# The Multiclass GI/PH/N Queue in the Halfin-Whitt Regime

*A. A. Puhalskii*
University of Colorado, Denver
and Institute for Problems in Information Transmission, Moscow

*M. I. Reiman*
Bell Labs, Lucent Technologies
Murray Hill, New Jersey 07974

March 27, 2000

**Abstract**

We consider a multiserver queue in the heavy-traffic regime introduced and studied by Halfin and Whitt [7] who investigated the case of a single customer class with exponentially distributed service times. Our purpose is to extend their analysis to a system with multiple customer classes, priorities, and phase-type service distributions. We prove a weak convergence limit theorem showing that a properly defined and normalized queue length process converges to a particular $K$ dimensional diffusion process, where $K$ is the number of phases in the service time distribution. We also show that a properly normalized waiting time process converges to a simple functional of the limit diffusion for the queue length.

## 1 Introduction and Summary

Multiserver queueing systems arise in many applications, and have generally proven to be more difficult to analyze than single server queues. In this paper we consider a GI/PH/N queue (renewal arrivals; 'phase type' service distribution; $N$ servers; infinite waiting room) with several customer classes. We also allow for two priority levels. We analyze this queueing system in the asymptotic regime pioneered by Halfin and Whitt [7], where $N \to \infty$ and $\rho_N \to 1$, with $\rho_N$ denoting the traffic intensity (defined precisely below) in the system with $N$ servers. More specifically, the Halfin-Whitt regime entails $\sqrt{N}(1 - \rho_N) \to \beta$ as $N \to \infty$, where $-\infty < \beta < \infty$.

We examine both the 'queue length' process as well as waiting times. We show that a properly centered and normalized version of a properly defined queue length process converges to a particular $K$ dimensional diffusion process, where $K$ is the number of phases in the service distribution. We also show that properly normalized, the waiting times converge to a simple functional of the diffusion limit for the queue length.

There is a substantial literature on multiserver queues. Our purpose here is not to provide a survey or comprehensive bibliography, so we restrict ourselves to papers that help put the contributions of this paper in perspective. We first describe some exact results. The simplest version of the queueing system under consideration is the M/M/N or Erlang C queue, first analyzed by Erlang [6], which has Poisson arrivals and exponential service times. The queue length process is a birth-death process, and the steady-state distribution has a simple form. The GI/PH/N queue was studied by Neuts [15], who showed that the steady state queue length distribution has a matrix geometric form, and described algorithms for calculating it. The numerical complexity grows quickly in the number of servers and/or phases. In particular, with $K$ phases and $N$ servers, the matrices that arise in the matrix geometric solution have dimension $\binom{N+K-1}{N}$.

The most closely related limit theorem is that of Halfin and Whitt [7], who considered the GI/M/N queue as $N \to \infty$ and $\rho_N \to 1$ such that $\sqrt{N}(1-\rho_N) \to \beta$ with $-\infty < \beta < \infty$. First restricting their attention to the M/M/N queue with $\rho_N < 1$, they showed that the steady-state probability that a customer must wait in the queue approaches a limit $\alpha$ with $0 < \alpha < 1$ as $N \to \infty$ if and only if $0 < \beta < \infty$. For the GI/M/N queue they showed that a properly centered and normalized version of the queue length process converges to a one-dimensional diffusion. We restrict our description of their result to the M/M/N case. Let $Q^N(t)$ denote the number of customers in the system with $N$ servers, let $\mu$ be the service rate (per server), and let $\lambda^N$ denote the arrival rate to the system with $N$ servers. (Then $\rho_N = \lambda^N/N\mu$.) Assume that $\lambda^N = N\mu - \beta\mu\sqrt{N}$, with $-\infty < \beta < \infty$, and let

$$X^N(t) = \frac{Q^N(t) - N}{\sqrt{N}}, \quad t \geq 0.$$

Let $X = (X(t), t \geq 0)$, be the unique strong solution to

$$X(t) = X(0) - \mu\beta t - \mu \int_0^t [X(s) \wedge 0]\, ds + \sqrt{2\mu} W(t),$$

where $W$ is a standard Wiener process that is independent of $X(0)$. Then $X$ is a one-dimensional diffusion with infinitesimal drift $m(x)$ given by

$$m(x) = \begin{cases} -\mu\beta, & x \geq 0 \\ -\mu(x+\beta), & x \leq 0 \end{cases}$$

and constant infinitesimal variance $2\mu$. This process can be viewed intuitively as 'piecing together' a Brownian motion with drift on $[0, \infty)$ and an Ornstein-Uhlenbeck process on $(-\infty, 0]$. Halfin and Whitt showed that, if $X^N(0) \xrightarrow{d} X(0)$ ($\xrightarrow{d}$ denotes convergence in distribution), then $X^N \xrightarrow{d} X$ on $D([0, \infty), R)$ (see Section 2 for a definition of $D([0, \infty), R)$).

Our results can be viewed as the natural generalization of the result of Halfin and Whitt to phase type service distributions and multiple customer classes. Halfin and Whitt briefly discussed the GI/$H_2$/N queue, where $H_2$ denotes the hyperexponential distribution with 2 phases (a mixture of 2 exponentials). They introduced a three dimensional process whose three components are: the number of customers in the queue (waiting for service), the number of customers in service in phase 1, and the number of customers in service in phase 2. This is a three dimensional random walk, and Halfin and Whitt showed that under the proper centering and normalization the infinitesimal drift and variance converge. But there is more that needs to be done here in order to conclude weak convergence of the process. In particular, the infinitesimal parameters are defined only on a two dimensional 'boundary' sub-manifold consisting of two perpendicular half planes: Either there are customers in the queue, in which case the total number of customers in service is $N$, or there are fewer than $N$ customers in service, in which case the queue is empty. It is not clear how to associate a unique diffusion process with these parameters, and it is also not clear that the convergence of the infinitesimal parameters will imply weak convergence of the process. (There are several general theorems in the latter spirit, such as Theorem 8.3.1 in [13] and Theorem IX.3.48 in [12]; it is not clear if they would apply here, since they require checking extra conditions.) The approach taken in the present paper involves constructing a $K$ (as opposed to $K + 1$) dimensional process for a system with $K$ phases in the service distribution.

It is worth contrasting the asymptotic regime of Halfin and Whitt with two other "heavy traffic" regimes. A heavy traffic limit theorem for the GI/G/N queue was proved by Iglehart and Whitt [10]. They considered $N$ fixed and took the limit as $\rho \to 1$. They showed that properly normalized versions of the queue length and waiting time processes converge to one dimensional reflected Brownian motions. (No centering is needed in this case.) One of the characteristics of this limiting regime is that the probability of having at least one customer in the queue converges to unity as $\rho \to 1$. Thus, although this limit theorem provides a simple limit process for multiserver queues with general service time distributions, it is not appropriate for systems with many servers where not all customers need to wait.

The GI/G/N queue with $N \to \infty$ in such a way that $(N\mu - \lambda^N)/\sqrt{N} \to \infty$, for different initial conditions and assumptions on the interarrival and service time distributions, was considered by Borovkov [2, 3], Iglehart [8, 9] and Whitt [18]. The growth condition on $N$ makes the system asymptotically equivalent to the infinite server queue GI/G/$\infty$. Phase-type service, which is the focus of our paper, was considered in Whitt [18]. There the process of interest is the $K$ dimensional process ($K$ is the number of phases in the service time distribution) whose $k^{\text{th}}$ component is the number of customers in service in phase $k$, $1 \leq k \leq K$. He showed that a properly centered and normalized queue length process converges to a $K$ dimensional Ornstein-Uhlenbeck process. It seems intuitively clear that the 'local' behavior of the GI/PH/N system when there are idle servers should be identical to the GI/PH/$\infty$ system. Our limit theorems bear this point out.

The final related reference we want to mention is Mandelbaum, Massey, and Reiman [14]. They provide fluid and diffusion limits for what they call 'Markovian service networks' with many servers. They allow multiple stations and time varying parameters. The results there cover the M/M/N queue in the Halfin-Whitt regime as well as a special case of the M/PH/N queue where there is exactly one 'initial phase' ($K' = 1$ in the notation introduced below in Section 2).

The rest of this paper is organized as follows. Section 2 provides a precise specification of the model as well as a statement of the main results. The proof of Theorem 1, with Poisson arrivals and one priority level is contained in Section 3. Section 4 contains a proof of Theorem 2, which covers a system with two priority levels. Section 5 covers the extension to renewal arrivals. Two technical lemmas needed in the proofs of the main results are proved in the appendix.

## 2 Statement of Main Results

We begin with the case of Poisson arrivals and one priority level. Consider an $N$-server queue with infinite wating room. Arrivals to the $N$-th queue are Poisson with rate $\lambda^N$, while service time distributions are held fixed with cumulative distribution function $F(x)$ of 'phase-type' [15]. Arrival and service processes are independent, and customers are taken into service on a first-come-first-served basis.

A phase-type (PH) distribution corresponds to the distribution of the life-time of a transient, continuous-time, finite-state Markov chain. Let $K$ denote the number of states,

$(p_1, \ldots, p_K)$ denote the initial distribution, $(p_{ij},\ 1 \le i,\ j \le K)$, denote the transition probabilities, $p_{i0}$, $1 \le i \le K$, the probabilities of absorbtion ($p_{i0} = 1 - \sum_{j=1}^{K} p_{ij}$), and $\mu_i^{-1}$, $1 \le i \le K$, the mean sojourn times in the states. It is worth remarking here that $p_{ii} = 0$, $1 \le i \le K$. A state of this Markov chain may also be called a phase, or a phase of service. Assume that the states are labeled in such a way that $p_i > 0$ for $1 \le i \le K'$, and $p_i = 0$ for $K' < i \le K$, with $1 \le K' \le K$. We associate a customer class with each possible initial state, so that there are $K'$ customer classes. The arrival rate of class $k$ is then $p_k \lambda^N$, and the service time distribution of a class $k$ customer is the phase-type distribution $F(\cdot)$ conditioned to have initial state $k$, $1 \le k \le K'$. Note that we are making no assumption on the structure of $P = (p_{ij}, 1 \le i, j \le K)$ other than transience. (Thus, although it is possible that each state is reachable from only one starting state, so that we can determine the initial state, and hence class, of a customer from its current state – we call such a Markov chain 'separated', this is not required. We use class only to keep track of the customers in the queue, keeping track of customers in service by state. If it is desirable to keep track of customers in service by class, and the original Markov chain is not separated, it can be modified, at the expense of increasing the number of states, into a separated Markov chain.)

Let us consider an auxiliary discrete-time Markov chain with $K + 1$ states indexed by $0, 1, \ldots, K$, so that the probability of going from state $i$ to state $j$ is $p_{ij}$ for $i = 1, \ldots, K$, $j = 0, 1, \ldots, K$, and is equal to $p_j$ for $i = 0$ and $j = 1, \ldots, K$. Obviously, this is an ergodic Markov chain. Let $(\eta_0, \eta_1, \ldots, \eta_K)$ denote its stationary distribution, i.e.,

$$\eta_0 p_i + \sum_{j=1}^{K} p_{ji} \eta_j = \eta_i,\ 1 \le i \le K,\ \sum_{i=0}^{K} \eta_i = 1, \eta_i > 0\,. \tag{2.1}$$

The mean service time is $\mu^{-1}$, where

$$\mu = \left( \sum_{i=1}^{K} \frac{\eta_i}{\eta_0 \mu_i} \right)^{-1} . \tag{2.2}$$

Let

$$q_i = \mu \frac{\eta_i}{\eta_0 \mu_i}, \quad 1 \le 1 \le K\,. \tag{2.3}$$

Then $(q_1, \ldots, q_K)$ is the stationary distribution of the continuous time Markov chain obtained from the $K + 1$ state auxiliary Markov chain by making state 0 instantaneous, and having the sojourn time in state $i$ exponential with rate $\mu_i$. We also denote

$$\mu_{ij} = p_{ij} \mu_i, \quad i = 1, \ldots, K,\ j = 0, 1, \ldots, K\,. \tag{2.4}$$

We assume that, for some $\beta$, with $-\infty < \beta < \infty$,

$$\lambda^N = \mu N - \mu\beta\sqrt{N}, \quad N \geq 1 . \tag{2.5}$$

The traffic intensity is defined as $\rho_N = \lambda_N / N\mu$. Thus $\rho_N = 1 - \beta/\sqrt{N}$, and $\sqrt{N}(1-\rho_N) = \beta$.

Let $\hat{Q}_i^N(t)$, $1 \leq i \leq K$ denote the number of customers over all the servers who are being served at time $t$ in phase $i$, $Q_0^N(t)$ denote the number of customers in the queue at time $t$ (in the system but not in service), $\tilde{Q}_i^N(t)$, $1 \leq i \leq K'$, denote the number of class $i$ customers in the queue at time $t$ (in the system but not in service), and $\tilde{Q}_i^N(t) = 0, K' < i \leq K$.

Let

$$Q_i^N(t) = \hat{Q}_i^N(t) + \tilde{Q}_i^N(t), \quad 1 \leq i \leq K . \tag{2.6}$$

Then $Q_i^N(t)$ is the total number of "phase $i$" customers in the system, i.e., those who either (are in the queue and) start service in phase $i$ or are currently being served in phase $i$. It is not difficult to see that $Q_0^N(t)$ can be recovered from $Q_i^N(t), 1 \leq i \leq K$, by the relation

$$Q_0^N(t) = \left( \sum_{i=1}^K Q_i^N(t) - N \right)^+ . \tag{2.7}$$

We also note that

$$Q_0^N(t) = \sum_{i=1}^K \tilde{Q}_i^N(t). \tag{2.8}$$

Let us introduce the following independent Poisson processes:

$$A^N = (A^N(t),\ t \geq 0) \text{ has rate } \lambda^N, \quad N = 1, 2, \ldots$$
$$S_{ij}^l = (S_{ij}^l(t),\ t \geq 0) \text{ has rate } \mu_{ij}, \quad i = 1, \ldots, K,\ j = 0, 1, \ldots, K,\ l = 1, 2, \ldots .$$

We interpret $A^N$ as the arrival process. The interpretation of $S_{ij}^l$ is a bit more involved. The process $S_{ij}^l$ corresponds to transitions from phase $i$ to phase $j$ in the $l^{\text{th}}$ server that is serving a customer in phase $i$. When there are fewer than $l$ customers being served in phase $i$ at the moment of a jump in $S_{ij}^l$, the jump has no effect on the system state. Similarly, $S_{i0}^l$ corresponds to service completions from phase $i$.

