A Queue with Correlated Arrivals

Ralph L. Disney[1] *Donald C. McNickle*[2] *Sun Hur*[3]
Paulo Renato de Morais[4] *Ryszard Szekli*[5]

[1]Department of Industrial Engineering
Texas A&M University, College Station, USA

[2]Department of Management
University of Canterbury, Christchurch, New Zealand

[3]Department of Industrial Engineering
Hanyang University, Ansan, Korea

[4]Valley of Paraiba University (UNIVAP)
Av. Shishima Hifumi, 2911
12244-000 Sao Jose dos Campos - SP, Brazil

[5]Department of Mathematics
Wrocław University, Wrocław, Poland

Abstract

This paper deals with a queue with a Markov renewal arrival process (MRP) which is autocorrelated. Our choice of models for the arrival process has been motivated by the need to keep the marginal distributions the same as far as possible. By doing so, it is possible to better expose the pure effects of the parameters in the arrival process on the correlation coefficient and thence on the mean queue length. We consider the various effects of 4 parameters: the "stickiness" of the underlying Markov chain (p), differences in the mean interarrival times of each type ($m_i - m_j$), the variance of the arrival times of each type (v_j), and the number of states (n). We show that p and $m_i - m_j$ interact in such a way that the rate of convergence of mean queue length to infinity is faster for large $m_i - m_j$ as a function of p. It is possible for the queue length process to be in steady state but the mean queue length to be arbitrary large solely due to correlations. We also show that decreases in v_j increase correlations but can decrease the mean queue length. Also, the number of states, acting through the correlation coefficient can have additional effects on the mean queue length especially in the case of "sticky" MRP's. It would appear that more attention should be paid to the correlations especially in situations where the traffic intensity is high and where correlations can be present and can be large.

Keywords: Markov renewal process, single server queue, autocorrelation.

1 Introduction

The theory and most published applications of queueing theory start with three major assumptions: the arrival process is a sequence of i.i.d. random variables; the service times are i.i.d.; the arrival sequence and service time sequence are independent. These are ideas that have served well for nearly 90 years. But new demands being put on queueing through quality control, reliability, manufacturing and communications are showing that these models are no longer adequate. The new areas are not calling for new ever more complicated inter-event distributions but rather are calling for new process structures which incorporate dependencies.

In computer/communications modeling there is increasing emphasis on using non-renewal processes in models (see, for example, Paxon and Floyd(1994)). Some kinds of traffic in these systems has been identified as self similar (see Willinger, Taqqu, Leland and Wilson(1995)). A SM/G/1 queue (similar to our model but in discrete time) has been used by Hasslinger and Adams (1996) in the performance analysis of ATM traffic.

In this paper we study a queue with a Markov renewal arrival process (MRP). This class of MRP's has been chosen so that some of the characteristics of the arrival process can be kept fixed while others (especially the correlational structure) are varied. For much of the paper we will concentrate on a two-state process. The numerical techniques we use can deal with higher-state processes. However, the number of parameters then becomes confusing. Also a frequently described characteristic of a non-renewal communications traffic is that it basically moves between two states–a quiet state and an extremely busy state. Thus we hope that even our simple two-state model may provide some insight into the characteristics of such systems.

Some of the theoretical results we use come from Szekli, Disney and Hur (1994a or 1994b). We summarize these results here so that this paper is self contained.

In Section 3 we need to consider the various effects of 4 parameters on the correlation coefficient and thence on the mean queue length. In Subsection 3.1 we look at the effect of p, the one step transition probability for the Markov chain underlying our basic Markov renewal arrival process. In Section 3.2 we consider the effect of the mean values of the separate interarrival processes. In Section 3.3 we look at the effects of the variances of the separate interarrival process. In Section 3.4 we consider the dimensionalilty of the problem and in Section 3.5 the effects of the correlations of all orders as they define the index of dispersion for intervals (IDI) as defined in Sriram and Whitt (1986). In these sections our investigations necessarily are numerical and graphical.

Our choice of models for the arrival process has been motivated by the need to keep the marginal distributions the same (as far as possible) in order to better expose the pure effects of the parameters. Because of condition (i) in section 3.4 we still have the same marginals as in Section 3.1. However, by the changes in sections 3.2, 3.3

and 3.5 we are basically changing the marginal distributions. Thus, comparisions due purely to changes in the correlation coefficient can no longer be made as they were in sections 3.1 and 3.4.

We are interested in the steady state mean number in the system seen by an arrival, EN^a, but since the continuous time queue length process has a mean related to EN^a as $L^t = EN^t = \rho(EN^a + 1)$, any comment that we make about one applies equally to the other. Similarly, by Little's result, any comment we make about the mean queue length applies equally well to the mean waiting times.

2 The MRP Arrival Queue

2.1 MR Arrival Processes

We consider the homogeneous MR/M/1 queue where the mean service time is μ. The Markov renewal arrival process has kernel $\mathbf{Q}(t)$ where

$$Q(i,j,t) = p(i,j)F(i,j,t) \quad i,j \in E, t \geq 0.$$

Here E is a countable state space of the process, $p(i,j)$ is the one step probability for the Markov chain embedded in the MRP. $F(i,j,t)$ is, for each i,j, a distribution of the sojourn time spent in state i given the next state is j.

