

An Overloaded Multiclass FIFO Queue with Abandonments

Otis B. Jennings

Fuqua School of Business, Duke University, Durham, NC 27708, otis.jennings@duke.edu

Josh E Reed

Leonard N. Stern School of Business, New York University, New York, NY 10012, jreed@stern.nyu.edu

In this paper, we consider a single-server queue fed by K independent renewal arrival streams, each representing a different job class. Jobs are processed in a FIFO fashion, regardless of class. The total amount of work arriving to the system exceeds the server's capacity. That is, the nominal traffic intensity of the system is assumed to be greater than one. Jobs arriving to the system grow impatient and abandon the queue after a random amount of time if service has not yet begun. Interarrival, service and abandonment times are assumed to be generally distributed and class specific.

We approximate this system using both fluid and diffusion limits. To this end, we consider a sequence of systems indexed by n in which the arrival and service rates are proportional to n ; the abandonment distribution remains fixed across the sequence. In our first main result, we show that in the limit as n tends to infinity, the virtual waiting time process for those jobs that do not abandon from the queue converges to a limiting deterministic process. This limit may be characterized as the solution to a first order ODE. Specific examples are then presented for which the ODE may be explicitly solved. In our second main result, we refine the deterministic fluid approximation by showing that the fluid centered and diffusion scaled virtual waiting time process weakly converges to an Ornstein-Uhlenbeck process whose drift and infinitesimal variance both vary over time. This process may also be solved for explicitly, thus yielding approximations to the transient as well as steady state behavior of the queue.

Key words: Queueing, Abandonment, Supercritical Loading, Overloaded Queue, Ornstein-Uhlenbeck Process, Diffusion Limit, Virtual Waiting Time Process

1. Introduction

Increasingly, queueing models have included abandonment in their formulations, and for good reason. From a modeling standpoint, abandonment is an undeniable reality in many industrial settings. Two examples of queues with abandonment are call centers and organ transplant waiting

lists. In the call center context, abandonment is straightforward. Individuals seeking to speak with an agent are placed on hold and may eventually lose patience, hanging up before the actual service can begin. See Brown et al. (4) for an extensive review of call centers and modeling complexities. For organ transplants, those for kidneys being a prime example, needy individuals are placed on a waiting list, remaining there until they reach the top of the list and a viable organ match is found. Abandonment in this case is either death or reaching a state of health for which transplantation is no longer an option. Overloaded queueing models in this context are particularly relevant, as the need for kidneys exceeds the supply of donors. For more on queueing models of organ transplant waiting lists, in particular for kidneys, see Su and Zenios (12, 13) and references therein.

The usual course of action taken when introducing abandonment into queueing models is to assume that a customer's patience – measured in how long one is willing to wait for service – is exponentially distributed. Analysis of models with exponential abandonment was first undertaken by Ancker and Gafarian (1). Although simplifying the analysis considerably, this assumption has been shown in many cases to be untenable. In this paper, we stray from the exponential abandonment assumption by allowing the jobs arriving to our system to have general abandonment distributions. In particular, we consider a single-server, multiclass first-in-first-out (FIFO) queue where each job class has a general abandonment distribution. Our main quantity of interest is the *virtual waiting time process*, which tracks the amount of time until all jobs currently in the queue depart, either due to a service completion or abandonment. One may think of the virtual waiting time as the amount of time that a hypothetical customer would wait before being processed, assuming she was sufficiently patient and eventually received service. By studying the virtual waiting time process, one is able to obtain key performance measures related to the operation of the system, such as the long run fraction of jobs who abandon or the average amount of time that jobs that eventually receive service will have waited.

Recently, others have begun to consider queueing systems with general abandonment distributions. In (?), Bacelli, Boyer and Hebuterne consider a $GI/GI/1 + GI$ queue and determine necessary and sufficient conditions for the system to be stable. Baccelli and Hebuterne (?) study

the $M/M/1 + GI$ queue and obtain closed form expressions for a variety of quantities of interest related to this system. More recently, authors have begun to provide an asymptotic analysis of queueing systems with general abandonment distributions in which the traffic intensity is close to one. Ward and Glynn (14) provide diffusion limits for the virtual waiting time process and the queue length process of the $GI/GI/1 + GI$ queue; these limits incorporate the density of the abandonment distribution at the origin. Building off of the work of Ward and Glynn and by scaling the hazard rate function of the abandonment distribution, Reed and Ward (11) provide diffusion limits for both the virtual waiting time process and the queue length process of the $GI/GI/1 + GI$ queue; these limits incorporate the entire abandonment distribution. In work similar to Ward and Glynn (14), but for the many-server “Halfin-Whitt” heavy-traffic regime, Zeltyn and Mandelbaum (10) provide an asymptotic analysis of the $M/M/N + GI$ queue which takes as its starting point the work of Bacelli and Hebuterne (?).

There is a growing body of literature studying multi-server queues in an overloaded regime in which the traffic intensity is strictly greater than one. This line of research was initiated by the work of Whitt (17) in which formal fluid limit approximations to the queue length process of the $M/M/N + GI$ in an overloaded regime were conjectured to hold and proven to be true in a discrete time setting. In subsequent work, the analysis conducted in (17) has been extended in several directions, see for instance (?) and (?). Although in this paper we concentrate predominately on the virtual waiting time process, we anticipate in future work using the analysis presented here as a stepping stone for the studying the queue length process as well, and perhaps verifying the conjecture originally set forth in (17).

As mentioned at the outset, in this paper we study the virtual waiting time process of the overloaded, multiclass, single-server queue. In the first of our two main results, Theorem 1, we provide a first order fluid limit for the virtual waiting time process which turns out to be the solution to a first order ODE. For certain instances, this ODE may be explicitly solved; some examples are provided for which this is the case. We also show that as t grows large the solution to the ODE converges monotonically to a stationary state w^* , a quantity dependent on the initial

conditions of the system. This result is consistent with other works in which the service capacity is significantly less than the offered load, such as Whitt (16), Garnett et al. (8) and de Véricourt and Jennings (?).

Our second main result, Theorem 2, shows that the stochastic fluctuations of the virtual waiting time process about its fluid limit may be well approximated by a time inhomogeneous Ornstein-Uhlenbeck process. The drift of this process at time t is equal to the negative of the process value multiplied by the traffic-intensity-weighted sum of the abandonment density functions, each evaluated at the value of the fluid limit at time t . Thus, our work is complementary to that of Ward and Glynn (14) and Zeltyn and Mandelbaum (?), who also obtain approximations in which the value of the abandonment density plays a role. Our result suggests that the steady state distribution of the virtual waiting time process may be well approximated by a normal distribution with mean w^* and a variance which may be explicitly calculated. This result is also complementary to the work of Ward and Glynn (14) who suggest using a truncated normal distribution as an approximation for the steady state distribution of the virtual waiting time process of the $GI/GI/1+GI$ queue in a critically loaded regime.

The remainder of this paper is organized as follows. In the following section, we present the basic model which is used for the remainder of the paper. In Section 3, we provide results on the long-run behavior of the virtual waiting time process as well the long-run fraction of time that the server allocates to each particular job class. In Section 4, we obtain fluid limit results for the virtual waiting time process. The main result in this section is Theorem 1, which provides a fluid limit for the virtual waiting time process in the overloaded regime described above. We then provide several examples for which our limiting process may be explicitly solved. In the following section, we center the virtual waiting time process by its deterministic fluid limit of Theorem 1 and then scale by an appropriate constant. Our main result in Section 5 is Theorem 2, which shows that the diffusion scaled virtual waiting time process converges to a time inhomogeneous Ornstein-Uhlenbeck process whose time varying drift and infinitesimal variance may be explicitly calculated. Finally, in Section 6 we provide closing remarks and directions for possible future research.

Unless stated otherwise, processes are assumed to be elements of $\mathcal{D}(\mathbb{R}_+, \mathbb{R})$, the space of right continuous left limit functions mapping \mathbb{R}_+ to \mathbb{R} , abbreviated \mathcal{D} . We assume the space \mathcal{D} is endowed with the usual Skorohod J_1 topology; see, e.g., (15). Let \Rightarrow denote convergence in distribution for both sequences of stochastic processes as well as sequences of random variables and \rightarrow^P denote convergence in probability. Convergence of processes is assumed to occur in \mathcal{D} .

The notation throughout is necessarily involved. When invoking the Skorohod Representation Theorem, our convention is to denote alternative probability spaces as $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$ and processes on these spaces similarly, e.g. \hat{X} . In several instances we define prelimit processes using the limiting virtual waiting time process W . Our convention is to denote such prelimit processes by apply the $\hat{\cdot}$ marker; e.g., \hat{M}^n . Similarly, we define some prelimit processes in terms of both the limiting virtual waiting time process and the sequence of waiting time processes W^n , employing the $\check{\cdot}$ marker, e.g., \check{M}^n . Lastly, we consider fluid scaled processes and diffusion scaled processes throughout. The former and their limits are designated by bars and the later and their limits a denoted by tildes; e.g., \bar{M}^n and \tilde{W}^n .

2. The model

In this section, we provide the model of the multiclass $GI/GI/1 + GI$ queue which will be used for the remainder of the paper. We begin by assuming that the multiclass queue is fed by K independent arrival streams, each of which is representative of a different job class. We index each job class by $k \in \{1, \dots, K\}$.

Letting $k \in \{1, \dots, K\}$, for each $i \geq 1$ we denote by $\xi_k(i)$ the time between the $(i-1)^{st}$ and i^{th} class k job arrivals to the system, where we assume that $\{\xi_k(i), i \geq 1\}$ is an i.i.d. sequence of random variables with mean λ_k^{-1} and variance $\sigma_{a,k}^2$. Thus, the total number of class k arrivals to the system by time $t \geq 0$ is given by

$$A_k(t) = \sup \left\{ i \geq 0 : \sum_{j=1}^i \xi_k(j) \leq t \right\}.$$

We also denote by

$$t_k(i) = \xi_k(1) + \dots + \xi_k(i), \tag{1}$$

the arrival time of the i^{th} class k job to the system. Notice the inverse relation: $A_k(t_k(i)) = i$.

The potential service time of the i^{th} class k job to arrive to the system is given by $\nu_k(i)$ where we assume that $\{\nu_k(i), i \geq 1\}$ is an i.i.d. sequence of positive random variables with mean μ_k^{-1} and variance $\sigma_{s,k}^2$. Note, however, that the service time $\nu_k(i)$ will not be realized if job i abandons the queue before its service can commence.

The quantity $d_k(i)$ represents the amount of time that the i^{th} class k job is willing to wait to receive service before abandoning from the system. We refer to this quantity as the *deadline* of the i^{th} class k job. Thus, if $d_k(i)$ time units pass before the servicing of job i can begin, then that job will abandon from the system. The cumulative distribution of $d_k(i)$ is denoted by F_k and its complement is given by $G_k = 1 - F_k$.

For each $k \in \{1, \dots, K\}$, the sequence of arrival, service and abandonment time random variables are each i.i.d. Moreover, these sequences are independent of each other and also independent between job classes. Finally, we assume that jobs are processed in a FIFO fashion, without regard to job class designation except possibly when two jobs arrive simultaneously, which we assume occurs only on a set of measure zero. This assumption is valid, for instance, if the first arrival time for each job class has a continuous distribution and is independent of all future arrivals.

In this paper, we assume that the queueing system is overloaded in the sense that the nominal load exceeds the system's capacity. Defining $\rho_k = \lambda_k / \mu_k$ to be the offered load of the k^{th} job class for $k = 1, \dots, K$, we assume that that overall offered load of the system is greater than one, that is

$$\rho = \sum_{k=1}^K \rho_k > 1. \quad (2)$$

Our primary performance measure of interest is the *virtual waiting time process*, $W = \{W(t), t \geq 0\}$, which tracks, as a function of time, the amount of time that a newly arriving job would have to wait before being served, if the job were sufficiently patient. Note that there are three main contributors to the dynamics of W . First, there is the initial value at time zero, denoted by $W(0)$. Next, as time progresses, if a job is being served, then the server works at rate 1, thereby decreasing the virtual waiting time at a unit rate. When no jobs are present, the server necessarily idles.

Finally, newly arriving jobs contribute to increases in W if they are to begin service before their deadline expires; i.e., they do not abandon. Jobs that eventually abandon do not contribute to the virtual waiting time.

Assuming that no two jobs arrive simultaneously, one may now express the virtual waiting time process as the unique solution to the equation

$$W(t) = W(0) + \sum_{k=1}^K \int_0^t \nu_k(A_k(s)) \cdot \mathbf{1}_{\{d_k(A_k(s)) > W(s-)\}} dA_k(s) - B(t), \quad t \geq 0, \quad (3)$$

where

$$B(t) = \int_0^t \mathbf{1}_{\{W(s) > 0\}} ds, \quad t \geq 0,$$

is the cumulative server busy time up to time t .

Equation (3) provides the fundamental equation for the virtual waiting time process. However, it is not in a form that is particularly well-suited to analysis. We therefore now provide an alternative representation of the virtual waiting time process which will facilitate proving our limit theorems in the sections that follow. For each $k \in \{1, \dots, K\}$, let

$$M_{\nu,k}(i) = \sum_{j=1}^i \left(\nu_k(j) - \frac{1}{\mu_k} \right) \mathbf{1}_{\{d_k(j) > W(t_k(j)-)\}}, \quad i \geq 0, \quad (4)$$

and

$$M_{d,k}(i) = \frac{1}{\mu_k} \sum_{j=1}^i \left(\mathbf{1}_{\{d_k(j) > W(t_k(j)-)\}} - G_k(W(t_k(j)-)) \right), \quad i \geq 0. \quad (5)$$

The process $M_{\nu,k} = (M_{\nu,k}(i), i \geq 1)$ tracks the centered service times of those jobs that do not abandon from the system, whereas $M_{d,k} = (M_{d,k}(i), i \geq 1)$ tracks the number of jobs that do not abandon centered by its expected value. Defining the filtration $\mathcal{F}^k = (\mathcal{F}_i^k, i \geq 1)$ by

$$\mathcal{F}_i^k = \sigma(\xi_j(l), \nu_j(l), d_j(l); l \geq 1, j \neq k; \xi_k(l), \nu_k(l), d_k(l), l = 1, \dots, i-1, \xi_k(i)), \quad i \geq 1,$$

it is clear from (4) and (5) that the processes $M_{\nu,k}$ and $M_{d,k}$ are martingales with respect to \mathcal{F}^k .

This fact will be useful later in the paper. In particular, it will allow us to take advantage of martingale convergence theorems when proving certain results.

By (3) we may express the virtual waiting time process as

$$W(t) = W(0) - t + I(t) + \sum_{k=1}^K \left[M_{\nu,k}(A_k(t)) + M_{d,k}(A_k(t)) + \frac{1}{\mu_k} \int_0^t G_k(W(s-)) dA_k(s) \right], \quad (6)$$

for $t \geq 0$, where

$$I(t) = t - B(t), \quad t \geq 0,$$

is the cumulative idle time of the server up until time t . We define $I = (I(t), t \geq 0)$ to be the idle time process.

Furthermore, for each $k \in \{1, \dots, K\}$, setting

$$\varepsilon_k(t) = \frac{1}{\mu_k} \left(\int_0^t G_k(W(s-)) dA_k(s) - \int_0^t G_k(W(s-)) d(\lambda_k s) \right), \quad t \geq 0, \quad (7)$$

it follows from (6) that

$$W(t) = W(0) - t + I(t) + \sum_{k=1}^K [M_{\nu,k}(A_k(t)) + M_{d,k}(A_k(t)) + \varepsilon_k^n(t)] + \int_0^t \sum_{k=1}^K \rho_k G_k(W(s)) ds, \quad t \geq 0, \quad (8)$$

where the final equality follows since

$$\frac{1}{\mu_k} \int_0^t G_k(W(s-)) d(\lambda_k s) = \int_0^t \rho_k G_k(W(s)) ds, \quad t \geq 0. \quad (9)$$

We next turn to analyzing the virtual waiting time process on a heuristic level.

3. Service allocation and virtual waiting time

In preparation for the results to follow in future sections, we now show that there exists a unique long run allocation of the server's effort amongst each particular job class. We begin with the following definition.

Definition 1 A vector $\alpha = (\alpha_1, \dots, \alpha_K)$ is called a feasible allocation if $\alpha_k \geq 0$ for each $k = 1, \dots, K$, and $\sum_k \alpha_k \leq 1$. If $\sum_k \alpha_k = 1$, then the allocation is called complete.

For each $k \in \{1, \dots, K\}$, one may interpret α_k as the long run proportion of time that the server dedicates to serving class k jobs. Hence, the qualifier “feasible” is superfluous in the sense that we do not consider would-be infeasible allocations in this paper. Also note that a “complete” allocation is representative of a fully utilized server.

Now suppose that the virtual waiting time has reached a limiting value w and remains constant thereafter. In such a scenario, it is clear that all jobs arriving to the queue will experience an identical virtual waiting time, namely w . Furthermore, the larger the value of w , the larger the fraction of jobs that will abandon from the system. In particular, for each delay $w \geq 0$, there is an associated *survival rate vector* $(\lambda_1 G_1(w), \dots, \lambda_K G_K(w))$ where, for each $k \in \{1, \dots, K\}$, $\lambda_k G_k(w)$ represents the long run rate of class k jobs arriving to the system that eventually enter service if the delay is always equal to w .