All the random processes are assumed to have right-continuous sample paths with left limits and, hence, are considered as random elements of appropriate Skorohod spaces $D([0,\infty), R^d)$ of $R^d$-valued right-continuous functions with left-hand limits. We endow $D([0,\infty), R^d)$ with the Skorohod-Prohorov-Lindvall metric, which turns it into a Polish space, Liptser and Shiryaev [13]. For $f = (f(t), t \geq 0) \in D([0,\infty), R^d)$ and $t > 0$, we denote by $f(t-)$ the left limit at $t$; we also set $f(0-) = f(0)$.

Let $\{\alpha_j, j \geq 1\}$ be i.i.d. random variables independent of the arrival and service processes that take values in the set $\{1,\ldots,K\}$ with probabilities $P(\alpha_j = i) = p_i$. (Though the range of the $\alpha_j$ is actually $\{1,\ldots,K'\}$, for notational purposes it is conveninet to consider them as taking values in $\{1,\ldots,K\}$.) Informally, $\alpha_j$ indicates the phase in which the $j$-th customer to enter service after time zero begins service.

We assume as given initial values $Q_0^N(0)$ and $(\hat{Q}_1^N(0),\ldots,\hat{Q}_K^N(0))$ that are independent of $A^N$, $\{S_{ij}^l,\ 1 \leq i \leq K,\ 0 \leq j \leq K,\ l \geq 1\}$ and $\{\alpha_j,\ j \geq 1\}$. In addition, we assume that

$$\tilde{Q}_i^N(0) = \sum_{j=1}^{Q_0^N(0)} 1(\alpha_j = i), \quad \text{and if } Q_0^N(0) > 0, \quad \text{then } \sum_{i=1}^{K} \hat{Q}_i^N(0) = N\,.$$

The limit theorems that we prove are for properly centered and normalized versions of the above processes. Let

$$X_i^N(t) = \frac{Q_i^N(t) - Nq_i}{\sqrt{N}}, \quad 1 \leq i \leq K,\ t \geq 0\,, \tag{2.9}$$

$$\tilde{X}_i^N(t) = \frac{\tilde{Q}_i^N(t)}{\sqrt{N}}, \quad 1 \leq i \leq K,\ t \geq 0\,, \tag{2.10}$$

and

$$X_0^N(t) = \frac{Q_0^N(t)}{\sqrt{N}}, \quad t \geq 0\,. \tag{2.11}$$

In addition, for $1 \leq i \leq K$, let $X_i^N = (X_i^N(t),\ t \geq 0)$, $\tilde{X}_i^N = (\tilde{X}_i^N(t),\ t \geq 0)$ and $X_0^N = (X_0^N(t),\ t \geq 0)$. Finally, let $X^N = (X_1^N,\ldots,X_K^N)$ and $\tilde{X}^N = (\tilde{X}_1^N,\ldots,\tilde{X}_K^N)$. We consider $X^N$ and $\tilde{X}^N$ as random elements of the Skorohod space $D([0,\infty), R^K)$. The process $X_0^N$ is similarly a random element of the Skorohod space $D([0,\infty),\ R)$.

We define $x^+ = x \vee 0$, and let $\xrightarrow{d}$ denote convergence in distribution on the appropriate space. Let $X(0) = (X_1(0),\ldots,X_K(0)) \in R^K$ and $\tilde{X}(0) = (\tilde{X}_1(0),\ldots,\tilde{X}_K(0)) \in R^K$ be random vectors.

**Theorem 1:** *Let the above conditions hold. If $X^N(0) \xrightarrow{d} X(0)$, then $X^N \xrightarrow{d} X = (X_1,\ldots,X_K)$, where $X_i = (X_i(t),\ t \geq 0)$, $1 \leq i \leq K$, is the solution of the equation*

$$\begin{aligned} X_i(t) \;=\; & X_i(0) - \mu\beta p_i t + \sum_{j=1, j\neq i}^{K} \mu_{ji} \int_0^t X_j(s)ds \\ & - \mu_i \int_0^t X_i(s)ds - \left( \sum_{j=1,j\neq i}^{K} p_j\mu_{ji} - p_i\mu_i \right) \int_0^t \left( \sum_{j=1}^{K} X_j(s) \right)^+ ds \\ & + \; Y_i(t), \quad 1 \leq i \leq K\,, \end{aligned} \tag{2.12}$$

*and the processes* $(Y_i(t),\ t \geq 0)$, $1 \leq i \leq K$, *are defined as*

$$Y_i(t) = \sqrt{p_i\mu}W_i(t) + \sum_{j=1,j\neq i}^{K} \sqrt{q_j\mu_{ji}}W_{ji}(t) - \sum_{j=0,j\neq i}^{K} \sqrt{q_i\mu_{ij}}W_{ij}(t),\ 1 \leq i \leq K,$$

*with* $(W_i(t),\ t \geq 0)$, $1 \leq i \leq K$, *and* $(W_{ij}(t),\ t \geq 0)$, $1 \leq i \leq K, 0 \leq j \leq K$, *being independent standard Wiener processes that are also independent of* $X(0)$.

**Remark 1:** Since the drifts in (2.12) are Lipshitz continuous in $X$, equation (2.12) has a unique strong solution [11].

**Remark 2:** *The processes* $Y_i$ *are driftless Wiener processes with covariances*

$$EY_i(t)^2 = 2q_i\mu_i t, \quad EY_i(t)Y_{i'}(t) = -(\mu_{ii'}q_i + \mu_{i'i}q_{i'})t,\ i \neq i'\,.$$

**Remark 3:** *The process* $X$ *is a* $K$ *dimensional diffusion with infinitesimal drift vector* $m(x) = (m_1(x), \ldots, m_K(x))$ *with*

$$m_i(x) = \begin{cases} -\mu\beta p_i + \displaystyle\sum_{j=1,j\neq i}^{K} x_j\mu_{ji} - \mu_i x_i, & \displaystyle\sum_{\ell=1}^{K} x_\ell \leq 0 \\ -\mu\beta p_i + \displaystyle\sum_{j=1,j\neq i}^{K} \left[x_j - p_j \sum_{\ell=1}^{K} x_\ell\right]\mu_{ji} & \\ \qquad\qquad -\mu_i\left[x_i - p_i\displaystyle\sum_{\ell=1}^{K} x_\ell\right], & \displaystyle\sum_{\ell=1}^{K} x_\ell \geq 0, \end{cases}$$

*and infinitesimal covariance matrix* $\Gamma = (\Gamma_{ij}, 1 \leq i,j \leq K)$ *with*

$$\Gamma_{ii} = 2q_i\mu_i, \quad 1 \leq i \leq K$$

*and*

$$\Gamma_{ij} = -q_i\mu_{ij} - q_j\mu_{ji}, \quad 1 \leq i \neq j \leq K\,.$$

By (2.2) and (2.3)

$$\sum_{i=1}^{K} q_i = 1\,, \tag{2.13}$$

so that by (2.7)

$$X_0^N(t) = \left(\sum_{i=1}^{K} X_i^N(t)\right)^+\,. \tag{2.14}$$

Let $X_0(t) = \left(\sum_{i=1}^{K} X_i(t)\right)^+$, and $X_0 = (X_0(t),\ t \geq 0)$. Then Theorem 1 and (2.14) yield the following.

**Corollary 1:** Under the conditions of Theorem 1 $X_0^N \xrightarrow{d} X_0$.

We show, in Lemma 3, that for $0 < T < \infty$, $\epsilon > 0$, $1 \le i \le K$,

$$\lim_{N\to\infty} P\left(\sup_{0\le t\le T}\left|\tilde{X}_i^N(t) - p_i X_0^N(t)\right| > \epsilon\right) = 0\,. \tag{2.15}$$

This state space collapse result, combined with Corollary 1, yields the following limit for the queue length of each class.

**Corollary 2:** *Under the conditions of Theorem 1* $\tilde{X}^N \xrightarrow{d} \tilde{X} = (\tilde{X}_1,\dots,\tilde{X}_K)$*, where* $\tilde{X}_i = (\tilde{X}_i(t), t \ge 0)$*,* $1 \le i \le K$*, are given by*

$$\tilde{X}_i(t) = p_i\left(\sum_{j=1}^K X_j(t)\right)^+ ,\ t \ge 0\,.$$

Theorem 1 also allows us to establish asymptotics of waiting-time processes. Let $w^N(t)$ denote the virtual waiting time at time $t$ and $w_i^N$, the waiting time of the $i$th customer, and let us introduce the processes $\hat{w}^N = (\sqrt{N}\, w^N(t), t \ge 0)$ and $\tilde{w}^N = (\sqrt{N}\, w^N_{\lfloor Nt\rfloor+1}, t \ge 0)$.

**Corollary 3:** *Under the conditions of Theorem 1 the processes* $\hat{w}^N$ *and* $\tilde{w}^N$ *converge in distribution in* $D([0,\infty), R)$ *to the respective processes* $(X_0(t)/\mu, t \ge 0)$ *and* $(X_0(t/\mu)/\mu, t \ge 0)$.

We next consider a system with priorities. Let us assume that in the above model customers have two priority levels, high and low. High priority customers are those that start service in phases 1 through $K'' < K'$, low priority customers start service in phases $K''+1$ through $K'$. The low priority customers are allowed to enter service only when there are no high priority customers in the queue. (This is a non-preemptive system.)

Let $A_H^N(t)$ and $A_L^N(t)$ denote, respectively, the number of high and low priority customers arrived in $[0,t]$; obviously, $A_H^N = (A_H^N(t), t \ge 0)$ and $A_L^N = (A_L^N(t), t \ge 0)$ are independent Poisson processes with respective rates $\lambda_H^N = p_H\lambda^N$ and $\lambda_L^N = p_L\lambda^N$, where $p_H = \sum_{i=1}^{K''} p_i$ and $p_L = \sum_{i=K''+1}^{K'} p_i$ are probabilities with which an arbitrary customer is a high or low priority customer, respectively. We also introduce analogs of the random variables $\alpha_i$: random variables $\alpha_{H,j}, j = 1,2,\dots$ assume values 1 through $K''$ with probabilities $p_{H,i} = p_i/p_H, 1 \le i \le K''$ and indicate the phases in which successive high priority customers start service; similarly, random variables $\alpha_{L,j}, j = 1,2,\dots$ assume values $K''+1$ through $K'$

with probabilities $p_{L,i} = p_i/p_L, K''+1 \le i \le K'$ and indicate the phases in which successive low priority customers start service. We assume as given values $Q^N_{H,0}(0)$ and $Q^N_{L,0}(0)$ of the initial quantities of high-priority and low-priority customers, respectively, in the queue, and assume that

$$\tilde{Q}_i^N(0) = \sum_{j=1}^{Q^N_{H,0}(0)} 1(\alpha_{H,j} = i),\ 1 \le i \le K'', \quad \tilde{Q}_i^N(0) = \sum_{j=1}^{Q^N_{L,0}(0)} 1(\alpha_{L,j} = i),\ K''+1 \le i \le K',$$

$$\text{and if } \ Q^N_{L,0}(0) + Q^N_{H,0}(0) > 0, \quad \text{then } \sum_{i=1}^{K} \hat{Q}_i^N(0) = N\,.$$

The random objects $(Q^N_{H,0}(0), Q^N_{L,0}(0), \hat{Q}_i^N(0), 1 \le i \le K)$, $(\alpha_{H,j}, j = 1, 2, \ldots)$, $(\alpha_{L,j}, j = 1, 2, \ldots)$, $A^N_H$, $A^N_L$, and $S^l_{ij}$, $1 \le i \le K$, $0 \le j \le K$, $l \ge 1$, are assumed to be mutually independent.

**Theorem 2:** *Under the conditions and notation of Theorem 1, the processes $X^N$ converge in distribution on $D([0,\infty), R^K)$ to the process $X = (X_1, \ldots, X_K)$, where $X_i = (X_i(t),\ t \ge 0)$, $1 \le i \le K$, that is the solution of the equation*

$$\begin{aligned}
X_i(t) &= X_i(0) - \mu\beta p_i t + \sum_{j=1}^{K} \mu_{ji} \int_0^t X_j(s)ds \\
&- \mu_i \int_0^t X_i(s)ds - \left( \sum_{j=K''+1}^{K'} p_{L,j}\mu_{ji} \right) \int_0^t \left( \sum_{j=1}^{K} X_j(s) \right)^+ ds \\
&+ \ Y_i(t), \quad 1 \le i \le K'', K'+1 \le i \le K\,, \\
X_i(t) &= X_i(0) - \mu\beta p_i t + \sum_{j=1}^{K} \mu_{ji} \int_0^t X_j(s)ds \\
&- \mu_i \int_0^t X_i(s)ds - \left( \sum_{j=K''+1, j\ne i}^{K'} p_{L,j}\mu_{ji} - p_{L,i}\mu_i \right) \int_0^t \left( \sum_{j=1}^{K} X_j(s) \right)^+ ds \\
&+ \ Y_i(t), \quad K''+1 \le i \le K'\,,
\end{aligned}$$

*where the processes $(Y_i(t),\ t \ge 0)$, $1 \le i \le K$, are defined as in Theorem 1.*

We now extend Theorem 1 to the case of renewal arrivals. Let us assume that $A^N = (A^N(t), t \ge 0)$, the arrival process to the $N$th system, is a renewal process with $\xi_i^N$, $i \ge 1$, denoting the times between arrivals. Stated another way, the $\xi_i^N$, $i \ge 1$, are nonnegative, i.i.d. and

$$A^N(t) = \max\left\{ k : \sum_{j=1}^{k} \xi_j^N \le t \right\}\,.$$

We moreover assume that

$$P(\xi_1^N > 0) = 1, \quad E(\xi_1^N)^2 < \infty\,. \tag{2.16}$$

Denote

$$\lambda^N = (E\xi_1^N)^{-1},\ (\sigma^N)^2 = \text{Var}\ \xi_1^N\,.$$

We assume that $\lambda^N$ is given by (2.5), that

$$\lim_{N\to\infty} N^2(\sigma^N)^2 = \sigma^2 > 0\,, \tag{2.17}$$

and the Lindeberg condition holds:

$$\lim_{N\to\infty} N^2 E(\xi_1^N)^2\ 1(\xi_1^N\sqrt{N} > \epsilon) = 0, \quad \epsilon > 0\,. \tag{2.18}$$

We preserve the independence assumptions of Theorem 1.