A Markov renewal process whose kernel has the form:

$$\mathbf{Q}(t) = \begin{bmatrix} pF_1(t) & \frac{1-p}{n-1}F_2(t) & \cdots & \frac{1-p}{n-1}F_n(t) \\ \frac{1-p}{n-1}F_1(t) & pF_2(t) & \cdots & \frac{1-p}{n-1}F_n(t) \\ \cdot & \cdot & \cdots & \cdot \\ \frac{1-p}{n-1}F_1(t) & \frac{1-p}{n-1}F_2(t) & \cdots & pF_n(t) \end{bmatrix} \tag{1}$$

first introduced in Szekli, Disney and Hur (1994a), is particularly suitable for investigating the effects of correlation in the arrival process, since, as we shall show below, the correlation can be altered without changing the marginal distributions of the arrival process.

Thus, the sojourn time in state i only depends on the next state to be visited. If $Q(i,j,t)$ depended only on i (and not j) then the arrival process would be the superposition of n independent renewal processes thereby negating the effect of our primary interest, i.e., correlation in the arrival process. Change-over time problems in manufacturing may come close to our model.

Let m_j and v_j be the mean and variance of the sojourns when the next state is j. Then,

$$\rho_j = \frac{1}{\mu m_j}$$

is the traffic intensity for arrivals from state j.

We need the following definition later. Consequences of this and many other of these stochastic comparision theorems can be found in Stoyan (1983).

Definition 1: A distribution function F is smaller with respect to the increasing convex ordering than the distribution function G (symbollically $F \leq_{icx} G$) if for all increasing convex functions f for which the integrals exist we have

$$\int_{-\infty}^{\infty} f(x)\, dF(x) \leq \int_{-\infty}^{\infty} f(x)\, dG(x). \tag{2}$$

Also, if $EX = EY$ then $Ef(X) = Ef(Y)$ for any convex function increasing or not. For our purposes, the important result is that if $F \leq_{icx} G$ and $EX \leq EY$ then $\mathrm{Var}X \leq \mathrm{Var}Y$. We will see how this applies to our MRP arrival queue in Section 3.3.

2.2 Preliminary Results

We have the following preliminary results.

(a) Let D_k be the time between arrival k and $k-1$ regardless of their types. The sequence $D = \{D_k : k = 1, 2, \cdots\}$ is the overall interarrival time process irrespective of the state of the arrival process. For any n-state homogeneous MRP of the type described above

$$P\left(D_k \leq t\right) = \frac{1}{n} \sum_{j=1}^{n} F_j(t),$$

independent of p. This result is important since it allows us to change p without changing the marginal interarrival distributions (compare this to Patuwo, Disney and McNickle (1993)).

(b) For any finite r, and any homogeneous MRP

$$\mathrm{Corr}(r) = \frac{E\left[D_k D_{k+r}\right] - E\left[D_k\right] E\left[D_{k+r}\right]}{\mathrm{Var}D_k},$$

depending only on r. For the class of MRP's in equation (1):

$$\mathrm{Corr}(r) = \frac{\frac{1}{n^2} \sum\limits_{i<j} \left(m_i - m_j\right)^2}{\frac{1}{n} \sum\limits_{j=1}^{n} v_j + \frac{1}{n^2} \sum\limits_{i<j} \left(m_i - m_j\right)^2} {\xi_n}^r \tag{3}$$

(See Szekli, Disney and Hur (1994b) for the proof). Here, as elsewhere $\xi_n = (np-1)/(n-1)$ is the subdominant eigenvalue of the transition matrix for the embedded Markov chain. The importance of the correlation function above is that it

depends, explicitily, on all of the parameters of the system. However, because of the construction of the MR arrival process, the correlation can be changed by changing p without changing the other parameters and, hence, the marginal interarrival distribution. This means that one can change the correlation in the arrival process without changing the interarrival times distributions.

(c) It is interesting to note that with an obvious manipulation, the correlation function in (3) can be written as

$$\mathrm{Corr}(r) = \frac{{\xi_n}^r}{\left(n \sum_{j=1}^{n} v_j / \sum_{i<j} (m_i - m_j)^2 + 1\right)},$$

where the denominator seems to suggest some type of mean coefficient of variation as occurs in other queueing problems (e.g., the Pollaczek-Khinchin formula). Unfortunately, we have yet to find a way to exploit this relation and in the remainder of this paper, we make no more of it.

For $p = 1/n$ the correlation vanishes for all r. More importantly, it follows from Szekli, Disney and Hur (1994b) that the arrival process is a non-delayed renewal process with an interrenewal hyperexponential distribution. For $p = 1$, the embedded Markov chain consists of closed sets of states and thus is not irreducible. Therefore we need $p < 1$. We will require positive correlation so we need $1/n < p < 1$.

(d) The queue length process has a traffic intensity

$$\rho = \left[\frac{1}{n} \sum_{j=1}^{n} \frac{1}{\rho_j}\right]^{-1},$$

the harmonic mean of the individual traffic intensities. A steady state queue length exists if $\rho < 1$ (see Szekli, Disney and Hur (1994b)).

(e) For the 2-state case let N^a be the steady state queue length with mean value EN^a as seen by an arrival. Then, (Szekli, Disney and Hur 1994b)

$$EN^a = \frac{\rho}{1-\rho} + \frac{1 - \frac{2(1-P_0)}{\rho_1+\rho_2}}{2\,(1-p)\,(1-\rho)}, \tag{4}$$

for the case where F_j is exponential. Here, P_0 is the probability that an arriving customer of either type finds the queue empty. If the arithmetic mean $(\rho_1 + \rho_2)/2 > 1$ then $EN^a \to \infty$ for $p \to 1$ even when $\rho < 1$. *That is, even in the steady state case the mean queue length can become arbitrarily large due to correlations in the arrival process.* This is a consequence of the correlation, not of the distributions otherwise.

