Telecommunication Systems 21:2–4, 149–171, 2002

Queue Lengths and Waiting Times for Multiserver Queues with Abandonment and Retrials

AVI MANDELBAUM * avim@tx.technion.ac.il
Technion Institute, Haifa, 32000, Israel

WILLIAM A. MASSEY wmassey@princeton.edu
Princeton University, Princeton, NJ 08544, USA

MARTIN I. REIMAN and ALEXANDER STOLYAR {marty;stolyar}@lucent.com
Bell Laboratories, Murray Hill, NJ 07974, USA

BRIAN RIDER rider@math.duke.edu
Duke University, Durham, NC 27708-0320, USA

Abstract. We consider a Markovian multiserver queueing model with time dependent parameters where waiting customers may abandon and subsequently retry. We provide simple fluid and diffusion approximations to estimate the mean, variance, and density for both the queue length and virtual waiting time processes arising in this model. These approximations, which are generated by numerically integrating only 7 ordinary differential equations, are justified by limit theorems where the arrival rate and number of servers grow large. We compare our approximations to simulations, and they perform extremely well.

Keywords: fluid approximations, diffusion approximations, multiserver queues, queues with abandonment, virtual waiting time, queues with retrials, nonstationary queues, call centers

1. Introduction

In this paper we continue our ongoing examination of a multiserver queue with time varying parameters where waiting customers may abandon and subsequently retry. The model we consider is a relatively simple special case of the class of models considered in [Mandelbaum et al., 4], which were termed *Markovian Service Networks.*

Our model, depicted in figure 1, consists of two nodes: a *service* node with n_t servers, and a *retrial pool* with an unlimited number of servers, where customers effectively serve themselves. New customers arrive to the service node as a nonhomogeneous Poisson process of rate λ_t. Customers arriving to find an idle server are taken into service with a duration that has a memoryless distribution of rate μ_t^1. Customers that find all servers busy join a queue, from which they are served in a FCFS manner. Each customer waiting in the queue abandons at rate β_t. An abandoning customer leaves the system with probability ψ_t or joins the retrial pool with probability $1 - \psi_t$. Each customer in the retrial pool leaves to enter the service node at rate μ_t^2. Upon entry to the

[*] Corresponding author.

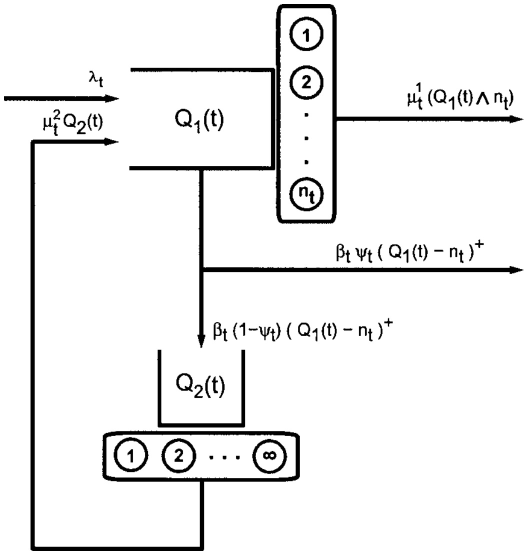
λ_t
$\mu_t^2 Q_2(t)$
$Q_1(t)$
1
2
n_t
$\mu_t^1 (Q_1(t) \wedge n_t)$
$\beta_t \psi_t (Q_1(t) - n_t)^+$
$\beta_t (1-\psi_t) (Q_1(t) - n_t)^+$
$Q_2(t)$
1
2
$\cdots$
∞

Figure 1. The multiserver queue with abandonment and retrials.

service node, these customers are treated the same as new customers. Our focus is the two-dimensional, continuous time Markov chain $\mathbf{Q}(t) = (Q_1(t), Q_2(t))$ where $Q_1(t)$ equals the *number of customers residing in the service node* (waiting or being served) and $Q_2(t)$ equals the *number of customers in the retrial pool.* We also consider the *virtual waiting time* $W(t)$, which is the time that an infinitely patient customer, arriving to the service node at time t, would have to wait before entering service.

This model, even with all parameters constant, is analytically intractable. We thus consider fluid and diffusion approximations for both the queue length and virtual waiting time processes. These approximations are justified by limit theorems where the arrival rate and number of servers grow large. Both the model and asymptotic regime are motivated by large telecommunication systems such as call centers, where abandonment and retrial occur naturally, and where time variability of parameters, specifically the arrival rate, cannot realistically be ignored. More discussion of this motivation is contained in [Mandelbaum et al., 5].

Fluid and diffusion limits for the (two-dimensional) queue length process arising in this model were proved in [Mandelbaum et al., 4]. In [Mandelbaum et al., 5] we compared the fluid limit with simulation results, and found that it provides an excellent approximation. Fluid and diffusion limits for the virtual waiting time are proved in [Mandelbaum et al., 7]. These results are described in [Mandelbaum et al., 6], where a single numerical example shows that the fluid approximation for the virtual waiting time is also excellent. In this paper we extend the previous results in several directions. First, we provide additional numerical examples for both the queue lengths and virtual waiting time, comparing the fluid approximations to simulations. Next, we provide new numerical results for the diffusion approximations. We also compare the simulated sample variance of the virtual waiting time to the variance of its diffusion approximation. Using equations originally obtained in [Mandelbaum et al., 4], we calculate the covariance matrix of the queue length diffusion, and compare it to simulations. Using a result from [Mandelbaum et al., 7] that provides conditions under which the queue length diffusion process is Gaussian, we also obtain a Gaussian approximation for the queue length density at the service node. We are similarly able to obtain an Gaussian approximation for the virtual waiting time density. These are also compared to simulations. In all of these comparisons our approximations are exceptionally good.

The rest of this paper is organized as follows. In section 2 we provide the equations for the queue length process and in section 3 we provide the same for the virtual waiting time process. We also state in both sections the relevant limit theorems that inspire our fluid and diffusion approximations. Section 4 contains numerical examples comparing our approximations with simulation results. Section 5 is an appendix that provides some background on Markovian service networks.

2. Queueing sample paths and asymptotics

In order to motivate our sample path construction of the multiserver queue with abandonment and retrials, we first present a brief description of the simpler $M_t/M_t/n_t$ queue. The $M_t/M_t/n_t$ queue length process $Q = \{Q(t) \mid t \geqslant 0\}$ is a continuous time Markov chain with time varying instantaneous transition rates. It consists of an arrival process that is time-inhomogeneous Poisson with rate function $\{\lambda_t \mid t \geqslant 0\}$, a deterministic schedule of servers $\{n_t \mid t \geqslant 0\}$ who each work for a service time that has an independent, memoryless distribution determined by the rate function $\{\mu_t \mid t \geqslant 0\}$. We assume that all these functions are locally integrable. Since the number of servers can vary in time, we use the convention of *preemptive-resume service*. When the number of servers suddenly drops below the number of customers currently in service, then the dropped customers are placed in the infinite buffer to resume service later.

The standard approach to constructing the sample path distribution for this $M_t/M_t/n_t$ queueing process is to state that its transition probabilities, i.e.

$$p_{i,j}(t) = \mathsf{P}\big\{Q(t) = j \mid Q(0) = i\big\}, \tag{2.1}$$

for all non-negative integers i and j, are the unique solutions to the forward equations

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{i,0}(t) = \mu_t p_{i,1}(t) - \lambda_t p_{i,0}(t) \tag{2.2}$$

and if $j \geqslant 1$,

$$\frac{\mathrm{d}}{\mathrm{d}t} p_{i,j}(t) = \lambda_t p_{i,j-1}(t) + \mu_t \min(j+1, n_t) p_{i,j+1}(t) - \big(\lambda_t + \mu_t \min(j, n_t)\big) p_{i,j(t)}. \tag{2.3}$$

where $p_{i,j}(0) = 1$ if and only if $i = j$ and $p_{ij}(0) = 0$ otherwise (for more details, see [Wolff, 9]).

The $M_t/M_t/n_t$ queueing process is the canonical example for a special family of CTMC's that we call *Markovian service networks* (see [Mandelbaum et al., 4] for details). They can be defined precisely by an alternative method to defining forward equations for their transition probabilities. Instead, we use an implicit definition to construct their random sample paths directly. The sample paths for the $M_t/M_t/n_t$ queueing process are the unique solution to the equation

$$Q(t) = Q(0) + \Pi^a\biggl(\int_0^t \lambda_s \,\mathrm{d}s\biggr) - \Pi^b\biggl(\int_0^t \mu_s \cdot \min\bigl(Q(s), n_s\bigr)\,\mathrm{d}s\biggr), \tag{2.4}$$

where $\Pi^a \equiv \{\Pi^a(t) \mid t \geqslant 0\}$ and $\Pi^b \equiv \{\Pi^b(t) \mid t \geqslant 0\}$ are two independent, standard (mean rate 1), Poisson processes.

