Queue Lengths and Waiting Times for Multiserver Queues with Abandonment and Retrials[1]

Avi Mandelbaum
Technion Institute
Haifa, 32000, ISRAEL
avim@tx.technion.ac.il

William A. Massey
Bell Laboratories
Murray Hill, NJ 07974, U.S.A.
will@research.bell-labs.com

Martin I. Reiman
Bell Laboratories
Murray Hill, NJ 07974, U.S.A.
marty@research.bell-labs.com

Brian Rider
Courant Institute
New York, NY 10012-1185, U.S.A.
riderb@cims.nyu.edu

Alexander Stolyar
Bell Laboratories
Murray Hill, NJ 07974, U.S.A.
stolyar@research.bell-labs.com

April 7, 2000

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 for both the queue length and virtual waiting time processes arising in this model.

These approximations, which are justified by limit theorems where the arrival rate and number of servers grow large, are compared to simulations, and perform extremely well.

Keywords: Call Centers, Fluid Approximations, Diffusion Approximations, Multiserver Queues, Queues with Abandonment, Virtual Waiting Time, Queues with Retrials, Nonstationary Queues.

[1]Submitted to the Selected Proceedings of the Fifth INFORMS Telecommunications Conference.

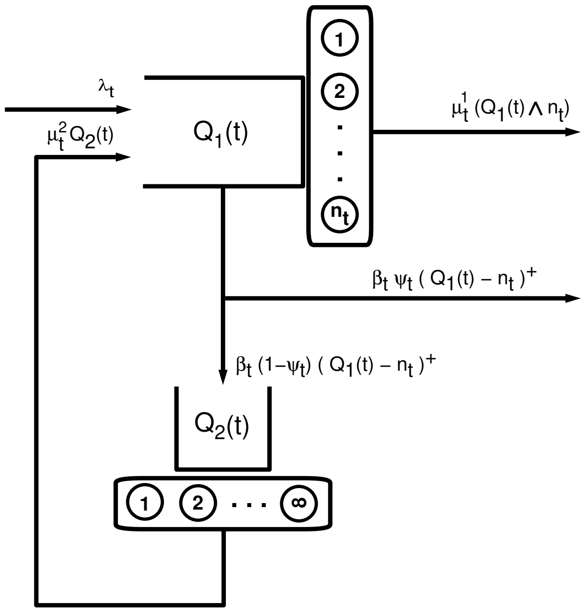

Figure 1: The abandonment queue with retrials.

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 [3], 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. (Customers effectively serve themselves at the retrial pool.) New customers arrive to the service node as a Poisson process of rate λ_t. Customers arriving to find an idle server are taken into service that has 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 service node, these customers are treated the same as new customers. Our focus is the two-dimensional, continuous time Markov chain $\mathbf{Q}(t) = \big(Q_1(t), Q_2(t)\big)$ 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), where W(t) is the time that an infinitely patient customer, arriving 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 the queue length and virtual waiting time process. 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 variablity of parameters, specifically the arrival rate, cannot realistically be ignored. More discussion of this motivation is contained in [4].

Fluid and diffusion limits for the queue length process arising in this model were proved in [3]. In [4] 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 [6]. (These results are described in [5], 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 length and virtual waiting time, comparing the fluid approximations to simulations. We next provide numerical results for the diffusion approximations. Using equations originally obtained in [3], we calculate the covariance matrix of the queue length diffusion, and compare it to simulations. Using a result from [6] that provides conditions under which the queue length diffusion is a Gaussian process we also obtain an approximation for the queue length distribution. We are similarly able to obtain an approximation for the virtual waiting time distribution. 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 and virtual waiting time processes. We also state the relevant limit theorems in a form that provide the information we need for our approximations. Section 3 contains numerical examples comparing our approximations with simulation results. Section 4 is an appendix that provides some background on Markovian service networks.

2 The Model and Limit Theorems

2.1 Basic Model and Queue Length Asymptotics

The sample paths of the queue length process $\mathbf{Q}(t) = (Q_1(t), Q_2(t))$ are uniquely determined by the relations

$$\begin{aligned} Q_1(t) &= Q_1(0) + \Pi_{21}^c\left(\int_0^t Q_2(s)\mu_s^2 ds\right) - \Pi_{12}^b\left(\int_0^t \left(Q_1(s)-n_s\right)^+\beta_s(1-\psi_s)ds\right) \\ &\quad +\Pi^a\left(\int_0^t \lambda_s ds\right) - \Pi^b\left(\int_0^t \left(Q_1(s)-n_s\right)^+\beta_s\psi_s ds\right) - \Pi^c\left(\int_0^t \left(Q_1(s)\wedge n_s\right)\mu_s^1 ds\right) \end{aligned} \tag{2.1}$$

and

$$Q_2(t) = Q_2(0) + \Pi_{12}^b\left(\int_0^t \left(Q_1(s)-n_s\right)^+\beta_s(1-\psi_s)ds\right) - \Pi_{21}^c\left(\int_0^t Q_2(s)\mu_s^2 ds\right), \tag{2.2}$$

where Π^a, Π^b, Π^c, Π_{12}^b, and Π_{21}^c are five given mutually independent, standard (mean rate 1), Poisson processes and $\lambda, \beta, \mu^1, \mu^2, \psi, n$ are locally integrable functions of time [3]. Here $x \wedge y = \min(x, y)$ and $x^+ = \max(x, 0)$ for all real x and y.

