

# Incorporating Autocorrelated Service Times in the Analysis of Delay Systems

**Mojtaba Araghi**

University of Toronto,  
Joseph L. Rotman School of Management  
105 St. George Street, Toronto, M5S 3E6, CANADA,  
*mojtaba.araghi08@rotman.utoronto.ca*

**Barış Balcioğlu**

University of Toronto,  
Department of Mechanical and Industrial Engineering  
5 King's College Rd., Toronto, ON M5S 3G8, CANADA,  
*baris@mie.utoronto.ca*

## Abstract

In the literature, it has been shown that the autocorrelation in service times has a dramatic impact on performance measures in single server queueing systems. Although some recent studies deal with systems having autocorrelated service times, a general methodology that can incorporate this information in the analytical models has not been previously designed. In this paper, we propose using renewal approximations to capture the characteristics of autocorrelated service times. Our observations indicate that a single approximating renewal service time distribution fails to accurately predict the performance measures of the original system with autocorrelated service times. Therefore, we propose incorporating the server utilization in the approximation procedure, which provides a family of approximating renewal service times. We test the proposed approximation in predicting the mean waiting times as well as optimizing make-to-stock queues when service times could be autocorrelated. The numerical results demonstrate that the proposed approximation, which has a broad application area, while being straightforward, is highly accurate.

**Keywords and Phrases:** Autocorrelation; Index of Dispersion; Mean Waiting Time; Queueing; Make-to-Stock Queues; MMPP

# 1 Introduction

While modeling delay systems, one of the most commonly made assumptions is that the service times are independent and identically distributed (i.i.d) random variables (r.v.). Yet, recently in the literature, there is an increasing number of problems studied where this assumption does not hold true since service times can be autocorrelated. More importantly, the presence of positive or negative autocorrelation in the service time data significantly influences the system performance. This necessitates incorporating autocorrelation information in the analysis instead of merely ignoring it. In this paper, we propose using a new renewal approximation technique to capture the autocorrelation in service time data while modeling single server queueing systems. The distinguishing feature of the proposed technique is that it fits a family of renewal service times as a function of server utilization. Once the approximating service times are characterized, the performance of the original system with autocorrelated service times can be accurately predicted using existing models operating on the i.i.d. service time assumption.

Livny, Melamed and Tsiolis (1993) study the impact of positive and negative autocorrelation in service times in a single server queue. In their example with service times having positive autocorrelation at lag-1, a subsequence of long service demands is followed by a subsequence of short service demands. In the example with service times having negative autocorrelation at lag-1, service times alternate between high and low values. Furthermore, they observe that when service times are positively autocorrelated the sample path of the workload process is more bursty than the one in its counterpart queue in which the service times are i.i.d. r.v.s. Their simulation results show that positively autocorrelated service times significantly lengthen the mean waiting time. In cases with negative autocorrelation, they also observe longer mean waiting times, yet this is attributed to positive autocorrelation one would have at even lags. In the examples with negatively autocorrelated service times we study in Sections 2 and 3, since autocorrelation vanishes beyond lag-1 we observe that mean delay can be shorter than the one in its counterpart queue having i.i.d. service times. As

will be discussed in more detail in Section 3.2, autocorrelated times to failure and/or repair times can induce autocorrelation in the effective service time, i.e., the process completion time. Via a simulation study, Altıok and Melamed (2001) point out that autocorrelated times to failure and autocorrelated repair times also affect the customer service levels.

Mi et al. (2007) present real data showing that in enterprise storage systems such as Web servers, e-mail servers or in traces from consumer electronics devices, disk service times can be autocorrelated. In their simulation study, they employ Markov-modulated Poisson processes (MMPP) to induce autocorrelation in service times. They consider multi-tiered systems with a closed loop and show that if autocorrelation exists in service times at a stage, it propagates across the entire loop and is present in the arrival stream of stages that precede that stage, significantly degrading their performance as well as end-to-end performance.

To shed light on computer systems where processors serve several applications in parallel, Mahabhashyam and Gautam (2005) study a single server system where customers arriving according to a Poisson process are served with Markov-modulated service rates (MMSR). In this model, the server rate changes when the environment changes. MMSR differs from MMPP since the Markov process representing the state of the environment can change even when there is no customer in the system. If MMPP models the service times, the state of the environment freezes with the customer leaving the server idle and may change only after the next arrival making the server busy again. The nature of MMSR makes the mean service time dependent on the Poisson arrival rate as demonstrated by Mahabhashyam and Gautam. As explained in Section 2, we assume autocorrelated service times to be independent of the customer arrival process, hence, our approximation technique cannot be used to study problems involving MMSR. Yet, we choose to mention about them here since one can observe high levels of autocorrelation in service times in systems with MMSR. For studies employing MMSR (possibly yielding autocorrelated service times) while modeling the traffic flow on a roadway link subject to incidents, we also refer the reader to Baykal-Gürsoy and Xiao (2004) and Baykal-Gürsoy, Xiao, and Ozbay (2009). An earlier model considering MMSR is by Yechiali and Naor (1971) who also include state dependent Poisson customer arrival

rates. Zhou and Gans (1999), on the other hand, slightly modify the MMPP to model systems with varying service times. In their model, the system state can change only after a service completion. We will extend their model in Section 3.1 while testing our renewal approximation in a make-to-stock queue setting. An extension to systems with varying service time distribution is that of Boxma and Kurkova (2001) who study an  $M/G/1$  queue alternating between two service speeds. They assume that the high-speed periods follow an exponential distribution while the low-speed periods have a general distribution and study the tail behavior of the workload distribution. An example where service times can become autocorrelated due to the superposition of arrival processes is given by Takine (2002). Takine considers a Markovian arrival process served on a first-come-first-served (FCFS) basis with service time distributions of non-Poisson arrival streams differing from one another and characterizes the joint probability generating function of the stationary queue length distribution.

As our brief literature survey shows, the importance of considering autocorrelation in service times has become more apparent with recent studies. Yet, constructing analytical models when customer arrivals are non-Poisson or state-dependent service times are non-exponential appears to be difficult. In such cases and in the presence of more general autocorrelated service times, researchers have resorted to simulation experiments to study the system behavior. Our initial idea to study problems involving autocorrelated service times has been to employ renewal approximations proposed for capturing the statistics of autocorrelated data. In the literature, two-parameter approximations by Whitt (1982) or three-parameter approximations by Jagerman et al. (2004), Balcioglu et al. (2008) and Araghi and Balcioglu (2008) have been used to approximate the autocorrelated interarrival times. The three-parameter renewal approximation has been further exploited to model positively autocorrelated times to failure by Balcioglu et al. (2007). One of the observations Balcioglu et al. (2007) make is that the autocorrelated process completion times (see Section 3.2 for an explanation) due to autocorrelated or non-exponential times to failure require a server utilization based renewal service time approximation. However, their study has been

restricted and the approximation they propose could only work if processing times in the absence of failures are deterministic and times to failure can be modeled by phase-type r.v.s. Hence, in this paper, we propose a procedure that can be applied to single server queues with renewal interarrival times (note that autocorrelated interarrival times can be approximated by renewal processes, see Balcioğlu et al. 2008 and the references therein) and more general autocorrelated service times than the above mentioned studies consider. We should add that Fendick, Saksena and Whitt (1989) propose using a two-parameter renewal service time approximation under the heavy-traffic regime. As will be discussed in Section 2, for medium to low utilizations this approximation can incur big estimation errors.

We have several contributions in this study. Although the degrading impact of autocorrelation in service times on mean response time has been shown, in addition to our examples concerning the mean waiting time in Section 3, we demonstrate that ignoring positive (negative) autocorrelation of service times results in underestimation (overestimation) of base-stock levels and system costs in make-to-stock queues. This observation extends the results by Zipkin (1995) who studies the  $M/GI/1$  make-to-stock queue and Sanajian and Balcioğlu (2009) who study the  $GI/GI/1$  make-to-stock queue while demonstrating the degrading impact of service time variability. We also show that even when autocorrelated service times are independent of the customer arrival process, the idle periods that are inserted between service times have a diluting effect, decreasing the impact of both positive and negative autocorrelation when the server utilization gets lower. This necessitates considering server utilization while obtaining a family of approximating renewal service times. The idea of incorporating server utilization has been suggested by Albin (1984) while approximating the mean waiting time in a queue with i.i.d. service times receiving the superposition of independent renewal arrival processes. Our study differs from hers since we consider autocorrelated service times while assuming a renewal customer arrival process (we assume that if necessary the superposition process can be well-approximated following Balcioğlu et al., 2008 and Araghi and Balcioğlu, 2008). We design a procedure to fit the renewal service time distributions. Accordingly, using the collected autocorrelated service time data in the

simulation of a queue receiving Poisson job arrivals, we fit a function of server utilization for the approximating service time distribution.

