Time Varying Multiserver Queues with Abandonment and Retrials

A. Mandelbaum[a], William A. Massey[b], Martin I. Reiman[c] and Brian Rider[d]

[a]Davidson Faculty of Industrial Engineering and Management, Technion Institute
Haifa, 32000, ISRAEL

[b]Bell Laboratories, Lucent Technologies
Office 2C-320, Murray Hill, NJ 07974, U.S.A.

[c]Bell Laboratories, Lucent Technologies
Office 2C-315, Murray Hill, NJ 07974, U.S.A.

[d]Courant Institute of Mathematical Sciences
New York, NY 10012-1185, U.S.A.

In this paper we consider a multiserver queueing model where waiting customers may abandon and subsequently retry. This model is of particular interest for analyzing performance and setting staffing levels in call centers. All of the parameters (arrival rate, service rate, etc.) are allowed to be time dependent. We propose a simple fluid approximation for the queue length process arising in this model. The fluid approximation, which is obtained as the solution of an intuitively appealing ordinary differential equation, is in fact asymptotically exact as the size of the system (arrival rate and number of servers) grows large. The fluid approximation is compared with simulations for several sets of parameters and performs extremely well.

1. Introduction

Time-varying analytical models of telecommunication systems are notorious for being intractable. This is an unfortunate state of affairs as telecommunication systems typically operate under time-varying conditions. The gap between this "demand" and "supply" has been traditionally circumvented, in practice, in one of two ways: either approximating time-varying behavior by piecewise-constant behavior, and then applying stationary analysis over intervals of "constancy"; or giving up analytical models altogether and carrying out performance analysis based on simulation. The goal of our paper is to demonstrate that time-variability is amenable to analysis, at least within specific regimes of operation. We do this through a model of a single service station that, we believe, is already of considerable practical importance. In fact, this model is a relatively simple special case of the class of models considered in [9], which includes complex service networks that are both time- and state-dependent. The analysis we provide is based on a fluid approximation, which is intuitively appealing, asymptotically exact (in a sense that we make precise), and surprisingly accurate.

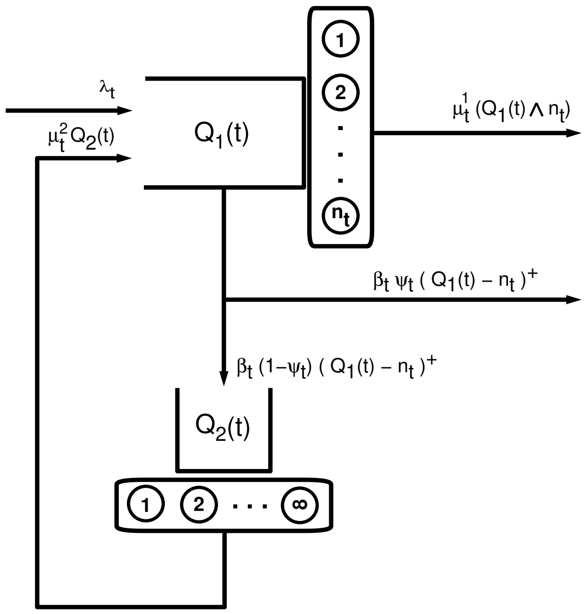

Figure 1. The abandonment queue with retrials.

Our model is a multi-server queue with time-varying parameters, in which customers are impatient and hence abandon after (subjectively) excessive wait. Moreover, obtaining service is important enough for some customers that they return and seek service after experiencing a "time-out". Formally, our model is depicted in Figure 1: there is a single "service" node with n_t, $t \geq 0$, servers. New customers arrive to the service node following 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 a 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. Time variability manifests itself through time-dependent rates for arrivals, abandonments and retrials, as well as a varying number of servers. (It is worth noting that, even if all of these parameters are constant, the model in Figure 1 is analytically intractable.)

We are motivated by the need to develop analytical tools that support performance analysis of *large* telecommunication systems. For convenience, however, we shall speak in terms of telephone *call centers*. These already constitute a multi-billion dollar industry that enjoys a 20% annual growth rate, see Brigandi, Dargon, Sheehan, and Spencer [1]. While abandonments and retrials arise naturally in call centers, they are also prevalent in many other telecommunication systems, as described for example in Boxma and de Waal [3] for abandonments, and Falin and Templeton [4] for retrials. Additional complementing references, that include numerous further leads, are Wolff [12], Grier, Massey, McKoy, and Whitt [6], Mandelbaum and Shimkin [10], and Garnet, Mandelbaum, and Reiman [5]. Articles that specifically address call centers with abandonments and retrials are Brandt, Brandt, Spahl, and Weber [2], Harris, Hoffman, and Saunders [7] and Sze [11].