We are interested 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 $Q_i^\eta(0) = \lceil \eta Q_i^{(0)}(0) + \sqrt{\eta} Q_i^{(1)}(0) \rceil + o(\sqrt{\eta})$ 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

$$\begin{aligned} Q_1^\eta(t) \;=\; & Q_1^\eta(0) + \Pi_{21}^c\left(\int_0^t Q_2^\eta(s)\mu_s^2 ds\right) - \Pi_{12}^b\left(\int_0^t \left(Q_1^\eta(s) - \eta n_s\right)^+ \beta_s(1-\psi_s)ds\right) \\ & + \Pi^a\left(\int_0^t \eta\lambda_s ds\right) - \Pi^b\left(\int_0^t \left(Q_1^\eta(s) - \eta n_s\right)^+ \beta_s\psi_s ds\right) - \Pi^c\left(\int_0^t \left(Q_1^\eta(s) \wedge (\eta n_s)\right)\mu_s^1 ds\right) \end{aligned} \tag{2.3}$$

and

$$Q_2^\eta(t) = Q_2^\eta(0) + \Pi_{12}^b\left(\int_0^t \left(Q_1^\eta(s) - \eta n_s\right)^+ \beta_s(1-\psi_s)ds\right) - \Pi_{21}^c\left(\int_0^t Q_2^\eta(s)\mu_s^2 ds\right). \tag{2.4}$$

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

$$\lim_{\eta\to\infty} \frac{1}{\eta}\mathbf{Q}^\eta(0) = \mathbf{Q}^{(0)}(0) \quad \text{a.s.}, \tag{2.5}$$

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

Theorem 2.1 *We have*

$$\lim_{\eta\to\infty} \frac{1}{\eta}\mathbf{Q}^\eta = \mathbf{Q}^{(0)} \quad \text{a.s.} \tag{2.6}$$

where the convergence is uniform on compact sets of t*. Moreover,* $\mathbf{Q}^{(0)} = \left\{ \mathbf{Q}^{(0)}(t) \,\middle|\, t \geq 0 \right\}$ *is uniquely determined by* $\mathbf{Q}^{(0)}(0)$ *and the autonomous differential equations*

$$\frac{d}{dt}Q_1^{(0)}(t) = \lambda_t + \mu_t^2 Q_2^{(0)}(t) - \mu_t^1\left(Q_1^{(0)}(t) \wedge n_t\right) - \beta_t\left(Q_1^{(0)}(t) - n_t\right)^+ \tag{2.7}$$

and

$$\frac{d}{dt}Q_2^{(0)}(t) = \beta_t(1-\psi_t)\left(Q_1^{(0)}(t) - n_t\right)^+ - \mu_t^2 Q_2^{(0)}(t). \tag{2.8}$$

This theorem states rigorously that $\mathbf{Q}^\eta \approx \eta\mathbf{Q}^{(0)}$ 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 $X \stackrel{d}{=} Y$. If $\{ X_n \mid n \geq 0 \}$ converges in distribution to Y, we denote this by $\lim_{n\to\infty} X_n \stackrel{d}{=} Y$. The fluid approximation can be refined using the following functional central limit theorem, as proved in [3]. We make the following assumptions for the initial conditions

$$\lim_{\eta\to\infty} \sqrt{\eta}\left(\frac{1}{\eta}\mathbf{Q}^\eta(0) - \mathbf{Q}^{(0)}(0)\right) \stackrel{d}{=} \mathbf{Q}^{(1)}(0), \tag{2.9}$$

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

Theorem 2.2 *We have*

$$\lim_{\eta\to\infty} \sqrt{\eta}\left(\frac{1}{\eta}\mathbf{Q}^{\eta} - \mathbf{Q}^{(0)}\right) \stackrel{d}{=} \mathbf{Q}^{(1)}. \tag{2.10}$$

where $\mathbf{Q}^{(1)} = \left\{ \mathbf{Q}^{(1)}(t) \,\middle|\, t \geq 0 \right\}$ *is a diffusion process and this is a convergence in distribution of the stochastic processes in an appropriate functional space [3].*

Moreover, if the set of time points $\left\{ t \geq 0 \,|\, Q_1^{(0)}(t) = n_t \right\}$ *has measure zero for the retrial model, then* $\left\{ \mathbf{Q}^{(1)}(t) \,\middle|\, t \geq 0 \right\}$ *is Gaussian. The mean vector for* $\mathbf{Q}^{(1)}$ *then solves the set of autonomous differential equations*

$$\frac{d}{dt}\mathsf{E}\left[Q_1^{(1)}(t)\right] = -\left(\mu_t^1 1_{\{Q_1^{(0)}(t)\leq n_t\}} + \beta_t 1_{\{Q_1^{(0)}(t)>n_t\}}\right)\mathsf{E}\left[Q_1^{(1)}(t)\right] + \mu_t^2 \mathsf{E}\left[Q_2^{(1)}(t)\right] \tag{2.11}$$

and

$$\frac{d}{dt}\mathsf{E}\left[Q_2^{(1)}(t)\right] = \beta_t(1-\psi_t)1_{\{Q_1^{(0)}(t)\geq n_t\}}\mathsf{E}\left[Q_1^{(1)}(t)\right] - \mu_t^2 \mathsf{E}\left[Q_2^{(1)}(t)\right]. \tag{2.12}$$

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

$$\begin{aligned}\frac{d}{dt}\mathsf{Var}\left[Q_1^{(1)}(t)\right] = &-2\left(\beta_t 1_{\{Q_1^{(0)}(t)>n_t\}} + \mu_t^1 1_{\{Q_1^{(0)}(t)\leq n_t\}}\right)\mathsf{Var}\left[Q_1^{(1)}(t)\right] + 2\mu_t^2\mathsf{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] \\ &+\lambda_t + \beta_t\left(Q_1^{(0)}(t) - n_t\right)^+ + \mu_t^1\left(Q_1^{(0)}(t)\wedge n_t\right) + \mu_t^2 Q_2^{(0)}(t),\end{aligned} \tag{2.13}$$

$$\begin{aligned}\frac{d}{dt}\mathsf{Var}\left[Q_2^{(1)}(t)\right] &= -2\mu_t^2\mathsf{Var}\left[Q_2^{(1)}(t)\right] + 2\beta_t(1-\psi_t)1_{\{Q_1^{(0)}(t)\geq n_t\}}\mathsf{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] \\ &\quad +\beta_t(1-\psi_t)\left(Q_1^{(0)}(t) - n_t\right)^+ + \mu_t^2 Q_2^{(0)}(t),\end{aligned} \tag{2.14}$$

and

$$\begin{aligned}\frac{d}{dt}\mathsf{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] &= \beta_t(1-\psi_t)1_{\{Q_1^{(0)}(t)\geq n_t\}}\mathsf{Var}\left[Q_1^{(1)}(t)\right] + \mu_t^2\mathsf{Var}\left[Q_2^{(1)}(t)\right] \\ &\quad - \left(\beta_t 1_{\{Q_1^{(0)}(t)>n_t\}} + \mu_t^1 1_{\{Q_1^{(0)}(t)\leq n_t\}} + \mu_t^2\right)\mathsf{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] \\ &\quad -\beta_t(1-\psi_t)\left(Q_1^{(0)}(t) - n_t\right)^+ - \mu_t^2 Q_2^{(0)}(t).\end{aligned} \tag{2.15}$$

This theorem states rigorously that $\mathbf{Q}^{\eta} \approx \eta\mathbf{Q}^{(0)} + \sqrt{\eta}\mathbf{Q}^{(1)}$ and we call $\mathbf{Q}^{(1)}$ the *diffusion approximation* for $\mathbf{Q}^{\eta}$.

