THE DYNAMIC ANALYSIS AND DESIGN OF A COMMUNICATION LINK WITH STATIONARY AND NONSTATIONARY ARRIVALS

WENHONG TIAN, HARRY PERROS
DEPARTMENT OF COMPUTER SCIENCE
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NC 27965
WTIAN@UNITY.NCSU.EDU, HP@CSC.NCSU.EDU

Abstract. Most research in queueing theory is typically based on the steady-state analysis. In today's dynamically changing traffic environment, the steady-state analysis may not provide enough information to operators regarding Quality of Service (QoS) requirements and dynamic design. In addition, the steady state analysis is not practical for nonstationary arrivals. In this paper, we consider the time-dependent behavior of a communication link depicted by an Erlang loss queue with stationary and nonstationary arrival rates. The time-dependent analysis for stationary arrival rates captures the dynamic nature of the system during its transient phase. The time-dependent analysis for nonstationary arrivals is of great interest since the arrival rate in most communication systems varies over time.

In this paper, we review and compare various methods that have been proposed in the literature for the time-dependent analysis of the nonstationary Erlang loss system for both stationary arrivals and nonstationary arrivals. We try to classify five practical methods into two categories: (1) closed-form exact solution and closed-form approximation; (2) numerical exact solution and numerical approximation. Our work tries to compare their computation complexity and accuracy. We apply some of these techniques to dimensioning dynamically a single communication system.

Key words. Markov chain, Stationary arrivals, Nonstationary arrivals, Blocking probability, Dynamic dimensioning

Introduction. In this paper, we consider the time-dependent behavior and design of a communication link with stationary and also with nonstationary arrival rates. Time-dependent (transient) analysis is motivated by the dynamic nature of the traffic. Riordan [14] introduced different methods for the transient analysis of a single service center. The necessary and sufficient conditions for a queueing network to have a transient product-form solution are provided by Taylor and Boucherie [15]. A new dimensioning approach for optical networks under nonstationary arrival rates was introduced by Nayak and Sivarajan [13].

Queueing models with nonstationary arrival rates have been studied extensively by Abdalla and Boucherie [1], Alnowibet [2], Alnowibet and Perros [3], Jagerman [6], Karagiannis et al. [7], Massey and Whitt [10], Massey [11], and Nayak and Sivarajan [13]. The nonstationary blocking probability for an Erlang loss queue was first obtained by Jagerman [6] using the modified offered load (MOL) approach. Massey and Whitt [10] developed analytical bounds on the error between the MOL approximation and the exact solution for an Erlang loss queue with a nonstationary arrival rate. A modified offered load approximate product-form solution was introduced by Abdalla and Boucherie [1] for mobile networks. A survey for the nonstationary analysis of the Erlang loss queue can be found in Alnowibet and Perros [3]. Mandjes and Ridder [9] proposed large deviation solutions for the transient analysis of the Erlang loss model with a stationary arrival rate. Massey [11] analyzed different queues with time-varying arrival rate for telecommunication models. Nayak and Sivarajan [13] introduced a dynamic dimensioning approach for optical networks under nonstationary arrival rates. Karagiannis et al. [7] showed that the traffic of the internet backbone network can be characterized by a nonstationary Poisson process.

In this paper, we review and compare various techniques that have been reported in the literature for the calculation of transient blocking probabilities of an Erlang loss queue assuming a stationary and nonstationary arrival rate. We also dimension a communication link, modelled by an Erlang loss queue for both stationary and nonstationary arrivals.

The paper is organized as follows. In section 1 we describe the behavior of an Erlang

loss queue as a function of time assuming that the arrival rate is either constant or nonstationary, i.e. a function of time t. We also show how an Erlang loss queue can be dimensioned using time-dependent blocking probabilities. In section 2, we review closed-form solutions of the transient behavior of an Erlang loss queue assuming constant and nonstationary arrival rates. Section 3 reviews an approximation method based on a property of truncated Markov processes, and section 4 describes a numerical procedure known as the fixed point approximation (FPA). An alternative approach to dimensioning a communication link, based on the method of large deviations, is presented in section 5. Numerical results are given in section 6. Finally the conclusions are summarized in section 7.

1. The Nonstationary Erlang Loss System. An Erlang loss queue is a system consisting of s servers and no waiting room. A customer is lost if it arrives at a time when all servers are busy. The loss queue is commonly used to model the telephone network. It has been extensively studied in the stationary case, i.e., assuming that the arrival process is a homogeneous Poisson process, or more generally, an Interrupted Poisson processes, and the service rate is exponentially distributed. (It has been shown that in a loss system, the blocking probability is insensitive to the service distribution but it only depends on its mean). The nonstationary loss queue, where the arrival rate is time-dependent is also of great interest.

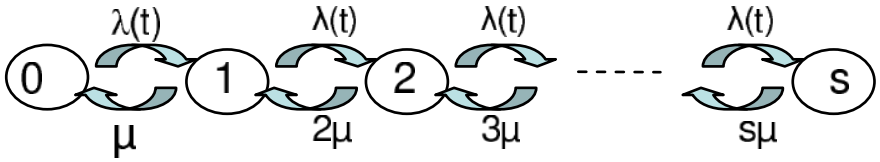

FIG. 1.1. *The Markov chain of Erlang loss model for a single service center*

The rate diagram of a loss queue with nonstationary arrivals $(M_t/M/s/s)$ is shown in FIG.1.1, where s is the total number of servers, $\lambda(t)$ is the time-dependent arrival rate and μ is the service rate. We say that the arrival process is stationary if it is time-independent, i.e., $\lambda(t) = \lambda$, and nonstationary if it is time-dependent (or time-varying). In this case $\lambda(t)$ is a single continuous or discrete function of time. We discuss these two cases in the following two subsections.