**Theorem 3:** *Let conditions* (2.16)–(2.18) *hold and the service time distribution be the same as in Theorem 1. Then the assertion of Theorem 1 holds with*

$$Y_i(t) = V_i(t) + \sum_{j=1,j\neq i}^{K} \sqrt{q_j\mu_{ji}}W_{ji}(t) - \sum_{j=0,j\neq i}^{K} \sqrt{q_i\mu_{ij}}W_{ij}(t),\ 1 \le i \le K, \tag{2.19}$$

*where $((V_1(t),\dots,V_K(t)),\ t \ge 0)$ is a $K$ dimensional Wiener process with zero drift and covariances $EV_i(t)^2 = (p_i(1-p_i)+p_i^2\mu^2\sigma^2)\mu t$, $EV_i(t)V_{i'}(t) = p_ip_{i'}(\mu^2\sigma^2-1)\mu t$; $(W_{ij}(t),\ t \ge 0)$, $1 \le i \le K, 0 \le j \le K$, are standard Wiener processes, and all the processes on the right of* (2.19) *are mutually independent and also independent of $X(0)$.*

**Remark 4:** *The processes $(Y_i(t), t \ge 0)$, $1 \le i \le K$, above are zero drift Wiener processes with covariances*

$$\begin{aligned} EY_i(t)^2 &= \left[2\mu_i q_i + (\mu^2\sigma^2 - 1)\mu p_i^2\right] t\,, \\ EY_i(t)Y_{i'}(t) &= \left[-\mu_{ii'}q_i - \mu_{i'i}q_{i'} + (\mu^2\sigma^2 - 1)\mu p_i p_{i'}\right] t\,, \quad i \neq i'\,. \end{aligned}$$

**Remark 5:** *Theorem 2 has a similar extension to the case of renewal arrivals.*

**Remark 6:** *Appropriate analogs of corollaries 1–3 carry over to the settings of Theorems 2 and 3. Thus, for example, the normalized waiting time for high priority customers converges to zero.* (*Lemma 4 shows that the normalized number of high priority customers waiting in the queue converges to zero.*)

## 3 The Proof of Theorem 1

The proof consists of two main steps. The first step, which is broken down into several smaller steps culminates (in Lemma 3) in proving (2.15), i.e., that the processes $\tilde{X}_i^N$ and $p_i X_0^N$ are asymptotically indistinguishable. (This is an example of state-space collapse.) In the second step we write a stochastic integral representation for the process $X^N$. This stochastic integral contains an error term that, by Lemma 3, is asymptotically negligible. The proof is then completed by using a martingale diffusion limit theorem.

One interesting aspect of the proof is worthy of mention. In developing semimartingale decompositions for $Q_0^N$ and $\hat{Q}_i^N$, $1 \leq i \leq N$, we use a filtration ($\mathbb{F}^N$) that does not 'know the identity' of customers in the queue. This is essential for the proper application of Lemma A1. For the second part of the proof we introduce a second filtration ($\hat{\mathbb{F}}^N$) that does 'know the identity' of customers in the queue. This second filtration is needed because the martingale terms appearing in the semimartingale decomposition for $Q^N$ are not $\mathbb{F}^N$ adapted but are $\hat{\mathbb{F}}^N$ adapted.

Our first step is to develop semimartingale decompositions for the processes $Q_0^N$ and $\hat{Q}_i^N, 1 \leq i \leq K$. Let

$$D_j^N(t) = \sum_{l=1}^{N} \int_0^t 1\left(\hat{Q}_j^N(s-) \geq l\right) 1\left(Q_0^N(s-) > 0\right) dS_{j0}^l(s)\,, 1 \leq j \leq K\,, \tag{3.1}$$

$$D^N(t) = \sum_{j=1}^{K} D_j^N(t)\,, \tag{3.2}$$

and

$$B^N(t) = D^N(t) + \int_0^t 1\left(\sum_{j=1}^{K} \hat{Q}_j^N(s-) < N\right) dA^N(s). \tag{3.3}$$

Then $D_j^N(t)$ represents the number of customers that left the queue during the interval $(0,t]$ due to a termination of service in the $j$-th phase; $D^N(t)$ the number of customers that left the queue and went into service during $(0,t]$; and $B^N(t)$ the number of customers that started service during $(0,t]$.

The processes $Q_0^N$, $\hat{Q}_i^N$, $1 \leq i \leq K$, obey the equations

$$Q_0^N(t) = Q_0^N(0) + \int_0^t 1\left(\sum_{j=1}^{K} \hat{Q}_j^N(s-) = N\right) dA^N(s) - D^N(t)\,, \tag{3.4}$$

and

$$\begin{aligned}\hat{Q}_i^N(t) &= \hat{Q}_i^N(0) + \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s-) < N\right) \, 1\left(\alpha_{B^N(s)} = i\right) dA^N(s) \\ &+ \sum_{j=1}^K \int_0^t 1\left(\alpha_{B^N(s)} = i\right) dD_j^N(s) \\ &+ \sum_{j=1, j\neq i}^K \sum_{l=1}^N \int_0^t 1\left(\hat{Q}_j^N(s-) \geq l\right) dS_{ji}^l(s) \\ &- \sum_{j=0, j\neq i}^K \sum_{l=1}^N \int_0^t 1\left(\hat{Q}_i^N(s-) \geq l\right) dS_{ij}^l(s), \quad 1 \leq i \leq K\,. \end{aligned} \tag{3.5}$$

We define the $\sigma$-algebra $\mathcal{F}^N(t)$ by

$$\begin{aligned}\mathcal{F}^N(t) = \sigma\{&Q_0^N(0), \hat{Q}_i^N(0), A^N(s), S_{ij}^l(s), \alpha_{B^N(s)}; 1 \leq i \leq K, \\ &0 \leq j \leq K,\ l \geq 1,\ 0 \leq s \leq t\} \vee \mathcal{N}\,,\end{aligned}$$

where $\mathcal{N}$ denotes the family of $P$-null sets, and introduce the filtration $\mathbb{F}^N = (\mathcal{F}^N(t),\ t \geq 0)$ (right-continuity of $\mathbb{F}^N$ follows from Brémaud [4]). Relations (3.1)– (3.5) show that the processes $(B^N(t), t \geq 0)$, $(D_j^N(t), t \geq 0), 1 \leq j \leq K$, $(D^N(t), t \geq 0)$, $(\alpha_{B^N(t)}, t \geq 0)$, $Q_0^N$, and $\hat{Q}_i^N$ are $\mathbb{F}^N$-adapted. Note that $\tilde{Q}_i^N$ is in general not $\mathbb{F}^N$ adapted.

In the following lemma we provide the desired semimartingale decompositions for $Q_0^N$ and $\hat{Q}_i^N$.

**Lemma 1:** *The processes* $Q_0^N = (Q_0^N(t), t \geq 0)$ *and* $\hat{Q}_i^N = (\hat{Q}_i^N(t), t \geq 0), 1 \leq i \leq K$, *admit the decompositions*

$$\begin{aligned}Q_0^N(t) &= Q_0^N(0) + \lambda^N \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s) = N\right) ds \\ &- \sum_{j=1}^K \mu_{j0} \int_0^t \hat{Q}_j^N(s) 1\left(Q_0^N(s) > 0\right) ds + M_0^N(t)\,, \end{aligned} \tag{3.6}$$

$$\begin{aligned}\hat{Q}_i^N(t) &= \hat{Q}_i^N(0) + p_i \lambda^N \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s) < N\right) ds \\ &+ p_i \sum_{j=1}^K \mu_{j0} \int_0^t \hat{Q}_j^N(s) 1(Q_0^N(s) > 0) ds + \sum_{j=1, j\neq i}^K \mu_{ji} \int_0^t \hat{Q}_j^N(s) ds \\ &- \mu_i \int_0^t \hat{Q}_i^N(s) ds + \hat{M}_i^N(t),\ 1 \leq i \leq K\,, \end{aligned} \tag{3.7}$$

*where $M_0^N = (M_0^N(t), t \geq 0)$ and $\hat{M}_i^N = (\hat{M}_i^N(t), t \geq 0), 1 \leq i \leq K$, are $\mathbb{F}^N$–locally square integrable martingales, whose respective predictable quadratic variations $\langle M_0^N \rangle = (\langle M_0^N \rangle(t), t \geq 0)$ and $\langle \hat{M}_i^N \rangle = (\langle \hat{M}_i^N \rangle(t), t \geq 0)$, for some $b > 0$ and all $t \geq 0, N = 1, 2, \ldots$, satisfy the inequalities*

$$\langle M_0^N \rangle(t) \leq bNt, \quad \langle \hat{M}_i^N \rangle(t) \leq bNt,\ 1 \leq i \leq K.$$

**Proof.** By a well-known fact (see, e.g., [11]), the Poisson processes $A^N$ and $S_{ij}^l$ admit the representations

$$A^N(t) = \lambda^N t + M^N(t)\,, \tag{3.8}$$

$$S_{ij}^l(t) = \mu_{ij} t + M_{ij}^l(t)\,, \quad 1 \leq i \leq K\,, \quad 0 \leq j \leq K\,, l \geq 1, \tag{3.9}$$

where $M^N = (M^N(t),\ t \geq 0)$ and $M_{ij}^l = (M_{ij}^l(t),\ t \geq 0)$ are independent locally square-integrable martingales relative to the associated natural filtrations with respective predictable quadratic variations, Liptser and Shiryaev [13]

$$\langle M^N \rangle(t) = \lambda^N t\,, \tag{3.10}$$

$$\langle M_{ij}^l \rangle(t) = \mu_{ij} t\,. \tag{3.11}$$

Since $\mathcal{F}^N(t) \subset \sigma\{Q_0^N(0), \hat{Q}_i^N(0), 1 \leq i \leq K\} \vee \sigma\{A^N(s), 0 \leq s \leq t\} \vee \sigma\{S_{ij}^l(s), 0 \leq s \leq t, 1 \leq i \leq K, 0 \leq j \leq K, l \geq 1\} \vee \sigma\{\alpha_i, i \geq 1\} \vee \mathcal{N}$, $(Q_0^N(0), \hat{Q}_i^N(0))$, $A^N$, $S_{ij}^l, 1 \leq i \leq K, 0 \leq j \leq K, l \geq 1$, and $\{\alpha_i, i \geq 1\}$ are mutually independent, and $A^N$ and $S_{ij}^l, 1 \leq i \leq K, 0 \leq j \leq K, l \geq 1$ are $\mathbb{F}^N$-adapted, the semimartingale decompositions (3.8) and (3.9) also hold relative to $\mathbb{F}^N$ in that $M^N$ and $M_{ij}^l, 1 \leq i \leq K, 0 \leq j \leq K, l \geq 1$, are orthogonal $\mathbb{F}^N$-locally square-integrable martingales with respective predictable quadratic variations given by (3.10) and (3.11).

Substituting $S_{j0}^l(t)$ from (3.9) into (3.1), we have

$$D_j^N(t) = \mu_{j0} \int_0^t \hat{Q}_j^N(s) 1(Q_0^N(s) > 0) ds + \sum_{l=1}^N \int_0^t 1\left(\hat{Q}_j^N(s-) \geq l\right) 1\left(Q_0^N(s-) > 0\right) dM_{j0}^l(s)\,,$$

which means that the process $D_j^N = (D_j^N(t), t \geq 0)$ has the $\mathbb{F}^N$–compensator $\left(\mu_{j0} \int_0^t \hat{Q}_j^N(s) 1(Q_0^N(s) > 0) ds, t \geq 0\right)$; hence, by Lemma A1 the process $\left(p_i \mu_{j0} \int_0^t \hat{Q}_j^N(s) 1(Q_0^N(s) > 0) ds, t \geq 0\right)$ is the $\mathbb{F}^N$-compensator of the process $\left(\int_0^t 1(\alpha_{B^N(s)} = i) dD_j^N(s), t \geq 0\right)$.

Similarly, by (3.8) and Lemma A1, $\left(p_i\lambda^N \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s) < N\right), t \geq 0\right) ds$ is the $F^N$-compensator of $\left(\int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s-) < N\right) 1(\alpha_{B^N(s)} = i) dA^N(s), t \geq 0\right)$ . Thus, we can write

$$\int_0^t 1\left(\alpha_{B^N(s)} = i\right) dD_j^N(s) = p_i\mu_{j0} \int_0^t \hat{Q}_j^N(s) 1(Q_0^N(s) > 0)\, ds + M_{D,ji}^N(t) \qquad (3.12)$$

and

$$\int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s-) < N\right) 1(\alpha_{B^N(s)} = i) dA^N(s) = \lambda^N p_i \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s) < N\right) ds + M_{A,i}^N(t) \qquad (3.13)$$

where $M_{D,ji}^N = (M_{D,ji}^N(t), t \geq 0)$ and $M_{A,i}^N = (M_{A,i}^N(t), t \geq 0)$ are $\mathbb{F}^N$–locally square integrable martingales with respective predictable quadratic variations

$$\langle M_{D,ji}^N \rangle(t) = p_i\mu_{j0} \int_0^t \hat{Q}_j^N(s) 1(Q_0^N(s) > 0)\, ds, \qquad (3.14)$$

$$\langle M_{A,i}^N \rangle(t) = p_i\lambda^N \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s) < N\right) ds\,. \qquad (3.15)$$

(The latter follows by the fact that the predictable quadratic variation of the locally square integrable martingale that appears in the semimartingale decomposition of a point process with a continuous compensator coincides with the compensator, Liptser and Shiryaev [13].)

