# The Impact of Autocorrelation on Queuing Systems

*Miron Livny †, Benjamin Melamed ‡ and Athanassios K. Tsiolis †*

† Department of Computer Sciences
University of Wisconsin-Madison
Madison, Wisconsin

‡ NEC USA, Inc.
C&C Research Laboratories
Princeton, New Jersey

*ABSTRACT*

The performance of single-server queues with independent interarrival intervals and service demands is well understood, and often analytically tractable. In particular, the M/M/1 queue has been thoroughly studied, due to its analytical tractability. Little is known, though, when autocorrelation is introduced into interarrival times or service demands, resulting in loss of analytical tractability. Even the simple case of an M/M/1 queue with autocorrelations does not appear to be well understood. Such autocorrelations do, in fact, abound in real-life systems, and worse, simplifying independence assumptions can lead to very poor estimates of performance measures. This paper reports the results of a simulation study of the impact of autocorrelation on performance in a FIFO queue. The study used two computer methods for generating autocorrelated random sequences, with different autocorrelation characteristics. The simulation results show that the injection of autocorrelation into interarrival times, and to a lesser extent into service demands, can have a dramatic impact on performance measures. From a performance viewpoint, these effects are generally deleterious, and their magnitude depends on the method used to generate the autocorrelated process. The paper discusses these empirical results and makes some recommendations to practitioners of performance analysis of queuing systems.

January 28, 1992

# The Impact of Autocorrelation on Queuing Systems

*Miron Livny †, Benjamin Melamed ‡ and Athanassios K. Tsiolis †*

† Department of Computer Sciences
University of Wisconsin-Madison
Madison, Wisconsin

‡ NEC USA, Inc.
C&C Research Laboratories
Princeton, New Jersey

## 1. INTRODUCTION

Most queuing models assume that customer interarrival times, service demands, and routing probabilities are independent, each being modeled as a renewal process (see, e.g., [Kell79]). These assumptions lead to models that are easy to simulate, and under suitable restrictions are analytically tractable. Unfortunately, these models are often poor representations of real-life systems where correlations do, in fact, abound. A case in point is the emerging technology of broadband ISDN (Integrated Services Digital Networks) which promises to be an important application domain for analysis of autocorrelated traffic. In particular, emerging transport mechanisms such as ATM (Asynchronous Transfer Mode) are slated to carry a broad mix of packet (cell) traffic, ranging from data to voice and video. File transfers and full motion video frames are known to be extremely bursty, and consequently the induced autocorrelations in arrival streams coupled with the enormous speeds of transmission are going to aggravate the problems of congestion control at the nodes and bandwidth allocation on the links.

Various studies have found that models that do not take autocorrelation into account can predict overly optimistic performance measures, such as line lengths and waiting times. Some representative references are now briefly surveyed.

Perhaps the earliest work on non-renewal queuing is due to J. Th. Runnenburg, some thirty years ago. Waiting time behavior is investigated in [Runn61] for a simple integer-valued Markovian arrival process offered to an exponential renewal server. A 2-state special case appears in [Runn62], where the waiting time distribution is studied, parameterized by the first autocorrelation; the difference relative to renewal

arrivals (Poisson traffic) is also illustrated. Recently, [Tin85] investigated the queue size and waiting time processes in a queuing system with independent exponential services and gamma interarrival intervals with a particular Markovian structure (see also references therein). More recently, [Patu93] studied queuing behavior engendered when Markov renewal traffic is offered to an exponential renewal server. A popular scheme is described in [Heff86], where the autocorrelations are generated by a Markov Modulated Poisson Process. This model has been widely used in telephony-based modeling to characterize source models of voice traffic. [Jaco80] analyzes a single-server queue with arrivals and services generated by a mixed exponential autoregressive moving average (EARMA) scheme. The arrival and service processes are positively autocorrelated and cross-correlated, and the performance measures are approximated under heavy traffic conditions. Another heavy traffic study of a single queue with correlated arrivals and services is reported in [Fend89]. This paper introduces a three dimensional measure called the Index of Dispersion for Intervals (IDI) which captures the variability of arrivals and services induced by autocorrelation and cross-correlation. A recurrent theme in these papers is the dramatic degradation of performance induced by increasing correlations.

In this paper we report the results of a study in which we investigate the impact of autocorrelated interarrival intervals and autocorrelated service demands on the performance of a single-queue single-server system under a variety of traffic loads. The study was motivated by two objectives: to profile the sensitivity of queuing models to the assumption that interarrival intervals and service demands are independent random variables, and to understand the power of correlated variates as a modeling tool that can capture the behavior of a broader range of systems including bursty arrival streams (see [Heff73, Heff80]). Our study is in the same vein as [Runn61, Runn61, Jaco80, Tin85, Fend89] in that it investigates a single server queue with exponential autocorrelated interarrivals and services. In effect, we start with an M/M/1 queue and study the mean waiting time performance as autocorrelations are increasingly injected into arrivals and services, but cross-correlations are kept at zero. Our study confirms the heavy traffic results in [Fend89] but it goes beyond those results to show the dramatic effect of correlations even in light and moderate traffic scenarios. We also conclude that positive autocorrelations always lead to degraded performance (in our case, mean waiting times), but the effect of negative correlations is not clear cut; its effect is

a more complicated function of the autocorrelation function, at least when successive autocorrelations alternate in sign.

Since analytical methods can only handle the analysis of quite restricted classes of autocorrelated time series, we used simulation to generate the results presented in this paper. We implemented two different methods for generating autocorrelated random sequences: the TES (Transfer-Expand-Sample) method [Mela91, Jage92, Jage92a] and the Minification method [Lewi91]. These methods give rise to the same marginal distribution, but their resultant autocorrelation functions have very different functional forms; more specifically, the autocorrelation tail of the Minification method decays to zero much faster than its TES counterpart. Furthermore, Minification generates monotone decreasing autocorrelation functions, whereas TES can generate both monotone and oscillatory ones. We, therefore, argue that our conclusions are more robust with respect to dependence, since obviously the autocorrelation structure does not generally determine the joint probability law of a stochastic process. Both generation methods were subsequently integrated with the DeNet simulation environment (see [Livn88]). This enabled us to generate autocorrelated variates with a wide variety of marginal distributions and autocorrelation structures, and to use them in simulation experiments.

The rest of this paper is organized as follows: Section 2 overviews the notions of correlation and autocorrelation. Section 3 describes briefly TES methods and Minification methods. Sections 4, 5 and 6 present qualitative descriptions of autocorrelation in the arrival process, in the service process and in both the arrival and service processes, respectively. Section 7 summarizes the results of a simulation study of the impact of autocorrelation on the waiting times of a FIFO server. Finally, Section 8 contains our conclusions and a brief discussion of our ongoing work in the area of autocorrelated queuing.

## 2. THE NOTIONS OF CORRELATION AND AUTOCORRELATION

The central theme of this paper is the effect of autocorrelation on performance measures in a queuing context. We, therefore, begin with a brief explanation of correlation and related concepts, mostly at an intuitive level.

Correlation is a measure of linear dependence between random variables (variates). The correlation

coefficient $\rho_{X,Y}$ of real variates $X$ and $Y$ is defined by:

$$\rho_{X,Y} = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y},$$

where $\sigma_X$ and $\sigma_Y$ are the standard deviations of $X$ and $Y$, respectively (we assume, of course, that the standard deviations and expectations are finite). Note that $\rho_{X,Y} = \rho_{Y,X}$ and that $\rho_{X,X} = 1$, whereas $\rho_{X,-X} = -1$. It can be shown that $-1 \le \rho_{X,Y} \le 1$. Furthermore, for the extremal cases ($\rho_{X,Y} = \pm 1$), $X$ and $Y$ are related by a linear equation

$$Y = aX + b$$

for some real constants $a \neq 0$ and $b$; for $a > 0$, we have $\rho_{X,Y} = 1$ and for $a < 0$, we have $\rho_{X,Y} = -1$.

Intuitively, the presence of correlation ($\rho_{X,Y} \neq 0$) means that $X$ and $Y$ tend to vary in the same linear direction ($\rho_{X,Y} > 0$), or in the opposite linear direction ($\rho_{X,Y} < 0$); the magnitude $|\rho_{X,Y}|$ is a measure of how strong the linear tendency is. This tendency is guaranteed for $\rho_{X,Y} = \pm 1$, but weakens progressively as $|\rho_{X,Y}|$ decreases to zero, until finally, this tendency disappears altogether. Note, though, that lack of correlation and lack of independence are distinct concepts. It is possible for $X$ and $Y$ to be uncorrelated ($\rho_{X,Y} = 0$) and yet be dependent. However, independence always implies lack of correlation. For this reason one can think of correlation as linear dependence, i.e., a restricted notion of dependence.

Operationally, suppose we sample pairs $(x_n, y_n)$ from the joint distribution of $X$ and $Y$. Under positive correlation, large values of $x_n$ will tend to occur with large values of $y_n$, and small $x_n$ with small $y_n$ (variation in the same direction). When the pairs $(x_n, y_n)$ are plotted in two dimensions, the human eye can discern a linear ascending pattern in the resulting picture (called a scattergram). Conversely, under negative correlation, large values of $x_n$ will tend to occur with small values of $y_n$, and small $x_n$ with large $y_n$; the resulting scattergram will then show a descending linear pattern. In the absence of correlation, any linear pattern vanishes and the scattergram looks like a random cloud or blob.

Consider now a stationary, real-valued time series $\{X_n\}$, $n = 0, \ldots, \infty$, in discrete time. By stationarity, the pairs $(X_n, X_{n+\tau})$ have the same distribution as $(X_0, X_\tau)$. The autocorrelation function $\rho_X(\tau)$ is the correlation coefficient of $X_0$ and $X_\tau$, that is,

$$\rho_X(\tau) = \rho_{X_0, X_\tau} = \frac{E[X_0, X_\tau] - \mu_X^2}{\sigma_X^2}, \quad \tau = 0, 1, \ldots$$

where $\mu_X$ is the common mean and $\sigma_X^2$ the common variance of the $X_n$. The argument $\tau$ is called the lag, since it denotes the time separation between the variates $X_0$ and $X_\tau$. Each number $\rho_X(\tau)$ is referred to as the lag-$\tau$ autocorrelation. In particular, $\rho_X(0) = 1$, always.