Time-varying queues alternate among phases of underloading, critical-loading, and overloading [2]. The set $\left\{ t \,|\, Q_1^{(0)}(t) = n_t \right\}$ corresponds to the times of critical-loading for the service node. The above differential equations must be modified for critical-loading, which is unnecessary here since the hypothesis of Theorem 2.2 applies to all the examples in the following section.

2.2 Virtual Waiting Time in Node 1: Marginal Distribution at a Given Time.

In this subsections we will consider asymptotics for the virtual waiting time process. To do that we need a few additional assumptions which, as we will see, are not very restrictive as far as applications of the results are concerned.

Assumption 2.1 In the interval $[0, \infty)$:

1. Function n_t is *continuously differentiable*;
2. Function μ_t^1 is continuous;
3. Functions μ_t^2 and β_t are bounded on bounded intervals.

Assumption 2.2 will be introduced below when the notations required are in place.

Suppose that we are interested in the waiting time of a *virtual customer* arriving at station 1 at a *fixed* time $\tau \geq 0$. 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 modification. *Suppose, that after time τ, there are no new exogenous arrivals into the system, and any customer departing any station i leaves the system.* In particular, station 1 has no new arrivals after time τ. It just serves 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". Namely, the following results follow directly from those two theorems (and their proofs in [3]).

Denote the arrival and departure processes for station 1 by

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

respectively. Let, by convention, the arrival process includes the customers in node 1 at time 0, so $A^\eta(0) = Q_1^\eta(0)$, $\Delta^\eta(0) = 0$, and $A^\eta(t) - \Delta^\eta(t) = Q_1^\eta(t), \ t \geq 0$.

We then obtain the following *fluid* limit result.

Theorem 2.3 *As a process we have*

$$\lim_{\eta \to \infty} \frac{1}{\eta} (\mathbf{Q}^\eta, A^\eta, \Delta^\eta) = (\mathbf{Q}^{(0)}, A^{(0)}, \Delta^{(0)}) \quad \text{a.s.} \tag{2.16}$$

and this convergence is uniform on compact sets of t. The fluid limit $Q_1^{(0)}(t)$ satisfies equation (2.7) for $t < \tau$. For $t \geq \tau$, we have the following properties:

$$\frac{d}{dt} Q_1^{(0)}(t) = -\mu_t^1 \Big(Q_1^{(0)}(t) \wedge n_t \Big) - \beta_t \Big(Q_1^{(0)}(t) - n_t \Big)^+, \tag{2.17}$$

$A^{(0)}(t) = A^{(0)}(\tau)$ *and* $\Delta^{(0)}$ *is a continuously differentiable non-decreasing function in* $[0, \infty)$.

We also obtain the following *diffusion* limit.

Theorem 2.4 *The following convergence in distribution holds:*

$$\lim_{\eta\to\infty}\sqrt{\eta}\big(\frac{1}{\eta}\mathbf{Q}^\eta-\mathbf{Q}^{(0)},\frac{1}{\eta}A^\eta-A^{(0)},\frac{1}{\eta}\Delta^\eta-\Delta^{(0)}\big)\stackrel{d}{=}(\mathbf{Q}^{(1)},A^{(1)},\Delta^{(1)}). \tag{2.18}$$

Moreover, if the set of time points $\left\{\ t\geq 0\,|\ \ Q_1^{(0)}(t)=n_t\ \right\}$ *has measure zero,* $\left\{\ Q_1^{(1)}(t)\ \middle|\ \ t\geq 0\right\}$ *is Gaussian and for* $t\geq\tau$*,* $\mathsf{Var}[Q_1^{(1)}(t)]$ *solves the differential equation*

$$\begin{aligned}\frac{d}{dt}\mathsf{Var}\left[Q_1^{(1)}(t)\right] &= -2\left(\beta_t 1_{\{Q_1^{(0)}(t)>n_t\}}+\mu_t^1 1_{\{Q_1^{(0)}(t)\leq n_t\}}\right)\mathsf{Var}\left[Q_1^{(1)}(t)\right] \\ &\quad+\beta_t\Big(Q_1^{(0)}(t)-n_t\Big)^+ +\mu_t^1\Big(Q_1^{(0)}(t)\wedge n_t\Big).\end{aligned} \tag{2.19}$$

It follows from the definitions and the above theorem that

$$Q_1^{(1)}(t)=A^{(1)}(t)-\Delta^{(1)}(t)\ . \tag{2.20}$$

Now, let us define the *potential service initiation* process D^η for node 1 by

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

Note that if $Q_1^\eta(t)<\eta n_t$, then $A^\eta(t)<D^\eta(t)$; so the potential service can be "ahead" of arrivals. It follows that

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

where the convergnece is uniform on compact sets of t and $D^{(0)}(t)=\Delta^{(0)}(t)+n_t,t\geq 0$. 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 will make an important (but not very restrictive in majority of applications) additional assumption.

Assumption 2.2. The function $D^{(0)}$ (of t) is continuously differentiable with *strictly positive derivative*, and

$$\lim_{t\to\infty}D^{(0)}(t)>A^{(0)}(\tau)\ . \tag{2.21}$$

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

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

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

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 (2.16) and (2.18) as follows (with all the functions being now defined for $t\geq -T$):

$$\lim_{\eta\to\infty}\frac{1}{\eta}(\mathbf{Q}^\eta,A^\eta,D^\eta)=(\mathbf{Q}^{(0)},A^{(0)},D^{(0)}) \tag{2.22}$$

and

$$\lim_{\eta\to\infty} \sqrt{\eta}(\frac{1}{\eta}\mathbf{Q}^\eta - \mathbf{Q}^{(0)}, \frac{1}{\eta}A^\eta - A^{(0)}, \frac{1}{\eta}D^\eta - D^{(0)}) \stackrel{d}{=} (\mathbf{Q}^{(1)}, A^{(1)}, D^{(1)}) , \tag{2.23}$$

