

THE NONSTATIONARY LOSS QUEUE: A SURVEY

KHALID ABDULAZIZ ALNOWIBET

*Department of Statistics and Operations Research, King Saud University,
Riyadh 1145 , Kingdom of Saudi Arabia*

HARRY PERROS

*Computer Science Department, NC State University
Raleigh, NC 27695-7534, USA*

The nonstationary loss queue is of great interest since the arrival rate in most communication systems varies over time. In view of the difficulty in solving the nonstationary loss queue, various approximation methods have been developed. In this paper, we review several of these approximation methods and present a new technique, the *fixed point approximation* (FPA) method. Numerical evidence points to the fact that the FPA method gives the exact solution.

1. Introduction

The loss queue is a queueing 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 also been recently used to model traffic-groomed optical networks and optical burst switching (OBS) networks, see for instance Washington and Perros [1] and Battestilli and Perros [2], and references within.

The loss queue has been extensively studied in the stationary case, i.e., assuming that the arrival rate and the service rate are time invariant. The

nonstationary loss queue, where the arrival rate is time-dependent is also of great interest, since in most communication systems the arrival rate varies over time.

Consider a nonstationary loss queue $M(t)/M/s/s$ with a Poisson arrival process with a time-dependent rate $\lambda(t)$. Each arrival requests a service that requires an exponential amount of time with mean $1/\mu$. The service requested by an arriving customer is performed by a single server. The system has s identical servers, and there is no waiting room. The probability that there are n , $n = 0, 1, \dots, s$, customers in the system at time t , $P_n(t)$, is represented by the following set of forward differential equations:

$$\begin{aligned} \frac{d}{dt} P_0(t) &= \mu P_1(t) - \lambda(t) P_0(t), \\ \frac{d}{dt} P_n(t) &= \lambda(t) P_{n-1}(t) + (n+1)\mu P_{n+1}(t) - (\lambda(t) + n\mu) P_n(t), \quad 0 < n < s, \\ \frac{d}{dt} P_s(t) &= \lambda(t) P_{s-1}(t) - s\mu P_s(t), \end{aligned} \quad (1)$$

where $P_0(t) + P_1(t) + P_2(t) + \dots + P_s(t) = 1$, $t \geq 0$, and $0 \leq P_n(t) \leq 1$ for $t \geq 0$ and $n = 0, 1, 2, \dots, s$, with initial conditions: $P_0(0) = 1$ and $P_n(0) = 0$; $n = 1, 2, 3, \dots, s$.

In the stationary case, where the arrival rate is constant, that is $\lambda(t) = \lambda$ for all t , we have that $(d/dt)P_n(t) = 0$ for all n as $t \rightarrow \infty$. The above system of differential equations reduces to a set of linear equations from which we can obtain the familiar closed-form solution for the probability P_n that there are n customers in the system:

$$P_n(t) = \lim_{t \rightarrow \infty} P\{Q(t) = n\} = \frac{\rho^n / n!}{\sum_{i=0}^s \rho^i / i!}, \quad n = 0, 1, 2, \dots, s \quad (2)$$

where $\rho = \lambda/\mu$, and $Q(t)$ is the number of customers in the system at time t . The probability of blocking BP is:

$$BP = \lim_{t \rightarrow \infty} P\{Q(t) = s\} = \frac{\rho^s / s!}{\sum_{i=0}^s \rho^i / i!} \quad (3)$$

and the average number of customers in the system (*i.e.* the average number of busy servers) is:

$$\lim_{t \rightarrow \infty} E[Q(t)] = E[Q] = (1 - BP)\rho. \quad (4)$$

The expression for $P_n(t)$ is independent of time, for sufficiently large t , due to the stationary arrival process.

In the case of the nonstationary arrival rate, $(d/dt)P_n(t)$ will converge to some value which is not necessarily zero at all time. In fact, $(d/dt)P_n(t)$ will converge to some function of time depending on the structure of $\square(t)$. Therefore, in order to obtain the queue-length distribution, one must solve the set of differential equations Eq. 1. The solution to these differential equations is complex even for fairly small systems with special arrival rate functions $\lambda(t)$. For example, let us consider the simplest case where there is only one server in the system. Let $\lambda(t)$ and μ be the arrival rate function and service rate respectively. Then, the forward equations Eq. 1 become:

$$\frac{d}{dt} P_0(t) = \mu P_1(t) - \lambda(t) P_0(t)$$

and

$$\frac{d}{dt} P_1(t) = \lambda(t) P_0(t) - \mu P_1(t)$$

where $P_0(0) = 1$ and $P_0(t) + P_1(t) = 1$, $t \geq 0$. This system can be reduced to solving a single differential equation:

$$\frac{d}{dt} P_1(t) + [\lambda(t) + \mu] P_1(t) = \lambda(t), \text{ with } P_1(0) = 0.$$

which can be done as follows. Multiplying both sides by $e^{\int \lambda(\eta) + \mu d\eta}$ we have:

$$\frac{d}{dt} P_1(t) e^{\int \lambda(\eta) + \mu d\eta} + [\lambda(t) + \mu] e^{\int \lambda(\eta) + \mu d\eta} P_1(t) = \lambda(t) e^{\int \lambda(\eta) + \mu d\eta}$$

or

$$\frac{d}{dt} \left(P_1(t) e^{\int \lambda(\eta) + \mu d\eta} \right) = \lambda(t) e^{\int \lambda(\eta) + \mu d\eta}$$

$$P_1(t) e^{-\int \lambda(u) + \mu du} = \int \lambda(t) e^{\int \lambda(u) + \mu du} dt + K$$

Given that $P_1(0) = 0$, we finally have the blocking probability, $BP(t)$:

$$BP(t) = P_1(t) = \int_0^t \lambda(u) e^{\int_0^u \lambda(\eta) + \mu d\eta} du e^{-\int_0^t \lambda(t) + \mu dt}$$

In view of the difficulty in solving the nonstationary loss queue, various approximation methods have been developed. In this paper, we review several of these approximation methods and we also present a new technique, the *fixed-point approximation* (FPA) method, which yields the mean number of customers and the blocking probability functions in a nonstationary loss queue. Numerical evidence points to the fact that the FPA method gives the exact solution.