The notion of autocorrelation can help describe visual properties of time series. For example, if $X_n$ is the $n$-th interarrival interval in an arrival process, then positive autocorrelations will tend to create alternating runs of short interarrival intervals and long interarrival intervals. Visually, the former would correspond to arrival bursts, whereas the latter would appear as light traffic periods.

## 3. METHODS FOR GENERATING AUTOCORRELATED VARIATES

Since analytical methods can only handle the analysis of quite restricted classes of autocorrelated time series, the need arises to supplement analytical methods with simulation. This in turn calls for devising computer methods for generating autocorrelated variates with a wide variety of marginal distributions and autocorrelation structures. In this section we briefly review two methods for generating autocorrelated variates with exponential marginals. Both methods are quite general in that they give rise to uniform variates which are then transformed to other distributions (in our case to exponentials) using the inversion method (see [Devr86, Brat87]). The methods surveyed are TES (Transform-Expand-Sample) methods, and Minification/Maxification methods. The reader is referred to [Mela91, Jage92, Jage92a, Lewi91], and to references therein for a survey on recent work in this area.

### 3.1. TES Methods

Throughout this paper, $\{ Z_n \}$ will denote a sequence of i.i.d. (independent identically distributed) Uniform(0,1) variates, i.e., a stream of pseudo-random numbers available on most computers. The sequence $\{ U_n \}$, usually with various superscripts, will denote a c.i.d (correlated identically distributed) Uniform(0,1) sequence obtained from $\{ Z_n \}$ via suitable transformations. The target exponential variates will be similarly denoted by $\{ X_n \}$. All stochastic sequences are indexed by 0, 1, . . . , unless otherwise stated. Finally, for any real number $x$, let $\lfloor x \rfloor = \max\{ n \text{ integer} : n \le x \}$ be the integral part of $x$, and $\langle x \rangle = x - \lfloor x \rfloor$ be the fractional part of $x$.

TES methods were introduced in [Mela91] and investigated in detail in [Jage92] and [Jage92a]. A basic TES method is parametrized by a real pair $(L, R)$ such that $-0.5 < L < R \leq 0.5$. TES methods come in two flavors: TES$^+(L, R)$ giving rise to a sequence $\{U_n^+\}$, and TES$^-(L, R)$ giving rise to a sequence $\{U_n^-\}$. The reason for devising two parametrized methods is the need to achieve coverage of the full range of feasible lag-1 autocorrelations (correlation coefficients); in fact, the TES$^+(L, R)$ methods cover the positive range [0,1] while the TES$^-(L, R)$ methods cover the negative range [−1,0]. In retrospect, this fact motivates the superscript notation of TES, its attendant sequences, and other associated mathematical objects such as the autocorrelation functions $\rho_U^+(\tau)$ and $\rho_U^-(\tau)$, where $\tau$ is the lag argument.

The definition of $\{U_n^+\}$ is given recursively by

$$U_n^+ = \begin{cases} Z_0 & \text{if } n = 0 \\ \langle U_{n-1}^+ + L + (R - L) Z_n \rangle & \text{if } n > 0 \end{cases} \tag{3.1}$$

The sequence $\{U_n^-\}$ is defined in terms of $\{U_n^+\}$ by

$$U_n^- = \begin{cases} U_n^+ & \text{if } n \text{ even} \\ 1 - U_n^+ & \text{if } n \text{ odd} \end{cases} \tag{3.2}$$

Notice that $\rho_U^-(\tau) = (-1)^\tau \rho_U^+(\tau)$, so $\rho_U^-(\tau)$ has alternating sign if $\rho_U^+(\tau)$ is always positive.

Since both $\{U_n^+\}$ and $\{U_n^-\}$ have uniform marginals, they can be transformed to a wide variety of other marginal distributions via the inversion method. More specifically, let $D(x)$ be a mapping from [0,1] to the reals, referred to as a distortion. In particular, to obtain Exponential($\lambda$) marginals (where $\lambda$ is the reciprocal of the mean), the proper distortion is $D(x) = (-1/\lambda) \ln(1-x)$. Thus, if $\{U_n\}$ is Uniform(0,1), then the sequence $\{X_n\}$, where $X_n = D(U_n)$ will be Exponential($\lambda$). The Laplace-Stieltjes Transform of $D(x)$ will be denoted by $\tilde{D}(s)$.

For representing the autocorrelation functions, it is more convenient to switch from the $(L, R)$ parametrization to the equivalent parametrization $(\alpha, \phi)$, where

$$\alpha = R - L, \quad \phi = \frac{R + L}{\alpha} \tag{3.3}$$

Now, let $\{X_n^+\}$ be defined by $X_n^+ = D(U_n^+)$ and $\{X_n^-\}$ by $X_n^- = D(U_n^-)$. The common variance of the $X_n^+$ and $X_n^-$ is denoted by $\sigma_X^2$, and $\sqrt{-1} = i$. It can be shown that the autocorrelation function of $\{X_n^+\}$ has the representation

$$\rho_X^+(\tau) = \frac{2}{\sigma_X^2} \sum_{\nu=1}^{\infty} \left( \frac{\sin(\pi\nu\alpha)}{\pi\nu\alpha} \right)^{\tau} \cos(\pi\nu\alpha\phi\tau) \left| \tilde{D}(i2\pi\nu) \right|^2, \quad \tau = 0, 1, 2, \ldots \tag{3.4}$$

and for $\{ X_n^- \}$ we have the representation

$$\rho_X^-(\tau) = \begin{cases} \rho_X^+(\tau) & \tau \text{ even} \\ \dfrac{2}{\sigma_X^2} \displaystyle\sum_{\nu=1}^{\infty} \left( \frac{\sin(\pi\nu\alpha)}{\pi\nu\alpha} \right)^{\tau} \cos(\pi\nu\alpha\phi\tau) \, \mathrm{Re}[\, \tilde{D}(i2\pi\nu)^2 \,] & \tau \text{ odd} \end{cases} \tag{3.5}$$

(see [Jage92] for details).

Both autocorrelation representations, as well as the generation algorithms of $\{ U_n^+ \}$ and $\{ U_n^- \}$, are amenable to fast numerical computation on a computer. One advantage of TES methods is that they give rise to autocorrelation functions with various shapes. In particular, whenever $\phi = 0$, the resultant autocorrelation function is monotone decreasing to zero in $\tau$; however, for $\phi \neq 0$, one gets an oscillatory autocorrelation function in $\tau$ with envelopes that converge to zero.

### 3.2. Minification/Maxification Methods

The class of Minification and Maxification methods is described in [Lewi91]. These methods derive their name from the mathematical operations of obtaining the minimum (or maximum) of a set of real numbers. Minification/Maxification methods are parametrized by a single parameter $C > 1$. The corresponding sequences, denoted respectively by $\{ U_n^{\min} \}$ and $\{ U_n^{\max} \}$, are both Uniform(0,1).

The Minification sequence $\{ U_n^{\min} \}$ is defined recursively by

$$U_n^{\min} = \begin{cases} Z_0 & \text{if } n = 0 \\ C \cdot \min\left( U_{n-1}^{\min}, \dfrac{Z_{n-1}}{C - 1 + Z_{n-1}} \right) & \text{if } n > 0 \end{cases} \tag{3.6}$$

Its autocorrelation function has a particularly simple form

$$\rho_U^{\min}(\tau) = c^{-\tau}. \tag{3.7}$$

The maxification sequence $\{ U_n^{\max} \}$ is defined recursively by

$$U_n^{\max} = \begin{cases} Z_0 & \text{if } n = 0 \\ \max\left( (U_{n-1}^{\max})^C, Z_{n-1}^{\frac{C}{C-1}} \right) & \text{if } n > 0 \end{cases} \tag{3.8}$$

with a somewhat more complicated autocorrelation function

$$\rho_U^{\max}(\tau) = \frac{3}{2C^{\tau} + 1}. \tag{3.9}$$

Observe that Minification/Maxification methods give rise to autocorrelation functions which are strictly monotone decreasing in the lag $\tau$; to obtain corresponding autocorrelation functions with alternating signs, we use the even/odd scheme of (3.2). Minification/Maxification methods are easily implemented on a computer, but they have a considerably higher time complexity than TES methods. This is due to the division and power operations required in their recursive definitions.

In our study we used Minification because it is faster than Maxification. The Uniform(0,1) sequence $\{ U_n^{\min} \}$ was then transformed to an exponential sequence through the logarithmic distortion $D(x)$ as described in the previous section. However, in the absence of formulas for the autocorrelation function of transformed Minification sequences, their autocorrelations can be engineered by trial and error.

### 3.3. Method Comparison

The TES method and the Minification method can generate time series with exponential and other marginals and a predefined lag-1 autocorrelation. Apart from these common properties, each time series will have different temporal characteristics as a consequence of the generation method. Specifically, although both methods give rise to autocorrelation functions with magnitudes decaying to zero, the decay rates may differ considerably. Note that a time series $\{ X_n \}$ with a rapid decay isolates the effect of $\rho_X(1)$ on performance measures, while a time series with slow decay incorporates the effect of the higher lag autocorrelations on those measures. Figure 3.1 clearly displays these differences. The decay of the autocorrelation function of a TES method is much slower than the decay of the corresponding Minification method. The decay rate of $\rho_U^+(\tau)$ determines the decay rate of $\rho_U^-(\tau)$ as well. As is demonstrated by the results presented in figure 3.2, the absolute value of $\rho_U^-(\tau)$ decays as $\rho_U^+(\tau)$, even though $\rho_U^-(\tau)$ oscillates between negative and positive values.