where

$$D^{(1)} = \Delta^{(1)} . \tag{2.24}$$

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 the Assumption 2.2 make the following processes well defined and finite with probability 1 for all sufficiently large η. Let us define, for all $t \geq -T$, the *first attainment* processes

$$S^\eta(t) = \inf\{s \geq -T : \; D^\eta(s) > A^\eta(t)\}$$

and

$$S^{(0)}(t) = \inf\{s \geq -T : \; D^{(0)}(s) > A^{(0)}(t)\}, \tag{2.25}$$

and the *attainment* waiting time processes

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

and

$$W^{(0)}(t) = S^{(0)}(t) - t . \tag{2.26}$$

Denote by $\hat{W}^\eta(\tau)$ the *virtual* waiting time at τ, i.e. the time a "test" customer (in the original non-modified system) arriving in node 1 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 $\hat{W}^\eta(\tau)$ and the attainment waiting time $W^\eta(\tau)$ is simply

$$\hat{W}^\eta(\tau) = W^\eta(\tau)^+ . \tag{2.27}$$

Indeed, note that $W^\eta(\tau)$ (and $W^{(0)}(\tau)$) may be negative. All this means is that $Q_1^\eta(\tau) < \eta n_\tau$, and therefore in this case $\hat{W}^\eta(\tau) = 0$. If $W^\eta(\tau)$ is non-negative, then its value is exactly equal to the virtual waiting time.

It follows directly from Theorem and Corollary in [7] that (2.22), (2.23), and Assumption 2.2, imply the following convergences.

Theorem 2.5 *We have*

$$\lim_{\eta\to\infty} (\frac{1}{\eta}\mathbf{Q}^\eta, \frac{1}{\eta}A^\eta, \frac{1}{\eta}D^\eta, W^\eta) = (\mathbf{Q}^{(0)}, A^{(0)}, D^{(0)}, W^{(0)}) . \tag{2.28}$$

$$\lim_{\eta\to\infty} \sqrt{\eta}(\frac{1}{\eta}\mathbf{Q}^\eta - \mathbf{Q}^{(0)}, \frac{1}{\eta}A^\eta - A^{(0)}, \frac{1}{\eta}D^\eta - D^{(0)}, W^\eta - W^{(0)}) \stackrel{d}{=} (Q^{(1)}, A^{(1)}, D^{(1)}, W^{(1)}) , \tag{2.29}$$

where

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

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

In particular, consider the non-trivial case $S^{(0)}(\tau) \geq \tau$ (which is equivalent to $Q_1^{(0)}(\tau) \geq n_\tau$). Moreover, assume that in $[0, \tau]$, the set of points $\left\{ t \mid \ Q_1^{(0)}(t) = n_t \right\}$ has measure zero. Then we obtain

$$\lim_{\eta \to \infty} W^\eta(\tau) = W^{(0)}(\tau)$$

and

$$\lim_{\eta \to \infty} \sqrt{\eta}(W^\eta(\tau) - W^{(0)}(\tau)) \stackrel{d}{=} W^{(1)}(\tau) = \frac{Q_1^{(1)}(S^{(0)}(\tau))}{d^{(0)}(S^{(0)}(\tau))} \ .$$

where $Q_1^{(1)}(S^{(0)}(\tau))$ is Gaussian with mean and variance computed as follows. Solving equation (2.17) for $Q_1^{(0)}(\cdot)$ in the interval $[\tau, \infty)$, we obtain

$$\frac{d}{dt} Q_1^{(0)}(t) = -\beta_t Q_1^{(0)}(t) + (\beta_t - \mu_t^1) n_t \ , \ t \geq \tau.$$

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

$$S^{(0)}(\tau) = \min\{t \geq \tau \mid Q_1^{(0)}(t) = n_t\} \ .$$

We then compute $\mathsf{Var}[Q_1^{(1)}(S^{(0)}(\tau))]$, where

$$\frac{d}{dt} \mathsf{Var}[Q_1^{(1)}(t)] = -2\beta_t \mathsf{Var}[Q_1^{(1)}(t)] + \beta_t (Q_1^{(0)}(t) - n_t) + \mu_t^1 n_t \quad t \geq \tau. \tag{2.30}$$

Finally, we have the closed formulas

$$Q_1^{(0)}(t) = Q_1^{(0)}(t) \exp\left(-\int_\tau^t \beta_s ds\right) + \int_\tau^t (\beta_s - \mu_s^1) n_s \exp\left(-\int_s^t \beta_r dr\right) ds \tag{2.31}$$

and

$$\begin{aligned} \mathsf{Var}\left[Q_1^{(1)}(S^{(0)}(\tau))\right] \quad &= \quad \mathsf{Var}\left[Q_1^{(1)}(\tau)\right] \exp\left(-\int_\tau^{S^{(0)}(\tau)} 2\beta_s ds\right) \qquad (2.32) \\ &+ \int_\tau^{S^{(0)}(\tau)} \left((Q_1^{(0)}(s) - \beta_s) n_s - \mu_s^1 n_s\right) \exp\left(-\int_s^{S^{(0)}(\tau)} 2\beta_r dr\right) ds. \end{aligned}$$

Remark. In this subsection we derived fluid and diffusion approximations of the marginal distribution of the attainment waiting time, which uniquely determines those for the virtual waiting time, in node 1 *at a given time* $\tau \geq 0$. However, it is shown in [6] that similar asymptotics hold for the attainment waiting time as a *random process* defined for $\tau \in [0, \infty)$. (See also [5] for the formal statement of the results.)

3 Numerical Examples

Our numerical examples cover the case of time-varying behavior only for the external arrival rate λ_t. We set $\mu^1 = 1$, $\mu^2 = 0.2$, and $Q_1(0) = Q_2(0) = 0$ but let n, β, and ψ range over a variety of different constants.

