On the Analysis of Exponential Queuing Systems with Randomly Changing Arrival Rates: Stability Conditions and Finite Buffer Scheme with a Resume Level *

Catherine Rosenberg

Département de Génie Electrique, Ecole Polytechnique, Case Postale 6079, Succ. A, Montréal, P.Q., Canada H3C 3A7

Ravi Mazumdar

INRS-Telecommunications, Université du Quebec, 3 Place du Commerce, Ile des Soeurs, P.Q., Canada H3E 1H6

Leonard Kleinrock

Department of Computer Science, University of California, Los Angeles, CA 90024-1596, USA

Received October 1988
Revised March 1990

This paper considers a single server exponential queue with random fluctuations in the intensity of the arrival process. The motivation being the modelling of random changes in traffic patterns. This random intensity model does not obey the independence assumption made in queuing theory. Necessary and sufficient conditions for the stability or ergodicity of the queuing process are obtained via analytic techniques using Jury's stability criteria, often used to study discrete time control systems. The effect of such fluctuations is then studied for a finite resume level queue which is often used in flow control. Exact performance measures are computed and are compared with existing results.

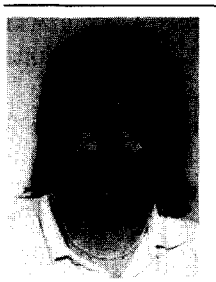
Keywords: Queues, Random Intensity, Independence, Jury's Stability Criteria, Resume Level Queues.

* The research was supported in part by a grant from NSERC No. OGP 0042023.

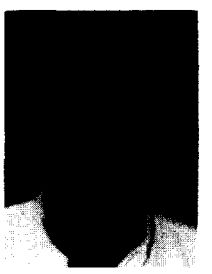
North-Holland
Performance Evaluation 11 (1990) 283–292

1. Introduction

Realistic considerations of certain stochastic service systems often warrant the introduction of *random* changes in the intensity of the input process. A typical example of a changing arrival rate system is a communication network with a sudden unpredictable increase in the traffic due to an external phenomenon. Such a situation can occur in a telephone network where a switching exchange node has crashed which results in an in-



Catherine Rosenberg was born in Cholet, France. She obtained the Diplome d'Ingénieur from ENST-Bretagne, Brest, France in 1983, the M.S. in Computer Science from the University of California, Los Angeles, USA in 1984 and the Doctorat en Sciences from Université de Paris XI, Orsay, France in 1986. From 1984–1986 she was an engineer with ALCATEL, Lannion, France. During the Fall of 1986 she was a Maitre de Conference Associée at Université de Paris, Orsay. From 1987–1988 she was a Member of Technical Staff at AT&T Bell Labs., Holmdel, NJ. Since July 1988 she has been an Assistant Professor in the Department of Electrical Engineering, Ecole Polytechnique, Montréal and is an Invited Professor at INRS-Telecommunications. Her research interests are queuing systems, modelling and performance evaluation of broadband integrated telecommunication networks.



Ravi Mazumdar was born in Bangalore, India. He obtained the B.Tech. in Electrical Engineering from the Indian Institute of Technology, Bombay, India in 1977, the M.Sc. DIC in Control Systems from Imperial College, London, UK in 1978 and the Ph.D. in Systems Science from the University of California, Los Angeles, USA in 1983. In 1978–1979 he was employed with GEC Electrical Projects Ltd., Rugby, UK. From April 1983–Oct. 1983 he was a Member of Technical Staff, AT&T Bell Laboratories, Holmdel, NJ. He held visiting appointments at UCLA (1983–1984) and Twente University, Enschede, Netherlands (Fall 1984). From 1985–1988 he was an Assistant Professor in the Department of Electrical Engineering, Columbia University, New York. In July 1988 he joined INRS-Telecommunications, Montreal, Canada where he is an Associate Professor. His research interests are in the modelling and analysis of stochastic systems.

crease in the traffic at the other nodes toward which the calls may be rerouted. One could then adapt or control other parameters of the system to respond to the random changes to maintain desired system performance.

Random intensity models were most comprehensively studied by Yechiali and Naor [14] and Neuts [8]. Their approach was to model the random fluctuations by an intensity rate indexed by a finite state Markov process extraneous to the queuing processes evolving independently of them. Conditioned on the sample path of the Markov process, the queuing process could be analysed in terms of well-known paradigms. In the monograph by Neuts [9] the model is discussed in the framework of phase type distributions and analysed via matrix-geometric techniques. The model proposed in this paper falls under the general class of models discussed above but differs in the assumption of independence of the sojourn time in a particular state and the arrival and service times. In particular, the sojourn time is assumed to be correlated with the arrival times differing from the usual Markov modulated arrival processes. A related problem with slow variations in arrival rates is discussed in [3].

We first study the infinite buffer case. Using purely analytic methods and an interesting application of Jury's stability criteria [4] we obtain necessary and sufficient conditions for the stability of the queue or the existence of an invariant measure for the Markov process. In particular, we obtain similar conditions as in [1] which were found via different techniques. We then study the effect of such random fluctuations in the performance of a finite buffer with a resume level since such schemes are often used for congestion control. The random fluctuations are useful to model situations with competing multiclass traffic and breakdowns. For a discussion of resume level queues in the M/M/1/K case see [10]. An extension of the results to the M/G/1/K case was done by Takagi [13]. Thus, the results can be considered as an extension to a particular case of a G/G/1/K queue.