In the same spirit, the random sample paths of the queue length process for the multiserver queue with abandonment and retrials $\mathbf{Q}(t) = (Q_1(t), Q_2(t))$ are uniquely determined by the relations

$$\begin{aligned} Q_1(t) = {} & Q_1(0) + \Pi_{21}^c\biggl(\int_0^t Q_2(s)\mu_s^2 \,\mathrm{d}s\biggr) - \Pi_{12}^b\biggl(\int_0^t \bigl(Q_1(s) - n_s\bigr)^+ \beta_s(1 - \psi_s)\,\mathrm{d}s\biggr) \\ & + \Pi^a\biggl(\int_0^t \lambda_s \,\mathrm{d}s\biggr) - \Pi^b\biggl(\int_0^t \bigl(Q_1(s) - n_s\bigr)^+ \beta_s \psi_s \,\mathrm{d}s\biggr) \\ & - \Pi^c\biggl(\int_0^t \bigl(Q_1(s) \wedge n_s\bigr)\mu_s^1 \,\mathrm{d}s\biggr) \end{aligned} \tag{2.5}$$

and

$$\begin{aligned} Q_2(t) = {} & Q_2(0) + \Pi_{12}^b\biggl(\int_t^0 \bigl(Q_1(s) - n_s\bigr)^+ \beta_s(1 - \psi_s)\,\mathrm{d}s\biggr) \\ & - \Pi_{21}^c\biggl(\int_0^t Q_2(s)\mu_s^2 \,\mathrm{d}s\biggr), \end{aligned} \tag{2.6}$$

where Π^a, Π^b, Π^c, Π_{12}^b, and Π_{21}^c are five given mutually independent, standard Poisson processes and λ, β, μ^1, μ^2, ψ, n are locally integrable functions of time [Mandelbaum et al., 4]. Here $x \wedge y = \min(x, y)$ and $x^+ = \max(x, 0)$ for all real x and y. Using the theory of strong approximations for Poisson processes, we can use the random sample path construction of our queueing processes to do an asymptotic sample path analysis and obtain our fluid and diffusion limit theorems.

We are interested in the asymptotic regime where we scale up the number of servers in response to a similar scaling up of the arrival rate by customers. More precisely, the asymptotic regime is as follows. In a system with index η, the only scaled parameters are: the initial conditions $Q_i^\eta(0) = \lceil \eta Q_i^{(0)}(0) + \sqrt{\eta} Q_i^{(1)}(0) \rceil + \mathrm{o}(\sqrt{n})$ for constants $Q_i^{(0)}(0)$ and $Q_i^{(1)}(0)$ $(i = 1, 2)$, the external arrival rate (i.e., the intensity of the Poisson arrival process), which is now $\eta\lambda_t$, and the number of servers, which is now ηn_t. (Actually, the latter should be the integer part of ηn_t, but to avoid trivial complications and simplify notation, we assume it's just ηn_t.) The scaled queue length process $\mathbf{Q}^\eta(t) = (Q_1^\eta(t), Q_2^\eta(t))$ is then uniquely determined by the relations

$$\begin{aligned} Q_1^\eta(t) = Q_1^\eta(0) &+ \Pi_{21}^c\left(\int_0^t Q_2^\eta(s)\mu_s^2\,\mathrm{d}s\right) - \Pi_{12}^b\left(\int_0^t \big(Q_1^\eta(s) - \eta n_s\big)^+\beta_s(1-\psi_s)\,\mathrm{d}s\right) \\ &+ \Pi^a\left(\int_0^t \eta\lambda_s\,\mathrm{d}s\right) - \Pi^b\left(\int_0^t \big(Q_1^\eta(s) - \eta n_s\big)^+\beta_s\psi_s\,\mathrm{d}s\right) \\ &- \Pi^c\left(\int_0^t \big(Q_1^\eta(s) \wedge (\eta n_s)\big)\mu_s^1\,\mathrm{d}s\right) \end{aligned} \tag{2.7}$$

and

$$Q_2^\eta(t) = Q_2^\eta(0) + \Pi_{12}^b\left(\int_0^t \big(Q_1^\eta(s) - \eta n_s\big)^+\beta_s(1-\psi_s)\,\mathrm{d}s\right) - \Pi_{21}^c\left(\int_0^t Q_2^\eta(s)\mu_s^2\,\mathrm{d}s\right). \tag{2.8}$$

Now we state the strong law of large numbers limit theorem for the retrial model. We make the following asymptotic assumptions for the initial conditions

$$\lim_{\eta\to\infty} \frac{1}{\eta}\mathbf{Q}^\eta(0) = \mathbf{Q}^{(0)}(0) \quad \text{a.s.}, \tag{2.9}$$

where $\mathbf{Q}^{(0)}(0)$ is a constant.

Theorem 2.1. We have

$$\lim_{\eta\to\infty} \frac{1}{\eta}\mathbf{Q}^\eta = \mathbf{Q}^{(0)} \quad \text{a.s.} \tag{2.10}$$

where the convergence is uniform on compact sets of t. Moreover, $\mathbf{Q}^{(0)} = \{\mathbf{Q}^{(0)}(t) \mid t \geqslant 0\}$ is uniquely determined by $\mathbf{Q}^{(0)}(0)$ and the autonomous differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}Q_1^{(0)}(t) = \lambda_t + \mu_t^2 Q_2^{(0)}(t) - \mu_t^1\big(Q_1^{(0)}(t) \wedge n_t\big) - \beta_t\big(Q_1^{(0)}(t) - n_t\big)^+ \tag{2.11}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}Q_2^{(0)}(t) = \beta_t(1-\psi_t)\big(Q_1^{(0)}(t) - n_t\big)^+ - \mu_t^2 Q_2^{(0)}(t). \tag{2.12}$$

This theorem states rigorously that $\mathbf{Q}^\eta \approx \eta\mathbf{Q}^{(0)}$ for large η and we call $\mathbf{Q}^{(0)}$ the *fluid approximation* for $\mathbf{Q}^\eta$.

If two random variables X and Y have the same distribution then we denote this by $X \overset{\mathrm{d}}{=} Y$. If $\{X_n \mid n \geqslant 0\}$ converges in distribution to Y, we denote this by $\lim_{n\to\infty} X_n \overset{\mathrm{d}}{=} Y$. The fluid approximation can be refined using the following functional central limit theorem, as proved in [Mandelbaum et al., 4]. We make the following assumptions for the initial conditions

$$\lim_{\eta\to\infty} \sqrt{\eta}\left(\frac{1}{\eta}\mathbf{Q}^{\eta}(0) - \mathbf{Q}^{(0)}(0)\right) \overset{\mathrm{d}}{=} \mathbf{Q}^{(1)}(0), \tag{2.13}$$

where $\mathbf{Q}^{(1)}(0)$ is a constant.

Theorem 2.2. We have

$$\lim_{\eta\to\infty} \sqrt{\eta}\left(\frac{1}{\eta}\mathbf{Q}^{\eta} - \mathbf{Q}^{(0)}\right) \overset{\mathrm{d}}{=} \mathbf{Q}^{(1)} \tag{2.14}$$

where $\mathbf{Q}^{1} = \{\mathbf{Q}^{1}(t) \mid t \geqslant 0\}$ is a diffusion process. This is a convergence in distribution of the stochastic processes in an appropriate functional space [Mandelbaum et al., 4].

Moreover, if the set of time points $\{t \geqslant 0 \mid Q_1^{(0)}(t) = n_t\}$ has measure zero for the multiserver queue with abandonment and retrial model, then $\{\mathbf{Q}^{(1)}(t) \mid t \geqslant 0\}$ is a Gaussian process. The mean vector for $\mathbf{Q}^{(1)}$ then solves the set of autonomous differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{E}\left[Q_1^{(1)}(t)\right] = -\left(\mu_t^1 1_{\{Q_1^{(0)}(t)\leqslant n_t\}} + \beta_t \mathbb{1}_{\{Q_1^{(0)}(t)>n_t\}}\right)\mathsf{E}\left[Q_1^{(1)}(t)\right] + \mu_t^2\,\mathsf{E}\left[Q_2^{(1)}(t)\right] \tag{2.15}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{E}\left[Q_2^{(1)}(t)\right] = \beta_t(1-\psi_t)\mathbb{1}_{\{Q_1^{(0)}(t)\geqslant n_t\}}\,\mathsf{E}\left[Q_1^{(1)}(t)\right] + \mu_t^2\,\mathsf{E}\left[Q_2^{(1)}(t)\right]. \tag{2.16}$$

Finally, the covariance matrix for $\mathbf{Q}^{(1)}$ solves the autonomous differential equations

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{Var}\left[Q_1^{(1)}(t)\right] = {}& -2\left(\beta_1\mathbb{1}_{\{Q_1^{(0)}(t)>n_t\}} + \mu_t^1\mathbb{1}_{\{Q_1^{(0)}(t)\leqslant n_t\}}\right)\mathsf{Var}\left[Q_1^{(1)}(t)\right]\\ &+ 2\mu_t^2\,\mathsf{Cov}\left[Q_1^{(1)}(t),\, Q_2^{(1)}(t)\right]\\ &+ \lambda_t + \beta_t\left(Q_1^{(0)}(t) - n_t\right)^+ + \mu_t^1\left(Q_1^{(0)}(t)\wedge n_t\right) + \mu_t^2 Q_2^{(0)}(t), \qquad (2.17)\\ \frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{Var}\left[Q_2^{(1)}(t)\right] = {}& -2\mu_t^2\,\mathsf{Var}\left[Q_2^{(1)}(t)\right] + 2\beta_t(1-\psi_t)\mathbb{1}_{\{Q_1^{(0)}(t)\geqslant n_t\}}\,\mathsf{Cov}\left[Q_1^{(1)}(t),\, Q_2^{(1)}(t)\right]\\ &+ \beta_t(1-\psi_t)\left(Q_1^{(0)}(t) - n_t\right)^+ + \mu_t^2 Q_2^{(0)}(t), \qquad (2.18)\end{aligned}$$

and

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{Cov}\left[Q_1^{(1)}(t), q_2^{(1)}(t)\right]\\ &\quad= \beta_t(1-\psi_t)\mathbb{1}_{\{Q_1^{(0)}(t)\geqslant n_t\}}\,\mathsf{Var}\left[Q_1^{(1)}(t)\right] + \mu_t^2\,\mathsf{Var}\left[Q_2^{(1)}(t)\right]\\ &\qquad - \left(\beta_t\mathbb{1}_{\{Q_1^{(0)}(t)>n_t\}} + \mu_t^1\mathbb{1}_{\{Q_1^{(0)}(t)\leqslant n_t\}} + \mu_t^2\right)\mathsf{Cov}\left[Q_1^{(1)}(t),\, Q_2^{(1)}(t)\right]\\ &\qquad - \beta_t(1-\psi_t)\left(Q_1^{(0)}(t) - n_t\right)^+ - \mu_t^2 Q_2^{(0)}(t). \end{aligned} \tag{2.19}$$