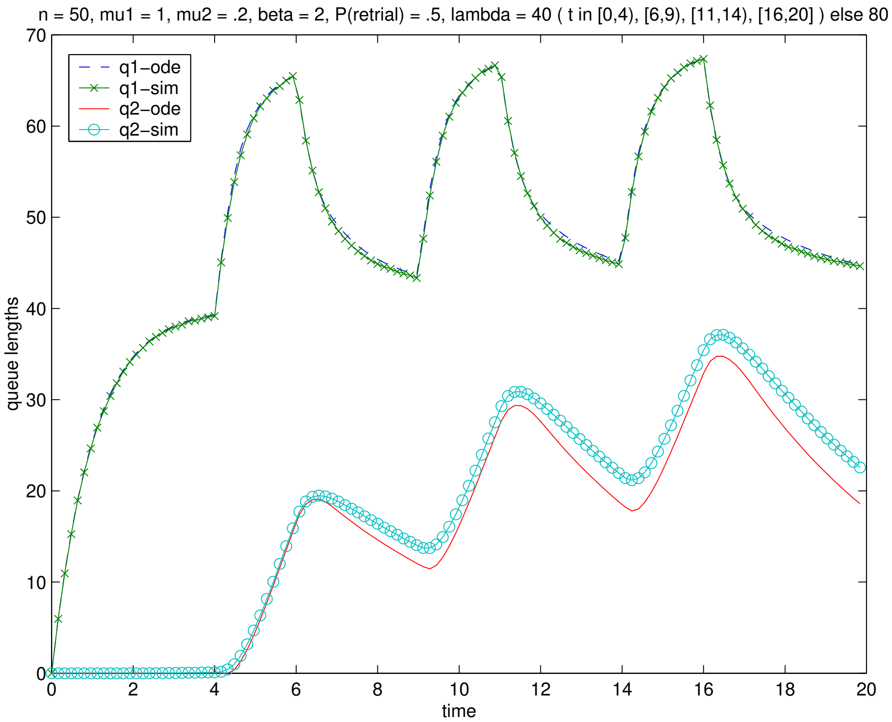

Figure 2: Numerical example: Empirical averages of $Q_1(t)$ and $Q_2(t)$ versus their fluid approximations for square wave case.

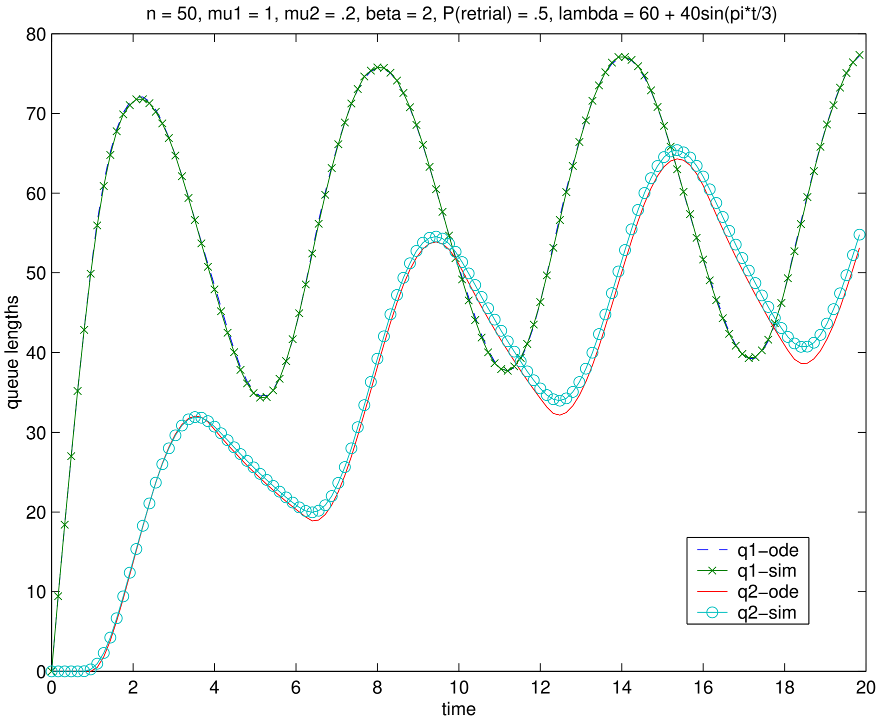

Figure 3: Numerical example: Empirical averages of $Q_1(t)$ and $Q_2(t)$ versus their fluid approximations for sine wave case.

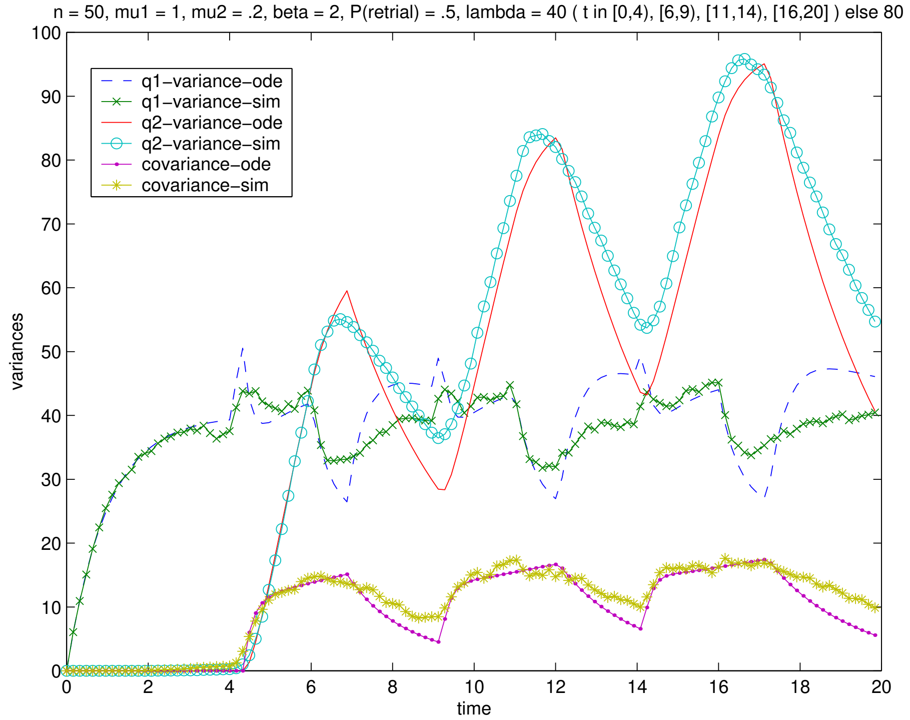

Figure 4: Numerical example: Empirical covariance matrix of queueing process versus the same from its diffusion approximation.