The paper is organized into 5 sections. In Section 2, the model is formulated. In Section 3, the case of infinite buffer is analysed and the stability conditions are obtained using Jury's criteria. In Section 4, the finite buffer scheme is analysed and several performance measures are calculated and compared with the resume level queue in the M/M/1 case.

Leonard Kleinrock is Professor of Computer Science at U.C.L.A. He received his Ph.D. from the M.I.T. His research interests focus on performance evaluation of high speed networks and parallel and distributed systems. He has had over 160 papers published and is the author of five books. He is a member of the National Academy of Engineering, is a Guggenheim Fellow, an IEEE Fellow, and a member of the Computer Science and Technology Board of the National Research Council. He has received numerous best paper and teaching awards, including the ICC Prize Winning Paper Award, the Lanchester Prize, the Communications Society Prize Paper Award, the C.C.N.Y. Townsend Harris Medal, the L. M. Ericsson Prize, and the Marconi International Fellowship Award.

2. Model formulation

As mentioned in the introduction the motivation for the model is to introduce the effect of random fluctuations in the intensity of the arrival process due to the effect of random phenomena or interfering classes of traffic.

In this paper the analysis is restricted to the case where the arrival rate takes two possible values, which may represent two message classes or different traffic conditions. It is also assumed that the basic queuing set-up is Markovian, i.e. the arrival and service processes are exponential. In particular it is assumed that the arrival rate of intensity λ takes values from the set $\{\lambda_0, \lambda_1\}$. Corresponding to the arrival rates the service process has the rate $\{\mu_0, \mu_1\}$.

Let $N_\lambda(t)$ denote the arrival process which is Poisson parametrized by the intensity λ. Let k, $k = 1, 2, 3, \ldots$, denote the arrival epoches of the process to the buffer.

The intensity is modelled as

$$\begin{aligned}\lambda(k+1) &= \left[\lambda_0 + (\lambda_1 - \lambda_0) I(\Theta_1 = 1)\right] I(\lambda(k) = \lambda_0) \\ &\quad + \left[\lambda_1 + (\lambda_0 - \lambda_1) I(\Theta_2 = 1)\right] I(\lambda(k) = \lambda_1)\end{aligned} \tag{2.1}$$

where Θ_1 and Θ_2 are two binary random variables with:

$$\Pr(\Theta_1 = 1) = (1 - p), \tag{2.2}$$

$$\Pr(\Theta_2 = 1) = (1 - r) \tag{2.3}$$

and $I(A)$ is the indicator function of the event A. The index k denotes the kth arrival instant.

The model for the intensity of the Poisson process given by (2.1)–(2.3) describes an intensity which fluctuates randomly between the values λ_0 and λ_1 with its sojourn in a particular state governed by a geometric distribution. It is however important to note that the Markov process governing the parameter fluctuations is now not independent of the queuing process. The above model can be readily extended to the case where $\lambda \in \{\lambda_0, \lambda_1, \ldots, \lambda_N\}$ but only the case with two values is considered here.

The model above assumes that the intensity changes at a random arrival instant. This results in a discretization of the model, and seems realistic since detection of changes can be made at arrival times as the interarrival times carry information about the stochastic process [7].

The above model is motivated by problems arising in the analysis of sources feeding into a network with time dependent characteristics. This class of models is proposed for packetized video with variable rate coding. Specifically, the coder can operate at various rates depending on the motion present [6]. When the coder is operating in the normal region packets arrive with a given rate and are served according to a given rate. When motion is present, the coder output rate increases and on the receipt of the packets in the queue the service rate is altered in order to maintain the fidelity of the transmission. This results in an adaptation of the service rate when the arrival rate changes. The length of time in each mode is assumed to be exponential, modelled by the parameters p and r in the model.

3. Infinite buffer case

3.1. Introduction

In this section, we study the case where the queue described above has an infinite buffer. This model does not obey the independence assumptions [5]; in particular, the sequence of interarrival times cannot be considered as independent and thus the standard results derived in classical Queuing Theory cannot be used. Thus, the method we will use is the basic Markovian one. We use the moment generating function approach to analyse this model and Jury's criteria [4] to obtain the stability conditions.

3.2. Analysis

3.2.1. The Markov chain

According to the formulation of the model, the description of the infinite two dimensional Markov chain is straightforward. A state of this chain can be viewed as the pair (k, λ_0) (resp. (k, λ_1)) where k represents the number of messages in system and λ_0 (resp. λ_1) represents the arrival rate of the last arriving message. The state-transition-rate diagram for this infinite state system is shown in Fig. 1 where the upper line corresponds to λ_0 and the lower to λ_1.

3.2.2. Equations

Let us study this Markov chain in steady state. Define:

$P_k = \Pr(k \text{ messages in system, } \lambda_0)$,

$Q_k = \Pr(k \text{ messages in system, } \lambda_1)$,

$N_k = \Pr(k \text{ messages in system}) = P_k + Q_k$.