Next, by (3.9), we also have

$$\sum_{l=1}^N \int_0^t 1\left(\hat{Q}_j^N(s-) \geq l\right) dS_{ji}^l(s) = \mu_{ji} \int_0^t \hat{Q}_j^N(s)\, ds + M_{S,ji}^N(t), \qquad (3.16)$$

where

$$M_{S,ji}^N(t) = \sum_{l=1}^N \int_0^t 1\left(\hat{Q}_j^N(s-) \geq l\right) dM_{ji}^l(s),\ j = 1, \ldots, K, i = 0, 1, \ldots, K. \qquad (3.17)$$

In view of (3.11) and the fact that the local martingales $M_{ij}^l, l = 1, 2, \ldots, i = 1, 2, \ldots, K, j = 0, 1, 2, \ldots, K$, are mutually independent, the predictable quadratic variations of the above processes are

$$\langle M_{S,ji}^N \rangle(t) = \mu_{ji} \int_0^t \hat{Q}_j^N(s)\, ds. \qquad (3.18)$$

Substituting (3.12), (3.13) and (3.16) into (3.5), and using

$$\sum_{j=0, j\neq i}^K \mu_{ij} = \mu_i, \qquad (3.19)$$

which follows by (2.4), yields the required representations for $\hat{Q}_i^N$ with

$$\hat{M}_i^N(t) = M_{A,i}^N(t) + \sum_{j=1}^K M_{D,ji}^N(t) + \sum_{j=1,j\neq i}^K M_{S,ji}^N(t) - \sum_{j=0,j\neq i}^K M_{S,ij}^N(t)\,. \tag{3.20}$$

The representation for $Q_0^N$ with

$$M_0^N(t) = \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s-) = N\right) dM^N(s) - \sum_{j=1}^K \sum_{i=1}^K M_{D,ji}^N(t), \tag{3.21}$$

follows by (3.4), (3.8), (3.12), and (3.2).

By construction, the processes $M_0^N = (M_0^N(t),\ t \geq 0)$ and $\hat{M}_i^N = (\hat{M}_i^N(t),\ t \geq 0)$, $1 \leq i \leq K$, are locally square-integrable martingales relative to $\mathbb{F}^N$. To estimate their predictable quadratic variations, we use the fact that, given two locally square-integrable martingales $Z_1 = (Z_1(t), t \geq 0)$ and $Z_2 = (Z_2(t), t \geq 0)$, their predictable covariation $\langle Z_1, Z_2\rangle = (\langle Z_1, Z_2\rangle(t), t \geq 0)$ satisfies the inequality $2\langle Z_1, Z_2\rangle(t) \leq \langle Z_1\rangle(t) + \langle Z_2\rangle(t)$, Liptser and Shiryaev [13, Probl. 1.8.9]. Applying this property to $M_0^N$ and $\hat{M}_i^N$ yields by (3.20) and (3.21)

$$\langle \hat{M}_i^N\rangle(t) \leq (3K-1)[\langle M_{A,i}^N\rangle(t) + \sum_{j=1,j\neq i}^K \langle M_{D,ji}^N\rangle(t) + \sum_{j=1,j\neq i}^K \langle M_{S,ji}^N\rangle(t) + \sum_{j=0,j\neq i}^K \langle M_{S,ij}^N\rangle(t)],$$

$$\langle M_0^N\rangle(t) \leq (K^2+1)[\langle M^N\rangle(t) + \sum_{j=1}^K \sum_{i=1}^K \langle M_{D,ji}^N\rangle(t)].$$

The required estimates follow now by (2.5), (3.10), (3.14), (3.15), (3.18) and the bound $\hat{Q}_j^N(s) \leq N$. □

We now introduce auxiliary processes $\bar{Q}_i^N = (\bar{Q}_i^N(t), t \geq 0)$ by

$$\bar{Q}_i^N(t) = p_i Q_0^N(t) + \hat{Q}_i^N(t),\ 1 \leq i \leq K\,, \tag{3.22}$$

and define the processes $\bar{X}_i^N = \big(\bar{X}_i^N(t),\ t \geq 0\big)$, $1 \leq i \leq K$, by

$$\bar{X}_i^N(t) = \frac{\bar{Q}_i^N(t) - Nq_i}{\sqrt{N}}\,. \tag{3.23}$$

Note that

$$\sum_{i=1}^K \bar{Q}_i^N(t) = \sum_{i=1}^K Q_i^N(t). \tag{3.24}$$

By (3.6) and (3.7) we obtain

$$\bar{Q}_i^N(t) = \bar{Q}_i^N(0) + p_i\lambda^N t + \sum_{j=1,j\neq i}^K \mu_{ji} \int_0^t \hat{Q}_j^N(s)ds$$
$$- \mu_i \int_0^t \hat{Q}_i^N(s)ds + \bar{M}_i^N(t)\,, \tag{3.25}$$

where

$$\bar{M}_i^N(t) = p_i M_0^N(t) + \hat{M}_i^N(t)\,. \tag{3.26}$$

We note that by (2.1), (2.2), (2.3) and (2.4), the $q_i$, $1 \le i \le K$, satisfy the equations

$$p_i\mu + \sum_{j=1, j\ne i}^K \mu_{ji} q_j - \mu_i q_i = 0\,. \tag{3.27}$$

Recalling (2.5), we then obtain by (3.25) the following equations for the processes $\bar{X}_i^N$ from (3.23):

$$\begin{aligned} \bar{X}_i^N(t) \;=\; & \bar{X}_i^N(0) - p_i\mu\beta t + \sum_{j=1, j\ne i}^K \mu_{ji} \int_0^t \hat{X}_j^N(s)ds \\ & -\mu_i \int_0^t \hat{X}_i^N(s)ds + \frac{1}{\sqrt{N}} \bar{M}_i^N(t)\,, \end{aligned} \tag{3.28}$$

where

$$\hat{X}_i^N(t) = \frac{\hat{Q}_i^N(t) - Nq_i}{\sqrt{N}}\,. \tag{3.29}$$

**Lemma 2:** *Under the conditions of Theorem 1, for every $T > 0$,*

*(i)* $\lim_{A\to\infty} \overline{\lim_{N\to\infty}}\, P\left(\sup_{t\le T} X_0^N(t) > A\right) = 0$ ,

*(ii)* $\lim_{A\to\infty} \overline{\lim_{N\to\infty}}\, P\left(\sup_{t\le T} |\hat{X}_i^N(t)| > A\right) = 0\,, 1 \le i \le K$ .

**Proof.** By (3.22), (2.7) and (3.24)

$$\hat{Q}_i^N(t) = \bar{Q}_i^N(t) - p_i \left(\sum_{j=1}^K \bar{Q}_j^N(t) - N\right)^+ ,$$

therefore, by (3.23), (2.13) and (3.29)

$$\hat{X}_i^N(t) = \bar{X}_i^N(t) - p_i \left(\sum_{j=1}^K \bar{X}_j^N(t)\right)^+ . \tag{3.30}$$

Substituting the latter into (3.28) yields

$$\begin{aligned} \bar{X}_i^N(t) \;=\; & \bar{X}_i^N(0) - p_i\mu\beta t + \sum_{j=1, j\ne i}^K \mu_{ji} \int_0^t \bar{X}_j^N(s)ds - \mu_i \int_0^t \bar{X}_i^N(s)ds \\ & + \left(p_i\mu_i - \sum_{j=1, j\ne i}^K p_j\mu_{ji}\right) \int_0^t \left(\sum_{j=1}^K \bar{X}_j^N(s)\right)^+ ds + \frac{1}{\sqrt{N}} \bar{M}_i^N(t)\,. \end{aligned}$$

This obviously implies that for some $0 < C < \infty$

$$\sum_{i=1}^{K} |\bar{X}_i^N(t)| \leq \sum_{i=1}^{K} |\bar{X}_i^N(0)| + |\mu\beta| t + \frac{1}{\sqrt{N}} \sum_{i=1}^{K} |\bar{M}_i^N(t)| + C \int_0^t \sum_{i=1}^{K} |\bar{X}_i^N(s)| ds \,. \tag{3.31}$$

Gronwall's inequality then yields

$$\sup_{t \leq T} \sum_{i=1}^{K} |\bar{X}_i^N(t)| \leq \left( \sum_{i=1}^{K} |\bar{X}_i^N(0)| + |\mu\beta| T + \frac{1}{\sqrt{N}} \sum_{i=1}^{K} \sup_{t \leq T} |\bar{M}_i^N(t)| \right) \cdot e^{CT} \,. \tag{3.32}$$

Since $X^N(0)$ converges in distribution to $X(0)$ as $N \to \infty$ by the assumptions of Theorem 1, we have using (2.14)

$$\lim_{A \to \infty} \overline{\lim_{N \to \infty}} \; P\left(X_0^N(0) > A\right) = 0 \,;$$

since $\tilde{Q}_i^N(0) \leq Q_0^N(0)$ by (2.8), it follows by (2.10) and (2.11) that

$$\lim_{A \to \infty} \overline{\lim_{N \to \infty}} \; P\left(\tilde{X}_i^N(0) > A\right) = 0 \,;$$

and, since by (3.22), (3.23), (2.6), (2.9), (2.10) and (2.11)

$$\bar{X}_i^N(t) = p_i X_0^N(t) + X_i^N(t) - \tilde{X}_i^N(t),$$

we conclude that

$$\lim_{A \to \infty} \overline{\lim_{N \to \infty}} \; P\left( \sum_{i=1}^{K} |\bar{X}_i^N(0)| > A \right) = 0 \,. \tag{3.33}$$

Next, since the process $\bar{M}_i^N = (\bar{M}_i^N(t),\ t \geq 0)$ is a locally square-integrable martingale with respect to $\mathbb{F}^N$ by (3.26), by the Lenglart–Rebolledo inequality, Liptser and Shiryaev [13], for $B > 0$,

$$P\left( \sup_{t \leq T} \frac{1}{\sqrt{N}} |\bar{M}_i^N(t)| > A \right) \leq \frac{B}{A^2} + P\left( \frac{1}{N} \langle \bar{M}_i^N \rangle(T) > B \right) \,. \tag{3.34}$$

By (3.26) and the inequality $2\langle M_0^N, \hat{M}_i^N \rangle(t) \leq \langle M_0^N \rangle(t) + \langle \hat{M}_i^N \rangle(t)$, we have that $\langle \bar{M}_i^N \rangle(t) \leq 2\left[ p_i^2 \langle M_0^N \rangle(t) + \langle \hat{M}_i^N \rangle(t) \right]$, so by Lemma 1

$$\langle \bar{M}_i^N \rangle(t) \leq rNt, \quad 1 \leq i \leq K \,, \tag{3.35}$$

for some $r > 0$. Therefore, by (3.34)

$$\lim_{A \to \infty} \overline{\lim_{N \to \infty}} \; P\left( \sup_{t \leq T} \frac{1}{\sqrt{N}} |\bar{M}_i^N(t)| > A \right) = 0, \tag{3.36}$$

which, combined with (3.32) and (3.33), allows us to conclude that

$$\lim_{A\to\infty} \overline{\lim_{N\to\infty}}\, P\left(\sup_{t\le T} \sum_{i=1}^{K} |\bar{X}_i^N(t)| > A\right) = 0\,. \tag{3.37}$$

The lemma follows by (3.30) and since, in view of (2.14),

$$X_0^N(t) = \left(\sum_{i=1}^{K} X_i^N(t)\right)^+ = \left(\sum_{i=1}^{K} \bar{X}_i^N(t)\right)^+ . \qquad \square$$

We now prove asymptotic indistinguishability of $\tilde{X}_i^N$ and $p_i X_0^N$, as well as of $\bar{X}_i^N$ and $X_i^N$.

**Lemma 3:** *Under the conditions of Theorem 1, for every $T > 0$, $\epsilon > 0$, and $1 \le i \le K$,*

*(i)* $\lim_{N\to\infty} P\left(\sup_{t\le T} |X_i^N(t) - \bar{X}_i^N(t)| > \epsilon\right) = 0\,.$

*(ii)* $\lim_{N\to\infty} P\left(\sup_{t\le T} \left|\tilde{X}_i^N(t) - p_i X_0^N(t)\right| > \epsilon\right) = 0\,.$

**Proof.** By (3.22) and (2.6)

$$\bar{Q}_i^N(t) - Q_i^N(t) = p_i Q_0^N(t) - \tilde{Q}_i^N(t)$$

so that, by (3.23), (2.9), (2.10) and (2.11),

$$\bar{X}_i^N(t) - X_i^N(t) = p_i X_0^N(t) - \tilde{X}_i^N(t).$$

Thus, parts (i) and (ii) are equivalent. We prove (ii) next.

We first note that the definition of $\tilde{Q}_i^N$ yields the representation

$$\tilde{Q}_i^N(t) = \sum_{j=Q_0^N(0)+A^N(t)-Q_0^N(t)+1}^{Q_0^N(0)+A^N(t)} 1(\alpha_j = i),\ 1 \le i \le K,\ t \ge 0\,. \tag{3.38}$$

Next, by (3.38), (2.10) and (2.11), for $C > 0$,

$$\begin{aligned} &P\left(\sup_{t\le T} |\tilde{X}_i^N(t) - p_i X_0^N(t)| > \epsilon\right) \\ &= P\left(\sup_{t\le T} \frac{1}{\sqrt{N}}\left|\sum_{j=Q_0^N(0)+A^N(t)-Q_0^N(t)+1}^{Q_0^N(0)+A^N(t)} [1(\alpha_j = i) - p_i]\right| > \epsilon\right) \\ &\le P\left(\frac{Q_0^N(0) + A^N(T)}{NT} > C\right) + P\left(\sup_{t\le T} X_0^N(t) > C\right) \\ &+ P\left(\sup_{\substack{0\le x\le C \\ x-\frac{C}{\sqrt{N}}\le y\le x}} \frac{1}{\sqrt{N}}\left|\sum_{j=\lfloor Ny\rfloor+1}^{\lfloor Nx\rfloor} [1(\alpha_j = i) - p_i]\right| > \epsilon\right) . \end{aligned} \tag{3.39}$$

Since $(A^N(t),\ t \geq 0)$ is a Poisson process with rate $\lambda^N$, equation (2.5) and the fact that $Q_0^N(0)/N \xrightarrow{P} 0$, which follows by Lemma 2(i), imply that

$$\lim_{C\to\infty} \overline{\lim_{N\to\infty}}\, P\left(\frac{Q_0^N(0) + A^N(T)}{NT} > C\right) = 0\,. \tag{3.40}$$

Also by Lemma 2(i), the second term on the right of (3.39) tends to 0 as $N \to \infty$ and $C \to \infty$. The proof is completed by proving that the third term on the right of (3.39) tends to 0 as $N \to \infty$.

