The time-dependent Erlang loss model with retrials

Nathaniel Grier [a], William A. Massey [a], Tyrone McKoy [b,*] and Ward Whitt [c]

[a] *Bell Laboratories, 700 Mountain Avenue, Office 2C-120, Murray Hill, NJ 07974, USA*
E-mail: will@research.bell-labs.com
[b] *NCR, 2 Choke Cherry Road, Rockville, MD 20850, USA*
E-mail: tyrone.mckoy@washingtondc.ncr.com
[c] *AT&T Laboratories, 700 Mountain Avenue, Office 2C-178, Murray Hill, NJ 07974, USA*
E-mail: wow@research.att.com

We consider a generalization of the classical Erlang loss model with both retrials of blocked calls and a time-dependent arrival rate. We make exponential-distribution assumptions so that the number of calls in progress and the number of calls in retry mode form a nonstationary, two-dimensional, continuous-time Markov chain. We then approximate the behavior of this Markov chain by two coupled nonstationary, one-dimensional Markov chains, which we solve numerically. We also develop an efficient method for simulating the two-dimensional Markov chain based on performing many replications within a single run. Finally, we evaluate the approximation by comparing it to the simulation. Numerical experience indicates that the approximation does very well in predicting the time-dependent mean number of calls in progress and the times of peak blocking. The approximation of the time-dependent blocking probability also is sufficiently accurate to predict the number of lines needed to satisfy blocking probability requirements.

1. Introduction

In this paper we consider a generalization of the classical Erlang loss model which incorporates *two* important features of real service systems: (i) retrials, and (ii) time-dependent arrival rates. There is a substantial literature on generalizations of the Erlang loss model which incorporate each of these features separately. First, early work on stationary loss models with retrials was done by Kosten [6] and Cohen [1]; see section 9.2.4 of Syski [9]. Accounts of more recent work on stationary loss models with retrials can be found in the surveys by Yang and Templeton [13] and Falin [2] and in chapter 7 of the textbook by Wolff [12]. Second, early work on nonstationary loss models without retrials was done by Palm [8] and Khintchine [5]. Accounts of more recent work on nonstationary loss models without retrials can be found in Jagerman [3], Taaffe and Ong [10] and Massey and Whitt [7]. However, we are unaware of any previous work on nonstationary loss models with retrials.

We make assumptions so that the nonstationary loss model with retrials

* Meyerhoff scholar from the University of Maryland at Baltimore County. This work was supported in 1994 by the AT&T Summer Research Program (SRP).

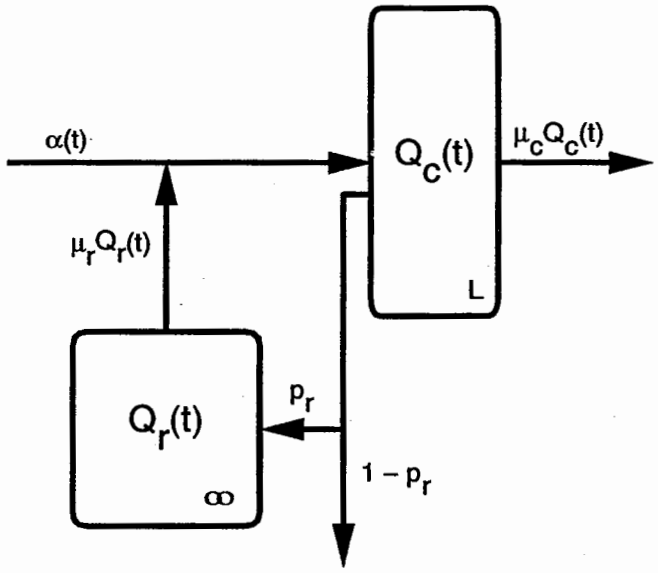

Figure 1. The call retry model.

can be represented as a two-dimensional *continuous-time Markov chain* (CTMC) $\{(Q_c(t), Q_r(t)):\ t \geqslant 0\}$, as depicted in figure 1. There are L lines (servers), $Q_c(t)$ is the number of calls in progress (i.e., the number of busy servers) at time t, and $Q_r(t)$ is the number of calls in retry mode (in orbit) at time t. We assume that the external arrival process is a nonstationary Poisson process with time-dependent intensity function $\alpha(t)$. We assume that the holding (service) times of successive calls to enter service are i.i.d. exponential random variables with mean μ_c^{-1}. Thus, the rate of service completion at time t is $\mu_c Q_c(t)$. Each arrival that finds all L lines busy is blocked. We assume that this call leaves the system with probability $1 - p_r$ and enters the retry mode with probability p_r, independent of previous history. Each call that enters the retry mode tries again after a random delay. We assume that the successive retry delays are i.i.d. exponential random variables with mean μ_r^{-1}. Thus the retry rate at time t is $\mu_r Q_r(t)$. Moreover, we assume that the arrival process, holding times and retry delays are all mutually independent. It is easy to see that these assumptions make $(Q_c(t), Q_r(t))$ a non-stationary CTMC on the state space $\{0, 1, \ldots, L\} \times \mathbb{Z}_+$, where $\mathbb{Z}_+$ is the set of nonnegative integers. We give the forward equations characterizing this CTMC in section 2. Our model has μ_c, p_r and μ_r constant and all time-variation in $\alpha(t)$, which seems to be the case of greatest interest, but we could also let μ_c, p_r and μ_r depend on t.

The time-dependent distributions $\mathrm{P}(Q_c(t) = j,\ Q_r(t) = k)$ can be obtained directly by numerically solving the forward equations if we modify the model to make the state space finite. For example, we can let the retrial probability be 0 instead of p_r when $Q_r(t) \geqslant R$ for a suitably large R. Then the total number of states is $(L + 1)(R + 1)$. Assuming that R is $\mathrm{O}(L)$, this makes the number of states, and thus the number of equations, $\mathrm{O}(L^2)$. Since typical cases of interest include $L = 100$ or $L = 1000$, the number of equations can be so large that computation is difficult.