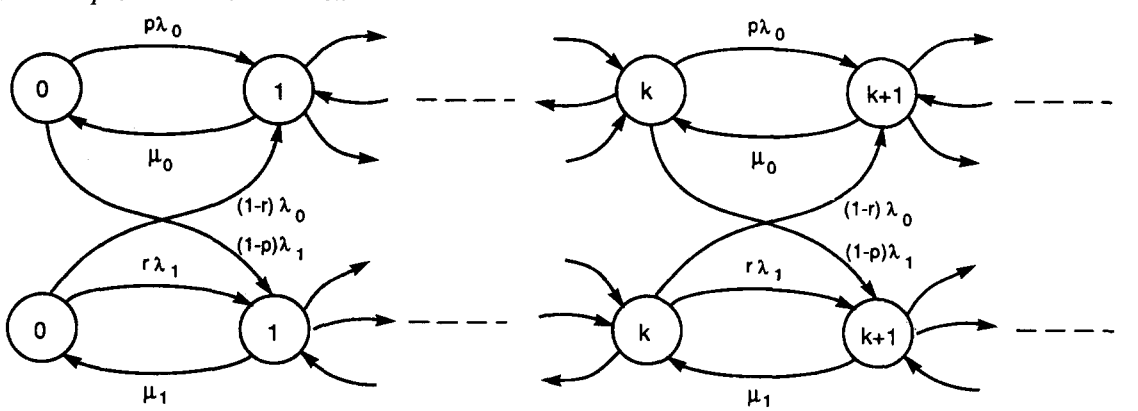


Fig. 1. The Markov chain: infinite buffer case.

and let the generating function for P_k (resp. Q_k and N_k) be

$$P(z) = \sum_{k=0}^{\infty} P_k z^k \quad \text{(resp. } Q(z) \text{ and } N(z)\text{)}.$$

We then find

$$N(z) = \frac{a_2 z^2 + a_1 z + a_0}{b_3 z^3 + b_2 z^2 + b_1 z + b_0} = \frac{N_u(z)}{D_e(z)} \tag{3.1}$$

with

$$P(1) = \frac{(1-r)\lambda_0}{(1-r)\lambda_0 + (1-p)\lambda_1}, \tag{3.2}$$

$$Q(1) = \frac{(1-p)\lambda_1}{(1-r)\lambda_0 + (1-p)\lambda_1}, \tag{3.3}$$

$$\begin{aligned}\mu_0 P_0 + \mu_1 Q_0 &= [\mu_0 - p\lambda_0 - (1-p)\lambda_1] P(1) \\ &\quad + [\mu_1 - r\lambda_1 - (1-r)\lambda_0] Q(1) \\ &= V \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} a_2 &= (-1+p+r)(\lambda_0 V + \mu_0(\lambda_1 - \lambda_0)P_0), \\ a_1 &= (-\mu_0 - p\lambda_0 - (1-p)\lambda_1)V \\ &\quad + \mu_0[(\mu_0 - \mu_1) - (-1+p+r)(\lambda_1 - \lambda_0)] \\ &\quad \times P_0, \\ a_0 &= \mu_0(V + (\mu_1 - \mu_0)P_0); \end{aligned} \tag{3.5}$$

$$\begin{aligned} b_3 &= \lambda_0\lambda_1(1-r-p), \\ b_2 &= \lambda_0\lambda_1(1-p-r+2pr) + r\mu_0\lambda_1 + p\lambda_0\mu_1 \\ &\quad + (1-p)r\lambda_1^2 + (1-r)p\lambda_0^2, \\ b_1 &= \lambda_0(-p\mu_1 - (1-r)\mu_0) \\ &\quad + \lambda_1(-r\mu_0 - (1-p)\mu_1) - \mu_0\mu_1, \\ b_0 &= \mu_0\mu_1. \end{aligned} \tag{3.6}$$

The problem is now to eliminate the last unknown constant, P_0.

3.2.3. Discussion

The solution for P_0 is obtained by noting that $N(z)$ is an analytic function in z of a probability distribution. First note that $N(1) = N_u(1)/D_e(1) = 1$ where $N_u(1) = a_2 + a_1 + a_0 < 0$ since $p, r < 1$.

This implies that $D_e(1)$ must also be negative which is just equivalent to the necessary condition for stability, i.e. $\bar{\lambda} < \bar{\mu}$, where $\bar{\lambda}$ and $\bar{\mu}$ are the mean arrival and the mean service rates respectively, that is

$$\begin{aligned}\bar{\lambda} &= (p\lambda_0 + (1-p)\lambda_1)P(1) + (r\lambda_1 + (1-r)\lambda_0) \\ &\quad \times Q(1)\end{aligned}$$

and

$$\bar{\mu} = \mu_0 P(1) + \mu_1 Q(1). \tag{3.7}$$

Noting that $N_u(0) = a_0 > 0$, this together with $N_u(1) < 0$ imply that the numerator has a real root in (0, 1]. Since $D_e(1) < 0$, $D_e(0) > 0$, there exists a real root for the denominator in (0, 1]. Since $N(z)$ is analytic, if $\bar{z}$ is a root of the numerator it must be a root of $D_e(z)$. Thus setting $N_u(\bar{z}) = 0$ enables us to compute P_0 exactly [11].

Let us now study the issue of stability.

3.2.4. Stability criteria

We know that a queuing system is stable if and only if the associated Markov chain is ergodic [5], in which case there exists an unique solution for the set of N_k where N_k satisfies $0 \leqslant N_k \leqslant 1$ and

$$\sum_{k=0}^{\infty} N_k = 1.$$

The uniqueness of N_k under these conditions is equivalent to the analyticity of $N(z)$ in the unit circle. It also implies that N_0 is unique and hence so must P_0 since

$$P_0 = N_0 - Q_0 = \frac{\mu_1 N_0 - V}{(\mu_1 - \mu_0)} \tag{3.8}$$

and V is unique (see Eq. (3.4)). Therefore, the system will be stable if $N(z)$ is analytic in the unit circle and P_0 is unique. This is equivalent to the denominator having a single root in the unit circle [11]. Hence necessary and sufficient conditions for stability are $D_e(1) < 0$ and D_e has a single root within the unit circle.