The paper is organized as follows. In sections 2 to 7 we describe the following approximation methods: the *simple stationary approximation* (SSA), the *stationary peakedness approximation* (PK), the *average stationary approximation* (ASA), the *closure approximation for nonstationary queues*, the *pointwise stationary approximation* (PSA), the *modified offered load approximation* (MOL). In section 8, we present the fixed point approximation (FPA) method, and finally the conclusions are given in section 9.

2. The simple stationary approximation (SSA) method

This method uses the average arrival rate of the nonstationary model to obtain the steady-state results. The average arrival rate for a cycle of length T is

$$\bar{\lambda} = \frac{1}{T} \int_0^T \lambda(t) dt . \quad (5)$$

Let $Q(t)$ be number of customers in the system at time t . Then, the steady-state distribution can be obtained using the expression:

$$P\{Q(t) = n\} = \frac{\rho^n / n!}{\sum_{i=0}^s \rho^i / i!} , \quad n = 0, 1, 2, \dots, s , \quad \text{where } \rho = \frac{\bar{\lambda}}{\mu}$$

and the blocking probability, $BP(t)$, at time t is the Erlang loss formula with parameter (s, ρ) , as follows:

$$BP(t) = P\{Q(t) = s\} = \frac{\rho^s / s!}{\sum_{i=0}^s \rho^i / i!}, \text{ for all } t.$$

The SSA method is simple and can be applied to a wide range of queueing systems. It provides a reasonable approximation for the nonstationary system with a weakly varying arrival rate. (An arrival rate is considered weakly varying over time if the arrival rate function remains within $\pm 10\%$ interval from the average arrival rate for all t .) However, this method noticeably underestimates the average performance measures of a nonstationary system with a strongly varying arrival rate. Green *et al.* [3] numerically investigated the level of nonstationarity at which this method provides a reasonable accuracy assuming a sinusoidal arrival rate function. The effect of nonstationarity with respect to amplitude, frequency of events and the size of the system (*i.e.* number of servers) was studied numerically. The authors showed that this method is applicable to relatively small systems (*e.g.* one or two servers) with small relative amplitude (*e.g.* less than 10%), and short cycle length (equivalently infrequent events).

Abdalla and Boucherie [4] used the SSA method to analyze a network of nonstationary loss queues. Consider a network of N independent loss queues, each with a time-dependent Poisson stream of external arrivals at rate $\lambda_i(t)$ and s_i servers, $i = 1, 2, \dots, N$. Upon service completion at queue i , a customer moves to node j with probability q_{ij} or it depart from the system with probability q_{i0} . Any external or internal arrival to queue i finds all servers busy is lost.

The arrival rate to each queue in the network is averaged over time using Eq. 5. Then the probability that the system is in state \mathbf{n} where $\mathbf{n} \in S$, $S = \{\mathbf{n} \in \mathbf{N}^N: 0 \leq n_i \leq s_i, i = 1, 2, \dots, N\}$ is:

$$P(\mathbf{n}; t) = \prod_{i=1}^N \frac{\rho_i^{n_i}}{n_i!} / \sum_{\mathbf{n} \in S} \prod_{j=1}^N \frac{\rho_j^{n_j}}{n_j!}.$$

The offered load of queue j , ρ_j , satisfies the solution of the following traffic equations:

$$0 = \bar{\lambda}_j + \sum_{i=1}^N \mu_i q_{ij} \rho_i - \mu_j \rho_j, \quad j = 1, 2, \dots, N.$$

It is worth mentioning that this method is an exact solution for loss networks with Markovian branching and stationary arrivals if and only the rates of queue j , $j=1, 2, \dots, N$, in the network satisfy the following conditions:

$$\lambda_j = q_{jj} \rho_j \text{ and } q_{ij} \rho_i = q_{ji} \rho_j, \quad j = 1, 2, \dots, N$$

The SSA underestimates the average performance measure of nonstationary systems even when the above two conditions are satisfied.

3. The stationary peakedness approximation (PK) method

The SSA method presented above does not consider the nonstationarity of the system. This can be done by using a non-Poisson stationary point process to approximate the time-dependent Poisson arrival process. Massey and Whitt [5] presented two approaches and used the heavy traffic peakedness to approximate the blocking probability of the nonstationary loss queue. (The peakedness is defined as the ratio of the variance to the mean of the steady-state number of customers in an infinite-server model with the same service time distribution and arrival process.)

To explain how this method works we consider a periodic Poisson arrival process with period T . The nonstationary arrival process is approximated by dividing the cycle T into n subintervals. It is assumed that the arrival rate at each subinterval is approximately constant. The arrival rate in any one subinterval is:

$$\lambda_k = \int_{(k-1)T/n}^{kT/n} \lambda(u) du, \quad 1 \leq k < n.$$

Then, the mean number of arrivals is:

$$\bar{\lambda}_n = \frac{1}{n} \sum_{k=1}^n \lambda_k = \frac{\bar{\lambda}T}{n}, \quad \text{where } \bar{\lambda} = \frac{1}{T} \int_0^T \lambda(u) du,$$

and its variance is:

$$\sigma_n^2 = \bar{\lambda}_n + \frac{1}{n} \sum_{k=1}^n (\lambda_k - \bar{\lambda}_n)^2 .$$

Based on the above analysis, the overall arrival process $M(t)$ in the interval $(0, T]$ is approximated by the stationary point process $\{N(t) : t \geq 0\}$ with mean and variance:

$$n\bar{\lambda}_n = \bar{\lambda}T \quad \text{and} \quad n\sigma_n^2 = \bar{\lambda}T + \sum_{k=1}^n (\lambda_k - \bar{\lambda}_n)^2 .$$