Call centers are constantly subject to time-varying conditions, and waiting customers in phone queues are unable to observe the state of the system. It follows that time-dependent modeling (as opposed to also state-dependent) is natural for call centers. In Figure 1, customers represent callers that seek service at the call center; servers are telephone agents (operators, CSR's = Customer Service Representatives). The rate λ_t represents the time-varying call arrival rate. (Our validation experiments, in Section 3, focus on two forms of behavior: continuously-changing load and sudden-peak.) Abandonment rates could depend on time as a response to broadcasted IVR-information (IVR = Interactive Voice Response). For example, announcements on predictable long queues by providers of 1-800 services, could be designed to encourage abandonments in order to reduce waiting costs; predictable long queues could arise, for example, as a consequence of a promotion whose timing is known in advance. Finally, the number of agents varies in response to time-variations of offered load (workforce management.)

The size of call centers varies from small (1-2 agents) to the very large (thousands of agents). The latter require delicate performance analysis, and here we demonstrate that asymptotic analysis suffices to provide significant insight, of both theoretical and practical importance. Specifically, our asymptotic analysis is in a regime where we scale up the number of servers in response to a similar scaling up of the arrival rate by customers. The outcome is first a fluid approximation (Theorem 1), followed by diffusion refinement (Theorem 2). The usefulness of our approximations is already apparent from a visual comparison with corresponding simulations (Section 3); see, for example, Figures 2–5.

2. The Model and Limit Theorems

The multiserver queue with abandonment and retrials, as illustrated in Figure 1, is characterized by the following set of parameters:

$$
\begin{aligned}
\lambda_t &= \text{external arrival rate to the calling node at time } t \ (\lambda_t \geq 0),\\
\beta_t &= \text{abandonment rate from the calling node at time } t \ (\beta_t \geq 0),\\
\mu_t^1 &= \text{service rate for the calling node at time } t \ (\mu_t^1 \geq 0),\\
\mu_t^2 &= \text{service rate for the retry node at time } t \ (\mu_t^2 \geq 0),\\
\psi_t &= \text{probability of no retrial at time } t \ (0 \leq \psi_t \leq 1),\\
n_t &= \text{number of calling servers at time } t \ (n_t = 0, 1, 2, \ldots).
\end{aligned}
$$

The above parameters suggest that $\mathbf{Q}^{(0)}(t) = \left(Q_1^{(0)}(t), Q_2^{(0)}(t)\right)$, the *fluid approximation* for $\mathbf{Q}(t) = \left(Q_1(t), Q_2(t)\right)$, solves the non-linear differential equations

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

and

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

Here $x \wedge y = \min(x,y)$ and $x^+ = \max(x,0)$ for all real x and y. Given an initial condition $\mathbf{Q}^{(0)}(0) = \Big(Q_1^{(0)}(0), Q_2^{(0)}(0)\Big)$, then $\mathbf{Q}^{(0)}(t)$ is uniquely determined [9]. The correspondence between Figure 1 and the fluid model (1) and (2) is clear: for example, the rate of change in $Q_1^{(0)}$, namely (1), consists of the exogenous input rate λ_t plus the outflow rate from the retrial pool $\mu_t^2 Q_2^{(0)}(t)$, from which one subtracts the departure rate from the network $\mu_t^1\Big(Q_1^{(0)}(t) \wedge n_t\Big)$, where $Q_1^{(0)}(t) \wedge n_t$ approximates the number of active servers at time t, as well as the abandonment rate $\beta_t\Big(Q_1^{(0)}(t) - n_t\Big)^+$, where $\Big(Q_1^{(0)}(t) - n_t\Big)^+$ approximates the the waiting customers that are vulnerable to abandon.