## 4. AUTOCORRELATED INTERARRIVAL INTERVALS

In this section we use three sample paths generated by TES methods to present a qualitative description of the impact autocorrelated interarrivals have on the arrival process. Using TES methods with $\phi = 0$, we generated sample paths for arrival processes with positive and negative $\rho_a$ (1), where the subscript '$a$'

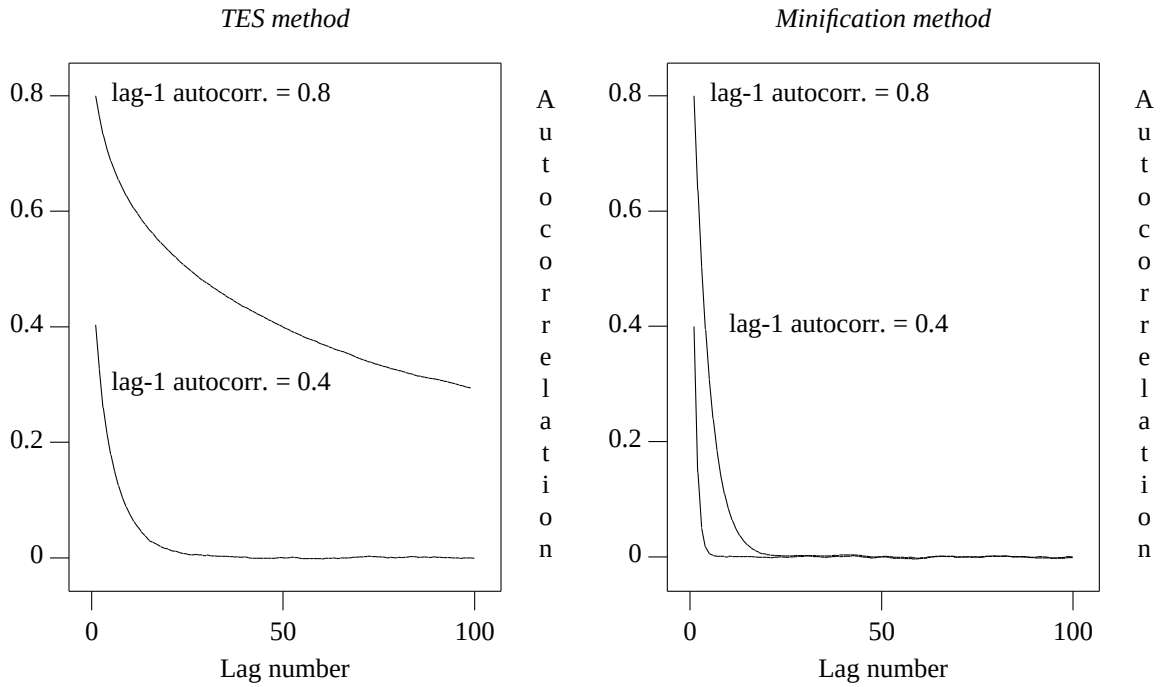


**Figure 3.1**: Measured autocorrelation versus lag number for the TES and the Minification methods. Positive lag-1 autocorrelation.

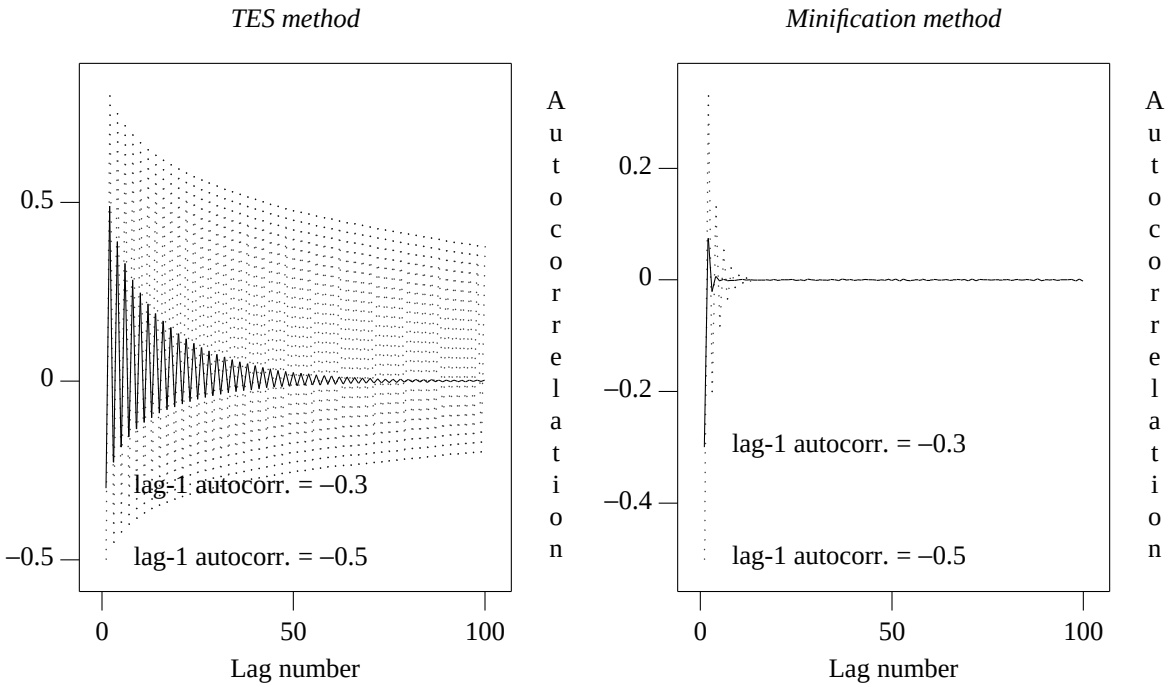


**Figure 3.2**: Measured autocorrelation versus lag number for the TES and the Minification methods. Negative lag-1 autocorrelation.

stands for arrival; in a similar vein, $\rho_s$ (1) denotes the lag-1 autocorrelation in a service process[1] (the

[1] all the presented results for the TES method use $\phi = 0$.

subscript ‘*s*’ stands for service). Interarrivals intervals in all cases follow a marginal exponential distribution with rate $\lambda$. TES was also used to generate sample paths for a Poisson process by selecting $\rho_a(1) = 0.0$. We point out that for $\alpha = 1$, TES generates independent variates (zero autocorrelations, for all $\tau$). From this point on, we will use the term autocorrelated to mean that the autocorrelations are not identically zero for all lags.

Figure 4.1 presents a graphical realization of the three sample paths: the first with $\rho_a(1) = 0.8$, the second with $\rho_a(1) = 0.0$, and the third with $\rho_a(1) = -0.5$. For each path, the arrivals in the interval [0, 100] and [500, 600] are displayed. An arrival of a job is represented by a vertical line at the arrival epoch of the job, so the separation between lines represents interarrival intervals. For $\rho_a(1) = 0.8$, we observe that the arrival pattern within an interval has a more regular appearance than the independent case. As expected, positive autocorrelation increases the likelihood that successive interarrival times will have similar durations. There is, however, a significant difference between the first and second interval. Whereas in the case of independent arrivals the two intervals have about the same number of arrivals, for $\rho_a(1) = 0.8$ the first interval has almost twice as many arrivals as the second interval. Note that the number of arrivals in both intervals is much smaller than 100, which is the expected number of arrivals in 100 time units. One can thus expect that for $\rho_a(1) = 0.8$, the number of arrivals in some intervals would be much larger than 100, and the resulting effect is one of burstiness.

Before we shift our attention to the case where $\rho_a(1) = -0.5$, we point out that the smallest attainable $\rho_a(1)$ for an exponentially distributed variable is about $-0.62$. Therefore, $\rho_a(1) = 0.0$ can be viewed as located about midway between $\rho_a(1) = -0.5$ and $\rho_a(1) = 0.8$. As in the case of positive autocorrelation, for $\rho_a(1) = -0.5$, the arrival pattern has a more regular appearance than in the independent case. Here the pattern involves consecutive interarrival times of alternating length: long, short, long, short ... As can be seen from Figure 4.2 (where we present a detailed view of the interval [0, 15]), most of the vertical lines in the case of negative lag-1 autocorrelation actually represent two arrivals. This case also exhibits a regular appearance across intervals. As in the independent case, the number of arrivals in each interval is close to 100.

The three sample paths depicted in Figure 4.1 indicate that changes in $\rho_a(1)$ entail changes in the