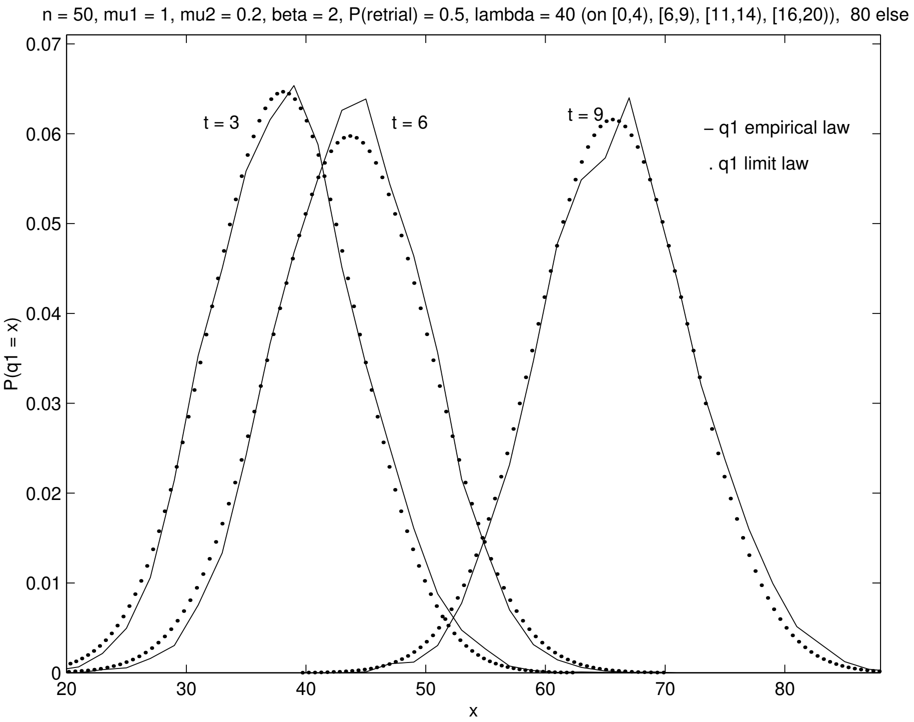

Figure 5: Numerical example: Empirical distribution of Q_1 versus the same from its diffusion approximation.

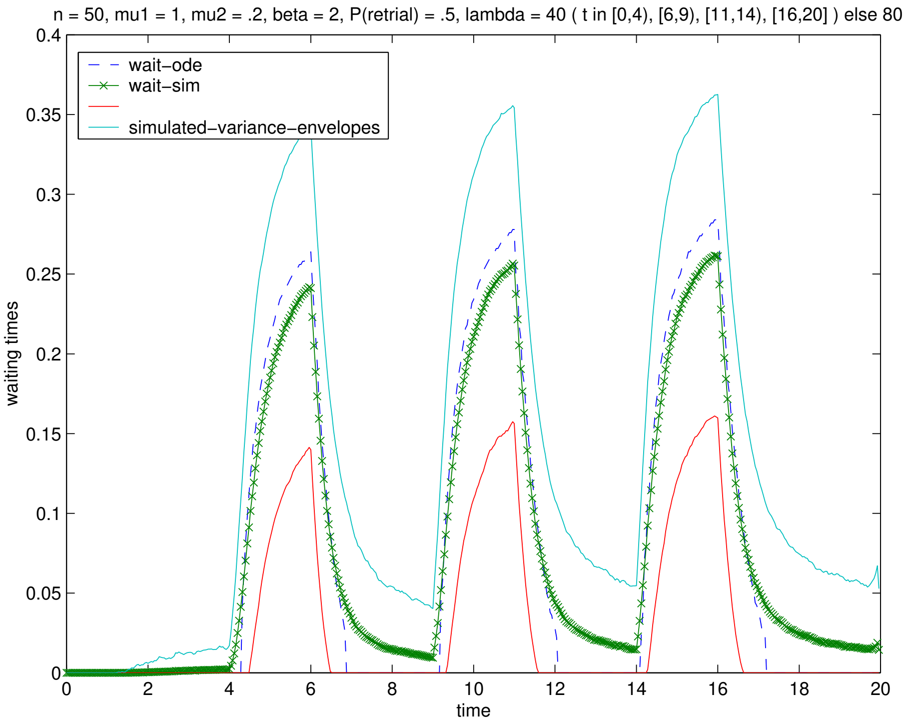

Figure 6: Numerical example: Empirical average of waiting time versus the same from its fluid approximation for square wave case.

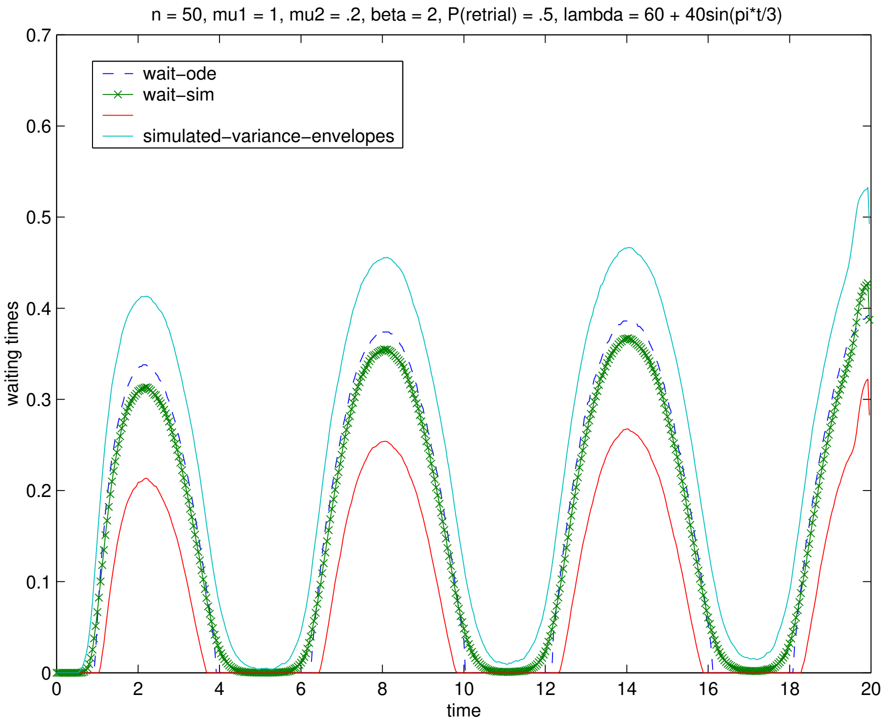

Figure 7: Numerical example: Empirical average of virtual waiting time versus the same from its fluid approximation for sine wave case.