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

$$\begin{aligned} Q_1(t) &= Q_1(0) + A_{21}^c\left(\int_0^t Q_2(s)\mu_s^2 ds\right) - A_{12}^b\left(\int_0^t \Big(Q_1(s) - n_s\Big)^+ \beta_s(1-\psi_s)ds\right) \\ &\quad + A^a\left(\int_0^t \lambda_s ds\right) - A^b\left(\int_0^t \Big(Q_1(s) - n_s\Big)^+ \beta_s\psi_s ds\right) - A^c\left(\int_0^t \Big(Q_1(s) \wedge n_s\Big)\mu_s^1 ds\right) \end{aligned} \tag{3}$$

and

$$Q_2(t) = Q_2(0) + A_{12}^b\left(\int_0^t \Big(Q_1(s) - n_s\Big)^+ \beta_s(1-\psi_s)ds\right) - A_{21}^c\left(\int_0^t \Big(Q_2(s)\Big)\mu_s^2 ds\right), \tag{4}$$

where A^a, A^b, A^c, A_{12}^b, and A_{21}^c are five given mutually independent, standard (mean rate 1), Poisson processes [9]. As described in the introduction, we are interested in scaling up both the arrival rate and the number of servers. To this end we introduce a *scaling parameter* η, $\eta \uparrow \infty$, and construct a *scaled* version $\mathbf{Q}^\eta(t) = \Big(Q_1^\eta(t), Q_2^\eta(t)\Big)$ of the process $\mathbf{Q}$, where

$$\begin{aligned} Q_1^\eta(t) &= Q_1^\eta(0) + A_{21}^c\left(\int_0^t Q_2^\eta(s)\mu_s^2 ds\right) - A_{12}^b\left(\int_0^t \Big(Q_1^\eta(s) - \eta n_s\Big)^+ \beta_s(1-\psi_s)ds\right) \\ &\quad + A^a\left(\int_0^t \eta\lambda_s ds\right) - A^b\left(\int_0^t \Big(Q_1^\eta(s) - \eta n_s\Big)^+ \beta_s\psi_s ds\right) - A^c\left(\int_0^t \Big(Q_1^\eta(s) \wedge \eta n_s\Big)\mu_s^1 ds\right) \end{aligned} \tag{5}$$

and

$$Q_2^\eta(t) = Q_2^\eta(0) + A_{12}^b\left(\int_0^t \Big(Q_1^\eta(s) - \eta n_s\Big)^+ \beta_s(1-\psi_s)ds\right) - A_{21}^c\left(\int_0^t \Big(Q_2^\eta(s)\Big)\mu_s^2 ds\right). \tag{6}$$

Theorem 2.1 (Strong Law of Large Numbers) *Using the scaling of (5) and (6), we have*

$$\lim_{\eta\to\infty} \frac{1}{\eta}\mathbf{Q}^\eta(t) = \mathbf{Q}^{(0)}(t), \quad \text{for all } t \geq 0, \text{ a.s.}, \tag{7}$$

where this is a convergence of sample-paths, which converge uniformly on compact sets of $t \geq 0$, almost surely.

We can refine this deterministic fluid approximation by deriving a stochastic *diffusion approximation* $\mathbf{Q}^{(1)}$, to the queueing model as follows:

$$\mathbf{Q}^{\eta}(t) \stackrel{d}{=} \eta\mathbf{Q}^{(0)}(t) + \sqrt{\eta}\mathbf{Q}^{(1)}(t) + o(\sqrt{\eta}), \tag{8}$$

where $\mathbf{Q}^{(1)}$ is formally defined through the following theorem.

Theorem 2.2 (Central Limit Theorem) *Using the scaling of (5) and (6) and the fluid approximation $\mathbf{Q}^{(0)}(t)$, we obtain the diffusion approximation $\mathbf{Q}^{(1)}(t) = \left(Q_1^{(1)}(t), Q_2^{(1)}(t)\right)$ by*

$$\lim_{\eta\to\infty} \sqrt{\eta}\left(\frac{\mathbf{Q}^{\eta}(t)}{\eta} - \mathbf{Q}^{(0)}(t)\right) \stackrel{d}{=} \mathbf{Q}^{(1)}(t), \quad \text{for all } t \geq 0, \tag{9}$$

where this is convergence in distribution of the corresponding stochastic processes in an appropriate functional space [9]. Here $\mathbf{Q}^{(1)}(t)$ is the solution to the integral equations