1.1. Stationary Arrivals. Let us consider a loss queue M/M/s/s with a time-independent Poisson arrival rate λ. Each arrival requests a service that requires an exponential amount of time with mean $1/\mu$, and it is performed by a single server. The queue has s identical servers and there is no waiting room. The probability that there are n, n=0, 1, ..., s, customers in the queue at time t, $P_n(t)$, is given by the following set of forward differential equations:

$$P_0'(t) = \mu P_1(t) - \lambda P_0(t) \tag{1.1}$$

$$P_n'(t) = \lambda P_{n-1}(t) + (n+1)\mu P_{n+1}(t) - (\lambda + n\mu)P_n(t), \tag{1.2}$$

$$P_s'(t) = \lambda P_{s-1}(t) - s\mu P_s(t) \tag{1.3}$$

where $P_0(t)+P_1(t)+P_2(t)+...+P_s(t)$=1, and $0 \leq P_n(t) \leq 1$, for $t \geq 0$ and n=$0, 1, 2, ..., s$, with initial conditions: $P_0(0)$=1 and $P_n(0)$=0, n=1, 2, 3, ..., s.
A numerical example of the time-dependent blocking probability is shown in FIG.1.2. These probabilities were obtained by solving equations (1.1)-(1.3) using an ordinary differential

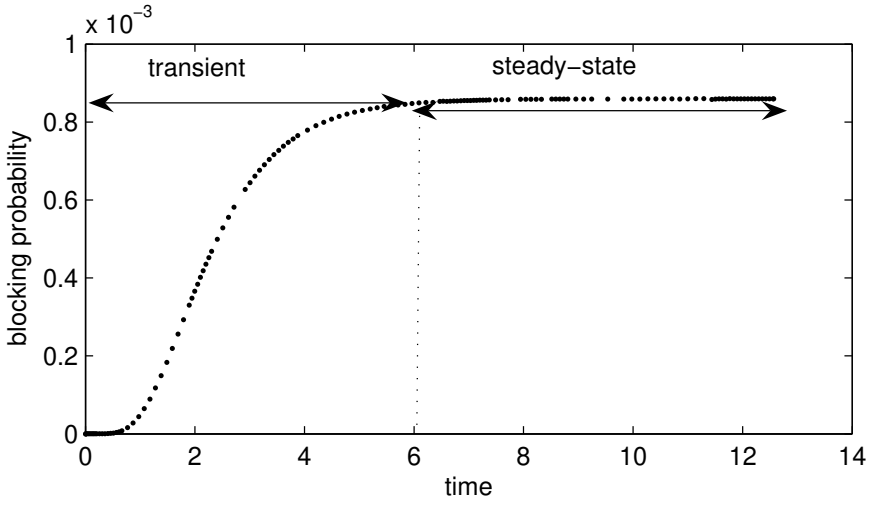

FIG. 1.2. *The time-dependent blocking probabilities of the stationary arrival for* $M/M/8/8$*, with offered load* $\rho = 2$

equation (ODE) solver. We note that the blocking probability reaches steady-state when $t = 6$. We also note that the steady-state probability can be very different from that during the transient state. Therefore, dimensioning the network based on the steady-state may result in overprovisioning during the transient period. This may have little impact if the duration of the transient period is very short. However, if the transient period is long, for example, a few months for optical networks, then it may be advantageous to dimension the network using the time-dependent blocking probability instead of the steady-state probability.

1.2. Time-varying Arrivals. Let us consider a loss queue M(t)/M/s/s with a time-dependent arrival rate $\lambda(t)$. Each arrival requests a service that requires an exponential amount of time with mean $1/\mu$. The probability that there are n, n=0, 1, ... , s, customers in the queue at time t, $P_n(t)$, is represented by the following forward differential equations:

$$P_0'(t) = \mu P_1(t) - \lambda(t) P_0(t) \tag{1.4}$$

$$P_n'(t) = \lambda(t) P_{n-1}(t) + (n+1)\mu P_{n+1}(t) - (\lambda(t) + n\mu) P_n(t), \tag{1.5}$$

$$P_s'(t) = \lambda(t) P_{s-1}(t) - s\mu P_s(t) \tag{1.6}$$

where $P_0(t) + P_1(t) + P_2(t) + ... + P_s(t)$=1, $t \geq 0$, and $0 \leq P_n(t) \leq 1$, for $t \geq 0$ and n=$0, 1, 2, ..., s$, with initial conditions: $P_0(0)$=1 and $P_n(0)$=0, n=1, 2, ..., s.

In FIG.1.3, we show a numerical example of the time-dependent blocking probability for a single link obtained assuming a periodic arrival rate function $\lambda(t) = 180{+}50sin(2(t{+}20))$. These probabilities were calculated numerically by solving equations (1.4)-(1.6) using an ODE solver. We note that the blocking probability in this case also has a transient period followed by repeating periods. The periodic behavior looks like the steady-state behavior of the stationary arrival case but the blocking probabilities follow a repeating pattern.

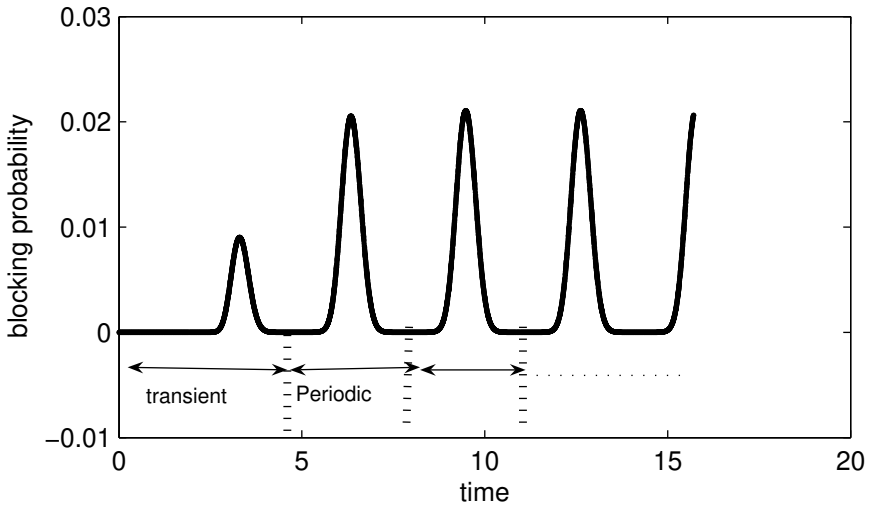

FIG. 1.3. *The time-dependent blocking probabilities of the nonstationary arrival where* $\lambda(t) = 180 + 50sin(2(t + 20))$ *and s=220*

1.3. Dimensioning A Single Link . A link can be dimensioned using the time-dependent blocking probability for both stationary and nonstationary arrivals. This is done by calculating the number of servers s so that the blocking probability is under a given threshold at any time t.