To address this problem, we propose an alternative approximation scheme that has only $L+2$ equations. The idea is to assume, as an approximation, that $Q_c(t)$ and $Q_r(t)$ can be approximated by random variables $\overline{Q}_c(t)$ and $\overline{Q}_r(t)$ that are *probabilistically independent*; i.e., we assume that

$$\mathsf{P}\big(\overline{Q}_c(t) = j, \overline{Q}_r(t) = k\big) = \mathsf{P}\big(\overline{Q}_c(t) = j\big)\mathsf{P}\big(\overline{Q}_r(t) = k\big) \tag{1.1}$$

for all t, j and k. This allows us to treat the evolution of the one-dimensional probabilities $\mathsf{P}(\overline{Q}_c(t) = j)$ and $\mathsf{P}(\overline{Q}_r(t) = k)$ *separately* via separate systems of forward equations. Moreover, since $\overline{Q}_r(t)$ corresponds to an infinite-server system, we can describe its behavior through a single equation involving its mean $\mathsf{E}[\overline{Q}_r(t)]$. This reduction makes the total number of equations $L + 2$.

Of course, it is important to approximately capture the important dependence between these probabilities. We do this by making the time-dependent transition rates in each system depend on the time-dependent distribution of the other component; i.e., when considering the evolution of $\mathsf{P}(\overline{Q}_c(t) = j)$, we let the arrival rate from retrials be $\mu_r \mathsf{E}\overline{Q}_r(t)$; and when considering the evolution of $\mathsf{P}(\overline{Q}_r(t) = k)$, we let the arrival rate from retrials by new arrivals be $\alpha(t)p_r\mathsf{P}(\overline{Q}_c(t) = L)$ and the departure rate from the retry mode be $\overline{Q}_r(t)\mu_r(1 - p_r\mathsf{P}(Q_c(t) = L))$. The term $\overline{Q}_r(t)\mu_r p_r\mathsf{P}(\overline{Q}_c(t) = L))$ represents the rate of retrials completing a retry delay that immediately retry again because all L lines are busy again.

The overall approximation scheme can be regarded as time-dependent analog of the reduced-load (or Erlang) fixed-point approximation for blocking probabilities in stationary loss models; see Whitt [11] and Kelly [4] and references therein. The analog of the independence assumption (1.1) above is the facility-independence assumption (5) on [11, p. 1814].

It is significant that our approximation scheme reduces the analysis to two coupled time-dependent systems that have been analyzed previously. In particular, the process $\overline{Q}_c(t)$ evolves as an $M_t/M/L/0$ loss model, while $\overline{Q}_r(t)$ evolves as an $M_t/M_t/\infty$ model, as depicted in figure 2. Hence approximations for these more elementary non-stationary models can be used to obtain even simpler approximations. For example, the pointwise stationary approximation (PSA) and modified-offered-load (MOL) approximation could be used for the $M_t/M/L/0$ loss model; see [7]. The MOL approximation reduces the number of equations for the $M_t/M/L/0$ model from $L + 1$ to 1, and thus reduce the overall number of equations to 2. However, we found that the PSA and MOL approximations performed significantly worse than the exact computation of the $M_t/M/L/0$ probabilities in the approximation. Hence, we do not carefully examine such further simplifying approximations here, but their availability should be noted. The weakness of PSA is in overestimating the blocking probabilities at peak times and in not computing the *time* of peak blocking accurately. The nature of MOL is to be at its best when approximating small probabilities, but we are interested in analyzing the retry model when the blocking probabilities are relatively large.

We evaluate our approximations by making numerical comparisons with simulations. Our approach to simulation itself is worth mention. We obtain simulation efficiency by performing multiple replications within a single run.

Here is how the rest of this paper is organized. In section 2 we write down the functional forward equations for the nonstationary CTMC and derive the approximate equations from them. In section 3 we describe our simulation methodology. Finally, in section 4 we compare the approximations to simulations for a few numerical examples.

2. The functional forward equations

We start by giving the functional version of the forward equations for the CTMC. We then use it to derive the approximation. Let f be any bounded, real-valued function on the state space of the call retrial model. Then

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\big[f\big(Q_c(t),Q_r(t)\big)\big]\\
&\quad=\alpha(t)\mathsf{E}\big[f\big(Q_c(t)+1,Q_r(t)\big)-f\big(Q_c(t),Q_r(t)\big);\ Q_c(t)<L\big]\\
&\qquad+\alpha(t)p_r\mathsf{E}\big[f\big(Q_c(t),Q_r(t)+1\big)-f\big(Q_c(t),Q_r(t)\big);\ Q_c(t)=L\big]\\
&\qquad+\mu_c\mathsf{E}\big[Q_c(t)\big(f\big(Q_c(t)-1,Q_r(t)\big)-f\big(Q_c(t),Q_r(t)\big)\big)\big]\\
&\qquad+\mu_r\mathsf{E}\big[Q_r(t)\big(f\big(Q_c(t)+1,Q_r(t)-1\big)-f\big(Q_c(t),Q_r(t)\big)\big);\ Q_c(t)<L\big]\\
&\qquad+\mu_r(1-p_r)\\
&\qquad\times\mathsf{E}\big[Q_r(t)\big(f\big(Q_c(t),Q_r(t)-1\big)-f\big(Q_c(t),Q_r(t)\big)\big);\ Q_c(t)=L\big].
\end{aligned}\tag{2.1}$$

The set of equations obtained by letting f vary constitutes the forward equations. To focus on the marginal distribution of $Q_c(t)$, let $f(x,y)=g(x)$. Then the equations reduce to

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\big[g\big(Q_c(t)\big)\big]&=\alpha(t)\mathsf{E}\big[g\big(Q_c(t)+1\big)-g\big(Q_c(t)\big);\ Q_c(t)<L\big]\\
&\quad+\mu_c\mathsf{E}\big[Q_c(t)\big(g\big(Q_c(t)-1\big)-g\big(Q_c(t)\big)\big)\big]\\
&\quad+\mu_r\mathsf{E}\big[Q_r(t)\big(g\big(Q_c(t)+1\big)-g\big(Q_c(t)\big)\big);\ Q_c(t)<L\big].
\end{aligned}\tag{2.2}$$