This theorem states rigorously that $\mathbf{Q}^{\eta} \approx \eta\mathbf{Q}^{(0)} + \sqrt{\eta}\mathbf{Q}^{(1)}$ for large η and we call $\mathbf{Q}^{(1)}$ the *diffusion approximation* for $\mathbf{Q}^{\eta}$. It should be pointed out that equations (2.17)–(2.19) are corrected versions of the covariance equations for the multiserver queue with abandonment and retrials given in [Mandelbaum et al., 4, Mandelbaum et al., 5]. The previous incorrect formulas do *not* affect the numerical results of papers [Mandelbaum et al., Mandelbaum et al., 5, Mandelbaum et al., 6] since those computational results focused only on the utility of the fluid approximation and not the diffusion approximation. To double check the validity of the diffusion covariance equations used here, we derive in the appendix the general differential equations of the diffusion covariance for the special case of a two-dimensional Markovian service network.

Time-varying queues alternate among three phases. For a given time t, we define the phases to be:

1. Underloaded or $Q_1^{(0)}(t) < n_t$,
2. Critically-loaded or $Q_1^{(0)}(t) = n_t$,
3. Overloaded or $Q_1^{(0)}(t) > n_t$.

Similar phases and transitions are discussed in great detail for the $M_t/M_t/1$ queue in [Mandelbaum and Massey, 3].

To guarantee the results of theorem 2.2, the fluid model for the service node is free to alternate between phases of underloading and overloading. We only require during these transitions that it does not "linger" too long in the critically loaded phase so that $\{t \mid Q_1^{(0)}(t) = n_t\}$ is a set of measure zero. As we show in our numerical examples in section 4, even though our examples satisfy the measure zero hypothesis for the times of critical loading, this lingering behavior does affect the quality of our approximations.

3. Virtual waiting time for the service node

In this section we consider asymptotics for the virtual waiting time process. To do that we need a few additional assumptions which are not very restrictive.

Assumption 3.1. In the interval $[0, \infty)$:

1. The function n_t is continuously differentiable;
2. The function μ_t^1 is continuous;
3. The functions μ_t^2 and β_t are bounded on compact intervals.

Assumption 3.2 is introduced below when the required notation is in place.

Suppose that we are interested in the waiting time of a *virtual customer* arriving to the service node at a *fixed* time $\tau \geqslant 0$. Since we have a system with abandonment, a convenient way to approach this problem is to consider the system that is obtained from the original one by the following modifications:

1. There are no new exogenous arrivals into the system after time τ.
2. Any customer departing any station i, after time τ, leaves the entire system.

In particular, the service node has no new arrivals (exogenous or retrial) after time τ. It only processes the remaining customers that are there at time τ. Theorems 2.1 and 2.2 still apply to the modified system; the only difference is that certain terms in the equations, corresponding to the arrivals after time τ, should be "zeroed out." The following results follow directly from these two theorems (and their proofs in [Mandelbaum et al., 4]).

Denote the arrival and departure processes for the service node by

$$A^{\eta} = \{A^{\eta}(t) \mid t \geqslant 0\} \quad \text{and} \quad \Delta^{\eta} = \{\Delta^{\eta}(t) \mid t \geqslant 0\},$$

respectively. By convention, let the arrival process include the customers in the service node at time 0, so $A^{\eta}(0) = \widehat{Q}^{\,\eta}_{1}(0)$, $\Delta^{\eta}(0) = 0$, and $A^{\eta}(t) - \Delta^{\eta}(t) = \widehat{Q}^{\eta}_{1}(t)$, $t \geqslant 0$. We then obtain the following *fluid* limit result.

Theorem 3.1. As a joint process we have

$$\lim_{\eta\to\infty} \frac{1}{\eta}\big(\widehat{\mathbf{Q}}^{\eta}, A^{\eta}, \Delta^{\eta}\big) = \big(\widehat{\mathbf{Q}}^{(0)}, A^{(0)}, \Delta^{(0)}\big) \quad \text{a.s.} \tag{3.1}$$

and this convergence is uniform on compact sets of t. The fluid limit $\widehat{Q}^{(0)}_{1}(t)$ satisfies equation (2.11) for $t < \tau$. For $t \geqslant \tau$, we have the following properties:

1. The future evolution of $\widehat{Q}^{(0)}_{1}(t)$ is governed by the differential equation
$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{Q}^{(0)}_{1}(t) = -\mu^{1}_{t}\big(\widehat{Q}^{(0)}_{1}(t) \wedge n_t\big) - \beta_t\big(\widehat{Q}^{(0)}_{1}(t) - n_t\big)^{+}. \tag{3.2}$$
2. There are no future arrivals, so that $A^{(0)}(t) = A^{(0)}(\tau)$.
3. The deterministic process $\Delta^{(0)}$ is a continuously differentiable nondecreasing function in $[0, \infty)$.

We also obtain the following *diffusion* limit.

Theorem 3.2. The following convergence in distribution holds:

$$\lim_{\eta\to\infty} \sqrt{\eta}\left(\frac{1}{\eta}\widehat{\mathbf{Q}}^{\eta} - \widehat{\mathbf{Q}}^{(0)}, \frac{1}{\eta}A^{\eta} - A^{(0)}, \frac{1}{\eta}\Delta^{\eta} - \Delta^{(0)}\right) \overset{\mathrm{d}}{=} \big(\widehat{\mathbf{Q}}^{(1)}, A^{(1)}, \Delta^{(1)}\big). \tag{3.3}$$

Moreover, if the set of time points $\{t \geqslant 0 \mid \widehat{Q}^{(0)}_{1}(t) = n_t\}$ has measure zero, $\{\widehat{Q}^{(1)}_{1}(t) \mid t \geqslant 0\}$ is a Gaussian process and for $t \geqslant \tau$, $\mathsf{Var}[\widehat{Q}^{(1)}_{1}(t)]$ solves the differential equation

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{Var}\big[\widehat{Q}^{(1)}_{1}(t)\big] = {} & -2\big(\beta_t \mathbb{1}_{\{\widehat{Q}^{(0)}_{1}(t) > n_t\}} + \mu^{1}_{t}\mathbb{1}_{\{\widehat{Q}^{(0)}_{1}(t) \leqslant n_t\}}\big)\mathsf{Var}\big[\widehat{Q}^{(1)}_{1}(t)\big] \\ & + \beta_t\big(\widehat{Q}^{(0)}_{1}(t) - n_t\big)^{+} + \mu^{1}_{t}\big(\widehat{Q}^{(0)}_{1}(t) \wedge n_t\big).\end{aligned} \tag{3.4}$$

It follows from the definitions and the above theorem that

$$\widehat{Q}_1^{(1)}(t) = A^{(1)}(t) - \Delta^{(1)}(t). \tag{3.5}$$

Now, let us define the *potential service initiation* process D^η for the service node by

$$D^\eta(t) = \Delta^\eta(t) + \eta n_t, \quad t \geqslant 0.$$

Note that if $\widehat{Q}_1^\eta(t) < \eta n_t$, then $A^\eta(t) < D^\eta(t)$; so the potential service can be "ahead" of arrivals. It follows that

$$\lim_{\eta\to\infty} \frac{1}{\eta} D^\eta(\cdot) = D^{(0)}(\cdot) \quad \text{a.s.},$$

where the convergence is uniform on compact sets of t and $D^{(0)}(t) = \Delta^{(0)}(t) + n_t$, $t \geqslant 0$. Since n_t is continuously differentiable by assumption and we know that $\Delta^{(0)}(t)$ is continuously differentiable, $D^{(0)}(t)$ is also continuously differentiable and we denote its derivative by $d^0(t)$. Now we make an important but not very restrictive (in the majority of applications) additional assumption.

Assumption 3.2. The function $D^{(0)}$ (of t) is continuously differentiable with *strictly positive derivative,* and

$$\lim_{t\to\infty} D^0(t) > A^{(0)}(\tau). \tag{3.6}$$

According to our definitions, both $A^\eta(\cdot)$ and $A^{(0)}(\cdot)$ are constant in the interval $[\tau, \infty)$.

Also, it is convenient to adopt the convention that all the processes we consider are defined in the interval $[-T, \infty)$, with

$$T = \frac{n_0}{d^{(0)}(0)}.$$

We make this extension by assuming that nothing is happening in the interval $[-T, 0)$ (no arrivals or departures) except the number of servers is increasing linearly from 0 to ηn_0 (for the unscaled process with index η).

We then can rewrite (3.1) and (3.3) as follows (with all the functions being now defined for $t \geqslant -T$):

$$\lim_{\eta\to\infty} \frac{1}{\eta}\big(\widehat{\mathbf{Q}}^\eta, A^\eta, D^\eta\big) = \big(\widehat{\mathbf{Q}}^{(0)}, A^{(0)}, D^{(0)}\big) \tag{3.7}$$

and

$$\lim_{\eta\to\infty} \sqrt{\eta}\left(\frac{1}{\eta}\widehat{\mathbf{Q}}^\eta - \widehat{\mathbf{Q}}^{(0)}, \frac{1}{\eta}A^\eta - A^{(0)}, \frac{1}{\eta}D^\eta - D^{(0)}\right) \overset{\mathrm{d}}{=} \big(\widehat{\mathbf{Q}}^{(1)}, A^{(1)}, D^{(1)}\big), \tag{3.8}$$

where

$$D^{(1)} = \Delta^{(1)}. \tag{3.9}$$

Note that processes $A^{(0)}$, $D^{(0)}$, $A^{(1)}$, $D^{(1)}$ are continuous and $D^{(0)}(-T) = D^{(1)}(-T) = 0$.