We now determine which values of w are achievable. Interestingly, one of the key factors determining a delay’s achievability is the existence of a feasible allocation by which the delay may be sustained in the long run. This in turn motivates the following definition.

Definition 2 For a given delay $w \geq 0$, a feasible allocation vector α is called

sufficient if $\alpha_k \mu_k \geq \lambda_k G_k(w)$ for each k ,

partially insufficient if $\alpha_k \mu_k < \lambda_k G_k(w)$ for some k ,

partially excessive if $\alpha_k \mu_k > \lambda_k G_k(w)$ for some k ,

critical if $\alpha_k \mu_k = \lambda_k G_k(w)$ for each k , and

perfect if it is critical and complete.

Making use of Definition 2, the following proposition now shows that there exists a unique range of virtual waiting times $[w_l^*, w_u^*]$ which may be sustained by a perfect allocation. Before stating this proposition, however, we need the following assumption.

Assumption (A1) For each job class $k \in \{1, \dots, K\}$, the deadline distribution F_k is continuous. Furthermore, $F_k = 0$ and $\lim_{x \rightarrow \infty} F_k(x) = 1$.

We may now state Proposition 1.

Proposition 1 *Under assumption (A1), there exists a unique perfect allocation α^* . Moreover, there exists a unique critical delay range quantity $[w_l^*, w_u^*]$, where $w_l^* \leq w_u^*$, associated with this perfect allocation such that*

- *if the delay quantity is less than w_l^* , then any feasible allocation will be partially insufficient,*
- and*
- *if the delay quantity exceeds w_u^* , then any complete allocation will be partially excessive.*

Furthermore, if for each $k \in \{1, \dots, K\}$ the abandonment distribution F_k is strictly increasing, then $w_l^ = w_u^*$ so that there exists a unique delay w^* . Equivalently, $F_k(w_l^*) = F_k(w_u^*)$ for each k .*

We start with the proof of uniqueness. Suppose there are two perfect allocations, α and ν . Then, without loss of generality, there exist delays $w_1 > w_2$ such that $\alpha_k = \lambda_k \mu_k^{-1} G_k(w_1)$ and $\nu_k = \lambda_k \mu_k^{-1} G_k(w_2)$ for each $k \in \{1, \dots, K\}$, and $\alpha_\ell < \nu_\ell$ for some $\ell \in \{1, \dots, K\}$. By the monotonicity and decreasing nature of the G_k 's and the statement that both α and ν are perfect allocations, we have that

$$1 = \sum_k \alpha_k < \sum_k \nu_k = 1,$$

a contradiction. Thus there is at most one perfect allocation.

We now show existence of a perfect allocation. For the remainder of the proof, consider the function $\sigma = (\sigma(w), w \geq 0)$, where $\sigma(w) = \sum_k \lambda_k \mu_k^{-1} G_k(w)$. By Assumption (A1), we have $\sigma(0) = \rho > 1$ and $\lim_{w \rightarrow \infty} \sigma(w) = 0$. Also by (A1), since each G_k is decreasing, monotonic and continuous, so is σ . It follows that there exists a $w^* > 0$ such that $\sigma(w^*) = 1$. Therefore there exists a perfect allocation $\alpha^* = (\alpha_k, k = 1, \dots, K)$, where $\alpha_k^* = \lambda_k \mu_k^{-1} G_k(w^*)$. The monotonicity of the G_k 's dictates

that any such $w > 0$ yielding $\sigma(w) = 1$ is also associated with this perfect allocation, because if $\sigma(w) = 1$ then $G_k(w) = G_k(w^*)$ for each k . Define

$$w_l^* = \inf\{w \geq 0 : \sigma(w) \geq 1\} \quad \text{and} \quad w_u^* = \sup\{w \geq 0 : \sigma(w) \leq 1\}. \quad (10)$$

By the continuity of σ we have $\sigma(w_l^*) = \sigma(w_u^*) = 1$. By the monotonicity of σ we have that $\sigma(w^*) = 1$ for all $w^* \in [w_l^*, w_u^*]$, and each of these delays are associated with the perfect allocation.

Now take any $w < w_l^*$ and some feasible allocation α . By the arguments above, we have $\sigma(w) = \sum_k \lambda_k \mu_k^{-1} G_k(w) > 1 \geq \sum_k \alpha_k$. Because all quantities are nonnegative, we have that $\alpha_k \mu_k < \lambda_k G_k(w)$ for some k and hence α is partially insufficient.

Likewise, take any $w > w_u^*$ and some complete allocation α . Again, because all quantities are nonnegative and $\sum_k \lambda_k G_k(w) < 1 = \sum_k \alpha_k$, there is at least one k such that $\alpha_k \mu_k > \lambda_k G_k(w)$. The allocation is partially excessive.

4. Limiting virtual waiting time process

In this section, we consider a sequence of queueing systems indexed by the parameter n in which the arrival rate and service rate become large but the deadline distribution remains fixed. The main result of this section is Theorem 1 which provides a first order ODE as the limit of the virtual waiting process as n tends to ∞ . We also show that as t tends to ∞ the solution to this ODE converges monotonically to a stationary point $w^* \in [w_l^*, w_u^*]$ which in general depends on the initial conditions of the system.

Our first step is to describe the sequence of queueing systems. Our convention is to indicate that random variables and processes are associated with the n^{th} system by appending the superscript n to them. We begin by assuming that the initial virtual waiting time in n^{th} system at time 0 is given by the random variable $W^n(0)$.

Next, for each $k \in \{1, \dots, K\}$, the number of class k jobs arriving to the n^{th} system by time $t \geq 0$ is given by

$$A_k^n(t) = \sup \left\{ i \geq 0 : \sum_{j=1}^i \xi_k^n(j) \leq t \right\},$$

where $\xi_k^n(i) = n^{-1}\xi_k(i)$ and $\{\xi_k(i), i \geq 1\}$ is the i.i.d. sequence of interarrival time random variables given in Section 2. It follows that class k jobs arrive to the n^{th} system according to a renewal arrival process with rate $n\lambda_k$. Furthermore, we set $t_k^n(i)$ equal to the arrival time of the i^{th} class k job. Again, we have the inverse relation: $A_k^n(t_k^n(i)) = i$.

In a similar manner, for each $i \geq 1$ we assume that the service time of the i^{th} class k job to arrive to the n^{th} system is given by $\nu_k^n(i) = n^{-1}\nu_k(i)$, where $\{\nu_k(i), i \geq 1\}$ is the i.i.d. sequence of potential service time random variables given in Section 2. Thus, the rate of service for class k jobs in the n^{th} system is given by $n\mu_k$.

Finally, we assume that the deadline time of the i^{th} class k job to arrive to the n^{th} system is given by $d_k(i)$, where $\{d_k(i), i \geq 1\}$ is the i.i.d. sequence of deadline time random variables with common distribution F_k given in Section 2. Note the absence of the superscript n on $d_k(i)$, implying that our sequence of deadline times is not changing as we index through n . To recapitulate, our overloaded regime is one in which the arrival and service rates grow proportionately to one another but the deadline time distributions and the offered load quantities ρ_1, \dots, ρ_K remain fixed.

Analogously to equation (8) in Section 2, we may express the virtual waiting time processes in the n^{th} system as

$$W^n(t) = W^n(0) - t + I^n(t) + \sum_{k=1}^K [M_{\nu,k}^n(A_k^n(t)) + M_{d,k}^n(A_k^n(t)) + \varepsilon_k^n(t)] + \int_0^t \sum_{k=1}^K \rho_k G_k(W^n(s)) ds, \quad t \geq 0, \quad (11)$$

where

$$M_{\nu,k}^n(i) = \sum_{j=1}^i \left(\nu_k^n(j) - \frac{1}{n\mu_k} \right) \mathbf{1}_{\{d_k(j) > W^n(t_k^n(j)-)\}}, \quad i \geq 0, \quad k \in \{1, \dots, K\}, \quad (12)$$

$$M_{d,k}^n(i) = \frac{1}{n\mu_k} \sum_{j=1}^i \left(\mathbf{1}_{\{d_k(j) > W^n(t_k^n(j)-)\}} - G_k(W^n(t_k^n(j)-)) \right), \quad i \geq 0, \quad k \in \{1, \dots, K\} \quad (13)$$

and

$$\varepsilon_k^n(t) = \frac{1}{n\mu_k} \left(\int_0^t G_k(W^n(s-)) dA_k^n(s) - \int_0^t G_k(W^n(s-)) d(\lambda_k s) \right), \quad t \geq 0, \quad (14)$$

are respective analogs of (4), (5) and (7), and

$$I^n(t) = \int_0^t 1_{\{W^n(s)=0\}} ds, \quad t \geq 0, \quad (15)$$

is the idle time process in the n^{th} system.

Defining for each job class $k \in \{1, \dots, K\}$ the fluid scaled quantities

$$\bar{A}_k^n(t) = n^{-1} A_k^n(t), \quad t \geq 0, \quad (16)$$

$$\bar{M}_{\nu,k}^n(t) = M_{\nu,k}^n(\lfloor nt \rfloor), \quad t \geq 0, \quad (17)$$

and

$$\bar{M}_{d,k}^n(t) = M_{d,k}^n(\lfloor nt \rfloor), \quad t \geq 0, \quad (18)$$

it follows from (11) that

$$\begin{aligned} W^n(t) = & W^n(0) - t + I^n(t) + \sum_{k=1}^K [\bar{M}_{\nu,k}^n(\bar{A}_k^n(t)) + \bar{M}_{d,k}^n(\bar{A}_k^n(t)) + \varepsilon_k^n(t)] \\ & + \int_0^t \sum_{k=1}^K \rho_k G_k(W^n(s)) ds, \quad t \geq 0. \end{aligned} \quad (19)$$

The representation above will be useful in proving the main result of this section, Theorem 1.

The following result is now a direct application of the Functional Strong Law of Large Numbers for renewal processes and partial sum processes. As such, we state it without proof. Letting $\bar{A}_k^n = (\bar{A}_k^n(t), t \geq 0)$, we now have the following.

Proposition 2 *For each $k \in \{1, \dots, K\}$, $\bar{A}_k^n \rightarrow \lambda_k e$ and $t_k^n(\lfloor ne \rfloor) \rightarrow \lambda_k^{-1} e$ almost surely, as $n \rightarrow \infty$.*

The next two propositions are required before the statement of our main result, Theorem 1; their proofs may be found in the Appendix. For each $k \in \{1, \dots, K\}$, let $\bar{M}_{\nu,k}^n = (\bar{M}_{\nu,k}^n(t), t \geq 0)$ and $\bar{M}_{d,k}^n = (\bar{M}_{d,k}^n(t), t \geq 0)$. We then have the following.

Proposition 3 *For each $k \in \{1, \dots, K\}$, $\bar{M}_{\nu,k}^n \Rightarrow 0$ as $n \rightarrow \infty$.*

Proposition 4 *For each $k \in \{1, \dots, K\}$, $\bar{M}_{d,k}^n \Rightarrow 0$ as $n \rightarrow \infty$.*

We next show that the sequence of virtual waiting time processes is tight, which in turn may be used to show that for each $k \in \{1, \dots, K\}$, the process $\varepsilon_k^n = (\varepsilon_k^n, t \geq 0)$ becomes vanishingly small in the limit as n tends to ∞ .

Proposition 5 *If $W^n(0) \rightarrow W(0)$ as $n \rightarrow \infty$, then the sequence $(W^n, n \geq 1)$ is tight.*

Proposition 6 *If $W^n(0) \rightarrow W(0)$ as $n \rightarrow \infty$, then, for each $k \in \{1, \dots, K\}$, $\varepsilon_k^n \Rightarrow 0$ as $n \rightarrow \infty$.*

We are now in a position to state the main result of this section.

Theorem 1 *Under Assumptions (A1), if $W^n(0) \rightarrow W(0)$ as $n \rightarrow \infty$, then $(W^n, I^n) \Rightarrow (W, 0)$ as $n \rightarrow \infty$, where $W = (W(t), t \geq 0)$ is the unique solution to the first order ODE,*

$$W(t) = W(0) - t + \int_0^t \sum_k \rho_k G_k(W(s)) ds, \quad t \geq 0. \quad (20)$$

Moreover, the process W is almost surely monotonic and the limiting value of W depends on its initial value, $W(0)$. In particular, if $W(0) < w_l^*$, then $W(t) \rightarrow w_l^*$ as $t \rightarrow \infty$, if $W(0) > w_u^*$, then $W(t) \rightarrow w_u^*$ as $t \rightarrow \infty$, and if $W(0) = w^* \in [w_l^*, w_u^*]$, then $W(t) = w^*$ for all $t \geq 0$.

PROOF OF THEOREM 1. First note that by equation (19) for the virtual waiting time process, we have by Proposition 3 of Reed and Ward (11), the representation

$$(W^n, I^n) = (\varphi_G, \psi_G) \left(W^n(0) - e + \sum_{k=1}^K [\bar{M}_{\nu,k}^n(\bar{A}_k^n(e)) + \bar{M}_{d,k}^n(\bar{A}_k^n(e)) + \varepsilon_k^n(e)] \right),$$

where $(\varphi_G, \psi_G) : D[0, \infty) \mapsto D^2[0, \infty)$ is a Lipschitz continuous function in the topology of uniform convergence on compact sets.

Next, it follows by Propositions 2, 3 and 4, the Random Time Change Theorem (3) and Proposition 6, that we have the weak convergence

$$\sum_{k=1}^K [\bar{M}_{\nu,k}^n(\bar{A}_k^n(e)) + \bar{M}_{d,k}^n(\bar{A}_k^n(e)) + \varepsilon_k^n(e)] \Rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Thus, by the assumption $W^n(0) \rightarrow W(0)$ as $n \rightarrow \infty$ in the statement of the theorem, the representation of (W^n, I^n) given above, the continuity of (φ_G, ψ_G) by Proposition 3 of Reed and Ward

(11) and the Continuous Mapping Theorem (15), it follows that $(W^n, I^n) \Rightarrow (\varphi_G, \psi_G)(W_0 - e)$ as $n \rightarrow \infty$, or equivalently, $(W^n, I^n) \Rightarrow (W, I)$ as $n \rightarrow \infty$, where W is the unique solution to the integral equation

$$W(t) = W(0) - t + I(t) + \int_0^t \sum_k \rho_k G_k(W(s)) ds, \quad t \geq 0, \quad (21)$$

and

$$\int_0^\infty \mathbf{1}_{\{W(s) > 0\}} dI(s) = 0. \quad (22)$$

We now prove the last statement of the theorem, which in turn shows that $W(t) > 0$ for all $t > 0$. Then, by the complimentary condition (22), we have that $I(t) = 0$ for all $t \geq 0$. First, note that

$$\frac{dW(t)}{dt} = -1 + \frac{dI(t)}{dt} + \sum_{k=1}^K \rho_k G_k(W(t)). \quad (23)$$

For any $w^* \in [w_l^*, w_u^*]$ we have

$$\sum_{k=1}^K \rho_k G_k(w^*) = 1.$$

For any t , if $W(t) = w^*$ then (22) implies $dI(t)/dt = 0$, so that $dW(t)/dt = 0$. Hence, if $W(s) = w^*$ for any $s \geq 0$ then $W(t) = w^*$ for all $t \geq s$. Now suppose that $0 \leq W(0) < w_l^*$. We first claim that $W(t) \leq w_l^*$ for all $t \geq 0$. This must be true since if $W(t) > w_l^*$, then, by the continuity of W and Bolzano's Intermediate Value Theorem (2), there must have existed some $s \leq t$ such that $W(s) = w_l^*$. But, since $w_l^* \in [w_l^*, w_u^*]$ is a fixed point of (21), this would then imply that $W(t) = w_l^*$ for all $t \geq s$, hence yielding a contradiction. Now note that if $W(t) < w_l^*$, then

$$\sum_{k=1}^K \rho_k G_k(t) > 1.$$

The fact that $dI(t)/dt \leq 1$ implies further that $dW(t)/dt > 0$, and hence W must be a strictly increasing process with an upper bound of w_l^* . Furthermore, for every $\varepsilon > 0$, there exists a $\delta_\varepsilon > 0$ such that $dW(t)/dt > \varepsilon$ for all $t < w_l^* - \delta_\varepsilon$. Hence, we have that $\lim_{t \rightarrow \infty} W(t) = w_l^*$. A similar argument may be used to show that if $W(0) > w_u^*$ then W decreases monotonically to w_u^* . This completes the proof.

Q.E.D.

The limiting process W given by (20) of Theorem 1 is a first order, separable ODE which may sometimes be solved for explicitly. First note that differentiating both sides of (20) with respect to t , we obtain the ODE

$$\frac{dW(t)}{dt} = -1 + \sum_{k=1}^K \rho_k G_k(W(t)), \quad t \geq 0,$$

so that

$$\frac{1}{-1 + \sum_{k=1}^K \rho_k G_k(W(t))} dW(t) = dt, \quad t \geq 0. \quad (24)$$

One may now attempt to integrate both sides of (24) in order to solve for W . We now present several examples in which this may be explicitly accomplished.