3 The Effects of the Parameters

The result (e) in Section 2.2 gives the behavior of EN^a for p, but this requires further explanation. That is the purpose of this section. From result (b) in Section 2.2 we see

that the correlation depends on 4 parameters (p, differences in the mean interarrival times of each type ($m_i - m_j$), the variance of the arrival times of each type (v_j), and the number of states n). In this section we show that these parameters interact so the p-effect that we have discussed is not the only effect on EN^a. In fact, we show that p and $m_i - m_j$ interact in such a way that the rate of convergence of EN^a to infinity is faster for large $m_i - m_j$ as a function of p. We also show that decreases in v_j *increase correlations* but can decrease EN^a.

3.1 The Effect of p

It should be noted that p is not only a measure of the correlation in (3), but p can also be thought of as a measure of the "stickiness" of the Markov chain. If p is near 1, that Markov chain will tend to stay in whatever state it finds itself at each step.

Consider the 2-state case. If $(\rho_1 + \rho_2)/2 < 1$, queues formed by either arrival process will have stationary distributions as will the overall queue since then $\rho < 1$. However, if $(\rho_1 + \rho_2)/2 \geq 1$, at least one of the arrival processes will produce a non-stationary queue (call this the fast arrival queue) while the other will still produce a queue with a stationary distribution. This is true even though the overall queueing process will produce a steady state queue. For example, $\rho < 1$ implies

$$\frac{\rho_1 + \rho_2}{2} > \rho_1 \rho_2.$$

Then, set $\rho_1 = 0.5$ and $\rho_2 = 10$ to get

$$\frac{\rho_1 + \rho_2}{2} = 5.25 \text{ but } \rho_1 \rho_2 = 5.$$

Then the system is in steady state ($\rho < 1$), the queue of type-1 arrivals is in steady state ($\rho_1 < 1$) but the queue of type-2 arrivals is transient. Therefore by 2.2(e) the mean queue length will be large even though the system is in steady state. Also from that section, larger values of p will generate larger values of the mean queue length in this case. If p is large, the embedded Markov chain will tend to stick in the fast arrival queue for long periods and the queue caused by those arrivals will increase without limit, even though the overall queue will still have a steady state distribution. Of course, the queue due to the slower arrivals compensates so as to make these results possible.

To illustrate this behaviour further, we investigate now the queue length as seen at the arrival of different types of customer. First note that from Neuts (1978), the state distribution P_k as seen by an arrival is:

$$P_k = \Pi(I - R)R^k,$$

where Π is the statioinary distribution of the embedded Markov chain and R is the solution of

$$R = \sum_{n=0}^{\infty} R^n A_n, \qquad A_n = \int_0^{\infty} \frac{e^{-\mu t}(\mu t)^n}{n!} d\mathbf{Q}(t).$$

For exponential distributions i.e., $F_j(t) = 1 - e^{-\lambda_j t}$ and two or three states these matrix equations can be solved numerically. Calculating to the precision available in MATLAB required approximating the infinite sum in the first equation by a finite sum of (about 100) terms. The solution then converges in about 100 iterations. The number of iterations required increases with the traffic intensity, and with the serial correlation of the arrival process. Checks on the numerical accuracy are available, for example if we set $p = 1/n$ this gives a hyperexponential arrival process.

Consider the state distribution at the end of a sequence of type-1 (say slow) arrivals. Since the type of the next arrival is decided randomly this is in fact the distribution as seen by an arbitrary type-1 arrival, so the k-th term is $\{P_k\}_1/\Pi(1)$.

From then on, type-2 arrivals occur as a renewal sequence until the type changes again. Regarding this system during this sequence as a GI/M/1 queue, the transition matrix for the imbedded Markov chain of the number in the system as seen by an arriving customer has the form:

$$\mathbf{P} = \begin{bmatrix} s & 1-s & 0 & 0 & \cdots \\ s^2 & (1-s)\,s & 1-s & 0 & \cdots \\ s^3 & (1-s)\,s^2 & (1-s)\,s & 1-s & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where $s = \mu/(\lambda_2 + \mu)$. Writing $\mathbf{p}^0 = \{p_0, p_1, \ldots\}$ for the state distribution as seen by the first type-2 arrival, then the state distribution as seen by the $(n+1)$st type-2 arrival is:

$$\mathbf{p}^n = \mathbf{p}^0 \mathbf{P}^n.$$

So

$$\begin{aligned} p_0^{'} &= p_0 s + p_1 s^2 + p_2 s^3 + \cdots \\ p_1' &= p_0\,(1-s) + p_1\,(1-s)\,s + p_2\,(1-s)\,s^2 + \cdots \\ &= p_0^{'}\left(\frac{1-s}{s}\right), \\ p_2^{'} &= p_1\,(1-s) + p_2\,(1-s)\,s + p_3\,(1-s)\,s^2 + \cdots \\ &= \left(p_1^{'} - p_0\,(1-s)\right)/s, \\ p_3^{'} &= p_2\,(1-s) + p_3\,(1-s)\,s + p_4\,(1-s)\,s^2 + \cdots \\ &= \left(p_2^{'} - p_1\,(1-s)\right)/s, \end{aligned}$$

and so on, which gives us a general algorithm to go from $\mathbf{p}^n$ to $\mathbf{p}^{n+1}$.

