLT for the queue-length process, where the limit is characterized as the pathwise unique solution to a stochastic integral equation with jumps. When the arrivals are renewal and the interruption cycle time is exponential, the limit is a Markov process, being a jump-diffusion process in the QED regime and an O-U process driven by a Levy process in the ED regime (and for infinite-server queues). A stochastic-decompostion property of the steady-state distribution of the limit process in the ED regime (and for infinite-server queues) is obtained.

Keywords: many-server queues, service interruptions, heavy-traffic limits, Skorohod M_1 topology, continuous mapping theorem, jump diffusion process, Levy-driven O-U process, stochastic-decomposition.

December 31, 2008

1. Introduction

The purpose of this paper is to study the impact of service interruptions on the performance of queueing systems. We consider exogenous regenerative service interruptions in many-server queues. We assume that some proportion of the servers cease functioning during the interruptions, focusing especially on the case of large proportions. We also assume that the arrivals continue and all customers remain in the system during the interruption. Customers in service when the interruption begins complete their remaining service after the interruption ends. However, we assume that the service times are exponentially distributed, so that the remaining service times are distributed the same as if the service started over when the interruption ends. Since customers do not leave in response to the interruption, this is a worst-case scenario with respect to the congestion impact upon other customers. As elaborated upon in Pang and Whitt (2008a), large scale makes the system more vulnerable to service interruptions when many servers are unable to function during the interruptions and, surprisingly, even infrequent short service interruptions can have a dramatic impact on congestion.

We quantify the performance impact of the service interruptions by establishing heavy-traffic stochastic-process limits for the queue-length (number-in-system) process in the $G/M/n+M$ model, allowing customer abandonment (the $+M$) and having general arrivals. We consider the many-server heavy-traffic limiting regime in which the number of servers, n, and the arrival rate, λ_n, go to infinity, while the service rate and abandonment rate remain unchanged. The stochastic-process limits we establish here are natural extensions of the conventional heavy-traffic stochastic-process limits for single-server systems with service interruptions in Kella and Whitt (1990), Chen and Whitt (1993) and §14.7 in Whitt (2002). However, the scalings for single-server systems and many-server systems are very different. In the conventional heavy-traffic limit, time is scaled so that both the times between interruptions and the lengths of the interruptions are allowed to grow. In contrast, for many-server systems, we do not scale time. Hence the times between service interruptions and the lengths of interruptions can remain unchanged in the heavy-traffic limit. In fact, the interruptions can even have a significant impact if the durations of the interruptions are asymptotically negligible. Then, in the many-server heavy-traffic limit, at each time the system is working with probability one and yet the interruptions have an impact through jumps in the limit of the scaled queue-length process. As usual with many-server heavy-traffic limits, we will consider three limiting regimes: quality-driven (QD), quality-and-efficiency-driven (QED), and efficiency-driven (ED); see Halfin and Whitt (1981) and Garnett et al. (2002). We have two types

of scalings for the service interruption durations: unscaled and asymptotically negligible, see §§2.2 and 2.3. With unscaled service interruptions, a stochastic fluid approximation for the queue-length process is obtained in Theorem 3.1 in all three regimes, which is in the same spirit as the previous fluid approximations for single-server systems in a random environment in Chen and Yao (1992), Kella and Whitt (1992), and Choudhury et al. (1997). Conditional on the service-interruption process and the functioning-server process, the fluid limit is deterministic and satisfies a nonlinear ordinary differential equation (ODE) in each of the two environment states (interruptions or no interruptions).

With asymptotically negligible service interruptions, we obtain the same deterministic fluid limit for the queue-length processes as without service interruptions in all three regimes (Theorem 3.2), but new refined stochastic limits for the queue-length processes in the QED and ED regimes (Theorems 3.3 and 3.4). The asymptotically negligible service interruptions produce unmatched jumps in the refined stochastic limits, which requires the Skorohod M_1 toplogy; see Chapter 12 of Whitt (2002). In order to apply the continuous mapping theorem, we need the mapping defined by the integral representation of the queue-length processes to be continuous in the Skorohod M_1 topology, which is established in Pang and Whitt (2008b). The continuity of this mapping also allows us to consider the more general arrival processes, e.g., heavy-tailed interarrival times; see the FCLT in the modified QED regime with asymptotically negligible service interruptions in Theorem 4.1.

The refined stochastic limits for the queue-length process are characterized by the pathwise unique solution to a stochastic integral equation with jumps. If the arrivals are renewal with interarrival times having a finite second moment and the cycle time of interruptions is exponential, then the limit processes are special Markov processes, and so are relatively tractable. The refined stochastic limit in the QED regime is a jump-diffusion process, while the refined stochastic limit in the ED regime (as a special case, for infinite-server queues) is an Ornstein-Uhlenbeck (O-U) process driven by a Levy process (a Brownian motion with drift plus a compound Poisson process). The size of jumps in the QED limit process depends on the service rate change, the proportion of non-functioning servers and the duration of the limit of the scaled down times, while the size of jumps in the ED limit process also depends on the customer abandonment rate change. The steady-state distribution of the limit queue-length process in the ED regime (and for infinite-server queues) is given explicitly by its characteristic function and can be decomposed into two independent random variables: one is that without service interruptions and the other represents the effect of the service interruptions. We refer to Jayawardene and Kella (1996), Baykal-Gursoy and Xiao

(2004), D'Auria (2007), Falin (2008), Tian and Zhang (2003a,b,c) and references therein for related work on the stochastic decomposition of the steady-state distribution of the queue-length process for many-server (infinite-server) queues with service interruptions or vacations. For conventional heavy-traffic limits, decompositions were discussed by Kella and Whitt (1990, 1991). As can be seen from Kella and Whitt (1990), this work contributes to a large body of literature on queues with service interruptions or vacations. For other related work on this topic, see White and Christie (1958), Mitrany and Avi-Itzhak (1968), O'Cinneide and Purdue (1986), Chao and Zhao (1997), Tian and Zhang (2003a,b,c), Altman and Uri (2006) and references therein.

Organization of the Paper

In §2, we start by more carefully describing the model, the many-server heavy-traffic limiting regimes, and the exogenous service interruptions. In §3, we state the main results: fluid approximations and their stochastic refinements in all three regimes for many-server queues and for infinite-server queues, and a stochastic-decomposition property of the steady-state distributions in the ED regime and for infinite-server queues. In §4, we state stochastic refinements for low impact interruptions and for busty arrivals in the QED regime. We give the proofs in §5 and conclude in §6.

2. Preliminaries and Assumptions

2.1. The Many-Server Heavy-Traffic Limiting Regimes

We consider a sequence of $G/M/n+M$ queueing models indexed by the number of servers, n, and let $n \to \infty$. For each $n \geq 1$, customers arrivals are general (the G), the n parallel homogenous servers have independent exponential service times (the M) and customers waiting in the queue have independent exponential patience times (the $+M$). We assume that arrival processes, service times and customer abandonments are mutually independent and that all the servers are functioning at time 0.

For each model, service interruptions occur exogenously, independent of the system described above. When interruptions occur, some or even all of the servers will stop working while arrivals keep joining the queue. If busy servers stop functioning because of an interruption, then the customers that were being served will be served to the extent possible by remaining functioning servers, while any excess customers remain to complete their service where they left off when the unavailable servers become available again. (We do not focus on the individual customer and server experience. Hence it suffices to assume that the customers selected to move to functioning servers

are chosen independently at random.)

We assume that the service rate of each server is $\mu_1 > 0$ and the customer abandonment rate is $\theta_1 \geq 0$ when there is no service interruption, and that the service rate is $\mu_2 \geq 0$ for functioning servers and the customer abandonment rate is $\theta_2 \geq 0$ during a service interruption. We assume that $\mu_1 \geq \mu_2$ and $\theta_1 \leq \theta_2$. (It is natural, but not really crucial. Inequalities among these parameters can be used to ensure that the jump terms in (3.3), (3.6) and (3.9) are all positive, but that is not required. It doesn't affect the correctness of the mathematical results if these conditions are not assumed.) For each $n \geq 1$, let $\{\eta_{n,k} : k \geq 1\}$ be a sequence of independent and identically distributed (i.i.d.) random variables, taking integer values from 0 to n, where for each k, $\eta_{n,k}$ is the number of functioning servers among the n servers when the k^{th} service interruption occurs. We assume that

$$\frac{\eta_{n,k}}{n} \Rightarrow \eta_k \quad \text{for all} \quad k \quad \text{as} \quad n \to \infty, \tag{2.1}$$

where $\Rightarrow$ denotes convergence in distribution and $\{\eta_k : k \geq 1\}$ is a sequence of i.i.d. random variables, taking values in $[0, 1]$. Here $\eta_k = 0$ means that all servers stop functioning when the interruption occurs, referred as *total-failure model*, while $\eta_k = 1$ means that all servers remain functioning but with a lower service rate $\mu_2 < \mu_1$. (This randomness assumption on the servers is different from that in Atar (2008), where the number of servers and their service rates are both assumed to be random with some structure.)

Let $D \equiv D([0, \infty), \mathbb{R})$ denote the function space of all right-continuous real-valued functions on $[0, \infty)$ with left-limits everywhere in $(0, \infty)$; see Billingsley (1999) and Whitt (2002) for background. We will make use of the Skorohod M_1 topology as well as the familiar Skorohod J_1 topology. Let $(D^k, M_1) \equiv (D, M_1) \times \cdots \times (D, M_1)$ be the k-fold product of (D, M_1) with the product topology. In contrast, let (D_k, M_1) be the space of $\mathbb{R}^k$-valued functions $D([0, \infty), \mathbb{R}^k)$ with the direct M_1 topology. Convergence in D_k implies convergence in D^k with any of the Skorohod topologies, but not conversely.

Let $A_n \equiv \{A_n(t) : t \geq 0\}$ be the arrival counting process in the n^{th} model with arrival rate $\lambda_n \equiv \lim_{t\to\infty} A_n(t)/t \in (0, \infty)$ and assume $\lambda_n/n \to \lambda \in (0, \infty)$ as $n \to \infty$. Let the associated fluid-scaled and diffusion-scaled arrival processes be $\bar{A}_n \equiv \{\bar{A}_n(t) : t \geq 0\}$ and $\hat{A}_n \equiv \{\hat{A}_n(t) : t \geq 0\}$, defined by

$$\bar{A}_n(t) = \frac{A_n(t)}{n}, \quad \hat{A}_n(t) \equiv \frac{A_n(t) - \lambda_n t}{\sqrt{n}}, \quad t \geq 0.$$

We assume that the arrival processes satisfy a functional central limit theorem (FCLT); i.e.,

$$\hat{A}_n \Rightarrow \hat{A} \quad \text{in} \quad (D, M_1) \quad \text{as} \quad n \to \infty. \tag{2.2}$$

Here we assume the weak convergence in the Skorohod M_1 topology to allow for more general arrival processes, as in §6.3 of Whitt (2002). When the arrival processes are renewal with interarrival times having a finite second moment, the limit process will be a Brownian motion (with time change), $\hat{A} \stackrel{\mathrm{d}}{=} \sqrt{\lambda c_a^2} B$, where the constant c_a^2 is the squared coefficient of variation (SCV, variance divided by the square of the mean) of an interarrival time and B is a standard Brownian motion. Since Brownian motion has continuous paths, the M_1 convergence is equivalent to uniform convergence on bounded subintervals.

The FCLT above implies that an associated functional weak law of large numbers (FWLLN) holds for the arrival processes; i.e.,

$$\bar{A}_n \Rightarrow \lambda e \quad \text{in} \quad D \quad \text{as} \quad n \to \infty, \tag{2.3}$$

where $e(t) = t$ for all $t \geq 0$.

Let $\rho_n \equiv \lambda_n / n\mu_1$ be the traffic intensity and assume that

$$\sqrt{n}(1 - \rho_n) \to \beta, \quad \text{as} \quad n \to \infty, \tag{2.4}$$

where β takes values in $\mathbb{R} \cup \{\pm\infty\}$. We obtain the quality-driven (QD) regime, the quality-and-efficiency-driven (QED) regime, and the efficiency-driven (ED) regime, respectively, when $\beta = +\infty$, $-\infty < \beta < +\infty$ and $\beta = -\infty$. The canonical examples for the QD and ED regimes are fixed traffic intensities, with $\rho_n = \rho < 1$ for all n with QD, and $\rho_n = \rho > 1$ for all n with ED, which is achieved by letting $\lambda_n = \lambda n$ for all n with $\lambda < \mu_1$ for QD and with $\lambda > \mu_1$ for ED. The FWLLN and FCLT for the queue-length processes in the QD and ED regime in Theorems 3.1, 3.2 and 3.4 assume these canonical examples.

For each $n \geq 1$, let $Q_n \equiv \{Q_n(t) : t \geq 0\}$ be the queue-length process, where for each $t \geq 0$, $Q_n(t)$ represents the number of customers in the n^{th} model at time t. Assume that the initial conditions $Q_n(0)$ are independent of the arrival process A_n, service times, customer patience times, service interruptions and $\{\eta_{n,k} : k \geq 0\}$. Define the fluid-scaled and diffusion-scaled processes $\bar{Q}_n \equiv \{\bar{Q}_n(t) : t \geq 0\}$ and $\hat{Q}_n \equiv \{\hat{Q}_n(t) : t \geq 0\}$ by

$$\bar{Q}_n(t) \equiv \frac{Q_n(t)}{n}, \quad \hat{Q}_n(t) \equiv \frac{Q_n(t) - n}{\sqrt{n}}, \quad t \geq 0. \tag{2.5}$$

2.2. Unscaled Exogenous Service Interruptions

We define the exogenous service-interruption process by the regenerative *up-down* (or *on-off*) cycles of the servers, specified by the sequence of independent random vectors $\{(u_{n,k}, v_{n,k}) : k \geq 1\}$, where $u_{n,k}$ and $v_{n,k}$ denote the k^{th} up (on) time and k^{th} down (off) time of the servers in the n^{th} queueing

model, respectively. We assume that $v_{n,k}, k \geq 1$, are i.i.d. and $u_{n,k}, k \geq 2$, are i.i.d., allowing $u_{n,1}$ to have a different distribution from $u_{n,k}, k \geq 2$. For simplicity, we assume that $u_{n,i}, v_{n,i} > 0$ for all i. The renewal times $\{T_{n,k} : k \geq 0\}$ are defined by

$$T_{n,k} \equiv \sum_{i=1}^{k} (u_{n,i} + v_{n,i}), \quad \text{for} \quad k \geq 1, \quad \text{and} \quad T_{n,0} = 0.$$

Thus the associated delayed renewal counting process $N_n \equiv \{N_n(t) : t \geq 0\}$ is defined by

$$N_n(t) \equiv \max\{k \geq 0 : T_{n,k} \leq t\}, \quad t \geq 0. \tag{2.6}$$

Define the availability process (random environment) of the servers, $U_n \equiv \{U_n(t) : t \geq 0\}$, by

$$U_n(t) \equiv \begin{cases} 1, & T_{n,k} \leq t < T_{n,k} + u_{n,k+1}, \quad \text{for} \quad k \geq 0, \\ 0, & T_{n,k} + u_{n,k+1} \leq t < T_{n,k+1}, \quad \text{for} \quad k \geq 0. \end{cases}$$

The cumulative up-time process $C_{U,n} \equiv \{C_{U,n}(t) : t \geq 0\}$ is defined by

$$C_{U,n}(t) \equiv \int_0^t U_n(s) ds, \quad t \geq 0.$$

The cumulative down-time process $C_{D,n} \equiv \{C_{D,n}(t) : t \geq 0\}$ is defined by $C_{D,n}(t) \equiv t - C_{U,n}(t)$ for each $t \geq 0$.