Example 1. Consider the case of $K = 1$ job class with exponentially distributed abandonment times with mean γ^{-1} . In this case, (24) becomes

$$\frac{1}{-1 + \rho e^{-\gamma W(t)}} dW(t) = dt, \quad t \geq 0,$$

which may be integrated to obtain

$$W(t) = \gamma^{-1} \ln(\rho + e^{-\gamma t} (e^{\gamma W(0) - \rho})), \quad t \geq 0. \quad (25)$$

Taking limits of both sides of the above as t goes to infinity, one also obtains that

$$\lim_{t \rightarrow \infty} W(t) = \gamma^{-1} \ln(\rho),$$

which corresponds to the limit of the virtual waiting time process as t goes to ∞ .

Example 2. Consider next the case of $K = 1$ job class with uniformly distributed abandonments time between a and b where $a \leq b$. In this case, we must treat the initial conditions of our system more carefully than in Example 1. First note that equation (24) becomes

$$\frac{1}{-1 + \rho(b-a)^{-1}(b - \max(W(t), a))^+} dW(t) = dt, \quad t \geq 0. \quad (26)$$

Now suppose that $W(0) \leq a$ in which case we obtain the solution $W(t) = W(0) + (\rho - 1)t$ for $0 \leq t \leq (a - W(0))/(\rho - 1)$ and

$$W(t) = b - \rho^{-1}(b - a)(1 + (\rho - 1)e^{-\rho(b-a)^{-1}t}), \quad t \geq (a - W(0))/(\rho - 1).$$

On the other hand, if $W(0) \geq b$, we obtain $W(t) = W(0) - t$ for $0 \leq t \leq b - W(0)$ and

$$W(t) = b - \rho^{-1}(b - a)(1 - e^{-\rho(b-a)^{-1}t}), \quad t \geq b - W(0).$$

Finally, if $W(0) \in [a, b]$, we obtain the solution

$$W(t) = b - \rho^{-1}(b - a)(1 + (\rho(b - a)^{-1}(b - W(0)) - 1)e^{-\rho(b-a)^{-1}t}), \quad t \geq 0.$$

Note also that taking limits as t goes to ∞ , we obtain in all three cases that

$$\lim_{t \rightarrow \infty} W(t) = b - \rho^{-1}(b - a). \quad (27)$$

which may be viewed as the limit for virtual waiting time process as t goes to ∞ .

Example 3. In this example, we consider the case of $K = 2$ job classes, where, for $k = 1, 2$, class k jobs have uniformly distributed deadlines between 0 and b_k , where, without loss of generality, we assume that $b_1 \leq b_2$. We then have that the ODE (24) becomes

$$\frac{1}{-1 + \rho_1 b_1^{-1}(b_1 - W(t))^+ + \rho_2 b_2^{-1}(b_2 - W(t))^+} dW(t) = dt, \quad t \geq 0. \quad (28)$$

Now let w^* be the solution to the equation

$$-1 + \rho_1 b_1^{-1}(b_1 - w^*)^+ + \rho_2 b_2^{-1}(b_2 - w^*)^+ = 0,$$

so that w^* will be the limiting point of the solution to (28). If $b_1 > ((\rho_2 - 1)/\rho_2)b_2$, then we have that $w^* < b_1$ so that at least some of the class 1 jobs will be served. On the other hand, if $b_1 \leq ((\rho_2 - 1)/\rho_2)b_2$, then the class 1 jobs are too impatient relative to the class 2 jobs and, as a consequence, $w^* \geq b_1$ so that none of the class 1 jobs are served. In either of these cases, one

may explicitly solve the ODE (28) in order to obtain the transient characteristics of the limiting virtual waiting time process as it tends to its limit point w^* .

To conclude this section, we provide fluid limit results for the processes that track of the number of jobs whose deadlines exceed prelimit and/or limiting versions of the virtual waiting time process. Before doing so, we must first prove the following technical result, whose proof may be found in the appendix. Let f be a real-valued function and for each $x \in D[0, \infty)$ define the function $\Gamma_f : D[0, \infty) \mapsto D[0, \infty)$ by

$$\Gamma_f(x)(t) = \int_0^t f(x(s))ds, \quad t \geq 0. \quad (29)$$

We then have the following.

Proposition 7 *If $f \in \mathcal{C}(\mathbb{R})$, then $\Gamma_f : D[0, \infty) \mapsto D[0, \infty)$ is continuous with respect to the Skorohod J_1 -metric.*

For each $k \in \{1, \dots, K\}$ and $n \geq 1$, let

$$R_k^n(i) = \sum_{j=1}^i \mathbf{1}_{\{d_k(j) > W^n(t_k^n(j-))\}}, \quad t \geq 0, \quad (30)$$

so that the quantity $R_k^n(A_k^n(t))$ counts the number of class k jobs that arrive to the system during $(0, t]$ and eventually receive service. Define $\bar{R}_k^n(t) = n^{-1}R_k^n(\lfloor nt \rfloor)$ to be the fluid scaled version and we set the corresponding process $\bar{R}_k^n = (\bar{R}_k^n(t), t \geq 0)$. Finally, define the fluid scaled vector $\bar{R}^n(t) = (\bar{R}_1^n(t), \dots, \bar{R}_K^n(t))$ and set $\bar{R}^n = (\bar{R}^n(t), t \geq 0)$. The following result now provides an asymptotic limit for the sequence of processes $(\bar{R}^n, n \geq 1)$ as n tends to ∞ .

Proposition 8 *If $W^n(0) \rightarrow W(0)$ as $n \rightarrow \infty$, then $\bar{R}^n \Rightarrow \bar{R}$ as $n \rightarrow \infty$, where for each $k \in \{1, \dots, K\}$,*

$$\bar{R}_k(t) = \int_0^{\lambda_k^{-1}t} G_k(W(s))d(\lambda_k s), \quad t \geq 0. \quad (31)$$

PROOF OF PROPOSITION 8. First note that since for each $k \in \{1, \dots, K\}$, the process $\bar{R}_k = (\bar{R}_k(t), t \geq 0)$ given above is deterministic, it follows by Theorem 3.9 of (3) that for each $k \in \{1, \dots, K\}$, it is sufficient to show that $\bar{R}_k^n \Rightarrow \bar{R}_k$ as $n \rightarrow \infty$.

Now note that for each $k \in \{1, \dots, K\}$ and $n \geq 1$, we have by (13) and (14), that for $t \geq 0$,

$$\begin{aligned} R_k^n(\lfloor nt \rfloor) &= \sum_{i=1}^{\lfloor nt \rfloor} \left(\mathbf{1}_{\{d_k(i) > W^n(t_k^n(i)-)\}} - G_k(W^n(t_k^n(i)-)) \right) \\ &\quad + \left(\sum_{i=1}^{\lfloor nt \rfloor} G_k(W^n(t_k^n(i)-)) - \int_0^{t_k^n(\lfloor nt \rfloor)} G_k(W^n(s)) d(n\lambda_k s) \right) \\ &\quad + \int_0^{t_k^n(\lfloor nt \rfloor)} G_k(W^n(s)) d(n\lambda_k s) \\ &= nM_{d,k}^n(\lfloor nt \rfloor) + n\mu_k \varepsilon_k^n(t_k^n(\lfloor nt \rfloor)) + \int_0^{t_k^n(\lfloor nt \rfloor)} G_k(W^n(s)) d(n\lambda_k s). \end{aligned}$$

Thus, by (18),

$$\bar{R}_k^n(t) = \bar{M}_{d,k}^n(t) + \mu_k \varepsilon_k^n(t_k^n(\lfloor nt \rfloor)) + \int_0^{t_k^n(\lfloor nt \rfloor)} G_k(W^n(s)) d(\lambda_k s), \quad (32)$$

for $t \geq 0$. Now, by Propositions 2, 4 and 6, and the Random Time Change Theorem (3), we have that $\bar{M}_{d,k}^n \Rightarrow 0$ and $\varepsilon_k^n \circ t_k^n(\lfloor ne \rfloor) \Rightarrow 0$ as $n \rightarrow \infty$, so that by (32) it remains to show that

$$\int_0^e G_k(W^n(s)) d(\lambda_k s) \Rightarrow \int_0^e G_k(W(s)) d(\lambda_k s) \quad \text{as } n \rightarrow \infty, \quad (33)$$

which, combined with the convergence of $t_k^n(\lfloor ne \rfloor)$ (Proposition 2) and the Random Time Change Theorem (3), will complete the proof.

In order to show (33), first note that by the definition of Γ_{G_k} in (29) above, we have that

$$\Gamma_{G_k}(x) = \int_0^e G_k(x(s)) d(\lambda_k s). \quad (34)$$

Furthermore, since the function G_k is bounded and by Assumption (A1) continuous, it follows by Proposition 7 that the function $\Gamma_{G_k} : D[0, \infty) \mapsto D[0, \infty)$ is continuous with respect to the Skorohod J_1 -metric. Recall that by Theorem 1 we have $W^n \Rightarrow W$ as $n \rightarrow \infty$. It follows by the Continuous Mapping Theorem (15) that

$$\int_0^e G_k(W^n(s)) d(\lambda_k s) = \Gamma_{G_k}(W^n) \Rightarrow \Gamma_{G_k}(W) = \int_0^e G_k(W) d(\lambda_k s) \quad \text{as } n \rightarrow \infty,$$

which completes the proof.

Q.E.D.

It will be beneficial later to have a version of R_k^n that is independent of both the virtual waiting time process as well as the arrival process. We can replace the virtual waiting time process by its limiting, deterministic process. Likewise, the arrival process can be replaced by a deterministic counting process whose jumps are evenly spaced. Specifically, we can replace the process A_k^n with $\hat{A}_k^n = (\hat{A}_k^n(t), t \geq 0)$, where $\hat{A}_k^n(t) = \lfloor n\lambda_k t \rfloor$ for each $t \geq 0$. We introduce the fluid scaled version $\bar{\hat{A}}_k^n = (\bar{\hat{A}}_k^n(t), t \geq 0)$, where $\bar{\hat{A}}_k^n(t) = n^{-1}\hat{A}_k^n(t)$ for each $t \geq 0$. The following is straightforward:

$$\bar{\hat{A}}_k^n \rightarrow \lambda_k e \quad \text{as } n \rightarrow \infty. \quad (35)$$

Next, for each $k \in \{1, \dots, K\}$ and $n \geq 1$, set

$$\hat{R}_k^n(i) = \sum_{j=1}^i \mathbf{1}_{\{d_k(j) > W(j/(n\lambda_k))\}}, \quad t \geq 0, \quad (36)$$

and

$$\check{R}_k^n(i) = \sum_{j=1}^i \mathbf{1}_{\{d_k(j) > \max(W^n(t_k^n(j)-), W(j/(n\lambda_k)))\}}, \quad t \geq 0. \quad (37)$$

The processes in (36) track those jobs that arrive to the system by time t and would eventually receive service if the prelimit virtual waiting process were replaced with the limiting waiting time process. Instead of evaluating the limiting waiting time process at the time of arrival of the i^{th} class k job, $t_k^n(i)$, we use the approximate arrival time $i/(n\lambda_k)$. For the processes in (37), each job's deadline is compared to the maximum of the limiting virtual waiting time and its prelimit; the prelimit process is evaluated at the times of the jobs' arrivals and the limiting process is evaluated at the approximate arrival times. Define $\bar{\hat{R}}_k^n(t) = n^{-1}\hat{R}_k^n(t)$ and $\bar{\check{R}}_k^n(t) = n^{-1}\check{R}_k^n(t)$ to be the corresponding fluid scaled quantities and set the corresponding processes $\bar{\hat{R}}_k^n = (\bar{\hat{R}}_k^n(t), t \geq 0)$ and $\bar{\check{R}}_k^n = (\bar{\check{R}}_k^n(t), t \geq 0)$. Finally, define the fluid scaled vectors $\bar{\hat{R}}^n(t) = (\bar{\hat{R}}_1^n(t), \dots, \bar{\hat{R}}_K^n(t))$ and $\bar{\check{R}}^n(t) = (\bar{\check{R}}_1^n(t), \dots, \bar{\check{R}}_K^n(t))$, and set $\bar{\hat{R}}^n = (\bar{\hat{R}}^n(t), t \geq 0)$ and $\bar{\check{R}}^n = (\bar{\check{R}}^n(t), t \geq 0)$. The following result is analogous to Proposition 8 and claims that $\bar{\hat{R}}^n$ and $\bar{\check{R}}^n$ have the same limit as \bar{R}^n .

Proposition 9 *If $W^n(0) \rightarrow W(0)$ as $n \rightarrow \infty$, then $(\bar{R}^n, \tilde{R}^n) \Rightarrow (\bar{R}, \bar{R})$, where \bar{R} is given in (31).*

PROOF OF PROPOSITION 9. As in the proof of Proposition 8, it is sufficient to show that, for each $k \in \{1, \dots, K\}$, $\bar{R}_k^n \Rightarrow \bar{R}_k$ as $n \rightarrow \infty$ and $\tilde{R}_k^n \Rightarrow \bar{R}_k$ as $n \rightarrow \infty$. Note that for each $k \in \{1, \dots, K\}$ and $n \geq 1$, because W is continuous, we have that, for $t \geq 0$,

$$\hat{R}_k^n(i) = B_k^n(i) + \int_0^{\lambda_k^{-1}i} G_k(W(s))d\hat{A}_k^n(s),$$

where

$$B_k^n(i) = \sum_{j=1}^i (\mathbf{1}_{\{d_k(j) > W(j/(n\lambda_k))\}} - G_k(W(j/(n\lambda_k))))$$

is analogous to $M_{d,k}^n$; the difference is that the former is defined in terms of W and the approximate arrival times, while the latter is defined using W^n and the actual arrival times. Defining

$$\bar{B}_k^n(t) = n^{-1}B_k^n(\lfloor nt \rfloor),$$

we have

$$\tilde{R}_k^n(t) = \bar{B}_k^n(t) + \int_0^{\lambda_k^{-1}t} G_k(W(s))d\tilde{A}_k^n(s) \tag{38}$$

for $t \geq 0$. We can see that B_k^n is a martingale and so, analogously to Proposition 4, we have for each $k \in \{1, \dots, K\}$ that $\bar{B}_k^n \Rightarrow 0$ as $n \rightarrow \infty$. Further, it follows by Lemma 8.1 in Dai and Dai (5) and the convergence (35) that

$$\int_0^{\lambda_k^{-1}e} G_k(W(s))d\tilde{A}_k^n(s) \Rightarrow \int_0^{\lambda_k^{-1}e} G_k(W(s))d(\lambda_k s), \text{ as } n \rightarrow \infty$$

so that $\tilde{R}_k^n \Rightarrow \bar{R}_k$, as $n \rightarrow \infty$. Finally, $\bar{R}^n \Rightarrow \bar{R}$, as $n \rightarrow \infty$ because, for each $k \in \{1, \dots, K\}$, \bar{R}_k is deterministic. The proof for convergence of \bar{R}^n is similar. Joint convergence follows because all limit points are deterministic. Q.E.D.

5. Limiting diffusion-scaled virtual waiting time process

In this section, we study the diffusion scaled virtual waiting time process. Our main approach is to first center the virtual waiting time process by its fluid limit of Section 4 and then scale by an appropriate sequence of constants. Through this approach we capture the stochastic fluctuations

of the virtual waiting time process about its mean. In our main result of this section, Theorem 2, we provide a diffusion limit for the sequence of centered and normalized virtual waiting time processes. The limit is an Ornstein-Uhlenbeck process with time varying drift and infinitesimal variance, a process which has highly tractable transient and steady state behavior. Using this limit, we are able to obtain approximations to both the transient as well as the steady state behavior of the virtual waiting time process.

For the remainder of this section, we make the following assumption on the class specific deadline distributions in addition to Assumption (A1).

Assumption (A2) For each job class $k \in \{1, \dots, K\}$, we may write

$$F_k(x) = \int_0^x f_k(u)du, \quad x \geq 0,$$

where $f \in \mathbb{C}_b(\mathbb{R})$. Further, for each $t \geq 0$ there exists at least one $k \in \{1, \dots, K\}$ such that $f_k(t) > 0$.

Under Assumption (A2) it follows that, for each $k \in \{1, \dots, K\}$, f_k is the density of F_k with respect to the Lebesgue measure. Moreover, the strictly positive condition implies that $w_l^* = w_u^*$. Under Assumption (A2), we refer to the unique equilibrium point as w^* .

