DECOMPOSITION APPROXIMATIONS FOR TIME-DEPENDENT MARKOVIAN QUEUEING NETWORKS

by

Ward Whitt[1]

AT&T Labs

November 11, 1997

Revision: November 19, 1998

Operations Research Letters 24 (1999) 97–103

[1]AT&T Labs, Room A117, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971; email: wow@research.att.com

Abstract

Motivated by the development of complex telephone call center networks, we present a general framework for decompositions to approximately solve Markovian queueing networks with time-dependent and state-dependent transition rates. The decompositions are based on assuming either full or partial product form for the time-dependent probability vectors at each time. These decompositions reduce the number of time-dependent ordinary differential equations that must be solved. We show how special structure in the transition rates can be exploited to speed up computation. There is extra theoretical support for the decomposition approximation when the steady-state distribution of the time-homogeneous version of the model has product form.

Keywords: time-dependent queues, time-dependent queueing networks, time-dependent Markov chains, decomposition approximations, systems of ordinary differential equations, product-form queueing networks, product-form approximations, telephone call centers, air traffic management

1. Introduction

This paper was motivated by the desire to analyze the time-dependent behavior of complex telephone call centers containing a network of interactive voice response systems and agent groups, where there may be *balking* (failure to join upon arrival if service cannot be provided immediately or if the queue is too long), *blocking* (rejection of new arrivals when there is no available capacity), *reneging* (customer abandonment after waiting) and *retrying* (customer reattempts later after experiencing balking, blocking or reneging); e.g., see Andrews and Parsons [1], Brigandi, Dargon, Sheehan and Spencer [2], Jennings, Mandelbaum, Massey and Whitt [5], Mehrota [12] and Rappaport [15]. We propose modeling such a system as a Markovian queueing network or a multidimensional birth-and-death process, with transition rates that are both state-dependent and time-dependent. Time-dependence and state-dependence are especially important to include in the model when dynamic routing strategies are being considered. For example, when one site is heavily congested, a portion of its new arrivals may be instantaneously routed to another site. Careful analysis is needed because it is hard to anticipate the consequences of such actions.

The Markovian structure makes it possible to obtain a time-dependent description of performance as the solution of a system of ordinary differential equations (ODEs), but the network structure causes there to be a very large number of equations, tending to make the analysis intractable. Hence there is a need for approximations. With suitable approximations, the analytical approach can provide an alternative to simulation that is convenient for rapidly examining many candidate scenarios. A modest number of ODEs can be solved quite efficiently using Runge-Kutta methods; e.g., see Green, Kolesar and Svoronos [3] and Taaffe and Ong [18]. Increasing computer power makes it now possible to solve substantially larger systems than previously.

In this paper we present a framework for decomposition approximations. Our proposed procedure is a generalization of the decomposition approach used by Grier, Massey, McKoy and Whitt [4] to analyze the time-dependent Erlang loss model with retrials. That model is a special case of a two-node queueing network. The numerical results there illustrate how such a decomposition procedure can perform. We leave to later work evaluating the effectiveness of other specific implementations. It is hoped that the discussion here will encourage more numerical studies.

The decomposition approach has also been considered previously by others, especially to analyze the time-dependent congestion in airports; see Koopman [7], Malone [8], Peterson, Bertsimas and Odoni [14] and Schmeiser and Taaffe [17]. For airports, each queue is naturally modelled as an $M_t/M_t/s$ queue for small s, e.g., $1 \leq s \leq 4$, corresponding to the number of runways. The idea here is essentially the same, but the framework is more general. As shown in this context, the decomposition approximation can be further accelerated by applying closure approximations for single queues; see Rothkopf and Oren [16] and Taaffe and Ong [18]. Then only one or two ODEs are needed for each queue. The general idea of a closure approximation is to assume a parametric form for the marginal distributions and then have a small number of ODEs to describe the time-dependent evolution of the characterizing parameters; e.g., these might be for the first two moments.