One may notice that the variance depends heavily on n . For example, for $n=1$, $n\sigma_n^2 = \bar{\lambda}T$; while $n\sigma_n^2 \rightarrow \bar{\lambda}T$ as $n \rightarrow \infty$. Therefore, n should be an intermediate point to capture the variability in arrival process. Next, the peakedness c^2 for number of customers in the infinite server system, $Q(t)$, is calculated :

$$c^2 = \frac{\text{Var}[N(T)]}{E[N(T)]} = 1 + \frac{1}{\bar{\lambda}T} \sum_{k=0}^n (\lambda_k - \bar{\lambda}_n)^2 .$$

Assuming that the arrival rate over each subinterval is constant, c^2 could be approximated as follows:

$$c^2 \approx 1 + \frac{1}{\bar{\lambda}T} \left(\frac{T}{n} \right) \int_0^T (\lambda(u) - \bar{\lambda})^2 du \approx 1 + \frac{1}{n\bar{\lambda}} \int_0^T (\lambda(u) - \bar{\lambda})^2 du .$$

It is always possible to rescale the problem to make the unit time equal to the mean service time. This means that $\mu = 1$. In this case, a good choice for n is to be equal to T . Thus,

$$c^2 \approx 1 + \frac{1}{\bar{\lambda}T} \int_0^T (\lambda(u) - \bar{\lambda})^2 du .$$

Then, c^2 is used to compute the heavy traffic peakedness of the nonstationary process. The heavy-traffic peakedness for an infinite-server system with exponential service distribution ($\mu = 1$) is:

$$z = 1 + \frac{c^2 - 1}{2} = 1 + \frac{1}{2\bar{\lambda}T} \int_0^T (\lambda(u) - \bar{\lambda})^2 du .$$

Finally, the approximate blocking probability for the $M(t)/M/s/0$ queue with time unit equals to the mean service time (*i.e.* $\mu = 1$) is given by the Erlang loss formula with updated number of servers s/z (s/z is integer) and updated offered load $\bar{\lambda}/z$ as follows:

$$BP(t) = P\{Q(t) = s\} = \frac{(\bar{\lambda}/z)^{s/z} / (s/z)!}{\sum_{i=0}^{s/z} (\bar{\lambda}/z)^i / i!} ,$$

The average number of customers is:

$$E[Q(t)] = (1 - BP(t)) \frac{\bar{\lambda}}{z}$$

This method is a stationary approximation of the original system. In other words, the PK method finds non-Poisson stationary parameters that better approximate the time-dependent arrival process. Therefore, the resulting approximation with the new parameters is a time reversible process. Although this approximation does not provide a solution for the system as a function time, it provides a better approximation than the SSA method for the average measure of performance of nonstationary Erlang loss models.

4. The average stationary approximation (ASA) method

This method was introduced by Whitt [6] for loss queues with periodic arrival rates. This approximation starts by dividing the arrival rate cycle T into sub-intervals each of length τ , where τ is proportional or equal to the mean service time. The arrival rate $\lambda(t)$ over subinterval $[t-\tau, t]$ is taken to be equal to the average arrival rate during $[t-\tau, t]$ as follows:

$$\bar{\lambda}_k(t) = \frac{1}{\alpha\mu^{-1}} \int_{t_k - \alpha\mu^{-1}}^{t_k} \lambda(u) du, \quad t \in [t_k - \alpha\mu^{-1}, t_k], \quad \alpha\mu^{-1} = \tau$$

The stationary results are used as a function of t and $\bar{\lambda}(t)$ to approximate the performance measures. Namely, the blocking probability, $BP(t,k)$, during sub-interval k is approximated as follows:

$$BP(t, k) = \frac{\rho_k^s / s!}{\sum_{i=0}^s \rho_k^s / i!}, \quad \rho_k = \frac{\bar{\lambda}_k(t)}{\mu}, \quad t \in [t_k - \alpha\mu^{-1}, t_k]$$

and the average number of customers, $E[Q(t,k)]$, during sub-interval k :

$$E[Q(t,k)] = (1 - BP(t, k)) \frac{\bar{\lambda}(t)}{\mu}, \quad \text{for } t \in [t_k - \alpha\mu^{-1}, t_k].$$

Obviously, the performance measures are going to be step functions due to the discretization of the arrival process. This method is simple to apply and it produces an insight into the behavior of the performance measures over time. In addition, this method provides an exact solution for the $M(t)/D/\infty$ queue when $\alpha = 1$. This method depends mainly on the choice of the subinterval length (τ) which is strongly related to the choice of α . If α is chosen to be small when it should not be, the approximation will pick up more variability from the arrival process than needed. In contrast, if α is chosen to be large then this method will approach the stationary approximation which kills the variability of the performance measures over time.

5. The closure approximation for nonstationary queues

This method reduces the number of differential equations needed to be solved by considering the differential equations of the mean and the variance of the number of customers in the system. In many systems, the equations for the mean and the variance involve more variables than the number of equations. Consequently, additional equations are required in order to obtain a unique solution.