We now begin our analysis by centering the virtual waiting time process in (19) by its deterministic fluid limit $W = (W(t), t \geq 0)$, as given in Theorem 1 of Section 4, and noting that

$$\begin{aligned} (W^n(t) - W(t)) &= (W^n(0) - W(0)) + I^n(t) \\ &+ \sum_{k=1}^K [\bar{M}_{\nu,k}^n(\bar{A}_k^n(t)) + \bar{M}_{d,k}^n(\bar{A}_k^n(t)) + \varepsilon_k^n(t)] \\ &+ \int_0^t \sum_{k=1}^K \rho_k (G_k(W^n(s)) - G_k(W(s))) ds, \quad t \geq 0. \end{aligned} \tag{39}$$

Next, setting

$$\delta_k^n(t) = \int_0^t \rho_k (G_k(W^n(s)) - G_k(W(s))) ds + \rho_k \int_0^t f_k(W(s))(W^n(s) - W(s)) ds,$$

for $t \geq 0$, it follows upon substitution into (39) that

$$(W^n(t) - W(t)) = (W^n(0) - W(0)) + I^n(t) \tag{40}$$

$$\begin{aligned}
 & + \sum_{k=1}^K [\bar{M}_{\nu,k}^n(\bar{A}_k^n(t)) + \bar{M}_{d,k}^n(\bar{A}_k^n(t)) + \varepsilon_k^n(t) + \delta_k^n(t)] \\
 & - \int_0^t \left(\sum_{k=1}^K \rho_k f_k(W(s)) \right) (W^n(s) - W(s)) ds, \quad t \geq 0.
 \end{aligned}$$

If we now define the diffusion scaled quantities,

$$\tilde{W}^n(t) = n^{1/2}(W^n(t) - W(t)), \quad t \geq 0, \quad (41)$$

$$\tilde{I}^n(t) = n^{1/2}I^n(t), \quad t \geq 0,$$

$$\tilde{M}_{\nu,k}^n(t) = n^{1/2}\bar{M}_{\nu,k}^n(t), \quad t \geq 0, \quad k \in \{1, \dots, K\},$$

$$\tilde{M}_{d,k}^n(t) = n^{1/2}\bar{M}_{d,k}^n(t), \quad t \geq 0, \quad k \in \{1, \dots, K\},$$

$$\tilde{\varepsilon}_k^n(t) = n^{1/2}\varepsilon_k^n(t), \quad t \geq 0, \quad k \in \{1, \dots, K\},$$

(42)

and

$$\tilde{\delta}_k^n(t) = n^{1/2}\delta_k^n(t), \quad t \geq 0, \quad k \in \{1, \dots, K\},$$

it then follows, multiplying equation (40) through by $n^{1/2}$ that

$$\begin{aligned}
 \tilde{W}^n(t) &= \tilde{W}^n(0) + \tilde{I}^n(t) \\
 &+ \sum_{k=1}^K [\tilde{M}_{\nu,k}^n(\bar{A}_k^n(t)) + \tilde{M}_{d,k}^n(\bar{A}_k^n(t)) + \tilde{\varepsilon}_k^n(t) + \tilde{\delta}_k^n(t)] \\
 &+ \int_0^t \left(\sum_{k=1}^K \rho_k f_k(W(s)) \right) \tilde{W}^n(s) ds, \quad t \geq 0.
 \end{aligned} \quad (43)$$

Equation (43) now provides the starting point for proving the main result of this section.

Define the function $\hat{f}: D[0, \infty) \mapsto D[0, \infty)$ for each $x \in D[0, \infty)$ by

$$\hat{f}(x)(t) = \sum_{k=1}^K \rho_k f_k(W(t))x(t), \quad t \geq 0.$$

We have that for $x, y \in D[0, \infty)$ and $T \geq 0$,

$$\|\hat{f}(x) - \hat{f}(y)\|_T = \sup_{0 \leq t \leq T} \left| \sum_{k=1}^K \rho_k f_k(W(t))x(t) - \sum_{k=1}^K \rho_k f_k(W(t))y(t) \right|$$

$$\leq \sup_{0 \leq t \leq T} \sum_{k=1}^K \rho_k f_k(W(t)) \|x - y\|_T,$$

so that \hat{f} is Lipschitz continuous in the topology of uniform convergence on compact intervals.

Thus, defining for each $n \geq 1$ the fluid centered and diffusion scaled virtual waiting time process $\tilde{W}^n = (\tilde{W}^n(t), t \geq 0)$ and for each $k \in \{1, \dots, K\}$ the processes $\tilde{M}_{\nu,k}^n = (\tilde{M}_{\nu,k}^n(t), t \geq 0)$, $\tilde{M}_{d,k}^n = (\tilde{M}_{d,k}^n(t), t \geq 0)$, $\tilde{\varepsilon}_k^n = (\varepsilon M_k^n(t), t \geq 0)$ and $\tilde{\delta}_k^n = (\tilde{\delta}_k^n(t), t \geq 0)$, it follows by Proposition 3 of Reed and Ward (11) that we have the representation

$$\tilde{W}^n = \mathcal{M}_{\hat{f}} \left(\tilde{W}^n(0) + \tilde{I}^n + \sum_{k=1}^K \left[\tilde{M}_{\nu,k}^n \circ \bar{A}_k^n + \tilde{M}_{d,k}^n \circ \bar{A}_k^n + \tilde{\varepsilon}_k^n + \tilde{\delta}_k^n \right] \right), \quad (44)$$

where the function $\mathcal{M}_{\hat{f}}: D[0, \infty) \mapsto D[0, \infty)$ is Lipschitz continuous with respect to the Skorohod- J_1 metric. This representation will be used in proving our main result.

Before stating the main result of this section, Theorem 2, we must first provide some auxiliary results. In our first result, setting $\tilde{I}^n = (\tilde{I}^n(t), t \geq 0)$ for $n \geq 1$, we show that the sequence of diffusion scaled idle time processes, $(\tilde{I}^n, n \geq 1)$ converges in distribution to 0 as n goes to ∞ .

Proposition 10 *If $W^n(0) \rightarrow W(0) > 0$ and $\tilde{W}^n(0) \Rightarrow \tilde{W}(0)$ as $n \rightarrow \infty$, then $\tilde{I}^n \Rightarrow 0$ as $n \rightarrow \infty$.*

We next prove convergence results for the sequences of diffusion scaled martingales $(\tilde{M}_{\nu,k}^n, n \geq 1)$ and $(\tilde{M}_{d,k}^n, n \geq 1)$. For each $k \in \{1, \dots, K\}$ and $n \geq 1$, define the approximating process

$$\hat{M}_{\nu,k}^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \left(\nu_k(i) - \frac{1}{\mu_k} \right) \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}}, \quad t \geq 0, \quad (45)$$

and set $\hat{M}_{\nu,k}^n = (\hat{M}_{\nu,k}^n(t), t \geq 0)$. Note that $\hat{M}_{\nu,k}^n$ is independent of each of the arrival processes A_j^n for $j = 1, \dots, K$ and of the sequences $\{v_j(i), i \geq 1\}$ and $\{d_j(i), i \geq 1\}$ for $j \neq k$. We now have the following result.

Proposition 11 *If $\tilde{W}^n(0) \Rightarrow \tilde{W}(0)$ as $n \rightarrow \infty$, then, for each $k \in \{1, \dots, K\}$, $(\tilde{M}_{\nu,k}^n, \hat{M}_{\nu,k}^n) \Rightarrow (\tilde{M}_{\nu,k}, \hat{M}_{\nu,k})$ as $n \rightarrow \infty$, where $\tilde{M}_{\nu,k} = (\tilde{M}_{\nu,k}(t), t \geq 0)$ is a Brownian motion with time varying infinitesimal variance given by*

$$\sigma_{\nu,k}^2(t) = \sigma_{s,k}^2 G_k(W(t/\lambda_k)), \quad t \geq 0.$$

Next, for each $k \in \{1, \dots, K\}$ and $n \geq 1$, define the approximating process

$$\hat{M}_{d,k}^n(t) = \frac{1}{\sqrt{n\mu_k}} \sum_{i=1}^{\lfloor nt \rfloor} (\mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} - G_k(W(i/(n\lambda_k))), \quad t \geq 0, \quad (46)$$

and set $\hat{M}_{d,k}^n = (\hat{M}_{d,k}^n(t), t \geq 0)$. Note that for each $k \in \{1, \dots, K\}$, $\hat{M}_{d,k}^n$ is independent of each of the arrival processes A_j^n and of the service time sequences $\{v_j(i), i \geq 1\}$ for $j = 1, \dots, K$. Moreover, $\hat{M}_{d,k}^n$ is independent of the deadline sequences $\{d_j(i), i \geq 1\}$ for each $j \neq k$. We now have the following result for the sequence of martingales $(\hat{M}_{d,k}^n, n \geq 1)$; it is analogous to Proposition 11.

Proposition 12 *If $\tilde{W}^n(0) \Rightarrow \tilde{W}(0)$ as $n \rightarrow \infty$, then, for each $k \in \{1, \dots, K\}$, $(\hat{M}_{d,k}^n, \hat{M}_{d,k}^n) \Rightarrow (\tilde{M}_{d,k}, \tilde{M}_{d,k})$ as $n \rightarrow \infty$, where $\tilde{M}_{d,k} = (\tilde{M}_{d,k}(t), t \geq 0)$ is a Brownian motion with time varying infinitesimal variance given by*

$$\sigma_{d,k}^2(t) = \mu_k^{-2} F_k(W(t/\lambda_k)) G_k(W(t/\lambda_k)), \quad t \geq 0.$$

Although the processes $M_{\nu,k}^n$ and $\hat{M}_{d,k}^n$ are not independent, they are asymptotically independent, as given in the following result.

Proposition 13 *If $\tilde{W}^n(0) \Rightarrow \tilde{W}(0)$ as $n \rightarrow \infty$, then, for each $k \in \{1, \dots, K\}$, $(\hat{M}_{\nu,k}^n, \hat{M}_{d,k}^n) \Rightarrow (\tilde{M}_{\nu,k}, \tilde{M}_{d,k})$ as $n \rightarrow \infty$, where $\tilde{M}_{\nu,k}$ and $\tilde{M}_{d,k}$ are independent.*

Fix $k \in \{1, \dots, K\}$. We next treat the sequence of processes $(\hat{\varepsilon}_k^n, n \geq 1)$ by showing that the sequence converges in distribution to a Brownian motion with time varying infinitesimal variance.

Define

$$\hat{\varepsilon}_k^n(t) = \frac{1}{\sqrt{n\mu_k}} \left(\int_0^t G_k(W(s)) dA_k^n(s) - \int_0^t G_k(W(s)) d(\lambda_k s) \right), \quad t \geq 0.$$

In a similar manner to the $\hat{M}_{d,k}^n$ processes defined above, we have that $\hat{\varepsilon}_k^n$ is independent of each of the arrival processes A_j^n for $j \neq k$ and independent of the sequence of service times and deadline times $\{v_j(i), i \geq 1\}$ and $\{d_j(i), i \geq 1\}$ for each $j \in \{1, \dots, K\}$. We then have the following result.

Proposition 14 *If $\tilde{W}^n(0) \Rightarrow \tilde{W}(0)$ as $n \rightarrow \infty$, then, for each $k \in \{1, \dots, K\}$, $(\tilde{\varepsilon}_k^n, \hat{\varepsilon}_k^n) \Rightarrow (\tilde{\varepsilon}_k, \tilde{\varepsilon}_k)$ as $n \rightarrow \infty$, where $\tilde{\varepsilon}_k = (\tilde{\varepsilon}_k(t), t \geq 0)$ is a Brownian motion with time varying infinitesimal variance given by*

$$\sigma_{\varepsilon,k}^2(t) = \lambda_k \rho_k^2 \sigma_{a,k}^2 G_k(W(t))^2, \quad t \geq 0.$$

Finally, before stating the main result of the section, we show that for each $k \in \{1, \dots, K\}$, the sequence of processes $(\tilde{\delta}_k^n, n \geq 1)$ converges in distribution to 0 as n tends to ∞ .

Proposition 15 *If $\tilde{W}^n(0) \Rightarrow \tilde{W}(0)$ as $n \rightarrow \infty$, then, for each $k \in \{1, \dots, K\}$, $\tilde{\delta}_k^n \Rightarrow 0$ as $n \rightarrow \infty$.*

We are now in a position to state the main result of this section, Theorem 2, which provides a weak limit for the sequence of diffusion scaled virtual waiting time processes, $(\tilde{W}^n, n \geq 1)$.

Theorem 2 *Under assumptions (A1) and (A2), if $W^n(0) \rightarrow W(0) > 0$ as $n \rightarrow \infty$, and $\tilde{W}^n(0) \Rightarrow \tilde{W}(0)$ as $n \rightarrow \infty$, then $\tilde{W}^n \Rightarrow \tilde{W}$ as $n \rightarrow \infty$, where $\tilde{W} = (\tilde{W}(t), t \geq 0)$ is the unique strong solution to the stochastic differential equation*

$$\tilde{W}(t) = \tilde{W}(0) + \int_0^t \tilde{\sigma}(s) d\tilde{B}(s) - \int_0^t \left(\sum_{k=1}^K \rho_k f_k(W(s)) \right) \tilde{W}(s) ds, \quad t \geq 0, \quad (47)$$

where $\tilde{B} = (\tilde{B}(t), t \geq 0)$ is a standard Brownian motion and

$$\tilde{\sigma}^2(t) = \sum_{k=1}^K \left[\sigma_{s,k}^2 G_k(W(t)) + \lambda_k \rho_k^2 \sigma_{a,k}^2 G_k(W(t))^2 + \mu_k^{-2} F_k(W(t)) G_k(W(t)) \right], \quad t \geq 0. \quad (48)$$

PROOF OF THEOREM 2. First recall that by (44) above, we have the representation

$$\tilde{W}^n = \mathcal{M}_{\tilde{f}} \left(\tilde{W}^n(0) + \tilde{I}^n + \sum_{k=1}^K \left[\tilde{M}_{\nu,k}^n \circ \bar{A}_k^n + \tilde{M}_{d,k}^n \circ \bar{A}_k^n + \tilde{\varepsilon}_k^n + \tilde{\delta}_k^n \right] \right), \quad (49)$$

where the function $\mathcal{M}_{\tilde{f}} : D[0, \infty) \mapsto D[0, \infty)$ is continuous with respect to the Skorohod J_1 -metric.

Thus, by the Continuous Mapping Theorem (15) and the definition of $\mathcal{M}_{\tilde{f}}$ in (17) of (?), it remains to show that

$$\tilde{W}^n(0) + \tilde{I}^n + \sum_{k=1}^K \left[\tilde{M}_{\nu,k}^n \circ \bar{A}_k^n + \tilde{M}_{d,k}^n \circ \bar{A}_k^n + \tilde{\varepsilon}_k^n + \tilde{\delta}_k^n \right] \Rightarrow \tilde{W}(0) + \int_0^e \tilde{\sigma}(s) \tilde{B}(s), \quad (50)$$

as $n \rightarrow \infty$, where $\tilde{B} = (\tilde{B}(t), t \geq 0)$ is a standard Brownian motion and $\tilde{\sigma} = (\tilde{\sigma}(t), t \geq 0)$ is as given by (48) in the statement of the proposition.

In order to show that (50) holds, first note that

$$\begin{aligned} & \tilde{W}^n(0) + \tilde{I}^n + \sum_{k=1}^K \left[\tilde{M}_{\nu,k}^n \circ \bar{A}_k^n + \tilde{M}_{d,k}^n \circ \bar{A}_k^n + \tilde{\varepsilon}_k^n + \tilde{\delta}_k^n \right] \\ &= \tilde{W}^n(0) + \tilde{I}^n + \sum_{k=1}^K \left[(\tilde{M}_{\nu,k}^n \circ \bar{A}_k^n - \hat{M}_{\nu,k}^n \circ \lambda_k e) + (\tilde{M}_{d,k}^n \circ \bar{A}_k^n - \hat{M}_{d,k}^n \circ \lambda_k e) \right. \\ & \quad \left. + (\tilde{\varepsilon}_k^n - \hat{\varepsilon}_k^n) + \tilde{\delta}_k^n \right] \\ & \quad + \sum_{k=1}^K [\hat{M}_{\nu,k}^n \circ \lambda_k e + \hat{M}_{d,k}^n \circ \lambda_k e + \hat{\varepsilon}_k^n]. \end{aligned} \quad (51)$$

We now claim that

$$\tilde{I}^n + \sum_{k=1}^K [(\tilde{M}_{\nu,k}^n \circ \bar{A}_k^n - \hat{M}_{\nu,k}^n \circ \lambda_k e) + (\tilde{M}_{d,k}^n \circ \bar{A}_k^n - \hat{M}_{d,k}^n \circ \lambda_k e) + (\tilde{\varepsilon}_k^n - \hat{\varepsilon}_k^n) + \tilde{\delta}_k^n] \Rightarrow 0, \quad (52)$$

as $n \rightarrow \infty$, after which it then remains to show that

$$\tilde{W}^n(0) + \sum_{k=1}^K [\hat{M}_{\nu,k}^n \circ \lambda_k e + \hat{M}_{d,k}^n \circ \lambda_k e + \hat{\varepsilon}_k^n] \Rightarrow \tilde{W}(0) + \int_0^e \tilde{\sigma}(s) d\tilde{B}(s), \quad (53)$$

as $n \rightarrow \infty$, in order to complete the proof.

First note that by Proposition 10, Proposition 14, Theorem 11.4.8 of (15) and Proposition 15, it follows that

$$\tilde{I}^n + \sum_{k=1}^K [(\tilde{\varepsilon}_k^n - \hat{\varepsilon}_k^n) + \tilde{\delta}_k^n] \Rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Next, since for each $k \in \{1, \dots, K\}$ we have that since $\lambda_k e$ is a deterministic process, it follows by Proposition 12 and Theorem 3.9 of (3) that $(\tilde{M}_{\nu,k}^n, \hat{M}_{\nu,k}^n, \bar{A}_k^n, \lambda_k e) \Rightarrow (\tilde{M}_{\nu,k}, \tilde{M}_{\nu,k}, \lambda_k e, \lambda_k e)$ as $n \rightarrow \infty$, and hence, by the Random Time Change Theorem (3), $(\tilde{M}_{\nu,k}^n \circ \bar{A}_k^n, \hat{M}_{\nu,k}^n \circ \lambda_k e) \Rightarrow (\tilde{M}_{\nu,k} \circ \lambda_k e, \tilde{M}_{\nu,k} \circ \lambda_k e)$ as $n \rightarrow \infty$. Thus, by Theorem 11.4.8 of (15), $\tilde{M}_{\nu,k}^n \circ \bar{A}_k^n - \hat{M}_{\nu,k}^n \circ \lambda_k e \Rightarrow 0$ as $n \rightarrow \infty$. A similar proof may also be used to show that for each $k \in \{1, \dots, K\}$, $\tilde{M}_{d,k}^n \circ \bar{A}_k^n - \hat{M}_{d,k}^n \circ \lambda_k e \Rightarrow 0$ as $n \rightarrow \infty$, which completes the proof of (52).