Define the process counting the number of functioning servers at each time, $\eta_n \equiv \{\eta_n(t) : t \geq 0\}$, by

$$\eta_n(t) \equiv \begin{cases} \eta_{n,N_n(t)+1}, & T_{n,k} + u_{n,k+1} \leq t < T_{n,k+1}, \quad \text{for} \quad k = N_n(t), \\ n, & T_{n,k} \leq t < T_{n,k} + u_{n,k+1}, \quad \text{for} \quad k \neq N_n(t). \end{cases} \tag{2.7}$$

In (2.7), $\eta_{n,N_n(t)+1}$ represents the number of functioning servers when a service interruption is occurring at time t, where that service interruption is the $(N_n(t)+1)^{\text{th}}$ service interruption, by the definition of N_n in (2.6).

For the *unscaled service interruptions*, we impose the following assumption on the down times of the service interruptions

$$\{(u_{n,k}, v_{n,k}) : k \geq 1\} \Rightarrow \{(u_k, v_k) : k \geq 1\} \quad \text{in} \quad (\mathbb{R}^2)^\infty \quad \text{as} \quad n \to \infty, \tag{2.8}$$

where $u_k, v_k > 0$ for each k with probability 1 (w.p.1). This implies that

$$\{T_{n,k} : k \geq 0\} \Rightarrow \{T_k : k \geq 0\} > T_{k-1} \quad \text{in} \quad \mathbb{R}^\infty \quad \text{as} \quad n \to \infty, \tag{2.9}$$

where

$$T_k \equiv \sum_{i=1}^{k} (u_i + v_i), \quad \text{for} \quad k \geq 1, \quad \text{and} \qquad T_0 \equiv 0. \tag{2.10}$$

As a consequence, we have the following elementary lemma. We not only have convergence in (D^3, J_1), but also in (D_3, J_1) because the discontinuity points of the component processes coincide.

Lemma 2.1. *For the unscaled service interruptions, satisfying* (2.8),

$$(N_n, U_n, n^{-1}\eta_n) \Rightarrow (N, U, \eta) \quad in \quad (D_3, J_1) \quad as \quad n \to \infty,$$

where the processes $N \equiv \{N(t) : t \geq 0\}$, $U \equiv \{U(t) : t \geq 0\}$ *and* $\eta \equiv \{\eta(t) : t \geq 0\}$ *are defined by*

$$N(t) \equiv \max\{k \geq 0 : T_k \leq t\}, \quad t \geq 0, \tag{2.11}$$

$$U(t) \equiv \begin{cases} 1, & T_k \leq t < T_k + u_{k+1}, \quad for \quad k \geq 0, \\ 0, & T_k + u_{k+1} \leq t < T_{k+1}, \quad for \quad k \geq 0. \end{cases}$$

and

$$\eta(t) \equiv \begin{cases} \eta_{N(t)+1}, & T_k + u_{k+1} \leq t < T_{k+1}, \quad for \quad k = N(t), \\ 1, & T_k \leq t < T_k + u_{k+1}, \quad for \quad k \neq N(t). \end{cases}$$

2.3. Asymptotically Negligible Service Interruptions

We now introduce an alternative limiting regime in which the down times decrease to 0. Instead of assumption (2.8), we assume that

$$\{(u_{n,k}, \sqrt{n}v_{n,k}) : k \geq 1\} \Rightarrow \{(u_k, v_k) : k \geq 1\} \quad \text{in} \quad (\mathbb{R}^2)^\infty \quad \text{as} \quad n \to \infty, \tag{2.12}$$

where again $u_k, v_k > 0$ for each k w.p.1. We refer to this assumption as *asymptotic negligible service interruptions*. This implies that (2.9) holds with

$$T_k \equiv \sum_{i=1}^{k} u_i > T_{k-1}, \quad \text{for} \quad k \geq 1, \quad \text{and} \qquad T_0 \equiv 0. \tag{2.13}$$

Let the distribution function of v_1 be G and assume that $E[v_1] = m_v < \infty$. Under this assumption, if in addition the random variables u_k, $k \geq 1$, are nonlattice, then $(U_n(t), n^{-1}\eta_n(t)) \Rightarrow (1, 1)$ in $\mathbb{R}^2$ as $n \to \infty$ for each $t \geq 0$, as if there were no interruptions at all. However, under this assumption, the processes $(U_n, n^{-1}\eta_n)$ will not converge to the deterministic processes (ω, ω) in D^2 with $\omega(t) = 1$ for all $t \geq 0$ in any of the Skorohod topologies; see Example 11.6.1 on p. 388 in [36]. Note that this is very different from the convergence of $(U_n, n^{-1}\eta_n)$ in Lemma 2.1.

Define the scaled cumulative down-time process of servers $V_n \equiv \{V_n(t) : t \geq 0\}$ by

$$V_n(t) \equiv \sqrt{n}C_{D,n}(t) = \sqrt{n}\int_0^t (1 - U_n(s))ds, \quad t \geq 0, \tag{2.14}$$

and the associated "lost service" process $R_n \equiv \{R_n(t) : t \geq 0\}$ by

$$R_n(t) \equiv \sqrt{n}\int_0^t \Big(1 - \frac{\eta_n(s)}{n}\Big)(1 - U_n(s))ds, \quad t \geq 0.$$

Lemma 2.2. *For the asymptotically negligible service interruptions,*

$$\big(N_n, V_n, R_n, C_{U,n}\big) \Rightarrow (N, V, R, e) \quad \textit{in} \quad (D, J_1) \times (D_3, M_1) \quad \textit{as} \quad n \to \infty, \tag{2.15}$$

where $V \equiv \{V(t) : t \geq 0\}$ *and* $R \equiv \{R(t) : t \geq 0\}$ *are defined by*

$$V(t) \equiv \sum_{k=1}^{N(t)} v_k, \quad R(t) \equiv \sum_{k=1}^{N(t)} v_k(1 - \eta_k), \quad t \geq 0,$$

N *is defined in* (2.11) *with* T_k *in* (2.13) *and* $e(t) \equiv t$ *for all* $t \geq 0$.

Remark. Convergence in (2.15) cannot be strengthened to (D_4, M_1) because, the limit processes N, V and R all have common discontinuities, but the converging processes N_n increase in jumps, whereas V_n and R_n have continuous sample paths. The convergence in (D_4, M_1) requires a single parametric representation of (N_n, V_n), but the requirements for the single time component are incompatible. There is no difficulty with the product topology because then N_n and V_n can have separate parametric representations. ∎

We remark that the T_k's in (2.10) and (2.13) are defined differently, but in the context it is easy to see which definition is used. We will not use the convergence of η_n/n for asymptotically negligible service interruptions in the proofs, but the convergence $R_n \Rightarrow R$ will play a key role. Note that the converging processes V_n and R_n have continuous sample paths, but their limit processes V and R have discontinuous sample paths. Thus, the weak convergence will not hold in the usual Skorohod J_1 topology because of the unmatched jumps, as discussed in Chapter 6 of Whitt (2002). However, since the processes $C_{U,n}$ and the limit process e are all continuous, the convergence of $C_{U,n}$ to e in the M_1 topology is actually equivalent to uniform convergence on compact intervals. We prove Lemma 2.2 in §5.4.

3. Main Results

3.1. Unscaled Service Interruptions

With the framework in §2.2, we can establish a fluid limit that is valid in all three limiting regimes. The proof is in §5.3. We use the conventional notations: $x^+ = \max\{x, 0\}$, $x \vee y = \max\{x, y\}$ and $x \wedge y = \min\{x, y\}$ for any $x, y \in \mathbb{R}$.

Theorem 3.1. (FWLLN with unscaled service interruptions) *Consider the* $G/M/n+M$ *model with unscaled service interruptions, in the QD, QED or ED regime. If there exists a random variable* $\bar{Q}(0)$ *such that* $\bar{Q}_n(0) \Rightarrow \bar{Q}(0)$ *as* $n \to \infty$, *then*

$$\bar{Q}_n \Rightarrow \bar{Q} \quad \textit{in} \quad (D, J_1) \quad \textit{as} \quad n \to \infty,$$

where $\bar{Q}_n$ *is defined in* (2.5) *and* $\bar{Q} \equiv \{\bar{Q}(t) : t \geq 0\}$ *is defined by the integral equation*

$$\begin{aligned}\bar{Q}(t) &= \bar{Q}(0) + \lambda t - \int_0^t \big[\mu_1(\bar{Q}(s) \wedge 1)U(s) + \mu_2(\bar{Q}(s) \wedge \eta(s))(1 - U(s)) \\ &\quad + \theta_1(\bar{Q}(s) - 1)^+ U(s) + \theta_2(\bar{Q}(s) - \eta(s))^+(1 - U(s))\big]ds, \quad t \geq 0. \end{aligned} \tag{3.1}$$

Conditional on the availability process U and the functioning-server process η, the process $\bar{Q}$ evolves deterministically with two different dynamics on the two alternating states of the servers. For $k \geq 0$, on the interval $[T_k + u_{k+1}, T_{k+1})$, when the service interruption happens, $\bar{Q}(t)$ evolves according to the nonlinear ordinary differential equation (ODE),

$$\frac{d}{dt}\bar{Q}(t) = \lambda - \mu_2(\bar{Q}(t) \wedge \eta(t)) - \theta_2(\bar{Q}(t) - \eta(t))^+,$$

starting from the point $\bar{Q}(T_k + u_{k+1}) = \bar{Q}((T_k + u_{k+1})-)$ and on the interval $[T_k, T_k + u_{k+1})$, when the servers are functioning, $\bar{Q}(t)$ evolves according to the nonlinear ODE,

$$\frac{d}{dt}\bar{Q}(t) = \lambda - \mu_1(\bar{Q}(t) \wedge 1) - \theta_1(\bar{Q}(t) - 1)^+,$$

starting from the point $\bar{Q}(T_k) = \bar{Q}(T_k-)$.

As a consequence, the limit process is nondecreasing w.p.1 for the $G/M/n$ total-failure model, without customer abandonment, in both the QED and ED regimes. Evidently, the queue length in the original $G/M/n$ system diverges to infinity as $t \to \infty$ if n is large enough with QED scaling. We verified that for the $M/M/n$ model in Theorem 3 of Pang and Whitt (2008a).

3.2. Asymptotically Negligible Service Interruptions

In this section, we state the FWLLN for the fluid-scaled queue-length processes in all three regimes and FCLT's for the diffusion-scaled queue-length processes in the QED and ED regimes with asymptotically negligible service interruptions. With asymptotically negligible service interruptions, the fluid limits in all three regimes are the same as if there were no interruptions at all. However, the proof requires some extra work since we cannot directly apply the continuous mapping theorem to the integral representation of the queue-length processes with the process U_n in it.

Theorem 3.2. (FWLLN with asymptotically negligible service interruptions) *Consider the* $G/M/n + M$ *model in the QD, QED or ED regime and assume asymptotically negligible service interruptions. If there exists some constant* $\bar{Q}(0) < \infty$ *such that* $\bar{Q}_n(0) \Rightarrow \bar{Q}(0)$ *as* $n \to \infty$*, then*

$$\bar{Q}_n \Rightarrow \bar{Q} \quad \textit{in} \quad (D, J_1) \quad \textit{as} \quad n \to \infty,$$

where $\bar{Q}(t)$ is deterministic and differentiable, and satisfies the integral equation

$$\bar{Q}(t) = \bar{Q}(0) + \lambda t - \int_0^t \big(\mu_1(\bar{Q}(s) \wedge 1) + \theta_1(\bar{Q}(s) - 1)^+\big)ds, \quad t \geq 0. \tag{3.2}$$

Moreover, as $t \to \infty$, $\bar{Q}(t) \to q$, where $q = \lambda/\mu_1 < 1$ in the QD regime, $q = 1$ in the QED regime and $q = 1 + (\lambda - \mu_1)/\theta_1 > 1$ in the ED regime. If $\bar{Q}(0) = q$, then $\bar{Q}(t) = q$ for all $t \geq 0$.

Since the steady-state limits as $t \to \infty$ in the fluid limits above (the q in Theorem 3.2) are the same as without service interruptions, the centering for the FCLTs will remain the same as without service interruptions. We remark that the fluid dynamics in Theorem 3.2 depends on the assumption on the initial conditions: $\bar{Q}_n(0) \Rightarrow \bar{Q}(0)$. If the assumption is changed to $\hat{Q}_n(0) \Rightarrow \hat{Q}(0)$ as $n \to \infty$, as in the following FCLT in the QED regime, the fluid dynamics will become $\bar{Q}_n \Rightarrow \omega$ as $n \to \infty$ where $\omega(t) = 1$ for all $t \geq 0$.

Theorem 3.3. (FCLT in the QED regime with asymptotically negligible service interruptions) *Consider the $G/M/n + M$ model in the QED regime, where $\lambda = \mu_1$, and assume asymptotically negligible service interruptions. If there is a random variable $\hat{Q}(0)$ such that $\hat{Q}_n(0) \Rightarrow \hat{Q}(0)$ as $n \to \infty$, the processes $\hat{A}$, V and R are independent of $\hat{Q}(0)$, and $\hat{A}$ and V have no simultaneous jumps w.p.1, then*

$$\hat{Q}_n \Rightarrow \hat{Q} \quad \text{in} \quad (D, M_1) \quad \text{as} \quad n \to \infty,$$

where $\hat{Q} \equiv \{\hat{Q}(t) : t \geq 0\}$ is defined as the pathwise unique solution to the following stochastic integral equation with jumps,

$$\hat{Q}(t) \;=\; \hat{Q}(0) - \mu_1\beta t + \hat{A}(t) - \sqrt{\mu_1}B(t) - \int_0^t \big(\mu_1(\hat{Q}(s) \wedge 0) + \theta_1(\hat{Q}(s) \vee 0)\big)ds + J(t), \tag{3.3}$$

for each $t \geq 0$, where

$$J(t) \equiv \sum_{k=1}^{N(t)} \big[\big((\mu_1 - \mu_2) + (\mu_2 - \theta_2)(1 - \eta_k)\big)v_k\big], \quad t \geq 0, \tag{3.4}$$

B is a standard Brownian motion, independent of $\hat{Q}(0)$, $\hat{A}$ and J.

Moreover, if $\hat{A} \stackrel{\mathrm{d}}{=} \sqrt{\mu_1 c_a^2}B$, as occurs if the arrival processes are time-scaled versions of a common renewal process with interarrival times having a finite second moment, and the process N is Poisson, then the limiting process $\hat{Q}$ is a jump-diffusion (Markov) process, given by the pathwise unique solution to the following stochastic differential equation (SDE) with jumps,

$$\hat{Q}(t) \;=\; \hat{Q}(0) - \mu_1\beta t + \sqrt{\mu_1(1 + c_a^2)}B(t) - \int_0^t \big(\mu_1(\hat{Q}(s) \wedge 0) + \theta_1(\hat{Q}(s) \vee 0)\big)ds + J(t), \tag{3.5}$$

for each $t \geq 0$, where $J(t)$ is defined in (3.4) and B is a standard Brownian motion, independent of $\hat{Q}(0)$ and J.