1.3.1. Stationary Arrivals. Let us consider an Erlang loss queue with $\lambda = 10$. The number of servers required so that the blocking probability is less than 0.01 is given in FIG.1.4. The solid line labelled 'steady-state dimensioning' gives the optimum number of servers calculated using the steady-state blocking probability of an Erlang loss model with $\lambda = 10$. The dotted line labelled 'time-dependent dimensioning' gives the result using the time-dependent blocking probability of the arrival rate, obtained using differential equations (1.1) to (1.3). We calculate the number of servers iteratively until the blocking probability is less than 0.01. Note that these two curves are the same after the transient phase is over. As we can see, the dimensioning results are quite different for these two scenarios with the time-dependent dimensioning requiring fewer servers.

1.3.2. Time-varying arrivals. We consider an example where the arrival rate varies as shown in FIG.1.5. We assume that time is divided into 12 periods, where each period for instance can be a month. During each period i, the arrival rate is constant. In FIG.1.5, the 12 arrival rates are: $\lambda(t)$=[8, 1, 3, 6, 2, 5, 12, 9, 11, 4, 7, 10]. We dimension the link so that at any time, the blocking probability is less than 0.01. The dimensioning results are also shown in FIG.1.5, and they were obtained assuming that all servers are free at time $t = 0$. These results were obtained as follows: we first calculate the number of servers for the first period using equations (1.4)-(1.6), assuming an empty system at time $t = 0$ and service rate $\mu = 1$, so that the nonstationary blocking probability is less than 0.01. This is done as before, in an iterative manner. We repeat this process for the second period assuming that at the beginning of the period the number of customers in it is equal to the average number of customers in the system.

This process is repeated until all 12 periods have been analyzed. We note that we solve

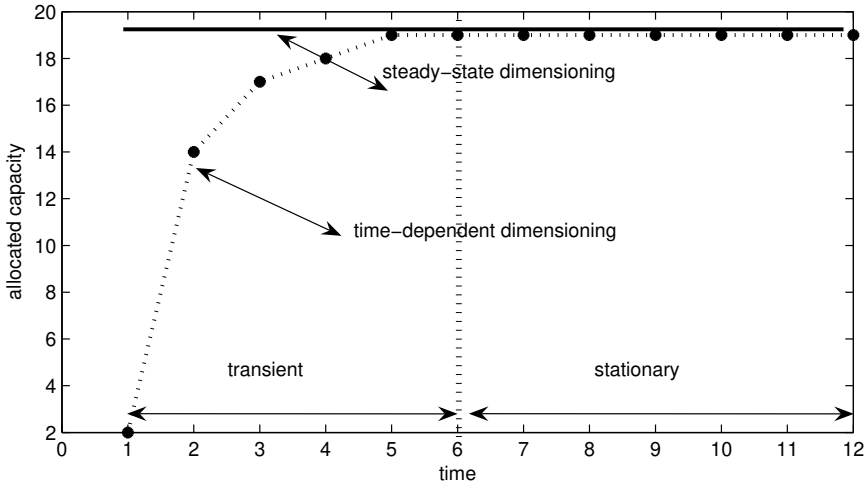

FIG. 1.4. *A dimensioning example of a single link with stationary arrival rates*

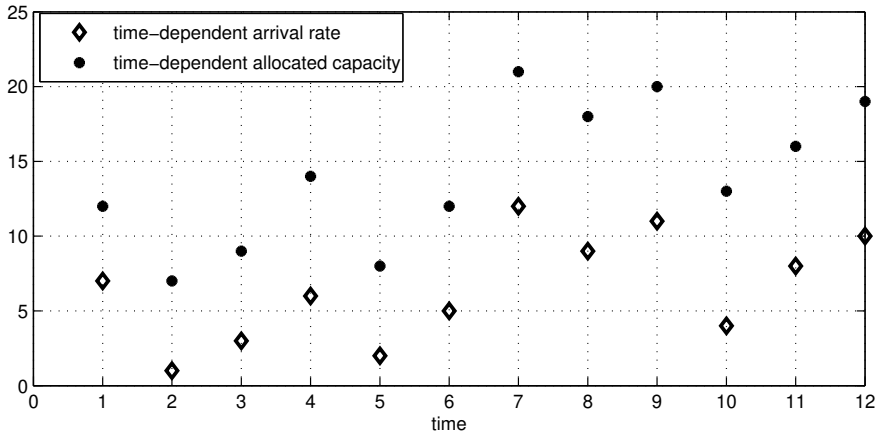

FIG. 1.5. *A dimensioning example of a single link with nonstationary arrival rates*

this problem by breaking it into periods and analyzing each period separately. Alternatively, we could solve equations (1.4)-(1.6) for the entire 12 period arrival process. In this case, we will not have to approximate the initial condition for each period. However, it is difficult to define $\lambda(t)$ as a single function over the entire 12 periods.

We note that the dimensioning results are very sensitive to the initial conditions and the arrival rates, and the required number of servers follows the arrival process.

2. Exact Closed-form Solutions for the Time-dependent Blocking Probability. In this section, we review closed-form solutions for the transient analysis of the Erlang loss queue under a constant and nonstationary arrivals.