We now use Jury's criteria [4] to find the conditions under which the denominator has a single root within the unit circle. This criteria has been widely used in the study of the stability of discrete time control systems and presents us with a convenient analytical tool to obtain the necessary and sufficient conditions for the Markov chain to be ergodic. For details see [4]. It suffices to mention that for an nth degree polynomial $p(z)$, we can define n products denoted Π_i, $i = 1, 2, \ldots, n$, with each Π_i being specified completely by the coeffi-

cients of the polynomial such that the number of negative products is equal to the number of zeros of the polynomial inside the unit circle. For the case of the polynomial of 3rd degree $D_e(z)$ we have

$$\Pi_1 = |b_0| - |b_3| = \mu_0\mu_1 - |1-p-r|\lambda_0\lambda_1,$$

$$\Pi_2 = \Pi_1\left(\left|b_0^2 - b_3^2\right| - |b_0b_2 - b_1b_3|\right),$$

$$\Pi_3 = \Pi_1\Pi_2 D_e(-1)D_e(1).$$

Hence by Jury's criteria $D_e(z)$ will have a single root within the unit circle if and only if exactly one of the products is negative.

Hence, a necessary and sufficient condition for stability is that exactly one of the conditions S_i, $i = 1, 2$, holds

$$S_1 = \{D_e(1) < 0,\ \mu_0\mu_1 > |1-p-r|\lambda_0\lambda_1\}, \tag{3.9}$$

$$S_2 = \{D_e(1) < 0,\ \mu_0\mu_1 < |1-p-r|\lambda_0\lambda_1,\ \Pi_2 > 0\}. \tag{3.10}$$

By further analysis using a convexity argument we show that [12] if $D_e(1) < 0$, then either S_1 or S_2 always holds and hence $D_e(1) < 0$ is both a necessary and sufficient condition for the stability of the queuing system.

To show that $D_e(1) < 0$ is necessary and sufficient is equivalent to showing that $D_e(1) < 0$ and $\Pi_1 < 0$ imply that $\Pi_2 > 0$. Two cases, i.e. $p + r < 1$ and $p + r > 1$, need only be considered. The case $p + r = 1$ is trivial.

Case 1: $p + r < 1$ ($b_3 > 0$). Then $\Pi_2 > 0$ when $D_e(1) < 0$ and $\Pi_1 < 0$ follows by simple algebra.

Case 2: $p + r > 1$ ($b_3 < 0$). Define $F(p, r) = b_3^2 - b_0^2 + b_0b_2 - b_3b_1$.

Then simple algebra gives $\Pi_2 > 0$ is equivalent to $F(p, r) < 0$ for $p + r > 1$. Let $p + r = a$ and consider the parametric equation $f_a(p) = F(p, a - p)$ for a in [1, 2). Then simple algebra shows that $f_a(p) = up^2 + vp + w$ where $u = \mu_1\mu_0(\lambda_0 - \lambda_1)^2 > 0$ is independent of a. Hence $f_a(p)$ is concave. It is easy to see that $(a-1, 1)$ and $(1, a-1)$ are the extreme points of this equation for a in [1, 2) and substituting for these it can be readily seen that $f_a(1)$ and $f_a(a-1) < 0$ and since $f_a(p)$ is concave, it implies that for any value of p in $[a-1, 1]$, $f_a(p) < 0$.

Another way of seeing this is that for $p = 1$ and $r = a - 1$, the system is equivalent after a finite time with an M/M/1 queue with parameters (λ_0, μ_0) and thus $D_e(1) < 0$ is necessary and sufficient.

3.2.5. Remarks about the stability conditions

(a) The above condition $D_e(1) < 0$ is equivalent to $\bar{\lambda} = E(\lambda) < E(\mu) = \bar{\mu}$. This condition agrees with the recent results of Baccelli and Makowski [1] in which they showed that a very large number of queues in random environments are stable iff the average arrival rate is less than the average service rate. Their techniques were based on Palm theory and Loynes' scheme. What we have obtained is a direct proof by purely analytic techniques without invoking probabilistic arguments.

(b) In the GI/GI/1 case where the arrival and service processes are independent of each other, the probability of the system being non empty is given by $N_0 = 1 - \rho$ where $\rho = \bar{\lambda}/\bar{\mu}$ is the utilization factor (see for example [2]).

Since the model we consider does not fall under this class, this result does not hold in general. In fact defining $X = N_0 - 1 + \rho$,

$$X = \frac{(\mu_1 - \mu_0)}{\bar{\mu}}(P_0Q(1) - Q_0P(1))$$

can be thought of as a measure fo the deviation from the GI/GI/1 results. Note that $X = 0$ if $\mu_1 = \mu_2$ or $P_0Q(1) = Q_0P(1)$.

Figures 2 and 3 show X as a function of μ_1 (for example). In Fig. 2 $(\lambda_0, \lambda_1, \mu_0, p, r) = (1, 4, 2, 0.2, 0.4)$ and $\mu_1 > \mu_0$ for stability whereas in Fig. 3 $(\lambda_0, \lambda_1, \mu_0, p, r) = (1, 3, 4, 0.3, 0.8)$ and the system is stable for all choices of μ_1.