In fact if $\lambda_2 \geq \mu$ these probabilities have to be calculated by direct summation, as the difference expressions do not converge. However it is easy enough to calculate the distribution by summing the expressions out to about 20 terms and hence calculate the mean queue length as seen by the first fast arrival, second fast arrival and so on.

A check on the calculation is possible in that provided $\lambda_2 < \mu$ as the mean number should converge to the appropriate M/M/1 value.

Fig. 3.1 is the best way of illustrating these effects that we have found. We start from the state distribution as seen by the last customer of type 1. This is a representation of the kind of thing we could expect to see. The graph then plots the mean number in the system as seen by the first, second, third etc. type-2 arrivals. Since $p = 0.8$ we could expect the arrival process to stay in a particular state for $1/(1-p) = 5$ arrivals, so we plot out for these 5 arrivals. Since $\rho_2 > 1$ the mean number in the system increases almost linearly. Now starting from the distribution of the number of customers as seen by the last fast arrival, we plot the mean number in the system as seen by the first 5 slow arrivals.

Not surprisingly the mean number as seen by the fifth (average number) of fast arrivals is almost exactly the mean number at the end of an average fast arrival time (note that there are two almost identical points at the peaks of the graph.) But of course the mean number after exactly 5 slow arrivals does not coincide with the mean number at the end of an average slow arrival period, since the response of of the overall mean queue length, L, is non-linear for $\rho_1 < 1$.

Also plotted on the graph is the mean number as seen by an arrival to an $H_2/M/1$ queue with parameters $\lambda_1 = 0.3,\ \lambda_2 = 2.0$, mixing probability 0.5 and $\mu = 1$.

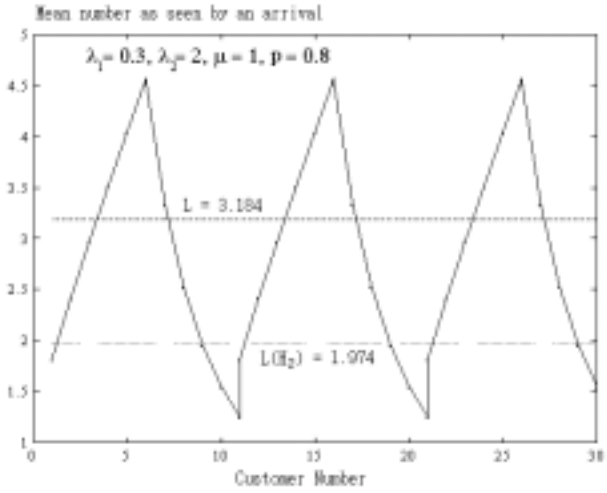

Fig. 3.1 - An illustration of the effect of fast and slow arrivals

The graph illustrates the conditions under which the mean number in the system can become unbounded as $p \to 1$. If both $\rho_1 < 1$ and $\rho_2 < 1$ then even during the

fast arrivals the mean number in the system is still bounded. The graph tends up towards the value of L for an M/M/1 queue with parameters λ_2 and μ during a fast arrival period, but cannot exceed it.

If, on the other hand one of ρ_1 or $\rho_2 > \mu$ (say $\rho_2 > 1$), then the average number of arrivals during a fast arrival period will increase without limit as $p \to 1$, so the upwards climb of the mean number in the system is also unlimited (at least in distribution) So the condition for the possibility of unbounded mean number in the system is that either ρ_1 or $\rho_2 > 1$. The stronger condition in the Theorem in Section 2.2(e), of course, implies this condition.

3.2 The Mean Value Effects

Without loss of generality, we hereafter assume $m_2 - m_1 > 0$. The condition $(\rho_1 + \rho_2)/2 > 1$ in result (e) in Section 2.2 can be restated in terms of the difference of mean values as the following theorem shows:

Theorem 1: For given $\rho < 1$, the condition $(\rho_1 + \rho_2)/2 > 1$ will occur when the difference $m_2 - m_1$is large enough.

Proof. We have $m_1 + m_2 = 2/(\mu\rho)$. Now, the condition $(\rho_1 + \rho_2)/2 > 1$ is equivalent to $\frac{1}{m_1} + \frac{1}{m_2} > 2\mu$ and in turn,

$$m_1 m_2 < \frac{m_1 + m_2}{2\mu} = \frac{1}{\mu^2 \rho}.$$

For convenience, let $m_2 - m_1 = d > 0$. Then $m_2 = \frac{1}{2}\left(\frac{2}{\mu\rho} + d\right)$ and $m_1 = \frac{1}{2}\left(\frac{2}{\mu\rho} - d\right)$. So the equation above becomes

$$\frac{1}{4}\left(\left(\frac{2}{\mu\rho}\right)^2 - d^2\right) < \frac{1}{\mu^2\rho}.$$