Define the scaled queueing length process in the ED regime, $\hat{Q}_n^{ED} \equiv \{\hat{Q}_n^{ED}(t) : t \geq 0\}$, by

$$\hat{Q}_n^{ED}(t) \equiv \sqrt{n}\Big(\bar{Q}_n(t) - \Big(1 + \frac{\lambda - \mu_1}{\theta_1}\Big)\Big), \quad t \geq 0.$$

In the ED regime, servers are always busy asymptotically. Service interruptions will not change that fact but will add jumps in the limit.

Theorem 3.4. (FCLT in the ED regime with asymptotically negligible service interruptions) *Consider the $G/M/n + M$ model in ED regime and assume asymptotically negligible service interruptions. If there is a random variable $\hat{Q}^{ED}(0)$ such that $\hat{Q}_n^{ED}(0) \Rightarrow \hat{Q}^{ED}(0)$ as $n \to \infty$, the processes $\hat{A}$, V and R are independent of $\hat{Q}^{ED}(0)$, and $\hat{A}$ and V have no simultaneous jumps w.p.1, then*

$$\hat{Q}_n^{ED} \Rightarrow \hat{Q}^{ED} \quad in \quad (D, M_1) \quad as \quad n \to \infty,$$

where $\hat{Q}^{ED} \equiv \{\hat{Q}^{ED}(t) : t \geq 0\}$ is defined as the pathwise unique solution to the following stochastic integral equation with jumps,

$$\hat{Q}^{ED}(t) \;=\; \hat{Q}^{ED}(0) + \hat{A}(t) - \sqrt{\lambda}B(t) - \theta_1 \int_0^t \hat{Q}^{ED}(s)ds + J^{ED}(t), \quad t \geq 0, \tag{3.6}$$

where

$$J^{ED}(t) \;\equiv\; \sum_{k=1}^{N(t)} \Big[\lambda\Big(1 - \frac{\theta_2}{\theta_1}\Big) + \Big(\frac{\theta_2}{\theta_1}\mu_1 - \mu_2\Big) + (\mu_2 - \theta_2)(1 - \eta_k)\Big]v_k, \quad t \geq 0, \tag{3.7}$$

B is a standard Brownian motion, independent of $\hat{Q}^{ED}(0)$, $\hat{A}$ and J^{ED}.

Moreover, if $\hat{A} \stackrel{\mathrm{d}}{=} \sqrt{\lambda c_a^2}B$, as occurs if the arrival processes are time-scaled versions of a common renewal process with interarrival times having a finite second moment, and the process N is Poisson, then the limiting process $\hat{Q}^{ED}$ is a non-Gaussian O-U process driven a Levy process (and thus a Markov process), given by the pathwise unique solution to the following SDE with jumps,

$$\hat{Q}^{ED}(t) \;=\; \hat{Q}^{ED}(0) + \sqrt{\lambda(1 + c_a^2)}B(t) - \theta_1 \int_0^t \hat{Q}^{ED}(s)ds + J^{ED}(t), \quad t \geq 0, \tag{3.8}$$

where $J^{ED}(t)$ is defined in (3.7), and B is a standard Brownian motion, independent of $\hat{Q}^{ED}(0)$ and J^{ED}.

We refer to Protter (2003), Cont and Tankov (2004) and Situ (2005) for jump-diffusion processes and O-U processes driven by a Levy process.

We remark that if the abandonment rates $\theta_1 = \theta_2$, $J(t)$ in (3.4) and $J^{ED}(t)$ in (3.7) are the same. In other words, the limit process $\hat{Q}$ in the QED regime is insensitive to the abandonment rate change

when a service interruption occurs, while the limit process $\hat{Q}^{ED}$ in the ED regime is sensitive to that. If $\mu_1 = \mu_2$ and $\mu_2 = \theta_2$, then $J(t) = 0$ for all $t \geq 0$ and the process $\hat{Q}$ in (3.5) becomes a diffusion process, just as if there were no interruptions. Similarly for the limit process in the ED regime if $\mu_1 = \mu_2$, $\theta_1 = \theta_2$ and $\mu_2 = \theta_2$. If $\eta_k = 1$ for all $k \geq 1$ w.p.1, i.e., all servers remain functioning when a service interruption occurs, then there is still a jump term, $J(t) = (\mu_1 - \mu_2)V(t)$, in the limit process $\hat{Q}$ provided $\mu_1 \neq \mu_2$ and a jump term, $J^{ED}(t) = \sum_{k=1}^{N(t)} \left[\lambda\big(1 - \frac{\theta_2}{\theta_1}\big) + \big(\frac{\theta_2}{\theta_1}\mu_1 - \mu_2\big)\right]v_k$, in the limit process $\hat{Q}^{ED}$ provided that $\theta_1 \neq \theta_2$ or $\mu_1 \neq \mu_2$. The jump size in the jump process $\{J(t) : t \geq 0\}$ in (3.4) depends on the service rate change, the proportion of non-functioning servers when an interruption occurs and the duration of the limit of the scaled down times. The jump size in the jump process $\{J^{ED}(t) : t \geq 0\}$ in (3.7) also depends on customer abandonment rate change.

3.3. Infinite-Server Queues

If we assume that $\theta_1 = \mu_1$ and $\theta_2 = \mu_2$ in the many-server-queue setting, which is often reasonable in applications, then the many-server queue with interruptions behaves as an infinite-server queue in a random environment. As special cases of the results above, we obtain the following limits for infinite-server queues. The regimes QD, QED and ED are no longer relevant.

We consider a sequence of infinite-server queueing models indexed by n and let $n \to \infty$. Let the arrival processes, service times and service interruptions be the same for many server queues, but now there is no customer waiting or abandonment.

Corollary 3.1. (infinite-server queues) *Consider a sequence of $G/M/\infty$ models. Theorems* 3.1-3.4 *hold with the limits taking a simple form:*

(i) For unscaled service interruptions, the deterministic fluid limit in a random environment $\bar{Q}$ in (3.1) *becomes*

$$\bar{Q}(t) = \bar{Q}(0) + \lambda t - \int_0^t \big(\mu_1 \bar{Q}(s)U(s) + \mu_2 \bar{Q}(s)(1 - U(s))\big)ds, \quad t \geq 0.$$

(ii) For asymptotically negligible service interruptions, the deterministic fluid limit $\bar{Q}$ in (3.2) *becomes*

$$\bar{Q}(t) = \bar{Q}(0) + \mu_1 t - \mu_1 \int_0^t \bar{Q}(s)ds, \quad t \geq 0.$$

(iii) For asymptotically negligible service interruptions, the limit process $\hat{Q}$ in (3.3) *and* (3.6) *of the scaled queue-length processes $\hat{Q}_n$ becomes*

$$\hat{Q}(t) = \hat{Q}(0) + \hat{A}(t) - \sqrt{\mu_1}B(t) - \mu_1 \int_0^t \hat{Q}(s)ds + (\mu_1 - \mu_2)V(t), \quad t \geq 0. \tag{3.9}$$

If $\hat{A} \stackrel{\mathrm{d}}{=} \sqrt{\lambda c_a^2} B$ *and the process* N *is Poisson, the limit process* $\hat{Q}$ *is a non-Gaussian O-U process driven by a Levy process.*

3.4. Stochastic-Decomposition Property of Steady-State Distributions

The steady-state distribution of O-U processes driven by a Levy process is well studied, for which we refer to Theorem 2 of Wolfe (1982) and Proposition 15.4 of Cont and Tankov (2004). For the general theory of jump-diffusion processes and O-U processes driven by a Levy process, we refer to Sato (1999), Protter (2003) and Situ (2005). By a direct application, we are able to establish the limiting steady-state distributions of the limit queue-length processes in the ED regime in Theorem 3.4 and for infinite server queues in Corollary 3.1. (The proof is omitted.) As will be seen, the limiting steady-state distribution can be decomposed into two independent random variables; one is the same as without any interruptions and the other represents the effect of interruptions. The form of such a decomposition is consistent with the stochastic-decomposition properties of the number of customers in $M/M/\infty$ queues in *light traffic* established in Jayawardene and Kella (1996), Baykal-Gursoy and Xiao (2004), D'Auria (2007) and Falin (2008). However, the variable representing the effect of interruptions in Falin (2008) is different from the others in light traffic. In heavy traffic, we obtain another different variable representing the effect of interruptions.

Theorem 3.5. (stochastic-decomposition property of the steady-state distributions in the ED regime and for infinite-server queues) *Assume that the process* N *is Poisson with rate* δ *and the service-interruption down-time distribution satisfies* $\int_{|x|\geq 1} G(dx) < \infty$, $\int_{|x|\leq 1} |x|^2 G(dx) < \infty$ *and* $\int_0^\infty \log(|x|+1) G(dx) < \infty$. *When the arrival processes are renewal, the process* $\hat{Q}^{ED}(t)$ *in* (3.8) *converges to* $\hat{Q}^{ED}(\infty) \stackrel{\mathrm{d}}{=} Z_1 + Z_2$ *in distribution as* $t \to \infty$, *and* $\hat{Q}^{ED}(\infty)$ *is self-decomposable, where* Z_1 *is independent of* Z_2, $Z_1 \stackrel{\mathrm{d}}{=} Normal\big(0, \frac{\lambda(1+c_a^2)}{2\theta_1}\big)$, *the steady-state distribution without any interruptions, and the distribution of* Z_2 *is given by its characteristic function*

$$\psi_{Z_2}(s) \equiv E[\exp(isZ_2)] = \exp\Big(\int_{\mathbb{R}} (e^{isy} - 1 - isy\mathbf{1}_{|y|<1}) \int_1^\infty \delta H(udy) \frac{1}{\theta_1 u} du \Big),$$

where $H(\cdot)$ *is the distribution function of the random variable* $\Big[(\mu_2 - \theta_2)(1-\eta_1) + \lambda\Big(1 - \frac{\theta_2}{\theta_1}\Big) + \Big(\frac{\theta_2}{\theta_1}\mu_1 - \mu_2\Big)\Big] v_1$. *Moreover, if the service interruption time* v_1 *is exponentially distributed with rate* m_v^{-1}, *and* $\theta_1 > \delta e^{-\frac{2}{\mu_1 m_v}}$, *then the mean and variance of* Z_2 *are given by*

$$E[Z_2] = \frac{\delta \mu_1 m_v}{\theta_1} e^{-\frac{1}{\mu_1 m_v}}, \quad Var[Z_2] = \frac{\delta \mu_1^2 m_v^2}{\theta_1} - \Big(\frac{\delta \mu_1 m_v}{\theta_1}\Big)^2 e^{-\frac{2}{\mu_1 m_v}}.$$

When the arrival processes are renewal, the limit queue-length process $\hat{Q}(t)$ *for infinite-server*

queues in (3.9) *converges to* $\hat{Q}(\infty) \stackrel{\mathrm{d}}{=} \hat{Q}^{ED}(\infty)$ *in distribution as* $t \to \infty$ *with* $\lambda = \mu_1 = \theta_1 = \theta_2$ *and* $\eta_1 = 1$ *w.p.*1.

The steady-state distribution for jump diffusion process $\hat{Q}$ in (3.5) can be characterized by its generator since it is a special Markov process, but here we conjecture that it also has a stochastic-decomposition property, which is left for future work. Moreover, the steady-state distribution of the number of customers in many-server queues with vacations in light traffic has a *conditional* stochastic-decomposition property as in Tian and Zhang (2008a,b,c) and references therein. We conjecture such a decomposition property also holds in heavy traffic.

4. Other Scalings

4.1. Low-Impact Interruptions

The interruptions will have less impact when relatively few servers are affected. In this section we establish a limit for the case in which the number of non-functioning servers during an interruption is of order $O(\sqrt{n})$ with n servers, and so constitutes only an asymptotically negligible proportion. We do that by considering another scaling of the random variables $\{\eta_{n,k} : k \geq 1\}$. In particular, assume that

$$\frac{\eta_{n,k} - n}{\sqrt{n}} \Rightarrow \hat{\eta}_k \quad \text{for all} \quad k \quad \text{as} \quad n \to \infty, \tag{4.1}$$

where $\{\hat{\eta}_k : k \geq 1\}$ is a sequence of i.i.d. random variables. As a consequence, for the continuous-time process in (2.7), we have

$$\frac{\eta_n(t) - n}{\sqrt{n}} \Rightarrow 0 \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty,$$

for each $t \geq 0$, but the process $(\eta_n - n)/\sqrt{n}$ will not converge in D, just as with $(U_n, n^{-1}\eta_n)$ in the setting of (2.12).

Under the assumptions in (4.1) and Theorem 3.3, the limit process $\hat{Q}$ in (3.3) becomes

$$\begin{aligned} \hat{Q}(t) &= \hat{Q}(0) - \mu_1\beta t + \hat{A}(t) - \sqrt{\mu_1}B(t) - \mu_1 \int_0^t \big(\hat{Q}(s) \wedge 0\big)ds - \theta_1 \int_0^t (\hat{Q}(s) \vee 0)ds \\ &\quad + (\mu_1 - \mu_2)V(t), \quad t \geq 0. \end{aligned} \tag{4.2}$$

This limit process is actually the same as $\hat{Q}$ in (3.3) when $\eta_k = 1$ for all k w.p.1. This is reasonable since under the assumption in (4.1) almost all servers will remain functioning when a service interruption occurs. The jump size in the limit process depends only on the service rate change and the duration of the limit of the scaled down times. Therefore, if $\mu_1 = \mu_2$, there is no surprising that the limit process $\hat{Q}$ will have no jumps caused by service interruptions, as revealed in (4.2).

4.2. Bursty Arrival Processes

We now establish a heavy-traffic limit with stronger scaling to cover bursty arrival processes, paralleling the result without service interruptions, Theorem 2.1 in Pang and Whitt (2008b). In particular, we establish the heavy traffic limits for the queue-length process with a nonstandard scaling of the space in the FCLT. Let $\{c_n : n \geq 1\}$ be a sequence of positive numbers such that $c_n \to \infty$, $n/c_n \to \infty$ and $\sqrt{n}/c_n \to 0$ as $n \to \infty$. For example, one can choose $c_n = n^{1/\alpha}$ for $1/2 < \alpha < 1$.

Define the scaled arrival processes $\hat{A}_n \equiv \{\hat{A}_n(t) : t \geq 0\}$ by

$$\hat{A}_n(t) \equiv c_n^{-1}(A_n(t) - \lambda_n t), \quad t \geq 0,$$

and assume that $\hat{A}_n$ satisfy the FCLT:

$$\hat{A}_n \Rightarrow \hat{A} \quad \text{in} \quad (D, M_1) \quad \text{as} \quad n \to \infty.$$

When A_n is a renewal process for each n, the limit process $\hat{A}$ will be a Levy process. Indeed, if A_n is a time-scaled version of a single renewal process, then $\hat{A}$ must be a stable (α-stable) process; i.e., the increments have stable laws

$$\hat{A}(t+s) - \hat{A}(s) \stackrel{\mathrm{d}}{=} S_\alpha(t^{1/\alpha}, \beta, 0) \stackrel{\mathrm{d}}{=} t^{1/\alpha} S_\alpha(1, \beta, 0)$$

for any $s, t \geq 0$ and for some α and β with $0 < \alpha \leq 2$ and $-1 \leq \beta \leq 1$; see §4.5.3 in Whitt (2002).