Similarly, to focus on the marginal distribution of $Q_r(t)$, let $f(x,y)=h(y)$. Then the equations reduce to

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\big[h\big(Q_r(t)\big)\big]\\
&\quad=\alpha(t)p_r\mathsf{E}\big[h\big(Q_r(t)+1\big)-h\big(Q_r(t)\big);\ Q_c(t)=L\big]\\
&\qquad+\mu_r\mathsf{E}\big[Q_r(t)\big(h\big(Q_r(t)-1\big)-h\big(Q_r(t)\big)\big);\ Q_c(t)<L\big]\\
&\qquad+\mu_r(1-p_r)\mathsf{E}\big[Q_r(t)\big(h\big(Q_r(t)-1\big)-h\big(Q_r(t)\big)\big);\ Q_c(t)=L\big].
\end{aligned}\tag{2.3}$$

Next, in (2.3) let $h(y) = y$. Then we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\big[Q_r(t)\big] = \alpha(t)p_r\mathsf{P}\big(Q_c(t) = L\big) - \mu_r\mathsf{E}\big[Q_r(t);\ Q_c(t) < L\big] - \mu_r(1-p_r)\mathsf{E}\big[Q_r(t);\ Q_c(t) = L\big]. \tag{2.4}$$

Now if we assume that $Q_c(t)$ and $Q_r(t)$ are approximately independent, then (2.2) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\big[g\big(Q_c(t)\big)\big] \approx \big(\alpha(t) + \mu_r\mathsf{E}\big[Q_r(t)\big]\big)\mathsf{E}\big[g\big(Q_c(t)+1\big) - g\big(Q_c(t)\big);\ Q_c(t) < L\big] + \mu_c\mathsf{E}\big[Q_c(t)\big(g\big(Q_c(t)-1\big) - g\big(Q_c(t)\big)\big)\big]. \tag{2.5}$$

Since any function g above is uniquely defined by the values it takes on the integers $\{0, 1, \ldots, L\}$, we see that equation (2.5) is equivalent to the set of forward equations for the $M_t/M/L/0$ queue with arrival rate function $(\alpha(t) + \mu_r\mathsf{E}[Q_r(t)])$ and service rate μ_c. Moreover, with the independence approximation, (2.4) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\big[Q_r(t)\big] \approx \alpha(t)p_r\mathsf{P}\big(Q_c(t) = L\big) - \mu_r\big(1 - p_r\mathsf{P}\big(Q_c(t) = L\big)\big)\mathsf{E}\big[Q_r(t)\big], \tag{2.6}$$

which is equivalent to the differential equation for the mean queue length of an $M_t/M_t/\infty$ system with arrival rate function $\alpha(t)p_r\mathsf{P}(Q_c(t) = L)$ and (instead of μ_r) service-rate function $\mu_r(1 - p_r\mathsf{P}(Q_c(t) = L))$.

Based on (2.5) and (2.6), we propose the following approximation. Consider the *joint* process $(\overline{Q}_c(t), \overline{Q}_r(t))$ where for all $1 \leqslant n \leqslant L-1$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{P}\big(\overline{Q}_c(t) = n\big) = \big(\alpha(t) + \mu_r\mathsf{E}\big[\overline{Q}_r(t)\big]\big)\mathsf{P}\big(\overline{Q}_c(t) = n-1\big) + (n+1)\mu_c\mathsf{P}\big(\overline{Q}_c(t) = n+1\big) - \big(\alpha(t) + \mu_r\mathsf{E}\big[\overline{Q}_r(t)\big] + n\mu_c\big)\mathsf{P}\big(\overline{Q}_c(t) = n\big), \tag{2.7}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{P}\big(\overline{Q}_c(t) = 0\big) = \mu_c\mathsf{P}\big(\overline{Q}_c(t) = 1\big) - \big(\alpha(t) + \mu_r\mathsf{E}\big[\overline{Q}_r(t)\big]\big)\mathsf{P}\big(\overline{Q}_c(t) = 0\big), \tag{2.8}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{P}\big(\overline{Q}_c(t) = L\big) = \big(\alpha(t) + \mu_r\mathsf{E}\big[\overline{Q}_r(t)\big]\big)\mathsf{P}\big(\overline{Q}_c(t) = L-1\big) - L\mu_c\mathsf{P}\big(\overline{Q}_c(t) = L\big), \tag{2.9}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\big[\overline{Q}_r(t)\big] = \alpha(t)p_r\mathsf{P}\big(\overline{Q}_c(t) = L\big) - \mu_r\big(1 - p_r\mathsf{P}\big(\overline{Q}_c(t) = L\big)\big)\mathsf{E}\big[\overline{Q}_r(t)\big]. \tag{2.10}$$

The $L+1$ differential equations in (2.7)–(2.9) constitute the forward equations for the $M_t/M/L/0$ queue with arrival rate function $\alpha(t) + \mu_r\mathsf{E}[\overline{Q}_r(t)]$ and service rate μ_c. The additional equation (2.10) is the differential equation for the mean queue length of an $M_t/M_t/\infty$ system with arrival rate function $\alpha(t)p_r\mathsf{P}(\overline{Q}_c(t) = L)$ and service rate function $\mu_r(1 - p_r\mathsf{P}(\overline{Q}_c(t) = L))$. Hence, we have the two coupled Markov chains shown in figure 2.