Rearranging both sides and noting $d > 0$, we have

$$d > \frac{2}{\mu\rho}\sqrt{1-\rho} = 2m\sqrt{1-\rho} \quad \text{where} \quad m = (m_1 + m_2)/2,$$

which completes the proof. □

Now we can explain the queueing behaviour in terms of $m_2 - m_1$ and ρ as follows. We say (m_1, m_2) is in a "critical region" if $m_2 - m_1 > c$ where $c = 2m\sqrt{1-\rho}$, in the sense that the mean queue length increases to infinity if ρ is close to one. Denote the critical region by

$$C = \left\{ m_2 - m_1 : m_2 - m_1 > 2m\sqrt{1-\rho} \right\}.$$

Since we assumed $m_2 > m_1$,

$$0 < m_1 < m < m_2 < 2m.$$

There is no chance for the mean queue length to be infinitely large if $\rho_1 < 1$ and $\rho_2 < 1$, that is,

$$m_1 > m\rho, \text{ and } m_2 > m\rho.$$

Combining the last two results we have

$$0 < m_2 - m_1 < (2m - m\rho) - m\rho = 2m\,(1-\rho)\,.$$

So we say (m_1, m_2) is in a "safety region" when $0 < m_2 - m_1 < 2m\,(1-\rho)$, in the sense that the mean queue length never grows to infinity. Denote the safety region by

$$S = \{m_2 - m_1 : 0 < m_2 - m_1 < 2m\,(1-\rho)\}\,.$$

We now use a numerical example to illustrate the joint effect of $m_2 - m_1$ and p and see how the critical and safety regions move as the traffic intensity varies. We consider an MR/M/1 queue where F_1 and F_2 are exponential with means m_1 and m_2, respectively. We fix the mean value m of the marginal interarrival times to be 5. Since $m_1 + m_2 = 10$, let $m_2 > m_1$, $5 < m_2 < 10$ and $0 < m_1 < 5$. Then $c = 10\sqrt{1-\rho}$ and thus the critical region is $C = \left\{m_2 - m_1 : m_2 - m_1 > 10\sqrt{1-\rho}\right\}$. The safety region is $S = \{m_2 - m_1 : 0 < m_2 - m_1 < 10\,(1-\rho)\}$.

The following table summarizes the critical and safety regions for different traffic intensities.

ρ	Safety Region (S)	Critical Region (C)
0.1	$0 < m_2 - m_1 < 9$	$9.49 < m_2 - m_1 < 10$
0.5	$0 < m_2 - m_1 < 5$	$7.07 < m_2 - m_1 < 10$
0.9	$0 < m_2 - m_1 < 1$	$3.16 < m_2 - m_1 < 10$

Therefore, if the traffic intensity gets larger, then the critical region C becomes wider, while the safety region S shrinks. On the contrary, if the traffic intensity gets smaller, we have a smaller critical region and a larger safety region.

Thus, we conclude that under heavy traffic, we have more chance for the queue length to be arbitrarily large, while under light traffic, the queue tends to be stable. Note that between the critical and the safety regions the behaviour of the queue has not been identified, which is believed to be possible after P_0 is found. We have shown then that the increase of the correlation coefficient in the arrival process via the parameter p, jointly with $m_2 - m_1$, the distance between mean values, can make the queue quite unstable.

3.3 The Variance Effects

As we have seen in the previous two sections, the parameter p and the differences in the mean values can have major effects on the mean queue length and the mean waiting times. Our purpose in this section is to show that large increases in mean queue length are not necessary consequences of the correlation coefficient. We demonstrate this by showing that increases in the correlation coefficient caused by decreases in the variances can cause decreases in the mean queue length. We state this as a theorem and follow it with a numerical example.

Consider two stationary Markov renewal arrival processes to a single server queue with i.i.d. exponentially distributed service times. Symbolize these arrival processes as $[A_n(p), \mathbf{F}]$ and $[A_n(p), \mathbf{F}']$, where $A_n(p)$ is the transition matrix for the embedded Markov chain (1) and $\mathbf{F} = [F_1, F_2, \cdots, F_n]$. Then we have

Theorem 2: Suppose that for each j $\in E$, $F_j \leq_{icx} F'_j$ and the corresponding expected values are the same (see section 2 for the definitions and consequences here) then

$$\text{Corr}[A_n(p), \mathbf{F}] \geq \text{Corr}[A_n(p), \mathbf{F}']$$

and

$$EW[A_n(p), \mathbf{F}] \leq EW[A_n(p), \mathbf{F}'].$$

Proof. The proof of the second part of the theorem is a direct consequence of Rolski (1983). The first part is a simple observation on (2) in Section 2.1. □

Consider the discussion of the increasing convex property in the beginning of section 2. There we showed that with the properties given to the distributions assumed here, the variances in the $[A_n(p), \mathbf{F}]$ are smaller than those in the $[A_n(p), \mathbf{F}']$ process. The result here simply says that decreasing the variance while increasing the correlation coefficient *decreases* the mean waiting time and consequently the mean queue length. Here's a numerical example of the variance effect.

Consider the Erlang density function

$$f(x) = k\lambda(k\lambda x)^{k-1} e^{-k\lambda x}/(k-1)!,\ \lambda > 0,\ k = 1, 2, \cdots$$

By keeping the λ fixed but increasing k the variance decreases. Now let the arrival process to a queue be a 2-state Markov renewal process with an underlying Markov chain whose one step matrix is $A_2(0.85)$, i.e.,

$$A = \begin{pmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{pmatrix}$$

and F_1 be the Erlang distribution function with $\lambda = 1/3,\ k = 1, 2, \cdots$ arbitrary. Let F_2 be the Erlang distribution function with $\lambda' = 1/7$ and the same k as in F_1. Let $\rho = 0.5$. Then the lag 1 correlation coefficient is

$$\text{Corr} = \frac{11.2k}{16k + 116}$$

which is increasing and concave in k. For $k^{'} \geq k$ we have the convexity property of Definition 1 and hence $EW(k) \geq EW(k^{'})$. Then by Little's result, $L^t(k) \geq L^t(k^{'})$, that is L^t decreases in k as the following Fig. 3.2 illustrates.