Our conventions together with assumption 3.2 make the following processes well defined and finite with probability 1 for all sufficiently large η. Let us define, for all $t \geqslant -T$, the *first attainment* processes

$$S^{\eta}(t) = \inf\{s \geqslant -T\colon\ D^{\eta}(s) > A^{\eta}(t)\}$$

and

$$S^{(0)}(t) = \inf\{s \geqslant -T\colon\ D^{(0)}(s) > A^{(0)}(t)\}. \tag{3.10}$$

Similarly, define the *attainment* waiting time processes to be

$$W^{\eta}(t) = S^{\eta}(t) - t$$

and

$$W^{(0)}(t) = S^{(0)}(t) - t. \tag{3.11}$$

Denote by $\widehat{W}^{\eta}(\tau)$ the *virtual* waiting time at τ, i.e. the time a "test" customer (in the original non-modified system) arriving to the service node at time τ would have to wait until its service starts, assuming this customer *does not abandon* while waiting. Then the relation between the virtual waiting time $\widehat{W}^{\eta}(\tau)$ and the attainment waiting time $W^{\eta}(\tau)$ is simply

$$\widehat{W}^{\eta}(\tau) = W^{\eta}(\tau)^{+}. \tag{3.12}$$

Indeed, note that $W^{\eta}(\tau)$ (and $W^{(0)}(\tau)$) may be negative. All this means is that $\widehat{Q}_1^{\eta}(\tau) < \eta n_{\tau}$, and therefore in this case $\widehat{W}^{\eta}(\tau) = 0$. If $W^{\eta}(\tau)$ is non-negative, then its value is exactly equal to the virtual waiting time.

It follows directly from the theorem and corollary in [Puhalskii, 8] that (3.7), (3.8), and assumption 3.2, imply the following convergences.

Theorem 3.3. We have

$$\lim_{\eta\to\infty}\left(\frac{1}{\eta}\widehat{\mathbf{Q}}^{\eta}, \frac{1}{\eta}A^{\eta}, \frac{1}{\eta}D^{\eta}, W^{\eta}\right) = \left(\widehat{\mathbf{Q}}^{(0)}, A^{(0)}, D^{(0)}, W^{(0)}\right) \quad \text{a.s.}, \tag{3.13}$$

$$\begin{aligned}&\lim_{\eta\to\infty}\sqrt{\eta}\left(\frac{1}{\eta}\widehat{\mathbf{Q}}^{\eta} - \widehat{\mathbf{Q}}^{(0)}, \frac{1}{\eta}A^{\eta} - A^{(0)}, \frac{1}{\eta}D^{\eta} - D^{(0)}, W^{\eta} - W^{(0)}\right)\\ &\quad\overset{\mathrm{d}}{=} \left(\widehat{Q}^{(1)}, A^{(1)}, D^{(1)}, W^{(1)}\right),\end{aligned} \tag{3.14}$$

where

$$W^{(1)}(t) = \frac{A^{(1)}(t) - D^{(1)}(S^{(0)}(t))}{d^{(0)}(S^{(0)}(t))} \quad \text{and}$$
$$S^{(0)}(t) = \inf\{s \geqslant -T\colon\ D^{(0)}(s) > A^{(0)}(t)\}.$$

Since the processes $A^{(1)}$, $D^{(1)}$, $\widehat{Q}^{(1)}$, $W^{(1)}$ are continuous with probability 1, we automatically obtain the convergence of finite-dimensional distributions.

In particular, consider the nontrivial case $S^{(0)}(\tau) \geqslant \tau$ (which is equivalent to $\widehat{Q}_1^{(0)}(\tau) \geqslant n_\tau$). Moreover, assume that in $[0, \tau]$, the set of points $\{t \mid \widehat{Q}_1^{(0)}(t) = n_t\}$ has measure zero. Then we obtain

$$\lim_{\eta\to\infty} W^\eta(\tau) = W^{(0)}(\tau) \quad \text{a.s.}$$

and

$$\lim_{\eta\to\infty} \sqrt{\eta}\big(W^\eta(\tau) - W^{(0)}(\tau)\big) \overset{\mathrm{d}}{=} W^{(1)}(\tau) = \frac{\widehat{Q}_1^{(1)}(S^{(0)}(\tau))}{d^{(0)}(S^{(0)}(\tau))},$$

where $\widehat{Q}_1^{(1)}(S^{(0)}(\tau))$ is Gaussian with mean and variance computed as follows. Solving equation (3.2) for $\widehat{Q}_1^{(0)}(\cdot)$ in the interval $[\tau, \infty)$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{Q}_1^{(0)}(t) = -\beta_t \widehat{Q}_1^{(0)}(t) + \big(\beta_t - \mu_t^1\big)n_t, \quad t \geqslant \tau.$$

We can find $S^{(0)}(\tau)$ from

$$S^{(0)}(\tau) = \min\big\{t \geqslant \tau \mid \widehat{Q}_1^{(0)}(t) = n_t\big\}.$$

We then compute $\mathsf{E}[\widehat{Q}_1^{(1)}(S^{(0)}(\tau))]$ and $\mathsf{Var}[\widehat{Q}_1^{(1)}(S^{(0)}(\tau))]$, where

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\big[\widehat{Q}_1^{(1)}(t)\big] = -\beta_t\,\mathsf{E}\big[\widehat{Q}_1^{(1)}(t)\big], \quad t \geqslant \tau. \tag{3.15}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{Var}\big[\widehat{Q}_1^{(1)}(t)\big] = -2\beta_t\,\mathsf{Var}\big[\widehat{Q}_1^{(1)}(t)\big] + \beta_t\big(\widehat{Q}_1^{(0)}(t) - n_t\big) + \mu_t^1 n_t, \quad t \geqslant \tau. \tag{3.16}$$

This yields the closed form formulas

$$\widehat{Q}_1^{(0)}(t) = \widehat{Q}_1^{(0)}(\tau)\exp\bigg(-\int_\tau^t \beta_s\,\mathrm{d}s\bigg) + \int_\tau^t \big(\beta_s - \mu_s^1\big)n_s \exp\bigg(-\int_\tau^t \beta_r\,\mathrm{d}r\bigg)\mathrm{d}s, \tag{3.17}$$

$$\mathsf{E}\big[\widehat{Q}_1^{(1)}(t)\big] = \mathsf{E}\big[\widehat{Q}_1^{(1)}(\tau)\big]\exp\bigg(-\int_\tau^t \beta_s\,\mathrm{d}s\bigg), \tag{3.18}$$

and

$$\begin{aligned}\mathsf{Var}\big[\widehat{Q}_1^{(1)}(S^{(0)}(\tau))\big] = {} & \mathsf{Var}\big[\widehat{Q}_1^{(1)}(\tau)\big]\exp\bigg(-\int_\tau^{S^{(0)}(\tau)} 2\beta_s\,\mathrm{d}s\bigg)\\ & + \int_\tau^{S^{(0)}(\tau)} \big(\big(\widehat{Q}_1^{(0)}(s) - n_s\big)\beta_s - \mu_s^1 n_s\big)\exp\bigg(-\int_s^{S^{(0)}(\tau)} 2\beta_r\,\mathrm{d}r\bigg)\mathrm{d}s.\end{aligned} \tag{3.19}$$

Finally, noting that $d^{(0)}(S^{(0)}(\tau)) = n_{S^{(0)}(\tau)}\mu_{S^{(0)}(\tau)}$ when $S^{(0)}(\tau) \geqslant \tau$, we obtain

$$\mathsf{Var}\left[W^{(1)}(\tau)\right] = \frac{\mathsf{Var}[\widehat{Q}^{(1)}(S^{(0)}(\tau))]}{(n_{S^{(0)}(\tau)}\mu_{S^{(0)}(\tau)})^2}. \tag{3.20}$$

Remark. In this section we derived fluid and diffusion approximations of the marginal distribution of the attainment waiting time, which uniquely determines those for the virtual waiting time at the service node for *at a given time* $\tau \geqslant 0$. However, it is shown in [Mandelbaum et al., 7] that similar asymptotics hold for the attainment waiting time as a *random process* defined for $\tau \in [0, \infty)$. (See also [Mandelbaum et al., 6] for the formal statement of the results.)

4. Numerical examples

Several examples indicating the accuracy of the fluid approximation for the queue length process were considered in [Mandelbaum et al., 5]. The first examples had constant arrival rate, and exhibited the approach to equilibrium. The next examples had a quadratic arrival rate, and the final examples involved a “spike” in the arrival rate. In all cases the fluid approximation was excellent. In [Mandelbaum et al., 6] the accuracy of the fluid approximation for the virtual waiting time was checked for one of the examples from [Mandelbaum et al., 5] with quadratic arrival rate. Although not as accurate as the fluid approximation for the queue length in the same example, the approximation for the virtual waiting time was nonetheless excellent.

Here we examine the performance of the fluid and diffusion approximations for both queue length and virtual waiting time in some new examples. Details of how the simulations are carried out are contained in [Mandelbaum et al., 5]. Here we merely point out that we use 5,000 independent replications in each of our experiments. By contrast, all the fluid and diffusion approximations used here come from numerically integrating 7 ordinary differential equations.