While the simulation experiments may be necessary for general autocorrelated service times, in the case of 2-state MMPP service times, following Yechiali and Naor (1971) we provide the analytical model to compute the mean waiting time exactly when customer arrivals are Poisson. Hence, in this case, the simulation step is not necessary while fitting a utilization dependent function for the approximating service times. We test the accuracy of our renewal approximation both in predicting the mean waiting times (Section 2) and finding the optimal base-stock level and system cost in make-to-stock queues (Section 3). Note that while mean waiting time prediction has been a common measure of performance for renewal approximations, the latter involving make-to-stock queues is more challenging since it requires the entire system size distribution to be captured accurately. Additionally, it is time consuming to employ a simulation-optimization approach to obtain the base-stock level and cost of a make-to-stock queue with autocorrelated service times. This indicates that even though our proposed renewal approximation might require simulated mean waiting times, once the approximating renewal service time family is fit it is more time-efficient than the simulation-optimization approach to optimize the make-to-stock queue. In summary, our proposed solution operates on a simple idea easy to implement while its application area is quite broad.

The rest of the paper is organized as follows: In Section 2, we propose a server utilization dependent renewal service time approximation. Here, we show the effect of autocorrelation in service times on the mean waiting time and demonstrate the accuracy of the proposed approximation. In Section 3, we design two make-to-stock queue problems, which can arise naturally in many settings. This gives us a chance to demonstrate the performance of our proposed approximation in capturing the system size distribution of the original system with autocorrelated service times. Finally, Section 4 concludes the paper and presents possible future research topics.

## 2 Utilization Dependent Renewal Approximation

In this section, we propose using server utilization dependent versions of the three-parameter Exponential Residual (ER) renewal approximation due to Jagerman et al. (2004) and Generalized Erlang (GE) renewal approximation due to Araghi and Balcioglu (2008) for autocorrelated service time data. Let  $X = \{X_n\}_{n=0}^{\infty}$  be a *second order stationary* or *weakly stationary* time process where  $X_n$  denotes the length of the  $n$ th customer's service time sampled from a delay system. In other words, the finite mean  $m_x$ , and the finite variance  $\sigma_x^2 = Var[X_n] < \infty$  of  $X$  are the same for all  $n$ , and the autocorrelation function between  $X_k$  and  $X_{k+j}$ ,  $\rho_x(j) = (E[X_k X_{k+j}] - m_x^2)/\sigma_x^2$ ,  $j = 0, 1, \dots$  does not depend on  $k$  but  $j$ . Let  $\mu_x = 1/m_x$  be the service rate.  $X$  is assumed to be non-lattice and independent of the customer arrival process.

Traditionally, the approximating renewal process  $R$  is expected to preserve or modify some key statistics of the approximated stream,  $X$ , as follows (see Whitt, 1982, Jagerman et al., 2004, and Araghi and Balcioglu, 2008):

- (a)  $\mu_R = \mu_x$ , so the mean service times of  $R$  and  $X$  are equal.
- (b)  $c_R^2 = I_x$ , where  $I_x$  is the limiting index of dispersion for counts of  $X$  given by (Cox and Lewis, 1968, p. 78 and Gusella, 1991)

$$I_x = c_x^2 \left[ 1 + 2 \sum_{j=1}^{\infty} \rho_x(j) \right],$$

and  $c_x^2 = \sigma_x^2/m_x^2$  is the squared-coefficient of variation of the marginal density function of the original process.

While setting  $\mu_R = \mu_x$  makes sense (since the impact of autocorrelation is not observed on the mean), forcing  $c_R^2 = I_x$  for autocorrelated service times can create problems. In order to demonstrate why renewal approximations for autocorrelated service times should be a function of server utilization, we present a simulation study in Table 1. Throughout this paper, we have used a commercial software package, Arena (Version 5), to conduct

our simulations. In all the simulation experiments, a warm-up period long enough to serve 1,000 customers guaranteed reaching the steady-state. Once in the steady-state in each of 20 replications, serving one million customers resulted in tight enough confidence intervals (CI). In Table 1, we consider a single server queueing system receiving Poisson arrivals where the service times are generated using a Transform-Expand-Sample (TES<sup>+</sup>) process, which we denote by the  $M/TES^+/1$  queue. We refer the reader to Appendix A for a summary of TES<sup>+</sup> processes. In this particular case, the TES<sup>+</sup> service times have been generated with parameters  $\mu_x = 0.8, L = -0.2, R = 0.2$  resulting in  $\rho_X(1) = 0.32$  and  $I_x = 3.5698$ . Here, the first column shows the server utilization  $\rho = \lambda/\mu_x$  with  $\lambda$  as the arrival rate, whereas the second column lists the simulated mean waiting times with their 95% CI in parenthesis. Since the TES<sup>+</sup> service times have a marginal exponential distribution, ignoring autocorrelation in service times implies approximating the original  $M/TES^+/1$  queue by an  $M/M/1$  queue with service rate 0.8 and the same  $\rho$  as listed in column 1. Using the standard formula (e.g. Eq. 2.20 in Gross and Harris, 1998, p. 65) one can easily obtain the mean waiting time in the  $M/M/1$  queue as listed in the third column. The prediction error of such an approximation is presented in the same column in parentheses. This helps us assess the impact of positive autocorrelation in service times on mean delay. Just like in problems involving positively autocorrelated interarrival time streams, we observe that as server utilization increases, positive autocorrelation in service times lengthens the mean delay more. This indicates the importance of incorporating autocorrelation in service times in analytical models.

Having discussed the renewal approximation techniques, since we have  $\mu_R = \mu_x$  and if we force  $c_R^2 = I_x$ , one could easily argue employing the PK formula,

$$\overline{W} = \frac{\rho(1 + c_R^2)}{2\mu_R(1 - \rho)} = \frac{\rho(1 + I_x)}{2\mu_x(1 - \rho)}, \quad (1)$$

to approximate the mean waiting time in the  $M/TES^+/1$  queue. The argument here is that after analyzing the autocorrelated service time streams, there will be a single approximating renewal service time to employ for all  $\rho$  values. Since the approximating system will be an

$M/GI/1$  queue with the same  $\rho$  as listed in column 1, all we need, then, should be  $\mu_x$  and  $I_x$  of the original service times since they become the rate and  $c_R^2$  of the approximating renewal service time, respectively. When Eq. (1) is employed, we obtain the estimates listed in column 4 of Table 1 with approximation errors provided in parentheses. As one can see, a single renewal service time overestimates the mean waiting time significantly (in absolute errors, worse than the  $M/M/1$  queue approximation) at lower utilization levels. Yet, the error decreases monotonically with increase in server utilization and gets low above  $\rho = 0.8$ .

This observation makes us believe that as far as autocorrelated service times are concerned, one has to have the approximating service time as a function of  $\rho$ . Since  $\mu_x$  does not change with  $\rho$ , having Poisson arrivals one only needs to characterize the second-moment (or the squared-coefficient of variation) of the approximating service time. By using the simulated mean waiting times listed in column 2 and Eq. (1), we solve for  $I(\rho)$  instead of a single  $c_R^2$  or  $I_x$ , which are given in the fifth column. The trend indicates that  $\lim_{\rho \rightarrow 1} I(\rho) \rightarrow I_x = 3.5698$ . It seems that the autocorrelated service time process is “diluted” more with idle times of the server when server utilization decreases and only when  $\rho = 1$  (with no idle time slots inserted between service times) it becomes the same as an autocorrelated same parameter  $TES^+$  interarrival time process and they have the same limiting index of dispersion. Acting on this observation, we fit a power function for  $I(\rho)$  minimizing the least squares error forcing  $I(1) = I_x = 3.5698$ , which yields  $I(\rho) = (I_x - 1)\rho^{1.366372} + 1$ .