Consider an $M(t)/M/1$ queue with arrival rate $\lambda(t)$ and service rate μ . The probability $Pn(t)$ of having n in the system at time t is given by the following set of differential equations:

$$\begin{aligned}\frac{d}{dt} P_0(t) &= -\lambda(t)P_0(t) + \mu P_1(t) \\ \frac{d}{dt} P_n(t) &= -(\lambda(t) + \mu)P_n(t) + \lambda(t)P_{n-1}(t) + \mu P_{n+1}(t), \quad n > 0\end{aligned}\quad (6)$$

Multiplying Eq. 6 by n and summing over all n gives:

$$\frac{d}{dt} E[n] = \sum_{n=0}^{\infty} n \frac{d}{dt} P_n(t) = \lambda(t) - \mu(1 - P_0(t)) \quad (7)$$

Multiplying Eq. 6 by n^2 and summing over all n gives:

$$\frac{d}{dt} E[n^2] = \sum_{n=0}^{\infty} n^2 \frac{d}{dt} P_n(t) = \lambda(t) - \mu(1 - P_0(t)) + 2 E[n](\lambda(t) - \mu)$$

Hence, the variance is as follows:

$$\frac{d}{dt} Var[n] = \frac{d}{dt} E[n^2] - \frac{d}{dt} E[n]^2 = \lambda(t) + \mu P_0(t)(2 E[n] + 1) \quad (8)$$

Equations Eq. 7 and Eq. 8 provide a system of two differential equations in three unknowns ($Var[n]$, $E[n]$ and $P_0(t)$). To obtain a unique solution using these equations an additional equation of $Var[n]$, $E[n]$ and $P_0(t)$ is required to bound the solution.

Rothkopf and Oren [7] consider the negative binomial distribution to provide a closure function for the M(t)/M(t)/s system. The negative binomial with probability of success q and parameters n and r has the form:

$$p_n(q, r) = \binom{r+n-1}{n} q^r (1-q)^n, \quad n = 0, 1, 2, \dots$$

with mean $r(1-q)^{-1}$ and variance is $r(1-q)q^{-2}$.

The negative binomial reduces to the geometric distribution if its parameters (q and r) are chosen such that:

$$Var[n] = E[n] (1 + E[n]). \quad (9)$$

The number of customers in an M/M/1 system has a geometric distribution. Therefore, the parameters q and r can be chosen as functions of the mean and variance of the system so that the resulting negative binomial satisfies property (9). The new negative binomial distribution will have the following parameters:

$$q(t) = \frac{E[n]}{Var[n]} \quad \text{and} \quad r(t) = \frac{E[n]^2}{Var[n] - E[n]},$$

where $Var[n]$ and $E[n]$ are functions of time. The closure function $P_0(t)$ of equations Eq. 6 is obtained by setting n equals to zero in the negative binomial with the parameters $q(t)$ and $r(t)$:

$$P_0(t) = p_0(q(t), r(t)) = q(t)^{r(t)}.$$

Similarly, the mean and variance of M(t)/ M/ s system are:

$$\frac{d}{dt} E[n] = \lambda(t) - \mu s + \mu \sum_{n=0}^{s-1} (s-n) P_n(t) \quad (10)$$

and

$$\frac{d}{dt} Var[n] = \lambda(t) + \mu s - \mu \sum_{n=0}^{s-1} (2E_t[n] + 1 - 2n)(s-n) P_n(t). \quad (11)$$

The closure functions ($P_n(t)$ for $n = 0, 1, 2, \dots, s-1$) of Eqs. 10 and 11 are obtained by evaluating the negative binomial distribution described earlier at $n = 0, 1, 2, \dots, s-1$.

This method provides an exact solution for the stationary M/M/1 queue and a very good approximation for the M(t)/M/1 queue due to the fact that the stationary system has a geometric steady-state solution. However, for the M(t)/M/s queue the error of the approximation increases very quickly as the number of servers increases. Rothkopf and Oren [7] provided an error correction term to improve the accuracy of the approximation for the M(t)/M/s/s queue.

6. The pointwise stationary approximation (PSA) method

This method is based on the idea that the nonstationary loss queue approximately behaves like a stationary model at each instance of time. Thus, the steady-state results of the stationary loss queue can be used to approximate the nonstationary loss queue at each point on time. This method was first introduced by Grassman [8] in 1983 as a way of constructing an upper bound on the expected number of customers in the queue. Green *et al.* [9] showed numerically that PSA gives an upper bound on the expected number of customers in the system and probability of delay, if the maximum traffic intensity is strictly less than one. In addition, Green and Kolesar [10] used the PSA method to approximate the steady-state average performance measures of the periodic M(t)/M/s/s queue.

Consider a stationary loss queue with arrival rate λ and service rate μ . Let $Q(t)$ be number of customers in the system at time t . Then, the probability P_n that there are n customers in the system is:

$$P_n = \lim_{t \rightarrow \infty} P\{Q(t) = n\} = \frac{\rho^n / n!}{\sum_{i=0}^s \rho^i / i!},$$

$$\rho = \frac{\lambda}{\mu}, \quad n = 0, 1, 2, \dots, s.$$

The probability of blocking BP is:

$$BP = \lim_{t \rightarrow \infty} P\{Q(t) = s\} = \frac{\rho^s / s!}{\sum_{i=0}^s \rho^i / i!},$$

and the average number in the system (*i.e.* average number of busy servers) is:

$$\lim_{t \rightarrow \infty} E[Q(t)] = E[Q] = (1 - BP) \rho.$$

In the PSA method, the time dependent-steady state distribution of the nonstationary Erlang loss system, given that the arrival rate is $\lambda(t)$ and service rate is μ , is calculated as follows:

$$P_n(t) = \frac{\rho(t)^n / n!}{\sum_{i=0}^s \rho(t)^i / i!}, \quad \rho(t) = \frac{\lambda(t)}{\mu} \quad \text{and} \quad n = 0, 1, 2, \dots, s.$$

the time-dependent steady-state blocking probability is

$$BP(t) = P_s(t) = \frac{\rho(t)^s / s!}{\sum_{i=0}^s \rho(t)^i / i!},$$

and the time-dependent steady-state average number in the system is:

$$E[Q(t)] = (1 - BP(t)) \rho(t).$$

where $\rho(t) = \lambda(t) / \mu$.

The PSA method can be easily generalized to most of the queueing systems, as long as $\rho < 1$ for all t is required for the stability of the equivalent stationary system.