The decomposition approach has also been a key way to develop approximations for the steady-state distribution of time-homogeneous queueing network models; e.g., see Whitt (1995) and references therein. There also are other candidate approximations for time-dependent behavior besides systems of ODEs. First, there are uniform acceleration asymptotic expansions, possibly combined with heavy-traffic and asymptotics; see Massey and Whitt [11] and Mandelbaum, Massey and Reiman [9]. For networks of infinite-server queues, exact solutions are possible; see Massey and Whitt [10]. In that context, there are exact-decomposition and exact-closure properties; i.e., (1) after determining the net arrival rates, the queues can be analyzed separately, (2) the marginal distributions are Poisson, so that it suffices to have a single ODE characterizing the mean for each queue; see Massey and Whitt [10] and Grier et al. [4]. Hence, if there tend to be ample servers, a network of infinite-server queues can be a convenient approximation. Here we are focusing on the case in which infinite-server models are not appropriate.

Here is how the rest of this paper is organized. In Section 2 we specify the model and display the exact solution as a system of ODEs. In Section 3 we briefly indicate how such models can be used for telephone call centers. In Section 4 we consider decomposition approximations based on assuming that the time-dependent probability vectors have product form. We consider partial product form as well as full product form. In Section 5 we show the computational advantage gained from assuming special structure in the transition rates.

In Section 6 we discuss the theoretical support provided by the steady-state distribution associated with the time-homogeneous version of a model having product form. Finally, in Section 7 we discuss non-exponential service-time distributions and non-Poisson arrival processes. The extension to phase-type service-time distributions and interarrival-time distributions is treated naturally within our framework using partial product form. This extension follows Ong and Taaffe [13] and Taaffe and Ong [18].

2. The Basic Model

Our model is a time-dependent and state-dependent generalization of a Jackson queueing network, also known as a migration process or a multidimensional birth-and-death process; e.g., see Chapters 2 and 3 of Kelly [6]. Let there be m queues. Let the system state at any time be the vector $\mathbf{n} \equiv (n_1, \ldots, n_m)$, where n_j is the number of customers at queue j. Let $\mathbf{e}_i$ be the m-dimensional vector with a 1 in the i^{th} place and 0's elsewhere.

Let there be external arrivals to queue i at rate $\lambda_i(t, \mathbf{n})$ when the system state is $\mathbf{n}$ at time t. Let there be departures from queue i at rate $\mu_i(t, \mathbf{n})$ when the system state is $\mathbf{n}$ at time t. Let a departure from queue i at time t be routed to queue j with probability $p_{ij}(t, \mathbf{n})$ when the system state (before the departure) is $\mathbf{n}$ at time t. These rates are understood to be 0 if any component n_j of $\mathbf{n}$ is negative. Moreover, $\mu_j(t, \mathbf{n}) = 0$ if $n_j = 0$.

Let $Q_j(t)$ be the number of customers at queue j at time t. We assume that the vector-valued process $\{(Q_1(t), \ldots, Q_m(t)) : t \geq 0\}$ is a nonhomogeneous continuous-time Markov chain (CTMC) with the transition intensities specified above. Let

$$q(t, \mathbf{n}) = P(Q_1(t) = n_1, \ldots, Q_m(t) = n_m) \tag{2.1}$$

for $\mathbf{n} \equiv (n_1, \ldots, n_m)$. Then the Markov assumption implies that the marginal (one-dimensional, i.e., for single t) probabilities in (2.1) satisfy a system of ordinary differential equations (ODEs), namely,

$$\begin{aligned} \dot{q}(t, \mathbf{n}) \equiv \frac{d}{dt} q(t, \mathbf{n}) &= \sum_{j=1}^{m} q(t, \mathbf{n} - \mathbf{e}_j) \lambda_j(t, \mathbf{n} - \mathbf{e}_j) \\ &+ \sum_{i=1}^{m} \sum_{j=1}^{m} q(t, \mathbf{n} + \mathbf{e}_j - \mathbf{e}_i) \mu_j(t, \mathbf{n} + \mathbf{e}_j - \mathbf{e}_i) p_{ji}(t, \mathbf{n} + \mathbf{e}_j - \mathbf{e}_i) , \end{aligned} \tag{2.2}$$

which can be solved given initial values $q(0, \mathbf{n})$ for all possible state vectors $\mathbf{n}$.