2.1. Stationary Arrivals. In this case, the time-dependent blocking probability can be obtained using equations (1.1)-(1.3). We have $P_n'(t) \to 0$ for all n as $t \to \infty$ in the dif-

ferential equations (1.1)-(1.3). The differential equations (1.1)-(1.3) reduce to a set of linear equations from which we can obtain the closed-form solution for the probability P_n that there are n customers in the system.

$$P_n = \lim_{t\to\infty} P\{q(t) = n\} = \frac{\rho^n/n!}{\sum_{i=0}^{s} \rho^i/i!}, n = 0, 1, ..., s \tag{2.1}$$

where ρ=λ/μ, and $q(t)$ is the number of customers in the system at time t. The probability of blocking B_p is:

$$B_p = \lim_{t\to\infty} P\{q(t) = s\} = \frac{\rho^s/s!}{\sum_{i=0}^{s} \rho^i/i!} \tag{2.2}$$

This is the well-known Erlang B formula. The average number of customers in the system (i.e. the average number of busy servers) is: $\lim_{t\to\infty} E[q(t)]$=E[q]=$(1 - B_p)\rho$.

The time-dependent blocking probability function at time t, can be obtained from the differential equations (1.1)-(1.3). We have:

$$P_s(t) = \beta e^{(Q^T)t}\alpha \tag{2.3}$$

where α is the initial state probability vector , β is an all-zero row vector except that the last entry is 1, and Q is the infinitesimal generator matrix of the underlying Markov chain, defined as:

$$\begin{bmatrix} -\lambda & \lambda & 0 & \cdots & 0 \\ \mu & -\mu-\lambda & \lambda & \cdots & 0 \\ 0 & 2\mu & -2\mu-\lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda & 0 \\ 0 & \cdots & (s-1)\mu & -(s-1)\mu-\lambda & \lambda \\ 0 & \cdots & 0 & s\mu & -s\mu \end{bmatrix} \tag{2.4}$$

The computation complexity of equation (2.3) is $O(s^3)$ using a scaling and squaring algorithm with a Pade approximation [12]. The solution to the differential equations (1.1)-(1.3) can be found using different methods. An elegant closed-form solution can be obtained using Sylvester matrix theorem. Let us assume that we start from an empty system and let Q^T be the transpose of Q, $P(t) = [P_0(t), ..., P_s(t)]^T$, and $P(0) = [1, 0, ..., 0]^T$. Now the system of differential equations (1.1)-(1.3) can be written in the following form:

$$P'(t) = Q^T P(t), \tag{2.5}$$

Solving this equation by applying the property of exponential function, we get

$$P(t) = e^{(Q^T)t} P(0), \tag{2.6}$$

Let $D = e^{Q^T t} = [d_0, d_1, ..., d_s]$ where d_i is the column vector of D. Then we have $P(t) = d_0$ and $P_i(t) = d_{0,i}$.

Theorem 1 : All the $(s+1)$ eigenvalues of Q are real and distinct and one eigenvalue of Q is zero .

Proof: We first change Q to a symmetric tridiagonal matrix A with positive subdiagonal elements by the similarity transform (this does not change the eigenvalues of Q). We have: A=$D^{-1}QD$, where D is a diagonal matrix. Let us set A^* as the conjugate transpose of matrix A. Then we know that A^*=A.
We now show that all eigenvalues of A are real as follow: let (λ, x) be the (eigenvalue, eigenvector) pair of matrix A.

$$Ax = \lambda x, \tag{2.7}$$

Multiplying both sides of equation (2.7) by x^*, we have

$$x^*Ax = \lambda x^*x, \tag{2.8}$$

Taking the transpose of equation (2.8) and noticing that $A^* = A$, we get:

$$x^*Ax = \lambda^* x^*x, \tag{2.9}$$

where λ^* is the conjugate transpose of λ. Comparing equation (2.8) and (2.9), we see that $\lambda = \lambda^*$, which means that all eigenvalues of A are real.
A is a tridiagonal matrix, so we can compute the characteristic polynomial with a three-term recurrence (just do a determinant expansion) to construct a Sturm sequence. Since off-diagonal elements are positive, the matrix can only have simple eigenvalues (by the property of Sturm sequence). Since det(Q)=0, then det(Q-0*I)=0, which means that zero is one of the eigenvalues of matrix Q. This can also be seen by the fact that row sum of matrix Q is zero.
Theorem 2: Sylvester's matrix theorem for distinct roots (eigenvalues) (see Frazer [5]): If $G(U)$ is any polynomial of the square matrix U, and if x_i represents one of the n distinct eigenvalues of U, then

$$G(U) = \sum_{i=1}^{n} \frac{G(x_i)Adj(x_iI - U)}{\prod_{j\neq i}(x_j - x_i)} \tag{2.10}$$

An important application of Sylverster's theorem is in finding a closed-form solution for the matrix exponential. The following explains how to use this theorem to get the closed-form solution to our $M/M/s/s$ transient analysis problem. Let $x_0, x_1, ..., x_s$ be the $(s+1)$ eigenvalues of Q. From Sylvester's theorem we have

$$D = e^{Q^T t} = \sum_{r=0}^{s} e^{x_r t} \frac{Adj(x_rI - Q^T)}{\prod_{i\neq r}(x_r - x_i)} \tag{2.11}$$

where Adj(U) is the Adjoint matrix of U. Especially, the probability that all servers are busy at time t can be further simplified as:

$$P_s(t) = \sum_{r=0}^{s} e^{x_r t} \frac{(-1)^{s+1} det(M)}{\prod_{i\neq r}(x_r - x_i)} \tag{2.12}$$

where $det(M) = \lambda^s$, λ is the average arrival rate and M is the submatrix of Q^T with size $s \times s$. We can obtain all the eigenvalues of matrix Q using the fast algorithm with complexity $O(s^2)$), reported in [4]. So the computation complexity of equation (2.12) is roughly $O(s^3)$. From $P_s(t)$ we can also know the steady-state probability. The steady-state probability is just the constant part (which corresponds to the zero eigenvalue of matrix Q) of $P_s(t)$. Other quantities of interest such as $P_n(t)$ and average number of busy servers at time t can also be calculated.