Let

$$U_i^N(t) = \frac{1}{\sqrt{N}} \sum_{j=1}^{\lfloor Nt \rfloor} \left[1(\alpha_j = i) - p_i\right]\,.$$

By Donsker's Theorem [1] the processes $(U_i^N(t),\ t \geq 0)$ converge in distribution as $N \to \infty$ to the process $(\sqrt{(1-p_i)p_i}\, W(t),\ t \geq 0)$, where $(W(t),\ t \geq 0)$ is a standard Wiener process. Therefore, by the almost sure continuity of the Wiener process, for any $\delta > 0$,

$$\begin{aligned}
&\overline{\lim_{N\to\infty}}\, P\left(\sup_{\substack{0\leq x,y\leq A\\ |x-y|\leq\delta}} \left|U_i^N(x) - U_i^N(y)\right| > \epsilon\right)\\
&\leq P\left(\sup_{\substack{0\leq x,y\leq A\\ |x-y|\leq\delta}} \sqrt{(1-p_i)p_i}\, |W(x) - W(y)| \geq \epsilon\right)\,.
\end{aligned}$$

By the almost sure continuity of the Wiener process again, the latter converges to 0 as $\delta \to 0$, so we deduce that the right-most term in (3.39) also converges to 0 as $N \to \infty$. □

**Proof of Theorem 1.** We introduce a $\sigma$-algebra $\hat{\mathcal{F}}^N(t)$, defined by

$$\begin{aligned}
\hat{\mathcal{F}}^N(t) = \sigma\{&Q_0^N(0),\ \hat{Q}_i^N(0),\ A^N(s),\ S_{ij}^l(s),\ \alpha_1, \ldots, \alpha_{A^N(t)+Q_0^N(0)};\ 1 \leq i \leq K,\\
&0 \leq j \leq K,\ l \geq 1,\ 0 \leq s \leq t\} \vee \mathcal{N}\,,
\end{aligned}$$

and introduce the filtration $\hat{\mathbb{F}}^N = (\hat{\mathcal{F}}^N(t),\ t \geq 0)$. The argument used in the proof of Lemma 1 to justify semimartingale decompositions (3.8) and (3.9) relative to $\mathbb{F}^N$ applies to $\hat{\mathbb{F}}^N$ as well to the effect that $M^N$ and $M_{ij}^l, 1 \leq i \leq K, 0 \leq j \leq K, l \geq 1$, are orthogonal $\hat{\mathbb{F}}^N$-locally square-integrable martingales with respective predictable quadratic variations given by (3.10) and (3.11).

Let

$$A_i^N(t) = \sum_{j=1}^{A^N(t)} 1(\alpha_{j+Q_0^N(0)} = i),\ 1 \leq i \leq K\,. \tag{3.41}$$

By Lemma A1, $A_i^N = (A_i^N(t), t \geq 0), 1 \leq i \leq K$, are Poisson processes with respective decompositions

$$A_i^N(t) = p_i \lambda^N t + M_i^N(t)\,, \tag{3.42}$$

where $M_i^N = (M_i^N(t),\ t \geq 0)$ are $\hat{\mathbb{F}}^N$-locally square-integrable martingales. (Note that $A_i^N$ are adapted to $\hat{\mathbb{F}}^N$ but are not adapted to $\mathbb{F}^N$.)

By (3.38), (3.41) and the fact that

$$Q_0^N(0) + A^N(t) - Q_0^N(t) = B^N(t), \tag{3.43}$$

we can write

$$\tilde{Q}_i^N(t) = \tilde{Q}_i^N(0) + A_i^N(t) - \int_0^t 1(\alpha_{B^N(s)} = i)\, dB^N(s).$$

Therefore, by (2.6), (3.3), (3.2), and (3.5), noting that by (3.41) and (3.43)

$$A_i^N(t) = \int_0^t 1(\alpha_{A^N(s)+Q_0^N(0)} = i)\, dA^N(s) = \int_0^t 1(\alpha_{B^N(s)+Q_0^N(s)} = i)\, dA^N(s), \tag{3.44}$$

and that in the first integral on the right of (3.5) we may replace $\alpha_{B^N(s)}$ by $\alpha_{B^N(s)+Q_0^N(s)}$ since $Q_0^N(s) = 0$ if $\sum_{j=1}^K \hat{Q}_i^N(s-) < N$, we have

$$\begin{aligned} Q_i^N(t) \ &= \ Q_i^N(0) + A_i^N(t) + \sum_{l=1}^N \sum_{j=1, j\neq i}^K \int_0^t 1\left(\hat{Q}_j^N(s-) \geq l\right) dS_{ji}^l(s) \\ &\quad - \sum_{l=1}^N \sum_{j=0, j\neq i}^K \int_0^t 1\left(\hat{Q}_i^N(s-) \geq l\right) dS_{ij}^l(s),\ 1 \leq i \leq K\,. \end{aligned}$$

By (3.9), (3.42) and (3.19),

$$\begin{aligned} Q_i^N(t) \ &= \ Q_i^N(0) + p_i \lambda^N t + \sum_{j=1, j\neq i}^K \mu_{ji} \int_0^t \hat{Q}_j^N(s) ds \\ &\quad - \mu_i \int_0^t \hat{Q}_i^N(s) ds + \tilde{M}_i^N(t),\ 1 \leq i \leq K\,, \end{aligned} \tag{3.45}$$

where

$$\tilde{M}_i^N(t) \ = \ M_i^N(t) + \sum_{j=1, j\neq i}^K M_{S,ji}^N(t) - \sum_{j=0, j\neq i}^K M_{S,ij}^N(t),\ 1 \leq i \leq K \tag{3.46}$$

(the $M_{S,ij}^N$ are defined in (3.17)).

Substituting (2.9) and (3.29) into (3.45) and using (2.5) and (3.27), we obtain

$$\begin{aligned} X_i^N(t) \ &= \ X_i^N(0) - p_i \mu \beta t + \sum_{j=1, j\neq i}^K \mu_{ji} \int_0^t \hat{X}_j^N(s) ds \\ &\quad - \mu_i \int_0^t \hat{X}_i^N(s) ds + \frac{1}{\sqrt{N}} \tilde{M}_i^N(t),\ 1 \leq i \leq K\,. \end{aligned} \tag{3.47}$$

Next, the processes $M_i^N = (M_i^N(t), t \geq 0)$ and $M_{S,ij}^N = (M_{S,ij}^N(t), t \geq 0)$ are orthogonal $\hat{\mathbb{F}}^N$–locally square integrable martingales, whose predictable quadratic variations, in view of (3.42) and (3.18), are given by

$$\langle M_i^N \rangle(t) = p_i \lambda^N t,\ \langle M_{S,ij}^N \rangle(t) = \mu_{ij} \int_0^t \hat{Q}_i^N(s)\, ds.$$

Also Lemma 2(ii) implies that

$$\sup_{t \leq T} \left| \frac{\hat{Q}_i^N(t)}{N} - q_i \right| \xrightarrow{P} 0,\ T > 0.$$

Therefore, by (2.5)

$$\left\langle \frac{1}{\sqrt{N}} M_i^N \right\rangle (t) \xrightarrow{P} p_i \mu t,\ \left\langle \frac{1}{\sqrt{N}} M_{S,ij}^N \right\rangle (t) \xrightarrow{P} \mu_{ij} q_i t\,,$$

and, e.g., by Theorem 8.3.1 in Liptser and Shiryaev [13], the processes $\{M_i^N/\sqrt{N}, M_{S,ij}^N/\sqrt{N}$, $i = 1, \dots, K,\ \ j = 0, \dots, K, j \neq i\}$ converge jointly in distribution to $\{\sqrt{p_i \mu} W_i, \sqrt{q_i \mu_{ij}} W_{i,j}$, $i = 1, \dots, K,\ j = 0, \dots, K, j \neq i\}$; so, by the continuous mapping theorem and (3.46) the processes $\{\tilde{M}_i^N/\sqrt{N}, 1 \leq i \leq K\}$ converge jointly in distribution to the processes $\{Y_i, 1 \leq i \leq K\}$.

In view of (3.30) and Lemma 3(i), we can rewrite (3.47) as

$$\begin{aligned} X_i^N(t) \ =\ & X_i^N(0) - p_i \mu \beta t + \sum_{j=1}^K \mu_{ji} \int_0^t X_j^N(s) ds - \mu_i \int_0^t X_i^N(s) ds \\ & + \left( p_i \mu_i - \sum_{j=1}^K p_j \mu_{ji} \right) \int_0^t \left( \sum_{j=1}^K X_j^N(s) \right)^+ ds \\ & + \frac{1}{\sqrt{N}} \tilde{M}_i^N(t) + \tilde{\epsilon}_i^N(t)\,, \end{aligned}$$

where

$$\sup_{t \leq T} \left| \tilde{\epsilon}_i^N(t) \right| \xrightarrow{P} 0,\ 1 \leq i \leq K\,.$$

Since the process $(X_i^N(t), t \geq 0)$ is a continuous function of the process $(X_i^N(0) + \tilde{M}_i^N(t)/\sqrt{N} + \tilde{\epsilon}_i^N(t), t \geq 0)$, the result follows by the continuous mapping theorem. □

**Remark 7:** *The latter argument shows that the processes* $(M_i^N/\sqrt{N}, X_i^N, 1 \leq i \leq K)$ *converge jointly in distribution on* $D(R_+, R^{2K})$ *to the processes* $(\sqrt{p_i \mu} W_i, X_i, 1 \leq i \leq K)$.

The formulas for the covariances of $Y_i$ in Remark 1 follow by the definition of $Y_i$ and (3.27).

Corollary 1 follows by (2.14) and the continuous mapping theorem.

Corollary 2 follows by Corollary 1, Lemma 3, and the converging together theorem.

Corollary 3 follows by Theorem 1 and Lemma A2 since, in the notation of the theorem, the processes $\left((A^N(t) - \lambda^N t)/\sqrt{N}, t \geq 0\right)$ and $(X_0^N(t), t \geq 0)$ jointly converge in distribution to the processes $\left(\sum_{i=1}^K \sqrt{p_i \mu} W_i(t), t \geq 0\right)$ and $(X_0(t), t \geq 0)$, which is implied by Remark 7, (3.42) and (2.14).

## 4 Non Preemptive Priorities

In this section we prove Theorem 2.

**Proof of Theorem 2.** The proof mostly repeats the proof of Theorem 1. We only mention the modifications that are required in the above proof. Let $Q_{H,0}^N(t)$ and $Q_{L,0}^N(t)$ denote the number of high and low priority customers in the queue at time $t$, respectively.

We introduce the analogs of the processes $D_j^N$, $D^N$ and $B^N$ from (3.1), (3.2) and (3.3) by

$$\begin{aligned}
D_{H,j}^N(t) &= \sum_{l=1}^N \int_0^t 1\left(\hat{Q}_j^N(s-) \geq l\right) 1\left(Q_{H,0}^N(s-) > 0\right) dS_{j0}^l(s)\,, 1 \leq j \leq K\,,\\
D_{L,j}^N(t) &= \sum_{l=1}^N \int_0^t 1\left(\hat{Q}_j^N(s-) \geq l\right) 1\left(Q_{H,0}^N(s-) = 0\right) 1\left(Q_{L,0}^N(s-) > 0\right) dS_{j0}^l(s)\,, 1 \leq j \leq K\,,\\
D_H^N(t) &= \sum_{j=1}^K D_{H,j}^N(t)\,, D_L^N(t) = \sum_{j=1}^K D_{L,j}^N(t)\,,\\
B_H^N(t) &= D_H^N(t) + \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s-) < N\right) dA_H^N(s),\\
B_L^N(t) &= D_L^N(t) + \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s-) < N\right) dA_L^N(s).
\end{aligned}$$

We define the $\sigma$-algebra $\mathcal{F}^N(t)$ by

$$\begin{aligned}
\mathcal{F}^N(t) = \sigma\{&Q_{H,0}^N(0),\ Q_{L,0}^N(0),\ \hat{Q}_i^N(0),\ A_H^N(s),\ A_L^N(s),\ S_{ij}^l(s),\ \alpha_{H,B_H^N(s)},\ \alpha_{L,B_L^N(s)};\\
&1 \leq i \leq K,\ 0 \leq j \leq K,\ l \geq 1,\ 0 \leq s \leq t\} \vee \mathcal{N}
\end{aligned}$$

where $\mathcal{N}$ denotes the family of $P$-null sets, and introduce the filtration $\mathbb{F}^N = (\mathcal{F}^N(t),\ t \geq 0)$.