From a computational perspective, a serious difficulty with (2.1) is that there can be a very large number of equations, one for each possible state $\mathbf{n}$. If n_j can assume any value between 0 and $N - 1$, with all combinations possible, then the number of equations is N^m. Hence there is motivation for introducing approximations that reduce the computational complexity.

3. Telephone Call Center Models

We briefly describe how a telephone call center can be modelled by the Markovian queueing network introduced in Section 2. Our goal here is to demonstrate general applicability; we leave to future work careful analysis of specific models.

Customers (callers) may enter the system at several different nodes (queues). More than one node might be used to represent different treatment depending on the calling or called number. There are nominal time-dependent customer arrival rates at each node, say $\alpha_i(t)$ at node i.

The nominal arrival rates are then reduced in a state-dependent fashion (still with time dependence) to account for balking and blocking upon arrival. The balking and blocking also generate corresponding arrivals to separate (typically infinite-server) retrial nodes. Having different balking and blocking retrial nodes lets us represent different retrial behavior for these two different experiences. The customers who previously balked or were blocked spend a random time in the retrial node and then transition to become new arrivals. Retrials from blocking were considered by Grier et al. [4].

To be more concrete, suppose that there are $s_i(t)$ servers (agents) and $n_i(t)$ extra waiting spaces (e.g., due to a limited number of available trunk lines) at node i at time t. Then blocking occurs when the number of customers at node i is $s_i(t) + n_i(t)$. Moreover, balking can occur when the number of customers is k, $s_i(t) \leq k < s_i(t) + n_i(t)$. Balking might then occur with probability $\beta_i(t, k)$. With these features, the arrival rate at node i becomes state-dependent as well as time-dependent. The arrivals that block or balk in turn may become arrivals at the retrial nodes.

Customers entering the network at node i to receive service may receive service and then be routed to another node j. For example, the first node might be a computerized integrated voice response (IVR) system, after which the call is routed to one of several sites each with many agents, or may leave the system. Consistent with actual systems, the routing within the network may be both state-dependent and time-dependent.

Customers may also renege after starting service. Hence the nominal service rates should be increased to reflect this possibility. Reneging customers are also routed to their own retrial queues. With k customers and $s_i(t)$ servers at node i at time t, the nominal departure (service completion) rate can be $\mu_i(t)[k \wedge s_i(t)]$, where $k \wedge s_i(t) \equiv \min\{k, s_i(t)\}$. With reneging, the total departure rate can be increased to $s_i(t)\mu_i(t) + (k - s_i(t))\eta_i(t)$ when $k > s_i(t)$, which represents a per-customer reneging rate of $\eta_i(t)$ while waiting. If reneging customers from node i retry with probability $\xi_i(t)$, then there is an arrival rate of $(k - s_i(t))\eta_i(t)\xi_i(t)$ into the reneging retrial queue when there are k $(> s_i(t))$ customers at node i.

Thus, the telephone call center can be modeled as a Markovian queueing network with external arrival rates $\lambda_i(t, \mathbf{n})$, service rates $\mu_i(t, \mathbf{n})$ and transition probabilities $p_{ij}(t, \mathbf{n})$. Moreover, these transitions rates often have special structure that can be exploited. (See Section 5 below.)

4. Product-Form Decomposition Approximations

We can reduce the computational burden of (2.2) by assuming, as an approximation, that the time-dependent probabilities $q(t, \mathbf{n})$ have special structure. There are different forms of special structure that can be considered. Here we assume that the time-dependent probabilities have product form. However, they need not have full product form. Instead, they could have only partial product form.