2.2. Nonstationary Arrivals. The closed-form solution to differential equations (1.4)-(1.6) is complex even for fairly small systems with special arrival rate function $\lambda(t)$, see Alnowibet [2]. An explicit solution is provided in Jagerman [6] by using the probability generating functions of the state probabilities and the corresponding binomial moments where the arrival rate function $\lambda(t)$ is considered to be continuous. Following Jagerman [6], we have that the probability of j calls arriving in the time interval (0, t) is given by

$$\frac{[\int_0^t a(u)du]^j}{j!} exp(-\int_0^t a(u)du) \tag{2.13}$$

where a(t) is Poisson-offered load given by a(t)=$\lambda(t)/\mu$. We normalize the service rate μ= 1, so that $a(t)$ is measured in Erlangs. Let us define the Volterra operator K_r

$$K_r f = \int_0^t K_r(t,\tau) f(\tau) d\tau, r = 0, 1, ..., N. \tag{2.14}$$

The time-dependent blocking probability that all servers are busy at time t is

$$P_s(t) = \frac{\gamma(t,0)}{s!} - K_s(t,\tau) P_s(t) \tag{2.15}$$

hence the explicit form of the solution is :

$$P_s(t) = \frac{\Lambda(t)^s}{s!} - K_s \frac{\Lambda(t)^s}{s!} + K_s^2 \frac{\Lambda(t)^s}{s!} - ..., \tag{2.16}$$

where

$$\Lambda(t) = e^{-t} \int_0^t e^u a(u) du \tag{2.17}$$

and K_s is a Volterra operator defined by the kernel

$$K_s(t,\tau) = \sum_{j=0}^{s-1} \frac{\gamma(t,0)^j}{j!} e^{-(s-j)(t-\tau)} \binom{N}{s-j-1} a(\tau) \tag{2.18}$$

where

$$\gamma(t,\tau) = e^{-t} \int_\tau^t e^u a(u) du \tag{2.19}$$

Note that $\Lambda(t) = \gamma(t,0)$. We see that the above explicit solution is quite complicated for an arbitrary arrival rate $\lambda(t)$. The computation complexity of equation (2.16) is approximately $O(s^3)$ depending on how many terms used in the series. In view of this, it is not useful in practice.

3. The Truncated Markov Process Approximation. The following Corollary holds for truncated reversible Markov process (see Kelly [8]).

Corollary 1: If a reversible Markov process X_t with state space S and equilibrium distribution $\Pi(j), j \in S$, is truncated to the set of $S_1 \subset S$, then the resulting Markov process Y_t is reversible in equilibrium and has the equilibrium distribution:

$$\Pi_1(j) = \frac{\Pi(j)}{\sum_{k \in S_1} \Pi(k)}, j \in S_1. \tag{3.1}$$

It is interesting to note that the equilibrium distribution of the truncated process is just the conditional (renormalized within the truncated state space) probability of the original process. An efficient way to obtain the stationary distribution of the M/M/s/s queue is to use the fact that the M/M/s/s queue is a truncated process of an $M/M/\infty$, which is a reversible Markov process. Therefore:

$$P\{q = n\} = P\{q_\infty = n | q_\infty < s\} = \frac{\rho^n/n!}{\sum_{i=0}^{s} \rho^i/i!} \tag{3.2}$$

For the time-dependent analysis of M/M/s/s queue, we also can apply this truncation property approximately. First let us consider the transient behavior of the $M/M/\infty$ queue Riordan [14]. We have the following differential equations:

$$P_n'(t) = -(\lambda + n\mu)P_n(t) + \lambda P_{n-1}(t) + (n+1)\mu P_{n+1}(t), \tag{3.3}$$

$$P_0'(t) = -\lambda P_0(t) + \mu P_1(t), \tag{3.4}$$

where

$$\sum_{i=0}^{\infty} P_i(t) = 1, t \geq 0 \tag{3.5}$$

and

$$0 \leq P_n(t) \leq 1, for\ t \geq 0\ and\ n = 0, 1, 2, ..., \infty, \tag{3.6}$$

with initial conditions: $P_0(0)$=1 and $P_n(0)$=0, n=1, 2, 3, ..., ∞. Applying the z-transform approach, we can obtain the transient distribution of $M/M/\infty$ Riordan [14] :

$$P_n^\infty(t) = \frac{m^n e^{-m}}{n!}, m \equiv m(t) = \rho(1 - e^{-\mu t}). \tag{3.7}$$

For the time-dependent analysis of an M/M/s/s queue with constant arrival rate, we can apply the truncation property approximately as follows:

$$P_n^s(t) \approx \frac{P_n^\infty(t)}{\sum_k P_k^\infty(t)}, k \in \{0, 1, ...s\} \tag{3.8}$$

Notice that this equation will be close to the exact solution as time increases for small blocking probabilities. Also it is a good approximation for low blocking probabilities. Despite its appealing closed-form solution, equation (3.8) is non-trivial to compute for large value of s since it involves factorial terms which may cause numerical instability problems such as overflows. So, we can adapt the well-known recursive formula from the steady-state as follows:

$$B(k+1, t) = \frac{m(t)B(k, t)}{k + 1 + m(t)B(k, t)} \tag{3.9}$$

where $B(s, t)$ is the probability that all s servers are busy at time t and $B(0, t) = 1$.