$$\begin{aligned} Q_1^{(1)}(t) &= Q_1^{(1)}(0) - B_{12}^b\left(\int_0^t \left(Q_1^{(0)}(s) - n_s\right)^+ \beta_s(1-\psi_s)ds\right) + B_{21}^c\left(\int_0^t \left(Q_2^{(0)}(s)\right)\mu_s^2 ds\right) \\ &\quad B^a\left(\int_0^t \lambda_s ds\right) + \int_0^t \Big[\left(\mu_s^1 1_{\{Q_1^{(0)}(s)\leq n_s\}} + \beta_s 1_{\{Q_1^{(0)}(s)>n_s\}}\right)Q_1^{(1)}(s)^- \\ &\quad -\left(\mu_s^1 1_{\{Q_1^{(0)}(s)<n_s\}} + \beta_s 1_{\{Q_1^{(0)}(s)\geq n_s\}}\right)Q_1^{(1)}(s)^+ + \mu_s^2 Q_2^{(1)}(s)\Big]\, ds \\ &\quad -B^b\left(\int_0^t \left(Q_1^{(0)}(s) - n_s\right)^+ \beta_s\psi_s ds\right) - B^c\left(\int_0^t \left(Q_1^{(0)}(s)\wedge n_s\right)\mu_s^1 ds\right), \end{aligned}$$

and

$$\begin{aligned} Q_2^{(1)}(t) &= Q_2^{(1)}(0) \\ &\quad + \int_0^t \Big[\left(Q_1^{(1)}(s)^+ 1_{\{Q_1^{(0)}(s)\geq n_s\}} - Q_1^{(1)}(s)^- 1_{\{Q_1^{(0)}(s)>n_s\}}\right)\beta_s(1-\psi_s) - \mu_s^2 Q_2^{(1)}(s)\Big]\, ds \\ &\quad + B_{12}^b\left(\int_0^t \left(Q_1^{(0)}(s) - n_s\right)^+ \beta_s(1-\psi_s)ds\right) + B_{21}^c\left(\int_0^t \left(Q_2^{(0)}(s)\right)\mu_s^2 ds\right), \end{aligned}$$

where B^a, B^b, B^c, B_{12}^b, and B_{21}^c are mutually independent, standard (the mean is zero and the variance at time t is t) Brownian motions.

The diffusion process $\mathbf{Q}^{(1)}$ provides us with confidence bounds for the fluid approximation. These are obtained by computing from a simple set of non-linear differential equations for the mean and variance of this diffusion, given by the following proposition [9].

Proposition 2.3 *Assume that the set of time points $\left\{\, t \geq 0 \mid \; Q_1^{(0)}(t) = n_t \right\}$ has measure zero. The mean vector for the diffusion approximation then solves the set of differential equations*

$$\frac{d}{dt}\mathsf{E}\left[Q_1^{(1)}(t)\right] = -\left(\mu_t^1 1_{\{Q_1^{(0)}(t)\leq n_t\}} + \beta_t 1_{\{Q_1^{(0)}(t)>n_t\}}\right)\mathsf{E}\left[Q_1^{(1)}(t)\right] + \mu_t^2\mathsf{E}\left[Q_2^{(1)}(t)\right] \tag{10}$$

and

$$\frac{d}{dt}\mathsf{E}\left[Q_2^{(1)}(t)\right] = \beta_t(1-\psi_t)\mathsf{E}\left[Q_1^{(1)}(t)\right]1_{\{Q_1^{(0)}(t)\geq n_t\}} - \mu_t^2\mathsf{E}\left[Q_2^{(1)}(t)\right]. \tag{11}$$