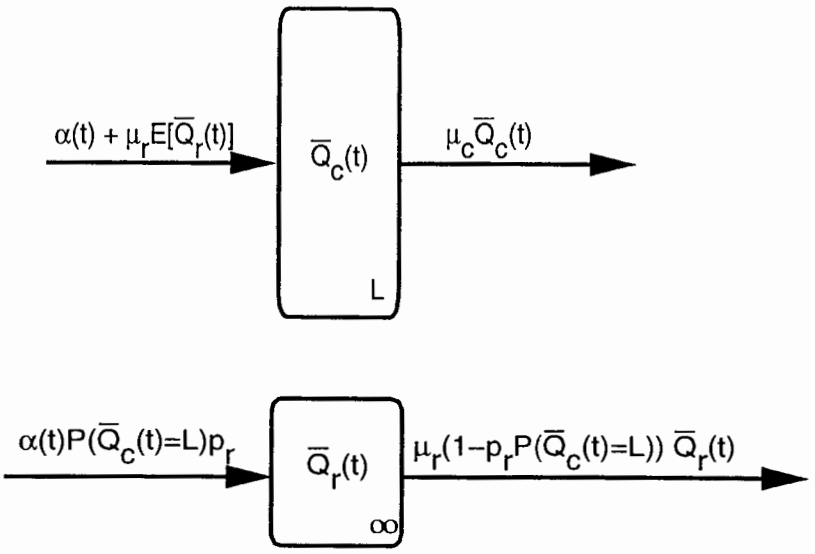

Figure 2. The approximate nonstationary model.

If we use an MOL approximation for $\overline{Q}_c$, then we can reduce the number of equations to 2. We first replace (2.7)–(2.9) by

$$\frac{\mathrm{d}}{\mathrm{d}t} m_c(t) = \alpha(t) + \mu_r \mathsf{E}\big[\overline{Q}_r(t)\big] - m_c(t)\mu_c. \tag{2.11}$$

We then use (2.10) after replacing $\mathsf{P}(\overline{Q}_c(t) = L)$ by $\beta_L(m_c(t))$, where $\beta_L(m)$ is the Erlang blocking formula with L servers and offered load m, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\big[\overline{Q}_r(t)\big] = \alpha(t)p_r\beta_L\big(m_c(t)\big) - \mu_r\big(1 - p_r\beta_L\big(m_c(t)\big)\big)\mathsf{E}\big[\overline{Q}_r(t)\big]. \tag{2.12}$$

It may be possible to understand the physics of the time-dependent retrial model better by investigating (2.11) and (2.12) analytically, but we do not pursue that here.

3. Simulation methodology

In addition to developing the new approximation described above, we develop a new efficient method for simulating the full nonstationary CTMC $(Q_c(t),\ Q_r(t))$ in order to calculate the exact time-dependent characteristics of interest. The simulation takes substantially longer than the approximation to obtain the characteristics of interest, but it is certainly viable.

Time-dependent queueing models have special features that invite nonstandard simulation methodology. There is no notion of steady-state, so that it is impossible to average over time. Thus, in order to achieve high precision, it is necessary to perform a very large number of independent replications. Also, the performance measures we seek are functions of time instead of a single number; e.g., we want to calculate the time-dependent blocking probability instead of a single long-run average blocking probability. Hence, the requirements are more demanding.

We have found that performing multiple replications within a single simulation run is an effective way to increase the efficiency of the simulation algorithm. For example, we might perform 1000 replications within one run. For this purpose, we divide the time interval of interest $[0, T]$ into T/δ short intervals each of length δ. We then update the states of all processes and compute the desired summary average statistics for each subinterval for all replications in one pass through the T/δ subintervals. A major advantage of this approach is that it reduces storage. With N separate replications, we would need to store N samples of the T/δ descriptive statistics which we would later average to produce a set of T/δ desired description statistics. In contrast, within one run, we only need to store the final T/δ averaged descriptive statistics. At each new time point $n\delta$ we update the state of all N processes. We also compute the average performance at that time. Since most of the N processes have no events in the small interval, we need only update the sum by the small number of non-null events. The average at that time point is more efficiently computed in this way.

We also exploit the Markovian structure in our generation of successive events in each process. For each replication of the nonstationary CTMC, we generate the sample path as follows. Let $\overline{\alpha}$ be a number greater than or equal to the supremum of the arrival rate $\alpha(t)$ over all time t of interest. If after an event occurs at time t, the system state is $Q_c(t) = j$ and $Q_r(t) = k$, then we let the time until the next potential event in the process have an exponential distribution with mean λ^{-1}, where $\lambda = \overline{\alpha} + j\mu_c + k\mu_r$. If the exponential variable takes the value s, then the next potential event occurs at time $t + s$. We let that event be a call completion with probability $j\mu_c/\lambda$, a retrial completion with probability $k\mu_r/\lambda$, an external arrival with probability $\alpha(t + s)/\lambda$, and a fictitious event corresponding to no state change with the remaining probability $(\overline{\alpha} - \alpha(t + s))/\lambda$. It is well known that this procedure produces sample paths with the correct distribution.

In summary, we created a simple C program to calculate $E[Q_c(t)]$, $E[Q_r(t)]$ and $P(Q_c(t) = L)$. For example, at any time t, we estimate $E[Q_c(t)]$ by

$$\widehat{Q}_c(t) = \frac{1}{N} \sum_{k=1}^{N} Q_c^{(k)}(t) \ ,$$

where N is the number of replications and $Q_c^{(k)}(t)$ is the number of calls in progress at time t in replication k.