Analogs of equations (3.4) and (3.5) look as follows:

$$\begin{aligned}
Q_{H,0}^N(t) &= Q_{H,0}^N(0) + \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s-) = N\right) dA_H^N(s) - D_H^N(t)\,, \\
Q_{L,0}^N(t) &= Q_{L,0}^N(0) + \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s-) = N\right) dA_L^N(s) - D_L^N(t)\,, \\
\hat{Q}_i^N(t) &= \hat{Q}_i^N(0) + \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s-) < N\right) 1(\alpha_{H,B_H^N(s)} = i) dA_H^N(s) \\
&+ \sum_{j=1}^K \int_0^t 1\left(\alpha_{H,B_H^N(s)} = i\right) dD_{H,j}^N(s) + \sum_{l=1}^N \sum_{j=1,j\neq i}^K \int_0^t 1\left(\hat{Q}_j^N(s-) \geq l\right) dS_{ji}^l(s) \\
&- \sum_{l=1}^N \sum_{j=0,j\neq i}^K \int_0^t 1\left(\hat{Q}_i^N(s-) \geq l\right) dS_{ij}^l(s)\ , 1 \leq i \leq K''\,, \\
\hat{Q}_i^N(t) &= \hat{Q}_i^N(0) + \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s-) < N\right) 1(\alpha_{L,B_L^N(s)} = i) dA_L^N(s) \\
&+ \sum_{j=1,j\neq i}^K \int_0^t 1\left(\alpha_{L,B_L^N(s)} = i\right) dD_{L,j}^N(s) \\
&+ \sum_{l=1}^N \sum_{j=1}^K \int_0^t 1\left(\hat{Q}_j^N(s-) \geq l\right) dS_{ji}^l(s) - \sum_{l=1}^N \sum_{j=0,j\neq i}^K \int_0^t 1\left(\hat{Q}_i^N(s-) \geq l\right) dS_{ij}^l(s)\,, \\
& K'' + 1 \leq i \leq K'\,, \\
\hat{Q}_i^N(t) &= \hat{Q}_i^N(0) + \sum_{l=1}^N \sum_{j=1,j\neq i}^K \int_0^t 1\left(\hat{Q}_j^N(s-) \geq l\right) dS_{ji}^l(s) \\
&- \sum_{l=1}^N \sum_{j=0,j\neq i}^K \int_0^t 1\left(\hat{Q}_i^N(s-) \geq l\right) dS_{ij}^l(s)\,, K' + 1 \leq i \leq K.
\end{aligned}$$

By an argument similar to the one used in the proof of Lemma 1 these equations allow us to derive the following analogs of (3.6) and (3.7)

$$\begin{aligned}
Q_{H,0}^N(t) &= Q_{H,0}^N(0) + \lambda_H^N \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s) = N\right) ds \\
&- \sum_{j=1}^K \mu_{j0} \int_0^t \hat{Q}_j^N(s) 1\left(Q_{H,0}^N(s) > 0\right) ds + M_{H,0}^N(t)\,, \qquad (4.1) \\
Q_{L,0}^N(t) &= Q_{L,0}^N(0) + \lambda_L^N \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s) = N\right) ds \\
&- \sum_{j=1}^K \mu_{j0} \int_0^t \hat{Q}_j^N(s) 1\left(Q_{L,0}^N(s) > 0\right) 1\left(Q_{H,0}^N(s) = 0\right) ds + M_{L,0}^N(t)\,,
\end{aligned}$$

$$
\begin{aligned}
\hat{Q}_i^N(t) \,=\, & \hat{Q}_i^N(0) + p_i\lambda^N \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s) < N\right) ds \\
& + p_{H,i}\sum_{j=1}^K \mu_{j0}\int_0^t \hat{Q}_j^N(s)1(Q_{H,0}^N(s) > 0)ds + \sum_{j=1, j\neq i}^K \mu_{ji}\int_0^t \hat{Q}_j^N(s)ds \\
& -\mu_i \int_0^t \hat{Q}_i^N(s)ds + \hat{M}_i^N(t),\ 1 \le i \le K'', \\
\hat{Q}_i^N(t) \,=\, & \hat{Q}_i^N(0) + p_i\lambda^N \int_0^t 1\left(\sum_{j=1}^K \hat{Q}_j^N(s) < N\right) ds \\
& + p_{L,i}\sum_{j=1}^K \mu_{j0}\int_0^t \hat{Q}_j^N(s)1(Q_{L,0}^N(s) > 0)1(Q_{H,0}^N(s) = 0)ds + \sum_{j=1, j\neq i}^K \mu_{ji}\int_0^t \hat{Q}_j^N(s)ds \\
& -\mu_i \int_0^t \hat{Q}_i^N(s)ds + \hat{M}_i^N(t),\ K''+1 \le i \le K', \\
\hat{Q}_i^N(t) \,=\, & \hat{Q}_i^N(0) + \sum_{j=1, j\neq i}^K \mu_{ji}\int_0^t \hat{Q}_j^N(s)ds - \mu_i\int_0^t \hat{Q}_i^N(s)ds + \hat{M}_i^N(t),\ K'+1 \le i \le K\,,
\end{aligned}
$$

where $M_{H,0}^N = (M_{H,0}^N(t), t \ge 0)$, $M_{L,0}^N = (M_{L,0}^N(t), t \ge 0)$ and $\hat{M}_i^N = (\hat{M}_i^N(t), t \ge 0), 1 \le i \le K$, are $\mathbb{F}^N$–locally square integrable martingales with respective predictable quadratic variations satisfying, for some $b' > 0$,

$$
\langle M_{H,0}^N\rangle(t) \le b'Nt, \quad \langle M_{L,0}^N\rangle(t) \le b'Nt, \quad \langle \hat{M}_i^N\rangle(t) \le b'Nt,\ 1 \le i \le K. \tag{4.2}
$$

The latter representations lead to equation (3.28), where the processes $\bar{X}_i^N = \big(\bar{X}_i^N(t),\ t \ge 0\big)$, and $\hat{X}_i^N = \Big(\hat{X}_i^N(t),\ t \ge 0\Big)$, $1 \le i \le K$, are still defined by (3.23) and (3.29), respectively, the difference being that processes $\bar{Q}_i^N = (\bar{Q}_i^N(t), t \ge 0)$ are defined as

$$
\begin{aligned}
\bar{Q}_i^N(t) \,&=\, p_{H,i}Q_{H,0}^N(t) + \hat{Q}_i^N(t),\ 1 \le i \le K''\,, && (4.3)\\
\bar{Q}_i^N(t) \,&=\, p_{L,i}Q_{L,0}^N(t) + \hat{Q}_i^N(t),\ K''+1 \le i \le K'\,, && (4.4)\\
\bar{Q}_i^N(t) \,&=\, \hat{Q}_i^N(t),\ K'+1 \le i \le K\,,
\end{aligned}
$$

and

$$
\begin{aligned}
\bar{M}_i^N(t) \,&=\, p_{H,i}M_{H,0}^N(t) + \hat{M}_i^N(t)\,, 1 \le i \le K'', \\
\bar{M}_i^N(t) \,&=\, p_{L,i}M_{L,0}^N(t) + \hat{M}_i^N(t)\,, K''+1 \le i \le K', \\
\bar{M}_i^N(t) \,&=\, \hat{M}_i^N(t)\,, K'+1 \le i \le K.
\end{aligned}
$$

We next prove the assertion of Lemma 2 by essentially the same argument as above except that in order to deduce (3.31) from (3.28) we have to use instead of (3.30) the inequalities

$$|\hat{X}_i^N(t)| \le |\bar{X}_i^N(t)| + \left( \sum_{j=1}^K \bar{X}_i^N(t) \right)^+ ,$$

which follow by (3.23) (3.29), (4.3) and (4.4).

The next step is to prove an analog of Lemma 3(ii). Let

$$X_{L,0}^N(t) = \frac{Q_{L,0}^N(t)}{\sqrt{N}}, \quad X_{H,0}^N(t) = \frac{Q_{H,0}^N(t)}{\sqrt{N}}.$$

Since, in analogy with (3.38),

$$\tilde{Q}_i^N(t) = \sum_{j=Q_{L,0}^N(0)+A_L^N(t)-Q_{L,0}^N(t)+1}^{Q_{L,0}^N(0)+A_L^N(t)} 1(\alpha_{L,j} = i), \ K''+1 \le i \le K', \ t \ge 0 \, ,$$

the same argument as in the proof of Lemma 3 shows that

$$\lim_{N\to\infty} P\left( \sup_{t\le T} \left| \tilde{X}_i^N(t) - p_{L,i} X_{L,0}^N(t) \right| > \epsilon \right) = 0, \ K''+1 \le i \le K' . \tag{4.5}$$

An analogous result holds for the high-priority customers as well, but we do not need it because of the following lemma.

**Lemma 4:** *For every $T > 0$, as $N \to \infty$,*

$$\sup_{t\le T} X_{H,0}^N(t) \xrightarrow{P} 0.$$

**Proof.** Let us denote

$$\begin{aligned} Z^N(t) &= Q_{H,0}^N(0) + \lambda_H^N \int_0^t 1\left( \sum_{j=1}^K \hat{Q}_j^N(s) = N \right) ds \\ &\quad - \sum_{j=1}^K \mu_{j0} \int_0^t \hat{Q}_j^N(s) ds + M_{H,0}^N(t), \\ \tilde{Z}^N(t) &= Q_{H,0}^N(0) + \lambda_H^N t - \sum_{j=1}^K \mu_{j0} \int_0^t \hat{Q}_j^N(s) ds + M_{H,0}^N(t), \\ \bar{Z}^N(t) &= Q_{H,0}^N(0) - \sum_{j=1}^K \mu_{j0} \int_0^t (\hat{Q}_j^N(s) - N q_j) ds + M_{H,0}^N(t). \end{aligned}$$

From (4.1) we can see that $Q_{H,0}^N$ is the Skorohod reflection of the process $(Z^N(t), t \ge 0)$. Define the process $\tilde{Q}_{H,0}^N$ as the Skorohod reflection of the process $(\tilde{Z}^N(t), t \ge 0)$. Since the process $(\tilde{Z}^N(t) - Z^N(t), t \ge 0)$ is nondecreasing, it follows that $Q^N_{H,0}(t) \le \tilde{Q}^N_{H,0}(t)$. Therefore, it is enough to prove that

$$\sup_{t \le T} \frac{\tilde{Q}^N_{H,0}(t)}{\sqrt{N}} \xrightarrow{P} 0. \tag{4.6}$$

Since

$$\frac{\bar{Z}^N(t)}{\sqrt{N}} = X^N_{H,0}(0) - \sum_{j=1}^{K} \mu_{j0} \int_0^t \hat{X}^N_j(s)ds + \frac{M^N_{H,0}(t)}{\sqrt{N}},$$

it follows from (4.2) with the use of the Lenglart-Rebolledo inequality and the analog of Lemma 2(ii) that the sequence of processes $(\bar{Z}^N(t)/\sqrt{N}, t \ge 0)$ is relatively compact in distribution. So, by taking a subsequence, if necessary, and using the principle of the common probability space we can assume that, for some continuous process $(Z_1(t), t \ge 0)$, as $N \to \infty$, $\bar{Z}^N(t)/\sqrt{N} \to Z_1(t)$ $P$-a.s. uniformly in $t$ over bounded intervals. Also, by hypotheses, $\lim_{N\to\infty} \lambda^N_H/N = \mu \sum_{i=1}^{K''} p_i < \mu = \sum_{j=1}^{K} q_j \mu_{j0}$, so that $(\lambda^N_H - N\sum_{j=1}^{K} q_j \mu_{j0})/\sqrt{N} \to -\infty$. Since

$$\frac{\tilde{Z}^N(t)}{\sqrt{N}} = \frac{\bar{Z}^N(t)}{\sqrt{N}} + \frac{\lambda^N_H - N\sum_{j=1}^{K} q_j \mu_{j0}}{\sqrt{N}}\, t$$

and the process $(\tilde{Q}^N_{H,0}(t)/\sqrt{N}, t \ge 0)$ is the Skorohod reflection of the process $(\tilde{Z}^N_{H,0}(t)/\sqrt{N}, t \ge 0)$, convergence (4.6) follows by Theorem 6.4(iii) in Whitt [17]. □

The rest of the proof of Theorem 2 is analogous to the proof of Theorem 1. The $\sigma$-algebra $\hat{\mathcal{F}}^N(t)$ is defined as

$$\begin{aligned} \hat{\mathcal{F}}^N(t) = \sigma\{&Q^N_{H,0}(0),\ Q^N_{L,0}(0),\ \hat{Q}^N_i(0),\ A^N_H(s),\ A^N_L(s),\ S^l_{ij}(s),\\ &\left(\alpha_{H,1}, \dots, \alpha_{A^N_H(t)+Q^N_{H,0}(0)}\right),\ \left(\alpha_{L,1}, \dots, \alpha_{A^N_L(t)+Q^N_{L,0}(0)}\right);\\ &1 \le i \le K,\ 0 \le j \le K,\ l \ge 1,\ 0 \le s \le t\} \vee \mathcal{N}, \end{aligned}$$

and $\hat{\mathbb{F}}^N = (\hat{\mathcal{F}}^N(t),\ t \ge 0)$. We observe that the introduced quantities satisfy equations (3.47) and (3.46). By the same argument as in the proof of Theorem 1, the processes $\{\tilde{M}^N_i, 1 \le i \le K\}$ converge jointly in distribution to the processes $\{Y_i, 1 \le i \le K\}$. The next step is to use (2.14), (4.3), (4.4), (4.5) and Lemma 4 to substitute into (3.47)

$$\begin{aligned} \hat{X}^N_i(t) &= X^N_i(t) + \hat{\epsilon}^N_i(t), 1 \le i \le K'',\\ \hat{X}^N_i(t) &= X^N_i(t) - p_{L,i}\left(\sum_{j=1}^{K} X^N_j(t)\right)^+ + \hat{\epsilon}^N_i(t), K''+1 \le i \le K',\\ \hat{X}^N_i(t) &= X^N_i(t), K'+1 \le i \le K, \end{aligned}$$

where

$$\sup_{t\le T}\left|\hat{\epsilon}_i^N(t)\right| \xrightarrow{P} 0,\ 1\le i\le K'.$$

We end the proof by applying the continuous mapping theorem. □

## 5 Renewal arrivals

We indicate here how the proof of Theorem 1 is extended to the context of renewal arrivals. Analogous arguments extend the proof of Theorem 2. The plan of the proof is the same: we first need to prove the assertion of Lemma 3 and then use a representation in the theme of (3.47). The techniques used are also similar. For example, we still need two different filtrations when proving Lemma 3 and the theorem itself. However, we no longer have decomposition (3.8) for the arrival process, so certain additional tools are required.