The usual definition of the QED regime needs to be modified. Now we assume that

$$c_n^{-1} n(1 - \rho_n) \to \beta, \quad -\infty < \beta < \infty.$$

The asymptotically negligible service interruptions are assumed to satisfy

$$\{(u_{n,k}, c_n^{-1} n v_{n,k}) : k \geq 1\} \Rightarrow \{(u_k, v_k) : k \geq 1\} \quad \text{in} \quad (\mathbb{R}^2)^\infty \quad \text{as} \quad n \to \infty,$$

where $u_k, v_k > 0$ for all $k \geq 1$ w.p.1.

Define the scaled cumulative down time process of services V_n and the associated "lost service" process R_n by

$$V_n(t) \equiv c_n^{-1} n C_{D,n} = c_n^{-1} n \int_0^t (1 - U_n(s)) ds, \quad t \geq 0,$$

and

$$R_n(t) \equiv c_n^{-1} n \int_0^t \Big(1 - \frac{\eta_n(s)}{n}\Big)(1 - U_n(s)) ds, \quad t \geq 0.$$

As in Lemma 2.2, we can show that

$$\big(N_n, V_n, R_n, C_{U,n}\big) \Rightarrow (N, V, R, e) \quad \text{in} \quad (D, J_1) \times (D_3, M_1) \quad \text{as} \quad n \to \infty,$$

where N, V, R, e are as defined in Lemma 2.2.

Now define the scaled queue-length processes $\bar{Q}_n$ and $\hat{Q}_n$ by

$$\bar{Q}_n(t) \equiv n^{-1} Q_n(t), \quad \hat{Q}_n(t) \equiv c_n^{-1}(Q_n(t) - n), \quad t \geq 0.$$

The following theorem can be proved by similar arguments in the proof of Theorem 3.3 and hence its proof is omitted. Given the independence of A and V, the simultaneous jumps will be avoided w.p.1 if the processes A and V have no common fixed discontinuity points, e.g., if one of them is continuous in probability.

Theorem 4.1. (FCLT in the modified QED regime with asymptotically negligible service interruptions) *Consider the $G/M/n + M$ model in the modified QED regime and assume asymptotically negligible service interruptions with the nonstandard scaling above. If there is a random variable $\hat{Q}(0)$ such that $\hat{Q}_n(0) \Rightarrow \hat{Q}(0)$ as $n \to \infty$, the processes $\hat{A}$, V and R are independent of $\hat{Q}(0)$, and $\hat{A}$ and V have no simultaneous jumps w.p.1, then*

$$\hat{Q}_n \Rightarrow \hat{Q} \quad \textit{in} \quad (D, M_1) \quad \textit{as} \quad n \to \infty,$$

where $\hat{Q} \equiv \{\hat{Q}(t) : t \geq 0\}$ is defined by the following stochastic integral equation with jumps,

$$\hat{Q}(t) \;=\; \hat{Q}(0) - \mu_1 \beta t + \hat{A}(t) - \int_0^t \big(\mu_1(\hat{Q}(s) \wedge 0) + \theta_1(\hat{Q}(s) \vee 0)\big) ds + J(t), \quad t \geq 0,$$

where $J(t)$ is defined in (3.4).

5. Proofs

5.1. Martingale Representation of the Queue-Length Processes

The proofs follow the martingale argument reviewed in Pang et al. (2007). By a simple conservation of flow, the queue-length at any time t equals the initial content plus flow in minus flow out; i.e.,

$$\begin{aligned} Q_n(t) \;=\; & Q_n(0) + A_n(t) - S_1\Big(\mu_1 \int_0^t (Q_n(s) \wedge n) U_n(s) ds\Big) \\ & - S_2\Big(\mu_2 \int_0^t (Q_n(s) \wedge \eta_n(s))(1 - U_n(s)) ds\Big) - L_1\Big(\theta_1 \int_0^t (Q_n(s) - n)^+ U_n(s) ds\Big) \\ & - L_2\Big(\theta_2 \int_0^t (Q_n(s) - \eta_n(s))^+ (1 - U_n(s)) ds\Big), \quad t \geq 0, \end{aligned} \tag{5.1}$$

where the processes $S_i \equiv \{S_i(t) : t \geq 0\}$ and $L_i \equiv \{L_i(t) : t \geq 0\}$, $i = 1, 2$, are independent Poisson processes with unit rate.

Applying the argument in Lemma 2.1 in Pang et al. (2007) and the scaling in (2.5), we obtain the following representations.

Theorem 5.1. *The queueing process Q_n in* (5.1) *is well defined as a random element of the space D. The scaled queueing processes $\bar{Q}_n$ and $\hat{Q}_n$ can be represented as*

$$\begin{aligned}\bar{Q}_n(t) &= \bar{Q}_n(0) + \bar{A}_n(t) - \bar{S}_{n,1}(t) - \bar{S}_{n,2}(t) - \bar{L}_{n,1}(t) - \bar{L}_{n,2}(t) \\ &\quad - \mu_1 \int_0^t \big(\bar{Q}_n(s) \wedge 1\big) U_n(s) ds - \mu_2 \int_0^t \Big(\bar{Q}_n(s) \wedge \frac{\eta_n(s)}{n}\Big)(1 - U_n(s)) ds \\ &\quad - \theta_1 \int_0^t (\bar{Q}_n(s) - 1)^+ U_n(s) ds - \theta_2 \int_0^t \Big(\bar{Q}_n(s) - \frac{\eta_n(s)}{n}\Big)^+ (1 - U_n(s)) ds, \qquad (5.2)\end{aligned}$$

and

$$\begin{aligned}\hat{Q}_n(t) &= \hat{Q}_n(0) + \hat{A}_n(t) - \hat{S}_{n,1}(t) - \hat{S}_{n,2}(t) - \hat{L}_{n,1}(t) - \hat{L}_{n,2}(t) + \frac{\lambda_n - n\mu_1}{\sqrt{n}} t \\ &\quad - \mu_1 \int_0^t \big(\hat{Q}_n(s) \wedge 0\big) U_n(s) ds - \theta_1 \int_0^t \hat{Q}_n(s)^+ U_n(s) ds - \mu_2 R_{n,1}(t) - \theta_2 R_{n,2}(t) \\ &\quad + (\mu_1 - \mu_2) V_n(t), \qquad (5.3)\end{aligned}$$

where

$$\begin{aligned}\hat{S}_{n,1}(t) &= \frac{1}{\sqrt{n}}\Big(S_1\Big(\mu_1 \int_0^t \big(Q_n(s) \wedge n\big) U_n(s) ds\Big) - \mu_1 \int_0^t \big(Q_n(s) \wedge n\big) U_n(s) ds\Big), \\ \hat{S}_{n,2}(t) &= \frac{1}{\sqrt{n}}\Big(S_2\Big(\mu_2 \int_0^t \big(Q_n(s) \wedge \eta_n(s)\big)(1 - U_n(s)) ds\Big) - \mu_2 \int_0^t \big(Q_n(s) \wedge \eta_n(s)\big)(1 - U_n(s)) ds\Big), \\ \hat{L}_{n,1}(t) &= \frac{1}{\sqrt{n}}\Big(L_1\Big(\theta_1 \int_0^t (Q_n(s) - n)^+ U_n(s) ds\Big) - \theta_1 \int_0^t (Q_n(s) - n)^+ U_n(s) ds\Big), \\ \hat{L}_{n,2}(t) &= \frac{1}{\sqrt{n}}\Big(L_2\Big(\theta_2 \int_0^t (Q_n(s) - \eta_n(s))^+ (1 - U_n(s)) ds\Big) - \theta_2 \int_0^t (Q_n(s) - \eta_n(s))^+ (1 - U_n(s) ds\Big),\end{aligned}$$

$$\begin{aligned}R_{n,1}(t) &= \sqrt{n} \int_0^t \Big((\bar{Q}_n(s) - 1) \wedge \Big(\frac{\eta_n(s)}{n} - 1\Big)\Big)(1 - U_n(s)) ds, \\ R_{n,2}(t) &= \sqrt{n} \int_0^t \Big(\bar{Q}_n(s) - \frac{\eta_n(s)}{n}\Big)^+ (1 - U_n(s)) ds,\end{aligned}$$

and

$$\bar{S}_{n,i}(t) = \frac{1}{\sqrt{n}} \hat{S}_{n,i}(t), \quad \bar{L}_{n,i}(t) = \frac{1}{\sqrt{n}} \hat{L}_{n,i}(t), \quad i = 1, 2, \quad t \geq 0,$$

and V_n is defined in (2.14).

The following is established just as Theorem 7.2 in Pang et al. (2007).

Lemma 5.1. *The processes* $(\hat{S}_{n,1}, \hat{S}_{n,2}, \hat{L}_{n,1}, \hat{L}_{n,2})$ *defined in Theorem 5.1 are square integrable martingales with respect to the filtration* $\mathbf{F}_n \equiv \{\mathcal{F}_n(t) : t \geq 0\}$ *where*

$$\begin{aligned}\mathcal{F}_n(t) \equiv\ & \sigma\Big\{Q_n(0), S_1\Big(\mu_1 \int_0^s \big(Q_n(u) \wedge n\big)U_n(u)du\Big), S_2\Big(\mu_2 \int_0^t \big(Q_n(s) \wedge \eta_n(s)\big)(1 - U_n(s))ds\Big),\\ & L_1\Big(\theta_1 \int_0^t (Q_n(s) - n)^+ U_n(s)ds\Big), L_2\Big(\theta_2 \int_0^t (Q_n(s) - \eta_n(s))^+(1 - U_n(s))ds\Big) :\\ & 0 \leq s \leq t\Big\} \vee \sigma\big(A_n(s), U_n(s), \eta_n(s) : s \geq 0\big) \vee \mathcal{N},\end{aligned}$$

and $\mathcal{N}$ *is the collection of all null sets. The predictable quadratic variation processes* $\langle \hat{S}_{n,i} \rangle \equiv \{\langle \hat{S}_{n,i} \rangle(t) : t \geq 0\}$ *and* $\langle \hat{L}_{n,i} \rangle \equiv \{\langle \hat{L}_{n,i} \rangle(t) : t \geq 0\}$, $i = 1, 2$, *are*

$$\langle \hat{S}_{n,1} \rangle(t) = \frac{\mu_1}{n} \int_0^t \big(Q_n(s) \wedge n\big)U_n(s)ds, \quad \langle \hat{S}_{n,2} \rangle(t) = \frac{\mu_2}{n} \int_0^t \big(Q_n(s) \wedge \eta_n(s)\big)(1 - U_n(s))ds,$$
$$\langle \hat{L}_{n,1} \rangle(t) = \frac{\theta_1}{n} \int_0^t \big(Q_n(s) - n\big)^+ ds, \quad \langle \hat{L}_{n,2} \rangle(t) = \frac{\theta_2}{n} \int_0^t (Q_n(s) - \eta_n(s))^+(1 - U_n(s)ds.$$

Proof. It is known that for $i = 1, 2$, the processes $M_{S_i} \equiv \{M_{S_i}(t) : t \geq 0\}$ and $M_{L_i} \equiv \{M_{L_i}(t) : t \geq 0\}$ defined by $M_{S_i}(t) = S_i(t) - t$ and $M_{L_i}(t) = L_i(t) - t$ for $t \geq 0$ are square integrable martingales with respect to the filtration generated by the processes S_i and L_i, with predictable quadratic variation processes $\langle M_{S_i} \rangle \equiv \{\langle M_{S_i} \rangle(t) : t \geq 0\}$ and $\langle M_{L_i} \rangle \equiv \{\langle M_{L_i} \rangle(t) : t \geq 0\}$ defined by $\langle M_{S_i} \rangle(t) = t$ and $\langle M_{L_i} \rangle(t) = t$.

As for the case without service interruptions in §7.1 in Pang et al. (2007), we will need to apply the optional stopping theorem for multiparameter random time change (see §§2.8 and 6.2 of Either and Kurtz (1986)).

Define the processes $\tau_{n,i} \equiv \{\tau_{n,i}(t) : t \geq 0\}$ $(i = 1, 2, 3, 4)$ by

$$\tau_{n,1}(t) = \mu_1 \int_0^t (Q_n(s) \wedge n)U_n(s)ds, \quad \tau_{n,2}(t) = \mu_2 \int_0^t (Q_n(s) \wedge \eta_n(s))^+(1 - U_n(s))ds,$$
$$\tau_{n,3}(t) = \theta_1 \int_0^t (Q_n(s) - n)^+ U_n(s)ds, \quad \tau_{n,4}(t) = \theta_2 \int_0^t (Q_n(s) - \eta_n(s))^+(1 - U_n(s))ds.$$

All the $\tau_{n,i}$'s have continuous nondecreasing nonnegative sample paths and $\big(\tau_{n,1}(t), \tau_{n,2}(t), \tau_{n,3}(t), \tau_{n,4}(t)\big)$ are stopping times with respect to the filtration $\mathbf{H}_n \equiv \{\mathcal{H}_n(t_1, t_2, t_3, t_4) : t_i \geq 0, i = 1, 2, 3, 4\}$ for each $t \geq 0$; i.e., for all $u_i \geq 0$ $(i = 1, 2, 3, 4)$,

$$\{\tau_{n,1}(t) \leq u_1, \tau_{n,2}(t) \leq u_2, \tau_{n,3}(t) \leq u_3, \tau_{n,4}(t) \leq u_4\} \in \mathcal{H}_n(u_1, u_2, u_3, u_4),$$

where

$$\begin{aligned}\mathcal{H}_n(t_1, t_2, t_3, t_4) \equiv\ & \sigma\big\{Q_n(0), S_1(s_1), S_2(s_2), L_1(s_3), L_2(s_4) : 0 \leq s_i \leq t_i, i = 1, 2, 3, 4\big\}\\ & \vee \sigma\big\{A_n(s), U_n(s), \eta_n(s) : s \geq 0\big\} \vee \mathcal{N}.\end{aligned}$$

We need to check the moment conditions for $\tau_{n,i}$'s are satisfied; i.e., $E[\tau_{n,i}(t)] < \infty$ for all i and

$$E[S_i(\tau_{n,i}(t))] < \infty, \quad \text{for} \quad i = 1, 2, \quad E[L_1(\tau_{n,3}(t))] < \infty, \quad \text{and} \quad E[L_2(\tau_{n,4}(t))] < \infty, \quad t \geq 0.$$

The moment conditions for $\tau_{n,1}$ and $\tau_{n,2}$ are obviously satisfied and we apply the crude inequality $Q_n(t) \leq Q_n(0) + A_n(t)$ for each $t \geq 0$ to obtain the moment conditions for $\tau_{n,3}$ and $\tau_{n,4}$. For each $t \geq 0$,

$$E\Big[\theta_1 \int_0^t (Q_n(s) - n)^+ U_n(s) ds\Big] \leq \theta_1 t\big(E[Q_n(0)] + E[A_n(t)] + n\big) < \infty, \quad t \geq 0,$$

and

$$\begin{aligned} E\Big[L_1\Big(\theta_1 \int_0^t (Q_n(s) - n)^+ U_n(s) ds\Big)\Big] &\leq E\Big[L_1\Big(\theta_1 t\big(E[Q_n(0)] + E[A_n(t)] + n\big)\Big)\Big] \\ &= \theta_1 t\big(E[Q_n(0)] + E[A_n(t)] + n\big) < \infty. \end{aligned}$$

This implies the moment conditions for $\tau_{n,3}$ hold if $E[Q_n(0)] < \infty$. Similar arguments hold for $\tau_{n,4}$.