Full product form is obtained by assuming that

$$q(t, \mathbf{n}) = \prod_{i=1}^{m} q_i(t, n_i) \quad \text{for all} \quad t \geq 0 \quad \text{and} \quad \mathbf{n} = (n_1, \ldots, n_m) \ , \tag{4.1}$$

where $q_i(t, n_i)$ is a function of time t and the single component n_i. *Partial product form* is obtained by assuming that a generalization of (4.1) holds for designated subsets of queues. It can be specified by considering the case of partitioning the queues into two disjoint subsets, say queues $1, \ldots, m_1$ and $m_1 + 1, \ldots, m$. Let $\mathbf{n}_1 = (n_1, \ldots, n_{m_1})$ and $\mathbf{n}_2 = (n_{m_1+1}, \ldots, n_m)$. Then partial product form is obtained by assuming that

$$q(t, \mathbf{n}) = q_1(t, \mathbf{n}_1)q_2(t, \mathbf{n}_2) \quad \text{for all} \quad t \geq 0 \ , \tag{4.2}$$

and all $\mathbf{n} = (\mathbf{n}_1, \mathbf{n}_2)$.

The key simplification provided by (4.1) and (4.2) is that we have coupled smaller systems of ODEs instead of one big system of ODEs. For example, with (4.1), we would calculate $q_i(t + \Delta t, n_i)$ for *each* i (separately) assuming that $q_i(s, n_i)$ is known for $s \leq t$ for *all* i. Thus, we simultaneously solve m systems of ODEs, each of which has say N equations, instead of solving one system of N^m ODEs.

Suppose that we consider (4.2). Let $\mathbf{e}_{1j}$ $(\mathbf{e}_{2j})$ be a m_1-dimensional $((m - m_1)$-dimensional) vector with a 1 in the j^{th} and 0's elsewhere. Then, instead of (2.2), we obtain

$$\dot{q}_1(t, \mathbf{n}_1) = \sum_{\mathbf{n}_2} \sum_{i=1}^{m_1} q_1(t, \mathbf{n}_1 - \mathbf{e}_{1i})q_2(t, \mathbf{n}_2)\lambda_i(t, \mathbf{n} - \mathbf{e}_i)$$

$$+ \sum_{\mathbf{n}_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_1} q_1(t, \mathbf{n} + \mathbf{e}_j - \mathbf{e}_i) q_2(t, \mathbf{n}_2) \mu_j(t, \mathbf{n} + \mathbf{e}_j - \mathbf{e}_i) p_{ji}(t, \mathbf{n}_1 + \mathbf{e}_j - \mathbf{e}_i)$$

$$+ \sum_{\mathbf{n}_2} \sum_{i=1}^{m_1} \sum_{j=m_1+1}^{m} q_1(t, \mathbf{n}_1 - \mathbf{e}_{1i}) q_2(t, \mathbf{n}_2 + \mathbf{e}_{2j}) \mu_j(t, \mathbf{n} + \mathbf{e}_j - \mathbf{e}_i) p_{ji}(t, \mathbf{n} + \mathbf{e}_j - \mathbf{e}_i) , \tag{4.3}$$

where the sum on $\mathbf{n}_2$ extends over all possible vectors $\mathbf{n}_2$. A similar system of ODEs holds for $q_2(t, \mathbf{n}_2)$; indeed it can be regarded as (4.3) if we switch the labels.

This flexibility beyond full product form can be very useful. For example, our interest may be focused on two queues, say queues 1 and 2, that are quite tightly coupled. Then we might want to *not* assume product form for these two queues, but otherwise assume product form. We could apply (4.2) with $m_1 = 2$, also exploiting the fact that $q_2(t, n_2)$ can be expressed in the full product form (4.1). The number of equations for calculating $q_1(t, \mathbf{n}_1)$ is the number of vectors (n_1, n_2). For the other components, we are calculating $q_j(t, n_j)$ for $j \geq 3$, where n_j ranges over the possible states of queue j only.