When it comes to approximating the mean waiting in the  $GI/TES^+/1$  queue with non-Poisson renewal customer arrivals, we need a service time distribution instead of its first and second moments only (the heavy-traffic formula can be used with the first and second moments but we have seen that this increases the approximation error, hence, we have chosen not to include them in our paper). Araghi and Balcioglu (2008) suggest using the ER renewal approximation for autocorrelated streams having  $I_x \geq 1$ . Since  $I(\rho) \geq 1$  for all  $\rho$  values, we can follow their suggestion for which the utilization dependent service time density function

will have the following Laplace transform (LT)

$$\tilde{b}(s, \rho) = \frac{\mu_x \alpha_E + (\mu_x + A_E(\rho) \alpha_E) s}{\mu_x \alpha_E + (\mu_x + \alpha_E + A_E(\rho) \alpha_E) s + s^2},$$

with  $\mu_x$  as the service rate and  $2A_E(\rho) + 1 = I(\rho)$  guaranteeing that the squared-coefficient of variation is  $I(\rho)$  (see Jagerman et al., 2004). However, our tests indicate that using the  $\alpha_E$  estimated from autocorrelated service times following Balcioğlu et al. (2008) might result in poor estimations. Although, there does not seem to be a method to estimate  $\alpha_E$  (or a function of  $\alpha_E(\rho)$ ) for autocorrelated service times, our extensive numerical experiments show that setting  $\alpha_E = \infty$  gives the most accurate mean waiting time estimates for all  $\rho$  values when customers arrive according to a renewal process. With  $\alpha_E = \infty$ , the LT of the approximating utilization dependent service time can be rewritten as

$$\tilde{b}(s, \rho) = \frac{A_E(\rho)}{1 + A_E(\rho)} + \frac{1}{1 + A_E(\rho)} \frac{\frac{\mu_x}{1 + A_E(\rho)}}{\frac{\mu_x}{1 + A_E(\rho)} + s}. \quad (2)$$

The LT in Eq. (2) is that of the *generalized exponential distribution* (Altıok, 1997, p. 45–46). Adapting to our problem, with probability  $1/(1 + A_E(\rho))$  the generalized exponential distribution is an exponential distribution with rate  $\mu_x/(1 + A_E(\rho))$ , otherwise it is 0.

In Table 2 (and later on in Tables 4, 6), we assess the accuracy of the proposed approximation by assuming that the interarrival times follow a 10-stage Erlang distribution with the density function LT of a  $k$ -stage Erlang (Erlang( $k, \lambda$ )) distribution given by

$$\tilde{a}(s) = \left(\frac{\lambda}{\lambda + s}\right)^k, \quad (3)$$

where  $\lambda = k\mu_x\rho$  with  $\rho$  values listed in column 1. The rationale behind choosing such an arrival process is due to its low squared-coefficient of variation (=0.1), which would magnify the approximation errors. If one considers only the marginal exponential distribution of TES<sup>+</sup> service times, i.e., ignores the autocorrelation in service times, one can attempt to approximate the original  $GI/TES^+/1$  queue by an  $GI/M/1$  queue with service rate 0.8 and the same  $\rho$  as listed in column 1. Using the standard formula (e.g. Eq. 5.65 in Gross and Harris, 1998, p. 263) we obtain the mean waiting time in the  $GI/M/1$  queue as listed in

the third column with prediction errors presented in parentheses. By comparing the rows with the same  $\rho$  values in Tables 1 and 2, we observe that at high utilization levels the degrading impact of positive autocorrelation in service times on mean delay is felt more when interarrival times have less variability than the Poisson arrivals.

Our approximation on the other hand considers a  $GI/GI/1$  queue with the same 10-stage Erlangian interarrival times and the service times with the LT given in Eq. (2). We have computed the mean waiting times in such a  $GI/GI/1$  queue numerically following Abate et al. (1993). These estimates with their approximation errors in parenthesis are listed in column 4 of Table 2. Although at very small utilization levels the errors of our approximation in absolute values are bigger than those of the  $GI/M/1$  queue approximation, for medium to high utilization levels, we conclude that it is highly accurate.

Next we analyze a system for which the  $I_x$  of autocorrelated service times is less than 1. To do this, we superposed mutually independent five Renewal A and five Renewal B streams and recorded the interarrival times in files, which were later on used as autocorrelated service times in our experiments presented in Tables 3 and 4. The characteristics of the superposed streams are as follows:

- **Renewal A:** The interarrival times follow an Erlang distribution with parameters  $k = 2, \lambda = 0.2$  resulting in an arrival rate of  $\lambda_A = 0.1$  and the squared-coefficient of variation  $c_A^2 = 0.5$ .
- **Renewal B:** The interarrival times follow an Erlang distribution with parameters  $k = 3, \lambda = 0.3$  resulting in an arrival rate of  $\lambda_B = 0.1$  and the squared-coefficient of variation  $c_B^2 = 0.334$ .

In Table 3, we consider a single server queueing system receiving Poisson arrivals where the service times are generated according to the procedure described above, which we denote by the  $M/G^-/1$  queue where  $G^-$  refers to the autocorrelated service times in this system having  $\rho_X(1) = -0.08$ . Moreover,  $I_x = 0.417$  in this problem. Here, the first column

shows the server utilization  $\rho$ , whereas the second column lists the simulated mean waiting times with their 95% CI in parenthesis. Irrespective of the server utilization, if a single approximating renewal service time is employed, since the approximating system will be an  $M/GI/1$  queue with the same  $\rho$  as listed in column 1, using  $\mu_x$  and  $I_x$  of the original service times in Eq. (1) we obtain the estimates listed in column 3 with approximation errors provided in parentheses. Although the approximation errors are smaller than the ones listed in Table 1, we see that a single renewal service time underestimates the mean waiting time significantly at lower utilization levels. Yet, the error decreases monotonically with increase in server utilization and gets low above  $\rho = 0.7$ .

Next we are going to construct the approximating service time as a function of  $\rho$ . By using the simulated mean waiting times listed in column 2 and Eq. (1), we solve for  $I(\rho)$ , which are given in the fourth column. The trend indicates that  $\lim_{\rho \rightarrow 1} I(\rho) \rightarrow I_x = 0.417$ . With higher server utilization the impact of negative autocorrelation is felt more and  $I(\rho)$  decreases. When  $\rho = 1$  (with no idle time slots inserted between service times), the autocorrelated service time process becomes the same as an autocorrelated interarrival time process in the form of the superposition of mutually independent five Renewal A and five Renewal B streams and they have the same limiting index of dispersion. Acting on this observation, we fit a power function for  $I(\rho)$  minimizing the least squares error forcing  $I(1) = I_x = 0.417$ , which yields  $I(\rho) = (I_x - 1)\rho^{0.83117} + 1$ .

Since  $I(\rho) \leq 1$  for all  $\rho$  values, following Araghi and Balcioglu (2008), we will employ the GE approximation accommodating the utilization dependent service times with the density function LT

$$\tilde{b}(\rho, s) = p(\rho)\left(\frac{\mu}{\mu + s}\right)^k + (1 - p(\rho))\frac{\mu}{\mu + s}, \quad (4)$$

with  $\mu = \mu_x \{p(\rho)k + (1 - p(\rho))\}$ , and

$$p(\rho) = \frac{-\{2(I(\rho) + 1) - k(k + 1) + 2\} + \sqrt{\{2(I(\rho) + 1) - k(k + 1) + 2\}^2 - 4(I(\rho)^2 - 1)(k - 1)^2}}{2(I(\rho) + 1)(k - 1)^2}.$$

Our numerical studies indicate that  $k = \lceil 1/I(\rho) \rceil$  (smallest integer greater than or equal to  $1/I(\rho)$ ) for all  $\rho$  values seems to minimize the estimation errors when the system receives

non-Poisson renewal arrivals. In Table 4, we consider such a system with interarrival times following a 10-stage Erlang distribution. The server utilizations and the simulated mean waiting times of the  $GI/G^-/1$  queue with 5 renewal A and 5 renewal B streams superposed as service times are listed in columns 1 and 2, respectively. Our approximation makes use of a  $GI/GI/1$  queue with the same 10-stage Erlangian interarrival times and the service times with the LT given in Eq. (4). These estimates with their approximation errors in parentheses are listed in column 3. The errors seem to fluctuate in sign and we do not see a declining trend with increase in server utilization, however, we consider the errors to be reasonably small.