An important factor that affects the accuracy of the PSA is the arrival rate function. The PSA method will provide a good approximation as the arrival rate increases. For example, consider two nonstationary loss queues with sinusoidal arrival rate function $\lambda(t) = \bar{\lambda} + \beta \sin(\gamma T)$ where $\bar{\lambda}$ is the average arrival rate, β is the amplitude and γ is the frequency set equal to $2\pi/T$, T being the cycle length. Let $\{\lambda(t) = 5 + 2.5 \sin(t), \mu = 0.5, s = 10\}$ and $\{\lambda(t) = 20 + 10 \sin(t), \mu = 2, s = 10\}$ be the parameters of loss queue 1 and 2 respectively. Then according to PSA, the time-dependent offered load for both systems is

$$\rho(t) = \frac{\lambda(t)}{\mu} = 10 + 5 \sin(t).$$

Although both systems have the same offered load, PSA will provide a better approximation for loss queue 2, since it needs shorter time to reach the steady state due to the higher arrival and service rates. This means that loss queue 2 will behave more like a stationary system within reasonably small interval of time than queue 1. In figure 1, we plot the exact time-dependent average number $E[Q(t)]$ for loss queue 1, 2 (labelled "Queue 1" and "Queue 2" respectively in figure 1) and the PSA values as a function of time t . As expected, PSA provides better approximation for loss queue 2 than for loss queue 1.

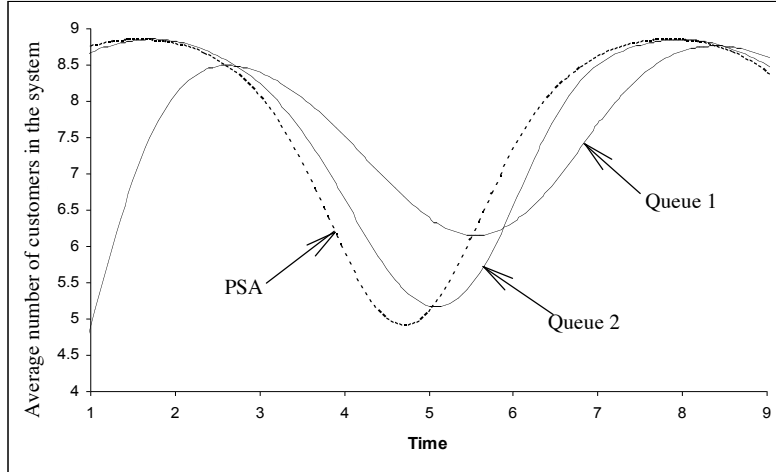


Figure 1: PSA and exact values of the average number of customers in loss queues 1 and 2

As can be seen in figure 1, the PSA method overestimates the peak of the average number of customers. In addition, the PSA peak lags the peak of the average number of customers. (The same also applies to other performance measures.) These two problems become negligible as the arrival and service rates increase. Whitt [11] showed that the PSA solution for an $M(t)/M(t)/s$ queue is asymptotically correct as the rates increase.

Green and Kolesar [12] proposed the *Simple Peak Hour Approximation* (SPHA) technique for the computation of average performance measures of periodic systems during the peak period. SPHA starts by obtaining the measure of interest, say $X(t)$, using the PSA method. Next, the peak time t^* at which the $X(t)$ achieves its maximum is determined. The average of $X(t)$ over the interval $[a, b]$ where t^* is the center of the interval is the SPHA value for $X(t)$.

SSA and PSA can be seen as two extreme cases of averaging out the arrival rate. The SSA method averages the arrival rate $\lambda(t)$ over a period of time equal to the cycle length T , whereas the PSA method uses the average arrival rate over an infinitesimally small interval.

Finally, the PSA method has also been used to analyze a network nonstationary loss queues, see Massey and Whitt [14] and Abdalla and Boucherie [4]. The same system of traffic equations is used as in the SSA method only the arrival rates are

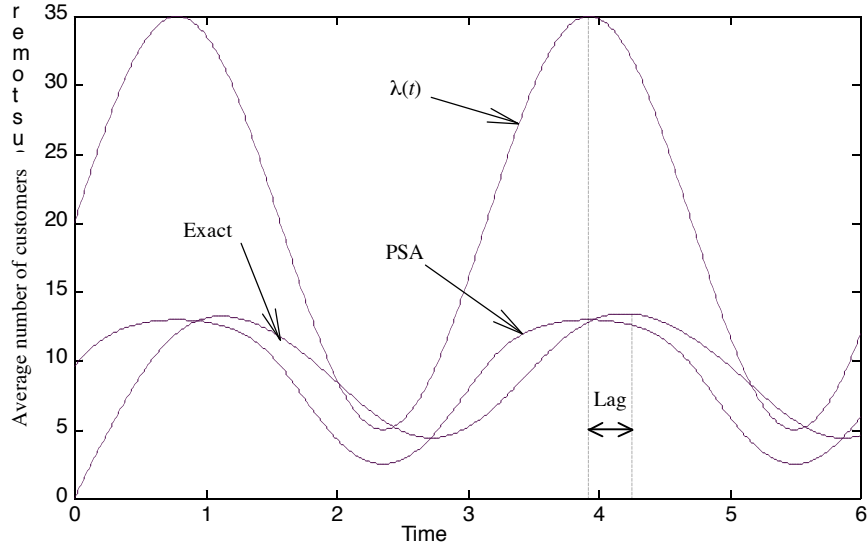


Figure 2: Exact and PSA values of the average number of customers

taken to be functions of time instead of averages. That is, the time-dependent offered loads $\rho_j(t)$ are obtained by solving the following system of traffic equations in time t :

$$0 = \lambda_j(t) + \sum_{i=1}^N \mu_i q_{ij} \rho_i(t) - \mu_j \rho_j(t); \quad j = 1, 2, \dots, N \quad (12)$$

Then, the probability, $P(\mathbf{n}; t)$, of having \mathbf{n} in the system, for $\mathbf{n} \in S$, $S = \{\mathbf{n} \in \mathbf{N}^N: 0 \leq n_i \leq s_i; i = 1, 2, \dots, N\}$, is

$$P(\mathbf{n}; t) = \frac{\prod_{i=1}^N \frac{\rho_i(t)^{n_i}}{n_i!}}{\sum_{\mathbf{n} \in S} \prod_{j=1}^N \frac{\rho_j(t)^{n_j}}{n_j!}} \quad (13)$$