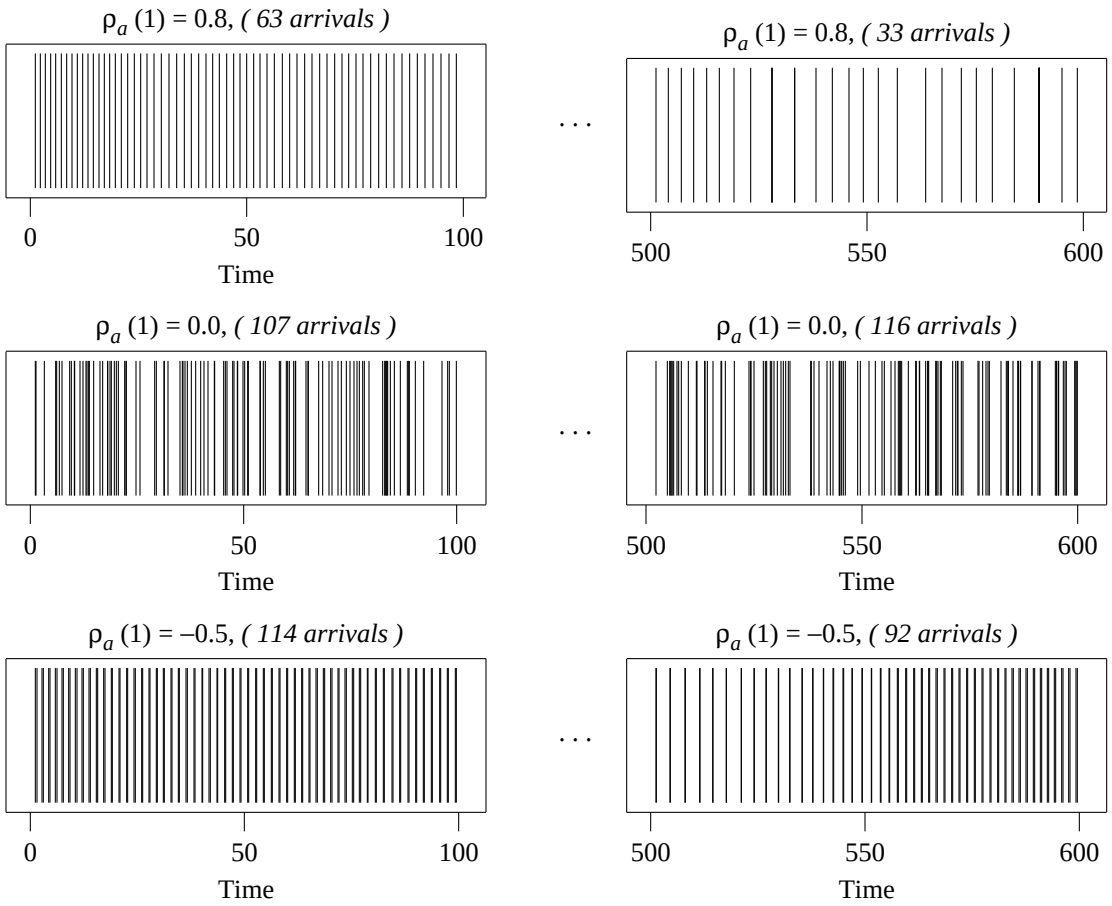


**Figure 4.1**: Sample Path Arrivals ( λ = 1.0 ).

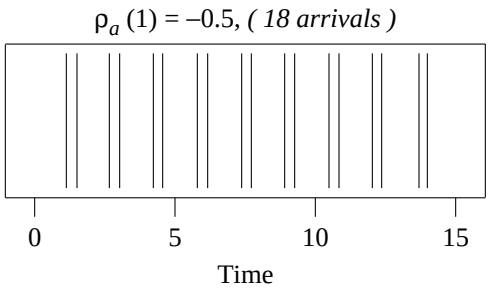


**Figure 4.2**: Sample Path Arrivals in the interval [0, 15].

amount of work that enters the system over small time intervals. In order to demonstrate this, we measured the workload that enters the system during 500 intervals of length 30 for each sample path. It was assumed that each arrival generates a service demand, independently drawn from an exponential distribution with mean 1.0. Figure 4.3 displays the results of these measurements. Each measurement was normalized by the interval length, and thus the figure displays the rate of work arrivals for each interval.

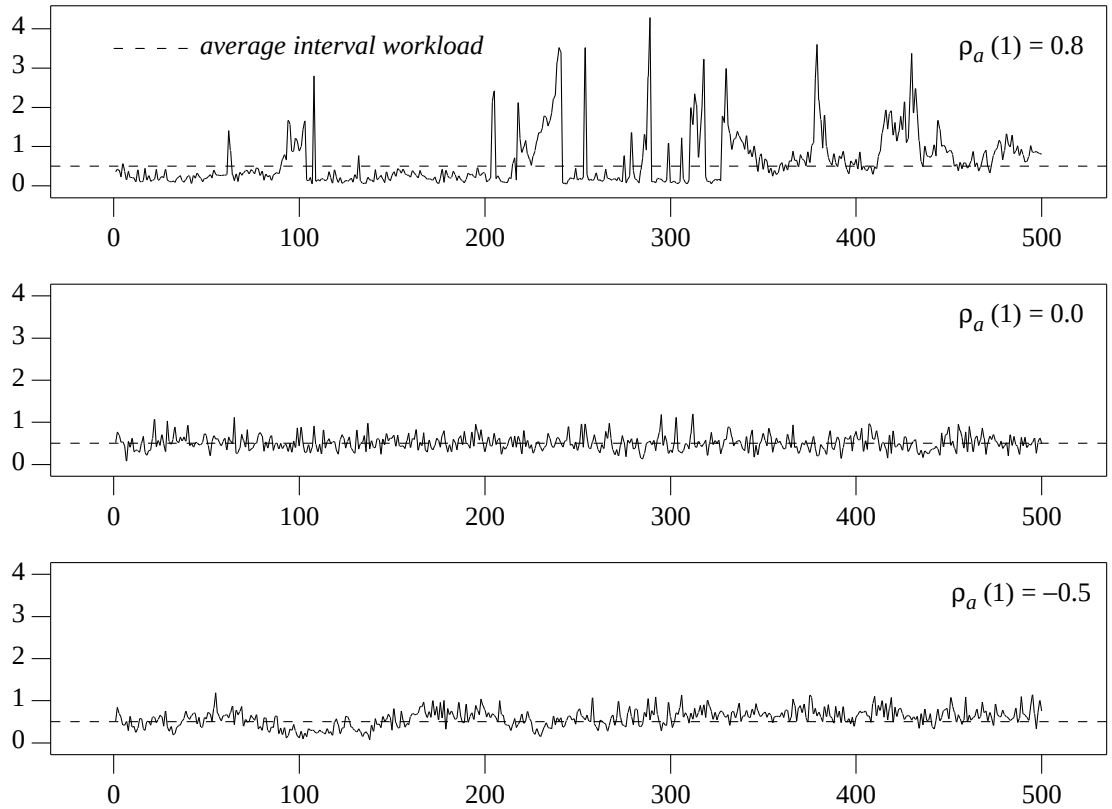


**Figure 4.3**: Workload entering the system during a time interval. TES method, $\rho_s$ (1) = 0.0, $\lambda$ = 0.5, interval length = 30 time units.

Figure 4.3 is a clear demonstration of how positive autocorrelation in interarrival intervals affects the system workload. When interarrival intervals are independent, the workload in each interval fluctuates in a narrow band straddling the average. On the other hand, for $\rho_a$ (1) = 0.8, the workload is punctuated with bursty arrivals and concomitant surges in the arrival rate of workload. From Figure 4.1 one would expect that for $\rho_a$ (1) < 0.0, the workload will be less bursty than the independent case. We do not, in fact, observe such a difference in Figure 4.3. Although we note a larger number of consecutive intervals with above/below average workload for $\rho_a$ (1) = –0.5 as compared to the independent case, the workload remains in the same narrow band straddling the average. In order to capture the impact of negative autocorrelation on workload burstiness, smaller intervals have to be considered. Figure 4.4 depicts the workload for intervals of length 5 for $\rho_a$ (1) = –0.5 and $\rho_a$ (1) = 0.0. As expected, we observe a smaller number of high traffic intervals when $\rho_a$ (1) is negative than in the case of independent arrivals.

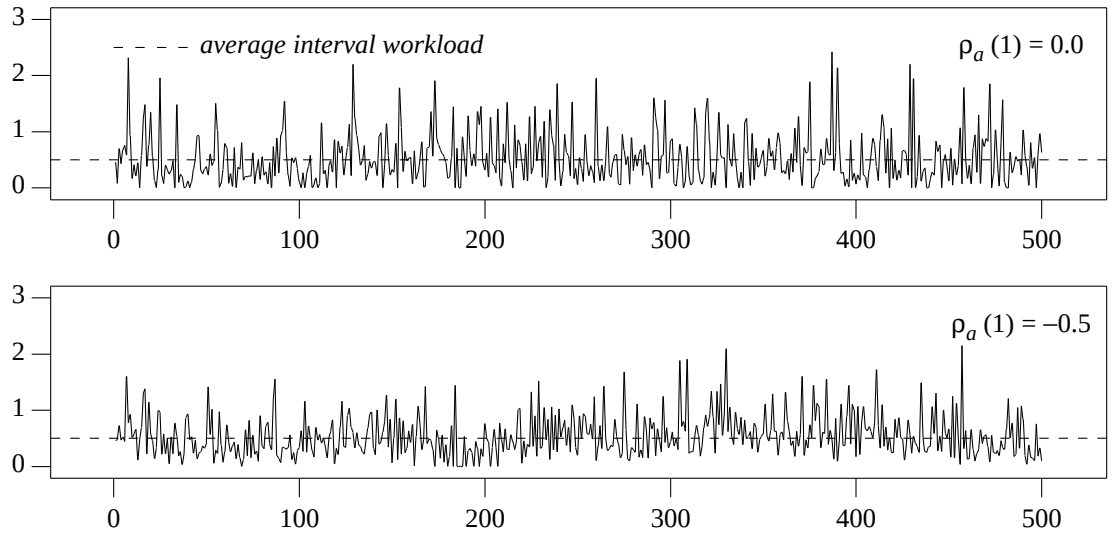


**Figure 4.4**: Workload entering the system during a time interval. TES method, $\rho_s$ (1) = 0.0, $\lambda$ = 0.5, interval length = 5 time units.

## 5. AUTOCORRELATED SERVICE DEMANDS

The trends identified in the previous section regarding the impact of autocorrelations on the arrival process are also valid for the service process. We would expect, however, to see a qualitative difference in the impact that positive autocorrelation in service demands has on the workload. A sequence of small (large) interarrival intervals has an opposite effect on the workload than a sequence of small (large) values of service demands. Small interarrival intervals increase the rate of workload arrivals, whereas small service demands increase the rate of workload disposals. Since the median of an exponential distribution is significantly smaller than the average, sequences of small values are likely to have a more profound impact on the workload than sequences of large values.

Figure 5.1 shows three sample paths of service demands with different lag-1 autocorrelations $\rho_s$ (1). For $\rho_s$ (1) = 0.8 (the upper part of Figure 5.1), three phases can be discerned in the sample path. At first, service demands are low; then comes a subsequence of high service demands which is followed by another subsequence of low service demands. Such a pattern is not present in the other two sample paths. Negative autocorrelation does, however, introduce phases in the envelope bounding the service demands.