Therefore, we outline the procedure of applying the renewal approximation for autocorrelated service times as follows:

- Step 1.** Following Jagerman et al. (2004) and Balcioglu et al. (2008) analyze the autocorrelated service times to estimate  $\mu_x$  and  $I_x$ .
- Step 2.** Use the autocorrelated service times in the simulation of a single server queue receiving Poisson arrivals by varying the server utilization to obtain the simulated mean waiting time estimates.
- Step 3.** Using Eq. (1), solve for  $I(\rho)$  values. Fit a function for  $I(\rho)$ . If  $I(\rho) \geq 1$ , use Eq. (2), otherwise use Eq. (4) as the density function LT of the approximating the service time.

Note that using a simulation approach in Step 2 can be justified as in Jagerman et al. (2004) where the  $I_x$  of the autocorrelated data is estimated via a different simulation study as well. For certain problems, using analytical models one can obtain the mean waiting time in a single server queue with autocorrelated service times when customers arrive according to a Poisson process. Here, we provide the analysis of such a system when the service time data come from (or can be well approximated by) a 2-state MMPP (see Appendix A for a description of the MMPP). We refer the reader to Meier-Hellstern (1987) for an algorithm on fitting a 2-state MMPP for observed autocorrelated data.

Let the service times come from a 2-state MMPP with parameters

$$Q = \begin{vmatrix} -\sigma_1 & \sigma_1 \\ \sigma_2 & -\sigma_2 \end{vmatrix}, \quad M = \begin{vmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{vmatrix},$$

where  $M$  corresponds to the diagonal matrix  $\Lambda$  in Appendix A with service rates  $\mu_1$  and  $\mu_2$  while the modulating process is in states 1 and 2, respectively. In a  $M/MMPP/1$  queue with Poisson arrival rate  $\lambda$  and 2-state MMPP service times as characterized above, we can use the model due to Yechiali and Naor (1971) by assuming a single Poisson arrival rate and not permitting state transitions when the system is idle. Yet, we summarize the analysis for the convenience of the reader. Accordingly, Eq.s 9a-9d of Yechiali and Naor become

$$\begin{aligned} \lambda P_{0,1} &= \mu_1 P_{1,1}, \\ \lambda P_{0,2} &= \mu_2 P_{1,2}, \\ (\lambda + \mu_1 + \sigma_1) P_{i,1} &= \mu_1 P_{i+1,1} + \lambda P_{i-1,1} + \sigma_2 P_{i,2}, \quad i > 0, \\ (\lambda + \mu_2 + \sigma_2) P_{i,2} &= \mu_2 P_{i+1,2} + \lambda P_{i-1,2} + \sigma_1 P_{i,1}, \quad i > 0, \end{aligned} \quad (5)$$

where  $P_{i,j}$  denotes the steady-state probability of having  $i$  customers in the system when the modulating process is in state  $j = 1, 2$ . Let  $G_j(z) = \sum_{i=0}^{\infty} z^i P_{i,j}$  be the probability generating functions of the system size when state is  $j$  with  $G(z) = G_1(z) + G_2(z)$  denoting the probability generating function of the system size. Multiplying both sides of Eq. (5) by  $z^i$  and summing over all  $i$  yield

$$\begin{aligned} G_1(z)[\lambda(1-z) + \sigma_1 + \mu_1(1-1/z)] - \sigma_2 G_2(z) &= P_{0,1}(\mu_1(1-1/z) + \sigma_1) - P_{0,2}\sigma_2, \\ G_2(z)[\lambda(1-z) + \sigma_2 + \mu_2(1-1/z)] - \sigma_1 G_1(z) &= P_{0,2}(\mu_2(1-1/z) + \sigma_2) - P_{0,1}\sigma_1, \end{aligned}$$

from which we can write

$$G(z) = \frac{P_{0,1}H_1(z) + P_{0,2}H_2(z)}{\lambda^2 z^3 - H_3(z) + H_4(z) - \mu_1 \mu_2}, \quad (6)$$

where

$$\begin{aligned}
H_1(z) &= -\lambda\mu_1z^2 + (\lambda\mu_1 + \mu_1\mu_2 + \mu_1\sigma_2 + \mu_2\sigma_1)z - \mu_1\mu_2, \\
H_2(z) &= -\lambda\mu_2z^2 + (\lambda\mu_2 + \mu_1\mu_2 + \mu_1\sigma_2 + \mu_2\sigma_1)z - \mu_1\mu_2, \\
H_3(z) &= \lambda(\lambda + \sigma_1 + \sigma_2 + \mu_1 + \mu_2)z^2, \\
H_4(z) &= (\lambda(\mu_1 + \mu_2) + \mu_1\sigma_2 + \mu_2\sigma_1 + \mu_1\mu_2)z.
\end{aligned}$$

Observing that  $G(1) = 1$  in Eq. (6), we have

$$P_{0,1} + P_{0,2} = \frac{\sigma_1(\mu_2 - \lambda) + \sigma_2(\mu_1 - \lambda)}{\mu_1\sigma_2 + \mu_2\sigma_1} = 1 - \frac{\lambda(\sigma_1 + \sigma_2)}{\mu_1\sigma_2 + \mu_2\sigma_1}. \quad (7)$$

Note that the denominator  $\lambda^2z^3 - H_3(z) + H_4(z) - \mu_1\mu_2$  of  $G(z)$  has only one root inside the unit circle for any real number  $z$ ,  $|z| \leq 1$ , hence,  $G(z)$  is analytical. If  $z_0$  is the root of  $\lambda^2z^3 - H_3(z) + H_4(z) - \mu_1\mu_2$ , it should be the root of the numerator  $P_{0,1}H_1(z) + P_{0,2}H_2(z)$ , i.e.,  $P_{0,1}H_1(z_0) + P_{0,2}H_2(z_0) = 0$ . Having this in hand and using Eq. (7), we can find  $P_{0,1}$  and  $P_{0,2}$ . This way we obtain  $G(z)$  from which the expected number of customers in the system is found

$$L = \left. \frac{dG(z)}{dz} \right|_{z=1}.$$

Then, using the Little's law the mean delay in this  $M/MMPP/1$  queue is

$$\bar{W} = \frac{L}{\lambda} - \frac{1}{\mu_x}, \quad (8)$$

where  $\mu_x = \lambda/(1 - P_{0,1} - P_{0,2}) = (\mu_1\sigma_2 + \mu_2\sigma_1)/(\sigma_1 + \sigma_2)$ .

In Table 5, we consider the  $M/MMPP/1$  queue where the service times are generated from a 2-state MMPP with parameters

$$Q = \begin{vmatrix} -0.03 & 0.03 \\ 0.16 & -0.16 \end{vmatrix}, \quad M = \text{diag}(0.2, 4),$$

The autocorrelated service times have  $\rho_X(1) = 0.33$  and  $I_X = 26.264$ . Here, the first and second columns show the server utilization and the mean waiting times found from Eq.

(8), respectively. Irrespective of the server utilization, if a single approximating renewal service time is employed, which entails using the same  $\rho$  as listed in column 1 and  $\mu_x$  and  $I_x$  of the original service times in Eq. (1), we obtain the estimates listed in column 3 with approximation errors provided in parentheses. Similar to Table 1, we see that a single renewal service time incurs big errors especially at lower utilization levels but decreases monotonically with increase in server utilization and gets small above  $\rho = 0.7$ .

To construct the approximating service time as a function of  $\rho$ , we use the calculated mean waiting times listed in column 2 and Eq. (1) and solve for  $I(\rho)$ , which are given in the fourth column. The trend indicates that  $\lim_{\rho \rightarrow 1} I(\rho) \rightarrow I_x = 26.264$ . One can repeat the same observation on the diluting impact of idle periods on autocorrelated service times when the server utilization decreases. Next, we fit a logarithmic function for  $I(\rho)$  minimizing the least squares error forcing  $I(1) = I_x = 26.264$ , which yields  $I(\rho) = I_x + 10.60787 \ln(\rho)$ . For  $\rho > 0.093$ ,  $I(\rho) > 1$  and one can use Eq. (2) (otherwise Eq. 4) as the LT of the utilization dependent approximating renewal service time. For instance, in Table 6, we consider such a system with interarrival times following a 10-stage Erlang distribution. The server utilizations and the simulated mean waiting times of  $GI/MMPP/1$  queue are listed in columns 1 and 2, respectively. Our approximation makes use of a  $GI/GI/1$  queue with the same 10-stage Erlangian interarrival times and the service times with the LT given in Eq. (2). These estimates with their approximation errors in parentheses are listed in column 3. Except for very small  $\rho$  values, we consider the errors to be quite small.