7. The modified offered load approximation (MOL) method

Let us first consider the stationary loss queue M/M/s/s and the stationary infinite server queue M/M/∞. Let $Q_{\infty}(t)$ be the number of customers in the infinite server queue at time t . The probability P_n that there are n customers in an M/M/∞ with an arrival rate λ and a service rate μ is:

$$P_n = \lim_{t \rightarrow \infty} P\{Q_{\infty}(t) = n\} = \frac{\rho^n}{n!} e^{-\rho},$$

where $\rho = \lambda/\mu$. Likewise, let $Q(t)$ be number of customers in the loss queue with s servers at time t . The probability P_n that there are n customers in an M/M/s/0 with an arrival rate λ and a service rate μ , is:

$$P_n = \lim_{t \rightarrow \infty} P\{Q(t) = n\} = \frac{\rho^n/n!}{\sum_{i=0}^s \rho^i/i!},$$

where $\rho = \lambda/\mu$ and $n = 0, 1, 2, \dots, s$

Another way to obtain the stationary distribution of the M/M/s/0 queue is to use the fact that the M/M/s/0 queue is a truncated process of an M/M/∞ queue which is a reversible Markov process. Since the arrival rates are time invariant we dropped the time variable from the random variables. Then, we have:

$$P\{Q = n\} = P\{Q_{\infty} = n \mid Q_{\infty} < s\} = \frac{e^{-\rho} (\rho^n/n!)}{e^{-\rho} \sum_{i=0}^s \rho^i/i!} = \frac{\rho^n/n!}{\sum_{i=0}^s \rho^i/i!}$$

In the M(t)/M/∞ queue, the rate of change in the average number of customers at time t is equal to the difference between the arrival rate and the departure rate due to the Markovian property. That is,

$$\frac{d}{dt} E[Q_{\infty}(t)] = \lambda(t) - \mu E[Q_{\infty}(t)].$$

Recall that there is always an idle server for each arriving customer to the M(t)/M/∞ queue. This means that no customers are lost and all customers in the system at time t are being served. Therefore, the average number of customers at time t is equal to

the average number of customers in the system at time t which equals to the offered load $\rho(t)$. We have the following differential equation for $\rho(t)$:

$$\frac{d}{dt} \rho(t) = \lambda(t) - \mu \rho(t) \quad (14)$$

Analogous to the stationary queues, one can approximate the $M(t)/M/s/0$ by truncating the $M(t)/M/\infty$ queue. This method is called the *modified offered load method* (MOL). The MOL approximation was first developed by Jagerman [15] in 1975. The probability $P_n(t)$ that there are n customers in the system using MOL is:

$$P_n(t) \approx \text{P}\{Q_\infty(t) = n \mid Q_\infty(t) < s\} = \frac{\rho(t)^n / n!}{\sum_{i=0}^s \rho(t)^i / i!}, \quad n = 0, 1, 2, \dots, s.$$

where $\rho(t)$ is determined from Eq. 14.

The truncated $M/M/\infty$ queue provides an exact solution to the $M/M/s/0$ queue due to the reversibility property. In the case of nonstationary arrival process, the reversibility property is lost and hence the truncated $M(t)/M/\infty$ will not provide an exact solution to the $M(t)/M/s/0$. Massy and Whitt [16] developed analytical bounds on the error between the MOL approximation and the exact solution of the $M(t)/M/s/0$ system.

The MOL method can be seen as averaging out the arrival rate over an interval that depends on the mean and the distribution of the service time. This is in contrast to the SSA method where the arrival rate $\lambda(t)$ is averaged over the cycle length T , and the PSA method where the arrival rate is averaged over an infinitesimally small interval.

The $M(t)/M/s/s$ behaves like $M(t)/M/\infty$ as the blocking probability gets smaller. In view of this, the MOL method provides a good approximation for the $M(t)/M/s/s$ as long as the system has a small blocking probability. 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. As expected, the MOL underestimates the blocking probability of a loss queue with a high load, *i.e.* when the exact blocking probability is high.

The MOL method provides a good estimation for the peak time for a loss queue with a small blocking probability. This is due to the fact that the MOL method is sensitive to the service process through its mean and distribution. The

PSA method depends on the service distribution only through its mean. As a result, PSA appears to lag the actual performance measures values of the system. Figure 3 shows the exact, PSA, and MOL values of the average number of customers in an $M(t)/M/s/s$ queue with $\lambda(t)=20+15 \sin(2t)$, $\mu = 2$ and $s = 15$.

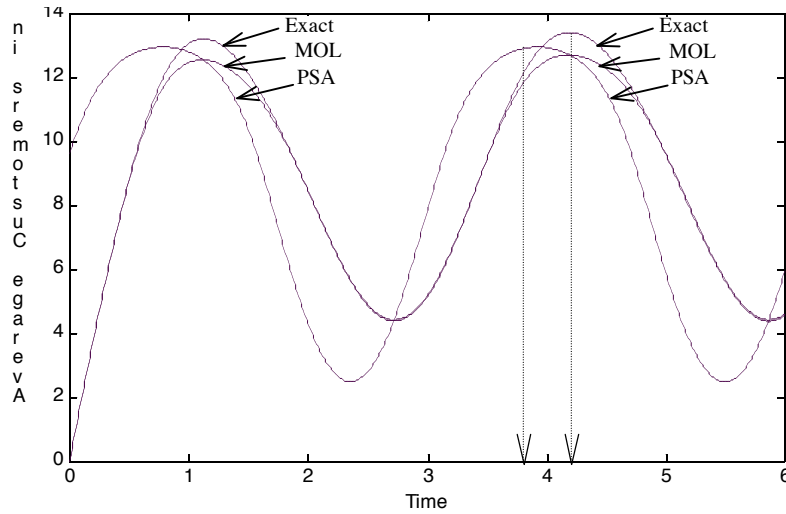


Figure 3: Exact, PSA and MOL values of the average number of customers

The MOL method has also been used to analyze networks of nonstationary loss queues. Whitt [17] described how the MOL method can be used in a decomposition algorithm for the analysis of a nonstationary loss network. Jennings and Massey [18] used the MOL method to analyze time-dependent circuit-switched networks. Abdalla and Boucherie [4] applied the MOL method to a mobile communication network with time-varying arrival rates and redialing. The authors established an exact expression for the error in the MOL approximation as well as bounds on the error.