(c) It can be easily seen that a system with two stable "subqueues" is stable, which justifies our

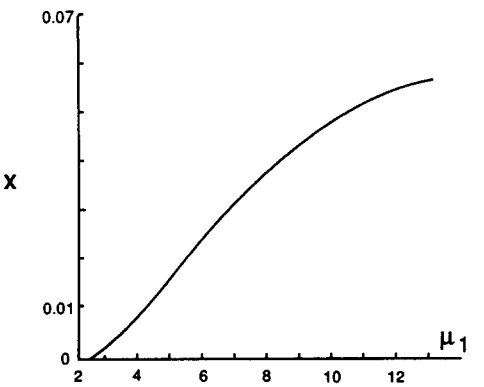


Fig. 2. X as a function of μ_1 with $(\lambda_0, \lambda_1, \mu_0, p, r) = (1, 4, 2, 0.2, 0.4)$.

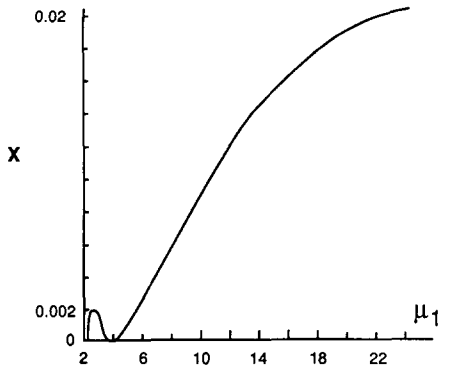


Fig. 3. X as a function of μ_1 with $(\lambda_0, \lambda_1, \mu_0, p, r) = (1, 3, 4, 0.3, 0.8)$.

intuition (a subqueue corresponds to the pair (λ_i, μ_i) $i = 0, 1$). However it may be possible for the entire system to be stable even if both the subqueues are unstable. This is due to the fact that the parameters can compensate for each other. As an example consider the case where $\lambda_0 > \mu_0 > \lambda_1 > \mu_1$ with $\lambda_0 = 100$, $\mu_0 = 90$, $\lambda_1 = 10$, $\mu_1 = 9$ and $p = r = 0.1$; then $D_e(1) < 0$ and so the stability condition is satisfied. This is due to the switching; the system will be most likely served at the highest rate, namely μ_0 (with probability $P(1) = 0.909$) whereas the arrival rate which will most likely occur is the lowest, namely λ_1 (with a probability $(1-p)P(1) + rQ(1) = 0.83$). It can also be easily seen that for the case $\lambda_i > \mu_j$ for $i, j = 0, 1$, the system is always unstable.

3.2.6. Performance measures

In this section some performance measures for the infinite buffer case are computed and compared to known results.

Figures 4–6 show the behaviour of the average number of messages N in the queue vs. r for different sets of parameters. These results are compared to the result $\bar{N}$ for a standard M/M/1 queue with the parameters $\bar{\lambda}$ and $\bar{\mu}$ as defined in Eq. (3.7). In Fig. 4, N and $\bar{N}$ have the same behaviour but N is always greater than $\bar{N}$. In Fig. 5, N and $\bar{N}$ have similar behaviour but N is not always greater than $\bar{N}$. In Fig. 6, N and $\bar{N}$ have different behaviour and N is always greater than $\bar{N}$. This implies that in general it is difficult to predict the performance by considering the "averaged" M/M/1 approximation.

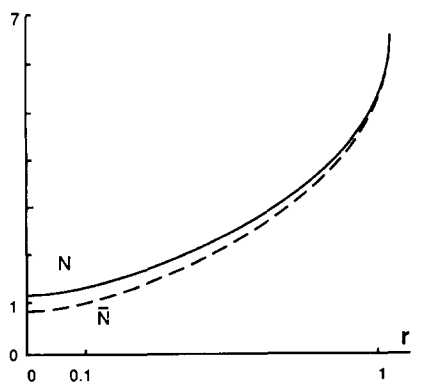


Fig. 4. $N(r)$ and $\bar{N}(r)$ for $(\lambda_0, \lambda_1, \mu_0, \mu_1, p) = (2, 7, 3, 8, 0.4)$.

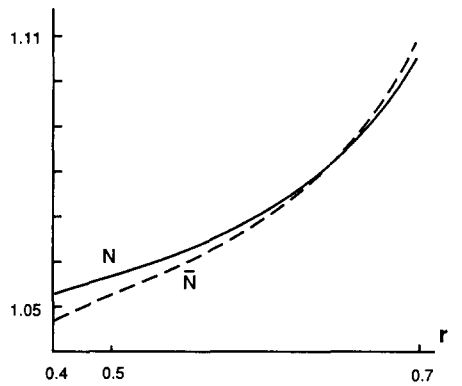


Fig. 5. $N(r)$ and $\bar{N}(r)$ for $(\lambda_0, \lambda_1, \mu_0, \mu_1, p) = (2, 3, 4, 5, 0.92)$.

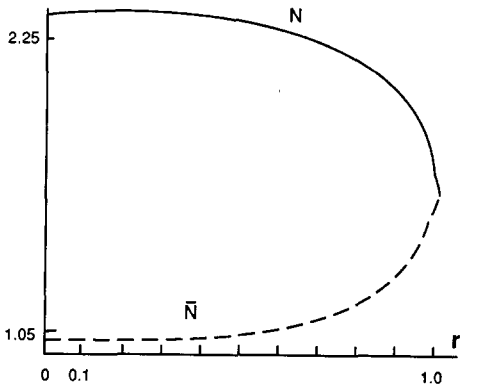


Fig. 6. $N(r)$ and $\bar{N}(r)$ for $(\lambda_0, \lambda_1, \mu_0, \mu_1, p) = (8, 2, 6, 3, 0.08)$.