### 3 Make-to-Stock Queues with Autocorrelated Service Times

In this section, we analyze two make-to-stock queues in which service times become autocorrelated. In Section 3.1, we study a production/inventory system that works with two suppliers whereas in Section 3.2 a system with production stoppages is considered. In both

problems, we assume that demands occur one at a time, following a renewal arrival process with rate  $\lambda$  and the inventory is controlled according to a base-stock policy with base-stock level  $S$ . Whenever there is available stock in the inventory, demand requests are satisfied right away. Unsatisfied demands are backordered. The system incurs a holding cost of  $h$  per unit in inventory per unit time and a backordering cost of  $b$  per unit backordered per unit time. In this framework, a single server queue with service rate  $\mu_x$  models the production stage. Whenever the inventory level drops below the base-stock level  $S$ , the server becomes busy, hence, the production starts. It stops and the server becomes idle only when the inventory level reaches  $S$  again.

As a result of this structure, each demand request results in a production order, which is mapped as a customer arrival in the underlying single server queueing system. Hence,  $N(t)$ , the number of customers present in the queueing system at time  $t$ , gives the shortfall from the base-stock level  $S$ . This implies that when  $N(t) \leq S$ , the inventory will be carrying  $S - N(t)$  units and when  $N(t) > S$ , the system will be running with  $N(t) - S$  units backordered. We assume that  $\rho = \lambda/\mu_x < 1$ .

Our approximation for the original make-to-stock queues makes use of the  $GI/GI/1$  queue studied by Sanajian and Balcioğlu (2009) employing the approximating service times with a LT denoted by  $\tilde{b}(s, \rho)$  obtained following the procedure presented in Section 2. In this  $GI/GI/1$  queue if one obtains the steady-state probability of having  $n$  customers in the system, namely  $p(n) = P(N = n)$ , the problem of finding the optimal  $S$  can be consequently expressed as follows:

$$\min_S \{C(S) = h \sum_{n=0}^S (S - n)p(n) + b \sum_{n=S+1}^{\infty} (n - S)p(n)\}. \quad (9)$$

Note that the right hand side of Eq. (9) is nothing but the expected inventory cost per unit time. Letting  $F(n) = P(N \leq n)$  denote the cumulative distribution function of the number in the system, the optimal base-stock level  $S^*$  minimizing Eq. (9) is found as (see Veatch

and Wein, 1996)

$$S^* = \min\{n : F(n) > b/(h + b)\}. \quad (10)$$

We refer the reader to Sanajian and Balcioglu (2009) on finding  $S^*$  and  $C(S^*)$  when demand arises according to a non-Poisson renewal process and general i.i.d. service times.

### 3.1 Make-to-Stock Queues with Two Suppliers

Consider a production/inventory system processing raw material that is received from two different suppliers, supplier F and supplier G, to produce a single unit. We assume that there is always abundant raw material in the receiving warehouse so that no production stoppages arise due to lack of raw material. Note that the infinite raw material is an implicitly yet commonly made assumption while studying make-to-stock queues. The production facility makes use of one piece of raw material (let us say a single set or a single bundle of raw material) for each product it produces. When a new product is going to be produced, the production stage pulls the next raw material bundle in line, i.e., ignores the supplier information. Yet, the quality of the suppliers might be different affecting the service times. Hence, if the raw material bundle picked up has come from Supplier F the production times will be denoted by a general i.i.d. r.v.  $G_1$ , which could be different from the general i.i.d. production times  $G_2$  if the bundle has come from supplier G. We assume that if the last bundle picked up was from Supplier F, the probability that the next bundle will be from Supplier F (Supplier G) is denoted by  $p_{FF}$  ( $p_{FG}=1-p_{FF}$ ). Similarly, if the last bundle picked up was from Supplier G, the probability that the next bundle will be from Supplier G (Supplier F) is denoted by  $p_{GG}$  ( $p_{GF}=1-p_{GG}$ ).

Analyzing such a system to obtain the system size distribution for Eq.s (9-10) is quite difficult. Only in the special case when the customers arrive according to a Poisson process and the service times are exponential, it reduces to the system analyzed by Zhou and Gans (1999). Note that Zhou and Gans refer to such a system as the  $M/MMPP/1$  queue, however,

the MMPP service times in their study is different from the MMPP discussed in Appendix A and Section 2 since in this case the state (i.e. the supplier who sent the raw material bundle) can change only at service completion instants.

However, when service times are non-exponential, the analytical model due to Zhou and Gans can incur big prediction errors. Consider Table 7 where  $p_{FG} = p_{GF} = 0.15$ . Service time r.v. is  $G_1$ , which is deterministic(=0.02), if the raw material has come from Supplier F, and  $G_2$ , which follows Erlang(5,40) distribution, otherwise. Since the customers arrive according to a Poisson process, we denote this system by the  $M/G_1 + G_2/1$  queue. The autocorrelated service times in this system have  $\rho_X(1) = 0.44$  and  $I_X = 3.388$  (estimated following Jagerman et al., 2004). In Table 7, the first column shows the server utilization  $\rho$ , whereas the second column lists the simulated mean waiting times with their 95% CI in parenthesis. If exponential service times with rates 50 and 8 are used instead of  $G_1$  and  $G_2$ , respectively, one can use the model due to Zhou and Gans to compute the mean waiting times presented in column 3. With exponential service times instead of  $G_1$  and  $G_2$ , service times become autocorrelated with an estimated  $\rho_X(1) = 0.18$ . This model could serve as an approximation to the original system, however, while the approximation errors tend to decrease with the increase in server utilization, using mean preserving exponential service times instead of  $G_1$  and  $G_2$  and applying the model due to Zhou and Gans incurs big error even when  $\rho = 0.9$ .

Hence, we propose to use the procedure outlined in Section 2 to obtain the approximating service time as a function of  $\rho$ . By using the simulated mean waiting times listed in column 2 and Eq. (1), we solve for  $I(\rho)$ , which are given in the fourth column. The trend indicates that  $\lim_{\rho \rightarrow 1} I(\rho) \rightarrow I_X = 3.388$ . Finally, we fit a power function for  $I(\rho)$  minimizing the least squares error forcing  $I(1) = I_X = 3.388$ , which yields  $I(\rho) = (I_X - 1)\rho^{1.7876} + 1$ . Since  $I(\rho) \geq 1$  for all  $\rho$  values, we will use Eq. (2) as the LT of the approximating renewal service times.

To test if our renewal approximation is accurate in capturing the behavior of a make-to-stock queue, we consider the  $GI/G_1 + G_2/1$  make-to-stock system in Table 8 where interar-

rival times between demand requests follow a 2-stage Erlang distribution. Here, columns 1 and 2 list the  $b/h$  ratio (with  $h = 1$  in all make-to-stock examples) and the server utilization, respectively. Since we do not have an analytic model for the  $GI/G_1 + G_2/1$  queue,  $S^*$  and  $C(S^*)$  (with its 95% CI in parenthesis) are found via simulation and presented in columns 3 and 4, respectively. Our approximation employs the  $GI/GI/1$  queue studied by Sanajian and Balcioglu (2009) where the customer interarrival times are the same as in the original system and the approximating renewal service times that we have characterized above. The approximated  $S^*$  and  $C(S^*)$  (with its approximation error in parenthesis) are listed in columns 5 and 6, respectively. As can be seen in Table 8, our approximation gives sub-optimal base-stock level only when  $\rho = 0.9$  while in other cases it correctly captures it. Additionally, the predicted system costs turn out to be very close to their corresponding optimal system costs.