Recall that $E[Q_n(0)] < \infty$ is not assumed in the statement of Theorem 3.1. As elaborated upon in §6.3 of Pang et al. (2007), we first consider bounded initial conditions that converge to the same limit and establish the limit under the modified initial condition, which is asymptotically equivalent to the original limit in probability. So without loss of generality, we could have assumed that $E[Q_n(0)] < \infty$ at the very beginning.

Therefore, by applying a variation of Lemma 3.2 in Pang et al. (2007) for multiparameter martingales together with the optional stopping theorem for multiparameter random time change, we obtain the desired result. ∎

5.2. Continuity of an Integral Representation

The following integral representation plays a key role in the proof of the heavy-traffic limit theorems in this paper. To apply the continuous mapping theorem, we establish the continuity of the mapping, based on a new characterization of the M_1 convergence, Theorem 1.2 in Pang and Whitt (2008b).

Lemma 5.2. *Consider the integral equation*

$$y(t) = x(t) + \int_0^t h(y(s), z(s)) q(s) ds + \int_0^t g(y(s)) b(s) ds, \quad t \geq 0, \tag{5.4}$$

where the function $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is Lipstchitz continuous in each coordinate, the function $g : \mathbb{R} \to \mathbb{R}$ is Lipstchitz continuous and the functions $q, b \in D$. The integral equation in (5.4) *has a unique solution $y \in D$ so that it gives a function $\psi : D_4 \to D$ mapping (x, z, q, b) into $y \equiv \psi(x, z, q, b)$.*

Moreover, the function ψ is continuous if the spaces D_4 and D are both endowed with either the Skorohod J_1 or M_1 topology.

Proof. Let $\Upsilon_n \equiv (x_n, z_n, q_n, b_n)$ and $\Upsilon \equiv (x, z, q, b)$. Fix $T > 0$. This proof is a minor modification of Theorem 4.1 in Pang et al. (2007) and Theorem 1.1 in Pang and Whitt (2008b). Since q and b are elements of D, they are bounded on $[0, T]$ for each $T > 0$. Let K be such that $||q||_T \vee ||b||_T \leq K$, where for any function $f \in D$, $||f||_T \equiv \sup_{0\leq t\leq T} |f(t)|$ and for simplicity, we write $||f||$ if $T = 1$. Suppose that $|h(w_1, l_1) - h(w_2, l_2)| \leq c_1(|w_1 - w_2| + |l_1 - l_2|)$ and $|g(w_1) - g(w_2)| \leq c_2|w_1 - w_2|$ for all $w_i, l_i \in \mathbb{R}$ and some $c_1, c_2 \in (0, \infty)$. We will only prove continuity property in the M_1 topology since the proof in the J_1 topology is similar and easier. Given that $d_{M_1}(\Upsilon, \Upsilon) \to 0$, by Theorem 1.2 in Pang and Whitt (2008b), let (u_n, r_n) and (u, r) be parametric representations of Υ_n and Υ, constructed such that r and r_n are absolutely continuous with respect to Lebesgue measure on $[0, 1]$ with derivatives r' and r'_n for all n satisfying

$$||r'_n - r'||_{L_1} \equiv \int_0^1 |r'_n(s) - r'(s)|ds \to 0 \quad \text{as} \quad n \to \infty, \quad ||r'|| < \infty \quad \text{and} \quad \sup_{n\geq 1}\{||r'_n||\} < \infty. \tag{5.5}$$

We will construct the associated parametric representations (u_{y_n}, r_{y_n}) and (u_y, r_y) for y_n and y. Since the jumps of y_n necessarily coincide with the jumps of x_n for each n and so do the jumps of y and x, we can let $r_y = r$ and $r_{y_n} = r_n$ for all n. So the time components r_y and r_{y_n} satisfy (5.5) and $||r_{y_n} - r_y|| \to 0$ as $n \to \infty$. We define the spatial components $u_y(s) = y(r(s))$ for $r(s) \in Disc(y)^c$ and $u_{y_n}(s) = y_n(r_n(s)) \in Disc(y_n)^c$ for all $n \geq 0$, where $Disc(y)$ is the set of the discontinuity points and $Disc(y)^c$ is its complements and define the remaining values by linear interpolation. Now we can write

$$\begin{aligned} u_{y_n}(s) &= u_n^{(1)}(s) + \int_0^s h(u_{y_n}(w), u_n^{(2)}(w))u_n^{(3)}(w)r'_n(w)dw + \int_0^s g(u_{y_n}(w))u_n^{(4)}(w)r'_n(w)dw, \\ u_y(s) &= u^{(1)}(s) + \int_0^s h(u_y(w), u^{(2)}(w))u^{(3)}(w)r'(w)dw + \int_0^s g(u_y(w))u^{(4)}(w)r'(w)dw, \end{aligned}$$

for $s \in [0, 1]$, where $u^{(i)}$ is the i^{th} component of u, and similarly for $u_n^{(i)}$. Now we have

$$|u_{y_n}(s) - u_y(s)| \leq |u_n^{(1)}(s) - u^{(1)}(s)| + \Delta_{n,1} + \Delta_{n,2},$$

where

$$\Delta_{n,1} \equiv \Big| \int_0^s h(u_{y_n}(w), u_n^{(2)}(w))u_n^{(3)}(w)r'_n(w)dw - \int_0^s h(u_y(w), u^{(2)}(w))u^{(3)}(w)r'(w)dw \Big|,$$

and

$$\Delta_{n,2} \equiv \Big| \int_0^s g(u_{y_n}(w))u_n^{(4)}(w)r'_n(w)dw - \int_0^s g(u_y(w))u^{(4)}(w)r'(w)dw \Big|.$$

Then

$$\begin{aligned}
\Delta_{n,1} &\le \Big|\int_0^s h(u_{y_n}(w),u_n^{(2)}(w))u_n^{(3)}(w)r_n'(w)dw - \int_0^s h(u_y(w),u^{(2)}(w))u_n^{(3)}(w)r_n'(w)dw\Big| \\
&\quad + \Big|\int_0^s h(u_y(w),u^{(2)}(w))u_n^{(3)}(w)r_n'(w)dw - \int_0^s h(u_y(w),u^{(2)}(w))u_n^{(3)}(w)r'(w)dw\Big| \\
&\quad + \Big|\int_0^s h(u_y(w),u^{(2)}(w))u_n^{(3)}(w)r'(w)dw - \int_0^s h(u_y(w),u^{(2)}(w))u^{(3)}(w)r'(w)dw\Big| \\
&\le Kc_1||r_n'||\int_0^s (|u_{y_n}(w)-u_y(w)| + |u_n^{(2)}(w)-u^{(2)}(w)|)dw \\
&\quad + ||h(y,z)||_T\Big(K\int_0^s |r_n'(w)-r'(w)|dw + ||r'||\cdot||u_n^{(3)}-u^{(3)}||\Big),
\end{aligned}$$

and

$$\begin{aligned}
\Delta_{n,2} &\le \Big|\int_0^s g(u_{y_n}(w))u_n^{(4)}(w)r_n'(w)dw - \int_0^s g(u_y(w))u_n^{(4)}(w)r_n'(w)dw\Big| \\
&\quad + \Big|\int_0^s g(u_y(w))u_n^{(4)}(w)r_n'(w)dw - \int_0^s g(u_y(w))u_n^{(4)}(w)r'(w)dw\Big| \\
&\quad + \Big|\int_0^s g(u_y(w))u_n^{(4)}(w)r'(w)dw - \int_0^s g(u_y(w))u^{(4)}(w)r'(w)dw\Big| \\
&\le Kc_2||r_n'||\int_0^s |u_{y_n}(w)-u_y(w)|dw \\
&\quad + ||g(y)||_T\Big(K\int_0^s |r_n'(w)-r'(w)|dw + ||r'||\cdot||u_n^{(3)}-u^{(3)}||\Big).
\end{aligned}$$

Hence,

$$\begin{aligned}
|u_{y_n}(s)-u_y(s)| &\le ||u_n^{(1)}-u^{(1)}|| + (||h(y,z)||_T + ||g(y)||_T)\big(K||r_n'-r'||_{L_1} + ||r'||\cdot||u_n^{(3)}-u^{(3)}||\big) \\
&\quad + Kc_1||r_n'||\cdot||u_n^{(2)}-u^{(2)}|| + K(c_1+c_2)||r_n'||\int_0^s |u_{y_n}(w)-u_y(w)|dw.
\end{aligned}$$

Now by Gronwall's inequality and (5.5), we obtain

$$\begin{aligned}
||u_{y_n}-u_y|| &\le \big(||u_n^{(1)}-u^{(1)}|| + (||h(y,z)||_T + ||g(y)||_T)\big(K||r_n'-r'||_{L_1} + ||r'||\cdot||u_n^{(3)}-u^{(3)}||\big) \\
&\quad + Kc_1||r_n'||\cdot||u_n^{(2)}-u^{(2)}||\big)\cdot e^{||r_n'||K(c_1+c_2)} \to 0, \quad \text{as} \quad n\to\infty.
\end{aligned}$$

Therefore we have proved that $d_{M_1}(y_n,y)\to 0$ as $n\to\infty$. ∎

5.3. Proof of Theorem 3.1

There are two possible scenarios for the random environment. In the first scenario, the random environment U_n is common and equal to U for all n. In the second scenario, the random environments are different for the models, but $U_n \Rightarrow U$ in the Skorohod J_1 topology as $n\to\infty$. In the first scenario, the proof is done in Theorem 1 of Pang and Whitt (2008a) for the case of total-failure models and is basically the same as without service interruptions except that conditioning

on the process U, we need to apply the continuous mapping theorem to the mapping defined by the integral representation in Lemma 5.2. Here we only consider the second scenario.

The following lemma can be proved by applying Lemmas 5.5, 5.8, 5.9 of Pang et al. (2007), with minor modifications. We state the result without proof.

Lemma 5.3. *Under the assumptions of Theorem 3.1, the sequence of processes* $\{(\hat{S}_{n,1}, \hat{S}_{n,2}, \hat{L}_{n,1}, \hat{L}_{n,2}) : n \geq 1\}$ *defined in Theorem 5.1 is stochastically bounded in the space* D_4*, i.e., for all* $\epsilon > 0$ *and* $T > 0$*, there exists a positive real number* K *such that*

$$P(||(\hat{S}_{n,1}, \hat{S}_{n,2}, \hat{L}_{n,1}, \hat{L}_{n,2})||_T \leq K) > 1 - \epsilon, \quad \textit{for} \quad n \geq 1.$$

The sequence of processes $\{\bar{Q}_n : n \geq 1\}$ *in Theorem 5.1 is stochastically bounded in* D *and we have the joint convergence*

$$(\bar{Q}_n(0), \bar{A}_n, \bar{S}_{n,1}, \bar{S}_{n,2}, \bar{L}_{n,1}, \bar{L}_{n,2}) \Rightarrow (\bar{Q}(0), \lambda e, \zeta, \zeta, \zeta, \zeta) \quad \textit{in} \quad \mathbb{R} \times (D_5, J_1) \quad \textit{as} \quad n \to \infty,$$

where $e(t) = t$ *and* $\zeta(t) = 0$ *for all* $t \geq 0$*.*

Proof of Theorem 3.1 We observe that the continuous mapping theorem cannot be applied directly to the integral representation of $\bar{Q}_n$ in (5.2), due to the process U_n in the integral, so we will prove the weak convergence of the processes $\bar{Q}_n$ conditional on the processes U_n to the process $\bar{Q}$ conditional on the process U. We notice that the process $\bar{Q}$ conditional on the process U is differentiable.

As a first step, given $U_n \Rightarrow U$ in (D, J_1) as $n \to \infty$, we apply the Skorohod representation theorem to get versions of U_n converging to U w.p.1. We can then restrict our attention to these and focus on a single sample point ω for which $||U_n(\omega) - U(\omega)||_T \to 0$ as $n \to \infty$. Fix T large as a continuity point of U. Since the random environment is independent of the queueing system, by fixing such an ω, we will consider the stochastic queue-length processes and partition the time domain into the time intervals $\left[T_{n,k}, T_{n,k} + u_{n,k+1}\right)$ and $\left[T_{n,k} + u_{n,k+1}, T_{n,k+1}\right)$ for $k \geq 0$.

By the representation of $\bar{Q}_n$ in Theorem 5.1, conditional on the processes U_n, for all $k \geq 0$, on the time intervals $\left[T_{n,k}, T_{n,k} + u_{n,k+1}\right)$

$$\begin{aligned}\bar{Q}_n(t) &= \bar{Q}_n(T_{n,k}-) + \bar{A}_n(t) - \bar{A}_n(T_{n,k}-) - \left(\bar{S}_{n,1}(t) - \bar{S}_{n,1}(T_{n,k}-)\right) \\ &\quad - \left(\bar{L}_{n,1}(t) - \bar{L}_{n,1}(T_{n,k}-)\right) - \mu_1 \int_{T_{n,k}}^{t} (\bar{Q}_n(s) \wedge 1) ds - \theta_1 \int_{T_{n,k}}^{t} (\bar{Q}_n(s) - 1)^+ ds,\end{aligned}$$

and on the time intervals $\big[T_{n,k}+u_{n,k+1}, T_{n,k+1}\big)$,

$$\begin{aligned}\bar{Q}_n(t) &= \bar{Q}_n((T_{n,k}+u_{n,k+1})-) + \bar{A}_n(t) - \bar{A}_n((T_{n,k}+u_{n,k+1})-) \\ &\quad - \Big(\bar{S}_{n,2}(t) - \bar{S}_{n,2}((T_{n,k}+u_{n,k+1})-)\Big) - \Big(\bar{L}_{n,2}(t) - \bar{L}_{n,2}((T_{n,k}+u_{n,k+1})-)\Big) \\ &\quad - \mu_2 \int_{T_{n,k}+u_{n,k+1}}^{t} \Big(\bar{Q}_n(s) \wedge \frac{\eta_n(s)}{n}\Big) ds - \theta_2 \int_{T_{n,k}+u_{n,k+1}}^{t} \Big(\bar{Q}_n(s) - \frac{\eta_n(s)}{n}\Big)^+ ds.\end{aligned}$$

We proceed the proof by induction on k. For $k=0$, conditional on U_n, on the time interval $\big[0, u_{n,1}\big)$,

$$\bar{Q}_n(t) = \bar{Q}_n(0) + \bar{A}_n(t) - \bar{S}_{n,1}(t) - \bar{L}_{n,1}(t) - \int_0^t [\mu_1(\bar{Q}_n(s)\wedge 1) + \theta_1(\bar{Q}_n(s)-1)^+] ds$$

This representation of the processes $\bar{Q}_n$ corresponds to the mapping defined in Lemma 5.2 with the function g defined by $g(x) = -\mu_1(x\wedge 1) - \theta_1(x-1)^+$ for all $x\in\mathbb{R}$, $b=1$ and $h=q=0$. By the continuous mapping theorem together with Lemma 5.3 applying to the addition mapping and the mapping defined in Lemma 5.2 and the assumptions on the unscaled service interruptions, we obtain the weak convergence of the processes $\bar{Q}_n$ conditional on the processes U_n restricted to the time interval $\big[0, u_{n,1}\big)$ to the process $\bar{Q}$ conditional on the process U restricted to the time interval $[0, u_1)$ in D as $n\to\infty$, where

$$\bar{Q}(t) = \bar{Q}(0) + \lambda t - \int_0^t [\mu_1(\bar{Q}(s)\wedge 1) + \theta_1(\bar{Q}(s)-1)^+] ds, \quad t\in[0,u_1).$$

On the time interval $\big[u_{n,1}, T_{n,1}\big)$,

$$\begin{aligned}\bar{Q}_n(t) &= \bar{Q}_n(u_{n,1}-) + \bar{A}_n(t) - \bar{A}_n(u_{n,1}-) - \Big(\bar{S}_{n,2}(t) - \bar{S}_{n,2}(u_{n,1}-)\Big) \\ &\quad - \Big(\bar{L}_{n,2}(t) - \bar{L}_{n,2}(u_{n,1}-)\Big) - \mu_2 \int_{u_{n,1}}^{t} \Big(\bar{Q}_n(s) \wedge \frac{\eta_n(s)}{n}\Big) ds \\ &\quad - \theta_2 \int_{u_{n,k+1}}^{t} \Big(\bar{Q}_n(s) - \frac{\eta_n(s)}{n}\Big)^+ ds.\end{aligned}$$

By the continuous mapping theorem together with Lemma 5.3 applied to the addition mapping and the integral mapping defined in Lemma 5.2 with the function h defined by $h(x,z) = -\mu_2(x\wedge z) - \theta_2(x-z)^+$ for all $x,z\in\mathbb{R}$, $q=1$ and $g=b=0$, we obtain the weak convergence of the processes $\bar{Q}_n$ conditional on the processes U_n restricted to the time interval $\big[u_{n,1}, T_{n,1}\big)$ to the process $\bar{Q}$ conditional on the process U restricted to the time interval $[u_1, T_1)$ in D, where

$$\begin{aligned}\bar{Q}(t) &= \bar{Q}(u_1-) + \lambda t - \lambda u_1 - \int_{u_1}^{t} \big[\mu_2(\bar{Q}(s)\wedge\eta(s)) - \theta_2(\bar{Q}(s)-\eta(s))^+\big] ds \\ &= \bar{Q}(0) + \lambda t - \int_0^t [\mu_1(\bar{Q}(s)\wedge 1) + \theta_1(\bar{Q}(s)-1)^+] ds \\ &\quad - \int_{u_1}^{t} \big[\mu_2(\bar{Q}(s)\wedge\eta(s)) - \theta_2(\bar{Q}(s)-\eta(s))^+\big] ds.\end{aligned}$$

Now the weak convergence for $k = 0$ is obtained. Suppose we have obtained the weak convergence for some $k > 0$. We want to show the weak convergence for $k + 1$.