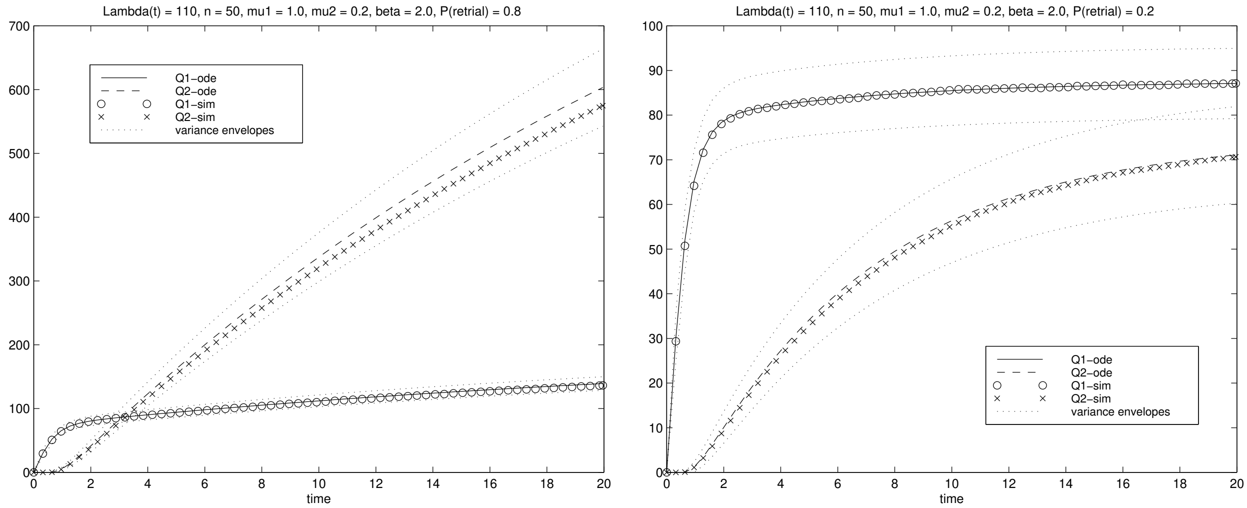

Figure 2. Numerical examples: Constant arrival rate cases.

Moreover, the covariance matrix for the diffusion approximation solves the differential equations

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

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

and

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

Time-varying queues alternate among phases of underloading, critical-loading, and overloading [8]. The set $\left\{ t \mid Q_1^{(0)}(t) = n_t \right\}$ 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 Proposition 2.3 applies to all the examples in the following section.

3. Numerical Examples

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

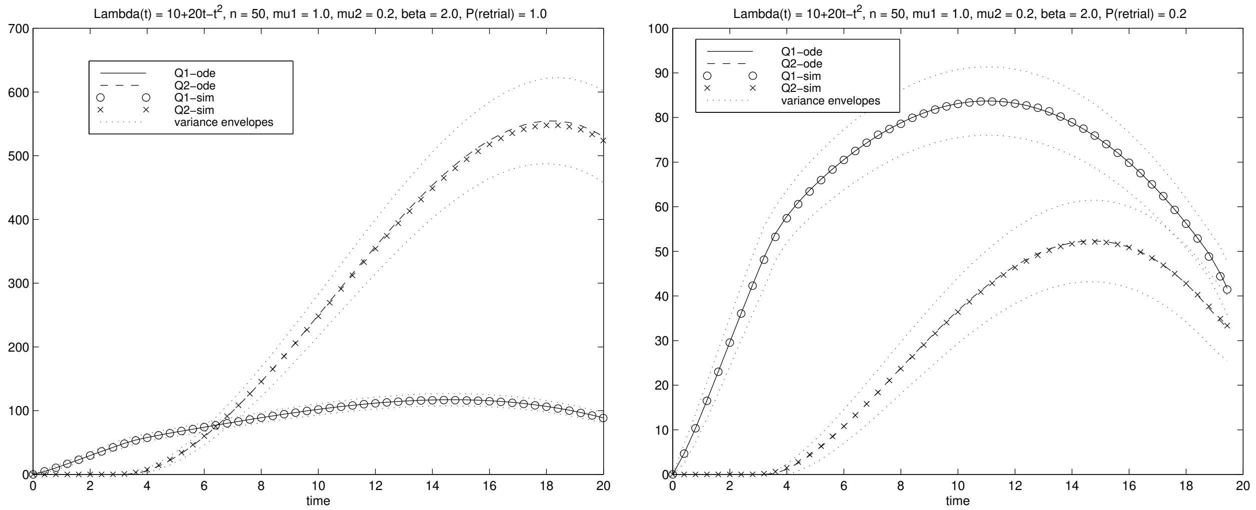

Figure 3. Numerical examples: $\psi_t = 0.0$ and 0.8.