4. Comparing the approximation to the simulation

We now evaluate our approximation by making comparisons with simulations of the nonstationary CTMC. We first consider an example with

$$\alpha(t) = 300 + 100 \sin(0.1\pi t), \tag{4.1}$$

which has peaks at times 5, 25, 45 and so on. We let $\mu_c = 1.0$, so that time is in units of mean call holding times. We let $\mu_r = 30.0$, making mean retrial delays

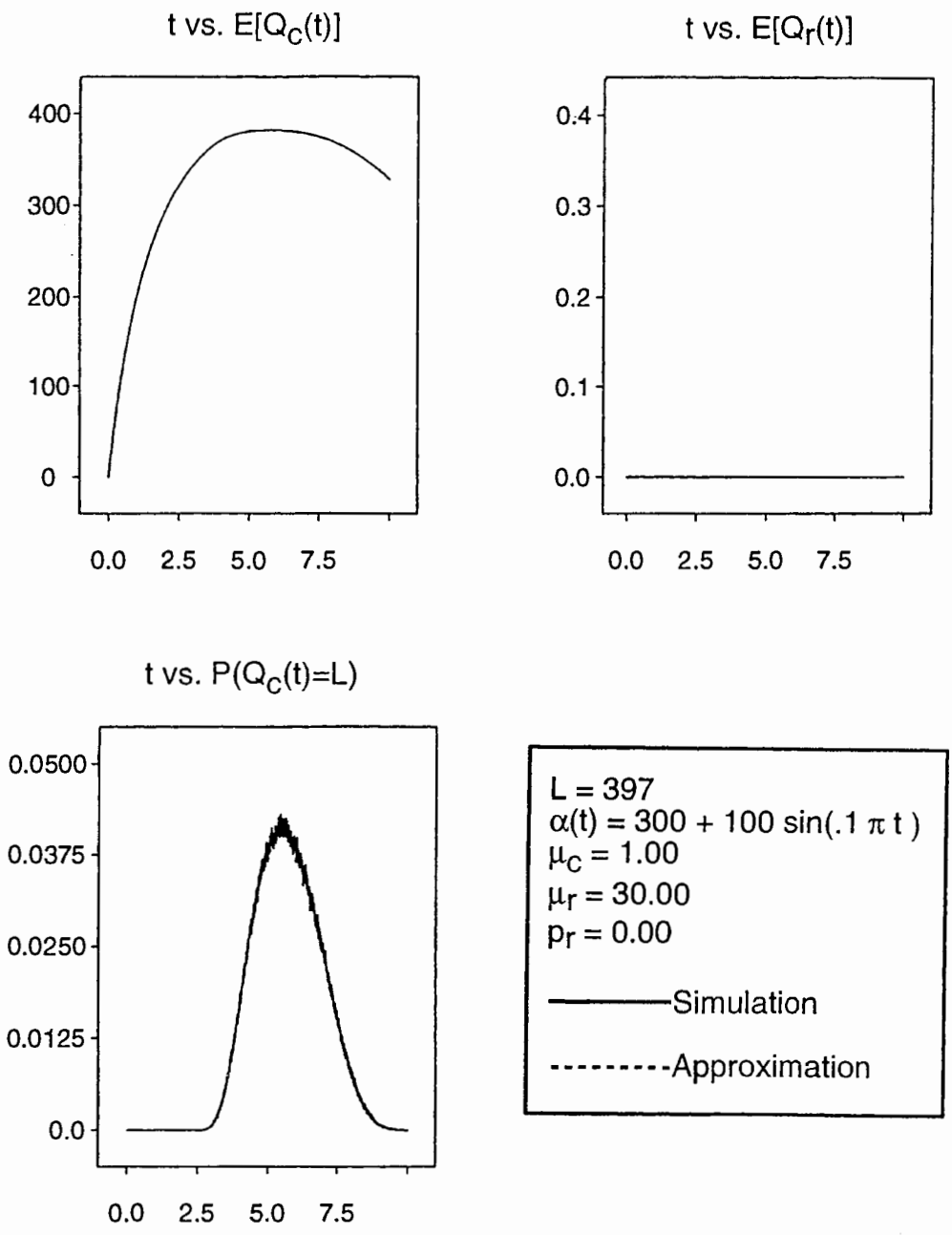

Figure 3. Numerical example of the simulations and approximations for the retry model.

substantially shorter than mean call holding times, as usually is the case. We let the number of lines be $L = 397$, which is slightly less than the peak offered load of 400. The value 397 was chosen to be the minimum number of lines such that the peak time-dependent blocking probability $\mathrm{P}(Q_c(t) = L)$ is less than or equal to 0.10 when the retrial probability p_r is 0.40.

In figures 3–5, we plot the time-dependent blocking probability $\mathrm{P}(Q_c(t) = L)$, the average number of calls in progress $\mathrm{E}[Q_c(t)]$, and the average number of blocked calls in retry mode $\mathrm{E}[Q_r(t)]$ as a function of time t for three values of the retrial