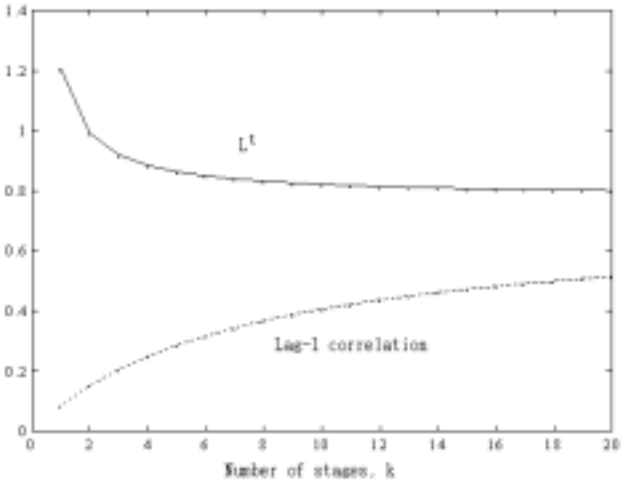

Fig. 3.2 - The behavior of mean queue length and lag-1 correlation as k changes

3.4 The Effect of the Number of States

We will now investigate the effect of the parameter n, the number of different types of arrivals (call this the *dimension* of the arrival process). If we change this dimension keeping the marginal interarrival times distributions fixed and look at the change of the correlation coefficient and performance measures (e.g., mean queue length), then we can see the pure effect of changing the dimension of the arrival process on the queueing properties. This effect may reflect the feasibility of grouping and reducing the number of different customer types to be considered in a system's design.

Consider two n and n' dimensional MRP's denoted by $[A_n(p), \mathbf{F}]$ and $[A_{n'}(p'), \mathbf{G}]$, where $\mathbf{F} = [F_1, \cdots, F_n]$ and $\mathbf{G} = [G_1, \cdots, G_{n'}]$. We need two conditions to extract pure effects due to the changing of the dimension. We'll call them (i) the same marginal condition and (ii) the same eigenvalue condition, to be explained below.

(i) Same Marginal Condition

We suppose the marginal distribution of the interarrival times of the two MRP's

are the same, that is for all k and $t \in R_+$

$$P(D_k \leq t) = \frac{1}{n}\sum_{i=1}^{n} F_i(t) = \frac{1}{n'}\sum_{i=1}^{n'} G_i(t).$$

We have not been able to satisfy this in general but if n' is an integer multiple of n (or vice versa), then we can achieve the condition by letting $\mathbf{F}$ be a mixture of $\mathbf{G}'$s. For example, if $n = 2$ and $n' = 6$, then let $F_1 = \frac{1}{3}(G_1 + G_2 + G_3)$ and $F_2 = \frac{1}{3}(G_4 + G_5 + G_6)$ to obtain

$$P(D_k \leq t) = \frac{1}{2}\left(F_1 + F_2\right)(t) = \frac{1}{6}\left(G_1 + \cdots + G_6\right)(t).$$

(ii) Same Eigenvalue Condition

To get the subdominant eigenvalues the same (see Section 2.2) we'll require

$$\xi_p(n) = \frac{np-1}{n-1} = \frac{n'p'-1}{n'-1} = \xi_{p'}(n').$$

This condition can be achieved by adjusting p and p', for given n and n'. Then we ask: Is $\mathrm{Corr}(n) < \mathrm{Corr}(n')$ when $n < n'$? The answer is "yes" whenever n' is an integer multiple of n (a proof is given in Hur(1993)). As a consequence if an MRP has more types of customers than another, under conditions (i) and (ii), the queue with more types is more correlated than the other. It is, however, not true that $\mathrm{Corr}(n) < \mathrm{Corr}(n')$ for arbitrary n and n'. Counterexamples can be constructed as the following illustrates. Let $n = 2$ and $n' = 3$ and the mean sojourn times be (4, 6) and (1, 6, 8) respectively. Then,

$$\frac{\mathrm{Corr}(2)}{\mathrm{Corr}(3)} = \left(\frac{3}{2}\right)^2 \frac{(4-6)^2}{(1-6)^2 + (6-8)^2 + (6-8)^2} = \frac{36}{312} < 1.$$

Thus, $\mathrm{Corr}(2) < \mathrm{Corr}(3)$, but with mean sojourn times (1, 9) and (1, 6, 8) the corresponding result is $\mathrm{Corr}(2) > \mathrm{Corr}(3)$. To avoid such a result in a study of pure effects of dimensionallity, conditions (i) and (ii) are necessary.

To further study the effects of dimensionality we turn to a numerical procedure. We start with a 48-dimensional MRP and keep mixing the distribution functions pairwise to get 24, 12, 6, and 3-dimensional MRP's. For the 48-dimensional process we take $F_j(t)$ to be an Erlang distribution with mean $m_j = 2j/49$ and $j = 1, 2, \ldots 48$. The m_j are chosen to make the overall mean value of the arrival process, m, to be 1. Then, the 24-dimensional MRP $W[A_{24}(p'), \mathbf{G}]$ is constructed by taking

$$G_j(t) = \left(F_j + F_{24+j}\right)/2, \quad \text{for} j = 1, 2, \cdots, 24$$

and p' so that

$$\xi(p) = \frac{48p-1}{47} = \frac{24p'-1}{23} = \xi(p')$$

so that conditions (i) and (ii) are satisfied.