As indicated in the introduction, a special case of the product-form decomposition approximation for a network with two queues was considered by Grier et al. [4]. They considered the time-dependent Erlang loss model with retrials. The customers "in orbit" waiting to retry at a later time constitute the second queue. Assuming that the number of servers in the loss model is L and the maximum possible number of customers in retry mode is R (as an approximation for infinity), the original number of ODE's is $(L+1)(R+1)$. (We might have $R = L$ and $L = 100$ or $L = 1000$.) In contrast, the product-form decomposition approximation has only $L+2$ equations, because only a single equation for the mean is needed for the infinite-server retry mode in isolation. The paper by Grier et al. [4] reports numerical results illustrating how the approximation performs. The approximation quite accurately describes the time-dependent mean number of busy servers and the time when the time-dependent blocking probability is highest, but it only roughly describes the time-dependent blocking probability and the time-dependent mean number of customers in retry mode. Even though the time-dependent blocking probability approximation is not very accurate, it can be sufficiently accurate for many engineering purposes, such as determining time-dependent staffing levels, as in Jennings et al. [5].

5. Special Structure in the Transition Rates

A product-form approximation is likely to perform better if the system of queues is relatively weakly coupled. A typical way weak coupling occurs is for the transition rates to depend only on the states of the queues involved.

Toward this end, first suppose that the external arrival rate $\lambda_i(t, \mathbf{n})$ to queue i is only a function of n_i, the number of customers at queue i; then we write $\lambda_i(t, n_i)$. Similarly, suppose that the departure rate $\mu_i(t, \mathbf{n})$ depends only on n_i; then we write $\mu_i(t, n_i)$. Finally, assume that $p_{ij}(t, \mathbf{n})$ is only a function of n_i and n_j; then we write $p_{ij}(t, n_i, n_j)$.

Now we observe that the computations with product-form approximations simplify. Note that we can sum over $\mathbf{n}_2$ in (4.3) to obtain

$$\begin{aligned}\dot{q}_1(t, \mathbf{n}_1) = &\sum_{i=1}^{m_1} q_1(t, \mathbf{n}_1 - \mathbf{e}_{1i})\lambda_i(t, n_i - 1) \\ + &\sum_{i=1}^{m_1}\sum_{j=1}^{n_1} q_1(t, \mathbf{n}_1 + \mathbf{e}_{1j} - \mathbf{e}_{1i})\mu_j(t, n_j + 1)p_{ji}(t, n_j + 1, n_i - 1) \\ + &\sum_{i=1}^{m_1}\sum_{j=m_1+1}^{m} q_1(t, \mathbf{n}_1 - \mathbf{e}_{1i})q_j(t, n_j + 1)\mu_j(t, n_j + 1)p_{ji}(t, n_j + 1, n_i - 1) \ , \end{aligned} \tag{5.1}$$

where $q_j(t, n_j)$ is the one-dimensional marginal distribution, i.e.,

$$q_j(t, n_j) = \sum_{\substack{\mathbf{n}_2: \\ n_{2j} = n_j}} q_j(t, \mathbf{n}_2) \ . \tag{5.2}$$

The system of ODEs in (5.1) becomes especially attractive when full product form is assumed within the second subset. Then the calculation (5.2) over all vectors $\mathbf{n}_2$ need not be performed. Then the ODEs for one queue j in the second subset, $m_1 + 1 \le j \le m$, become

$$\begin{aligned}\dot{q}_j(t, n_j) \quad = \quad & q_j(t, n_j - 1)\lambda_j(t, n_j - 1) + q_j(t, n_j)\mu_j(t, n_j)p_{jj}(t, n_j, n_j) \\ & + q_j(t, n_j - 1) \sum_{\substack{k=m_1+1 \\ k \ne j}}^{m} \sum_{n_k} q_k(t, n_k + 1)\mu_k(t, n_k + 1)p_{kj}(t, n_k + 1, n_j - 1) \\ & + q_j(t, n_j - 1) \sum_{\mathbf{n}_1} q_1(t, \mathbf{n}_1 + \mathbf{e}_{1i})\mu_i(t, n_i + 1)p_{ij}(t, n_i + 1, n_j - 1) \end{aligned} \tag{5.3}$$