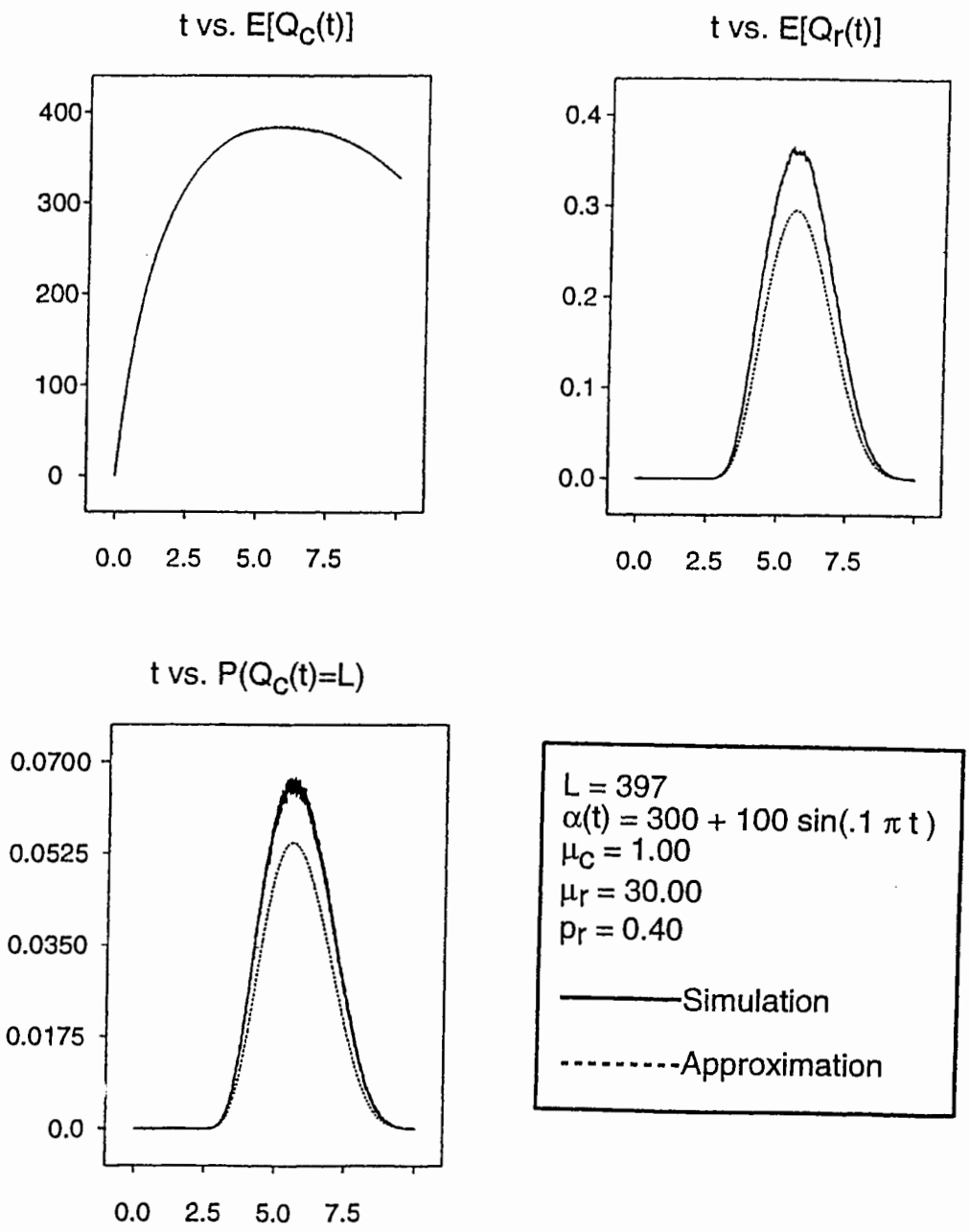

Figure 4. Numerical example of the simulations and approximations for the retry model.

probability p_r: $p_r = 0$, $p_r = 0.4$ and $p_r = 0.8$. In these cases the simulation had 10,000 replications. These plots illustrate both how the approximation compares to a simulation of the exact retrial model, as well as the effect that increasing the retry probability p_r has on the performance of the retry system. In figure 3, we have $p_r = 0$, so that there are no retries. In this case the call retry model and the approximate model are identical. The simulation and approximation are validated by the fact that the three graphs in figure 3 are each plots of *two* curves that are sitting on top of each other, with exceptions due only to numerical error. This example helps to gauge how well

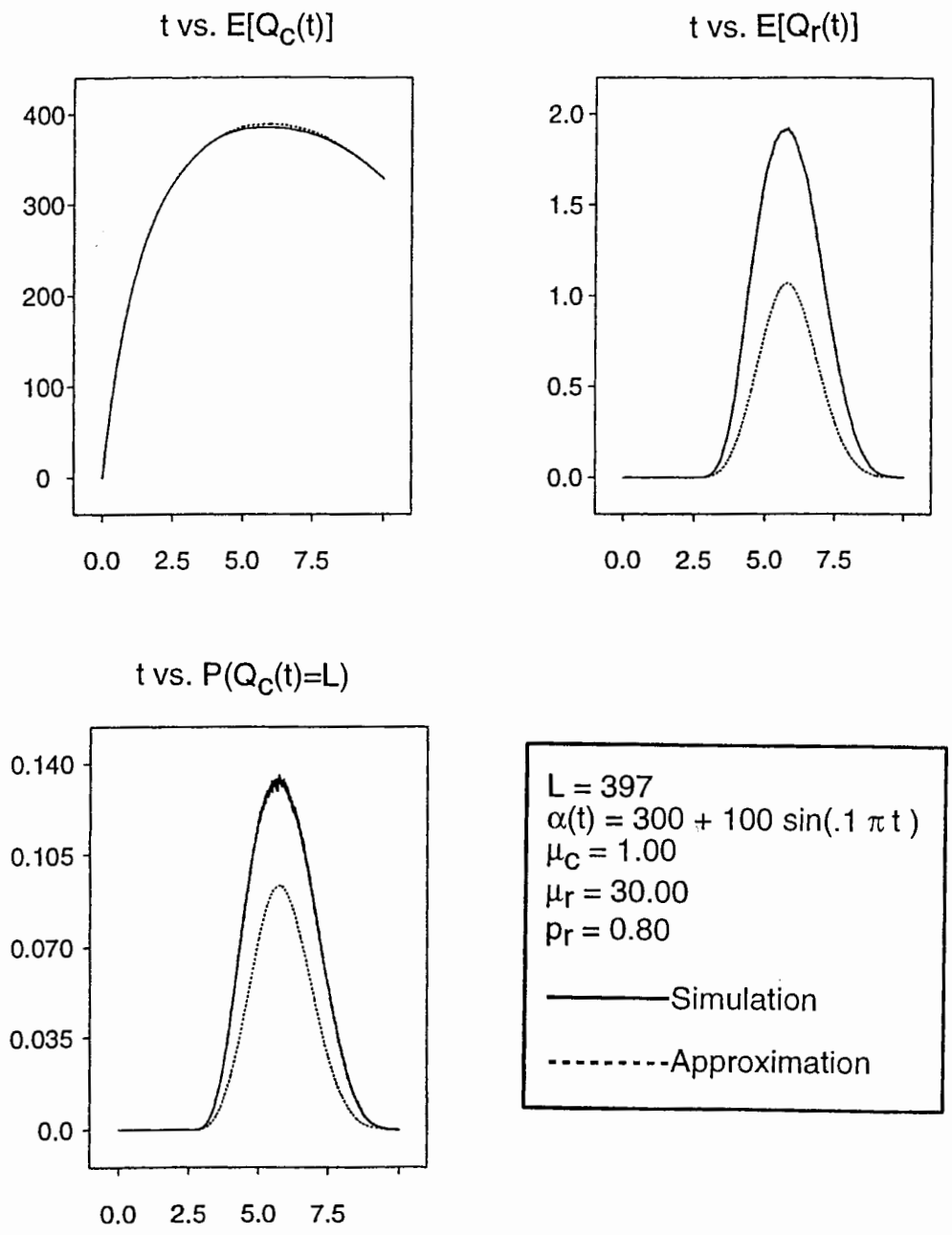

Figure 5. Numerical example of the simulations and approximations for the retry model.

the simulation is computing $\mathsf{E}[Q_c(t)]$, $\mathsf{E}[Q_r(t)]$, and $\mathsf{P}(Q_c(t) = L)$.

