THE TIME-DEPENDENT ERLANG LOSS MODEL WITH RETRIALS

by

Nathaniel Grier, William A. Massey, Tyrone McKoy[1] and Ward Whitt

AT&T Bell Laboratories
Murray Hill, NJ 07974-0636

June 27, 2000

Abstract

Keywords: performance analysis, traffic theory, nonstationary queues, time-dependent arrival rates, retrials, infinite-server queues, nonstationary Erlang loss model, Markovian simulation.

[1]Tyrone McKoy was supported by the AT&T Summer Research Program.

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 with 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 can be represented as a two-dimensional *continuous-time Markov chain* (CTMC) $\{(Q_c(t), Q_r(t)) : t \geq 0\}$, 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 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. 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 $\{0, 1, \ldots, L\} \times Z_+$, where 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 $P(Q_c(t) = j,\ Q_r(t) = k)$ 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 $Q_r(t) \geq R$ for a suitably large R. Then the total number of states is $(L+1)(R+1)$. Assuming that R is $O(L)$, this makes the number of states, and thus the number of equations, $O(L^2)$. 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 $\bar{Q}_c(t)$ and $\bar{Q}_r(t)$ that are *probabilistically independent*; i.e., we assume that

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

for all t, j and k. This allows us to treat the evolution of the one-dimensional probabilities $P(\bar{Q}_c(t) = j)$ and $P(\bar{Q}_r(t) = k)$ *separately* via separate systems of forward equations. Moreover, since $\bar{Q}_r(t)$ corresponds to an infinite-server system, we can describe its behavior through a single equation involving its mean $E\bar{Q}_r(t)$. 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 $P(\bar{Q}_c(t) = j)$, we let the arrival rate from retrials be $\mu_r E\bar{Q}_r(t)$; and when considering the evolution of $P(\bar{Q}_r(t) = k)$, we let the arrival rate from retrials by new arrivals be $\alpha(t)p_r P(\bar{Q}_c(t) = L)$ and the departure rate from the retry mode be $\bar{Q}_r(t)\mu_r(1 - p_r P(Q_c(t) = L))$. The term $\bar{Q}_r(t)\mu_r p_r P(\bar{Q}_c(t) = L))$ 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]. The analog of the independence assumption (1.1) above is the facility-independence assumption (5) on p. 1814 of [11].

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 $\bar{Q}_c(t)$ evolves as an $M_t/M/L$ loss model, while $\bar{Q}_r(t)$ evolves as an $M_t/M_t/\infty$ model. Hence approximations for these more elementary nonstationary 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 $M_t/M/L$ loss model; see [7]. The MOL approximation reduces the number of equations for the $M_t/M/L$ 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 $M_t/M/L$ probabilities in the approximation. Hence, we do not discuss 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 most 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 seems 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. In Section 3 we describe our simulation methodology. In Section 4 we compare the approximations to simulations for a few numerical examples. Finally, in Section 5 we state our conclusions.

References

[1] Cohen, J. W. (1957) Basic Problems of Telephone Traffic Theory and the Influence of Repeated Calls. *Philips Telecom. Rev.* **18**, 49–100.

[2] Falin, G. (1990) A Survey of Retrial Queues. *Queueing Systems* **7**, 127–167.

[3] Jagerman, D. L. (1975), Nonstationary Blocking in Telephone Traffic, *Bell System Technical Journal*, **54**, 625–661.

[4] Kelly, F. P. (1991) Loss Networks, *Ann. Appl. Prob.* **1**, 319–378.

[5] Khintchine, A. Y. (1955) *Mathematical Methods in the Theory of Queueing*, Trudy Math. Inst. Steklov 49 (in Russian, English translation by Charles Griffin and Co., London, 1960).

[6] Kosten, L. (1947) On the Influence of Repeated Calls in the Theory of Probabilities of Blocking, *De Ingenieur* **59**, 1–25 (in Dutch).

[7] Massey, W. A. and Whitt, W. (1994) An Analysis of the Modified Offered Load Approximation for the Nonstationary Erlang Loss Model, *Ann. Appl. Prob.* **4**, to appear.

[8] Palm, C. (1943) Intensity Variations in Telephone Traffic, *Ericsson Technics*, **44**, 1–189 (in German). (English translation by North-Holland, Amsterdam, 1988).

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

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

[11] Whitt, W. (1985) Blocking When Service Is Required from Several Facilities Simultaneously, *AT&T Tech. J.* **64**, 1807–1856.

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

[13] Yang, T. and Templeton, J. G. C. (1987) A Survey on Retrial Queues. *Queueing Systems* **2**, 201–233.