4. The finite buffer scheme with a resume level

4.1. Introduction

The resume level scheme is now analysed as an exponential queuing system with a finite waiting room K, subject to arrival and service processes discussed above. When the buffer is full, the input flow is shut down until the contents of the buffer

fall below a level H, called the resume level or hysteresis, after which the input is restored.

Below the model is analysed via the moment generating function approach as in Section 3.2.2.

4.2. Analysis

4.2.1. The Markov chain

From the formulation of the model it is straightforward to see that the queue forms a continuous time Markov chain whose state is parametrized by the triplet (n, λ_i, x) where n represents the number of messages in the buffer, λ_i represents the arrival rate of the last arriving message and x is a binary variable with $x = 1$ representing non-blocked input flow implying that the buffer has not saturated while $x = 0$ implies that the buffer has saturated but not dropped below H indicating blocked input flow. The state transition diagram for the process is shown in Fig. 7. The chain has four phases. The topmost (bottommost) corresponds to $x = 0$, $\lambda = \lambda_0$ (resp. λ_1). Phase 2 (resp. phase 3) corresponds to $x = 1$ and $\lambda = \lambda_0$ (resp. λ_1).

4.2.2. Equations

The Markov chain is now analysed to obtain the steady-state or invariant distributions for the queuing process (the existence is trivial since the process forms a finite irreducible Markov chain).

Define:

$$P_k = \Pr(k \text{ messages in buffer}, \lambda_0, x = 1), \quad (4.1)$$

$$Q_k = \Pr(k \text{ messages in buffer}, \lambda_1, x = 1), \quad (4.2)$$

$$T_k = \Pr(k \text{ messages in buffer}, \lambda_0, x = 0), \quad (4.3)$$

$$L_k = \Pr(k \text{ messages in buffer}, \lambda_1, x = 0), \quad (4.4)$$

and

$$N_k = P_k + Q_k + T_k + L_k. \quad (4.5)$$

And let $P(z)$, $Q(z)$, $L(z)$, $T(z)$ and $N(z)$ denote the moment generating functions of $P.$, $Q.$, $L.$, $T.$ and $N.$ respectively, i.e.

$$P(z) = \sum_{n=0}^{K-1} P_n z^n, \quad (4.6)$$

$$Q(z) = \sum_{n=0}^{K-1} Q_n z^n, \quad (4.7)$$

$$T(z) = \sum_{n=0}^{K} T_n z^n = A(z)(K - H)Pb_1 \quad (4.8)$$

where

$$T_m = \rho_0(pP_{K-1} + (1-r)Q_{K-1}) = Pb_1 \quad \text{for } H < n \leqslant K,$$

$$\rho_0 = \frac{\lambda_0}{\mu_0} \quad \text{and} \quad A(z) = \sum_{n=H+1}^{K} \frac{z^n}{K-H}. \quad (4.9)$$

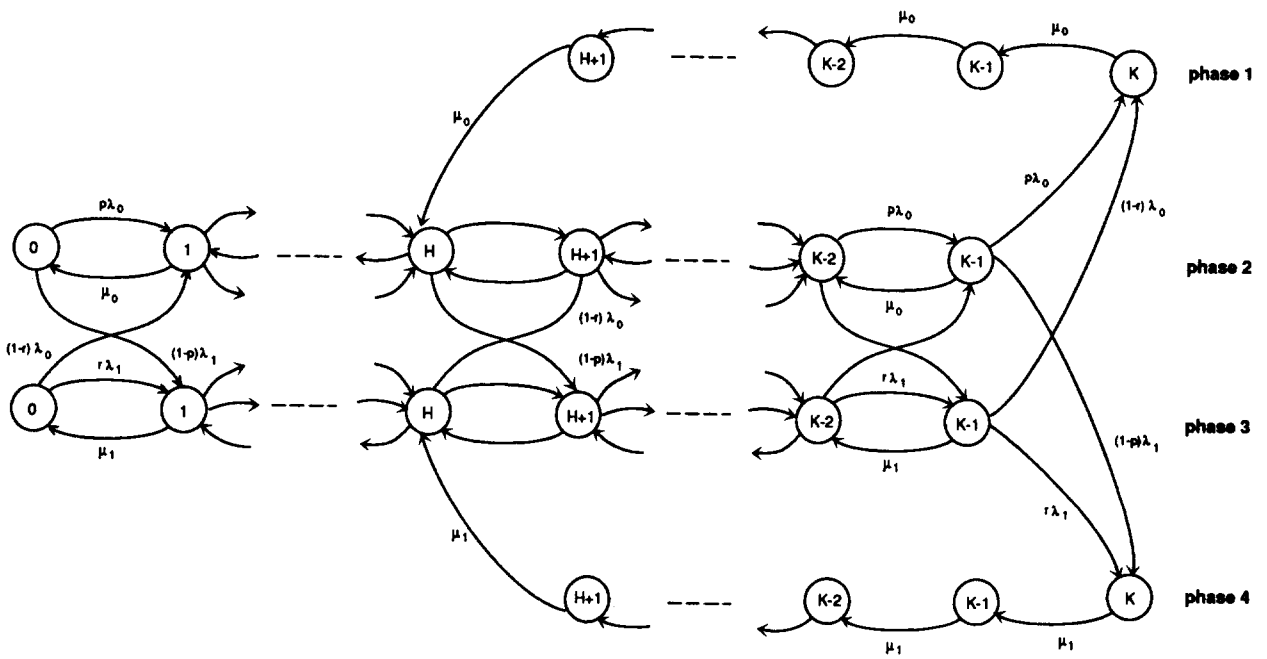


Fig. 7. The Markov chain: finite buffer with a resume level.