Our numerical examples cover the case of time-varying behavior only for the external arrival rate λ_t. The type of time varying behavior used is that of a periodic square wave, oscillating between two values (starting with the smaller value) and the duration of each value is 2 time units for a total time interval of 20 time units. The 20/100 *case* will have λ_t oscillating between the values of 20 and 100 and the 40/80 *case* will have λ_t oscillating between the values of 40 and 80. For both cases, we set $\mu_t^1 = 1$, $\mu_t^2 = 0.2$, $Q_1^\eta(0) = Q_2^\eta(0) = 0$, $n_t = 50$, $\beta_t = 2$, and $\psi_t = 0.5$ for all $t \geqslant 0$ and $\eta > 0$.

The graphs are ordered by pairing the 20/100 case first (the top graph) followed by the 40/80 case (the bottom graph) for the following numerical plots:

1. Empirical averages of $Q_1(t)$ and $Q_2(t)$ versus their fluid approximations (figure 2).
2. Empirical covariance matrix of $Q_1(t)$ and $Q_2(t)$ versus the covariance matrix of their joint diffusion approximation (figure 3).
3. Empirical density of $Q_1(t)$ versus its Gaussian approximation (figure 4).

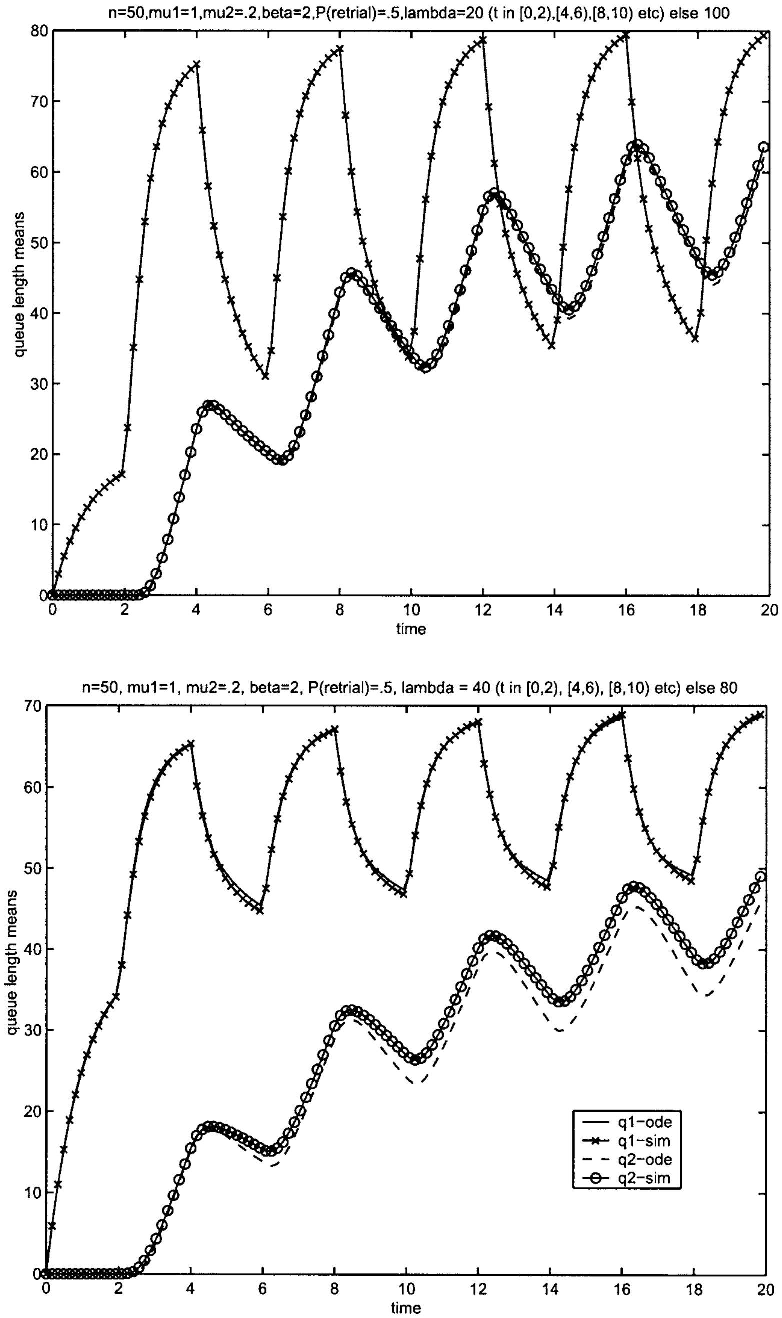

Figure 2. Numerical example: Empirical averages of $Q_1(t)$ and $Q_2(t)$ versus their fluid approximations for the 20/100 and 40/80 square wave cases.

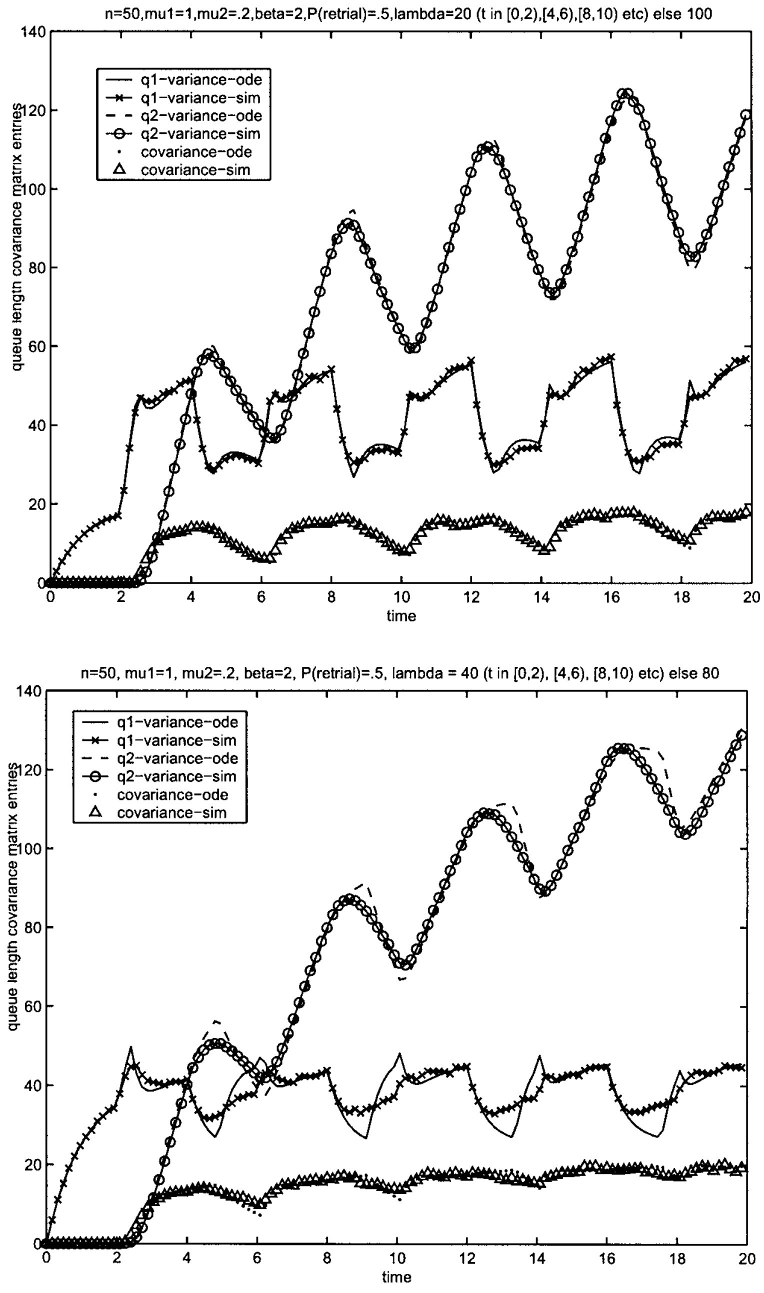

Figure 3. Numerical example: Empirical covariance matrix of $Q_1(t)$ and $Q_2(t)$ versus the same from its diffusion approximation for the 20/100 and 40/80 square wave cases.