Thus, we can produce a sequence of 5 Markov renewal cases, each with the same marginal distributions and the same subdominant eigenvalue. From the results in Section 2.2, the correlation is changing only through the change in the number of states.

The mean number in the system at an arbitrary time in equillibrium was calculated as described in Section 3.1. For low traffic intensities L^t increases weakly with dimensionality. However, when the traffic intensity is high, the effects are more striking. Fig. 3.3 plots L^t against the number of states for traffic intensity of 0.9. The five curves are for values of the subdominant eigenvalue ξ =0.1, 0.3, 0.5, 0.7, 0.9. Hence, they correspond, in terms of the probability p to a range from "very sticky" ($\xi = 0.9$) to "not very sticky" ($\xi = 0.1$).

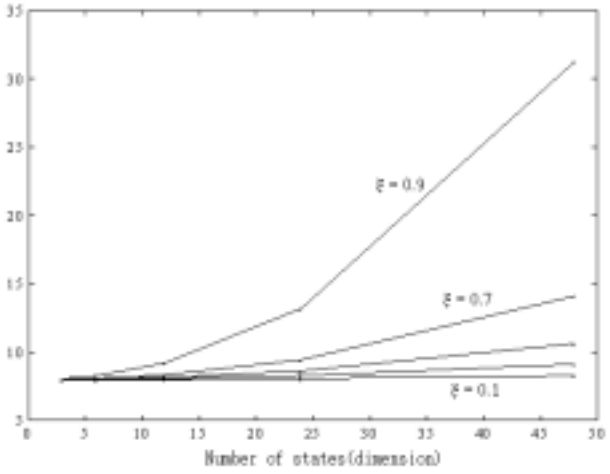

Fig. 3.3 - Mean queue length as the dimension increases

Thus, dimensionality, acting through the correlation coefficient can have additional effects on the mean queue length especially in the case of "sticky" (large values of ξ) of MRP's.

3.5 The Influence of the Index of Dispersion for Intervals (IDI)

The Index of Dispersion for Intervals (IDI) is defined as

$$C_\infty^2 = \frac{\mathrm{Var}(D_r)}{[E(D_r)]^2}(1 + 2\sum_{r=1}^{\infty} \mathrm{Corr}(r)).$$

Thus, the IDI takes account of the lagged correlation coefficient for all lags (see Sriram and Whitt(1986)). In our case this is easy to compute due to the special structure of the correlations as given in Section 2.2(b). From those results we have that the IDI is

$$C_\infty^2 = \frac{2(\lambda_1^2 + \lambda_2^2)}{(1-p)(\lambda_1 + \lambda_2)} - \frac{p}{1-p}$$

In the following graphs (Fig. 3.4-3.6) we have computed the value of L^t against the IDI for $\lambda_2 = 0.3, 0.6, 2.1$ and λ_1 has been chosen to keep the traffic intensity at 0.2. The values of p are of the form 0.3, 0.4, etc.

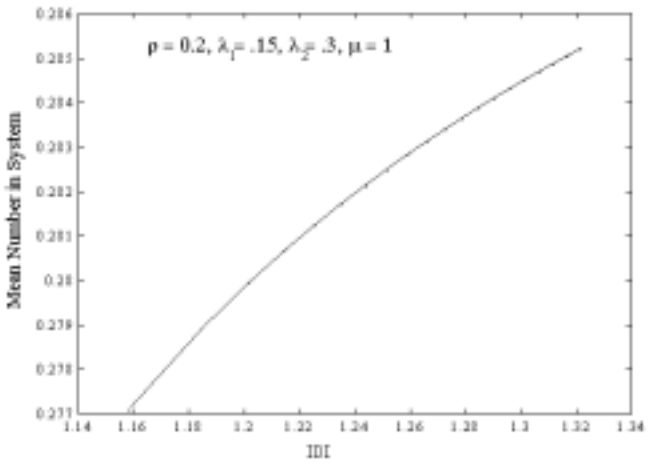

Fig. 3.4 - The behavior of maen queue length versus IDI ($\lambda_2 = 0.3$)

What is clear from here is that the near linearity of the mean queue length versus the IDI is due to either ρ_1 or $\rho_2 > 1$ and is not due to a large traffic intensity as previously supposed (Patuwo, Disney and McNickle (1993)).

4 Comments

There is a number of conclusions that one can draw from this study. Perhaps the most important is that correlations can have major effects on queueing properties as reasonable as the mean queue length.