Figure 5.2 presents a portion of the sample path of $\rho_s$ (1) = –0.5 and demonstrates more clearly the alternations between high and low values. This portion is enclosed in a rectangle in Figure 5.1. Figures 5.3 and 5.4 depict sample paths of the normalized workload over equal-length consecutive time intervals. To

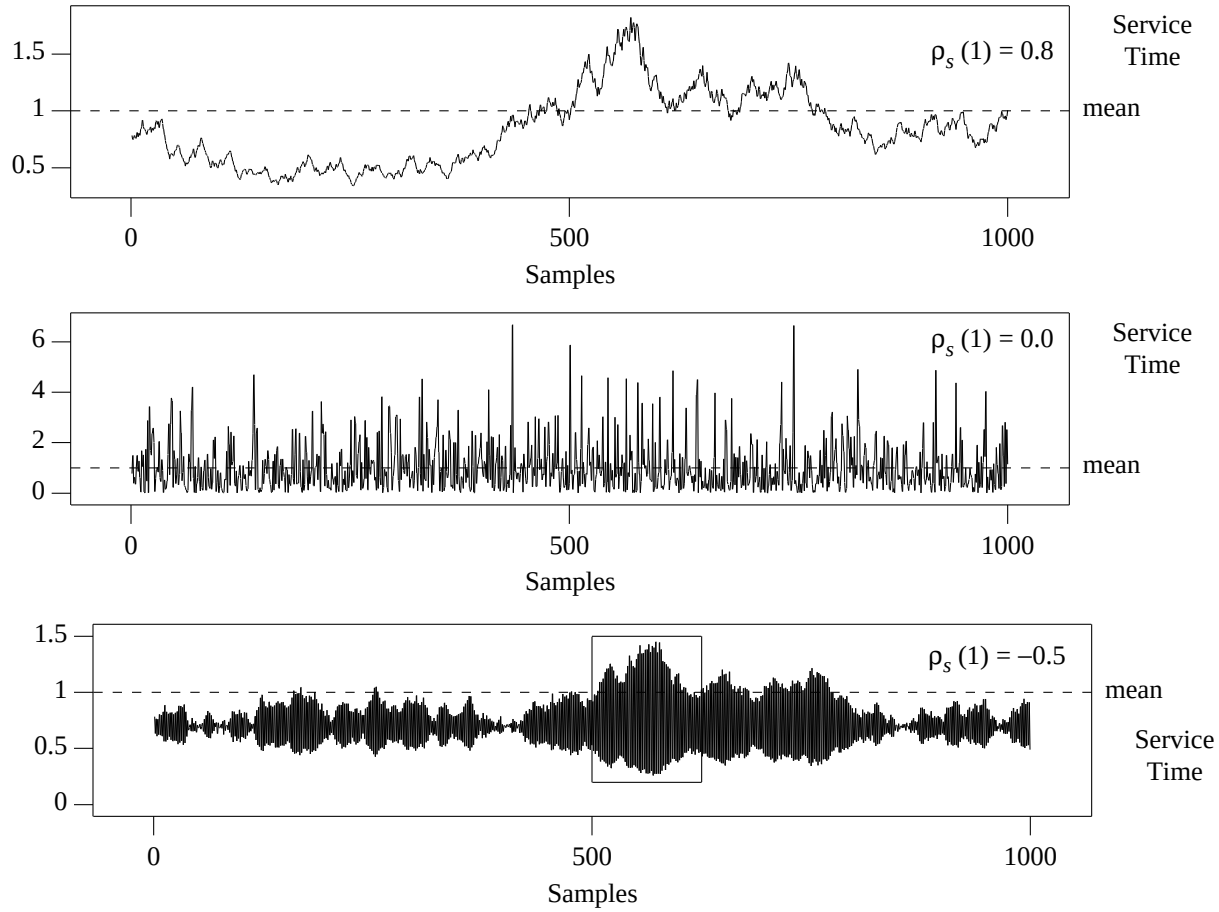


**Figure 5.1**: Service sample paths. TES method, mean = 1.0.

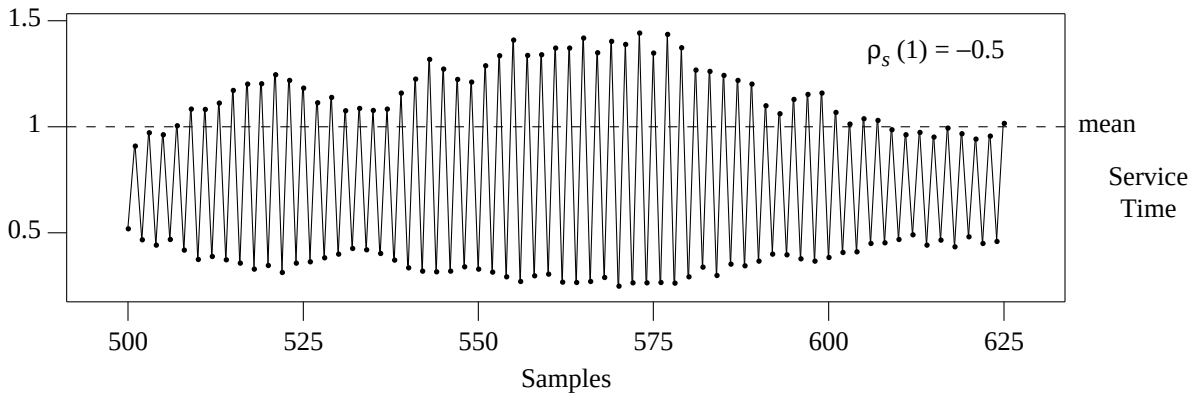


**Figure 5.2**: Service sample paths for the interval [ 500, 625 ].

compute the workload we assumed that interarrival intervals were independently drawn from an exponential distribution with rate 0.5.

When compared to the sample paths presented in Figure 4.3, the sample path of the workload process with positively autocorrelated service demands is more bursty than its independent counterpart. It is, how-

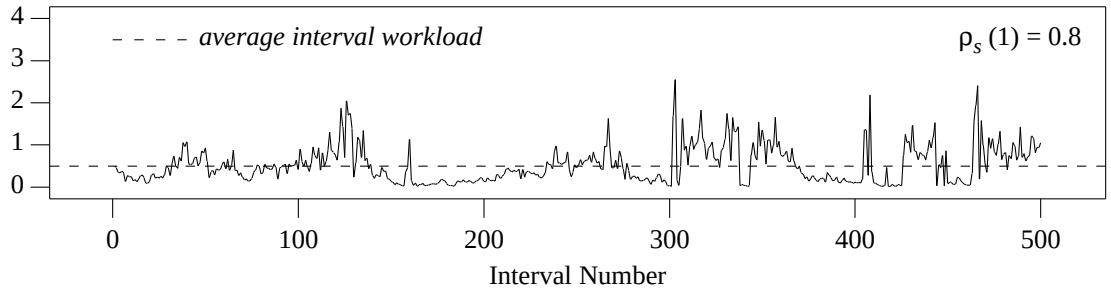


**Figure 5.3**: Workload entering the system during a time interval. TES method, $\rho_a$ (1) = 0.0, $\lambda$ = 0.5, interval length = 30 time units.

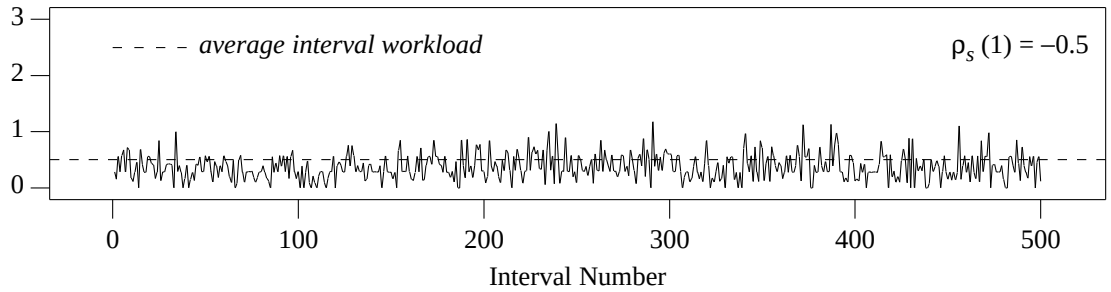


**Figure 5.4**: Workload entering the system during a time interval. TES method, $\rho_a$ (1) = 0.0, $\lambda$ = 0.5, interval length = 5 time units.

ever, less bursty than the workload generated by an arrival process with $\rho_a$ (1) = 0.8. We observe shorter periods of heavy workload and smaller amplitudes in the case of positively autocorrelated service demands, as compared to positively autocorrelated interarrival intervals.

## 6. AUTOCORRELATED ARRIVALS AND SERVICES

Having gained a measure of qualitative understanding of how autocorrelation in interarrivals or service demands affects the workload in the system, we now proceed to consider the case where both the interarrival intervals process and the service demand process are, respectively, autocorrelated.

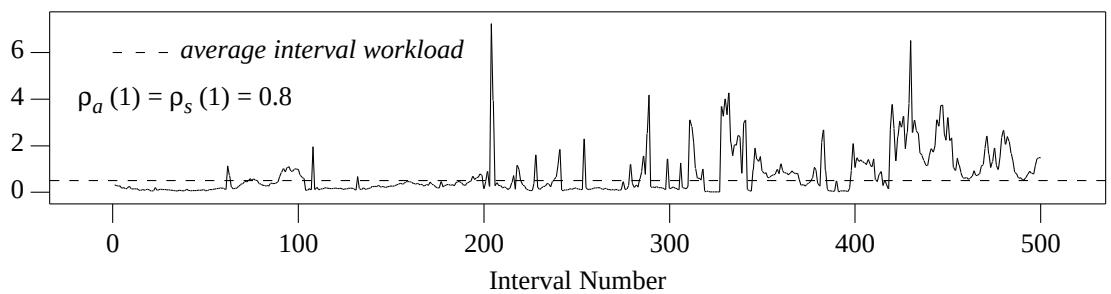


**Figure 6.1**: Workload entering the system during a time interval. TES method, $\lambda$ = 0.5, interval length = 30 time units.

However, we assume that these processes are mutually independent and, in particular, they are not cross-correlated. Thus, when $\rho_a(1) > 0$ and $\rho_s(1) > 0$, we expect half of the heavy workload periods to consist of jobs with high service demands. This leads to periods of very heavy workload surges as evidenced by Figure 6.1 which displays a sample path of the workload for $\rho_a(1) = \rho_s(1) = 0.8$. Some of the periods in Figure 6.1 generate almost three times the workload than their counterparts in Figure 5.3.