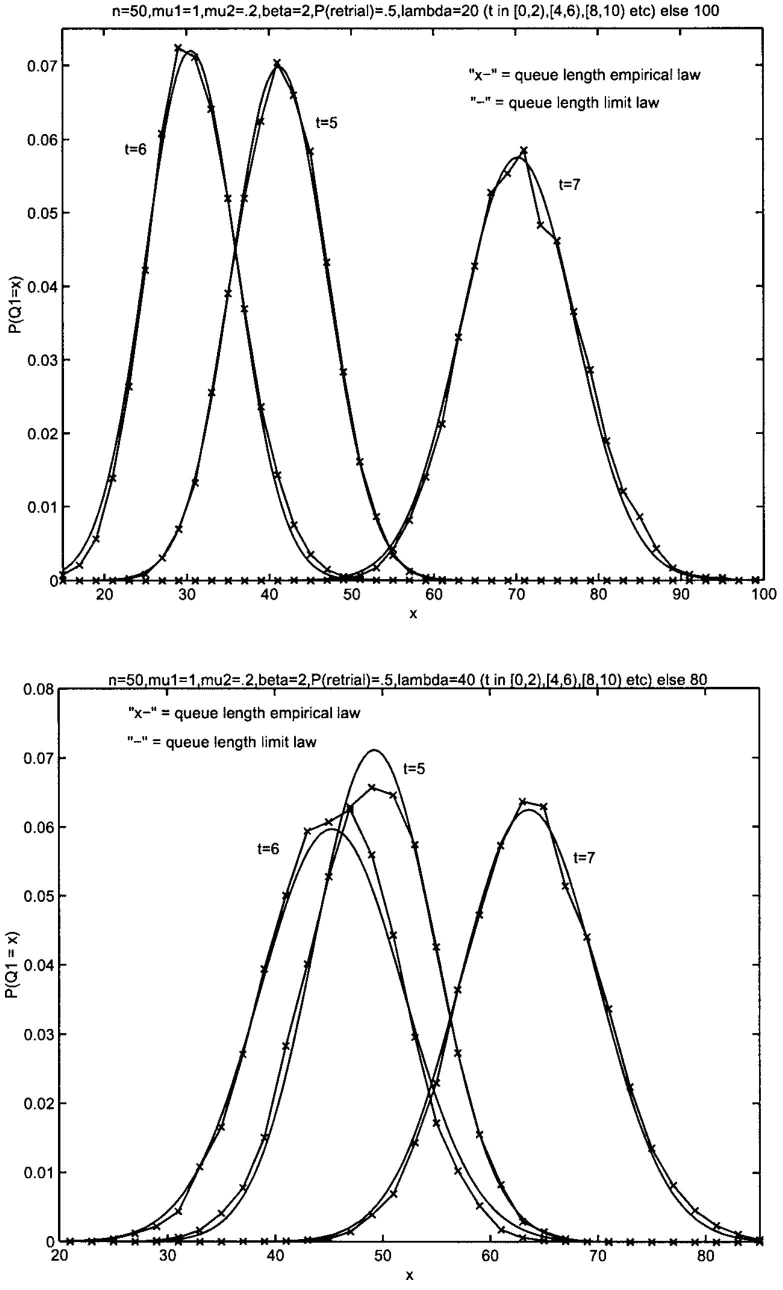

Figure 4. Numerical example: Empirical density of $Q_1(t)$ at times $t = 5, 6, 7$ versus the same from its diffusion approximation for the 20/100 and 40/80 square wave cases.

4. Empirical average of the virtual waiting time versus its fluid approximation (figure 5).
5. Empirical variance of the virtual waiting time versus the variance of its diffusion approximation (figure 6).
6. Empirical density of the virtual waiting time versus its Gaussian approximation (figure 7).

We see that all our approximations for the queue length processes are very good for both cases 20/100 and 40/80. However, in figures 5 and 6, describing the waiting time at the service node, readers can easily notice the following two features:

(a) For the underloaded time intervals the approximation formulas for both the mean and variance of the waiting time $W_1(t)$ are equal to 0. The simulation results for the 20/100 case do agree with this approximation. In the 40/80 case however, the mean and variance, although small indeed, clearly stay away from 0.

(b) At the time points when the service node enters an overloaded interval, there is a strange "spike" in the theoretical variance of the waiting time.

Both features are due to the same simple fact that our approximations for each time t have a *different form* depending on whether t is underloaded or overloaded. The approximations for the underloaded t implicitly assume that the probability of nonzero waiting time is negligible; and the approximations for the overloaded t assume that this probability is close to 1. These assumptions are indeed *asymptotically* correct, as the system scale (the number of servers and the input rate) increases to infinity. However, for a system of a fixed size, the *closer* the system is at time t to the critically loaded phase (when $Q_1^{(0)}(t)$ is equal to n_t), the worse those assumptions are.

Therefore, the feature (a) is explained by the fact that in the 40/80 case, $Q_1^{(0)}(t)$ remains "too close" to $n_t = 50$, while in the 20/100 case it does not. This rule of thumb is supported by the fact that equations (2.15)–(2.19) are not the *general set* of differential equations for the mean and covariance of the diffusion process. We only obtain these autonomous differential equations when the condition $Q_1^{(0)}(t) = n_t$ holds for a set of time points that have measure zero. For example, if this condition does not hold, then equations (2.15) and (2.16) are really of the form

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{E}\big[Q_1^{(1)}(t)\big] = {}&\big(\mu_t^1 \mathbb{1}_{\{Q_1^{(0)}(t)\leqslant n_t\}} + \beta_t \mathbb{1}_{\{Q_1^{(0)}(t)>n_t\}}\big)\,\mathsf{E}\big[Q_1^{(1)}(t)^-\big] \\ &- \big(\mu_t^1 \mathbb{1}_{\{Q_1^{(0)}(t)<n_t\}} + \beta_t \mathbb{1}_{\{Q_1^{(0)}(t)\geqslant n_t\}}\big)\,\mathsf{E}\big[Q_1^{(1)}(t)^+\big] + \mu_t^2\,\mathsf{E}\big[Q_2^{(1)}(t)\big]\end{aligned} \tag{4.1}$$

and

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{E}\big[Q_2^{(1)}(t)\big] = {}&\beta_t(1-\psi_t)\big(\mathsf{E}\big[Q_1^{(1)}(t)^+\big]\mathbb{1}_{\{Q_1^{(0)}(t)\geqslant n_t\}} - \mathsf{E}\big[Q_1^{(1)}(t)^-\big]\mathbb{1}_{\{Q_1^{(0)}(t)>n_t\}}\big) \\ &- \mu_t^2\,\mathsf{E}\big[Q_2^{(1)}(t)\big]\end{aligned} \tag{4.2}$$

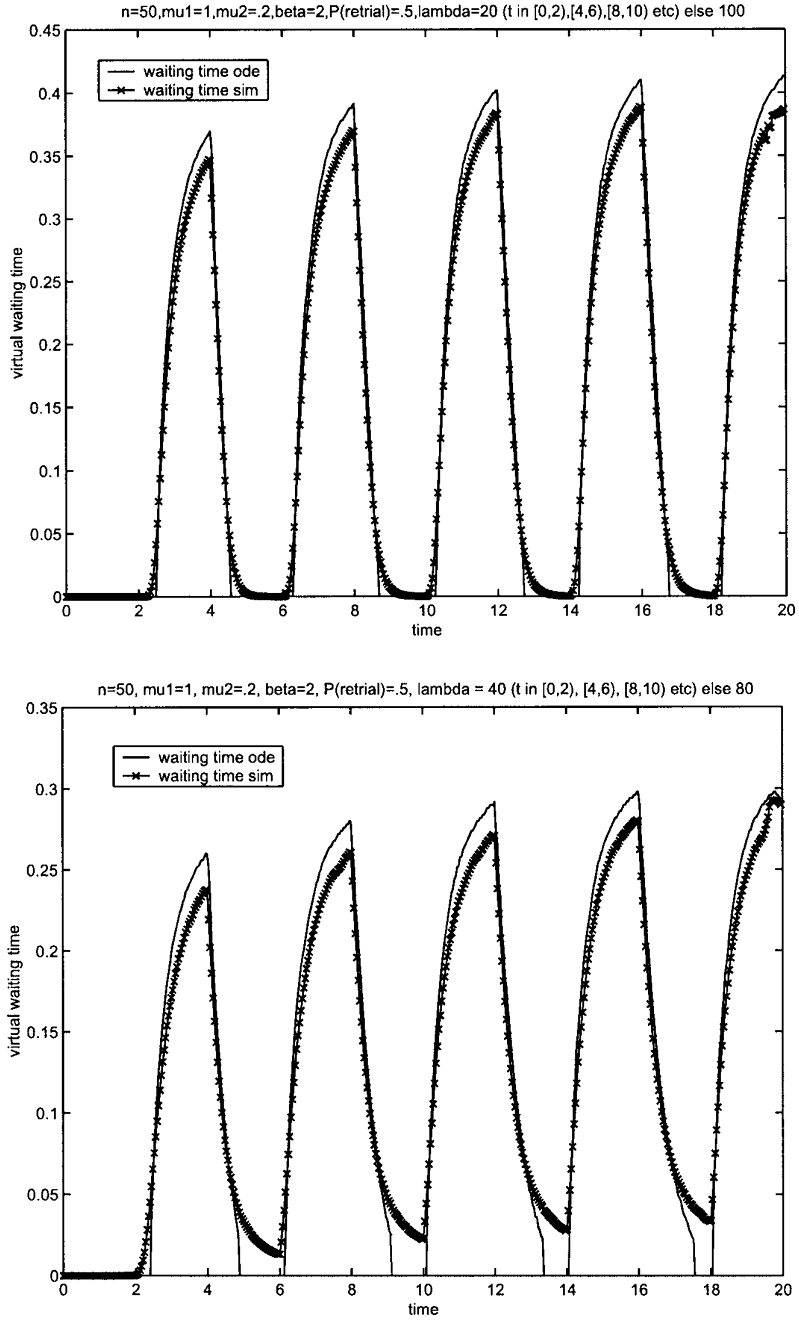

Figure 5. Numerical example: Empirical average of the virtual waiting time versus its fluid approximation for the 20/100 and 40/80 square wave cases.