In this numerical example, we would like to see how erroneous it would be to ignore the autocorrelation in service times. To do this, we have collected the service time data from simulation and assuming that they were i.i.d. r.v.s, we have fit a service time that has the LT given in Eq. (4) with parameters  $p = 0.4286$ ,  $\mu = 19.70499$  and  $k = 2$ . We have tested the goodness-of-fit of this distribution using the Kolmogorov-Smirnov test at  $\alpha = 0.1$  and failed to reject it and concluded that it fits the data well if autocorrelation is ignored. Using this service time in the  $GI/GI/1$  queue with the same interarrival time distribution, we find  $S^*$  and  $C(S^*)$  values listed in columns 7 and 8, respectively. We see that ignoring positive autocorrelation results in base-stock levels and system costs to be underestimated significantly. This is an example of why autocorrelation (if any) in service times should be incorporated in the analysis while determining optimal base-stock level and cost in a make-to-stock queue.

### 3.2 Make-to-Stock Queues with Production Interruption

In this section, we study a production system subject to interruptions due to failures. Processing times in the absence of failures would be general i.i.d. r.v.s. The production stage encounters interruptions provided that there are production orders in the system. The production is interfered throughout repair time that starts immediately after a failure. Repair times are assumed to be general i.i.d. r.v.s that are independent of the failure process. Once the repair is completed, production resumes operation from the point of interruption. Such systems have been studied in the literature by carrying out a *process completion time* analysis (Gaver, 1962 and Avi-Itzhak and Naor, 1963) assuming that times to failure are exponentially distributed. The process completion time is defined as the total time a job spends in production until its production is completed, hence, it includes repair times that a job has to wait for. When times to failure are not exponentially distributed the process completion time (PCT) sequence can become autocorrelated. Especially if more than one interruption during the processing time of a job can be observed, the autocorrelation in PCT can become significant. In the literature, general times to failure have been considered with restrictive assumptions. When customers arrive according to a Poisson process, Federgruen and Green (1986) propose approximations for the steady-state distribution of the jobs in the system. Later, with Poisson job arrivals and repair times following Normal and Gamma distributions, Federgruen and Green (1986) provide the exact analysis if times to failure follow phase-type distributions. Balcioğlu et al. (2007) provide the exact analysis when customers arrive according to a Poisson process considering general repair times with phase-type times to failures. Then they propose an accurate approximation if customer arrive according to a renewal process. Yet, in their analysis the processing times are restricted to be deterministic.

This discussion shows that although production stoppages are commonly observed, exact solutions for the PCT analysis is quite difficult. However, with our utilization based renewal service time approximation this problem can be handled easily. In Table 9, we consider a system subject to interruptions. Processing times are Erlang(5,5) r.v.s while times to failure

follow a Weibull distribution with parameters  $\alpha = 1.667$  and  $\beta = 0.5$  (Weibull( $\alpha, \beta$ )) with the density function

$$f_w(t) = \frac{\beta}{\alpha} \left(\frac{t}{\alpha}\right)^{\beta-1} \exp\left(-\left(\frac{t}{\alpha}\right)^\beta\right), t \geq 0,$$

which makes the autocorrelated PCT's have a  $\rho_X(1) = -0.14326$  and  $I_x = 0.2844$  (estimated following Jagerman et al., 2004). In Table 9, the first column shows the server utilization  $\rho$ , whereas the second column lists the simulated mean waiting times with their 95% CI in parenthesis. Next we employ the procedure outlined in Section 2 to obtain the approximating service time as a function of  $\rho$ . By using the simulated mean waiting times listed in column 2 and Eq. (1), we solve for  $I(\rho)$ , which are given in the third column. The trend indicates that  $\lim_{\rho \rightarrow 1} I(\rho) \rightarrow I_x = 0.2844$ . Finally, we fit a power function for  $I(\rho)$  minimizing the least squares error forcing  $I(1) = I_x = 0.2844$ , which yields  $I(\rho) = (I_x - 1)\rho^{0.154254} + 1$ . Since  $I(\rho) \leq 1$  for all  $\rho$  values, we will use Eq. (4) as the LT of the approximating renewal service times.

To test if our renewal approximation is accurate in capturing the behavior of a make-to-stock queue, we consider the  $GI/G/1$  make-to-stock system in Table 10 where interarrival times between demand requests follow a 2-stage Erlang distribution. Here, columns 1 and 2 list the  $b/h$  ratio and the server utilization, respectively. Similar to Section 3.1,  $S^*$  and  $C(S^*)$  (with its 95% CI in parenthesis) are found via simulation and presented in columns 3 and 4, respectively. Our approximation yields  $S^*$  and  $C(S^*)$  (with its approximation error in parenthesis) as listed in columns 5 and 6, respectively. As can be seen in Table 10, our approximation gives the optimal  $S^*$  in all the cases and incurs little error in predicting optimal system costs.

In this numerical example, we would like to see how erroneous it would be to ignore the autocorrelation in service times. To do this, we have collected the service time data from simulation and assuming that they were i.i.d. r.v.s, we have fit a service time that has the LT given in Eq. (4) with parameters  $p = 0.82734$ ,  $\mu = 1.76978$  and  $k = 3$ . We have tested the

goodness-of-fit of this distribution using the Kolmogorov-Smirnov test at  $\alpha = 0.1$  and failed to reject it and concluded that it fits the data well if autocorrelation is ignored. Using this service time in the  $GI/GI/1$  queue with the same interarrival time distribution, we obtain  $S^*$  and  $C(S^*)$  values listed in columns 7 and 8, respectively. We see that ignoring negative autocorrelation results in base-stock levels and system costs to be overestimated significantly especially when server utilization is high and  $b/h = 80$ . This is an example showing that optimal base-stock levels and system costs could be less than one would predict if negative autocorrelation in PCT were ignored.

## 4 Conclusion

In this paper, we have proposed adapting two renewal approximations from the literature to incorporate autocorrelated service times in analytical models. The empirical observations are that instead of a single approximating renewal service time, a family of service times as a function of server utilization should be fit. Although we assume that the autocorrelated service times are independent of the customer arrival process, the idle periods inserted between sampled service times, which is a byproduct of the traffic intensity, decreases the impact of autocorrelation at lower server utilization levels. We have tested our approximation both in predicting the mean delay and the system size distribution of the original systems with autocorrelated service times, which proved to be quite accurate.

We note that this is the first attempt to approach this general problem in a systematic way. Our promising results assure us that further research is necessary in many directions. For instance, unlike the restrictive forms for the service time function we consider, more flexible forms can increase the accuracy more. One of the more important questions worth pursuing is whether the step of the proposed procedure, which requires a simulation analysis, can be bypassed for any autocorrelated service time data, similar to what we have done with the MMPP service times. Additionally, the utilization dependent renewal approximation idea can be studied for approximating autocorrelated arrival processes, which helps approximating

the system size distribution accurately, a goal that has been avoided in earlier research.

## Acknowledgements

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

## References

- Abate, J., G. L. Choudhury, and W. Whitt. 1993. "Calculation of the  $GI/G/1$  Waiting-Time Distribution and Its Cumulants from Pollaczek's Formulas", *Archiv für Elektronik und Übertragungstechnik*, Vol. 47, No. 5/6, 311–321.
- Albin, S. 1984. "Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes to Queues", *Operations Research*, Vol. 32, No. 5, 1133–1162.
- Altıok, T. 1997. *Performance Analysis of Manufacturing Systems*, Springer-Verlag, NY.
- Altıok, T. and B. Melamed. 2001. "The Case for Modeling Correlation in Manufacturing Systems," *IIE Transactions*, Vol. 33, 779–791.
- Araghi, M. and B. Balcioglu. 2008. "A New Renewal Approximation for Certain Autocorrelated Processes", *OR Letters*, Vol. 36, No. 1, 133–139.
- Avi-Itzhak, B., and P. Naor. 1963. "Some Queueing Problems with the Service Station Subject to Breakdown," *Operations Research*, Vol. 11, 303–320.
- Balcioglu, B., D. Jagerman and T. Altıok. 2007. "Approximate Mean Waiting Time in a  $GI/D/1$  Queue with Autocorrelated Times to Failures", *IIE Transactions*, Vol. 39, No. 10, 985–996.
- Balcioglu, B., D. L. Jagerman, and T. Altıok. 2008. "Merging and Splitting Autocorrelated Arrival Processes and Impact on Queueing Performance", *Performance Evaluation* Vol. 65, 653–669.