The first two examples, see Figure 2, that we consider actually have the arrival rate λ equal to a constant 110, with $n = 50$, $\beta = 2.0$, and $\psi = 0.2$ and 0.8. This is an overloaded system, see [8], i.e. $Q_1^{(0)}(t) > n$ for large enough t, and equations (1) and (2) indicate that $Q_1^{(0)}(t) \rightarrow q_1$ and $Q_1^{(0)}(t) \rightarrow q_2$ as $t \rightarrow \infty$. Setting $\frac{d}{dt}Q_1^{(0)}(t) = \frac{d}{dt}Q_1^{(0)}(t) = 0$ as $t \rightarrow \infty$, then q_1 and q_2 solve the linear equations

$$\lambda + \mu^2 q_2 - \mu^1 n - \beta(q_1 - n) = 0 \tag{12}$$

and

$$\beta(1-\psi)(q_1 - n) - \mu^2 q_2 = 0. \tag{13}$$

These equations can be easily solved to yield

$$q_1 = n + \frac{\lambda - \mu^1 n}{\beta\psi} \quad \text{and} \quad q_2 = \frac{\beta(1-\psi)}{\mu^2}\frac{\lambda - \mu^1 n}{\beta\psi}. \tag{14}$$

Substituting in $\psi = 0.2$ and the other parameters indicated above yields $q_1 = 200$, $q_2 = 1200$. This case corresponds to the graph of the left in Figure 2 and indicates that this system is still far from equilibrium at time 20. With $\psi = 0.8$ (so the probability of retrials is equal to 0.2) we obtain $q_1 = 87.5$ and $q_2 = 75$. This case corresponds to the graph on the right in Figure 2. Here it appears that $Q_1^{(0)}$ has essentially reached equilibrium by the time $t = 20$, while $Q_2^{(0)}$ has a bit more to go.

In general, the accuracy for the computation of the fluid approximation can be checked by a simple test that only requires a visual inspection of the graphs. The test is the fact that the time for a local maximum (or local minimum) of a continuously differentiable function must be one where the derivative of the function is zero. Consider the right hand plot in Figure 4, where $\beta_t = 1$, $\psi_t = 0.5$, $n_t = 50$, and $\lambda_t = 10 + 20t - t^2$. The graph of

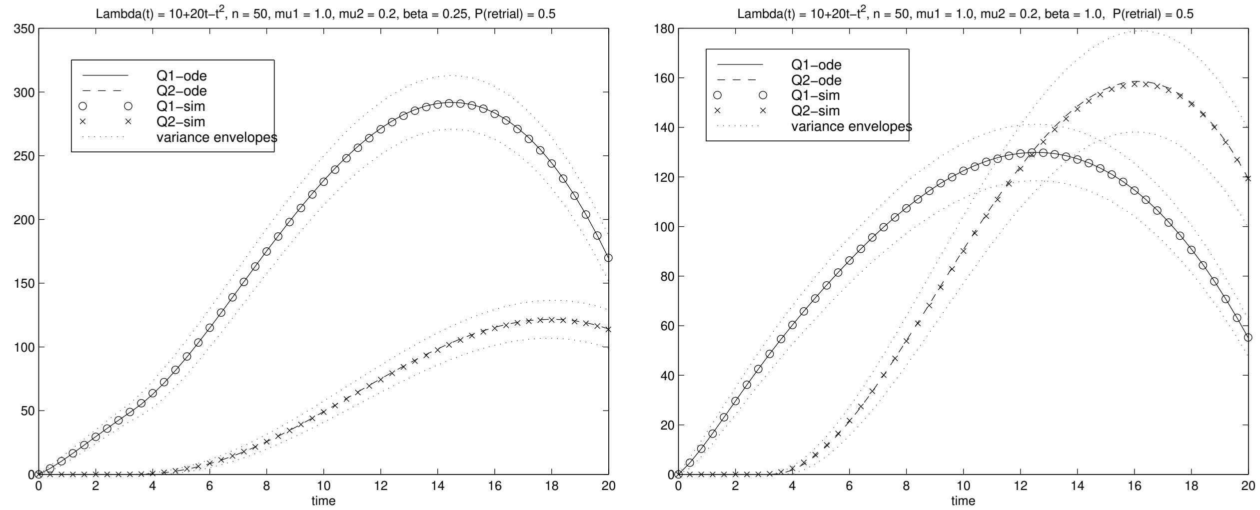

Figure 4. Numerical examples: $\beta_t = 0.25$ and 1.0.

$Q_1^{(0)}$ appears to peak roughly at the value 130 at time $t \approx 12$. Since the derivative at a local maximum is zero, then equation (1) becomes

$$\lambda_t + \mu_t^2 Q_2^{(0)}(t) \approx \mu_t^1 \Big(Q_1^{(0)}(t) \wedge n_t \Big) + \beta_t \Big(Q_1^{(0)}(t) - n_t \Big)^+ \tag{15}$$

when $t \approx 12$, as well as $Q_1^{(0)}(t) \approx Q_2^{(0)}(t) \approx 130$. The left hand side of (15) equals $106 + .2 \cdot 130 = 132$ which is roughly the value of the right hand side of (15), which is $50 + 80 = 130$.