It now remains to show the stated convergence in (53) in order to complete the proof. First note that by definition each of the processes appearing on the lefthand side of (53) above are either

independent of one another – or at least asymptotically independent of each other (see Proposition 13) – as well as independent from $W^n(0)$. Thus, since by the assumption of the Theorem, $W^n(0) \Rightarrow W(0)$ as $n \rightarrow \infty$ in conjunction with Propositions 11, 12, 13 and 14, it follows that

$$\tilde{W}^n(0) + \sum_{k=1}^K [\hat{M}_{\nu,k}^n \circ \lambda_k e + \hat{M}_{d,k}^n \circ \lambda_k e + \hat{\varepsilon}_k^n] \Rightarrow W(0) + \sum_{k=1}^K [\tilde{M}_{\nu,k} \circ \lambda_k e + \tilde{M}_{d,k} \circ \lambda_k e + \tilde{\varepsilon}_k],$$

as $n \rightarrow \infty$. Furthermore, since each of the limiting processes on the righthand side above are mutually independent, one may now use Propositions 11, 12, 13 and 14 to verify that

$$\sum_{k=1}^K [\tilde{M}_{\nu,k} \circ \lambda_k e + \tilde{M}_{d,k} \circ \lambda_k e + \tilde{\varepsilon}_k]$$

is indeed a Brownian motion with time-varying infinitesimal variance $\tilde{\sigma} = (\tilde{\sigma}(t), t \geq 0)$. The proof is now complete.

Q.E.D.

The limiting process of Theorem 2, \tilde{W} , may be characterized as an Ornstein-Uhlenbeck processes with both time varying drift and infinitesimal variance. Consequently, by standard techniques one may solve for \tilde{W} directly in terms of the Brownian motion \tilde{B} . Specifically, letting

$$f(x) = \sum_{k=1}^K \rho_k f_k(x), \quad x \geq 0,$$

a standard proof may be used to show that the solution to the stochastic differential equation given by Theorem 2 above is given by

$$\tilde{W}(t) = \tilde{W}(0)e^{-\int_0^t f(W(s))ds} + \int_0^t e^{-\int_s^t f(W(u))du} \tilde{\sigma}(s) d\tilde{B}(s), \quad t \geq 0.$$

Furthermore, for each $0 \leq s < t$, one may also show that the transitions densities of \tilde{W} are given by

$$\mathbb{P}(\tilde{W}(t) \in dy | \tilde{W}(s) = x) = \frac{1}{\sqrt{2\pi\tilde{\sigma}^2(s,t)}} e^{-(y-xe^{-\int_s^t f(W(u))du})^2/2\tilde{\sigma}^2(s,t)} dy,$$

where $\tilde{\sigma}^2(s,t) = \int_s^t e^{-2\int_u^t f(W(x))dx} \tilde{\sigma}^2(u) du$. These results may then be used in order to obtain insights into to the transient dynamics of the virtual waiting time process in the original system. In particular, recalling from (41) the relationship

$$\tilde{W}^n(t) = n^{1/2}(W^n(t) - W(t)), \quad t \geq 0,$$

and since by Theorem 2, $\tilde{W}^n \Rightarrow \tilde{W}$ as $n \rightarrow \infty$, it follows that

$$W^n(t) = W(t) + n^{-1/2}\tilde{W}(t) + o(n^{-1/2}), \quad t \geq 0. \quad (54)$$

Ignoring the $o(n^{-1/2})$ term in (54), one then has the approximation

$$W^n(t) \approx W(t) + n^{-1/2}\tilde{W}(t), \quad t \geq 0. \quad (55)$$

which, together with Theorem 2 and the discussion above may be used to obtain an approximation to the transient behavior of W^n . In particular, we have that

$$\begin{aligned} & \mathbb{P}(W^n(t) \in dy | W^n(s) \in dx) \\ & \approx \mathbb{P}(W(t) + n^{-1/2}\tilde{W}(t) \in dy | W(s) + n^{-1/2}\tilde{W}(s) \in dx) \\ & = \mathbb{P}(\tilde{W}(t) \in d(n^{1/2}(y - W(t))) | \tilde{W}(s) \in d(n^{1/2}(x - W(s))))), \end{aligned}$$

for $0 \leq s < t$.

In addition to the transient behavior, one may also analyze the stationary behavior on the limiting virtual waiting time process \tilde{W} . Specifically, first note that in the case where $W(0) = w^*$, we have that the fluid limit W of Theorem 1 is constant and hence \tilde{W} is a time-homogeneous Ornstein-Uhlenbeck process. In this case, well known results show that $\tilde{W}(t) \Rightarrow \tilde{W}(\infty)$ as $t \rightarrow \infty$, where $\tilde{W}(\infty)$ is a normal random variable with mean 0 and variance

$$\tilde{\sigma}^2(0, \infty) = \sum_{k=1}^K [\sigma_{s,k}^2 G_k(w^*) + \lambda_k \rho 2_k \sigma_{a,k}^2 G_k(w^*)^2 + \mu_k^{-2} F_k(w^*) G_k(w^*)] / (2 \sum_{k=1}^K f_k(w^*)).$$

Furthermore, it may also be shown that the above result extends to the time-inhomogeneous case in which $W(0) \notin [w_l^*, w_u^*]$ so long as the densities f_k are each continuous.

In a similar fashion to the transient analysis performed previously, the above results may now also be used to analyze the steady-state behavior of the virtual waiting time process $W^n = (W^n(t), t \geq 0)$ in original system. In particular, taking limits as t goes to ∞ in (55), one obtains by the above discussion the approximation

$$W^n(\infty) \approx w^* + n^{-1/2}\tilde{W}(\infty),$$

where $\tilde{W}(\infty)$ is a normal random variable with mean zero and variance $\tilde{\sigma}^2(0, \infty)$.

6. Closing remarks and extensions

In this paper, we have studied the overloaded, multiclass, $GI/GI/1+GI$ queue operating under the FIFO service discipline. In our first main result, Theorem 1, we provide a first order approximation to the transient behavior of the virtual waiting time process when the mean of the abandonment times is large relative to both the interarrival and service times. Our first order approximation is a first order ODE which may be solved for by the method of separation of variables. In our second main result, Theorem 2, after first centering the virtual waiting time process by its first order fluid limit and then normalizing by an appropriate constant, we have provided a second order stochastic approximation to the virtual waiting time process. Our stochastic approximation is a Ornstein-Uhlenbeck process with time varying drift and infinitesimal variance. Due to the tractability of the Ornstein-Uhlenbeck process, we are able to solve for its transient behavior explicitly in addition to showing that it tends to a weak limit as time tends to ∞ .

There are many directions for future research on this problem. We would first like to investigate the system operating under a different service discipline. In particular, it would perhaps be interesting to impose a specific cost structure on this problem and then determine under which service disciplines the system would be operating in a close to optimal manner with respect to the cost structure imposed. One potential candidate service discipline would be the static priority discipline. Under this service discipline, there may potentially be some job classes that do not abandon at all due to their high priority and others job classes that might never be processed at all because they are too impatient and have a low priority. An example of such jobs might be patients with superficial wounds that visit a very busy emergency room.

A second direction for future research would be to study the queue length process of this system. In general, the queue length process will not admit as simple a representation as the virtual waiting time process (3) and more sophisticated techniques might perhaps have to be employed.

Finally, it would be interesting to extend the results in this paper to the many-server heavy traffic regime. Recently, Whitt and several co-authors (?) (?) (17) have set forth conjectures

on the evolution of the queue length process in the overloaded many-server heavy traffic regime. We feel that the analysis presented in this paper could provide a first step in proving that these conjectures are indeed true.

7. Acknowledgements

The authors would like to thank Rishi Talreja for numerous helpful comments and suggestions on an earlier version of this paper.

Appendix. I

In the Appendix, we prove several propositions which were stated in the main body of the paper. We begin with the following technical lemma which is taken from a portion of the proof of Theorem 3.1 of Chapter 7 of (6).

Lemma 1 *Let $\{X_i, i \geq 1\}$ be an i.i.d. sequence of random variables with finite variance σ^2 . Then, for each $T \geq 0$,*

$$E \left[n^{-1/2} \sup_{i=1, \dots, nT} |X_i| \right] \rightarrow 0 \text{ as } n \rightarrow \infty.$$

PROOF OF LEMMA 1. We have

$$\begin{aligned} E \left[n^{-1/2} \sup_{i=1, \dots, nT} |X_i| \right] &= \int_0^\infty \mathbb{P} \left(n^{-1/2} \sup_{i=1, \dots, nT} |X_i| > x \right) dx \\ &\leq \varepsilon + \int_\varepsilon^\infty (nT) \mathbb{P}(|X_i| > n^{1/2}x) dx \\ &\leq \varepsilon + T\varepsilon^{-1} E[X_1^2 1(|X_1| > n^{1/2}\varepsilon)] \\ &\rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

This completes the proof.

Q.E.D.

We now present the proof of Proposition 3.

PROOF OF PROPOSITION 3. For this proof, we use the martingale central limit theorem as stated in Theorem 1.4 of Section 7 of (6) by which it is sufficient to show that for each $T \geq 0$,

$$E \left[\sup_{0 \leq t \leq T} |\bar{M}_{\nu,k}^n(t) - \bar{M}_{\nu,k}^n(t-)| \right] \rightarrow 0 \text{ as } n \rightarrow \infty,$$

and, for each $t \geq 0$, that $[\bar{M}_{\nu,k}^n, \bar{M}_{\nu,k}^n](t) \Rightarrow 0$ as $n \rightarrow \infty$.

Let $T \geq 0$ and note that by (12), (17) and Lemma 1 we have

$$\begin{aligned} E \left[\sup_{0 \leq t \leq T} |\bar{M}_{\nu,k}^n(t) - \bar{M}_{\nu,k}^n(t-)| \right] &= E \left[n^{-1} \sup_{i=1, \dots, nT} \left| \nu_k(i) - \frac{1}{\mu_k} \right| \right] \\ &\rightarrow 0 \text{ as } n \rightarrow \infty, \end{aligned}$$

and so that the first part of condition (a) of the theorem is satisfied.

Next, note that by Remark 1.5 below the theorem, it follows that

$$\begin{aligned} [\bar{M}_{\nu,k}^n, \bar{M}_{\nu,k}^n](t) &= \frac{1}{n^2} \sum_{i=1}^{\lfloor nt \rfloor} \left(\nu_k(i) - \frac{1}{\mu_k} \right)^2 \mathbf{1}_{\{d_k(i) > W^n(t_k^n(i))\}} \\ &\leq \frac{1}{n^2} \sum_{i=1}^{\lfloor nt \rfloor} \left(\nu_k(i) - \frac{1}{\mu_k} \right)^2 \\ &\Rightarrow 0 \text{ as } n \rightarrow \infty, \end{aligned}$$

where the last convergence follows since by the Weak Law of Large Numbers,

$$\frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} \left(\nu_k(i) - \frac{1}{\mu_k} \right)^2 \Rightarrow t \sigma_{s,k}^2 \text{ as } n \rightarrow \infty, \quad (56)$$

which completes the proof. Q.E.D.

We next present the proof of Proposition 4 which is similar in spirit to the proof of Proposition 3.

PROOF OF PROPOSITION 4. We again use the martingale central limit theorem as given by Theorem 1.4 of Section 7 of (6) by showing that for each $T \geq 0$,

$$E \left[\sup_{0 \leq t \leq T} |\bar{M}_{d,k}^n(t) - \bar{M}_{d,k}^n(t-)| \right] \rightarrow 0 \text{ as } n \rightarrow \infty,$$

and, for each $t \geq 0$, $[\bar{M}_{d,k}^n, \bar{M}_{d,k}^n](t) \Rightarrow 0$ as $n \rightarrow \infty$.

Let $T \geq 0$ and note that by (13) and (18) we have,

$$\begin{aligned} E \left[\sup_{0 \leq t \leq T} |\bar{M}_{d,k}^n(t) - \bar{M}_{d,k}^n(t-)| \right] &= \frac{1}{n\mu_k} E \left[\sup_{i=1, \dots, \lfloor nT \rfloor} \left| \mathbf{1}_{\{d_k(i) > W^n(t_k^n(i)-)\}} - G_k(W^n(t_k^n(i)-)) \right| \right] \\ &\leq \frac{1}{n\mu_k} \\ &\rightarrow 0 \text{ as } n \rightarrow \infty, \end{aligned}$$

so that that first part of condition (a) of Theorem 1.4 of Section 7 of (6) is shown.

Next, by Remark 1.5 below the theorem, we have that for each $t \geq 0$,

$$\begin{aligned} [\bar{M}_{d,k}^n, \bar{M}_{d,k}^n](t) &= \frac{1}{n^2\mu_k^2} \sum_{i=1}^{\lfloor nt \rfloor} \left(\mathbf{1}_{\{d_k(i) > W^n(t_k^n(i)-)\}} - G_k(W^n(t_k^n(i)-)) \right)^2 \\ &\leq \frac{t}{n\mu_k^2} \\ &\rightarrow 0 \text{ as } n \rightarrow \infty, \end{aligned}$$

which completes the proof.

Q.E.D.

We next present the proof of Proposition 5.

PROOF OF PROPOSITION 5. By Theorem 16.8 in Billingsley (3), we must verify that the following two conditions are satisfied:

1. For each $T \geq 0$,

$$\lim_{a \rightarrow \infty} \limsup_n P \left(\sup_{0 \leq t \leq T} W^n(t) \geq a \right) = 0.$$

2. For each $\varepsilon > 0$ and $T \geq 0$,

$$\lim_{\delta \rightarrow 0} \limsup_n P \left(\sup_{0 \leq t \leq T-\delta} w(W^n, [t, t+\delta]) \geq \varepsilon \right) = 0,$$

where, for $x \in D[0, \infty)$ and any set $S \subset [0, T]$,

$$w(x, S) = \sup_{u, v \in S} |x(u) - x(v)|.$$

Defining

$$X^n(t) = W^n(0) + \sum_{k=1}^K \sum_i^{A_k^n(t)} v_k^n(i)$$

for all $t \geq 0$, we have that $W^n(t) \leq X^n(t)$ for all $t \geq 0$. By the functional strong law of large numbers, Proposition 2 and the fact that $W^n(0) \rightarrow W(0)$ as $n \rightarrow \infty$ we have that, for all $t \geq 0$,

$$X^n(t) \rightarrow W(0) + \rho t, \quad \text{as } n \rightarrow \infty, \quad (57)$$

uniformly over compact intervals. It follows that, for each $T \geq 0$, since X^n is increasing,

$$\lim_{a \rightarrow \infty} \limsup_n P \left(\sup_{0 \leq t \leq T} W^n(t) \geq a \right) \leq \lim_{a \rightarrow \infty} \limsup_n P(X^n(T) \geq a) = 0,$$

and so condition 1 is proven.

As for condition 2, notice that W^n increases when the service times are added to the virtual waiting time, and decreases at most at rate 1 between arrivals. It follows then that for any $t \geq 0$ and δ ,

$$w(W^n, [t, t + \delta]) \leq w(X^n, [t, t + \delta]) + w(B^n, [t, t + \delta]) = X^n(t + \delta) - X^n(t) + \delta.$$

It follows from the uniform convergence of (57) that

$$\begin{aligned} & \lim_{\delta \rightarrow 0} \limsup_n P \left(\sup_{0 \leq t \leq T - \delta} w(W^n, [t, t + \delta]) \geq \varepsilon \right) \\ & \leq \lim_{\delta \rightarrow 0} \limsup_n P \left(\sup_{0 \leq t \leq T - \delta} X^n(t + \delta) - X^n(t) + \delta \geq \varepsilon \right) \\ & = 0, \end{aligned}$$

which completes the proof.

Q.E.D.

We next present the proof of Proposition 6.

PROOF OF PROPOSITION 6. Let $k \in \{1, \dots, K\}$. The expression in (14) may be equivalently expressed as

$$\begin{aligned} \varepsilon_k^n(t) &= \frac{1}{\mu_k} \left(\int_0^t G_k(W^n(s)) d\bar{A}_k^n(s) - \int_0^t G_k(W^n(s)) d(\lambda_k s) \right) \\ &\quad + \frac{1}{n\mu_k} \sum_{i=1}^{A_k^n(t)} (G_k(W^n(t_k^n(i)-)) - G_k(W^n(t_k^n(i)))) \end{aligned} \quad (58)$$

for each $n \geq 1$ and $t \geq 0$. We now show that each of the two terms on the righthand side of (58) above converge in distribution to 0 as n goes to ∞ , which completes the proof.