- Baykal-Gürsoy, M. and W. Xiao. 2004. “Stochastic Decomposition in  $M/M/\infty$  Queues with Markov Modulated Service Rates”, *Queueing Systems*, Vol. 48, 75–88.
- Baykal-Gürsoy, M., W. Xiao, and K. Ozbay. 2009. “Modeling traffic flow interrupted by incidents”, *European Journal of Operational Research*, Vol. 195, 127–138.
- Boxma, O. J. and I. A. Kurkova. 2001. “The  $M/G/1$  Queue with Two Service Speeds,” *Advance Applied Probability*, Vol. 33, 520–540.
- Cox, D. R. and P. A. W. Lewis. 1968. *The Statistical Analysis of Series of Events*, Methuen, London, 78.
- Federgruen, A., and L. Green, 1986. “Queueing Systems with Service Interruptions,” *Operations Research*, Vol. 34, 752–768.
- Federgruen, A., and L. Green. 1988. “Queueing Systems with Service Interruptions II,” *Naval Research Logistics*, Vol. 35, 345–358.
- Fendick, K. W., V. R. Saksena and W. Whitt. 1989. “Dependence in Packet Queues,” *IEEE Trans. on Comm.*, Vol. 37, 1173–1183.
- Fischer, W. and K. Meier-Hellstern. 1992. “The Markov-modulated Poisson process (MMPP) cookbook”, *Performance Evaluation*, Vol. 18, 149–171.
- Gaver, D. P. 1962. “A Waiting Line with Interrupted Service Including Priorities,” *Journal of the Royal Statistical Society*, B24, 73–90.
- Gross, D. and C. M., Harris. 1998. *Fundamentals of Queueing Theory*, John Wiley & Sons, New York.
- Gusella, R. 1991. “Characterizing the Variability of Arrival Processes With Indexes of Dispersion”, *IEEE J. on Selected Areas in Communications*, Vol. 9, No. 2, 203–211.
- Jagerman D. L., and B. Melamed. 1992a. “The Transition and Autocorrelation Structure of TES Processes Part I: General Theory”, *Stochastic Models* **8(2)**, 193–219.
- Jagerman D. L., and B. Melamed. 1992b. “The Transition and Autocorrelation Structure of TES Processes Part II: Special Cases”, *Stochastic Models* **8(3)**, 499–527.

- Jagerman, D. L., B. Balcioglu, T. Altiok and B. Melamed. 2004. “Mean Waiting Time Approximations in the  $G/G/1$  Queue”, *Queueing Systems*, Vol. 46, No. 3, 481–506.
- Livny, M., B. Melamed and A. K. Tsolis. 1993. “The Impact of Autocorrelation on Queuing Systems,” *Management Science*, Vol. 39, No. 3, 322–339.
- Mahabhashyam, S. and N. Gautam. 2005. “On Queues with Markov Modulated Service Rates”, *Queueing Systems*, Vol. 51, 89–113.
- Meier-Hellstern, K. S. 1987. “A Fitting Algorithm for Markov-modulated Poisson Processes Having Two Arrival Rates,” *European Journal of Operational Research*, Vol. 29, 370–377.
- Melamed, B. 1991. “TES: A Class of Methods for Generating Autocorrelated Uniform Variates”, *ORSA J. on Computing* **3**, 317–329.
- Melamed, B. 1993. “An Overview of TES Processes and Modeling Methodology”, in *Performance Evaluation of Computer and Communications Systems* (L. Donatiello and R. Nelson, Editors), 359–393, *Lecture Notes in Computer Science*, Springer-Verlag.
- Mi. N., Q. Zhang, A. Riska, E. Smirni, and E. Riedel. 2007. “Performance impacts of autocorrelated flows in multi-tiered systems”, *Performance Evaluation*, Vol. 64, 1082–1101.
- Sanajian, N., and, B. Balcioglu. 2009. “The Impact of Production Time Variability on Make-to-Stock Queue Performance,” *European Journal of Operational Research*, Vol. 194, 847–855.
- Takine, T. 2002. “Queue Length Distribution in a FIFO Single-Server Queue with Multiple Arrival Streams Having Different Service Time Distributions,” *Queueing Systems*, Vol. 39, 349–375.
- Veatch, M., and L. M. Wein. 1996. “Scheduling a Make-to-Stock Queue: Index Policies and Hedging Points”, *Operations Research*, Vol. 44, No. 4, 634–647.
- Whitt, W. 1982. “Approximating a Point Process by a Renewal Process, I: Two Basic Methods”, *Operations Research*, Vol. 30, No. 1, 125–147.
- Yechiali, U. and P. Naor. 1971. “Queuing Problems with Heterogeneous Arrivals and Service”, *Operations Research*, Vol. 19, No. 3, 722–734.

Zhou, Y-P. and N. Gans. 1999. "A Single-Server Queue with Markov Modulated Service Times", *working paper*, The Wharton School, University of Pennsylvania.

Zipkin, P. H. 1995. "Performance Analysis of a Multi-Item Production-Inventory System under Alternative Policies," *Management Science*, Vol. 41, No. 4, 690–703.

## Appendix A Autocorrelated Processes

We have employed the Transform-Expand-Sample (TES<sup>+</sup>) process and Markov-modulated Poisson processes (MMPP) in generating autocorrelated service times in Section 2. In this section, we will summarize these processes.

The TES<sup>+</sup> considered in this table has marginal exponential distribution characterized by a parameter triplet  $(L, R, \mu_x)$  (Jagerman and Melamed, 1992a, 1992b). While  $\mu_x$  is the rate of the exponentially distributed service times,  $[L, R)$  subject to  $-0.5 \leq L \leq R < 0.5$ , is named as the support interval. The decrease in the length of the support interval induces more positive autocorrelation in the interruption process (Melamed, 93). In particular, the lag-1 autocorrelation,  $\rho_x(1)$ , of the service times ranges from 0 (for  $L = -0.5$  and  $R = 0.5$ ) to 1 (for  $L = R = 0$ ) (Melamed, 1991).

The MMPP have been widely used in modeling arrival processes with arrival rates varying randomly over time and autocorrelations between the interarrival times (we refer the reader to Fischer and Meier-Hellstern (1992) for a thorough investigation of the MMPP).

An  $m$ -state MMPP is a doubly stochastic Poisson process whose arrival rate at a certain time  $t$  depends on the state of an  $m$ -state modulating Markov process. It is defined by an  $m \times m$  infinitesimal generator  $Q$  of the underlying process and an  $m \times m$  diagonal matrix  $\Lambda$  with diagonal elements  $\lambda_i, i = 1, \dots, m$ . When the modulating Markov chain is in state  $i$ , arrivals will follow a Poisson process with rate  $\lambda_i$ . The length of time that the Markov process spends in state  $i$  before making a transition out of that state is exponentially distributed with rate  $-Q_{ii}$ . The probability that the Markov process enters state  $j$  after leaving state  $i$

is given by  $-Q_{ij}/Q_{ii}$ .

The steady-state probability vector  $\Pi$  of the Markov chain is found by solving

$$\Pi Q = 0, \quad \Pi e = 1,$$

where  $e = (1, \dots, 1)^T$  is the column vector of length  $m$ . The mean arrival rate is  $\lambda_X = \Pi \Lambda e$ .