On the time interval $\big[T_{n,k+1}, T_{n,k+1} + u_{n,k+2}\big)$

$$\begin{aligned}\bar{Q}_n(t) &= \bar{Q}_n(T_{n,k+1}-) + \bar{A}_n(t) - \bar{A}_n(T_{n,k+1}-) - \Big(\bar{S}_{n,1}(t) - \bar{S}_{n,1}(T_{n,k+1}-)\Big) \\ &\quad - \Big(\bar{L}_{n,1}(t) - \bar{L}_{n,1}(T_{n,k+1}-)\Big) - \mu_1 \int_{T_{n,k+1}}^{t} (\bar{Q}_n(s) \wedge 1) ds - \theta_1 \int_{T_{n,k+1}}^{t} (\bar{Q}_n(s) - 1)^+ ds.\end{aligned}$$

This representation of the processes $\bar{Q}_n$ corresponds to the mapping defined in Lemma 5.2 with the function g defined by $g(x) = -\mu_1(x \wedge 1) - \theta_1(x-1)^+$ for all $x \in \mathbb{R}$, $b = 1$ and $h = q = 0$. By the continuous mapping theorem together with Lemma 5.3 applying to the addition mapping and the mapping defined in Lemma 5.2 and the assumptions on the unscaled service interruptions, we obtain the weak convergence of the processes $\bar{Q}_n$ conditional on the processes U_n restricted to the time interval $\big[T_{n,k+1}, T_{n,k+1} + u_{n,k+2}\big)$ to the process $\bar{Q}$ conditional on the process U restricted to the time interval $[T_{k+1}, T_{k+1} + u_{k+2})$ in D as $n \to \infty$, where

$$\begin{aligned}\bar{Q}(t) &= \bar{Q}(T_{k+1}-) + \lambda t - \lambda T_{k+1} - \int_{T_{k+1}}^{t} [\mu_1(\bar{Q}(s) \wedge 1) + \theta_1(\bar{Q}(s) - 1)^+] ds \\ &= \bar{Q}(0) + \lambda T_{k+1} - \sum_{j=0}^{k} \int_{T_j}^{T_j + u_{j+1}} \big[\mu_1(\bar{Q}(s) \wedge 1) + \theta_1(\bar{Q}(s) - 1)^+\big] ds \\ &\quad - \sum_{j=0}^{k} \int_{T_j + u_{j+1}}^{T_{j+1}} \big[\mu_2(\bar{Q}(s) \wedge \eta(s)) + \theta_2(\bar{Q}(s) - \eta(s))^+\big] ds \\ &\quad + \lambda t - \lambda T_{k+1} - \int_{T_{k+1}}^{t} [\mu_1(\bar{Q}(s) \wedge 1) + \theta_1(\bar{Q}(s) - 1)^+] ds \\ &= \bar{Q}(0) + \lambda t - \sum_{j=0}^{k} \int_{T_j}^{T_j + u_{j+1}} \big[\mu_1(\bar{Q}(s) \wedge 1) + \theta_1(\bar{Q}(s) - 1)^+\big] ds \\ &\quad - \int_{T_{k+1}}^{t} [\mu_1(\bar{Q}(s) \wedge 1) + \theta_1(\bar{Q}(s) - 1)^+] ds \\ &\quad - \sum_{j=0}^{k} \int_{T_j + u_{j+1}}^{T_{j+1}} \big[\mu_2(\bar{Q}(s) \wedge \eta(s)) + \theta_2(\bar{Q}(s) - \eta(s))^+\big] ds.\end{aligned}$$

On the time interval $\big[T_{n,k+1} + u_{n,k+2}, T_{n,k+2}\big)$,

$$\begin{aligned}\bar{Q}_n(t) &= \bar{Q}_n((T_{n,k+1} + u_{n,k+2})-) + \bar{A}_n(t) - \bar{A}_n((T_{n,k+1} + u_{n,k+2})-) \\ &\quad - \Big(\bar{S}_{n,2}(t) - \bar{S}_{n,2}((T_{n,k+1} + u_{n,k+2})-)\Big) - \Big(\bar{L}_{n,2}(t) - \bar{L}_{n,2}((T_{n,k+1} + u_{n,k+2})-)\Big) \\ &\quad - \mu_2 \int_{T_{n,k+1} + u_{n,k+2}}^{t} \Big(\bar{Q}_n(s) \wedge \frac{\eta_n(s)}{n}\Big) ds - \theta_2 \int_{T_{n,k+1} + u_{n,k+2}}^{t} \Big(\bar{Q}_n(s) - \frac{\eta_n(s)}{n}\Big)^+ ds.\end{aligned}$$

By the continuous mapping theorem together with Lemma 5.3 applied to the addition mapping and the integral mapping defined in Lemma 5.2 with the function h defined by $h(x, z) = -\mu_2(x \wedge$

$z)-\theta_2(x-z)^+$ for all $x, z \in \mathbb{R}$, $q = 1$ and $g = b = 0$, we obtain the weak convergence of the processes $\bar{Q}_n$ conditional on the processes U_n restricted to the time interval $\big[T_{n,k+1} + u_{n,k+2}, T_{n,k+2}\big)$ to the process $\bar{Q}$ conditional on the process U restricted to the time interval $[T_{k+1} + u_{k+2}, T_{k+2})$ in D as $n \to \infty$, where

$$\begin{aligned}
\bar{Q}(t) &= \bar{Q}((T_{k+1} + u_{k+2})-) + \lambda t - \lambda(T_{k+1} + u_{k+2}) \\
&\quad - \int_{T_{k+1}+u_{k+2}}^{t} \big[\mu_2(\bar{Q}(s) \wedge \eta(s)) + \theta_2(\bar{Q}(s) - \eta(s))^+\big] ds \\
&= \bar{Q}(0) + \lambda t - \sum_{j=0}^{k+1} \int_{T_j}^{T_j+u_{j+1}} \big[\mu_1(\bar{Q}(s) \wedge 1) + \theta_1(\bar{Q}(s) - 1)^+\big] ds \\
&\quad - \sum_{j=0}^{k} \int_{T_j+u_{j+1}}^{T_{j+1}} \big[\mu_2(\bar{Q}(s) \wedge \eta(s)) + \theta_2(\bar{Q}(s) - \eta(s))^+\big] ds \\
&\quad - \int_{T_{k+1}+u_{k+2}}^{t} \big[\mu_2(\bar{Q}(s) \wedge \eta(s)) + \theta_2(\bar{Q}(s) - \eta(s))^+\big] ds.
\end{aligned}$$

So the induction steps are valid and the weak convergence of the processes $\bar{Q}_n$ conditional on U_n to the process $\bar{Q}$ conditional on the process U holds for the intervals $[T_j, T_{j+1})$, $0 \le j \le k+1$ in D. Therefore, the weak convergence of the processes $\bar{Q}_n$ conditional on U_n to the process $\bar{Q}$ conditional on the process U holds in D and without conditioning the processes $\bar{Q}_n$ converge weakly to the process $\bar{Q}$ in D. ∎

5.4. Proof of Lemma 2.2

We apply Skorohod representation theorem to obtain random variables $\{(u_{n,k}, v_{n,k}) : k \ge 1\}$ such that (2.12) holds w.p.1. Fix $k \ge 1$ such that $T_k < T < T_{k+1}$ for some continuity point T of V. Consider the interval $[0, T]$.

Let Γ and Γ_n be the complete graphs of V and V_n defined by

$$\Gamma = \big\{(z, t) : z = \alpha V(t) + (1-\alpha)V(t-) \ge 0, t \in [0, T], \alpha \in [0, 1]\big\},$$

and

$$\Gamma_n = \big\{(z, t) : z = V_n(t) \ge 0, t \in [0, T]\big\}.$$

Let (a_i, b_i) be a pair of positive numbers such that $0 < a_i < b_i < a_{i+1} < b_{i+1} < 1$. Define the parametric representations of V and V_n by $(u, r) : [0, 1] \to \Gamma$ and $(u_n, r_n) : [0, 1] \to \Gamma_n$, respectively,

where

$$
\begin{aligned}
&r(0) = r_n(0) = 0, \quad r(1) = r_n(1) = T, \\
&r(a_i) = r(b_i) = T_i, \quad r_n(a_i) = T_{n,i-1} + u_{n,i}, \quad r_n(b_i) = T_{n,i}, \\
&u(a_i) = v_0 + ... + v_{i-1} \quad u(b_i) = v_1 + ... + v_i, \\
&u_n(a_i) = \sqrt{n}(v_{n,0} + ... + v_{n,i-1}), \quad u_n(b_i) = \sqrt{n}(v_{n,1} + ... + v_{n,i}),
\end{aligned}
$$

for each $i \geq 1$ with $v_0 = v_{n,0} = 0$, and the values of u, r, u_n, r_n at the remaining points are determined by linear interpolation.

Thus we have

$$
\begin{aligned}
||r_n - r||_T &= \max_{i \leq k} |T_{n,i-1} + u_{n,i} - T_i|, \\
||u_n - u||_T &= \max_{i \leq k} \big\{ |\sqrt{n}(v_{n,0} + ... + v_{n,i-1}) - (v_0 + ... + v_{i-1})| \\
&\qquad \vee |\sqrt{n}(v_{n,1} + ... + v_{n,i}) - (v_1 + ... + v_i)| \big\}.
\end{aligned}
$$

By the assumptions in (2.12) and $\lambda_n, \lambda < \infty$, we obtain

$$
||u_n - u||_T \vee ||r_n - r||_T \to 0, \quad \text{as} \quad n \to \infty.
$$

So by Theorem 12.5.1 (i) in Whitt (2002), the weak convergence of V_n to V in (D, M_1) is proved.

For the convergence of R_n to R, we can follow the same argument above by replacing $v_{n,i}$ by $(1 - \eta_{n,i}/n)v_{n,i}$ and v_i by $(1 - \eta_i)v_i$. Since $\eta_{n,i}$ and $v_{n,i}$ are independent for all i, and so are η_i and v_i, the convergence follows from the assumptions on $\eta_{n,i}$ in (2.1) and $v_{n,i}$ in (2.12). Since we can use the same time components in their parametric representations, we obtain the joint convergence $(V_n, R_n) \Rightarrow (V, R)$ in (D_2, M_1).

For the convergence of $C_{U,n}$ to e, it suffices to the prove the uniform convergence on compact time intervals. Since $V_n \Rightarrow V$ as $n \to \infty$, we have $V_n(T) \Rightarrow V(T)$ as $n \to \infty$ and

$$
\sup_{0 \leq t \leq T} \big| C_{U,n}(t) - t \big| = \sup_{0 \leq t \leq T} \int_0^t \mathbf{1}_{\{U_n(s)=0\}} ds = \frac{1}{\sqrt{n}} V_n(T) \Rightarrow 0 \quad \text{as} \quad n \to \infty.
$$

Hence, $C_{U,n} \Rightarrow e$ in D is proved and thus we have the joint convergence $(V_n, R_n, C_{U,n}) \Rightarrow (V, R, e)$ in (D_3, M_1).

The convergence $N_n \Rightarrow N$ in (D, J_1) and (D, M_1) is straightforward, but for M_1, the time component of the parametric representation of N_n must be different from that of V_n and R_n. Therefore, as noted in the remark after Lemma 2.2, we need to use the weaker M_1 product topology for the joint convergence of N_n and the others; i.e., $(N_n, V_n, R_n, C_{U,n}) \Rightarrow (N, V, R, e)$ in $(D, J_1) \times (D_3, M_1)$ as $n \to \infty$. That completes the proof of Lemma 2.2.