First note that since by Proposition 5 the sequence $(W^n, n \geq 1)$ is tight, it follows by Prohorov's Theorem (3) that $(W^n, n \geq 1)$ is relatively compact as well. Thus, for every sequence $\{n_k\}$, there exists a subsequence $\{n'_k\}$ along which $(W^n, n \geq 1)$ converges in distribution to some limit point W^* . Without loss of generality, we now relabel the sequence $\{n'_k\}$ by $\{n\}$ and note that since by Proposition 2, $\bar{A}_k^n \Rightarrow \lambda_k e$ as $n \rightarrow \infty$, where $\lambda_k e$ is a deterministic function, it follows by Theorem 3.9 of (3) that $(W_k^n, \bar{A}_k^n) \Rightarrow (W^*, \lambda_k e)$ as $n \rightarrow \infty$.

By the Skorohod Representation, there now exists an alternate probability space, $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$ on which are defined a sequence of processes $((\hat{W}^n, \hat{A}_k^n), n \geq 1)$ and a limit process $(\hat{W}^*, \lambda_k e)$ such that $(\hat{W}^n, \hat{A}_k^n) =^D (W^n, \bar{A}_k^n)$ for $n \geq 1$ and $(\hat{W}^*, \lambda_k e) =^D (W^*, \lambda_k e)$ and such that $(\hat{W}^n, \hat{A}_k^n) \rightarrow (\hat{W}^*, \lambda_k e)$, $\hat{\mathbb{P}}$ -a.s. as $n \rightarrow \infty$. Thus, since by assumption (A1), $G_k \in C_b(\mathbb{R})$, it follows by Lemma 8.1 in Dai and Dai (5), that, as $n \rightarrow \infty$,

$$\begin{aligned} &\int_0^t G_k(\hat{W}^n(s)) d\hat{A}_k^n(s) - \int_0^t G_k(\hat{W}^n(s)) d(\lambda_k s) \\ &\rightarrow \int_0^t G_k(\hat{W}^*(s)) d(\lambda_k s) - \int_0^t G_k(\hat{W}^*(s)) d(\lambda_k s) \\ &= 0, \end{aligned} \quad (59)$$

uniformly on compact sets, $\hat{\mathbb{P}}$ -a.s. However, since for each $n \geq 1$,

$$\begin{aligned} &\int_0^t G_k(\hat{W}^n(s)) d\hat{A}_k^n(s) - \int_0^t G_k(\hat{W}^n(s)) d(\lambda_k s) \\ &=^D \int_0^t G_k(W^n(s)) d\bar{A}_k^n(s) - \int_0^t G_k(W^n(s)) d(\lambda_k s), \end{aligned}$$

and almost sure convergence implies convergence in distribution, it follows by (59) that

$$\int_0^e G_k(W^n(s)) d\bar{A}_k^n(s) - \int_0^e G_k(W^n(s)) d(\lambda_k s) \Rightarrow 0 \text{ as } n \rightarrow \infty,$$

which, since the sequence $\{n'_k\}$ was arbitrary, shows the first term on the righthand side of (58) converges in distribution to 0 as n goes to ∞ .

Next, note that for each $T \geq 0$,

$$\begin{aligned} & \sup_{0 \leq t \leq T} \left| \frac{1}{n} \sum_{i=1}^{A_k^n(t)} (G_k(W^n(t_k^n(i)-)) - G_k(W^n(t_k^n(i)))) \right| \\ &= \frac{1}{n} \sum_{i=1}^{A_k^n(T)} (G_k(W^n(t_k^n(i)-)) - G_k(W^n(t_k^n(i)))) \\ &\leq \bar{A}_k^n(T) \max_{i=1, \dots, A_k^n(T)} (G_k(W^n(t_k^n(i)-)) - G_k(W^n(t_k^n(i)))) \\ &= \bar{A}_k^n(T) \max_{i=1, \dots, A_k^n(T)} (G_k(W^n(t_k^n(i)-)) - G_k(W^n(t_k^n(i)-) + v_k^n(i))) \\ &\leq \bar{A}_k^n(T) \max_{i=1, \dots, A_k^n(T)} \sup_{t \geq 0} (G_k(t) - G_k(t + v_k^n(i))) \\ &= \bar{A}_k^n(T) \sup_{t \geq 0} \left(G_k(t) - G_k \left(t + \max_{i=1, \dots, A_k^n(T)} v_k^n(i) \right) \right) \\ &\Rightarrow 0 \text{ as } n \rightarrow \infty, \end{aligned}$$

where the last convergence follows since by assumption (A1) the function G_k is continuous and by Lemma 3.3 of Iglehart and Whitt (9), Proposition 2 and the Random Time Change Theorem (3),

$$\max_{i=1, \dots, A_k^n(T)} v_k^n(i) = \frac{1}{n} \max_{i=1, \dots, n\bar{A}_k^n(T)} v_k(i) \Rightarrow 0 \text{ as } n \rightarrow \infty. \quad (60)$$

Thus,

$$\frac{1}{n} \sum_{i=1}^{A_k^n} (G_k(W^n(t_k^n(i)-)) - G_k(W^n(t_k^n(i)))) \Rightarrow 0 \text{ as } n \rightarrow \infty,$$

which, since the sequence $\{n'_k\}$ was arbitrary, completes the proof.

Q.E.D.

We next provide a proof of Proposition 7.

PROOF OF PROPOSITION 7. In order to show that the mapping Γ_h is continuous with respect to the Skorohod J_1 -metric, first suppose that $x_n \rightarrow x$ as $n \rightarrow \infty$ with respect to the Skorohod J_1 -metric and let $T \geq 0$ be a continuity point of x . It then follows that there exists a sequence $(\lambda^n, n \geq 1)$ of increasing, absolutely continuous, homeomorphisms of $[0, T]$ with derivatives $(\dot{\lambda}^n, n \geq 1)$ such that $\|x^n \circ \lambda^n - x\|_T \vee \|\dot{\lambda}^n - 1\|_T \rightarrow 0$ as $n \rightarrow \infty$. Also note that since by standard results, we have that for each $T \geq 0$, the sequence $\{\|x^n\|_T\}$ is bounded, it follows by the continuity of h that $\{\|h(x^n \circ \lambda^n)\|_T\}$ is bounded and hence there exists some $B \geq 0$ such that $\|h(x^n \circ \lambda^n)\|_T \leq B$ for sufficiently large n . We then have that for n large,

$$\begin{aligned}
 \|\gamma_h(x^n) \circ \lambda^n - \gamma_h(x)\|_T &= \left\| \int_0^{\lambda^n} h(x^n(t)) d(\lambda_k t) - \int_0^e h(x(t)) d(\lambda_k t) \right\|_T \\
 &= \left\| \int_0^e h(x^n(\lambda^n(t))) \dot{\lambda}^n(t) d(\lambda_k t) - \int_0^e h(x(t)) d(\lambda_k t) \right\|_T \\
 &\leq \left\| \int_0^e (h(x^n(\lambda^n(t))) - h(x(t))) d(\lambda_k t) \right\|_T \\
 &\quad + \left\| \int_0^e h(x^n(\lambda^n(t))) (\dot{\lambda}^n(t) - 1) d(\lambda_k t) \right\|_T \\
 &\leq \int_0^T |h(x^n(\lambda^n(t))) - h(x(t))| d(\lambda_k t) \\
 &\quad + \int_0^T h(x^n(\lambda^n(t))) |\dot{\lambda}^n(t) - 1| d(\lambda_k t) \\
 &\leq \int_0^T |h(x^n(\lambda^n(t))) - h(x(t))| d(\lambda_k t) + B \lambda_k(T) \|\dot{\lambda}^n - 1\|_T \\
 &\rightarrow 0 \text{ as } n \rightarrow \infty,
 \end{aligned}$$

where the final convergence follows by the Bounded Convergence Theorem (7) since by the continuity of h it follows that $|h(x^n(\lambda^n(t))) - h(x(t))| \rightarrow 0$ as $n \rightarrow \infty$ for each $t \geq 0$. Thus, $\|\gamma_h(x^n) \circ \lambda^n - \gamma_h(x)\|_T \vee \|\dot{\lambda}^n - 1\|_T \rightarrow 0$ as $n \rightarrow \infty$ and the proof is complete. Q.E.D.

We next provide a proof of Proposition 10.

PROOF OF PROPOSITION 10.

First note that in order to show that $\tilde{I}^n \Rightarrow 0$ as $n \rightarrow \infty$, it suffices to show that for each $T \geq 0$, $\sup_{0 \leq t \leq T} \tilde{I}^n(t) = \tilde{I}^n(T) \Rightarrow 0$ as $n \rightarrow \infty$. Moreover, since convergence in probability implies convergence

in distribution, it suffices to show that for each $\varepsilon > 0$, we have $P(\tilde{I}^n(T) < \varepsilon) \rightarrow 1$ as $n \rightarrow \infty$. Now note that since by (15) and (42), we have

$$\tilde{I}^n(t) = n^{1/2} \int_0^t 1_{(W^n(s)=0)} ds, \quad t \geq 0.$$

It follows that $P(\tilde{I}^n(T) < \varepsilon) \geq P(\tilde{I}^n(T) = 0) \geq P(\inf_{0 \leq t \leq T} W^n(t) > 0)$. By Theorem 1, $W^n \Rightarrow W$ as $n \rightarrow \infty$, where $W = (W(t), t \geq 0)$ is a continuous monotone function with $W(t) > 0$ for $t \geq 0$. Furthermore, the above convergence occurs in the topology of uniform convergence on compact sets. It follows that $P(\inf_{0 \leq t \leq T} W^n(t) > 0) \rightarrow 1$ as $n \rightarrow \infty$, which completes the proof. Q.E.D.

We now provide a proof of Proposition 11.

PROOF OF PROPOSITION 11. For each $k \in \{1, \dots, K\}$, we have $\tilde{M}_{\nu,k}^n(t) = \hat{M}_{\nu,k}^n(t) + \check{M}_{\nu,k}^n(t)$, where we define $\check{M}_{\nu,k}^n(t) = \tilde{M}_{\nu,k}^n(t) - \hat{M}_{\nu,k}^n(t)$. By (45) it is clear that $\hat{M}_{\nu,k}^n = (\hat{M}_{\nu,k}^n(t), t \geq 0)$ is clearly a martingale with respect to the filtration $\mathcal{F}^{n,k} = (\mathcal{F}_{\lfloor nt \rfloor}^k, t \geq 0)$. It follows that $\check{M}_{\nu,k}^n = (\check{M}_{\nu,k}^n(t), t \geq 0)$, being the difference of two martingales, is a martingale with respect to $\mathcal{F}^{n,k}$ as well.

The proof now proceeds in two parts. First, we show that $\hat{M}_{\nu,k}^n \Rightarrow \tilde{M}_{\nu,k}$ as $n \rightarrow \infty$, where $\tilde{M}_{\nu,k}$ is the Gaussian process given in the statement of the proposition. Next, we show that $\check{M}_{\nu,k}^n \Rightarrow 0$ as $n \rightarrow \infty$, whereby, from the decomposition above and Theorem 11.4.8 of (15), we obtain the joint convergence $(\tilde{M}_{\nu,k}^n, \hat{M}_{\nu,k}^n) \Rightarrow (\tilde{M}_{\nu,k}, \tilde{M}_{\nu,k})$ as $n \rightarrow \infty$, which completes the proof.

As in the proofs of Proposition 3 and 4, we again use the martingale central limit theorem of Theorem 1.4 of Chapter 7 of Ethier and Kurtz (6). First note that in a manner similar to the proof of Proposition 3, we have by Lemma 1 above that

$$E \left[\sup_{0 \leq t \leq T} \left| \hat{M}_{\nu,k}^n(t) - \hat{M}_{\nu,k}^n(t-) \right| \right] \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

and so the first part of condition (a) of the Theorem from Ethier and Kurtz (6) is shown.

We require a renumbering of the jobs. Recall that the abandonment time of the i^{th} class k job is typically compared to the waiting time. However, for the process $\hat{W}_{\nu,k}$, we compare the abandonment time with the limiting waiting time process, evaluated at the approximate arrival time, $W(i/(n\lambda_k))$. Let $\hat{j}_k^n(i)$ denote the index of the i^{th} class k job whose abandonment time is greater than the limiting waiting time process evaluated at the job's approximate arrival time. Now note that by Remark 1.5 below the statement of Theorem 1.4 of Chapter 7 of (6), we have that

$$\begin{aligned}
[\hat{M}_{\nu,k}^n, \hat{M}_{\nu,k}^n](t) &= n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \left(\nu_k(i) - \frac{1}{\mu_k} \right)^2 \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} \\
&= n^{-1} \sum_{i=1}^{\hat{R}_k^n(\lfloor nt \rfloor)} \left(\nu_k(\hat{j}_k^n(i)) - \frac{1}{\mu_k} \right)^2 \\
&= n^{-1} \sum_{i=1}^{n\tilde{R}_k^n(t)} \left(\nu_k(\hat{j}_k^n(i)) - \frac{1}{\mu_k} \right)^2 \\
&\Rightarrow \sigma_{s,k}^2 \int_0^{t/\lambda_k} G_k(W(s)) d(\lambda_k s) \quad \text{as } n \rightarrow \infty, \tag{61}
\end{aligned}$$

where the final convergence above follows by the Functional Strong Law of Large Numbers, Proposition 9, the Random Time Change Theorem (3) and the fact that $(\nu_k(\hat{j}_k^n(i)), i \geq 1)$ is an i.i.d. sequence of random variables with mean $1/\mu_k$ and variance $\sigma_{s,k}^2$. However, since convergence in distribution to a constant implies convergence in probability, it now follows by Theorem 1.4 of Chapter 7 of Ethier and Kurtz (6) that $\hat{M}_{\nu,k}^n \Rightarrow \tilde{M}_{\nu,k}$ as $n \rightarrow \infty$, where $\tilde{M}_{\nu,k}$ is a Gaussian process as stated in the proposition.

We now proceed to show that $\tilde{M}_{\nu,k}^n \Rightarrow 0$ as $n \rightarrow \infty$. First note that, as in the proof of Proposition 3, we have by Lemma 1 above that

$$E \left[\sup_{0 \leq t \leq T} |\tilde{M}_{\nu,k}^n(t) - \tilde{M}_{\nu,k}^n(t-)| \right] \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

and so the first part of condition (a) of the Theorem from Ethier and Kurtz is shown. Furthermore, we have

$$\begin{aligned}
&[\tilde{M}_{\nu,k}^n, \tilde{M}_{\nu,k}^n](t) \\
&= n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \left(\nu_k(i) - \frac{1}{\mu_k} \right)^2 \left(\mathbf{1}_{\{d_k(i) > W^n(t_k^n(i-))\}} - \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} \right)^2
\end{aligned}$$

$$\begin{aligned}
&= \left[\hat{M}_{\nu,k}^n, \hat{M}_{\nu,k}^n \right] (t) \\
&\quad - 2n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \left(\nu_k(i) - \frac{1}{\mu_k} \right)^2 \mathbf{1}_{\{d_k(i) > \max(W^n(t_k^n(i)-), W(i/(n\lambda_k)))\}} \\
&\quad + n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \left(\nu_k(i) - \frac{1}{\mu_k} \right)^2 \mathbf{1}_{\{d_k(i) > W^n(t_k^n(i)-)\}}
\end{aligned} \tag{62}$$

Now note that by Remark 1.5 below the statement of Theorem 1.4 of Chapter 7 of (6), we have, letting $\tilde{j}_k^n(i)$ be the index of the i^{th} class k job that does not abandon, that

$$\begin{aligned}
&n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \left(\nu_k(i) - \frac{1}{\mu_k} \right)^2 \mathbf{1}_{\{d_k(i) > W^n(i/(n\lambda_k))\}} \\
&= n^{-1} \sum_{i=1}^{R_k^n(\lfloor nt \rfloor)} \left(\nu_k(\tilde{j}_k^n(i)) - \frac{1}{\mu_k} \right)^2 \\
&= n^{-1} \sum_{i=1}^{n\bar{R}_k^n(t)} \left(\nu_k(\tilde{j}_k^n(i)) - \frac{1}{\mu_k} \right)^2 \\
&\Rightarrow \sigma_{s,k}^2 \int_0^{t/\lambda_k} G_k(W(s)) d(\lambda_k s) \text{ as } n \rightarrow \infty,
\end{aligned} \tag{63}$$

where the final convergence above follows by the Functional Strong Law of Large Numbers, Proposition 8, the Random Time Change Theorem (3) and the fact that $(\nu_k(\tilde{j}_k^n(i)), i \geq 1)$ is an i.i.d. sequence of random variables with mean $1/\mu_k$ and variance $\sigma_{s,k}^2$.

By similar arguments, and letting $\check{j}_k^n(i)$ be the index of the i^{th} class k job whose deadline upon arrival exceeds both the virtual waiting time (evaluated at the time of arrival) as well as the limiting waiting time (evaluated at the approximate arrival time), we have

$$\begin{aligned}
&n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \left(\nu_k(i) - \frac{1}{\mu_k} \right)^2 \mathbf{1}_{\{d_k(i) > \max(W^n(t_k^n(i)-), W(i/(n\lambda_k)))\}} \\
&= n^{-1} \sum_{i=1}^{\check{R}_k^n(\lfloor nt \rfloor)} \left(\nu_k(\check{j}_k^n(i)) - \frac{1}{\mu_k} \right)^2 \\
&= n^{-1} \sum_{i=1}^{n\check{R}_k^n(t)} \left(\nu_k(\check{j}_k^n(i)) - \frac{1}{\mu_k} \right)^2 \\
&\Rightarrow \sigma_{s,k}^2 \int_0^{t/\lambda_k} G_k(W(s)) d(\lambda_k s) \text{ as } n \rightarrow \infty.
\end{aligned} \tag{64}$$

Hence, by (33), (61), (62), (63) and (64), we have that as $n \rightarrow \infty$,

$$[\check{M}_{\nu,k}^n, \check{M}_{\nu,k}^n](t) \Rightarrow 0.$$

Since convergence in distribution to a constant implies convergence in probability, it follows by Theorem 1.4 of Chapter 7 of Ethier and Kurtz (6) that $\check{M}_{\nu,k}^n \Rightarrow 0$ as $n \rightarrow \infty$, which completes the proof. Q.E.D.