Similarly, the graph of $Q_2^{(0)}$ appears to peak roughly at the value 155 at time $t \approx 16.5$ which also implies $Q_1^{(0)}(t) \approx 110$ and equation (2) becomes

$$\beta_t(1 - \psi_t)\Big(Q_1^{(0)}(t) - n_t \Big)^+ \approx \mu_t^2 Q_2^{(0)}(t). \tag{16}$$

The left hand side of (16) is $0.5 \cdot 60 = 30$ and the right hand side of (16) is about the same or $0.2 \cdot 155 = 31$.

The reader should be convinced of the effectiveness of the fluid approximation after an examination of Figures 2 through 5. Here we compare the numerical solution (via forward Euler) of the system of ordinary differential equations for $\mathbf{Q}^{(0)}(t)$ given in (1) and (2) to a simulation of the real system. These quantities are denoted in the legends as $Q1$-ode, $Q2$-ode, $Q1$-sim, and $Q2$-sim. Throughout, the term "variance envelopes" refers to

$$Q_i^{(0)}(t) \pm \sqrt{\mathsf{Var}\left[Q_i^{(1)}(t) \right]} \tag{17}$$

for $i = 1, 2$, where $\mathsf{Var}\left[Q_1^{(1)}(t) \right]$ and $\mathsf{Var}\left[Q_2^{(1)}(t) \right]$ are the numerical solutions, again by forward Euler, of the differential equations determining the covariance matrix of the diffusion approximation $\mathbf{Q}^{(1)}$ (see Proposition 2.3). Setting $Q_1^{(1)}(0) = Q_1^{(1)}(0) = 0$ yields by

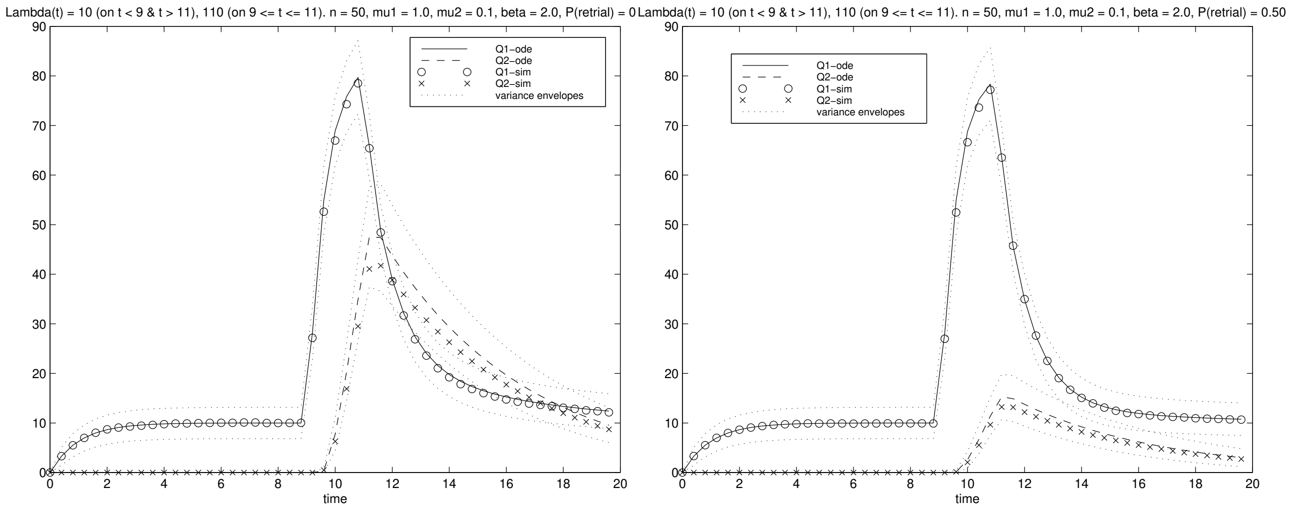

Figure 5. Numerical examples: Spike at time interval $(9, 11)$ and $\psi_t = 0.25, 0.75$.

equations (10) and (11) that $\mathsf{E}[Q_1^{(1)}(t)] = \mathsf{E}[Q_1^{(1)}(t)] = 0$ for all $t \geq 0$. Otherwise two additional ordinary differential equations would be needed to compute the variance envelopes. As all the numerical examples demonstrate, the fluid approximation tracks the simulation well.