Several examples indicating the accuracy of the fluid approximation for the queue length process were considered in [4]. 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 [5] the accuracy of the fluid approximation for the virtual waiting time was checked for one of the examples from [4] 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 approximation for both queue length and virtual waiting time in some new examples, and also examine the performance of the diffusion approximation for the queue length process.

Details of how the simulations are carried out are contained in [4]. Here we merely point out that we use 5,000 independent replications in each of our experiments.

We first examine the performance of the fluid approximation for the queue length process in two new examples that have square wave and sine wave arrival rates respectively. The square wave, which is not periodic, has $\lambda_t = 40$ for $0 \le t \le 4$, $6 \le t < 9$, $11 \le t < 14$, and $16 \le t \le 20$ with $\lambda_t = 80$ otherwise. Figure 2 contains plots of the fluid approximation for the queue lengths as well as the sample mean, for $0 \le t \le 20$. The fluid approximation for Q_1 is better than that of Q_2, but both are quite good. The sine wave have period 6, with $\lambda_t = 60 + 40\sin(\pi t/3)$. Figure 3 presents the results for this case; the results are similar to those for the square wave.

We consider two quantities associated with the diffusion approximation for queue lengths: variances and distributions. In Figure 4 we present plots of the queue length variances (and covariance) from the diffusion approximation and from the simulation in the square wave example. The accuracy here is not so good as the fluid approximation, particularly for the variance of Q_1. As pointed out in Theorem 2.3, $\mathbf{Q}^{(1)}$ is a Gaussian process. Thus its mean vector and covariance matrix are sufficient to calculate its distribution. Using the distribution to obtain a diffusion approximation for the distribution of $Q_1(t)$, we plot the approximation and empirical distribution for $Q_1(t)$ at $t = 3$, $t = 6$, and $t = 9$ in Figure 5. This approximation is startlingly good.

We now examine the performance of the fluid approximation for the virtual waiting time process in the square and sine wave examples considered above. In Figure 6 we compare the fluid approximation to the simulation average for the square wave, while Figure 7 contains the results for the sine wave. These results are excellent.

4 Appendix

4.1 Markovian Service Networks

Our model is a special case of a *Markovian service network* (see [3]). 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 $\{\, \mathbf{Q}(t) \mid t \geq 0 \,\}$ by the equation

$$\mathbf{Q}(t) = \mathbf{Q}(0) + \sum_{i \in I} \Pi_i \left(\int_0^t \alpha_s\Big(\mathbf{Q}(s); i\Big) ds \right) \mathbf{v}_i, \tag{4.1}$$

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

$$\mathbf{Q}^\eta(t) = \mathbf{Q}^\eta(0) + \sum_{i \in I} \Pi_i \left(\int_0^t \eta \alpha_s\Big(\frac{1}{\eta}\mathbf{Q}^\eta(s); i\Big) ds \right) \mathbf{v}_i. \tag{4.2}$$

In [3], we proved the following functional strong law of large numbers limit theorem

Theorem 4.1 *If* $\lim_{\eta \to \infty} \frac{1}{\eta} \mathbf{Q}^\eta(0) = \mathbf{Q}^{(0)}(0)$ *holds a.s., then*

$$\lim_{\eta \to \infty} \frac{1}{\eta} \mathbf{Q}^\eta = \mathbf{Q}^{(0)} \quad \text{a.s.} \tag{4.3}$$

where the convergence is uniform on compact sets of t*,* $\mathbf{Q}^{(0)} = \left\{\, \mathbf{Q}^{(0)}(t) \,\Big|\, t \geq 0 \right\}$ *is uniquely determined by* $\mathbf{Q}^{(0)}(0)$ *and the autonomous differential equation*

$$\frac{d}{dt}\mathbf{Q}^{(0)}(t) = \boldsymbol{\alpha}_t\left(\mathbf{Q}^{(0)}(t)\right) \tag{4.4}$$

with

$$\boldsymbol{\alpha}_t(\mathbf{x}) \equiv \sum_{i \in I} \alpha_t(\mathbf{x}; i)\mathbf{v}_i. \tag{4.5}$$

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

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

$$\mathbf{x} \otimes \mathbf{y} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \cdots & \vdots \\ x_n y_1 & x_n y_2 & \cdots & x_n y_n \end{bmatrix} \tag{4.6}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ and $\mathbf{y} = [y_1, y_2, \ldots, y_n]$. 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 $\mathbf{X} = [X_1, X_2, \ldots, X_n]$ and $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_n]$ to be

$$\mathsf{Cov}[\mathbf{X}, \mathbf{Y}] = \mathsf{E}[\mathbf{X} \otimes \mathbf{Y}] - \mathsf{E}[\mathbf{X}] \otimes \mathsf{E}[\mathbf{Y}], \tag{4.7}$$

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

If $\mathbf{A}$ and $\mathbf{B}$ 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 $\mathbf{A}$ and $\mathbf{B}$. The object $\mathbf{A} \otimes \mathbf{B}$ is a linear transformation on the family of rank two tensors into themselves where

$$\mathbf{x} \otimes \mathbf{y} \mapsto (\mathbf{x}\mathbf{A}) \otimes (\mathbf{y}\mathbf{B}) \tag{4.8}$$

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 $\mathbf{C}$, then in terms of matrix multiplication we have

$$(\mathbf{x} \otimes \mathbf{y}) \circ (\mathbf{A} \otimes \mathbf{B}) = (\mathbf{x}\mathbf{A}) \otimes (\mathbf{y}\mathbf{B}) = \mathbf{A}^{\mathsf{T}}\mathbf{C}\mathbf{B}, \tag{4.9}$$

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

Now we state the general functional central limit theorem.