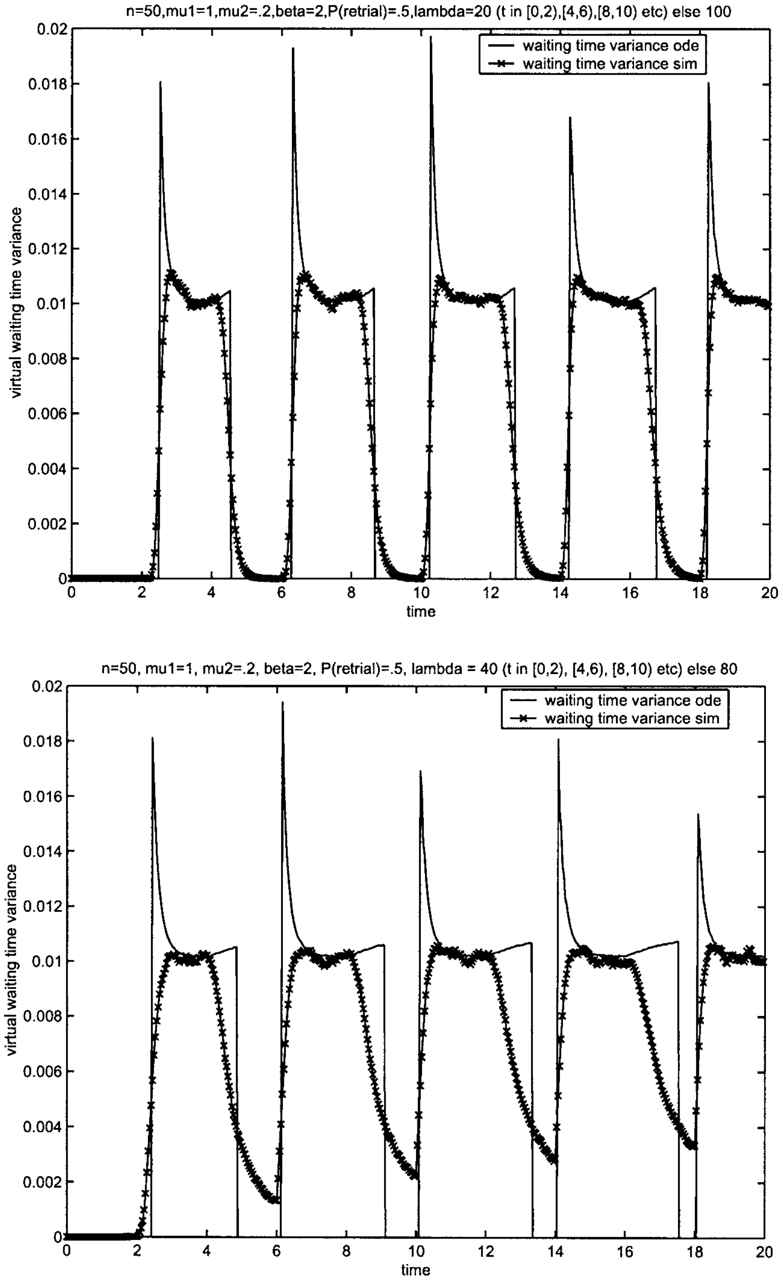

Figure 6. Numerical example: Empirical variance of the virtual waiting time versus the same from its diffusion approximation for the 20/100 and 40/80 square wave cases.

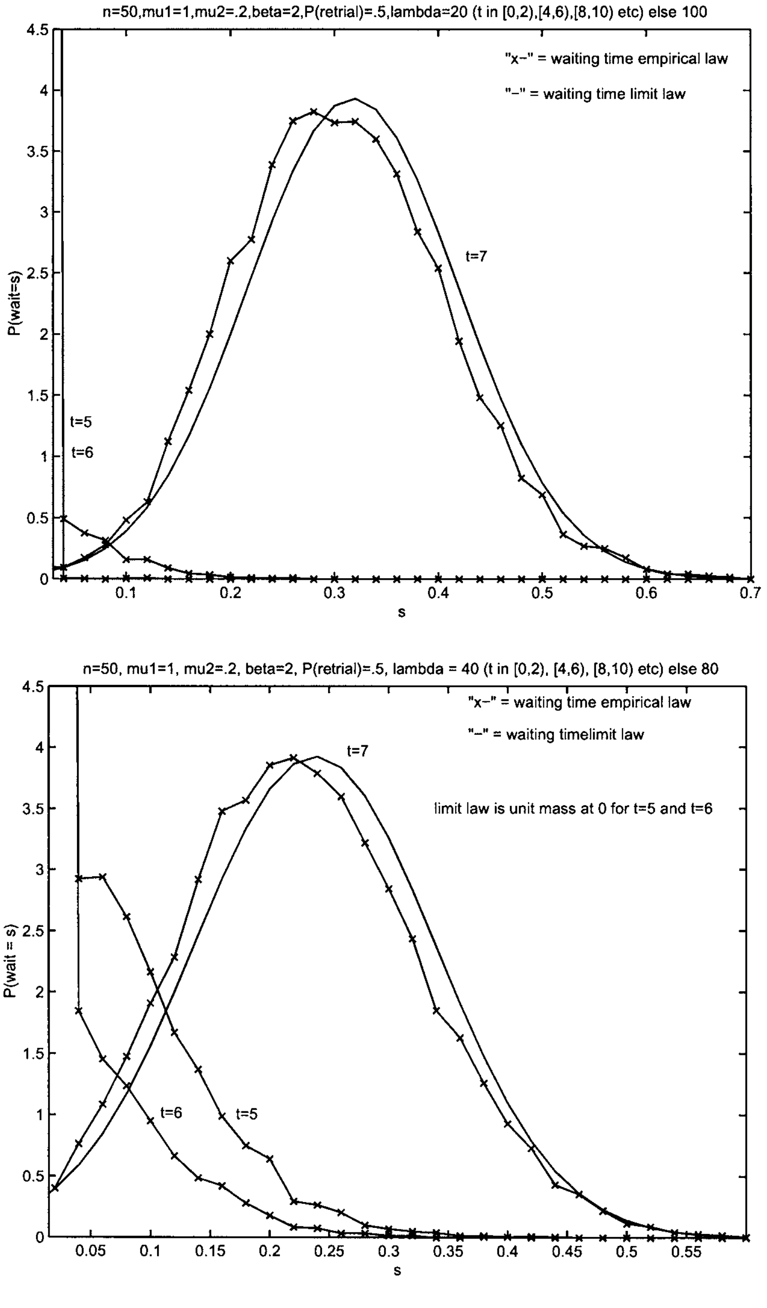

Figure 7. Numerical example: Empirical density of the virtual waiting time versus the same from its diffusion approximation for the 20/100 and 40/80 square wave cases.

and the equations for the covariance matrix have a similar form. Therefore, when $Q_1^{(0)}$ "lingers" to close to n, we see that the autonomous differential equations may not be capturing the true mean and covariance behavior of the diffusion approximation. The behavior described in (b) can also be explained by the "breakdown" of the approximation assumptions for time points in the vicinity of the critically loaded phase. The spike in the variance would indeed be observed if the scale of the system were larger.

Appendix. Markovian service networks

Our model is a special case of a *Markovian service network* (see [Mandelbaum et al., 4]). Given a finite dimensional vector space $\mathbb{V}$ that contains our state space, a finite index set I, transition vectors $\mathbf{v}_i$, rate functions $\alpha_t(\cdot; i)$ that are Lipschitz functions of $\mathbb{V}$ and locally integrable functions of time, we can uniquely define the Markov process $\{\mathbf{Q}(t) \mid t \geqslant 0\}$ by the equation

$$\mathbf{Q}(t) = \mathbf{Q}(0) + \sum_{i \in I} \Pi_i \left(\int_0^t \alpha_s\big(\mathbf{Q}(s); i\big)\, \mathrm{d}s \right) \mathbf{v}_i, \tag{A.1}$$

where the Π_i are an i.i.d. family of standard Poisson processes. Given $\eta > 0$ we can now define $\mathbf{Q}^\eta$ to be a scaled version of this process where

$$\mathbf{Q}^\eta(t) = \mathbf{Q}^\eta(0) + \sum_{i \in I} \Pi_i \left(\int_0^t \eta\alpha_s \left(\frac{1}{\eta} \mathbf{Q}^\eta(s); i \right) \mathrm{d}s \right) \mathbf{v}_i. \tag{A.2}$$

In [Mandelbaum et al., 4], we proved the following functional strong law of large numbers limit theorem.

Theorem A.1. If $\lim_{\eta\to\infty} (1/\eta)\mathbf{Q}^\eta(0) = \mathbf{Q}^{(0)}(0)$ holds a.s., then

$$\lim_{\eta\to\infty} \frac{1}{\eta}\mathbf{Q}^\eta = \mathbf{Q}^{(0)} \quad \text{a.s.} \tag{A.3}$$

where the convergence is uniform on compact sets of t, $\mathbf{Q}^{(0)} = \{\mathbf{Q}^{(0)}(t) \mid t \geqslant 0\}$ is uniquely determined by $\mathbf{Q}^{(0)}(0)$ and the autonomous differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Q}^{(0)}(t) = \boldsymbol{\alpha}_t\big(\mathbf{Q}^{(0)}(t)\big) \tag{A.4}$$

with

$$\boldsymbol{\alpha}_t(\mathbf{x}) \equiv \sum_{i \in I} \alpha_t(\mathbf{x}; i)\mathbf{v}_i \tag{A.5}$$

for all $\mathbf{x} \in \mathbb{V}$

For the diffusion limit, we first need to define the *tensor product* of vectors **x** and **y** in $\mathbb{V}$ to be

$$\mathbf{x} \otimes \mathbf{y} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_n \\ \vdots & \vdots & \dots & \vdots \\ x_n y_1 & x_n y_2 & \dots & x_n y_n \end{bmatrix} \tag{A.6}$$

where $\mathbf{x} = [x_1, x_2, \dots, x_n]$ and $\mathbf{y} = [y_1, y_2, \dots, y_n]$. Vectors are rank one tensors and the above array is a rank two tensor. The vector space of rank two tensors is the finite linear sum of all products $\mathbf{x} \otimes \mathbf{y}$. We can use the tensor product to define the *covariance matrix* of two random vectors $\mathbf{X} = [X_1, X_2, \dots, X_n]$ and $\mathbf{Y} = [Y_1, Y_2, \dots, Y_n]$ to be

$$\mathsf{Cov}[\mathbf{X}, \mathbf{Y}] = \mathsf{E}[\mathbf{X} \otimes \mathbf{Y}] - \mathsf{E}[\mathbf{X}] \otimes \mathsf{E}[\mathbf{Y}], \tag{A.7}$$

where we define $\mathsf{Cov}[\mathbf{X}] = \mathsf{Cov}[\mathbf{X}, \mathbf{X}]$.