It should come as no surprise that the fluid approximation is a far more efficient means of computation than simulation. For the experiments pictured, the simulation was carried out by the well known method of uniformization for the external arrival rate. That is, let λ_{max} be the supremum of λ_t over the time interval $[0, T]$ of interest. If after an event occurs at time t the system is in state $Q_1(t) = j$ and $Q_2(t) = k$, then a set of exponentially distributed random variablies X_1,X_2, X_3, and X_4 with means given by the inverses of $\lambda_{max}, \mu_1 j, \mu_2 k, \beta(k-n)^+$ are generated. These correspond to a potential external arrival, call completion, retrial completion, or abandonment respectively. Uniformization dictates that an external arrival is considered possible, given that X_1 has value s, with probability $\lambda(t+s)/\lambda_{max}$. Using the Markovian property of the model, the smallest value of these random variables is taken as the time until the next event. In this way we step through the time interval $[0, T]$ to complete a single simulation. We used 5000 replications in all of our experiments. In each simulation we record the state of the system in time by splitting the interval into T/δ intervals of length δ. These bins are then averaged over the total number of simulations. For most of our examples a δ of order 10^{-2} proved sufficient for stabilization. That is, decreasing δ further would produce no noticeable effect in the simulation results. However for certain parameter regimes, namely high retrial probabilities (the left graph in Figure 3) it was neccessary to use δ's on order 10^{-4}.

The need for such small δ's greatly increases computation time. The simulations in Figure 3 required as long as 20 minutes of elapsed time on a MIPS 4400/4000 processor running at 150 MHz, while a naive forward Euler integration of the fluid system is essentially instantaneous.

4. Conclusions

We have provided a simple intuitive approximation for a time varying queueing system of practical interest. Our numerical results show that the fluid approximation is also quite accurate.

The present paper is a necessary and significant first step in a natural progression of further experiments and research. The goal is the development of practical approximations for time-dependent queueing systems. Specifically, we have in mind the analysis of waiting times (in addition to the present queue-length); staffing strategies (optimizing n_t in response to a varying λ_t); analysis of complex periodic systems; diffusion refinements during critical loading; parameter estimation, for example abandonment rates (from ACD data) and retrial rates (from ANI data; ANI = Automatic Number Identification) all of which culminates in a visualization-animation-based tool that supports performance analysis, both theoretically and practically.

REFERENCES

1. Brigandi, A. J., Dargon, D. R., Sheehan, M. J., and Spencer, T. AT&T's Call Processing Simulator (CAPS) Operational Design for Inbound Call Centers, Interfaces 24:1, 1994, 6–28.
2. Brandt, A., Brandt, M., Spahl, G., Weber, D., Modeling and optimization of call distribution systems, ITC-15, 133-144.
3. Boxma and de Waal, Multiserver queues with impatient customers, ITC-14, 743-756.
4. Falin, G. I. and Templeton, J. G. C., *Retrial Queues*, Chapman and Hall, 1997.
5. Garnet, O., Mandelbaum, M. and Reiman, M., Designing a telephone call center with impatient customers, in preparation.
6. Grier, N., Massey, W. A., McKoy, T., and Whitt, W., The time-dependent Erlang loss model with retrials, Telecommunication Systems 7, 1997, 253-265.
7. Harris, C. M., Hoffman, C. L., Saunders, P. B., Modeling the IRS telephone taxpayer information system, Operations Research 35, 504-523, 1987.
8. Mandelbaum, A. and Massey, W. A., Strong approximations for time dependent queues, Mathematics of Operations Research, 20:1, 1995, pp. 33–64.
9. Mandelbaum, A., Massey, W. A., and Reiman, M. I., Strong approximations for Markovian service networks, Queueing Systems 30, 1998, 149–201.
10. Mandelbaum, M. and Shimkin, N., A model for rational abandonments from invisible queues, submitted.
11. Sze, D.Y., A queueing model for telephone operator staffing, Operations Research, 32, 229-249, 1984.
12. Wolff, R. W. *Stochastic modeling and the theory of queues.* Prentice-Hall, Inc., 1989.