The MOL method is used to analyze a network of nonstationary loss queues as follows. Consider a network consisting of N independent loss queues, each with a time-dependent Poisson stream of external arrivals at rate $\lambda_i(t)$ and s_i servers, $i = 1, 2, \dots, N$. Upon service completion at queue i , a customer moves to queue j with probability q_{ij} or it depart from the system with probability q_{i0} . Any external or internal arrival to queue i finds all servers busy is lost. The traffic equations of the equivalent network consisting of nonstationary infinite server queues are first solved in order to obtain the time-dependent offered loads $\rho_i(t)$, $i= 1, 2, \dots, N$. This system of

traffic equations is the same as Eqs. 12 used in the PSA method except that the $(d/dt)\rho_i(t)$, $i=1,2,\dots,N$, are not taken to be zero. That is, the time dependent offered loads $\rho_j(t)$ are obtained by solving the following system of differential equations in time t :

$$\frac{d}{dt}\rho_j(t) = \lambda_j(t) + \sum_{i=1}^N \mu_i q_{ij} \rho_i(t) - \mu_j \rho_j(t); \quad j=1,2,\dots,N$$

Then, for $\mathbf{n} \in \{\mathbf{n} \in \mathbf{N}^N: 0 \leq n_i \leq s_i; i=1,2,\dots,N\}$

$$P(\mathbf{n}; t) = \prod_{i=1}^N \frac{\rho_i(t)^{n_i}}{n_i!} / \sum_{\mathbf{n} \in \mathcal{S}} \prod_{j=1}^N \frac{\rho_j(t)^{n_j}}{n_j!}$$

It is worth mentioning that the $M_t/M/s/0$ behaves like $M_t/M/\infty$ as the blocking probability gets smaller. Then, it is expected that the MOL method will provide a good approximation for the nodes in the system that has a small blocking probability.

8. The fixed point approximation (FPA) method

The *fixed point approximation* (FPA) method was proposed by Alnowibet and Perros [19]. This method calculates numerically the time-dependent mean number of customers and blocking probability functions in a nonstationary multi-rate loss queue. Experimental results showed that the FPA algorithm provides an exact solution. The FPA method has also been extended to nonstationary queueing networks of multi-rate loss queues, see Alnowibet and Perros [19], and nonstationary queueing networks with population constraints, see Alnowibet and Perros [20]. In this paper, we describe the FPA method for the analysis of the nonstationary (single-class) loss queue. Consider a loss queue $M(t)/M/s/s$ with a Poisson arrival process with time-dependent rate $\lambda(t)$. The time-dependent average number of customers $E[Q(t)]$ in an $M(t)/M/s/s$ queue can be expressed as the difference between the effective arrival rate and the departure rate at time t . We have:

$$\frac{d}{dt} E[Q(t)] = \lambda(t) (1 - BP(t)) - [\mu P\{Q(t)=1\} + 2\mu P\{Q(t)=2\} + \dots + s\mu P\{Q(t)=s\}]$$

where $BP(t)$ is the blocking probability at time t . The above equation can be written as follows:

$$\frac{d}{dt} E[Q(t)] = \lambda(t) (1-BP(t)) - \mu E[Q(t)] \quad (15)$$

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

$$\rho(t) = \frac{E[Q(t)]}{1-BP(t)}. \quad (16)$$

Using expressions Eq. 15, 16, and 17, we can calculate the blocking probability iteratively. The steps of the algorithm are 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 $t = 0, \Delta t, 2\Delta t, \dots, T_f$.
4. Start with an initial blocking probability $BP^0(t) = 0, t = 0, \Delta t, 2\Delta t, \dots, T_f$.
5. Set the iteration counter $k = 0$.
6. Solve numerically for $E[Q^k(t)]$ using the following equation:

$$E[Q^k(t+\Delta t)] = E[Q^k(t)] + \lambda(t) (1-BP^k(t))\Delta t - \mu E[Q^k(t)]\Delta t.$$
7. Calculate
$$\rho^k(t) = \frac{E[Q^k(t)]}{1-BP^k(t)}, \quad t = 0, \Delta t, 2\Delta t, \dots, T_f.$$
8. Update blocking probability
$$BP^{k+1}(t) = \frac{[\rho^k(t)]^s / s!}{\sum_{i=0}^s [\rho^k(t)]^i / i!}, \quad t = 0, \Delta t, \dots, T_f$$
9. If $\|BP^k(t) - BP^{k+1}(t)\| < \varepsilon$, then $BP^k(t)$ has converged and the algorithm stops. Else, set $k = k + 1$ and go to step 6.

The algorithm does not require a closed-form expression for the arrival rate function. It only needs that the arrival rate function be defined at time points equally spaced by Δt . In view of this, any periodic arrival rate function can be used irrespective of 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.

We also note that the algorithm can be easily extended to the case where the service rate is also time-dependent by simply defining the service rate as a vector corresponding to the same time points used for the arrival rate function.