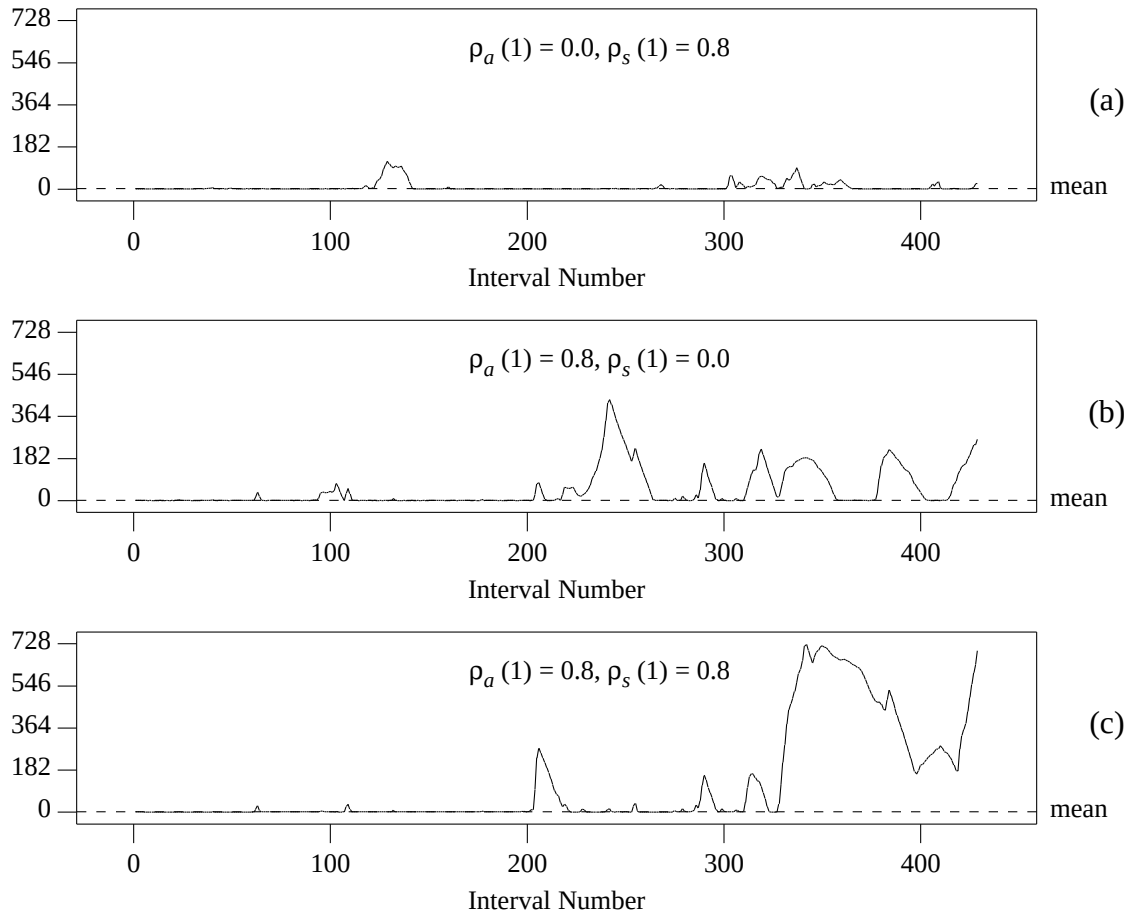


**Figure 6.2**: Mean Sojourn Times. TES method, FIFO server discipline, $\lambda = 0.5$, interval length = 30 time units.

These short but very intense periods have a dramatic impact on the sojourn time[2] in the system. Figure 6.2 depicts the impact of positive autocorrelation on the mean sojourn time through a FIFO server for the following three cases: (a) positively autocorrelated service demands and independent interarrival intervals, (b) independent service demands and positively autocorrelated interarrival intervals, and (c) positively autocorrelated service demands and positively autocorrelated interarrival intervals. For each interval, the

[2] sojourn time is the total time spent in the system by a transaction, from arrival to departure.

mean sojourn time of the jobs that arrived during that interval are displayed. By comparing the first two graphs, one can see that positive autocorrelation in interarrival intervals affects the mean sojourn times more than positive autocorrelation in service demands. When we introduce positive autocorrelation in both arrivals and services, the mean sojourn times increase dramatically; specifically, note that the magnitudes in Figure 6.2(c) are almost 1000% of those in Figure 6.2(a). The reason for this behavior is mainly the cumulative effect of bursty workload entering the system.

## 7. SIMULATION STUDY

To study the effect of autocorrelation on the mean waiting time[3] in a queuing system, we used the DeNet simulation language (see [Livn88]) to build a simulator of a First In First Out (FIFO) server. In addition to the server module, the simulator includes a source module that generates a sequence of jobs with exponential interarrival intervals and exponential service demands. The source can use either a TES method or a Minification method to introduce autocorrelation into the arrival and/or service demand processes. Each simulation experiment is characterized by the following parameters:

1. Autocorrelation generation method for the job interarrival intervals.
2. Arrival rate ( $\lambda$ ).
3. Lag-1 autocorrelation coefficient ( $\rho_a$ (1) ) for the job interarrival intervals process.
4. Autocorrelation generation method for service demands process.
5. Mean service demand ( $1/\mu$ ).
6. Lag-1 autocorrelation coefficient ( $\rho_s$ (1) ) for the job service demands process.

In each experiment the same method was used to generate both the interarrival interval and the service demand variates. The lag-1 autocorrelation coefficients for the interarrival interval and the service demand processes were drawn from the set: $P = \{ -0.55, -0.40, -0.25, 0.00, 0.25, 0.50, 0.75, 0.85 \}$ ( $(\rho_a (1), \rho_s (1)) \in P \times P$ ). Recall that the asymmetry between the positive and negative values of the lag-1 autocorrelation coefficient is due to the fact that the smallest attainable negative lag-1 autocorrelation for

[3] waiting time is defined as sojourn time excluding service time(s).

exponentially distributed variates is about −0.62, while positive lag-1 autocorrelation can approach 1.00. For all experiments that use a TES method, we used the $TES^+$ method to generate variates with positive lag-1 autocorrelation, and the $TES^-$ method to generate variates with negative lag-1 autocorrelation. To generate Minification variates with negative lag-1 autocorrelation, we used the scheme of (3.2) applied to Minification variates with positive lag-1 autocorrelation.

The mean job service time was kept constant at one time unit. For the mean interarrival intervals, we used the following values: 4.0, 2.0, 1.5 and 1.25, resulting in server utilizations of 0.25, 0.50, 0.66 and 0.80, respectively. All experiments were simulated for 20,000,000 time units.

| | **Mean** | **Distribution** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Interarrival Time** | 4.0 | Exponential, TES method | | | | | | |
| **Service Time** | 1.0 | Exponential, TES method | | | | | | |
| **Utilization** | 0.250 | | | | | | | |
| *Interarrival Time Correlation Coefficient* | *Service Time Correlation Coefficient* | | | | | | | |
| | **-0.55** | **-0.40** | **-0.25** | **0.00** | **0.25** | **0.50** | **0.75** | **0.85** |
| **-0.55** | -21.0% | -27.3% | -30.8% | -35.6% | -33.3% | -26.6% | 29.8% | 258.5% |
| **-0.40** | -19.4% | -23.1% | -26.8% | -31.5% | -29.3% | -22.2% | 29.8% | 188.8% |
| **-0.25** | -12.9% | -16.2% | -19.4% | -24.6% | -21.8% | -14.0% | 28.5% | 115.8% |
| **0.00** | 6.2% | 3.9% | 3.1% | 0.3317 | 12.5% | 35.2% | 88.3% | 148.1% |
| **0.25** | 86.3% | 72.7% | 65.1% | 61.2% | 90.7% | 149.7% | 274.6% | 390.2% |
| **0.50** | 441.5% | 344.0% | 287.9% | 260.9% | 333.3% | 519.4% | 979.5% | 1364% |
| **0.75** | 3737% | 2507% | 2177% | 2047% | 2220% | 2816% | 5497% | 8604% |
| **0.85** | 14214% | 10235% | 9235% | 9156% | 9368% | 10333% | 16066% | 25374% |

**Table 7.1**: Waiting times for a single-queue, single-server system. FIFO server discipline. Utilization 25%.

Tables 7.1 - 7.8 summarize the simulation results for the FIFO server. For each set of simulation parameters $(\rho_a(1), \rho_s(1))$, we measured the mean waiting time $\overline{W}(\rho_a(1), \rho_s(1))$. Table entries $D(\rho_a(1), \rho_s(1))$ corresponding to $(\rho_a(1), \rho_s(1))$ display the percentage of relative discrepancy from the measured benchmark mean waiting time at $\rho_a(1) = \rho_s(1) = 0.0$; that is

$$D(\rho_a(1), \rho_s(1)) = \frac{\overline{W}(\rho_a(1), \rho_s(1)) - \overline{W}(0, 0)}{\overline{W}(0, 0)} \times 100.$$

Note, however, the exception for entries indexed by (0.00, 0.00) which are the mean waiting times $\overline{W}(0, 0)$ for uncorrelated (actually independent) arrivals and services in an M/M/1 queue; the

| | **Mean** | **Distribution** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Interarrival Time** | 2.0 | Exponential, TES method | | | | | | |
| **Service Time** | 1.0 | Exponential, TES method | | | | | | |
| **Utilization** | 0.500 | | | | | | | |
| *Interarrival Time Correlation Coefficient* | *Service Time Correlation Coefficient* | | | | | | | |
| | **-0.55** | **-0.40** | **-0.25** | **0.00** | **0.25** | **0.50** | **0.75** | **0.85** |
| **-0.55** | 2498% | 138.3% | 7.3% | -32.2% | 12.2% | 265.4% | 2784% | 11276% |
| **-0.40** | 799.2% | 94.7% | 4.6% | -31.0% | 10.7% | 222.3% | 1626% | 5419% |
| **-0.25** | 375.3% | 42.9% | -1.6% | -28.8% | 7.1% | 150.8% | 1054% | 4191% |
| **0.00** | 332.2% | 47.6% | 15.9% | 1.0013 | 31.3% | 138.7% | 933.1% | 4019% |
| **0.25** | 504.8% | 164.5% | 101.9% | 79.3% | 126.4% | 284.6% | 1181% | 4378% |
| **0.50** | 1266% | 648.8% | 511.1% | 466.1% | 539.1% | 782.6% | 2058% | 5628% |
| **0.75** | 7376% | 4934% | 4430% | 4332% | 4422% | 4998% | 8081% | 14954% |
| **0.85** | 26708% | 21123% | 20781% | 20410% | 20810% | 21721% | 26180% | 39816% |