(a) We have shown that correlation alone as determined by p can have a major effect (2.2(e)). It would appear that more attention should be paid to the correlations especially in situations where the traffic intensity is high and where correlations can be present and can be large. One can suppose that the true effect here is dependence and correlation is a poor measure of this dependence in these

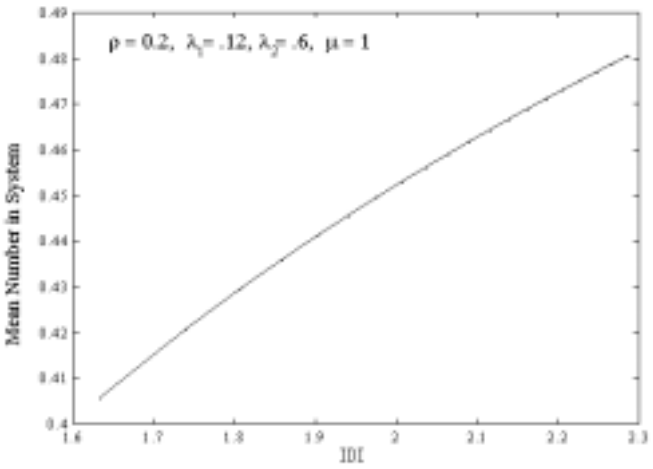

Fig. 3.5 - The behavior of maen queue length versus IDI ($\lambda_2 = 0.6$)

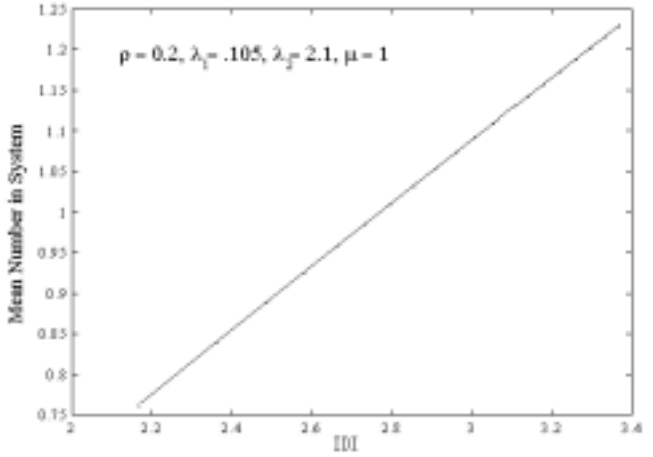

Fig. 3.6 - The behavior of maen queue length versus IDI ($\lambda_2 = 2.1$)

non-linear systems. Perhaps other measures of dependence are needed.

(b) Since the harmonic mean is always greater than the arithmetic mean, it is possible for the queue length process to be in steady state but the mean queue length to be arbitrary large solely due to correlations (because of p)(2.2(e)).

(c) In Section 2.2(e) EN^a appears to be made up of a term that is the mean queue length for an M/M/1 queue plus a term that depends on the parameter p. That extra term needs to be investigated. We have not been able to find P_0 explicitly so we do not know analytically how the extra term responds to changes in the system's parameters.

(d) The MRP arrival process depends on 4 parameters (p, differences in the mean interarrival times of each type ($m_i - m_j$), the variance of the arrival times of each type (v_j), and the number of states (n). Early investigations show that these parameters interact so the p-effect that we've discussed is not the only effect. In fact, p and $m_i - m_j$ interact in such a way that the rate of convergence of EN^a to infinity is faster for large $m_i - m_j$ as a function of p. But we can also show that decreases in v_j increase correlations but can decrease EN^a.

(e) The linear dependence of the queue characteristics on the IDI, which we had previously characterised as a heavy traffic property, rather appears to depend only on the property that the queue will occasionally move to a busy state.

Acknowledgements

The first author was supported during the preparation of this paper by an Erskine Foundation Fellowship at the University of Canterbury, Christchurch, New Zealand and by a grant from the Department of Statistics, Massey University, Palmerston North, New Zealand. Part of the paper was done under a Work-Study Leave Program of his home university.

The fourth author was supported in part by CNPq grant #500.756/90-2.

The fifth author was supported in part by grants from the Kosciuszko Foundation and KBN.

5 References

[1] Hasslinger, G. and M. Adam, "Modelling and performance analysis of traffic in ATM networks including autocorrelations", *Proceedings of IEEE INFOCOM,* (1996), 1460-1467.

[2] Hur, S., *The effect of positively correlated arrivals on the single server queue,* Ph.D. Disseration, Dept of Industrial Engineering, Texas A&M University, College Station, TX., 1993.

[3] Neuts, M.F., "Markov chains with applications in queueing theory, which have a matrix-geometric invariant probability vector", *Adv. Appl. Probab.*, 10, (1978), 185-212.

[4] Patuwo, B.E., R.L. Disney and D.C. McNickle, "The effect of correlated arrivals on queues", *IIE Transactions*, 25, (1993), 105-110.

[5] Paxson, V. and S. Floyd, "Wide-area traffic: the failure of Poisson modelling", *Proceedings of ACM SIGCOM* (London), (1994), 257-268.

[6] Rolski, T., "Comparison theorems for queues with dependent interarrival times: Modelling and Performance Evaluation", In: *Proceeding of the International Seminar*, Paris, (1983), 42-67.

[7] Sriram, K. and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data", *IEEE J. on Selected Areas in Comm.*, SAC-4, (1986), 833-846.

[8] Stoyan, D., *Comparision Theorems for Queues and Other Stochastic Models,* John Wiley and Sons, Berlin, 1983.

[9] Szekli, R., R.L. Disney, and S. Hur, "On performance comparison of MR/G/1 queues", *QUESTA*, 17, (1994a), 451-470.

[10] Szekli, R., R.L. Disney, and S. Hur, "MR/GI/1 queues with positively correlated arrival streams", *J. Appl. Probab.*, 31, (1994b), 497-514.

[11] Willinger, W., M. Taqqu, W. Leland and D. Wilson, "Self-similarity in high-speed packet traffic: analysis and modelling of ethernet traffic measurements", *Statistical Science,* 10, (1995), 67-85.