In all the experiments the FPA results were very close to the exact numerical results or within the simulation confidence intervals. This lead us to the conjecture that Eq. 17 for the nonstationary blocking probability used in the FPA method is in fact correct. However, due to the discretization process, the FPA and the exact numerical results never matched, which prevented from establishing beyond doubt the correctness of Eq. 17. As an example, let us consider a loss queue with a sinusoidal arrival rate function $\lambda(t) = \bar{\lambda} + \beta \sin(\gamma t)$, where $\bar{\lambda} = 20$, $\beta = 15$ and $\gamma = 2$, $s = 10$, and the service rate $\mu = 1$. The FPA method was applied with tolerance $\varepsilon = 0.01$ for different values of Δt . The average absolute error between the exact and the FPA solutions for the blocking probability and the average number of customers are given in table 1. As can be seen the absolute error decreases as Δt decreases. Due to the CPU and memory limitations, it was not possible to consider Δt values less than 0.0001.

Table 1: Average absolute error of FPA as $\Delta t \rightarrow 0$

	$\Delta t =$ 0.1	$\Delta t =$ 0.05	$\Delta t =$ 0.01	$\Delta t =$ 0.005	$\Delta t =$ 0.001	$\Delta t =$ 0.0005	$\Delta t =$ 0.0001
$BP(t)$	0.269	0.0165	0.0123	0.0118	0.0115	0.0114	0.0114
$E[Q(t)]$	0.0992	0.692	0.0507	0.0474	0.0447	0.0445	0.0445

9. Conclusions

The loss queue has been extensively studied in the stationary case, i.e., assuming that the arrival rate and the service rate are time invariant. The nonstationary loss queue, where the arrival rate is time-dependent is also of interest, since the arrival rate in most communication systems varies over time. In view of the difficulty in solving the nonstationary loss queue, various approximation methods have been developed. In this paper, we reviewed the following approximation methods: the simple stationary approximation (SSA), the stationary peakedness approximation (PK), the average stationary approximation (ASA), the closure approximation for nonstationary queues, the pointwise stationary approximation (PSA), and the modified offered load approximation (MOL). We also presented a new technique, referred to as the *fixed-point approximation* (FPA) method, which yields the mean

number of customers and the blocking probability functions in a nonstationary loss queues. Numerical evidence points to the fact that the FPA method gives the exact solution.

References

1. Washington, A.N. and Perros, H. G., " Call blocking probabilities in a traffic groomed tandem optical network", Special issue dedicated to the memory of Professor Olga Casals, Blondia and Stavrakakis (Eds.) Journal of Computer Networks, Vol 45 (2004).
2. Battestilli, L. and Perros, H. G., "End-to-End Burst Probabilities in an OBS Network with Simultaneous Link Possession" Workshop on OBS, BroadNets 2004. (<http://www.csc.ncsu.edu/faculty/perros/recentpapers.html>).
3. Green, L.V., Kolesar, P. J. and Svoronos, A., "Some Effects of Nonstationarity On Multiserver Markovian Queues Systems", Operations Research, 39(1991), 502–511.
4. Abdalla, N. and Boucherie R. J., "Blocking Probabilities in Mobile Communications Networks with Time-Varying Rates and Redialing Subscribers", Department of Applied Mathematics Internal Reports, University of Twente, The Netherlands, 1596 (2001).
5. Massey, W. A. and Whitt W., "Stationary-Process Approximation for the Nonstationary Erlang Loss Model", Operations Research, 44(1996), 976–983.
6. Whitt W., "The Pointwise Stationary Approximation for $M_t/M_t/s$ Queues is Asymptotically Correct as the Rates Increase", Management Sciences, 37(1991), 307–314.
7. Rothkopf, M. H. and Oren, S. S., " A Closure Approximation for the Nonstationary $M/M/s$ Queue", Management Sciences, 25(1979), 522–534.
8. Grassmann, W., "The 'Convexity of the Mean Queue Size of the $M/M/c$ Queue with Respect to the Traffic Intensity", Journal of Applied Probability, 20(1983), 916-919.
9. Green, L.V., Kolesar, P. J. and Svoronos, A., "Some Effects of Nonstationarity On Multiserver Markovian Queues Systems", Operations Research, 39(1991), 502–511.
10. Green, L.V. and Kolesar, P. J., "The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals", Management Sciences, 37(1991), 84–97

11. Whitt W., "The Pointwise Stationary Approximation for $Mt/Mt/s$ Queues is Asymptotically Correct as the Rates Increase", *Management Sciences*, 37(1991), 307–314.
12. Green, L.V. and Kolesar, P. J., "The Lagged PSA for Estimating Peak Congestion in Multiserver Markovian Queues with Periodic Arrival Rates", *Management Sciences*, 43(1997), 80–87.
13. Green, L.V. and Kolesar, P. J., "On the Accuracy of the Simple Peak Hour Approximation for Markovian Queues", *Management Sciences*, 41(1995), 1353–1370.
14. Massey, W. A. and Whitt W., "Networks of Infinite-Server Queues with Nonstationary Poisson Input", *Queueing Systems*, 13(1993), 183–251.
15. Jagerman, D. L., "Nonstationary Blocking in Telephone Traffic", *The Bell System Technical Journal*, 54(1975), 626–661.
16. Massey, W. A. and Whitt W., "An Analysis of the Modified Offered-load Approximation for the Nonstationary Loss Model", *Annals of Applied Probability*, 4(1994), 1145–1160.
17. Whitt, W., "Decomposition Approximation for Time-Dependent Markovian Queueing Networks", *Operations Research Letters*, 24(1999), 97–103.
18. Jennings, O. B. and Massey, W. A., "A Modified Offered Load Approximation for Nonstationary Circuit Switched Networks", *Telecommunication Systems*, 7(1997), 229–251.
19. Alnowibet, K. and Perros, H.G., "Nonstationary Loss Queues and Loss Queueing Networks", (<http://www.csc.ncsu.edu/faculty/perros/recentpapers.html>)
20. Alnowibet, K. and Perros, H., "Nonstationary Analysis of Circuit-Switched Communication networks" (<http://www.csc.ncsu.edu/faculty/perros>)