Finally, if we assume full product form as in (4.1) and the special transition structure, then we get

$$\begin{aligned}\dot{q}_i(t, n_i) \quad = \quad & q_i(t, n_i - 1)\lambda_i(t, n_i - 1) + q_i(t, n_i)\mu_i(t, n_i)p_{ii}(t, n_i, n_i) \\ & + \sum_{\substack{k=1 \\ k \ne i}}^{m} \sum_{n_k} q_i(t, n_i - 1)q_k(t, n_k + 1)\mu_k(t, n_k + 1)p_{ki}(t, n_k + 1, n_i - 1) \ . \end{aligned} \tag{5.4}$$

If n_i runs from 0 to $N - 1$ for each i, then the sum in (5.4) has mN terms and there are mN ODEs to solve.

6. The Steady-State Consistency Check

The product-form approximation has additional theoretical justification if the steady-state distribution of the associated time-homogeneous CTMC has a product-form distribution. Then the time-dependent probability vectors approximately have product form when the transition rates change slowly, because then the time-dependent distribution has approximately the steady-state distribution. (This property is formalized and refined by UA approximations as in Massey and Whitt [11].) For example, this is the case when, in addition to the special structure we introduced in Section 4, $\lambda_i(t, n_i)$ and $p_{ji}(t, n_j, n_i)$ are independent of n_i; e.g., see p. 49 of Kelly [6]. Note that we can allow dependence of $\mu_j(t, n_j)$ and $p_{ji}(t, n_j, n_i)$ upon n_j. Hence this case includes a network of $M_t/M_t/s$ queues with time-dependent (but not state-dependent) Markovian routing. There can be any number of servers at each queue. Moreover, the model can be extended to allow general state-dependent service, so that queue-length dependent reneging can be included as well.

On the other hand, the time-dependent Erlang loss model with retrials considered by Grier, Massey, McKay and Whitt [4] fails to satisfy this consistency condition. It has arrivals at the queue of customers in retry mode only when the main queue has all servers busy. In particular, the transition rates for that model are

$$\begin{aligned}\lambda_1(t, \mathbf{n}) &= \begin{cases} \alpha(t), & n_1 < L \\ 0, & n_1 \geq L \end{cases} \\ \lambda_2(t, \mathbf{n}) &= \begin{cases} 0, n_1 < L \\ p_r\alpha(t), & n_1 \geq L \end{cases}\end{aligned} \tag{6.1}$$

$$\mu_1(t, \mathbf{n}) = n_1\mu_c \ , \quad \mu_2(t, \mathbf{n}) = n_2\mu_r \tag{6.2}$$

$$p_{12}(t, \mathbf{n}) = 0 \ , \quad p_{21}(t, \mathbf{n}) = 1 \ . \tag{6.3}$$

Intuitively, it is clear that the arrival process to the retry-mode queue is in fact more bursty than predicted by the model, which explains why the approximation tends to underestimate the number of customers in retry model. Exact and approximate steady-state distributions are natural to consider for making refinements to time-dependent decomposition approximations.

When the model satisfies the steady-state consistency condition and the transition rates do not change too quickly, we can anticipate that the approximation will perform well. In other cases, the approximations may nevertheless help.