Let $\mathcal{G}_k^N$ be the $\sigma$-algebra generated by the random variables $\xi_1^N,\dots,\xi_k^N$. Then $A^N(t)+1$ is a stopping time with respect to the flow $G^N=(\mathcal{G}_{\lfloor t\rfloor}^N,\ t\ge 0)$ and the $\sigma$-algebra $\mathcal{G}_{A^N(t)+1}^N$ is well defined [5]. In contrast with the proof of Theorem 1, we define the $\sigma$-algebra $\mathcal{F}^N(t)$ as

$$\mathcal{F}^N(t)=\mathcal{G}^N_{A^N(t)+1}\vee\sigma\{Q_0^N(0),\ \hat{Q}_i^N(0),\ S_{ij}^l(s),\ B^N(s),\alpha_{B^N(s)};\quad 1\le i\le K,\ 0\le j\le K,\ l\ge 1,\ 0\le s\le t\}\vee\mathcal{N}$$

and introduce the filtration $\mathbb{F}^N=(\mathcal{F}^N(t),t\ge 0)$. The process $A^N$ is $\mathbb{F}^N$-predictable [5].

Our first goal is to prove the assertion of Lemma 2. Let processes $\bar{X}_i^N=(\bar{X}_i^N(t),t\ge 0)$, $1\le i\le N$, be defined by (3.23) and (3.22). As in the proof of Lemma 2, it suffices to prove (3.37). Algebraic manipulations similar to those that led to (3.28) lead to the equation

$$\bar{X}_i^N(t) = \bar{X}_i^N(0)+p_i\frac{A^N(t)-\lambda^N t}{\sqrt{N}}-p_i\mu\beta t+\sum_{j=1,j\ne i}^K \mu_{ji}\int_0^t \hat{X}_j^N(s)ds -\mu_i\int_0^t \hat{X}_i^N(s)ds+\frac{1}{\sqrt{N}}\bar{M}_i^N(t)\,, \tag{5.1}$$

where $\hat{X}_i^N(t)$ are defined by (3.29),

$$\bar{M}_i^N(t) = \bar{M}_{A,i}^N(t)+\sum_{j=1}^K M_{D,ji}^N(t)+\sum_{j=1,j\ne i}^K M_{S,ji}^N(t)-\sum_{j=0,j\ne i}^K M_{S,ij}^N(t) \tag{5.2}$$

$$-p_i\sum_{j=1}^K\sum_{k=1}^K M_{D,jk}^N(t), \tag{5.3}$$

$M^N_{D,ji}(t)$ and $M^N_{S,ji}(t)$ are defined by (3.12) and (3.16), respectively, and

$$\bar{M}^N_{A,i}(t) = \int_0^t 1\left(\sum_{j=1}^K \hat{Q}^N_j(s-) < N\right)\left(1(\alpha_{B^N(s)} = i) - p_i\right) dA^N(s). \tag{5.4}$$

In analogy with the derivation of (3.32), we obtain from (5.1)

$$\sup_{t\le T}\sum_{i=1}^K |\bar{X}^N_i(t)| \le \left(\sum_{i=1}^K |\bar{X}^N_i(0)| + \sup_{t\le T}\left|\frac{A^N(t) - \lambda^N t}{\sqrt{N}}\right| + |\mu\beta|T + \frac{1}{\sqrt{N}}\sum_{i=1}^K \sup_{t\le T} |\bar{M}^N_i(t)|\right)\cdot e^{C'T}. \tag{5.5}$$

Therefore, (3.37) would follow by (3.36) and

$$\lim_{A\to\infty}\overline{\lim_{N\to\infty}}\, P\left(\sup_{t\le T}\frac{1}{\sqrt{N}}|A^N(t) - \lambda^N t| > A\right) = 0.$$

The latter holds by the fact that the $A^N$ satisfy a functional central limit theorem (see, e.g., [1]). Limit (3.36) is proved as in the proof of Lemma 2 by using the Lenglart-Rebolledo inequality (3.34) except that we do not have the bound (3.35). However, since by (5.4) and Lemma A1 (recall that as in the proof of Theorem 1 we can replace subscript $B^N(s)$ in the integral on the right of (5.4) by $A^N(s) + Q^N(0)$ and $A^N$ is $\mathbb{F}$-predictable),

$$\langle \bar{M}^N_{A,i}\rangle(t) = p_i(1-p_i)\int_0^t 1\left(\sum_{j=1}^K \hat{Q}^N_j(s-) < N\right) dA^N(s) \le A^N(t),$$

the hypotheses of Theorem 3 easily imply that

$$\frac{A^N(t)}{N} \xrightarrow{\mathrm{P}} \mu t\,, \tag{5.6}$$

and by (3.14) and (3.18), for some $r' > 0$,

$$\langle M^N_{D,ji}\rangle(t) \le r'Nt, \quad \langle M^N_{S,ji}\rangle(t) \le r'Nt,$$

it follows by (5.2) that

$$\lim_{A\to\infty}\overline{\lim_{N\to\infty}}\, P\left(\frac{1}{N}\langle \bar{M}^N_i\rangle(t) > A\right) = 0.$$

Though the latter is weaker than (3.35), it is enough to deduce (3.36) from (3.34). This completes the proof of (3.37) and Lemma 2. Lemma 3 follows from Lemma 2 by the same argument as above.

In the rest of the proof we use the new $\sigma$-algebra

$$\hat{\mathcal{F}}^N(t) = \mathcal{G}^N_{A^N(t)+1} \vee \sigma\{Q_0^N(0), \hat{Q}_i^N(0), S^l_{ij}(s), \alpha_1, \dots, \alpha_{A^N(t)+Q_0^N(0)};$$
$$1 \le i \le K;\ 0 \le j \le K,\ l \ge 1,\ 0 \le s \le t\} \vee \mathcal{N},$$

and filtration $\hat{\mathbb{F}}^N = \{\hat{\mathcal{F}}^N(t),\ t \ge 0\}$. Defining $A_i^N(t)$ as in (3.41) we let

$$M^{N,i}(t) = A_i^N(t) - p_i A^N(t)\,. \tag{5.7}$$

Representation (3.44) allows us to apply the second part of Lemma A1 to conclude that the $M^{N,i} = (M^{N,i}(t),\ t \ge 0)$, $1 \le i \le K$ are $\hat{\mathbb{F}}^N$-locally square-integrable martingales with predictable quadratic covariations

$$\langle M^{N,i}\rangle(t) \,=\, p_i(1-p_i)A^N(t). \tag{5.8}$$

Since by the same lemma $\langle M^{N,i} + M^{N,i'}\rangle(t) = (p_i + p_{i'})(1 - p_i - p_{i'})A^N(t)$ and $\langle M^{N,i}, M^{N,i'}\rangle(t) = 1/2(\langle M^{N,i} + M^{N,i'}\rangle(t) - \langle M^{N,i}\rangle(t) - \langle M^{N,i'}\rangle(t))$, we have that

$$\langle M^{N,i}, M^{N,i'}\rangle(t) = -p_i p_{i'} A^N(t), \quad i \ne i'. \tag{5.9}$$

Let

$$M^N(t) = \sum_{j=2}^{A^N(t)+1} (1 - \lambda^N \xi_j^N).$$

Since we can equivalently write

$$M^N(t) = \int_0^t (1 - \lambda^N \xi^N_{A^N(s)+1})\, dA^N(s),$$

by Lemma A1 the process $M^N = (M^N(t), t \ge 0)$ is an $\mathbb{F}^N$-locally square-integrable martingales with predictable quadratic variation

$$\langle M^N\rangle(t) = (\lambda^N)^2(\sigma^N)^2 A^N(t). \tag{5.10}$$

(Another proof of this result can be found in [5]).

In analogy with the derivation of the predictible covariation of $M^{N,i}$ and $M^{N,i'}$, it is easy to see that

$$\langle M^N, M^{N,i}\rangle(t) = 0. \tag{5.11}$$

Therefore, we can write

$$A_i^N(t) = p_i \lambda^N t + M_i^N(t) + \gamma_i^N(t),\ 1 \le i \le K\,, \tag{5.12}$$

where $M_i^N = (M_i^N(t), t \geq 0)$ is the $\hat{\mathbb{F}}^N$-locally square-integrable martingale given by

$$M_i^N(t) = M^{N,i}(t) + p_i M^N(t) \tag{5.13}$$

and

$$\gamma_i^N(t) = p_i \lambda^N \left( \sum_{j=2}^{A^N(t)+1} \xi_j^N - t \right) . \tag{5.14}$$

We note that since $0 < \sum_{j=1}^{A^N(t)+1} \xi_j^N - t \leq \xi_{A^N(t)+1}^N$, $\lambda^N/N \to \mu$, (2.18) and (5.6) hold,

$$\sup_{t \leq T} \frac{|\gamma_i^N(t)|}{\sqrt{N}} \xrightarrow{P} 0. \tag{5.15}$$

By (5.13), (5.8), (5.9), (5.10), and (5.11), the predictable covariations of the $M_i^N$ are

$$\langle M_i^N \rangle(t) = \left(p_i(1-p_i) + p_i^2(\lambda^N)^2(\sigma^N)^2\right) A^N(t) , \tag{5.16}$$

and

$$\langle M_i^N, M_{i'}^N \rangle(t) = p_i p_{i'} \left((\lambda^N)^2(\sigma^N)^2 - 1\right) A^N(t) . \tag{5.17}$$

With representation (5.12) playing the role of (3.42) in the proof of Theorem 1, we can now write, in analogy with (3.47),

$$\begin{aligned} X_i^N(t) \; &= \; X_i^N(0) - p_i \mu \beta t + \sum_{j=1, j \neq i}^{K} \mu_{ji} \int_0^t \hat{X}_j^N(s) ds \\ &- \mu_i \int_0^t \hat{X}_i^N(s) ds + \frac{\tilde{M}_i^N(t)}{\sqrt{N}} + \frac{\gamma_i^N(t)}{\sqrt{N}}, \; 1 \leq i \leq K , \end{aligned} \tag{5.18}$$

where $\tilde{M}_i^N(t)$ are defined by (3.46), using $M_i^N(t)$ from (5.13).

Using (5.16), and (5.17) we obtain

$$\frac{\langle M_i^N \rangle(t)}{N} \xrightarrow{P} \left(p_i(1-p_i) + p_i^2 \mu^2 \sigma^2\right) \mu t \, , \; \frac{\langle M_i^N, M_{i'}^N \rangle}{N}(t) \xrightarrow{P} p_i p_{i'}(\mu^2 \sigma^2 - 1)\mu t. \tag{5.19}$$

As in the proof of Theorem 1, in view of (5.19), the processes $\{M_i^N/\sqrt{N},\ M_{S,ij}^N/\sqrt{N},\ i = 1, \ldots, K, j = 0, \ldots, K, j \neq i\}$ converge jointly in distribution to the processes $\{V_i, \sqrt{q_i \mu_{ij}} W_{i,j}$, $i = 1, \ldots, K, j = 0, \ldots, K, j \neq i\}$ so that by the continuous mapping theorem the processes $\{\tilde{M}_i^N/\sqrt{N}, 1 \leq i \leq K\}$ converge jointly in distribution to the processes $\{Y_i, 1 \leq i \leq K\}$. In view of (5.15) and the assertion of Lemma 3, the proof is completed by the same argument as in Theorem 1.

## Appendix

In this appendix we state and prove a lemma about thinnings of compound point processes and a lemma about deriving waiting-time asymptotics from queue-length asymptotics.

**Lemma A1:** *Let $A = (A(t), t \geq 0)$ and $D = (D(t), t \geq 0)$ be point processes on $(\Omega, \mathcal{F}, P)$ adapted to filtration $\mathbb{F} = (\mathcal{F}(t), t \geq 0)$ such that jumps of $A$ occur only at the times of jumps of $D$. Let $\beta_1, \beta_2, \ldots$ be identically distributed, nonnegative and integrable random variables such that, for every $t > 0$, the random variables $\beta_1, \ldots, \beta_{D(t)}$ are $\mathcal{F}(t)$–measurable, and the random variables $\beta_{D(t)+1}, \beta_{D(t)+2}, \ldots$ are independent of $\mathcal{F}(t)$.*

*Let $\tilde{A} = (\tilde{A}(t), t \geq 0)$ denote the $\mathbb{F}$–compensator of $A$ and $\tau_i$ denote the times of jumps of $A$. Then the process $B = (B(t), t \geq 0)$, defined by*

$$B(t) = \int_0^t \beta_{D(s)}\, dA(s) := \sum_{i=1}^{A(t)} \beta_{D(\tau_i)},$$

*is $\mathbb{F}$-adapted, and has $\mathbb{F}$-compensator $\tilde{B} = (\tilde{B}(t), t \geq 0)$ of the form*

$$\tilde{B}(t) = E\beta_1\, \tilde{A}(t).$$

*If, in addition, $E\beta_1^2 < \infty$, then the process $L = (L(t), t \geq 0)$, defined by*

$$L(t) = \int_0^t (\beta_{D(s)} - E\beta_1)\, dA(s),$$

*is an $\mathbb{F}$-locally square integrable martingale with the predictable quadratic variation process*

$$\langle L \rangle(t) = Var\beta_1\, \tilde{A}(t).$$

**Proof.** The fact that $B$ is $\mathbb{F}$–adapted easily follows from the hypotheses. To derive its $\mathbb{F}$–compensator, let us first introduce filtration $\mathbb{H} = (\mathcal{H}(t), t \geq 0)$ by $\mathcal{H}(t) = \mathcal{F}(t) \vee \mathcal{G}(t)$, where $\mathcal{G}(t)$ is the $\sigma$-algebra generated by $\beta_{D(t)+1}, \beta_{D(t)+2}, \ldots$, and prove that $B$ has the $\mathbb{H}$–compensator

$$B'(t) = \int_0^t \beta_{D(s-)+1}\, d\tilde{A}(s). \tag{A.1}$$

The process $(\beta_{D(s-)+1}, s \geq 0)$ is $\mathbb{H}$–adapted and left-continuous, hence, it is $\mathbb{H}$–predictable. Since the $\sigma$–algebra $\mathcal{G}(t)$ is independent of $\mathcal{F}(t)$, the process $\tilde{A}$, being the $\mathbb{F}$–compensator of $A$, is also the $\mathbb{H}$–compensator of $A$. Also, since $D$ jumps when $A$ does, we can rewrite $B(t)$ as

$$B(t) = \int_0^t \beta_{D(s-)+1}\, dA(s).$$

Now the claim follows from properties of stochastic integrals (see, e.g., Liptser and Shiryaev [13]).