Theorem 4.2 *If* $\lim_{\eta\to\infty} \sqrt{\eta}(\frac{1}{\eta}\mathbf{Q}^{\eta}(0) - \mathbf{Q}^{(0)}(0)) = \mathbf{Q}^{(1)}(0)$ *holds, where* $\mathbf{Q}^{(1)}(0)$ *is a constant, then*

$$\lim_{\eta\to\infty} \sqrt{\eta}\left(\frac{1}{\eta}\mathbf{Q}^{\eta} - \mathbf{Q}^{(0)}\right) \stackrel{d}{=} \mathbf{Q}^{(1)}. \tag{4.10}$$

where $\mathbf{Q}^{(1)} = \left\{ \mathbf{Q}^{(1)}(t) \,\middle|\, t \geq 0 \right\}$ *is a diffusion process and this is a convergence in distribution of the stochastic processes in an appropriate functional space [3].*

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*

$$\frac{d}{dt}\mathsf{E}\left[\mathbf{Q}^{(1)}(t)\right] = \mathsf{E}\left[\mathbf{Q}^{(1)}(t)\right] D\boldsymbol{\alpha}_t\left(\mathbf{Q}^{(0)}(t)\right), \tag{4.11}$$

and

$$\frac{d}{dt}\mathsf{Cov}\left[\mathbf{Q}^{(1)}(t)\right] = \mathsf{Cov}\left[\mathbf{Q}^{(1)}(t)\right] \circ \left(D\boldsymbol{\alpha}_t\left(\mathbf{Q}^{(0)}(t)\right) \otimes \mathbf{I} + \mathbf{I} \otimes D\boldsymbol{\alpha}_t\left(\mathbf{Q}^{(0)}(t)\right)\right) + \boldsymbol{\alpha}_t\left(\!\left(\mathbf{Q}^{(0)}(t)\right)\!\right) \tag{4.12}$$

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

$$\boldsymbol{\alpha}_t(\!(\mathbf{x})\!) \equiv \sum_{i\in I} \alpha_t(\mathbf{x}; i)\mathbf{v}_i \otimes \mathbf{v}_i. \tag{4.13}$$

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

$$\frac{d}{dt}\mathsf{Cov}\left[\mathbf{Q}^{(1)}(s), \mathbf{Q}^{(1)}(t)\right] = \mathsf{Cov}\left[\mathbf{Q}^{(1)}(s), \mathbf{Q}^{(1)}(t)\right] \circ \left(\mathbf{I} \otimes D\boldsymbol{\alpha}_t\left(\mathbf{Q}^{(0)}(t)\right)\right). \tag{4.14}$$

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

$$\frac{d}{dt}\mathsf{E}\left[\mathbf{Q}^{(1)}(t)\right] = \mathsf{E}\left[\mathbf{Q}^{(1)}(t)\right] \mathbf{A}_t \tag{4.15}$$

and

$$\frac{d}{dt}\mathsf{Cov}\left[\mathbf{Q}^{(1)}(t)\right] = \mathsf{Cov}\left[\mathbf{Q}^{(1)}(t)\right] \mathbf{A}_t + \mathbf{A}_t^{\mathsf{T}}\mathsf{Cov}\left[\mathbf{Q}^{(1)}(t)\right] + \mathbf{B}_t, \tag{4.16}$$

where

$$\mathsf{Cov}\left[\mathbf{Q}^{(1)}(t)\right] = \begin{bmatrix} \mathsf{Var}\left[Q_1^{(1)}(t)\right] & \mathsf{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] \\ \mathsf{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] & \mathsf{Var}\left[Q_2^{(1)}(t)\right] \end{bmatrix}, \tag{4.17}$$

$$\mathbf{A}_t = \begin{bmatrix} a_t^{11} & a_t^{12} \\ a_t^{21} & a_t^{22} \end{bmatrix}, \quad \mathbf{B}_t = \begin{bmatrix} b_t^{11} & b_t^{12} \\ b_t^{12} & b_t^{22} \end{bmatrix}. \tag{4.18}$$

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

$$\frac{d}{dt}\mathsf{E}\left[Q_1^{(1)}(t)\right] = a_t^{11}\mathsf{E}\left[Q_1^{(1)}(t)\right] + a_t^{21}\mathsf{E}\left[Q_2^{(1)}(t)\right] \tag{4.19}$$

$$\frac{d}{dt}\mathsf{E}\left[Q_2^{(1)}(t)\right] = a_t^{12}\mathsf{E}\left[Q_1^{(1)}(t)\right] + a_t^{22}\mathsf{E}\left[Q_2^{(1)}(t)\right] \tag{4.20}$$

and finally

$$\frac{d}{dt}\mathsf{Var}\left[Q_1^{(1)}(t)\right] = 2a_t^{11}\mathsf{Var}\left[Q_1^{(1)}(t)\right] + 2a_t^{21}\mathsf{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] + b_t^{11} \tag{4.21}$$

$$\frac{d}{dt}\mathsf{Var}\left[Q_2^{(1)}(t)\right] = 2a_t^{22}\mathsf{Var}\left[Q_2^{(1)}(t)\right] + 2a_t^{12}\mathsf{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] + b_t^{22} \tag{4.22}$$

$$\begin{aligned} \frac{d}{dt}\mathsf{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] &= a_t^{12}\mathsf{Var}\left[Q_1^{(1)}(t)\right] + a_t^{21}\mathsf{Var}\left[Q_2^{(1)}(t)\right] \\ &\quad + (a_t^{11} + a_t^{22})\mathsf{Cov}\left[Q_1^{(1)}(t), Q_2^{(1)}(t)\right] + b_t^{12}. \end{aligned} \tag{4.23}$$

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] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus (Second Edition)*. Springer-Verlag, New York, 1991.

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

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

[4] A. Mandelbaum, W. A. Massey, M. I. Reiman, B. Rider. Time Varying Multiserver Queues with Abandonment and Retrials. *ITC-16*, Edinburgh, Scotland, (1999).

[5] A. Mandelbaum, W. A. Massey, M. I. Reiman, A. L. Stolyar. Waiting Time Asymptotics for Time Varying Multiserver Queues with Abandonment and Retrials. *Proceedings of the Allerton Conference*, (1999).

[6] A. Mandelbaum, W. A. Massey, M. I. Reiman, A. L. Stolyar. Waiting Time Asymptotics for Multiserver, Nonstationary Jackson Networks with Abandonment. *In preparation.*

[7] A. Puhalskii. On the Invariance Principle for the First Passage Time. *Mathematics of Operatioins Research*, Vol. 19, (1994), pp. 946-954.

[8] Wolff, R. W. *Stochastic modeling and the theory of queues.* Prentice-Hall, Inc., 1989.