7. Non-Exponential Service Times and Non-Poisson Arrival Processes

The model considered so far only naturally covers exponential service-time distributions (and time-dependent generalizations) and Poisson external arrival processes. However, other phase-type service-time distributions and interarrival-time distributions can be considered by inserting extra queues to represent exponential phases, as in Ong and Taaffe [13] and Taaffe and Ong [18]. For example, an Erlang E_k distribution can be represented by k queues in series, while a hyperexponential distribution can be represented by k queues in parallel, with probabilistic routing to the queues. The transition rates then depend on the set of queues associated with a service-time distribution. For example, for the series of queues representing an $M_t/M_t/1$ node, only one customer can be in the last $k-1$ queues, and the service rate can be positive only for the one customer in service.

The main point is that the model can be extended to cover non-exponential service times and non-Poisson arrival processes by enlarging the network, so that it all fits within the given framework. Then it is natural to use a partial product form in which all the queues representing the service-time distribution at an original queue are kept intact; i.e., we do not assume product form within these artificial queues, but even that could be modified.

References

[1] B. H. Andrews and H. L. Parsons, Establishing telephone agent staffing levels through economic optimization, *Interfaces* **23** (1993), 14–20.

[2] A. J. Brigandi, D. R. Dargon, M. J. Sheehan and T. Spencer, III, AT&T's call processing simulator (CAPS): operational design for inbound call centers, *Interfaces* **24** (1994), 6–28.

[3] L. V. Green, P. J. Kolesar and A. Svoronos, Some effects of nonstationarity on multiserver Markovian queueing systems, *Opns. Res.* **39** (1991), 502–511.

[4] N. Grier, W. A. Massey, T. McKoy and W. Whitt, The time-dependent Erlang loss model with retrials, *Telecommunication Systems* **7** (1997), 253–265.

[5] O. B. Jennings, A. Mandelbaum, W. A. Massey and W. Whitt, Server staffing to meet time-varying demand, *Management Sci.* **42** (1996), 1383–1394.

[6] F. P. Kelly, *Reversibility and Stochastic Networks*, Wiley, New York, 1979.

[7] B. Koopman, Air terminal queues under time-dependent conditions, *Opns, Res.* **20** (1972), 1089–1114.

[8] K. M. Malone, *Dynamic Queueing Systems: Behavior and Approximations for Individual Queues*, Ph.D. dissertation, Operations Research Center, MIT, 1995.

[9] A. Mandelbaum, W. A. Massey and M. I. Reiman, Strong approximations for Markovian service networks, *Queueing Systems*, to appear.

[10] W. A. Massey and W. Whitt, Networks of infinite-server queues with nonstationary Poisson input, *Queueing Systems* **13** (1993), 183–250.

[11] W. A. Massey and W. Whitt, Uniform acceleration expansions for Markov chains with time-varying rates, *Ann. Appl. Prob.*, 1998, to appear.

[12] V. Mehrota, Ringing up big business, *OR/MS Today* **24** (1997), 18–24.

[13] K. L. Ong and M. R. Taaffe, Nonstationary queues with interrupted Poisson arrivals and unreliable/repairable servers, *Queueing Systems* **4** (1989), 27–46.

[14] M. D. Peterson, D. J. Bertsimas and A. R. Odoni, Decomposition algorithms for analyzing transient phenomena in multiclass queueing networks in air transportation, *Opns. Res.* **43** (1995), 995–1011.

[15] D. M. Rappaport, Key role of integration in call centers, *Business Communications Review*, July 1996, 44–48.

[16] M. H. Rothkopf and S. S. Oren, A closure approximation for the nonstationary $M/M/s$ queue, *Management Sci.* **25** (1979), 522–534.

[17] B. W. Schmeiser and M. R. Taaffe, Time-dependent queueing network approximations as simulation external control variates, *Opns. Res. Letters* **16** (1994), 1–9.

[18] M. R. Taaffe and K. L. Ong, Approximating $Ph(t)/M(t)/S/C$ queueing systems, *Ann. Oper. Res.* **8** (1987), 103–116.

[19] W. Whitt, Variability functions for parametric-decomposition approximations of queueing networks, *Management Sci.* **41** (1995), 1704–1715.