5.5. Proofs of Theorems 3.2 and 3.3

In this section, we will prove Theorems 3.2 and 3.3. Since their proofs are similar, we will prove Theorem 3.3 by proving Theorems 5.2 and 5.3 first and then sketch the proof of Theorem 3.2. We observe that the continuous mapping theorem cannot be applied directly to the integral representation of the fluid-scaled queue-length processes $\bar{Q}_n$ and the diffusion-scaled queue-length processes $\hat{Q}_n$ in Theorem 5.1, since the mapping defined by the integral to obtain the processes $\bar{Q}_n$ and $\hat{Q}_n$ is not continuous in the Skorohod topologies due to the process U_n in the integral. We instead consider the queue-length processes $Q_n^0 \equiv \{Q_n^0(t) : t \geq 0\}$ without interruptions, represented by

$$Q_n^0(t) \equiv Q_n(0) + A_n(t) - S_1\Big(\mu_1 \int_0^t (Q_n^0(s) \wedge n)ds\Big) - L_1\Big(\theta_1 \int_0^t (Q_n^0(s) - n)^+ ds\Big), \quad t \geq 0, \quad (5.6)$$

and the associated scaled processes $\bar{Q}_n^0 \equiv \{\bar{Q}_n^0(t) : t \geq 0\}$ and $\hat{Q}_n^0 \equiv \{\hat{Q}_n^0(t) : t \geq 0\}$, defined by $\bar{Q}_n^0 \equiv n^{-1}Q_n^0(t)$, and

$$\hat{Q}_n^0(t) \equiv \sqrt{n}(\bar{Q}_n^0(t) - 1) + (\mu_1 - \mu_2)V_n(t) - \mu_2 R_{n,1}^0(t) - \theta_2 R_{n,2}^0(t), \quad t \geq 0, \quad (5.7)$$

where

$$R_{n,1}^0(t) \equiv \sqrt{n} \int_0^t \Big((\bar{Q}_n^0(s) - 1) \wedge \Big(\frac{\eta_n(s)}{n} - 1\Big)\Big)(1 - U_n(s))ds, \quad t \geq 0,$$

and

$$R_{n,2}^0(t) \equiv \sqrt{n} \int_0^t \Big(\bar{Q}_n^0(s) - \frac{\eta_n(s)}{n}\Big)^+ (1 - U_n(s))ds, \quad t \geq 0.$$

Note that $Q_n^0(0) = Q_n(0)$, $\bar{Q}_n^0(0) = \bar{Q}_n(0)$ and $\hat{Q}_n^0(0) = \hat{Q}_n(0)$. Now the processes $\bar{Q}_n^0$ and $\hat{Q}_n^0$ can be represented as

$$\begin{aligned} \bar{Q}_n^0(t) &= \bar{Q}_n(0) + \bar{A}_n(t) - \bar{S}_{n,1}^0(t) - \bar{L}_{n,1}^0(t) \\ &\quad - \int_0^t \big[\mu_1(\bar{Q}_n^0(s) \wedge 1) + \theta_1(\bar{Q}_n^0(s) - 1)^+\big]ds, \end{aligned} \quad (5.8)$$

and

$$\begin{aligned} \hat{Q}_n^0(t) &= \hat{Q}_n(0) + \hat{A}_n(t) - \hat{S}_{n,1}^0(t) - \hat{L}_{n,1}^0(t) - \frac{n\mu_1 - \lambda_n}{\sqrt{n}}t - \mu_2 R_{n,1}^0(t) - \theta_2 R_{n,2}^0(t) \\ &\quad + (\mu_1 - \mu_2)V_n(t) - \int_0^t \big[\mu_1(\hat{Q}_n^0(s) \wedge 0) + \theta_1\big(\hat{Q}_n^0(s) \vee 0\big)\big]ds, \end{aligned} \quad (5.9)$$

where

$$\hat{S}_{n,1}^0(t) \equiv \frac{1}{\sqrt{n}}\Big(S_1\Big(n\mu_1 \int_0^t (\bar{Q}_n^0(s) \wedge 1)ds\Big) - n\mu_1 \int_0^t (\bar{Q}_n^0(s) \wedge 1)ds\Big),$$

$$\hat{L}_{n,1}^0(t) \equiv \frac{1}{\sqrt{n}}\Big(L_1\big(n\theta_1 \int_0^t (\bar{Q}_n^0(s) - 1)^+)ds\big) - n\theta_1 \int_0^t (\bar{Q}_n^0(s) - 1)^+)ds\Big),$$

and

$$\bar{S}^0_{n,1}(t) \equiv \frac{1}{\sqrt{n}}\hat{S}^0_{n,1}(t), \quad \bar{L}^0_{n,1}(t) \equiv \frac{1}{\sqrt{n}}\hat{L}^0_{n,1}(t), \quad t \geq 0,$$

and V_n is defined in (2.14).

We will sketch the proof of the processes $\hat{Q}^0_n$ converging weakly to the limit process $\hat{Q}$ defined in Theorem 3.3 in (D, M_1) since it is essentially the same as Theorem 7.1 of Pang et al. (2007). Then we prove that the processes $\hat{Q}^0_n$ and $\hat{Q}_n$ are asymptotically equivalent in D.

Theorem 5.2. *Under the assumptions of Theorem 3.3,*

$$\hat{Q}^0_n \Rightarrow \hat{Q} \quad in \quad (D, M_1) \quad as \quad n \to \infty,$$

where $\hat{Q}$ is defined in (3.3).

Proof (sketch). As in §7.1 of Pang et al. (2007), the processes $(\hat{S}^0_{n,1}, \hat{L}^0_{n,1})$ in (5.9) are square integrable martingales with respect to the filtration $\mathbf{F}^0_n \equiv \{\mathcal{F}^0_n(t) : t \geq 0\}$ where

$$\begin{aligned}\mathcal{F}^0_n(t) \equiv\ & \sigma\Big\{Q_n(0), S_1\Big(n\mu_1 \int_0^s \big(\bar{Q}^0_n(u) \wedge 1\big)du\Big), L_1\big(n\theta_1 \int_0^t (\bar{Q}^0_n(s) - 1)^+)ds\big) : 0 \leq s \leq t\Big\} \\ & \vee \sigma\Big(A_n(s), U_n(s), \eta_n(s) : s \geq 0\Big) \vee \mathcal{N},\end{aligned}$$

and $\mathcal{N}$ is the collection of all null sets. The predictable quadratic variation processes $\langle \hat{S}^0_{n,1} \rangle$ and $\langle \hat{L}^0_{n,1} \rangle$ are defined by

$$\langle \hat{S}^0_{n,1} \rangle(t) \equiv \mu_1 \int_0^t \big(\bar{Q}^0_n(s) \wedge 1\big)ds, \quad \langle \hat{L}^0_{n,1} \rangle(t) \equiv \theta_1 \int_0^t (\bar{Q}^0_n(s) - 1)^+ ds, \quad \text{for all} \quad t \geq 0.$$

Moreover, the sequence of martingale processes $\{(\hat{S}^0_{n,1}, \hat{L}^0_{n,1}) : n \geq 1\}$ is stochastically bounded in the space D_2. Next, by (5.8), we have the crude bound $0 \leq \bar{Q}^0_n(t) \leq \bar{Q}_n(0) + \bar{A}_n(t)$ for any $t \geq 0$. So we obtain

$$R^0_{n,1}(t) \leq \sqrt{n} \int_0^t (1 + \bar{Q}_n(0) + \bar{A}_n(s))(1 - U_n(s))ds + 2V_n(t) \leq (3 + \bar{Q}_n(0) + \bar{A}_n(t))V_n(t),$$

and

$$R^0_{n,2}(t) \leq \sqrt{n} \int_0^t (1 + \bar{Q}_n(0) + \bar{A}_n(s))(1 - U_n(s))ds \leq (1 + \bar{Q}_n(0) + \bar{A}_n(t))V_n(t).$$

By the assumption on the initial condition $Q_n(0)$, (2.3) and Lemma 2.2, the sequence of processes $\{(R^0_{n,1}, R^0_{n,2}) : n \geq 1\}$ is stochastically bounded in D_2. Hence, by Lemmas 3.3, 5.8 and 6.2 of Pang et al. (2007), the sequence of processes $\{\hat{Q}^0_n : n \geq 1\}$ is stochastically bounded in D.

Now by (5.7) and applying the FWLLN for the stochastic bounded sequences of processes in D in Lemma 5.9 of Pang et al. (2007), we obtain the FWLLN: $\bar{Q}_n^0 \Rightarrow \omega$ in D as $n \to \infty$ where $\omega(t) = 1$ for $t \geq 0$. Then, by applying continuous mapping theorem to the function $\phi : D \to D_2$ defined by

$$\phi(x)(t) = \Big(\int_0^t \mu_1(x(s) \wedge 1)ds, \int_0^t \theta_1(x(s) - 1)^+ ds \Big), \quad t \geq 0,$$

we obtain $\big(\langle \hat{S}_{n,1}^0 \rangle, \langle \hat{L}_{n,1}^0 \rangle\big) \Rightarrow \big(\mu_1 e, \zeta\big)$ in D_2 as $n \to \infty$, where $e(t) = t$ and $\zeta(t) = 0$ for all $t \geq 0$. By the martingale FCLT (Theorem 7.1 in Either and Kurtz (1986), Whitt (2007)), we obtain the weak convergence of the processes $(\hat{S}_{n,1}^0, \hat{L}_{n,1}^0) \Rightarrow (B \circ \mu_1 e, \zeta)$, where B is a standard Brownian motion.

By Lemma 2.2 and $\bar{Q}_n^0 \Rightarrow \omega$ in D as $n \to \infty$, we have

$$(R_{n,1}^0, R_{n,2}^0) \Rightarrow (-R, R) \quad \text{in} \quad (D_2, M_1) \quad \text{as} \quad n \to \infty.$$

So we have the joint convergence

$$\big(\hat{Q}_n(0), \hat{A}_n, \hat{S}_{n,1}^0, \hat{L}_{n,1}^0, V_n, R_{n,1}^0, R_{n,2}^0\big) \Rightarrow \big(\hat{Q}(0), \hat{A}, B \circ \mu_1 e, \zeta, V, -R, R\big) \quad \text{as} \quad n \to \infty,$$

in $\mathbb{R} \times (D, M_1) \times (D_2, J_1) \times (D_3, M_1)$, where $e(t) = t$ and $\zeta(t) = 0$ for all $t \geq 0$.

The representation of the processes $\hat{Q}_n^0$ in (5.9) corresponds to the the mapping in (5.4) with the functions g defined by $g(w) = -\mu_1(w \wedge 0) - \theta_1(w \vee 0)$ for all $w \in \mathbb{R}$, $b = 1$ and $q = h = 0$. By the continuous mapping theorem applying to the addition operation and the mapping in (5.4), we obtain the weak convergence of the processes $\hat{Q}_n^0$ to the process $\hat{Q}$ in (D, M_1). When we apply continuous mapping theorem to the addition operation, we need the assumption that the processes $\hat{A}$, V and R have no simultaneous jumps w.p.1. ∎

Theorem 5.3. *Under the assumptions of Theorem 3.3, the processes $\hat{Q}_n$ and $\hat{Q}_n^0$ are asymptotically equivalent as $n \to \infty$, so that $\hat{Q}_n \Rightarrow \hat{Q}$ in (D, M_1) as $n \to \infty$.*

Proof. By the representation of the processes $\hat{Q}_n$ and $\hat{Q}_n^0$, we have

$$
\begin{aligned}
& |\hat{Q}_n(t) - \hat{Q}_n^0(t)| \\
\leq\ & |\hat{S}_{n,1}(t) - \hat{S}_{n,1}^0(t)| + |\hat{L}_{n,1}(t) - \hat{L}_{n,1}^0(t)| + |\hat{S}_{n,2}(t)| + |\hat{L}_{n,2}(t)| + \mu_2|R_{n,1}(t) - R_{n,1}^0(t)| \\
& + \theta_2|R_{n,2}(t) - R_{n,2}^0(t)| + \mu_1\Big|\int_0^t \big(\hat{Q}_n(s) \wedge 0\big)U_n(s)ds - \int_0^t \big(\hat{Q}_n^0(s) \wedge 0\big)ds\Big| \\
& + \theta_1\Big|\int_0^t (\hat{Q}_n(s) \vee 0)U_n(s)ds - \int_0^t (\hat{Q}_n^0(s) \vee 0)ds\Big| \\
\leq\ & |\hat{S}_{n,1}(t) - \hat{S}_{n,1}^0(t)| + |\hat{L}_{n,1}(t) - \hat{L}_{n,1}^0(t)| + |\hat{S}_{n,2}(t)| + |\hat{L}_{n,2}(t)| + \mu_2|R_{n,1}(t) - R_{n,1}^0(t)| \\
& + \theta_2|R_{n,2}(t) - R_{n,2}^0(t)| + (\mu_1 + \theta_1)\int_0^t \big|\hat{Q}_n(s)\big|(1 - U_n(s))ds \\
& + \mu_1\int_0^t \Big|\big(\hat{Q}_n(s) \wedge 0\big) - \big(\hat{Q}_n^0(s) \wedge 0\big)\Big|ds + \theta_1\int_0^t \Big|(\hat{Q}_n(s) \vee 0) - (\hat{Q}_n^0(s) \vee 0)\Big|ds \\
\leq\ & |\hat{S}_{n,1}(t) - \hat{S}_{n,1}^0(t)| + |\hat{L}_{n,1}(t) - \hat{L}_{n,1}^0(t)| + |\hat{S}_{n,2}(t)| + |\hat{L}_{n,2}(t)| + \mu_2|R_{n,1}(t) - R_{n,1}^0(t)| \\
& + \theta_2|R_{n,2}(t) - R_{n,2}^0(t)| + (\mu_1 + \theta_1)\int_0^t \big|\hat{Q}_n(s)\big|(1 - U_n(s))ds \\
& + (\mu_1 \vee \theta_1)\int_0^t \Big|\hat{Q}_n(s) - \hat{Q}_n^0(s)\Big|ds, \quad \text{for all} \quad t \geq 0.
\end{aligned}
$$

We will apply the result of Problem 1.5.25 (solution on p. 45) in Karatzas and Shreve (1991), which says that for a sequence of continuous local martingales $\{M^{(n)} : n \geq 1\}$ with filtration $\mathbf{F}$ and any stopping time T of the same filtration, if $\langle M^{(n)}\rangle_T \to 0$ in probability as $n \to \infty$, then $\max_{0\leq t\leq T} |M_t^{(n)}| \to 0$ in probability as $n \to \infty$.

We can define the augmented filtration $\mathbf{F}_n^1 = \mathbf{F}_n \vee \mathbf{F}_n^0$ such that such that $\hat{S}_{n,1} - \hat{S}_{n,1}^0$ and $\hat{L}_{n,1} - \hat{L}_{n,1}^0$ are square integrable martingales with respect to the filtration $\mathbf{F}_n^1$ and their predictable quadratic variation processes are given by

$$
\begin{aligned}
\langle \hat{S}_{n,1} - \hat{S}_{n,1}^0\rangle(t) &= \mu_1\int_0^t [(\bar{Q}_n(s) \wedge 1)U_n(s) - (\bar{Q}_n^0(s) \wedge 1)]ds \\
&\leq \mu_1\int_0^t [(\bar{Q}_n(s) \wedge 1) - (\bar{Q}_n^0(s) \wedge 1)]ds, \quad t \geq 0,
\end{aligned}
$$

and

$$
\langle \hat{L}_{n,1} - \hat{L}_{n,1}^0\rangle(t) = \theta_1\int_0^t [(\bar{Q}_n(s) - 1)^+ - (\bar{Q}_n^0(s) - 1)^+]ds, \quad t \geq 0.
$$

By Lemma 5.1, the processes $\hat{S}_{n,2}$ and $\hat{L}_{n,2}$ are also square integrable martingales with respect to the filtration $\mathbf{F}_n^1$ and their predictable quadratic variation processes are bounded by

$$
\langle \hat{S}_{n,2}\rangle(t) \leq \mu_2\int_0^t (1 - U_n(s))ds = \mu_2 C_{D,n}(t), \quad t \geq 0,
$$

and

$$
\langle \hat{L}_{n,2}\rangle(t) \leq \theta_2\int_0^t (|\bar{Q}_n(s)| + 1)(1 - U_n(s))ds, \quad t \geq 0.
$$

Analogous to the proof of the stochastic boundedness of the sequence of processes $\{\hat{Q}_n^0 : n \geq 1\}$ in D and the FWLLN: $\bar{Q}_n^0 \Rightarrow \omega$ in D as $n \to \infty$ where $\omega(t) = 1$ for $t \geq 0$, we can prove that the sequence of processes $\{\hat{Q}_n : n \geq 1\}$ is stochastically bounded in D and the FWLLN holds: $\bar{Q}_n \Rightarrow \omega$ in D as $n \to \infty$. so we have $\langle \hat{S}_{n,1} - \hat{S}_{n,1}^0 \rangle(T) \Rightarrow 0$ and $\langle \hat{L}_{n,1} - \hat{L}_{n,1}^0 \rangle(T) \Rightarrow 0$ for any $T > 0$ as $n \to \infty$, which implies that $||\hat{S}_{n,1} - \hat{S}_{n,1}^0||_T \Rightarrow 0$ and $||\hat{S}_{n,1} - \hat{S}_{n,1}^0||_T \Rightarrow 0$ as $n \to \infty$. Also by Lemma 2.2 and the stochastic boundedness of $\bar{Q}_n$, we have $\langle \hat{S}_{n,2} \rangle(T) \Rightarrow 0$ and $\langle \hat{L}_{n,2} \rangle(T) \Rightarrow 0$ for any $T > 0$ as $n \to \infty$, which implies that $||\hat{S}_{n,2}||_T \Rightarrow 0$ and $||\hat{L}_{n,2}||_T \Rightarrow 0$ as $n \to \infty$. By Lemma 2.2 and the convergence of $\bar{Q}_n \Rightarrow \omega$ in D as $n \to \infty$, we have $(R_{n,1}, R_{n,2}) \Rightarrow (-R, R)$ in (D_2, M_1) as $n \to \infty$. So $||R_{n,1} - R_{n,1}^0||_T \Rightarrow 0$ and $||R_{n,2} - R_{n,2}^0||_T \Rightarrow 0$ for any $T > 0$ as $n \to \infty$.