In a similar way it can be seen that

$$L(z) = A(z)(K-H)Pb_2 \tag{4.10}$$

where

$$L_m = \rho_1(rQ_{K-1} + (1-p)P_{K-1}) = Pb_2$$
$$\text{for } H < n \leqslant K,$$
$$\rho_1 = \frac{\lambda_1}{\mu_1}. \tag{4.11}$$

Then $N(z)$ is given by

$$N(z) = P(z) + Q(z) + L(z) + T(z)$$
$$= P(z) + Q(z) + \text{Pb}\, A(z). \tag{4.12}$$

Remark. Note that $\text{Pb} = (\text{Pb}_1 + \text{Pb}_2)(K-H)$ is the blocking probability of the system.

Hence taking transforms of the equilibrium probability equations and after straightforward algebra it can be easily seen that $N(z)$ can be obtained in terms of the system parameters and the four unknowns (P_0, Q_0, Pb_1, Pb_2) as

$$N(z) = \big(a_2z^2 + a_1z + a_0 + A(z)(K-H)$$
$$\times\big(a_{K-1}z^2 + a_{K-2}z + a_{K-3}\big)\big)$$
$$\times\big(b_3z^3 + b_2z^2 + b_1z + b_0\big)^{-1} + \text{Pb}\, A(z) \tag{4.13}$$

with

$$a_2 = (p + r - 1)(\lambda_1\mu_0P_0 + \lambda_0\mu_1Q_0),$$
$$a_1 = -\mu_1(\mu_0 + p\lambda_0 + (1-p)\lambda_1)Q_0$$
$$-\mu_0(\mu_1 + r\lambda_1 + (1-r)\lambda_0)P_0,$$
$$a_0 = \mu_0\mu_1(P_0 + Q_0).$$

The parameters a_{K-1}, a_{K-2}, a_{K-3} can be obtained by replacing P_0 by Pb_1 and Q_0 by Pb_2 in the expressions for a_2, a_1, a_0 respectively.

Remark. The denominator is identical to the denominator in the case of an infinite buffer (see (3.6)) and is completely characterized by the known probabilities, arrival and service rates p, r, λ_0, λ_1, μ_0, μ_1, respectively.

The unknowns can now be eliminated by using the fact that $N(z)$ must be analytic being a moment generating function of a probability distribution.

From the expression for $N(z)$ it can be seen that it is of the form

$$N(z) = \text{Pb}\, A(z) + \frac{P_u(z)}{D_e(z)} \tag{4.14}$$

where $P_u(z)$ is a polynomial in z of order $K+2$. Since P_k, Q_k, L_k, T_k are zero for $k > K$ it implies that $N(z)$ must be a polynomial of order K. This is only possible for all values of the parameters λ, μ, p, r if and only if the roots of the denominator polynomial $D_e(z)$ are also roots of $P_u(z)$. This yields three equations in order to eliminate the four unknowns. The fourth relation is obtained from the fact that $N(z)$ being a moment generating function must be equal to 1 at $z = 1$. Thus, $N(z)$ can be written as a Kth degree polynomial in z whose coefficients can be completely determined and hence the various probabilities can be solved for.

Remark. If $K \to \infty$, then $A(z)$ tends to zero and the results of Section 3 can be recovered.

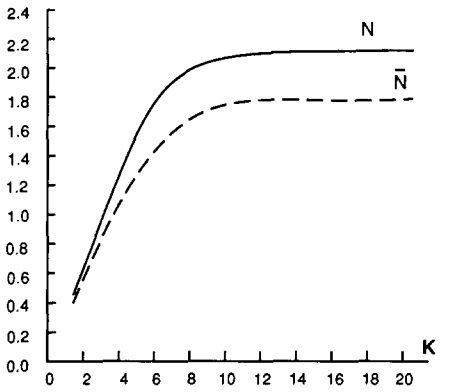


Fig. 8. $N(K)$ and $\bar{N}(K)$ for $H = K-1$ and (λ_0, λ_1, μ_0, μ_1, p, r) = (8, 2, 7, 3, 0.08, 0.5).

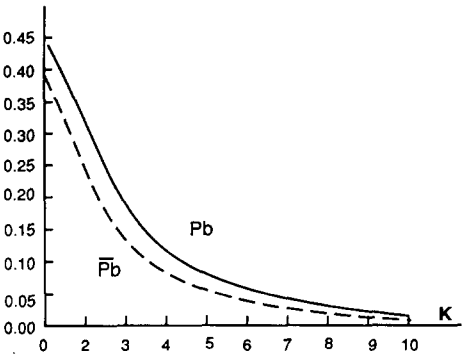


Fig. 9. $\text{Pb}(K)$ and $\overline{\text{Pb}}(K)$ for $H = K-1$ and (λ_0, λ_1, μ_0, μ_1, p, r) = (8, 2, 7, 3, 0.08, 0.5).

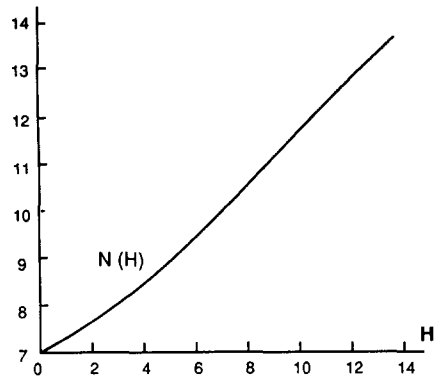


Fig. 10. $N(H)$ for $(\lambda_0, \lambda_1, \mu_0, \mu_1, p, r, K) = (10, 50, 11, 20, 0.2, 0.5, 15)$.