**Table 7.2**: Waiting times for a single-queue, single-server system. FIFO server discipline. Utilization 50%.

| | **Mean** | **Distribution** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Interarrival Time** | 1.5 | Exponential, TES method | | | | | | |
| **Service Time** | 1.0 | Exponential, TES method | | | | | | |
| **Utilization** | 0.667 | | | | | | | |
| *Interarrival Time Correlation Coefficient* | *Service Time Correlation Coefficient* | | | | | | | |
| | **-0.55** | **-0.40** | **-0.25** | **0.00** | **0.25** | **0.50** | **0.75** | **0.85** |
| **-0.55** | 10219% | 1212% | 249.9% | 63.6% | 250.7% | 1131% | 7441% | 21649% |
| **-0.40** | 3505% | 437.7% | 112.7% | 21.8% | 127.8% | 541.2% | 3281% | 12580% |
| **-0.25** | 2724% | 203.6% | 38.8% | -6.8% | 58.0% | 313.6% | 2612% | 11675% |
| **0.00** | 2638% | 165.5% | 32.6% | 1.9986 | 46.1% | 260.2% | 2491% | 11568% |
| **0.25** | 2833% | 281.9% | 119.4% | 80.8% | 135.1% | 370.3% | 2664% | 11805% |
| **0.50** | 3807% | 805.4% | 545.7% | 490.3% | 560.8% | 864.9% | 3381% | 12510% |
| **0.75** | 10549% | 5804% | 5281% | 5201% | 5277% | 5770% | 9312% | 19594% |
| **0.85** | 33021% | 24115% | 23209% | 23041% | 23248% | 24029% | 27888% | 41205% |

**Table 7.3**: Waiting times for a single-queue, single-server system. FIFO server discipline. Utilization 66.6%.

corresponding discrepancy is always zero and therefore we chose to display the explicit benchmark values. The first four tables are for the TES method, whereas the other four are for the Minification method.

Table 7.2 shows results of simulation experiments for the TES method corresponding to a 50% system utilization. The theoretical mean waiting time value for a FIFO single-server queue with

| | **Mean** | **Distribution** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Interarrival Time** | 1.25 | Exponential, TES method | | | | | | |
| **Service Time** | 1.0 | Exponential, TES method | | | | | | |
| **Utilization** | 0.800 | | | | | | | |
| *Interarrival Time Correlation Coefficient* | *Service Time Correlation Coefficient* | | | | | | | |
| | **-0.55** | **-0.40** | **-0.25** | **0.00** | **0.25** | **0.50** | **0.75** | **0.85** |
| **-0.55** | 17026% | 5993% | 4699% | 4450% | 4795% | 5797% | 13316% | 29360% |
| **-0.40** | 7782% | 793.3% | 343.4% | 246.0% | 366.1% | 816.3% | 4495% | 17629% |
| **-0.25** | 6809% | 407.2% | 83.8% | 27.0% | 101.4% | 431.8% | 3924% | 16818% |
| **0.00** | 6820% | 351.4% | 50.7% | 3.9974 | 58.4% | 363.5% | 3823% | 16665% |
| **0.25** | 6966% | 454.9% | 133.2% | 78.5% | 142.6% | 456.6% | 3938% | 16948% |
| **0.50** | 7850% | 946.9% | 561.9% | 503.0% | 573.9% | 916.9% | 4499% | 17317% |
| **0.75** | 14015% | 6026% | 5387% | 5315% | 5395% | 5895% | 9782% | 23740% |
| **0.85** | 34893% | 24614% | 23656% | 23463% | 23642% | 24683% | 29170% | 42364% |

**Table 7.4**: Waiting times for a single-queue, single-server system. FIFO server discipline. Utilization 80%.

$\rho_a$ (1) = $\rho_s$ (1) = 0.0 is 1.0 time unit, which is very close to the measured value. The entries in the column labeled (0.00) show how a FIFO server with independent exponential service times behaves when autocorrelation is introduced into the arrival process. The first three negative entries in the column indicate that negative lag-1 autocorrelation in the arrival process reduces mean waiting times. For $\rho_a$ (1) = –0.55, we observe a reduction of 32%. The effect of positive lag-1 autocorrelation in the interarrival intervals is much more pronounced. The smallest positive autocorrelation simulated already caused an increase of almost 80%. For $\rho_a$ (1) = 0.85, we measured a mean waiting time that was more than 200 times larger than the benchmark case $\rho_a$ (1) = $\rho_s$ (1) = 0.0.

The results displayed in the row labeled (0.00) in Table 7.2 convey a different message. Here we observe that when negative autocorrelation is introduced into services in a system with independent interarrival intervals, the mean waiting times can increase rather then decrease. Note that in this case, only odd lags have negative autocorrelation, while even lags have positive autocorrelation. The first three entries of the row display a monotonic increase in mean waiting times as $\rho_s$ (1) decreases. This observation depends, however, on the method employed to generate the autocorrelated variates; as the values in the corresponding row in Table 7.6 indicate, negative lag-1 autocorrelation in service demands can lead to a reduction in mean waiting times. Thus, the sign of $\rho_s$ (1) is not a sufficient predictor of the effect of autocorrelated

service demands on mean waiting times. Evidently, higher autocorrelations (sign and magnitude) must be taken into account. The columns labeled (0.00) in Tables 7.4 and 7.8 support this conclusion when considering $\rho_a$ (1). As we move to higher utilizations, negative lag-1 autocorrelation in interarrival intervals generated by the TES method leads to an increase rather than a decrease in mean waiting times.

The results displayed in the eight tables clearly indicate that positive lag-1 autocorrelation in either the arrival process or the service process leads to increased mean waiting times. The extent to which such autocorrelations affect queuing performance depends on the autocorrelation structure and the utilization of the server. Positively autocorrelated variates generated by the TES method have a significantly larger impact on mean waiting times than positively autocorrelated variates generated by the Minification method. The impact of positively autocorrelated variates seems to be monotonically increasing as a function of system utilization. For both methods, the effect increases nonlinearly in $\rho_a$ (1) and $\rho_s$ (1).

| | **Mean** | **Distribution** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Interarrival Time** | 4.0 | Exponential, Minification method | | | | | | |
| **Service Time** | 1.0 | Exponential, Minification method | | | | | | |
| **Utilization** | 0.250 | | | | | | | |
| *Interarrival Time Correlation Coefficient* | *Service Time Correlation Coefficient* | | | | | | | |
| | **-0.55** | **-0.40** | **-0.25** | **0.00** | **0.25** | **0.50** | **0.75** | **0.85** |
| **-0.55** | -33.6% | -33.5% | -32.5% | -28.7% | -18.3% | -2.2% | 37.6% | 84.6% |
| **-0.40** | -25.6% | -25.0% | -23.2% | -17.2% | -5.9% | 11.4% | 52.3% | 98.6% |
| **-0.25** | -19.2% | -18.7% | -16.1% | -8.4% | 4.8% | 24.6% | 69.9% | 120.0% |
| **0.00** | -13.1% | -12.6% | -9.5% | 0.3339 | 15.5% | 39.0% | 91.7% | 147.9% |
| **0.25** | -8.3% | -7.9% | -4.6% | 6.2% | 23.8% | 50.7% | 111.2% | 174.2% |
| **0.50** | 4.3% | 4.7% | 8.9% | 22.6% | 44.1% | 77.0% | 150.9% | 226.5% |
| **0.75** | 46.8% | 46.5% | 53.2% | 73.7% | 104.4% | 150.6% | 253.5% | 357.5% |
| **0.85** | 68.6% | 67.8% | 75.6% | 98.9% | 133.0% | 184.3% | 298.1% | 413.1% |

**Table 7.5**: Waiting times for a single-queue, single-server system. FIFO server discipline. Utilization 25%.

The impact of negatively autocorrelated variates on the performance of a FIFO server is not clear-cut. When the Minification method was used, a decrease in either $\rho_a$ (1) or $\rho_s$ (1) led to a decrease in mean waiting times. The higher the utilization, the larger the decrease. For TES the situation is more complex; for negatively autocorrelated variates, the direction of change in mean waiting times depends on the utilization and whether it is used to drive the arrival process or the service process. For low utilization, mean

| | **Mean** | **Distribution** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Interarrival Time** | 2.0 | Exponential, Minification Method | | | | | | |
| **Service Time** | 1.0 | Exponential, Minification Method | | | | | | |
| **Utilization** | 0.500 | | | | | | | |
| *Interarrival Time Correlation Coefficient* | *Service Time Correlation Coefficient* | | | | | | | |
| | **-0.55** | **-0.40** | **-0.25** | **0.00** | **0.25** | **0.50** | **0.75** | **0.85** |
| **-0.55** | -50.3% | -51.1% | -48.7% | -35.2% | -11.6% | 30.1% | 153.3% | 321.4% |
| **-0.40** | -41.1% | -41.4% | -38.4% | -26.1% | -3.8% | 36.4% | 158.1% | 324.9% |
| **-0.25** | -31.3% | -31.5% | -27.7% | -14.9% | 7.6% | 48.4% | 172.4% | 339.1% |
| **0.00** | -19.3% | -19.7% | -15.0% | 1.0002 | 24.2% | 67.7% | 198.3% | 366.8% |
| **0.25** | -8.7% | -9.5% | -4.0% | 13.1% | 39.9% | 86.9% | 226.8% | 400.0% |
| **0.50** | 18.0% | 16.9% | 23.7% | 44.1% | 74.3% | 130.3% | 286.2% | 460.3% |
| **0.75** | 102.9% | 101.1% | 111.7% | 141.6% | 182.3% | 252.4% | 421.6% | 613.8% |
| **0.85** | 227.1% | 225.3% | 237.4% | 275.9% | 324.7% | 397.3% | 579.6% | 790.2% |