Now let $\epsilon > 0$ and $\delta > 0$ be given and fix $T > 0$. Consider $K > 0$ such that $P\big(||\hat{Q}_n||_T > K\big) < \epsilon$. On $\{||\hat{Q}_n||_T > K\}$,

$$\begin{aligned} |\hat{Q}_n(t) - \hat{Q}_n^0(t)| \quad \leq \quad & |\hat{S}_{n,1}(t) - \hat{S}_{n,1}^0(t)| + |\hat{L}_{n,1}(t) - \hat{L}_{n,1}^0(t)| + |\hat{S}_{n,2}(t)| + |\hat{L}_{n,2}(t)| \\ & + \mu_2 |R_{n,1}(t) - R_{n,1}^0(t)| + \theta_2 |R_{n,2}(t) - R_{n,2}^0(t)| + (\mu_1 + \theta_1) K C_{D,n}(t) \\ & + (\mu_1 \vee \theta_1) \int_0^t \Big| \hat{Q}_n(s) - \hat{Q}_n^0(s) \Big| ds. \end{aligned}$$

By Gronwall's inequality, on $\{||\hat{Q}_n||_T > K\}$,

$$\begin{aligned} |\hat{Q}_n(t) - \hat{Q}_n^0(t)| \quad \leq \quad & \Big(|\hat{S}_{n,1}(t) - \hat{S}_{n,1}^0(t)| + |\hat{L}_{n,1}(t) - \hat{L}_{n,1}^0(t)| + |\hat{S}_{n,2}(t)| + |\hat{L}_{n,2}(t)| \\ & + \mu_2 |R_{n,1}(t) - R_{n,1}^0(t)| + \theta_2 |R_{n,2}(t) - R_{n,2}^0(t)| \\ & + (\mu_1 + \theta_1) K C_{D,n}(t) \Big) e^{(\mu_1 \vee \theta_1) T}. \end{aligned}$$

By Lemma 2.2 and the above analysis, we can find $n_0 \equiv n_0(\epsilon, \delta, T)$ such that

$$||\hat{Q}_n - \hat{Q}_n^0||_T \leq \delta \quad \text{for all} \quad n \geq n_0.$$

Hence

$$P\big(||\hat{Q}_n - \hat{Q}_n^0||_T > \delta\big) \leq \epsilon \quad \text{for all} \quad n \geq n_0(\epsilon, \delta, T).$$

Therefore, $||\hat{Q}_n - \hat{Q}_n^0||_T \Rightarrow 0$ as $n \to \infty$. ∎

Proof of Theorem 3.2. As in the proof of Theorem 3.3, the proof has two steps. The first step is to show $\bar{Q}_n^0 \Rightarrow \bar{Q}$ in D as $n \to \infty$, which is similar to the argument of Theorem 5.2, so we omit its proof. The second step is to show that the processes $\bar{Q}_n$ and $\bar{Q}_n^0$ are asymptotically equivalent as $n \to \infty$, for which we only highlight the following key equation and the rest of the proof is the

same as in Theorem 5.3.

$$
\begin{aligned}
& |\bar{Q}_n(t) - \bar{Q}_n^0(t)| \\
\le\ & |\bar{S}_{n,1}(t) - \bar{S}_{n,1}^0(t)| + |\bar{L}_{n,1}(t) - \bar{L}_{n,1}^0(t)| + |\bar{S}_{n,2}(t)| + |\bar{L}_{n,2}(t)| \\
& + \mu_1 \Big| \int_0^t \big(\bar{Q}_n(s) \wedge 1\big) U_n(s) ds - \int_0^t \big(\bar{Q}_n^0(s) \wedge 1\big) ds \Big| \\
& + \theta_1 \Big| \int_0^t (\bar{Q}_n(s) - 1)^+ U_n(s) ds - \int_0^t (\bar{Q}_n^0(s) - 1)^+ ds \Big| \\
& + \mu_2 \Big| \int_0^t \big(\bar{Q}_n(s) \wedge \frac{\eta_n(s)}{n}\big)(1 - U_n(s)) ds \Big| + \theta_2 \Big| \int_0^t \big(\bar{Q}_n(s) - \frac{\eta_n(s)}{n}\big)^+ (1 - U_n(s)) ds \Big| \\
\le\ & |\bar{S}_{n,1}(t) - \bar{S}_{n,1}^0(t)| + |\bar{L}_{n,1}(t) - \bar{L}_{n,1}^0(t)| + |\bar{S}_{n,2}(t)| + |\bar{L}_{n,2}(t)| \\
& + \mu_1 \int_0^t \big|\bar{Q}_n(s) \wedge 1\big| (1 - U_n(s)) ds + \theta_1 \int_0^t \big|(\bar{Q}_n(s) - 1)^+\big| (1 - U_n(s)) ds \\
& + \mu_2 \Big| \int_0^t \big(\bar{Q}_n(s) \wedge \frac{\eta_n(s)}{n}\big)(1 - U_n(s)) ds \Big| + \theta_2 \Big| \int_0^t \big(\bar{Q}_n(s) - \frac{\eta_n(s)}{n}\big)^+ (1 - U_n(s)) ds \Big| \\
& + (\mu_1 \vee \theta_1) \int_0^t |\bar{Q}_n(s) - \bar{Q}_n^0(s)| ds \quad \text{for all} \quad t \ge 0. \quad \blacksquare
\end{aligned}
$$

5.6. Proof of Theorem 3.4

First of all, by the fluid limit in Theorem 3.2, for any $\epsilon \in (0, (\lambda - \mu_1)/\theta_1)$ and each $T > 0$, there exists some n_0 large enough such that for all $n \ge n_0$,

$$\inf_{0 \le t \le T} Q_n(s) \ge n(1 + \frac{\lambda - \mu_1}{\theta_1} - \epsilon) > n.$$

This will simplify the martingale representation of $\hat{Q}_n^{ED}$ in (5.3) for large n,

$$
\begin{aligned}
\hat{Q}_n^{ED}(t) \ =\ & \hat{Q}_n^{ED}(0) + \hat{A}_n(t) - \hat{S}_{n,1}(t) - \hat{S}_{n,2}(t) - \hat{L}_{n,1}(t) - \hat{L}_{n,2}(t) \\
& - \theta_1 \int_0^t \hat{Q}_n^{ED}(s) U_n(s) ds - \theta_2 \int_0^t \hat{Q}_n^{ED}(s)(1 - U_n(s)) ds + (\mu_2 - \theta_2) R_n(t) \\
& + \Big(\lambda\Big(1 - \frac{\theta_2}{\theta_1}\Big) + \Big(\frac{\theta_2}{\theta_1}\mu_1 - \mu_2\Big)\Big) V_n(t).
\end{aligned}
$$

The proof is basically the same as that in the QED regime except that the martingale processes $\hat{L}_{n,1} \Rightarrow B \circ (\lambda - \mu_1)e$ in D as $n \to \infty$ where B is a standard Brownian motion and $\lambda > \mu_1$.

6. Conclusion

We have established fluid limits and refined stochastic limits for the queue-length process in a many-server queueing model with exponential service and patience times, subject to exogenous regenerative service interruptions. A highlight is the FCLT in Theorem 3.3 showing that even asymptotically negligible service interruptions can have a significant performance impact through

unmatched jumps in the limit process. There are many further research topics worth pursuing. First, it remains to establish the stochastic-process limits for the queue-length process for nonexponential service and patience distributions. Second, the conjecture that the steady-state distribution of the limit process $\hat{Q}$ in (3.5) has a stochastic-decomposition property remains to be proved. Third, for multiclass queueing models, it would be interesting to see how service interruptions affect the asymptotically optimal scheduling policies established in Atar (2005), Gurvich and Whitt (2008) and references therein. Fourth, stochastic-process limits for waiting times remain to be proved; see Talreja and Whitt (2008) for limit without service interruptions.

7. Acknowledgements.

This research was supported by NSF Grant DMI-0457095.

8. *

References

[1] Altman, E. and Y. Uri. 2006. Analysis of Customers' Impatience in Queues with Server Vacations. *Queueing Systems*. 52, 261-279.

[2] Atar, R. 2005. Scheduling Control for Queueing Systems with Many Servers: Asymptotic Optimality in Heavy Traffic. *The Annals of Applied Probability.* Vol. 15, No. 4, 2606-2650.

[3] Atar, R. 2008. Central Limit Theorem for a Many-server Queue with Random Service Rates. *The Annals of Applied Probability.* Vol. 18, No. 4, 1548-1568.

[4] Baykal-Gursoy, M. and W. Xiao. 2004. Stochastic Decomposition in $M/M/\infty$ Queues with Markov Modulated Service Rates. *Queueing Systems.* 48: 75-88.

[5] Billingsley, P. 1999. *Convergence of Probability Measures.* Second Edition. Wiley, New York.

[6] Boxma, O.J., J.A. Weststrate and U. Yechiali. 1993. A Globally Gated Polling System with Server Interruptions, and Applications to the Repairman Problem. *Probability in the Engineering and Informational Sciences*. 7, 187-208.

[7] Chao X. and Y.Q. Zhao. 1997. Analysis of Multi-Server Queues with Station and Server Vacations. *EJOR.*

[8] Chen, H. and W. Whitt. 1993. Diffusion Approximations for Multiclass Feedforward Queueing Networks. *Ann. Appl. Prob.* 10, 828-876.

[9] Chen, H. and D.D. Yao. 1992. A Fluid Model for Systems with Random Disruptions. *Operations Research*. Vol. 40, 239-247.

[10] Choudhury, G., A. Mandelbaum, M.I. Reiman and W. Whitt. 1997. Fluid and Diffusion Limits for Queues in Slowly Changing Random Environments. *Stochastic Models*. Vol. 13, No. 1, 121-146.

[11] Cont, R. and P. Tankov. 2004 *Financial Modelling with Jump Processes*. Chapman and Hall/CRC.

[12] D'Auria, B. 2007. Stochastic Decomposition of the $M/G/\infty$ Queue in a Random Environment. *Operations Research Letters*. Vol. 35, Issue 6, 805-812.

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

[14] Falin G. 2008. The $M/M/\infty$ Queue in a Random Environment. *Queueing Systems*. 58, 65-76.

[15] Garnett, O., A. Mandelbaum and M.I. Reiman. 2002. Designing a Call Center with Impatient Customers. *Manufacturing and Service Operations Management*. 4, 208-227.

[16] Gurvich, I. and W. Whitt. 2008. Scheduling Flexible Servers with Convex Delay Costs in Many-Server Service Systems. *Manufacturing and Service Operations Management*. Forthcoming.

[17] Halfin, S. and W. Whitt. 1981. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research*. Vol. 29, No. 3, 567-588.

[18] Jayawardene, A.K. and O. Kella. 1996. $M/G/\infty$ with Alternating Renewal Breakdowns. *Queueing Systems*. 22, 79-95.

[19] Karatzas, I. and S.E. Shreve. 1991. *Brownian Motion and Stochastic Calculus*. Second Edition. Springer.

[20] Kella, O. and W. Whitt. 1990. Diffusion Approximations for Queues with Server Vacations. *Adv. Appl. Prob.* 22, 706-729.

[21] Kella, O. and W. Whitt. 1991. Queues with Server Vacations and Levy Processes with Secondary Jump Input. *Annals of Applied Probability*. Vol. 1, No. 1, 104-117.

[22] Kella, O. and W. Whitt. 1992. A Storage Model With a Two-Stage Random Environment. *Operations Research*. Vol. 40, No. 2, 257-262.

[23] Mitrany, I.L. and B. Avi-Itzhak. 1968. A Many Server Queue with Service Interruptions. *Operations Research*. Vol. 16, 628-638.

[24] O'Cinneide, C. and P. Purdue. 1986. The $M/M/\infty$ Queue in a Random Environment. *J. Appl. Probab.* 23, 175-184.

[25] Pang, G., R. Talreja and W. Whitt. 2007. Martingale Proofs of Many-Server Heavy-Traffic Limits for Markovian Queues. *Probability Surveys*. 4, 193–267.

[26] Pang, G. and W. Whitt. 2008a. Service Interruptions in Large-Scale Service Systems. Submitted to *Management Science*.

[27] Pang, G. and W. Whitt. 2008b. Continuity of a Queueing Integral Representation in the M_1 Topology. Submitted to *Annals of Applied Probability*.

[28] Protter, P. 2003. *Stochastic Integration and Differential Equations.* Second Edition. Springer, New York.

[29] Sato, K. 1999. *Levy Processes and Infinitely Divisible Distributions.* Cambridge University Press.

[30] Situ, R. 2005. *Theory of Stochastic Differential Equations with Jumps and Applications.* Springer, New York.

[31] Talreja R. and W. Whitt. 2008. Heavy-Traffic Limits for Waiting Times in Many-Server Queues with Abandonment. *Submitted to Annals of Applied Probability*.

[32] Tian, N. and Z.G. Zhang. 2003a. Stationary Distributions of $GI/M/c$ Queues with PH type Vacations. *Queueing Systems*. 44, 183-202.

[33] Tian, N. and Z.G. Zhang. 2003b. Analysis of Queueing Systems with Synchronous Single Vacations for Some Servers. *Queueing Systems*. 45, 161-175.

[34] Tian, N. and Z.G. Zhang. 2003c. Analysis on Queueing Systems with Synchronous Vacations of Partial Servers. *Performance Evaluation*. 52, 269-282.

[35] White, H.C. and L.S. Christie. 1958. Queueing with Preemptive Priorities or with Breakdowns. *Operations Research*. 6, 79-95.

[36] Whitt, W. 2002. *Stochastic-Process Limits*, Springer.

[37] Whitt, W. 2004. Efficiency-Driven Heavy traffic Approximations for Many Server Queues with Abandonments. *Management Science*. Vol. 50, No. 10, 1449-1461.

[38] Whitt, W. 2007. Proofs of the Martingale Functional Central Limit Theorem: A Review. *Probability Surveys*. Vol. 4, 268-302. 2007.

[39] Wolfe, S.J. 1982. On A Continuous Analogue of the Stochastic Difference Equation $X_n = \rho X_{n-1} + B_n$. *Stochastic Processes and their Applications*. 12, 301-312.