4.2.3. Performance measures

In this section some performance measures for the buffer scheme are computed and compared to known results. Figure 8 shows the behaviour of the average number of messages N in the queue vs. the buffer capacity assuming that there is no resume level. This is compared to the results $\overline{N}$ for a standard M/M/1/K case with the parameters $\bar{\lambda}$ and $\bar{\mu}$ as defined in Eq. (3.7).

Figure 9 shows the behaviour of the blocking probability Pb vs. K, the capacity of the queue. $\overline{\text{Pb}}$ represents the blocking probability in the M/M/1/K case or the Erlang loss function.

It can be seen that while the behaviour is the same, the average model gives optimistic results, i.e. lower average number of messages and lower blocking probability than the model proposed here.

For a finite buffer scheme with a resume level a measure of the oscillations in the system due to the on/off nature of the control is given by F_s, the frequency of the on/off cycles (see [10]);

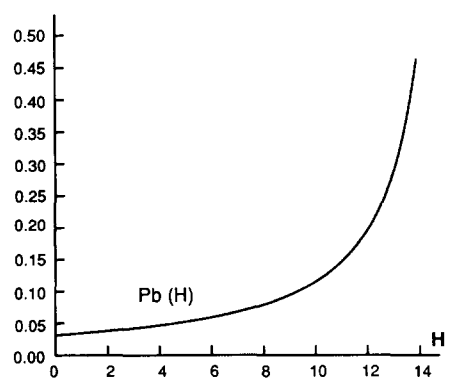


Fig. 11. Pb(H) for the same parameters as in Fig. 10.

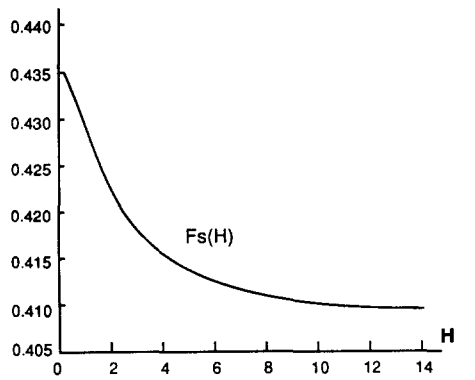


Fig. 12. $F_s(H)$ for the same parameters as in Fig. 10.

$$F_s = \frac{\text{Pb}}{K - H}.$$

Figures 10, 11, 12 show the behaviour of the average number of messages, the blocking probability and the frequency F_s vs. the resume level, H, for a finite capacity queue with $K = 15$ for typical values of the parameters. These curves compare very well with the "average" curves for the M/M/1/15 queue with the average values $\bar{\lambda}$, $\bar{\mu}$ defined earlier. This is because the average sojourn time for the processes in the second phase is appreciable and the average loading is high reducing the effect of the fluctuations in the parameters. The effect of the fluctuations is felt when the loading is low or if the sojourn time in the overload region is small and the parameters are vastly different. Thus under heavy traffic conditions the random parameters may be replaced by their averages and the conclusions of Reiser [10] are valid. However, these are only for the first order statistics. The second-order statistics can be readily obtained since the moment generating functions are known exactly.

The analysis of the curves permits the determination of a suitable resume level in order to obtain a good compromise between an average number of customers in the system and a small F_s so that the buffer could be designed to minimize the average number of resume cycles.

References

[1] F. Baccelli and A. Makowski, Stability and bounds for single server queues in random environment, *Comm. Statist.: Stochast. Models* **II** (2) (1986).

[2] E. Gelenbe and G. Pujolle, Introduction aux réseaux de files d'attente, Collection CNET-ENST, Eyrolles, 1982.

[3] E. Gelenbe and C. Rosenberg, Queues with slowly varying arrival and service process, *Manag. Sci.*, to appear.

[4] E.I. Jury, *Theory and Application of the Z-Transform Method* (Wiley, New York, 1964) 90–100.

[5] L. Kleinrock, *Queueing Systems, Vol. I: Theory* (Wiley-Interscience, New York, 1975).

[6] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson and J. Robbins, Performance models of statistical multiplexing in packet video communications, *IEEE Trans. Comm.* (July 1988).

[7] R. Mazumdar, Quickest detection of disorders in point processes – An application to computer communication networks, in: *Proc. Conference on Information Sciences and Systems*, Princeton (1985).

[8] M.F. Neuts, A queue subject to extraneous phase changes, *Adv. Appl. Probab.* **3** (1971) 78–119.

[9] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models* (Johns Hopkins Univ. Press, Baltimore, MD, 1981).

[10] M. Reiser, Queueing and delay analysis of a buffer pool with a resume level, in: *Performance '83* (North-Holland, Amsterdam, 1983).

[11] C. Rosenberg, Exponential queueing systems with randomly changing arrival and service rates, M.S. Thesis, Dept. of Computer Science, UCLA, 1984.

[12] C. Rosenberg, Non-stationnarité dans les files d'attente Markoviennes, Thèse de Doctorat en Science, Dép. d'informatique, Orsay, 1986.

[13] H. Takagi, Analysis of a finite capacity M/G/1 queue with resume level, *Perform. Eval.* **5** (1985).

[14] U. Yechiali and P. Naor, Queueing problems with heterogeneous arrivals and service, *Op. Res.* **19** (1971) 722–734.