**Table 7.6**: Waiting times for a single-queue, single-server system. FIFO server discipline. Utilization 50%.

| | **Mean** | **Distribution** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Interarrival Time** | 1.5 | Exponential, Minification Method | | | | | | |
| **Service Time** | 1.0 | Exponential, Minification Method | | | | | | |
| **Utilization** | 0.667 | | | | | | | |
| *Interarrival Time Correlation Coefficient* | *Service Time Correlation Coefficient* | | | | | | | |
| | **-0.55** | **-0.40** | **-0.25** | **0.00** | **0.25** | **0.50** | **0.75** | **0.85** |
| **-0.55** | -53.7% | -55.6% | -50.4% | -32.7% | -4.9% | 46.0% | 193.6% | 389.2% |
| **-0.40** | -47.1% | -48.7% | -44.3% | -28.1% | -1.4% | 48.2% | 194.4% | 389.4% |
| **-0.25** | -37.1% | -38.6% | -33.9% | -17.8% | 8.8% | 58.5% | 205.1% | 400.0% |
| **0.00** | -21.5% | -23.2% | -17.6% | 1.9991 | 27.6% | 78.9% | 228.0% | 424.7% |
| **0.25** | -6.0% | -7.9% | -1.4% | 18.2% | 48.3% | 102.8% | 257.9% | 460.1% |
| **0.50** | 30.5% | 28.5% | 36.2% | 58.3% | 91.2% | 149.6% | 312.2% | 520.7% |
| **0.75** | 142.8% | 140.4% | 150.0% | 176.4% | 215.5% | 282.0% | 460.3% | 682.8% |
| **0.85** | 293.0% | 290.3% | 301.3% | 331.0% | 375.8% | 449.1% | 642.3% | 878.7% |

**Table 7.7**: Waiting times for a single-queue, single-server system. FIFO server discipline. Utilization 66%.

waiting times decrease (Table 7.1). As utilization increases, we eventually observe an increase in mean waiting times (Table 7.4). In all cases, however, $\rho_s$ (1) precipitates an increase in mean waiting times. As in the case of Minification, the increase is monotonic in the server utilization; the higher the utilization, the larger the increase.

When we compare the values of an $(x, y)$ entry to its $(y, x)$ counterpart in each of the TES tables (Tables 7.1 - 7.4), we notice that positively autocorrelated arrivals have a stronger impact on waiting times

| | **Mean** | **Distribution** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Interarrival Time** | 1.25 | Exponential, Minification Method | | | | | | |
| **Service Time** | 1.0 | Exponential, Minification Method | | | | | | |
| **Utilization** | 0.800 | | | | | | | |
| *Interarrival Time Correlation Coefficient* | *Service Time Correlation Coefficient* | | | | | | | |
| | **-0.55** | **-0.40** | **-0.25** | **0.00** | **0.25** | **0.50** | **0.75** | **0.85** |
| **-0.55** | -53.2% | -55.6% | -48.9% | -29.5% | 0.0% | 55.6% | 218.5% | 435.1% |
| **-0.40** | -50.3% | -53.0% | -47.2% | -28.8% | 0.0% | 54.9% | 216.4% | 432.1% |
| **-0.25** | -41.0% | -43.7% | -37.9% | -19.4% | 8.9% | 63.5% | 225.1% | 441.2% |
| **0.00** | -23.1% | -25.7% | -19.3% | 3.9974 | 28.7% | 84.1% | 247.0% | 463.9% |
| **0.25** | -1.9% | -5.1% | 2.3% | 22.9% | 54.5% | 112.5% | 280.1% | 500.8% |
| **0.50** | 43.3% | 39.8% | 48.0% | 70.2% | 103.9% | 164.7% | 337.2% | 561.9% |
| **0.75** | 180.2% | 176.4% | 185.9% | 210.8% | 247.9% | 313.7% | 497.1% | 731.3% |
| **0.85** | 363.6% | 359.3% | 369.7% | 396.5% | 435.5% | 504.6% | 697.7% | 941.2% |

**Table 7.8**: Waiting times for a single-queue, single-server system. FIFO server discipline. Utilization 80%.

than positively autocorrelated services. In other words, the entry $(x, y)$ is smaller than $(y, x)$ for $x \geq 0.0$ and $y \geq 0.0$. This is not, however, the case for Minfication (Tables 7.5 - 7.8). Here we observe that entry $(x, y)$ is larger than entry $(y, x)$ for $x \geq 0.0$ and $y \geq 0.0$. We conclude that the two methods differ considerably in the relative impact of autocorrelation in the arrival and service processes, on the mean waiting times.

It is clear from Tables 7.1 - 7.8, that TES and Minification methods also have different absolute impacts on the mean waiting time. The fact that TES entries from Tables 7.1 - 7.4 consistently dominate their minification counterparts from Tables 7.5 - 7.8, appears to be due to the fact that TES autocorrelations dominate Minification autocorrelations in magnitude. Since TES autocorrelations persist to higher lags than Minification autocorrelations (the former have "fatter" tails than the latter), our study can be viewed as an illustration of the importance of high-lag autocorrelations on queuing performance measures.

## 8. CONCLUSIONS

In this paper we studied the impact of autocorrelated interarrival intervals and service demands on performance measures of a FIFO server. The general dramatic effects on mean sojourn times that were demonstrated by the simulation results lead us to make a number of recommendations to practitioners of queuing-oriented performance analysis. Specifically, we strongly recommend that analysts add routine

sensitivity analysis of system performance with respect to autocorrelation. Furthermore, we recommend that autocorrelation structure be included in the workload profile of a system. When a conservative benchmark is needed for positive or negative (alternating signs) autocorrelations, we suggest a TES method. We found both TES and Minification methods to be easy to implement and very fast to compute, though TES is considerably faster than Minification. Another advantage of TES methods is that they are quite versatile, as they can generate both monotone and oscillatory autocorrelation functions. They also provide a more conservative benchmark in that their higher (tail) autocorrelations decay more slowly than Minification autocorrelations.

It should be emphasized that there is a glaring need to complement simulation (empirical) studies of autocorrelation in queuing systems with analytical work. The analytical studies cited earlier, such as [Runn61, Runn62, Jaco80, Tin85, Patu93] clearly attest to the mathematical difficulties that result from introducing even simple (Markovian) dependence into arrivals and services. Nevertheless, more fundamental insight into the impact of autocorrelation on queuing systems can only be gained by analytical methods, and such effort is well worth it. This need is made all the more pressing with the advent of increasingly complex systems spawned by rapidly evolving technologies such as telecommunications, manufacturing, etc., where autocorrelations in arrival and service processes are both common and significant.

The study of the FIFO server reported here is part of an ongoing effort to develop computer methods for generating autocorrelated random sequences and to understand the modeling power of autocorrelation as a practical characterization of dependence in real-life systems. The impact of autocorrelation on other queuing disciplines and queuing networks is the focus of our current work. In the future, we plan to address the question of how to model a real-life bursty workload as a computer-generated autocorrelated time series.

**REFERENCES**

[Brat87].
P. Bratley, B.L. Fox, and L.E. Schrage, *A Guide to Simulation*, Springer-Verlag, 1987.

[Devr86].
L. Devroye, *Non-Uniform Random Variate Generation*, Springer-Verlag, 1986.

[Fend89].
K.W. Fendick, V.R. Saksena, and W. Whitt, ''Dependence in Packet Queues,'' *IEEE Transactions on Communications*, vol. 37, pp. 1173-1183, 1989.

[Heff73].
H. Heffes, ''Analysis of First-Come First-Served Queuing Systems with Peaked Inputs,'' *Bell System Technical Journal*, vol. 52, pp. 1215-1228, 1973.

[Heff80].
H. Heffes, ''A Class of Data Traffic Processes — Covariance Function Characterization and Related Queuing Results,'' *Bell System Technical Journal*, vol. 59, pp. 897-929, 1980.

[Heff86].
H. Heffes and Lucantoni D. M., ''A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance,'' *IEEE Journal on Selected Areas in Communications*, vol. SAC-4, pp. 856-868, 1986.

[Jaco80].
P.A. Jacobs, ''Heavy Traffic Results for Single Server Queues with Dependent (EARMA) Service and Interarrival Times,'' *Advances in Applied Probability*, vol. 12, pp. 517-529, 1980.

[Jage92].
D.L. Jagerman and B. Melamed, ''The Transition and Autocorrelation Structure of TES Proceses Part I: General Theory,'' *Stochastic Models*, 1992. (to appear)

[Jage92a].
D.L. Jagerman and B. Melamed, ''The Transition and Autocorrelation Structure of TES Processes Part II: Special Cases,'' *Stochastic Models*, 1992. (to appear)

[Kell79].
F.P. Kelly, *Reversibility and Stochastic Networks*, Wiley, 1979.

[Lewi91].
P.A.W. Lewis and E. McKenzie, ''Minification Processes and Their Transformations,'' *J. Appl. Prob.*, vol. 28, pp. 45-57, 1991.

[Livn88].
M. Livny, ''DeNet User's Guide'', 1988. Version 1.0, Computer Sciences Dept., Univ. of Wisconsin-Madison.

[Mela91].
B. Melamed, ''TES: A Class of Methods for Generating Autocorrelated Uniform Variates,'' *ORSA J. on Computing*, vol. 3, no. 4, pp. 317-329, 1991.

[Patu93].
B.E. Patuwo, R.L. Disney, and D.C. McNickle, ''The Effect of Correlated Arrivals on Queues,'' *IIE Transactions*, 1993. (to appear)

[Runn61].
J.Th. Runnenburg, ''An Example Illustrating the Possibilities of Renewal Theory and Waiting-Time Theory for Markov-Dependent Arrival Intervals,'' *Proc. Ser. A. Kon. Neder. Akad. Weten.*, vol. 64, pp. 560-576, 1961.

[Runn62].
J.Th. Runnenburg, ''Some Numerical Results on Waiting-Time Distributions for Dependent Arrival-Intervals,'' *Statistica Neerlandica*, vol. 16, no. 4, pp. 337-348, 1962.

[Tin85].
Tin, P., ‘‘A Queueing System with Markov Dependent Arrivals,’’ *J. Appl. Prob.*, vol. 22, pp. 668-677, 1985.