We next provide the proof of Proposition 12 which is similar to the proof of Proposition 11.

PROOF OF PROPOSITION 12. Let $k \in \{1, \dots, K\}$ and first note that as in the proof of Proposition 11, we have the representation $\check{M}_{d,k}^n(t) = \hat{M}_{d,k}^n(t) + \check{M}_{d,k}^n(t)$ for $t \geq 0$, where we set $\check{M}_{d,k}^n(t) = \tilde{M}_{d,k}^n(t) - \hat{M}_{d,k}^n(t)$. Also note that $\check{M}_{d,k}^n = (\check{M}_{d,k}^n(t), t \geq 0)$ is clearly a martingale with respect to the filtration $\mathcal{F}^{n,k} = (\mathcal{F}_{[nt]}^k, t \geq 0)$ since it is the difference of two martingales with respect to $\mathcal{F}^{n,k}$.

We now proceed to show that $\hat{M}_{d,k}^n \Rightarrow \tilde{M}_{d,k}$ as $n \rightarrow \infty$ and $\check{M}_{d,k}^n \Rightarrow 0$ as $n \rightarrow \infty$, where $\tilde{M}_{d,k}$ is the Gaussian process as given in the statement of the proposition. From the decomposition above and Theorem 11.4.8 of (15), this then implies the joint convergence $(\check{M}_{d,k}^n, \hat{M}_{d,k}^n) \Rightarrow (\tilde{M}_{d,k}, \tilde{M}_{d,k})$ as $n \rightarrow \infty$, which completes the proof.

In order to prove the two claims above, we again use the martingale central limit theorem of Theorem 1.4 of Chapter 7 of Ethier and Kurtz (6). First note that in a manner similar to the proof of Proposition 4 we have

$$E \left[\sup_{0 \leq t \leq T} \left| \hat{M}_{d,k}^n(t) - \hat{M}_{d,k}^n(t-) \right| \right] \rightarrow 0 \text{ as } n \rightarrow \infty.$$

The first part of condition (a) of Theorem 1.4 of Chapter 7 of Ethier and Kurtz (6) is satisfied.

Next, recalling the definition of \hat{R}_k^n in (36) and $\bar{R}_k^n(t) = n^{-1} \hat{R}_k^n(t)$, we have that

$$\begin{aligned} \mu_k^2 [\hat{M}_{d,k}^n, \hat{M}_{d,k}^n](t) &= n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} (\mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} - G_k(W(i/(n\lambda_k))))^2 \\ &= n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} (\mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} - 2 \cdot \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} G_k(i/(n\lambda_k))) \\ &\quad + G_k(W(i/(n\lambda_k)))^2 \\ &= \bar{R}_k^n(t) - 2 \int_0^t G_k(W(s/\lambda_k)) d\bar{R}_k^n(s) + \int_0^{\lambda^{-1}t} G_k(W(s))^2 d\bar{A}_k^n(s). \end{aligned} \tag{65}$$

By Proposition 9, (35) and twice applying Lemma 8.1 in Dai and Dai (5) we have that, as $n \rightarrow \infty$,

$$\int_0^t G_k(W(s/\lambda_k)) d\tilde{R}_k^n(s) \Rightarrow \int_0^t G_k(W(s/\lambda_k)) d\bar{R}_k(s) = \int_0^t G_k(W(s/\lambda_k))^2 ds$$

and

$$\int_0^{\lambda^{-1}t} G_k(W(s))^2 d\tilde{A}_k^n(s) \Rightarrow \int_0^{\lambda^{-1}t} G_k(W(s))^2 d(\lambda_k s) = \int_0^t G_k(W(s/\lambda_k))^2 ds.$$

By Proposition 9, and (65), it follows that as $n \rightarrow \infty$

$$\begin{aligned} \mu_k^2 [\hat{M}_{d,k}^n, \hat{M}_{d,k}^n] (t) &\Rightarrow \int_0^t G_k(W(s/\lambda_k)) ds - 2 \int_0^t G_k(W(s/\lambda_k))^2 ds + \int_0^t G_k(W(s))^2 ds \\ &= \int_0^t G_k(W(s/\lambda_k))(1 - G_k(W(s/\lambda_k))) ds \\ &= \int_0^t G_k(W(s/\lambda_k)) F_k(W(s/\lambda_k)) ds. \end{aligned}$$

Since the limiting process on the righthand side above is deterministic, it now follows that

$$[\hat{M}_{d,k}^n, \hat{M}_{d,k}^n] (t) \xrightarrow{P} \mu_k^{-2} \int_0^t G_k(W(s/\lambda_k)) F_k(W(s/\lambda_k)) ds \text{ as } n \rightarrow \infty,$$

and hence, by Theorem 1.4 of Chapter 7 of Ethier and Kurtz (6), we have that $\hat{M}_{d,k}^n \Rightarrow \tilde{M}_{d,k}$ as $n \rightarrow \infty$, where $\tilde{M}_{d,k}$ is the Gaussian process as stated in the proposition.

We now show that $\check{M}_{d,k}^n \Rightarrow 0$ as $n \rightarrow \infty$, which completes the proof. First, as in the previous portion of the proof above, it follows directly that for each $T \geq 0$,

$$E \left[\sup_{0 \leq t \leq T} |\check{M}_{d,k}^n(t) - \check{M}_{d,k}^n(t-)| \right] \rightarrow 0 \text{ as } n \rightarrow \infty,$$

and so the first part of condition (a) of Theorem 1.4 of Chapter 7 of Ethier and Kurtz is shown.

Next, we have for $t \geq 0$

$$\begin{aligned} [\check{M}_{d,k}^n, \check{M}_{d,k}^n] (t) &= \frac{1}{n\mu_k^2} \sum_{i=1}^{\lfloor nt \rfloor} \left(\left(\mathbf{1}_{\{d_k(i) > W^n(t_k^n(i)-)\}} - G_k(W^n(t_k^n(i)-)) \right) \right. \\ &\quad \left. - \left(\mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} - G_k(W(i/(n\lambda_k))) \right) \right)^2 \\ &= \frac{1}{n\mu_k^2} \sum_{i=1}^{\lfloor nt \rfloor} \left(\mathbf{1}_{\{d_k(i) > W^n(t_k^n(i)-)\}} - \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} \right)^2 \\ &\quad + \frac{1}{n\mu_k^2} \sum_{i=1}^{\lfloor nt \rfloor} \left(G_k(W^n(t_k^n(i)-)) - G_k(W(i/(n\lambda_k))) \right)^2 \end{aligned} \tag{66}$$

$$-\frac{2}{n\mu_k^2} \sum_{i=1}^{\lfloor nt \rfloor} \left(\mathbf{1}_{\{d_k(i) > W^n(t_k^n(i)-)\}} - \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} \right) \times \\ (G_k(W^n(t_k^n(i)-)) - G_k(W(i/(n\lambda_k))))).$$

We now show that each of the three summations on the righthand side of (66) above converges in distribution to 0 as n goes to ∞ , which will complete the proof.

First note that for $t \geq 0$,

$$\begin{aligned} & n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \left(\mathbf{1}_{\{d_k(i) > W^n(t_k^n(i)-)\}} - \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} \right)^2 \\ &= n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \mathbf{1}_{\{d_k(i) > W^n(t_k^n(i)-)\}} + n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} \\ &\quad - n^{-1} 2 \sum_{i=1}^{\lfloor nt \rfloor} \mathbf{1}_{\{d_k(i) > \max(W^n(t_k^n(i)-), W(i/(n\lambda_k)))\}} \\ &= \bar{R}_k^n(t) + \tilde{R}_k^n(t) - 2\check{R}_k^n(t) \\ &\Rightarrow 0, \end{aligned} \tag{67}$$

where convergence follows from Proposition 8 and Proposition 9. We have shown that the first sum on the righthand side of (66) converges in distribution to 0 as n goes to ∞ .

By Theorem 1, we have that $W^n \Rightarrow W$ as $n \rightarrow \infty$, where convergence is uniform over compact sets and the limit process W is continuous. It follows that for each $t > 0$ and $\delta > 0$, there exists a K_δ such that $\mathbb{P}(\sup_{0 \leq s \leq t} W^n(s) \geq K_\delta) < \delta$ for n sufficiently large. By Proposition 2 we have that $t_k^n(\lfloor ne \rfloor) \rightarrow \lambda_k^{-1}$ almost surely as $n \rightarrow \infty$. Moreover, the convergence is uniform over compact intervals, so that $\mathbb{P}(\sup_{i \leq \lfloor nt \rfloor} |t_k^n(i) - i/(n\lambda_k)| > \delta) < \delta$, for sufficiently large n . Thus, for n sufficiently large, for each $\varepsilon > 0$ and $\delta > 0$,

$$\begin{aligned} & \mathbb{P} \left(n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} (G_k(W^n(t_k^n(i)-)) - G_k(W(i/(n\lambda_k))))^2 > \varepsilon \right) \\ & \leq \delta + \mathbb{P} \left(n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} |G_k(W^n(t_k^n(i)-)) - G_k(W(i/(n\lambda_k)))| > \sqrt{\varepsilon} \left| \sup_{0 \leq s \leq t} W^n(s) \leq K_\delta \right. \right) \\ & \leq 2\delta + \mathbb{P} \left(n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \left(\sup_{0 \leq s \leq K_\delta} |f_k(s)| \right) |W^n(t_k^n(i)-) - W(i/(n\lambda_k))| > \sqrt{\varepsilon} \left| \sup_{i \leq \lfloor nt \rfloor} |t_k^n(i) - i/(n\lambda_k)| > \delta \right. \right) \end{aligned} \tag{68}$$

$$\begin{aligned}
&\leq 2\delta + \mathbb{P} \left(t \sup_{s \leq \delta + t/\lambda_k} |W^n(s-) - W(s)| > \sqrt{\varepsilon}/2 \sup_{0 \leq s \leq K_\delta} |f_k(s)| \right) \\
&\quad + \mathbb{P} \left(t \sup_{s \leq t/\lambda_k, s_0 \leq \delta} |W(s) - W(s + s_0)| > \sqrt{\varepsilon}/2 \sup_{0 \leq s \leq K_\delta} |f_k(s)| \right). \\
&\rightarrow 2\delta,
\end{aligned}$$

where convergence in the last line follows from the uniform convergence of W^n and the continuity of W . Thus, since the choice of ε and δ above was arbitrary, this then implies that

$$n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} (G_k(W^n(t_k^n(i)-)) - G_k(W(i/(n\lambda_k))))^2 \xrightarrow{P} 0 \text{ as } n \rightarrow \infty,$$

and so, since convergence in probability implies convergence in distribution, we then have that second summation on the righthand side of (66) above converges in distribution to 0 as n tends to ∞ .

Finally, note that

$$\begin{aligned}
&|2n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} (\mathbf{1}_{\{d_k(i) > W^n(t_k^n(i)-)\}} - \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}}) (G_k(W^n(t_k^n(i)-)) - G_k(W(i/(n\lambda_k))))| \\
&\leq 2n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \left| \mathbf{1}_{\{d_k(i) > W^n(t_k^n(i)-)\}} - \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} \right| |G_k(W^n(t_k^n(i)-)) - G_k(W(i/(n\lambda_k)))| \\
&\leq 2n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \left| \mathbf{1}_{\{d_k(i) > W^n(t_k^n(i)-)\}} - \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} \right| \\
&\Rightarrow 0 \text{ as } n \rightarrow \infty,
\end{aligned}$$

where the final convergence follows by (67) above. Thus, the third summation on the righthand side of (66) converges in distribution to 0 as n tend to ∞ .

It now follows by (66) that $[\check{M}_{d,k}^n, \check{M}_{d,k}^n](t) \Rightarrow 0$ as $n \rightarrow \infty$ and so, since convergence in distribution to a constant implies convergence in probability, by Theorem 1.4 of Chapter 7 of Ethier and Kurtz (6) we have that $\check{M}_{d,k}^n \Rightarrow 0$ as $n \rightarrow \infty$, which completes the proof.

Q.E.D.

We next prove Proposition 13 which extends Propositions 11 and 12, demonstrating that the convergences therein occur jointly.

PROOF OF PROPOSITION 13. Let $k \in \{1, \dots, K\}$, we will show that $(\hat{M}_{\nu,k}^n, \hat{M}_{d,k}^n) \Rightarrow (\tilde{M}_{\nu,k}, \tilde{M}_{d,k})$ as $n \rightarrow \infty$, where $\tilde{M}_{\nu,k}$ and $\tilde{M}_{d,k}$ are independent. The individual convergences $\hat{M}_{\nu,k}^n \rightarrow \tilde{M}_{\nu,k}$ as $n \rightarrow \infty$ and $\hat{M}_{d,k}^n \rightarrow \tilde{M}_{d,k}$ as $n \rightarrow \infty$ are given by Proposition 11 and Proposition 12, respectively. To show joint convergence to independent processes, by Theorem 1.4 of Chapter 7 of Ethier and Kurtz (6), it suffices to show that, as $n \rightarrow \infty$,

$$\left[\hat{M}_{\nu,k}^n, \hat{M}_{d,k}^n \right] (t) \rightarrow^P 0. \quad (69)$$

First note that

$$\begin{aligned} & \left[\hat{M}_{\nu,k}^n, \hat{M}_{d,k}^n \right] (t) \quad (70) \\ &= n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \left(\nu_k(i) - \frac{1}{\mu_k} \right) \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} \left(\mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} - G_k(W(i/(n\lambda_k))) \right) \\ &= n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \left(\nu_k(i) - \frac{1}{\mu_k} \right) \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} F_k(W(i/(n\lambda_k))) \\ &= n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \nu_k(i) \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} F_k(W(i/(n\lambda_k))) - \frac{1}{\mu_k} \int_0^t F_k(W(s/\lambda_k)) d\tilde{R}_k^n(s). \quad (71) \end{aligned}$$

As for the last term on the righthand side of (71) we have, by Lemma 8.3 of Dai and Dai (5) and Proposition 9, that, as $n \rightarrow \infty$,

$$\int_0^t F_k(W(s/\lambda_k)) d\tilde{R}_k^n(s) \rightarrow^P \int_0^t F_k(W(s/\lambda_k)) d\bar{R}_k(s) = \int_0^t F_k(W(s/\lambda_k)) G_k(W(s/\lambda_k)) ds. \quad (72)$$

As for the first term on the righthand side of (71), we split $\{1, \dots, \lfloor nt \rfloor\}$ into subsets over which the deviations in $W(\cdot/(n\lambda_k))$ are arbitrarily small. For any arbitrary $\epsilon > 0$ define

$$H_1^n(\epsilon) = \sum_{j=0}^{\lfloor \epsilon^{-1} \rfloor} F_k(W(j\epsilon t/\lambda_k)) H_1^n(\epsilon, j) \quad \text{and} \quad H_2^n(\epsilon) = \sum_{j=0}^{\lfloor \epsilon^{-1} \rfloor} F_k(W((j+1)\epsilon t/\lambda_k)) H_2^n(\epsilon, j),$$

where

$$H_1^n(\epsilon, j) = \frac{1}{n} \sum_{i=\lfloor j\epsilon nt \rfloor + 1}^{\lfloor \min((j+1)\epsilon, 1)nt \rfloor} \nu_k(i) \mathbf{1}_{\{d_k(i) > W((j+1)\epsilon t/\lambda_k)\}}, \quad j = 0, 1, \dots, \lfloor \epsilon^{-1} \rfloor,$$

and

$$H_2^n(\epsilon, j) = \frac{1}{n} \sum_{i=\lfloor j\epsilon nt \rfloor + 1}^{\lfloor \min((j+1)\epsilon, 1)nt \rfloor} \nu_k(i) \mathbf{1}_{\{d_k(i) > W(j\epsilon t/\lambda_k)\}}, \quad j = 0, 1, \dots, \lfloor \epsilon^{-1} \rfloor.$$

Assume for now that W is non-decreasing. For any $j \in \{0, \dots, \lfloor \epsilon^{-1} \rfloor\}$ and $i \in \{\lfloor j\epsilon n t \rfloor + 1, \dots, \lfloor \min((j+1)\epsilon, 1)nt \rfloor\}$ we note that, because F_k is increasing,

$$F_k(W(j\epsilon t/\lambda_k)) \leq F_k(W(i/(n\lambda_k))) \leq F_k(W((j+1)\epsilon t/\lambda_k))$$

and

$$\mathbf{1}_{\{d_k(i) > W((j+1)\epsilon t/\lambda_k)\}} \leq \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} \leq \mathbf{1}_{\{d_k(i) > W(j\epsilon t/\lambda_k)\}}.$$

It follows then that

$$H_1^n(\epsilon) \leq n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \nu_k(i) \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} F_k(W(i/(n\lambda_k))) \leq H_2^n(\epsilon).$$

If W is non-increasing then the inequalities are reversed. Without loss of generality, assume that W is non-decreasing.