The M(t)/M/s/s can also be approximated by truncating the $M(t)/M/\infty$ queue. This method is referred to as the modified offered load (MOL) method and it was first proposed by

Jagerman [6]. For the $M(t)/M/\infty$ queue, the time-dependent blocking probability is given by:

$$P_n^{\infty}(t) = \frac{\rho(t)^n e^{-\rho(t)}}{n!}. \tag{3.10}$$

where $\rho(t)$=$e^{-t}\int_0^t \lambda(u)e^u du$. For the M(t)/M/s/s queue, the probability $P_n(t)$ that there are n customers in the system is:

$$P_n^{\infty}(t) \approx P\{q_{\infty}(t) = n | q_{\infty}(t) < s\} = \frac{\rho(t)^n/n!}{\sum_{i=0}^{s}\rho(t)^i/i!} \tag{3.11}$$

In this case, the following recursive formula can be used:

$$B(k+1,t) = \frac{\rho(t)B(k,t)}{k+1+\rho(t)B(k,t)} \tag{3.12}$$

with $B(0,t)$=1 as the initial condition. $B(s,t)$ is the probability that all s servers are busy at time t, $P_s^s(t)$. The truncated $M/M/\infty$ provides an exact solution to an M/M/s/s queue in the steady-state due to the reversibility property. However, this property is lost in the nonstationary case [2]. Hence the truncated $M(t)/M/\infty$ provides an approximate solution to the M(t)/M/s/s. Massy and Whitt [10] developed analytical bounds on the error between the MOL approximation and the exact solution of the M(t)/M/s/s queue. The computation complexity of equation (3.10) and (3.12) is $O(s)$.

Experiments showed that the actual blocking probability of the M(t)/M/s/s queue should be less than 0.1 in order for the MOL to provide a good approximation, see Alnowibet and Perros [3]. As expected, the MOL underestimates the blocking probability of a loss queue with high load, i.e. when the exact blocking probability is high.

4. Numerical Solutions.

4.1. Differential Equations Solver. Equations (1.1)-(1.3) and (1.4)-(1.6) can be solved numerically using an ODE (ordinary differential equation) solver. The numerical results in section 2 were obtained using the ODE solver of Matlab 6.5, which can solve efficiently an Erlang loss queue with a few hundreds servers. However, as reported in Moler et al. [12], ODE solver may be very expensive. We also found that ODE solver is not suitable for dimensioning.

4.2. The Fixed Point Approximation (FPA) Method . The fixed point approximation (FPA) method was proposed by Alnowibet and Perros [3]. This method calculates numerically the time-dependent mean number of customers and blocking probability functions in a nonstationary loss queue. The FPA method was also extended to nonstationary queueing networks of multi-rate loss queues and nonstationary queueing networks with population constraints, see Alnowibet and Perros [3]. The main idea of the FPA method is as follows:

Given a loss queue M(t)/M/s/s with time-dependent rate $\lambda(t)$, the time-dependent average number of customers $E[Q(t)]$ can be expressed as the difference between the effective arrival rate and the departure rate at time t. That is:

$$E'[Q(t)] = \lambda(t)(1 - B_p(t)) - \mu E[Q(t)] \tag{4.1}$$

We note that the time-dependent mean number of customers is given by the expression: $E[Q(t)]$=$\rho(t)(1-B_p(t))$, from which we can develop the following expression for the offered load $\rho(t)$:

$$\rho(t) = E[Q(t)]/(1 - B_p(t)). \tag{4.2}$$

where

$$B_p(t) = \frac{\rho(t)^s/s!}{\sum_{i=0}^{s} \rho(t)^i/i!} \tag{4.3}$$

Using equations (4.1)–(4.3), we can calculate the blocking probability iteratively as follows.

1). Choose an appropriate Δt, final time T_f and tolerance ϵ.

2). Choose initial conditions for E[Q(t)]. Set E[Q(0)]=0.

3). Evaluate $\lambda(t)$ at t=0,$\Delta t, 2\Delta t, ..., T_f$.

4). Start with an initial blocking probability $B_p^0(t)$=0, t=0,Δt,2Δt,...,T_f.

5). Set the iteration counter k=0.

6). Solve numerically for $E[Q^k(t)]$ using the following equation:

$E[Q^k(t + \Delta t)]$=$E[Q^k(t)]$+$\lambda(t)(1 - B_p^k(t))\Delta t - \mu E[Q^k(t)]\Delta t$.

7). Calculate $\rho^k(t)$=$E[Q^k(t)]/(1 - B_p^k(t))$,t=0,Δ t,2Δ t,...,T_f.

8). Calculate the blocking probability $B_p^{k+1}(t)$=$[\rho^k(t)]^s/s!/(\sum_{i=0}^{s}[\rho^k(t)]^i/i!)$,t=0,$\Delta$ t,2Δ t,...,T_f.

9). If $\|(B_p^k(t) - B_p^{k+1}(t))\| < \epsilon$, then $B_p^k(t)$ has converged and the algorithm stops. Else, set $k = k + 1$, and go to step 6).

The FPA algorithm does not require a closed-form expression for the arrival rate function. It only requires that the arrival rate function be defined at time points equally spaced by Δt. In view of this, any arrival rate function can be used despite whether we know its closed-form or not. Since this algorithm discretizes the arrival rate function, the continuity and differentiability properties of the arrival rate function are not necessary. The computation complexity of this algorithm to find blocking probability is $O(sT_f/\Delta t)$. In all the experiments the FPA results were very close to the exact numerical results or within the simulation confidence intervals. This leads to the conjecture that the blocking probability obtained by FPA is exact (see, Alnowibet and Perros [2]). For dimensioning purpose, we need to slightly modify the algorithm in order to obtain the capacity for any time t given a blocking probability threshold. This can be done by adding an iterative procedure into the main algorithm.

5. The Large Deviation Approach. In this section, we obtain an expression for dimensioning the Erlang loss queue using the large deviation method. For stationary arrivals, Mandjes and Ridder [9] have obtained approximate expressions for $P_n(t)$, the probability of having n customers at time t. This expression is extended in the case of nonstationary arrivals. The large deviation theory is similar to the Central Limit theory (CLT). The CLT governs random fluctuations only near the mean, which are of the order of $\delta/\sqrt{n}$, where δ is the standard deviation. Fluctuations which are of the order of δ are, relative to typical fluctuations, much bigger: they are large deviations from the mean. They happen only rarely, and so the large deviation theory is often described as the theory of rare events, that is, events which take place away from the mean, out in the tails of the distribution. The main idea of the large deviation approach for the nonstationary Erlang loss queue is as follows. An asymptotic regime is obtained by scaling the arrival process. This is done by replacing $\lambda(t)$ with $n\lambda(t)$. The number of sources active at time t are partitioned into the sources that were active at time 0 and are still active at time t, and the sources that became active in (0,t) and are still active at time t. We then can apply Cramer's theorem and Chernoff's formula to obtain the result.