If **A** and **B** are defined to be square matrices that map $\mathbb{V}$ into itself, then we define $\mathbf{A} \otimes \mathbf{B}$ to be the *Kronecker product* of **A** and **B** (see [Horn and Johnson, 1]). The object $\mathbf{A} \otimes \mathbf{B}$ is a linear transformation on the family of rank two tensors into themselves where

$$\mathbf{x} \otimes \mathbf{y} \mapsto (\mathbf{xA}) \otimes (\mathbf{yB}) \tag{A.8}$$

which we will denote as $(\mathbf{x} \otimes \mathbf{y}) \circ (\mathbf{A} \otimes \mathbf{B})$. If we view $\mathbf{x} \otimes \mathbf{y}$ as a matrix C, then in terms of matrix multiplication we have

$$(\mathbf{x} \otimes \mathbf{y}) \circ (\mathbf{A} \otimes \mathbf{B}) = (\mathbf{xA}) \otimes (\mathbf{yB}) = \mathbf{A}^{\mathrm{T}}\mathbf{CB}, \tag{A.9}$$

where $\mathbf{A}^{\mathrm{T}}$ is the matrix transpose of **A**.

Now we state the general functional central limit theorem.

Theorem A.2. If $\lim_{\eta\to\infty} \sqrt{\eta}((1/\eta)\mathbf{Q}^{\eta}(0) - \mathbf{Q}^{(0)}(0)) = \mathbf{q}^{(1)}(0)$ holds, where $\mathbf{Q}^{(1)}(0)$ is a constant,

$$\lim_{\eta\to\infty} \sqrt{\eta}\left(\frac{1}{\eta}\mathbf{Q}^{\eta} - \mathbf{Q}^{(0)}\right) \overset{\mathrm{d}}{=} \mathbf{Q}^{(1)}. \tag{A.10}$$

where $\mathbf{Q}^{(1)} = \{\mathbf{Q}^{(1)}(t) \mid t \geqslant 0\}$ is a diffusion process and this is a convergence in distribution of the stochastic processes in an appropriate functional space [Mandelbaum et al., 4].

Moreover, if $\boldsymbol{\alpha}_t(\cdot)$ is differentiable at $\mathbf{Q}^{(0)}(t)$ for almost all t, then $\mathbf{Q}^{(1)}$ is a Gaussian process and its mean vector and covariance matrix are the unique solutions to the autonomous differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{E}\left[\mathbf{Q}^{(1)}(t)\right] = \mathsf{E}\left[\mathbf{Q}^{(1)}(t)\right] D\boldsymbol{\alpha}_t\left(\mathbf{Q}^{(0)}(t)\right), \tag{A.11}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{Cov}\left[\mathbf{Q}^{(1)}(t)\right] = \mathsf{Cov}\left[\mathbf{Q}^{(1)}(t)\right] \circ \left(D\boldsymbol{\alpha}_t\big(\mathbf{Q}^{0}(t)\big) \otimes \mathbf{I} + \mathbf{I} \otimes D\boldsymbol{\alpha}_t\big(\mathbf{Q}^{(0)}(t)\big)\right) + \boldsymbol{\alpha}_t\big(\!\big(\mathbf{Q}^{(0)}(t)\big)\!\big) \tag{A.12}$$

where $D\boldsymbol{\alpha}_t\ (\mathbf{Q}^{(0)}(t))$ is the Jacobian of $\boldsymbol{\alpha}_t(\cdot)$ when differentiated at $\mathbf{Q}^{(0)}(t)$ and

$$\boldsymbol{\alpha}_t((\mathbf{x})) \equiv \sum_{i \in I} \alpha_t(\mathbf{x};\, i)\mathbf{v}_i \otimes \mathbf{v}_i \tag{A.13}$$

for all $\mathbf{x} \in \mathbb{V}$. Finally, for all $s < t$

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{Cov}\left[\mathbf{Q}^{(1)}(s), \mathbf{Q}^{(1)}(t)\right] = \mathsf{Cov}\left[\mathbf{Q}^{(1)}(s), \mathbf{Q}^{(1)}(t)\right] \circ \left(\mathbf{I} \otimes D\boldsymbol{\alpha}_t\big(\mathbf{Q}^{(0)}(t)\big)\right). \tag{A.14}$$

Proof of theorem 2.2. The formulas follow from the general theorems for Markovian service networks. Here we write out these general equations for the two-dimensional case. Viewing $\mathbf{Q}^{(1)}$ as a two-dimensional row vector, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{E}\left[\mathbf{Q}^{(1)}(t)\right] = \mathsf{E}\left[\mathbf{Q}^{(1)}(t)\right]\mathbf{A}_t \tag{A.15}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{Cov}\left[\mathbf{Q}^{(1)}(t)\right] = \mathsf{Cov}\left[\mathbf{Q}^{(1)}(t)\right]\mathbf{A}_t + \mathbf{A}_t^{\mathrm{T}}\,\mathsf{Cov}\left[\mathbf{Q}^{(1)}(t)\right] + \mathbf{B}_t, \tag{A.16}$$

where

$$\mathsf{Cov}\left[\mathbf{Q}^{(1)}(t)\right] = \begin{bmatrix} \mathsf{Var}\left[Q_1^{(1)}(t)\right] & \mathsf{Cov}\left[Q_1^{(1)}(t),\, Q_2^{(1)}(t)\right] \\ \mathsf{Cov}\left[Q_1^{(1)}(t),\, Q_2^{(1)}(t)\right] & \mathsf{Var}\left[Q_2^{(1)}(t)\right] \end{bmatrix}, \tag{A.17}$$

$$\mathbf{A}_t = \begin{bmatrix} a_t^{11} & a_t^{12} \\ a_t^{21} & a_t^{22} \end{bmatrix}, \qquad \mathbf{B}_t = \begin{bmatrix} b_t^{11} & b_t^{12} \\ b_t^{12} & a_t^{22} \end{bmatrix}. \tag{A.18}$$

Note that $\mathbf{A}_t$ is not necessarily a symmetric matrix but $\mathbf{B}_t$ always is. Writing these differential equations out explicitly gives us

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{E}\left[Q_1^{(1)}(t)\right] = a_t^{11}\,\mathsf{E}\left[Q_1^{(1)}(t)\right] + a_t^{21}\mathsf{E}\left[Q_2^{(1)}(t)\right], \tag{A.19}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{E}\left[Q_2^{(1)}(t)\right] = a_t^{12}\,\mathsf{E}\left[Q_1^{(1)}(t)\right] + a_t^{22}\mathsf{E}\left[Q_2^{(1)}(t)\right], \tag{A.20}$$

and finally,

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{Var}\left[Q_1^{(1)}(t)\right] = 2a_t^{11}\,\mathsf{Var}\left[Q_1^{(1)}(t)\right] + 2a_t^{21}\,\mathsf{Cov}\left[Q_1^{(1)}(t),\, Q_2^{(1)}(t)\right] + b_t^{11}, \tag{A.21}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{Var}\left[Q_2^{(1)}(t)\right] = 2a_t^{11}\,\mathsf{Var}\left[Q_2^{(1)}(t)\right] + 2a_t^{12}\,\mathsf{Cov}\left[Q_1^{(1)}(t),\, Q_2^{(1)}(t)\right] + b_t^{22}, \tag{A.22}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{Cov}\big[Q_1^{(1)}(t),\,Q_2^{(1)}(t)\big]=a_t^{12}\,\mathsf{Var}\big[Q_1^{(1)}(t)\big]+a_t^{21}\,\mathsf{Var}\big[Q_2^{(1)}(t)\big] \\ +\big(a_t^{11}+a_t^{22}\big)\,\mathsf{Cov}\big[Q_1^{(1)}(t),\,Q_2^{(1)}(t)\big]=b_t^{12}. \quad \text{(A.23)}$$

Finally, to tailor this central limit theorem to the retrial model, observe that functions like $f(x) = x \wedge n$ and $g(x) = (x - n)^+$ are differentiable everywhere, except when $x = n$. □

References

[1] R.A. Horn and C.R. Johnson, *Matrix Analysis* (Cambridge Univ. Press, New York, 1985).

[2] I. Karatzas and S.E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed. (Springer, New York, 1991).

[3] A. Mandelbaum and W.A. Massey, Strong approximations for time dependent queues, Mathematics of Operations Research 20(1) (1995) 33–64.

[4] A. Mandelbaum, W.A. Massey and M.I. Reiman, Strong approximations for Markovian service networks, Queueing Systems 30 (1998) 149–201.

[5] A. Mandelbaum, W.A. Massey, M.I. Reiman and B. Rider, Time varying multiserver queues with abandonment and retrials, in: *ITC-16,* Edinburgh, Scotland, 1999.

[6] A. Mandelbaum, W.A. Massey, M.I. Reiman and A.L. Stolyar, Waiting time asymptotics for time varying multiserver queues with abandonment and retrials, in: *Proc. of the Allerton Conference*, 1999.

[7] A. Mandelbaum, W.A. Massey, M.I. Reiman and A.L. Stolyar, Waiting time asymptotics for multiserver, nonstationary Jackson networks with abandonment, in preparation.

[8] A. Puhalskii, On the invariance principle for the first passage time, Mathematics of Operations Research 19 (1994) 946–954.

[9] R.W. Wolff, *Stochastic Modeling and the Theory of Queues* (Prentice-Hall, Englewood Cliffs, NJ, 1989).