Since $\mathcal{F}(t) \subset \mathcal{H}(t)$ and the process $B$ is $\mathbb{F}$-adapted, its $\mathbb{F}$–compensator is the same as the $\mathbb{F}$–compensator of $B'$. Since $\tilde{B}$ is obviously $\mathbb{F}$–predictable and nondecreasing, to prove that $\tilde{B}$ is the $\mathbb{F}$–compensator of $B'$, we have to check that $B' - \tilde{B}$ is an $\mathbb{F}$–local martingale. Let $\tau_n^D$ denote the times of jumps of $D$ and let stopping times $\tau_n'$ be defined by $\tau_n' = \tau_n^D$ if $\lim_{t\to\infty} D(t) \geq n$ and $\tau_n' = \infty$ if $\lim_{t\to\infty} D(t) < n$. Since $\tau_n' \uparrow \infty$ as $n \to \infty$, we can take this sequence as a localising sequence; so, it is enough to check that the processes $(B'(t \wedge \tau_m') - \tilde{B}(t \wedge \tau_m'), t \geq 0), m = 1, 2, \ldots$ are $\mathbb{F}$–uniformly integrable martingales, which, in view of (A.1), the choice of $\tau_m'$ and the definition of $\tilde{B}$, is equivalent to checking that, for any $\mathbb{F}$–stopping time $\tau$,

$$E \int_0^{\tau\wedge\tau_m'} \beta_{D(s-)+1}\, d\tilde{A}(s) = E \int_0^{\tau\wedge\tau_m'} E\beta_1\, d\tilde{A}(s).$$

The argument of the proof of Lemma 3.12 in Jacod and Shiryaev [12] shows, in view of $\mathbb{F}$-predictability of $\tilde{A}$, that the above equality holds if for every $\mathbb{F}$-predictable stopping time $\sigma$

$$E\left[\beta_{D(\sigma-)+1}\, 1(\sigma \leq \tau \wedge \tau_m')1(\sigma < \infty)\right] = E\left[E\beta_1 1(\sigma \leq \tau \wedge \tau_m')1(\sigma < \infty)\right]. \tag{A.2}$$

Since $\{\sigma = 0\} \in \mathcal{F}_0$ and $\beta_1$ is independent of $\mathcal{F}_0$, $E[\beta_1 1(\sigma = 0)] = E[E\beta_1 1(\sigma = 0)]$ so that, by the monotone convergence theorem, (A.2) would follow if for every $N = 1, 2, \ldots$

$$E\left[\beta_{D(\sigma-)+1} 1(0 < \sigma \leq \tau \wedge \tau_m' \wedge N)\right] = E\left[E\beta_1 1(0 < \sigma \leq \tau \wedge \tau_m' \wedge N)\right].$$

Since $\sigma$ is a predictable stopping time, there exists a sequence $\{\sigma_n\}$ of $\mathbb{F}$-stopping times, which a.s. monotonically converges to $\sigma$ and is such that $\sigma_n < \sigma$ a.s. on the set $\{\sigma > 0\}$. Therefore, by the dominated convergence theorem the above equality would follow from

$$E\left[\beta_{D(\sigma_n)+1} 1(\sigma_n \wedge N < \tau \wedge \tau_m' \wedge N)\right] = E\left[E\beta_1 1(\sigma_n \wedge N < \tau \wedge \tau_m' \wedge N)\right]. \tag{A.3}$$

We prove (A.3) by approximating the $\sigma_n$ by piecewise constant stopping times. Let, for $k = 1, 2, \ldots$,

$$\sigma_{n,k} = \sum_{i=1}^{kN} \frac{i}{k}\, 1\Big(\frac{i-1}{k} < \sigma_n \leq \frac{i}{k}\Big).$$

Then the $\sigma_{n,k}$ are $\mathbb{F}$-stopping times and

$$\begin{aligned} &E\left[\beta_{D(\sigma_{n,k})+1} 1(\sigma_{n,k} < \tau \wedge \tau_m' \wedge N)\right] \\ &= \sum_{i=1}^{kN} E\left[\beta_{D(i/k)+1}\, 1\Big(\frac{i-1}{k} < \sigma_n \leq \frac{i}{k}\Big)\, 1\Big(\frac{i}{k} < \tau \wedge \tau_m' \wedge N\Big)\right]. \end{aligned}$$

Next, since $1((i-1)/k < \sigma_n \le i/k) 1(i/k < \tau \wedge \tau'_m \wedge N)$ is $\mathcal{F}(i/k)$-measurable by the fact that $\sigma_n, \tau$ and $\tau'_m$ are $\mathbb{F}$-stopping times,

$$\begin{aligned}
&E\left[\beta_{D(i/k)+1}\, 1\left(\frac{i-1}{k} < \sigma_n \le \frac{i}{k}\right) 1\left(\frac{i}{k} < \tau \wedge \tau'_m \wedge N\right)\right] \\
&= E\left[1\left(\frac{i-1}{k} < \sigma_n \le \frac{i}{k}\right) 1\left(\frac{i}{k} < \tau \wedge \tau'_m \wedge N\right) E(\beta_{D(i/k)+1} | \mathcal{F}(i/k))\right] \\
&= E\left[1\left(\frac{i-1}{k} < \sigma_n \le \frac{i}{k}\right) 1\left(\frac{i}{k} < \tau \wedge \tau'_m \wedge N\right)\right] E\beta_1,
\end{aligned}$$

where the latter equality holds since $\beta_{D(i/k)+1}$ and $\mathcal{F}(i/k)$ are independent. Thus,

$$\begin{aligned}
&E\left[\beta_{D(\sigma_{n,k})+1} 1(\sigma_{n,k} < \tau \wedge \tau'_m \wedge N)\right] \\
&= \sum_{i=1}^{kN} E\left[1\left(\frac{i-1}{k} < \sigma_n \le \frac{i}{k}\right) 1\left(\frac{i}{k} < \tau \wedge \tau'_m \wedge N\right)\right] E\beta_1 \\
&= P\left(\sigma_{n,k} < \tau \wedge \tau'_m \wedge N\right) E\beta_1.
\end{aligned}$$

Since $\sigma_{n,k} \downarrow \sigma_n \wedge N$ as $k \to \infty$, it follows by the dominated convergence theorem that

$$E\left[\beta_{D(\sigma_n)+1} 1(\sigma_n \wedge N < \tau \wedge \tau'_m \wedge N)\right] = P\left(\sigma_n \wedge N < \tau \wedge \tau'_m \wedge N\right) E\beta_1,$$

proving (A.3) and, hence, (A.2). The first part of the lemma is proved.

The fact that $L$ is an $\mathbb{F}$-local martingale follows by the part we have just proved. Next, by Ito's formula, the definition of $L$ and the fact that jumps of $A$ are of size 1

$$L(t)^2 = 2\int_0^t L(s-)\, dL(s) + \sum_{0<s\le t} (\Delta L(s))^2 = 2\int_0^t L(s-)\, dL(s) + \int_0^t (\beta_{D(s)} - E\beta_1)^2\, dA(s).$$

The first integral on the rightmost side is an $\mathbb{F}$-local martingale; the $\mathbb{F}$-compensator of the second integral by the first part of the lemma is equal to $(E(\beta_1 - E\beta_1)^2 \tilde{A}(t), t \ge 0)$, which ends the proof. □

In the next lemma we consider a sequence of queueing systems with a single input stream and a single output stream of customers, and the FIFO service discipline. For the $N$th system, we denote by $Q^N(t)$ the queue length at $t$, by $w^N(t)$, the virtual waiting time at $t$, by $w_i^N$, the waiting time of the $i$th customer, by $A^N(t)$, the number of arrivals by $t$, and by $D^N(t)$, the number of departures by $t$. We introduce the processes $K^N = (K^N(t), t \ge 0)$ and $L^N = (L^N(t), t \ge 0)$ by $K^N(t) = Q^N(t)/\sqrt{N}$ and $L^N(t) = (A^N(t) - \lambda^N t)/\sqrt{N}$, where $\lambda^N$ are real numbers. We assume that $A^N(0) = D^N(0) = 0$.

**Lemma A2:** *If the processes* $(K^N, L^N)$ *converge jointly in distribution on* $D([0,\infty), R^2)$ *to processes* $(K, L)$, *where* $K = (K(t), t \geq 0)$ *and* $L = (L(t), t \geq 0)$, *and* $\lambda^N/N \to \lambda > 0$, *then the processes* $(\sqrt{N} w^N(t), t \geq 0)$ *and* $(\sqrt{N} w^N_{\lfloor Nt \rfloor + 1}, t \geq 0)$ *converge in distribution on* $D([0,\infty), R)$ *to the respective processes* $(K(t)/\lambda, t \geq 0)$ *and* $(K(t/\lambda)/\lambda, t \geq 0)$.

**Proof.** The proof follows the argument used in the Example considered in [16]. Since

$$Q^N(t) = Q^N(0) + A^N(t) - D^N(t),$$

the hypotheses of the lemma imply that the processes $(U^N, V^N)$, where $U^N = (U^N(t), t \geq 0)$ and $V^N = (V^N(t), t \geq 0)$ are defined by $U^N(t) = (D^N(t) - \lambda^N t)/\sqrt{N}$ and $V^N(t) = L^N(t) + K^N(0)$, converge in distribution in $D([0,\infty), R^2)$ to processes $(U, V)$, where $U = (U(t), t \geq 0)$, $V = (V(t), t \geq 0)$, $U(t) = L(t) + K(0) - K(t)$, and $V(t) = L(t) + K(0)$. Since

$$w^N(t) + t = \inf(s \geq 0 : D^N(s) \geq A^N(t) + Q^N(0)), \ \frac{D^N(t)}{N} \xrightarrow{P} \lambda t, \ \frac{Q^N(0) + A^N(t)}{N} \xrightarrow{P} \lambda t,$$

an application of the Corollary in [16] shows that $(\sqrt{N} w^N(t), t \geq 0)$ converges in distribution to $(K(t)/\lambda, t \geq 0)$. The second convergence follows by the equality $w^N_{\lfloor Nt \rfloor + 1} = w^N\left(\tau^N_{\lfloor Nt \rfloor + 1}\right)$, where $\tau^N_i$ is the arrival time of the $i$th customer, the fact that $\tau^N_{\lfloor Nt \rfloor + 1} \xrightarrow{P} t/\lambda$ and the random time change theorem. □

## References

[1] Billingsley, P. (1968), *Convergence of Probability Measures*, John Wiley & Sons, NY.

[2] Borovkov, A.A. (1967), On limit laws for service processes in multi–channel systems (in Russian), *Siberian Math. J.* **8**, pp. 746–763.

[3] Borovkov, A.A. (1980), *Asymptotic Methods in Queueing Theory*, Nauka, Moscow (in Russian). English translation: Wiley, 1984.

[4] Brémaud, P. (1981), *Point Processes and Queues. Martingale Dynamics*, Springer.

[5] Coffman, Jr., E. G., Puhalskii, A. A. and Reiman, M. I. (1991), Storage-Limited Queues in Heavy Traffic, *Prob. Engin. Inform. Sci.* **5**, pp. 499–522.

[6] Erlang, A. K. (1917), Solutions of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges, *Electroteknikeren* (Danish) **13**, pp. 5–13. English translation: *P.O. Elec. Eng. J.* **10**, 189–197, 1917–1918.

[7] Halfin, S. and Whitt, W. (1981), Heavy-Traffic Limits for Queues with Many Exponential Servers, *Oper. Res.* **29**, pp. 567–588.

[8] Iglehart, D.L. (1965), Limit diffusion approximations for the many server queue and the repairman problem, *J. Appl. Prob.* **2**, pp. 429–441.

[9] Iglehart, D.L. (1973), Weak convergence of compound stochastic processes, *Stoch. Proc. Appl.* **1**, pp. 11–31.

[10] Iglehart, D. L. and Whitt, W. (1970), Multiple Channel Queues in Heavy Traffic, II: Sequences, Networks, and Batches, *Adv. Appl. Probab.* **2**, pp. 355–364.

[11] Ikeda, N. and Watanabe, S. (1989), *Stochastic Differential Equations and Diffusion Processes*, $2^{\mathrm{nd}}$ Ed., North Holland, Amsterdam.

[12] Jacod, J. and Shiryaev, A. N. (1987), *Limit Theorems for Stochastic Processes*, Springer.

[13] Liptser, R. Sh. and Shiryaev, A. N. (1989), *Theory of Martingales*, Kluwer.

[14] Mandelbaum, A., Massey, W. A. and Reiman, M. (1998), Strong Approximations for Markovian Service Networks, *Queueing Systems: Theory and Applications* (*QUESTA*) **30**, pp. 149–201.

[15] Neuts, M. F. (1981), *Matrix-Geometric Solutions in Stochastic Models*, Johns Hopkins Univ. Press.

[16] Puhalskii, A. (1994), On the Invariance Principle for the First Passage Time, *Math. Oper. Res.* **19**, pp. 946–954.

[17] Whitt, W. (1980), Some Useful Functions for Functional Limit Theorems, *Math. Oper. Res.* **5**, pp. 67–85.

[18] Whitt, W. (1982), On the Heavy-traffic Limit Theorem for GI/G/$\infty$ Queues, *Adv. Appl. Probab.* **14**, pp. 171–190.