Figures 4 and 5 genuinely evaluate the approximation. From figures 4 and 5, we see that the mean number of calls in progress, $\mathsf{E}[Q_c(t)]$ is approximated spectacularly well, much better than the other two quantities $\mathsf{E}[Q_r(t)]$ and $\mathsf{P}(Q_c(t) = L)$. Upon reflection, this is to be anticipated since $\mathsf{E}[Q_c(t)]$ is a much larger quantity. Moreover, both $\mathsf{E}[Q_r(t)]$ and $\mathsf{P}(Q_c(t) = L)$ depend on the tail of the distribution of $Q_r(t)$.

Also, the approximation does an excellent job of locating the *time* of peak congestion, which lags after the peak in $\alpha(t)$ at $t = 5$. In all curves in figures 3–5, the times

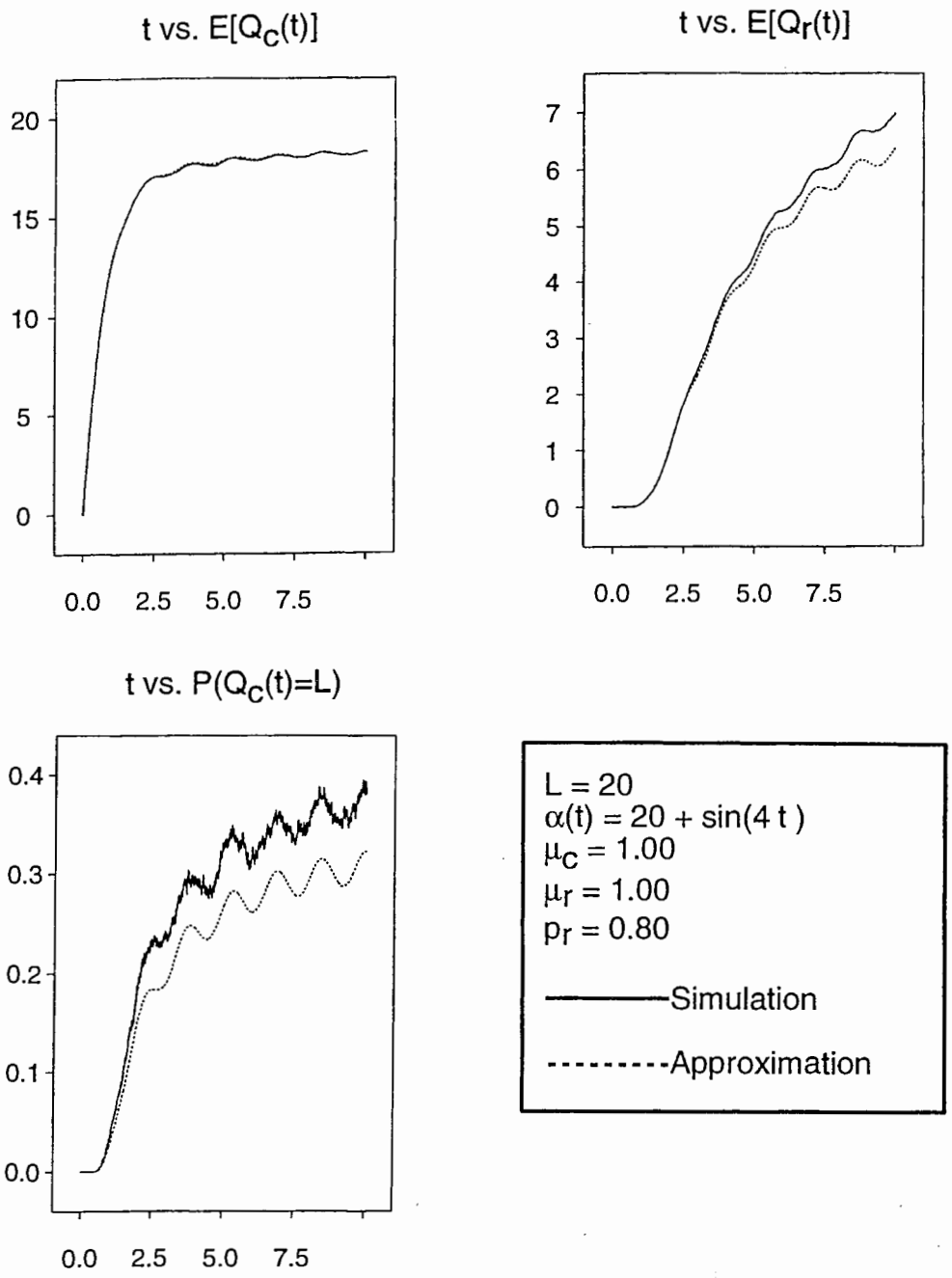

Figure 6. Numerical example of the simulations and approximations for the retry model.

for attaining maximum values are greater than 5.0 and the lag is more pronounced as the retry probability p_r increases.

The time-dependent blocking probability $P(Q_c(t) = L)$ itself is not especially well estimated by the approximation. Indeed, the true blocking probability is consistently *underestimated* by the approximation. This consistent ordering may make the approximation easier to interpret in applications. The direction of the error could be anticipated because the approximation fails to capture certain stochastic fluctuations.