5.1. Stationary Arrivals. Assuming exponential service time distribution, Mandjes and Ridder [9] obtained the following expression:

$$P_s(t) \approx e^{s(ln(\gamma(t))-\gamma(t)+1)}, \gamma(t) = \lambda_1/\mu(1 - e^{-t}) \tag{5.1}$$

where $\lambda = s\lambda_1$ and s is the total number of servers.

$P_s(t)$ can be better approximated using Bahadur-Rao approximation [9]. We have:

$$P_s(t) \approx \frac{1}{\sqrt{2\pi s}\delta\theta} e^{s(log(\gamma(t))-\gamma(t)+1)}, \gamma(t) = \lambda_1/\mu(1-e^{-t}), \tag{5.2}$$

where $\theta = -log(\gamma(t))$, $\delta^2 = \frac{M^{''}(\theta)}{M(\theta)}$ and $M(\theta) = e^{\gamma(t)(e^{\theta}-1)}$.

Mandjes and Ridder [9], pointed out that the Bahadur-Rao approximation (5.2) is more accurate than (5.1). We note that the large deviation theory yields simple expressions of the time-dependent blocking probability for stationary arrivals.

5.2. Nonstationary Arrivals. Let us assume that the service time is exponentially distributed with unit mean for nonstationary arrivals. Then expression (5.1) can be extended as follows:

$$P_s(t) \approx e^{s(log(\gamma(t))-\gamma(t)+1)} \tag{5.3}$$

where $\gamma(t) = e^{-t}\int_0^t \lambda_1(u)e^u du$,$\lambda(t) = s\lambda_1(t)$.

The Bahadur-Rao approximation given by (5.2) can be extended as follows:

$$P_s(t) \approx \frac{1}{\sqrt{2\pi s}\delta\theta} e^{s(log(\gamma(t))-\gamma(t)+1)}, \tag{5.4}$$

where $\gamma(t) = e^{-t}\int_0^t \lambda_1(u)e^u du$, $\theta = -log(\gamma(t))$, $\delta^2 = \frac{M^{''}(\theta)}{M(\theta)}$ and $M(\theta) = e^{\gamma(t)(e^{\theta}-1)}$. The computation complexity of equation (5.2) and (5.4) is $O(1)$.

For both stationary and nonstationary cases, the number of servers C for which the blocking probability is ϵ can be obtained using an iterative procedure: starting by the candidate allocation $C = n_0$, the candidate allocation is increased by one until the blocking probability is below the threshold ϵ.

6. Numerical Results. In FIG.6.1, we give the blocking probability calculated at a specific time $t = 4.8$ using four different methods for $\lambda_1(t) = 0.7 + 0.2sin(2\pi t)$. The time t=4.8 was chosen because by that time the system is out of the transient state for the given periodic arrival rate function. The initial condition was set to an empty system. The blocking probability is plotted in the logarithmic scale against the total number of servers s. The graph labelled "LD" shows the results obtained from the large deviation theory using equation (5.3), the graph labelled "BR" gives the results obtained using the Bahadur-Rao equation (5.4), and, the graph labelled "TR" gives the results obtained using equation (3.8) from the truncated Markov process approximation. The exact solution is obtained using the fixed point approximation (FPA).

Running more examples by varying the arrival rate with (high load, medium load, low load), we note that the truncated Markov process approximation provides a very good approximation to the exact solution but underestimates the blocking probabilities. The large deviations approximation differs considerably from the exact blocking probability and is less accurate than the Bahadur-Rao approximation. Because of page limit, we do not provide all the examples here.

In FIG.6.2, we show a dimensioning example for a communication link over 20 observation periods. We assume that the arrival rate is constant during each period. The values of the arrival rates are given in FIG.6.2. We calculated the capacity, i.e., the number of servers, iteratively so that at any time the blocking probability is less than 0.005. The dimensioning results are shown in FIG.6.2, where we assume that all servers are free at time t=0. 'BR' represents the results obtained using the Bahadur-Rao equation (5.4), and 'FPA' gives the results

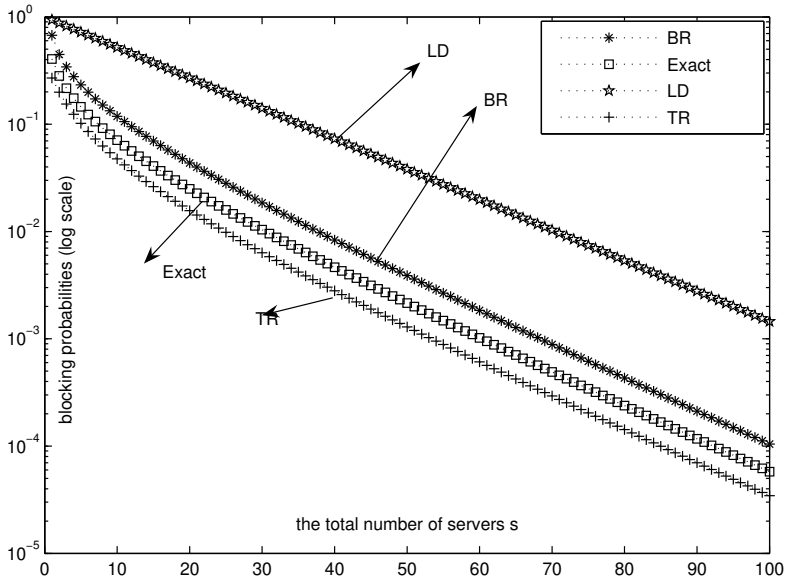

FIG. 6.1. *A comparison among the exact solution (Exact), the truncated Markov process approximation (TR), the Bahadur-Rao approximations(BR) and the large deviations approximations (LD)*

obtained using the fixed point approximation (see, Alnowibet and Perros [3]). The results obtained using the other two methods, i.e. the truncated Markov process approximation and equation (5.3) from the large deviation theory, are not shown because they have a large error. We also carried out a variety of numerical results under different loads, and here we only show a representative sample of these results, since they are all similar. We observed that Bahadur-Rao (BR) approximation is very close to the exact results but it overestimates the capacity. We note that both BR and exact allocated capacity follow the pattern of the arrival rates.