$\rho$	Simulation $\bar{W}$	$M/M/1 \bar{W}$	PK $\bar{W}$	$I(\rho)$
0.2	0.3525(+/-0)	0.3125(-11%)	0.7140(103%)	1.256
0.3	0.6547(+/-0)	0.5357(-18%)	1.2241(87%)	1.4442
0.4	1.1178(+/-0)	0.8333(-25%)	1.9041(70%)	1.6827
0.5	1.857(+/-0.01)	1.25(-33%)	2.8561(54%)	1.9712
0.6	3.0888(+/-0.01)	1.875(-39%)	4.2842(39%)	2.2947
0.7	5.3054(+/-0.03)	2.9167(-45%)	6.6643(26%)	2.6379
0.8	9.942(+/-0.07)	5(-50%)	11.4245(15%)	2.9768
0.9	24.2024(+/-0.26)	11.25(-54%)	25.7051(6%)	3.3026

Table 1: Finding  $I(\rho)$  for the  $M/TES^+/1$  queue with  $TES^+$  service times with  $L = -0.2$ ,  $R = 0.2$ ,  $\mu_x = 0.8$  and the limiting index of dispersion  $I_x = 3.5698$

$\rho$	Simulation $\bar{W}$	$GI/M/1 \bar{W}$	Approximation $\bar{W}$
0.2	0.0249(+/-0)	0.0234(-6%)	0.037(49%)
0.3	0.1074(+/-0)	0.0892(-17%)	0.162(51%)
0.4	0.3007(+/-0)	0.2088(-31%)	0.429(43%)
0.5	0.7168(+/-0)	0.4053(-43%)	0.922(29%)
0.6	1.569(+/-0.01)	0.7231(-54%)	1.803(15%)
0.7	3.2669(+/-0.02)	1.2702(-61%)	3.45(6%)
0.8	6.9598(+/-0.06)	2.3953(-66%)	7.008(1%)
0.9	18.438(+/-0.22)	5.7556(-69%)	18.21(-1%)

Table 2: Comparison of equilibrium mean waiting times in the  $GI/TES^+/1$  queue with  $TES^+$  service times with  $L = -0.2$ ,  $R = 0.2$ ,  $\mu_x = 0.8$  and 10-stage Erlang interarrival times

$\rho$	Simulation $\bar{W}$	PK $\bar{W}$	$I(\rho)$
0.2	0.2245(+/-0)	0.1771(-21%)	0.796
0.3	0.3783(+/-0)	0.3036(-20%)	0.7654
0.4	0.5768(+/-0)	0.4723(-18%)	0.7304
0.5	0.8454(+/-0)	0.7085(-16%)	0.6908
0.6	1.2343(+/-0)	1.0628(-14%)	0.6457
0.7	1.8611(+/-0)	1.6532(-11%)	0.595
0.8	3.0791(+/-0.01)	2.8340(-8%)	0.5395
0.9	6.6497(+/-0.04)	6.3765(-4%)	0.4777

Table 3: Finding  $I(\rho)$  for the  $M/G^-/1$  queue with 5 renewal A and 5 renewal B streams superposed as service times with rate  $\mu_x = 1$  and the limiting index of dispersion  $I_x = 0.417$

$\rho$	Simulation $\bar{W}$	Approximation $\bar{W}$
0.2	0.0103(+/-0)	0.012(17%)
0.3	0.0462(+/-0)	0.044(-5%)
0.4	0.1146(+/-0)	0.102(-11%)
0.5	0.2232(+/-0)	0.191(-14%)
0.6	0.3892(+/-0)	0.328(-16%)
0.7	0.6551(+/-0)	0.556(-15%)
0.8	1.1496(+/-0)	1.003(-13%)
0.9	2.5164(+/-0.01)	2.314(8%)

Table 4: Comparison of equilibrium mean waiting times in the  $GI/G^-/1$  queue with 5 renewal A and 5 renewal B streams superposed as service times with service rate  $\mu = 1$  and 10-stage Erlang interarrival times

$\rho$	$\bar{W}$	PK $\bar{W}$	$I(\rho)$
0.2	1.614	4.26(164%)	9.3296
0.3	3.714	7.3029(97%)	12.8656
0.4	7.222	11.36(57%)	16.3328
0.5	12.596	17.04(35)	19.1536
0.6	20.932	25.56(22%)	21.3275
0.7	35.012	39.76(14%)	23.0082
0.8	63.329	68.16(8%)	24.3316
0.9	148.466	153.36(3%)	25.3939

Table 5: Finding  $I(\rho)$  for the  $M/MMPP/1$  queue with 2-state MMPP service times with service rate  $\mu_x = 0.8$  and marginal Index of dispersion  $I_x = 26.264$

$\rho$	Simulation $\bar{W}$	Approximation $\bar{W}$
0.2	0.7938(+/-0)	0.993(25%)
0.3	2.6086(+/-0.02)	3.128(20%)
0.4	6.0091(+/-0.05)	6.403(7%)
0.5	11.2933(+/-0.11)	11.341(0%)
0.6	19.509(+/-0.22)	19.09(-2%)
0.7	33.4402(+/-0.47)	32.38(-3%)
0.8	61.5222(+/-1.21)	59.441(-3%)
0.9	145.32(+/-5.75)	141.457(-3%)

Table 6: Comparison of equilibrium mean waiting times in the  $GI/MMPP/1$  queue with 2-state MMPP service times with service rate  $\mu_x = 0.8$  and and 10-stage Erlang interarrival times

$\rho$	Simulation $\bar{W}$	Zhou-Gans Approximation $\bar{W}$	$I(\rho)$
0.2	0.0186 (+/-0.00)	0.0309 (66.28%)	1.048
0.3	0.0344 (+/-0.00)	0.0564 (63.89%)	1.215
0.4	0.0587 (+/-0.00)	0.0983 (67.52%)	1.428
0.5	0.0976 (+/-0.00)	0.1507 (54.41%)	1.692
0.6	0.1631 (+/-0.00)	0.2415 (48.07%)	1.999
0.7	0.2816 (+/-0.00)	0.3993 (41.80%)	2.329
0.8	0.5297 (+/-0.00)	0.7235 (36.59%)	2.653
0.9	1.2761 (+/-0.02)	1.7167 (34.53%)	2.911

Table 7: Finding  $I(\rho)$  for the  $M/G_1+G_2/1$  queue with service rate  $\mu_x = 13.793$  and marginal Index of dispersion  $I_x = 3.388$

$b/h$	$\rho$	Simulation		Approximation		Ignoring Autocorrelation	
		$S^*$	$C(S^*)$	$S^*$	$C(S^*)$	$S^*$	$C(S^*)$
10	0.7	8	8.3533 (+/-0.040)	8	8.3518 (-0.02%)	4	4.3677 (-47.71%)
	0.8	15	15.777 (+/-0.130)	15	15.5077 (-1.71%)	7	6.9962 (-55.66%)
	0.9	36	37.791 (+/-0.818)	37	38.1682 (1.00%)	15	14.9117 (-60.54%)
80	0.7	15	15.333 (+/-0.109)	15	15.3556 (0.15%)	8	7.9547 (-48.12%)
	0.8	28	29.006 (+/-0.479)	28	28.4864 (-1.85%)	13	12.8135 (-55.82%)
	0.9	68	69.164 (+/-2.751)	69	69.9845 (1.19%)	27	27.3153 (-60.51%)

Table 8: Comparison of  $S^*$  and  $C(S^*)$  in a  $GI/G_1+G_2/1$  make-to-stock system with service rate  $\mu_x = 13.793$  and 2-stage Erlang interarrival times

$\rho$	Simulation $\bar{W}$	$I(\rho)$
0.2	0.2670 (+/-0.00)	0.4239
0.3	0.4518 (+/-0.00)	0.4056
0.4	0.6929 (+/-0.00)	0.3858
0.5	1.0260 (+/-0.00)	0.3679
0.6	1.5187 (+/-0.00)	0.3499
0.7	2.3296 (+/-0.01)	0.3312
0.8	3.9438 (+/-0.01)	0.3146
0.9	8.7648 (+/-0.09)	0.2985

Table 9: Finding  $I(\rho)$  for the  $M/G/1$  queue with Erlang(5,5) processing times, Weibul(1.667,0.5) times to failure and Uniform(17/12,23/12) repair times in which PCT's have a rate  $\mu_x = 0.6667$  and the limiting index of dispersion  $I_x = 0.2844$

$b/h$	$\rho$	Simulation		Approximation		Ignoring Autocorrelation	
		$S^*$	$C(S^*)$	$S^*$	$C(S^*)$	$S^*$	$C(S^*)$
10	0.7	3	2.8571 (+/-0.008)	3	2.8720 (0.52%)	3	3.3680 (17.88%)
	0.8	5	4.4094 (+/-0.014)	5	4.4615 (1.18%)	5	5.2405 (18.85%)
	0.9	9	8.9996 (+/-0.121)	9	9.1181 (1.32%)	11	10.8736 (20.82%)
80	0.7	5	5.1156 (+/-0.028)	5	5.2131 (1.91%)	6	5.9072 (15.47%)
	0.8	8	7.9337 (+/-0.074)	8	8.0521 (1.49%)	10	9.4524 (19.14%)
	0.9	17	16.474 (+/-0.434)	17	16.6534 (1.09%)	20	19.9006 (20.80%)

Table 10: in a  $GI/G/1$  make-to-stock system with service rate  $\mu_x = 0.6667$  and 2-stage Erlang interarrival times