This can easily be seen in the case of a stationary model. Then, in steady-state, the approximation has the net arrival processes to the two component subsystems be Poisson processes, when in fact they should be somewhat more variable, because of the overflow phenomenon. It is natural to consider refined approximations that try to capture the extra variability, perhaps exploiting peakedness, as in section 1.9 of Whitt [11]. However, we leave such investigations to future research.

Although the time-dependent blocking probability $P(Q_c(t) = L)$ is not predicted exceptionally well in figures 4 and 5, the accuracy tends to be sufficient for many engineering purposes. In particular, the approximation tends to do an excellent job in determining the minimum number of lines needed so that the peak blocking probability remains below a specified threshold. For example, in the setting of figure 4, the approximation yields the correct minimum number $L = 397$ so that the time-dependent blocking probability stays below 0.10. Thus, the approximation can be used effectively together with simulation. A few simulation runs can be used to verify or refine settings determined by several runs of the approximation.

Figure 6 shows a different situation, with much lower arrival rate and faster fluctuations. In particular, the arrival rate here is

$$\alpha(t) = 20 + \sin(4t). \tag{4.2}$$

The retrial rate is slower as well, here being 1.0 instead of 30.0. Here the number of lines is equal to the average offered load 20, so that the peak rate 30 substantially exceeds the number of available lines. Figure 6 shows that the approximation performs reasonably well.

References

[1] J.W. Cohen, Basic problems of telephone traffic theory and the influence of repeated calls, Philips Telecom. Rev. 18 (1957) 49–100.

[2] G. Falin, A survey of retrial queues, Queueing Systems 7 (1990) 127–167.

[3] D.L. Jagerman, Nonstationary blocking in telephone traffic, Bell System Technical Journal 54 (1975) 625–661.

[4] F.P. Kelly, Loss networks, Ann. Appl. Probab. 1 (1991) 319–378.

[5] A.Y. Khintchin, Mathematical methods in the theory of queueing, Trudy Math. Inst. Steklov. 49 (1955) (in Russian). English translation by Charles Griffin and Co., London (1960).

[6] L. Kosten, On the influence of repeated calls in the theory of probabilities of blocking, De Ingenieur 59 (1947) 1–25 (in Dutch).

[7] W.A. Massey and W. Whitt, An analysis of the modified offered load approximation for the nonstationary Erlang loss model, Ann. Appl. Probab. 4 (1994) 1145–1160.

[8] C. Palm, Intensity variations in telephone traffic, Ericsson Technics 44 (1943) 1–189 (in German). English translation by North-Holland, Amsterdam (1988).

[9] R. Syski, *Introduction to Congestion Theory in Telephone Systems* (North-Holland, Amsterdam, 2nd ed., 1986).

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

[11] W. Whitt, Blocking when service is required from several facilities simultaneously, AT&T Tech. J. 64 (1985) 1807–1856.

[12] R.W. Wolff, *Stochastic Modeling and the Theory of Queues* (Prentice-Hall, Englewood Cliffs, NJ, 1989).
[13] T. Yang and J.G.C. Templeton, A survey on retrial queues, Queueing Systems 2 (1987) 201–233.

Nathaniel Grier received the B.S. degree in electrical engineering in 1981 from Howard University (Summa Cum Laude), the M.S. degree in computer engineering from Carnegie-Mellon University in 1982 and the M.S. degree in electrical engineering from the University of Pennsylvania while studying as a fellow in the Bell Labs Cooperative Research Fellowship Program. He has worked at Bell Laboratories since 1982 in the Telecommunications IC Group. From 1989 to 1991, he was a visiting professor in the Electrical Engineering Department at Prairie View A&M University. Since 1991 he has worked in the Mathematical Sciences Research Center and the Wide Area Network IC Group at Bell Laboratories. He is currently the marketing manager for SONET/SDH IC's in the Transmission IC Group.

William A. Massey received the A.B. degree in 1977 from Princeton University in mathematics (Magna Cum Laude, Phi Beta Kappa, and Sigma Xi) and was awarded a Bell Labs Cooperative Research Fellowship to attend graduate school at Stanford University. In 1981, he received his Ph.D. degree from Stanford in mathematics. Since 1981, he has been a member of technical staff in the Mathematical Science Research Center at Bell Laboratories, Murray Hill, NJ (which is now the research and development organization for Lucent Technologies). He has written papers on nonstationary queues, stochastic ordering, queueing networks, database theory, and wireless communications. His research interests include queuing theory, applied probability, and performance modelling of telecommunication systems. Dr. Massey is a member of INFORMS and the American Mathematical Society (AMS).

Tyrone McKoy received the B.S. degree in mathematics from the University of Maryland Baltimore County (UMBC) in 1995. He is currently completing the M.S. degree, which is scheduled to be awarded in 1997, in mathematics at the University of Maryland Graduate School Baltimore (UMGSB). While attending UMBC in the Meyerhoff Scholarship Program, he had the opportunity to develop his mathematics background and engage in state-of-the-art research at IBM, The Johns Hopkins School of Medicine, the Space Telescope Science Institute (a subsidiary of NASA), and AT&T Bell Laboratories. He worked on the research in this paper at AT&T Bell Laboratories while participating in the Bell Labs Summer Research Program. He is currently working as a Professional Services Consultant at NCR. His research interests include performance modeling of telecommunication systems, applied probability, and queueing theory.

Ward Whitt received the A.B. degree in mathematics from Dartmouth College in 1964 and the Ph.D. degree in operations research from Cornell University in 1969. He taught in the Department of Operations Research at Stanford University in 1968–1969 and in the Department of Administrative Sciences at Yale University from 1969 to 1977. Since 1977, he has been employed by AT&T. He currently works in the Network Mathematics Research Department of AT&T Labs-Research. His research interests include queueing theory, stochastic processes, stochastic models in telecommunications and numerical inversion of transforms. Dr. Whitt is a member of INFORMS and the Institute of Mathematical Statistics.