7. Conclusions. In this work, we reviewed and compared various time-dependent analysis techniques of a single loss queue with stationary and nonstationary arrivals. The aim of time-dependent analysis is to dimensioning a link in a more efficient way.

It is difficult to answer the question "Which method is the best ? ". One method maybe preferable over the others when considering computation complexity and accuracy of the results. We have the following observations:

1) For stationary arrivals, the exact closed-form solution and the truncated Markov process approximation are CPU efficient and easy to implement for medium size systems whereas the large deviation approach is preferred if the system is large.

2)For the nonstationary arrivals:

If the arrival rate is a single continuous function, then the truncated Markov process approximation (MOL) and FPA method will be a good choice for medium size systems, and the large deviation approach is a better choice if the system is large.

If the arrival rate is not a continuous function, the FPA method is a better choice.

The FPA method can work for both stationary arrivals and nonstationary arrivals and it can be used for medium size systems. For large systems, we may consider using the Bahadur-

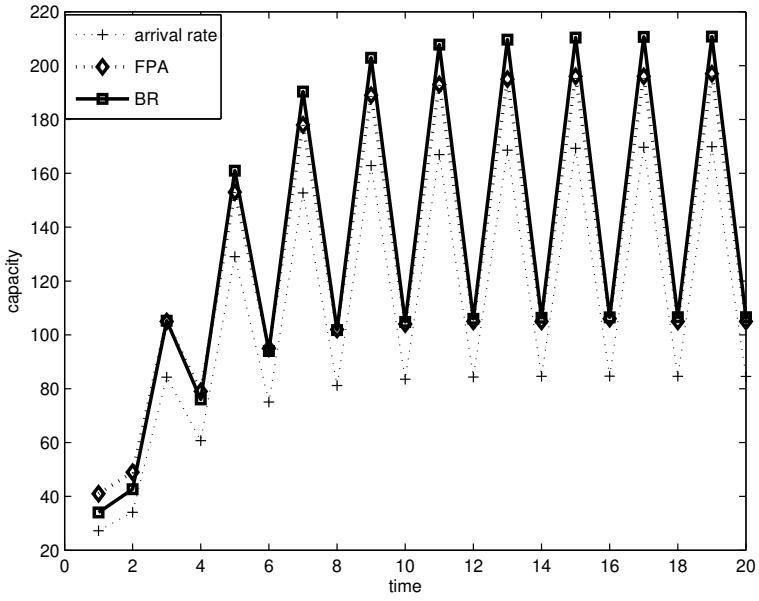

FIG. 6.2. *A dimensioning example of a single link with nonstationary arrival rates*

Rao approximation for dimensioning purposes for both stationary and nonstationary cases since it is the fastest and provides a tight upper bound.

The size of a medium and a large system is relative to the computer used. Our results were obtained on a Pentum(R)4 PC with a 3GHz CPU and a RAM of 504MB. In this context, a medium size system has less than two hundred servers, and a large size system has more than two hundred servers.

For future work we will develop efficient methods for the dynamic dimensioning of an entire network.

Acknowledgments. The authors would like to thank the three anonymous reviewers for their useful comments and suggestions.

REFERENCES

[1] A. Abdalla and R. J. Boucherie, *Blocking probabilities in mobile communications networks with time-varying rates and redialing subscribers* , Annals of Operations Research 112 (2002), pp. 15-34.

[2] K. Alnowibet, *Nonstationary Erlang Loss Queues and Networks*, Ph.D Dissertaion, North Carolina State University, 2004.

[3] K. Alnowibet and H. Perros, *The nostationary loss queue: A survey*, in: Modelling to Computer Systems and Networks, Ed: J. Barria, Imperial College Press, 2005.

[4] I. S. Dhillon, *A New $O(n^2)$ Algorithm for the Symmetric Tridiagonal Eigenvalue/Eigenvector Problem.*, Ph.D. thesis, Computer Science Division (EECS), University of California at Berkeley, 1997.

[5] R. A. Frazer et al. , *Elementary matrices*, Oxford, Cambridge Univ. Press, 1938, UK.

[6] D. L. Jagerman, *Nonstationary Blocking in Telephone Traffic*, The Bell System Technical Journal, 54 (1975), pp. 626-661.

[7] T. Karagiannis, M. Molle , M. Faloutsos and A. Broido, *A nonstationary Poisson view of Internet traffic*, In the Proceedings of INFOCOM 2004, Vol. 3, pp. 1558-1569.

[8] F. Kelly, *Markov Processes and Reversibility*,Wiley Chichester, 1979.
[9] M. Mandjes and Ad. Ridder, *A large deviations approach to the transient of the Erlang loss model*, Performance Evaluation, Vol. 43 (2001), pp. 181-198.
[10] W. A. Massey and W. Whitt, *An Analysis of the Modified Offered-load Approximation for the Nonstationary Loss Model*, Annals of Applied Probability, 4 (1994), pp. 1145-1160.
[11] W. A. Massey, *The Analysis of Queues with Time-Varying Rates for Telecommunication Models*, Telecommunication Systems, Vol. 21:2-4 (2002), pp. 173-204.
[12] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later.* , SIAM Rev., 45(1): pp. 3-49 (electronic), 2003.
[13] T. K. Nayak and K. N. Sivarajan, *Dimensioning Optical Networks Under Traffic Growth Models*, IEEE/ACM Transactions on Networking, Vol. 11, No. 6, December 2003.
[14] J. Riordan, *Stochastic Service Systems*, John Wiley & Sons, Inc., 1962.
[15] P. G. Taylor and R. Boucherie, *Transient product form distributions in queueing networks*, Discrete Event Dynamic Systems, Vol. 3 (1993), pp. 375-395.