It follows from the Functional Weak Law of Large Numbers, the Random Time Change Theorem (3) and the fact that the summands of $H_1^n(\epsilon, j)$ are i.i.d. with mean $\mu_k^{-1} G_k(W((j+1)\epsilon t/\lambda_k))$, that as $n \rightarrow \infty$, $H_1^n(\epsilon, j) \rightarrow^P \epsilon t \mu_k^{-1} G_k(W((j+1)\epsilon t/\lambda_k))$ for $j < \lfloor \epsilon^{-1} \rfloor$, and $H_1^n(\epsilon, j) \rightarrow^P (1 - \epsilon \lfloor \epsilon^{-1} \rfloor) t \mu_k^{-1} G_k(W((j+1)\epsilon t/\lambda_k))$ for $j = \lfloor \epsilon^{-1} \rfloor$. Moreover, we have that, as $n \rightarrow \infty$,

$$\begin{aligned} H_1^n(\epsilon) &\rightarrow^P H_1(\epsilon) \\ &= \epsilon t \mu_k^{-1} \sum_{j=0}^{\lfloor \epsilon^{-1} \rfloor - 1} F_k(W(j\epsilon t/\lambda_k)) G_k(W((j+1)\epsilon t/\lambda_k)) \\ &\quad + (1 - \epsilon \lfloor \epsilon^{-1} \rfloor) t \mu_k^{-1} F_k(W(\epsilon \lfloor \epsilon^{-1} \rfloor t/\lambda_k)) G_k(W((\epsilon \lfloor \epsilon^{-1} \rfloor + 1)t/\lambda_k)). \end{aligned}$$

One can likewise show that

$$\begin{aligned} H_2^n(\epsilon) &\rightarrow^P H_2(\epsilon) \\ &= \epsilon t \mu_k^{-1} \sum_{j=1}^{\lfloor \epsilon^{-1} \rfloor} F_k(W((j+1)\epsilon t/\lambda_k)) G_k(W(j\epsilon t/\lambda_k)) \\ &\quad + (1 - \epsilon \lfloor \epsilon^{-1} \rfloor) t \mu_k^{-1} F_k(W(t/\lambda_k)) G_k(W(\epsilon \lfloor \epsilon^{-1} \rfloor t/\lambda_k)). \end{aligned}$$

Finally because

$$\lim_{\epsilon \downarrow 0} H_1(\epsilon) = \lim_{\epsilon \downarrow 0} H_2(\epsilon) = \frac{1}{\mu_k} \int_0^t F_k(W(s/\lambda_k)) G_k(W(s/\lambda_k)) ds$$

it follows that

$$n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} \nu_k(i) \mathbf{1}_{\{d_k(i) > W(i/(n\lambda_k))\}} F_k(W(i/(n\lambda_k))) \rightarrow^P \frac{1}{\mu_k} \int_0^t F_k(W(s/\lambda_k)) G_k(W(s/\lambda_k)) ds. \quad (73)$$

By (71), (72) and (73) we have shown that (69) holds. This concludes the proof. Q.E.D.

We next prove Proposition 14.

PROOF OF PROPOSITION 14. Let $k \in \{1, \dots, K\}$ and note that integrating by parts, we obtain the representations

$$\tilde{\varepsilon}_k^n(t) = \frac{1}{\mu_k} \tilde{A}_k^n(t) G_k(W^n(t)) - \frac{1}{\mu_k} \int_0^t \tilde{A}_k^n(s) dG_k(W^n(s)), \quad t \geq 0, \quad (74)$$

and

$$\tilde{\varepsilon}_k^n(t) = \frac{1}{\mu_k} \tilde{A}_k^n(t) G_k(W(t)) - \frac{1}{\mu_k} \int_0^t \tilde{A}_k^n(s) dG_k(W(s)), \quad t \geq 0. \quad (75)$$

Next, since by Theorem 1, $W^n \Rightarrow W$ as $n \rightarrow \infty$, where W is a deterministic process and furthermore, since by Donsker's Theorem for renewal processes, $\tilde{A}_k^n \Rightarrow \tilde{A}_k$ as $n \rightarrow \infty$, where \tilde{A}_k is a Brownian motion with infinitesimal variance $\sigma_{a,k}^2 \lambda_k^3$, it follows by Theorem 3.9 of (3) that we have the joint convergence $(\tilde{A}_k^n, W^n) \Rightarrow (\tilde{A}_k, W)$ as $n \rightarrow \infty$.

By the Skorohod Representation Theorem (15), there now exists an alternate probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$ on which are defined a sequence of processes $((\hat{A}_k^n, \hat{W}^n), n \geq 1)$ such that $(\hat{A}_k^n, \hat{W}^n) =^D (\tilde{A}_k^n, W^n)$ for $n \geq 1$ and where $(\hat{A}_k^n, \hat{W}^n) \rightarrow (\hat{A}_k, \hat{W})$, $\hat{\mathbb{P}}$ -a.s. as $n \rightarrow \infty$. Setting

$$\hat{\varepsilon}_k^n(t) = \frac{1}{\mu_k} \left(\hat{A}_k^n(t) G_k(\hat{W}^n(t)) - \int_0^t \hat{A}_k^n(s) dG_k(\hat{W}^n(s)) \right), \quad t \geq 0, \quad (76)$$

and

$$\hat{\varepsilon}_k^n = \frac{1}{\mu_k} \left(\hat{A}_k^n(t) G_k(\hat{W}(t)) - \int_0^t \hat{A}_k^n(s) dG_k(\hat{W}(s)) \right), \quad t \geq 0, \quad (77)$$

we now claim that

$$\hat{\varepsilon}_k^n(t) \rightarrow \frac{1}{\mu_k} \left(\hat{A}_k(t) G_k(\hat{W}(t)) - \int_0^t \hat{A}_k(s) dG_k(\hat{W}(s)) \right) \quad (78)$$

and

$$\hat{\varepsilon}_k^n(t) - \varepsilon_k^n(t) \rightarrow 0,$$

uniformly on compact sets, \mathbb{P} -a.s. as $n \rightarrow \infty$, which, since $(\varepsilon_k^n, \hat{\varepsilon}_k^n) =^D (\tilde{\varepsilon}_k^n, \hat{\varepsilon}_k^n)$ implies that $(\tilde{\varepsilon}_k^n, \hat{\varepsilon}_k^n) \Rightarrow (\tilde{\varepsilon}_k, \hat{\varepsilon}_k)$ as $n \rightarrow \infty$, which will complete the proof.

Let $TV(G_k(W(\cdot)), [0, T])$ denote the total variation of $G_k(W(\cdot))$ on $[0, T]$. We then have by assumption (A2) and the form of the virtual waiting time process given by (3) that

$$\begin{aligned} TV(G_k(W^n(\cdot)), [0, T]) &= \sup_{P \in \mathcal{P}} \sum_{i=0}^{n_p-1} |G_k(W^n(t_{i+1})) - G_k(W^n(t_i))| \\ &\leq \sup_{P \in \mathcal{P}} \sum_{i=0}^{n_p-1} \sup_{0 \leq u \leq W^n(T)} f_k(u) |W^n(t_{i+1}) - W^n(t_i)| \\ &\leq \sup_{0 \leq u \leq W^n(T)} f_k(u) \sup_{P \in \mathcal{P}} \sum_{i=0}^{n_p-1} |W^n(t_{i+1}) - W^n(t_i)| \\ &\leq \sup_{0 \leq u \leq W^n(T)} f_k(u) \left(T + \sum_{i=1}^K \sum_{i=1}^{A_k^n(T)} v_k^n(i) \right) \\ &= \sup_{0 \leq u \leq W^n(T)} f_k(u) \left(T + \sum_{i=1}^K \frac{1}{n} \sum_{i=1}^{n\bar{A}_k^n(T)} v_k(i) \right), \end{aligned} \quad (79)$$

where the supremum is taken over all partitions $P = \{[0, t_1], [t_1, t_2], \dots, [t_{n_p-1}, t]\}$ of $[0, T]$. However, since by Theorem 1, $W^n(T) \rightarrow W(T)$ as $n \rightarrow \infty$ and since by the Functional Weak Law of Large Numbers, Proposition 2, and the Random Time Change Theorem (3), we have that

$$\sum_{i=1}^K \sum_{i=1}^{A_k^n(T)} v_k^n(i) = \sum_{i=1}^K \frac{1}{n} \sum_{i=1}^{n\bar{A}_k^n(T)} v_k(i) \Rightarrow \sum_{i=1}^K \mu_k(\lambda_k t), \quad \text{as } n \rightarrow \infty,$$

it follows that the term on the righthand side of (79) is \mathbb{P} -a.s. bounded as $n \rightarrow \infty$. Furthermore, since the function G_k is bounded and by Assumption (A1) continuous, it follows by a slight modification of Lemma 8.3 of Dai and Dai (5) – i.e., one that accounts for functions of bounded variation as opposed to simply increasing functions – that the convergence in (78) holds uniformly on compact sets, \mathbb{P} -a.s. as $n \rightarrow \infty$. It follows then that $\hat{\varepsilon}_k^n(t) - \varepsilon_k^n(t) \rightarrow 0$, uniformly on compact sets, \mathbb{P} -a.s. as $n \rightarrow \infty$, which completes the proof.

Q.E.D.

We next prove Proposition 15.

PROOF OF PROPOSITION 15. We first show that for each $T \geq 0$, the sequence $\left\{ \sup_{0 \leq t \leq T} |\tilde{W}^n(t)|, n \geq 1 \right\}$ is tight. Note that from equation (39) for the virtual waiting time process and recalling the definition of \tilde{W}^n from (41), we have that for each $T \geq 0$,

$$\begin{aligned}
 & \sup_{0 \leq t \leq T} |\tilde{W}^n(t)| \tag{80} \\
 &= \sup_{0 \leq t \leq T} \left| \tilde{W}^n(0) + \tilde{I}^n(t) + \sum_{k=1}^K \left[\tilde{M}_{\nu,k}^n(\bar{A}_k^n(t)) + \tilde{M}_{d,k}^n(\bar{A}_k^n(t)) + \tilde{\varepsilon}_k^n(t) \right] \right. \\
 & \quad \left. + \int_0^t \sum_{i=1}^K n^{1/2} \rho_k (G_k(W^n(s)) - G_k(W(s))) ds \right| \\
 &\leq \sup_{0 \leq t \leq T} \left| \tilde{W}^n(0) + \tilde{I}^n(t) + \sum_{k=1}^K \left[\tilde{M}_{\nu,k}^n(\bar{A}_k^n(t)) + \tilde{M}_{d,k}^n(\bar{A}_k^n(t)) + \tilde{\varepsilon}_k^n(t) \right] \right| \\
 & \quad + \sup_{0 \leq t \leq T} \left| \int_0^t \sum_{i=1}^K n^{1/2} \rho_k (G_k(W^n(s)) - G_k(W(s))) ds \right| \\
 &\leq \sup_{0 \leq t \leq T} \left| W^n(0) + I^n(t) + \sum_{k=1}^K \left[\tilde{M}_{\nu,k}^n(\bar{A}_k^n(t)) + \tilde{M}_{d,k}^n(\bar{A}_k^n(t)) + \tilde{\varepsilon}_k^n(t) \right] \right| \\
 & \quad + \sup_{0 \leq t \leq T} \left| \int_0^t \sum_{i=1}^K \rho_k \sup_{0 \leq u \leq \max(W^n(T), W(T))} f_k(u) \tilde{W}^n(s) ds \right| \\
 &\leq \sup_{0 \leq t \leq T} \left| \tilde{W}^n(0) + \tilde{I}^n(t) + \sum_{k=1}^K \left[\tilde{M}_{\nu,k}^n(\bar{A}_k^n(t)) + \tilde{M}_{d,k}^n(\bar{A}_k^n(t)) + \tilde{\varepsilon}_k^n(t) \right] \right| \\
 & \quad + \int_0^t \sum_{i=1}^K \rho_k \sup_{0 \leq u \leq \max(W^n(T), W(T))} f_k(u) \sup_{0 \leq u \leq s} |\tilde{W}^n(s)| ds.
 \end{aligned}$$

Thus, by Gronwall's inequality (6), it follows that

$$\begin{aligned}
 \sup_{0 \leq t \leq T} |\tilde{W}^n(t)| &\leq \sup_{0 \leq t \leq T} \left| \tilde{W}^n(0) + \tilde{I}^n(t) + \sum_{k=1}^K \left[\tilde{M}_{\nu,k}^n(\bar{A}_k^n(t)) + \tilde{M}_{d,k}^n(\bar{A}_k^n(t)) + \tilde{\varepsilon}_k^n(t) \right] \right| \\
 &\quad \times \exp \left(T \sum_{i=1}^K \rho_k \sup_{u \geq 0} f_k(u) \right). \tag{81}
 \end{aligned}$$

However, since by the Continuous mapping Theorem (3) and Propositions 10, 11, 12 and 14, the sequence

$$\left\{ \sup_{0 \leq t \leq T} \left| \tilde{W}^n(0) + \tilde{I}^n(t) + \sum_{k=1}^K \left[\tilde{M}_{\nu,k}^n(\bar{A}_k^n(t)) + \tilde{M}_{d,k}^n(\bar{A}_k^n(t)) + \tilde{\varepsilon}_k^n(t) \right] \right|, n \geq 1 \right\}$$

is tight, it follows from (81) that $\left\{ \sup_{0 \leq t \leq T} |\tilde{W}^n(t)|, n \geq 1 \right\}$ is tight as desired.

Now let $k \in \{1, \dots, K\}$ and note that by Assumption (A2) we have that for any $x \geq 0$ and $\delta \in \mathbb{R}$, $G_k(x + \delta) - G_k(x) = -f_k(x)\delta + o(\delta)$, where $o(\delta)/\delta \rightarrow 0$ as $\delta \rightarrow 0$, uniformly in x . It therefore follows from the definition of δ_k^n in (40), that

$$\begin{aligned} \sup_{0 \leq t \leq T} |\tilde{\delta}_k^n(t)| &= n^{1/2} \sup_{0 \leq t \leq T} \left| \int_0^t \rho_k ((G_k(W^n(s)) - G_k(W(s))) - (f_k(W(s))(W^n(s) - W(s)))) ds \right| \\ &= n^{1/2} \sup_{0 \leq t \leq T} \left| \int_0^t \rho_k o(W^n(s) - W(s)) ds \right| \\ &\leq T \rho_k \sup_{0 \leq t \leq T} n^{1/2} |o(W^n(t) - W(t))| \\ &= T \rho_k \sup_{0 \leq t \leq T} |\tilde{W}^n(t) (o(W^n(t) - W(t)) / (W^n(t) - W(t)))| \\ &\leq T \rho_k \sup_{0 \leq t \leq T} |\tilde{W}^n(t)| \sup_{0 \leq t \leq T} |o(W^n(t) - W(t)) / (W^n(t) - W(t))| \\ &\Rightarrow 0 \text{ as } n \rightarrow \infty, \end{aligned} \tag{82}$$

since the sequence $\left\{ \sup_{0 \leq t \leq T} |\tilde{W}^n(t)|, n \geq 1 \right\}$ is tight and by Theorem 1 and the Continuous Mapping Theorem, $\sup_{0 \leq t \leq T} |W^n(t) - W(t)| \Rightarrow 0$ as $n \rightarrow \infty$. This completes the proof. Q.E.D.

References

- [1] C. J. Ancker Jr. and A. Gafarian. Queueing with impatient customers who leave at random. *Journal of Industrial Engineering*, 13:84–90, 1962.
- [2] R. Bartle. *The Elements of Real Analysis*. Wiley, New York, 1976.
- [3] P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 2nd edition, 1999.
- [4] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing science perspective. 2002. Working paper.
- [5] J. G. Dai and W. Dai. A heavy traffic limit theorem for a class of open queueing networks with finite buffers. *Queueing Systems: Theory and Applications*, pages 5–40, 1999.

-
- [6] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, New York, 1986.
- [7] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, New York, 1999.
- [8] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4:208–227, 2002.
- [9] D. L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic, I. *Adv. Appl. Probab.*, (2):150–177, 1970.
- [10] A. Mandelbaum and S. Zeltyn. The impact of customers’ patience on delay and abandonment: Some empirically-driven experiments with the $M/M/n + G$ queue. *Operations Research*, 2005.
- [11] J. Reed and A. R. Ward. Approximating the $GI/GI/1+GI$ queue with a nonlinear drift diffusion. *Mathematics of Operations Research*, 2007. Forthcoming.
- [12] X. Su and S. A. Zenios. Patient choice in kidney allocation: the role of the queueing discipline. *Manufacturing & Service Operations Management*, 2004.
- [13] X. Su and S. A. Zenios. Recipient choice can address the efficiency-equity trade-off in kidney transplantation: A mechanism design model. 2005. Working paper.
- [14] A. R. Ward and P. W. Glynn. A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. 2005. Working paper.
- [15] W. Whitt. *Stochastic-Process Limits*. Springer, New York, 2002.
- [16] W. Whitt. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, 50:1449–1461, 2004.
- [17] W. Whitt. Fluid models for many-server queues with abandonments. *Operations Research*, 2005.