Queues with Dependency between Interarrival and Service times using Mixture of Bivariates

Srikanth K. Iyer, Indian Institute of Science, Bangalore
D. Manjunath, Indian Institute of Technology, Bombay

January 17, 2005

Abstract

We analyze queueing models where the joint density of the interarrival time and the service time is described by a mixture of joint densities. These models occur naturally in multiclass populations serviced by a single server through a single queue. Other motivations for this model are to model the dependency between the interarrival and service times and consider queue control models. Performance models with component heavy tailed distributions that arise in communication networks are difficult to analyze. However, long tailed distributions can be approximated using a finite mixture of exponentials. Thus, the models analyzed here provide a tool for the study of performance models with heavy tailed distributions. The joint density of A and X, the interarrival and service times respectively, $f(a,x)$, will be of the form $f(a,x) = \sum_{i=1}^{M} p_i f_i(a,x)$ where $p_i > 0$ and $\sum_{i=1}^{M} p_i = 1$. We derive the Laplace Stieltjes Transform of the waiting time distribution. We also present and discuss some numerical examples to describe the effect of the various parameters of the model.

1 Introduction

Let $\{A_n, X_n\}$ be the input sequence to a queue where A_n is the interarrival time between the n^{th} and $(n+1)^{\text{th}}$ customer and X_n is the service time of the n^{th} customer. Traditional queueing theory assumes X_n and A_n to be independent random variables. However, there has been an abiding interest in queues for which the X_n and A_n are not independent and there have been regular studies in this direction. Relaxing the assumption about independence of X_n and A_n is especially useful in the study of network subsystems in packet communication networks which has been an important application of queueing models. Among the first studies that relaxed the independence assumption of A_n and X_n are papers by Conolly [4], Conolly and Hadidi [5] and Conolly and Choo [6]. Fendick, Saksena and Whitt [9] give a review of the various types of dependencies that exist in packet queues in packet communication networks and their impact on system performance. With similar motivation, Cidon *et. al* [3], study a family of queues with random proportional dependence and an independent, additive component in describing the relation between A_n and X_n. In a previous study [2], Cidon *et al* also considered a model where A_{n-1} and X_n were dependent. More recently, Müller [13] generalized the result of Chao [1] and showed that stronger dependence between the A_n and X_n decreases the waiting times in the convex ordering sense. We refer the reader to [2, 1, 3, 13] for detailed discussions and further references on queues where the interarrival and service times are dependent. As can be seen from these references and the references therein, there has been a continuing interest in analyzing single server queues with cross correlations between the A_n and X_n. However, a generalized model with explicit expressions for waiting time distributions has been elusive.

In this paper we consider a finite mixture of bivariate distributions for the i.i.d. sequence $\{A_n, X_n\}$. Specifically, we consider models where the joint density of A_n and X_n, $f(a,x)$, is of the form

$$f(a,x) = \sum_{i=1}^{M} p_i f_i(a,x), \tag{1}$$

where $p_i > 0$ and $\sum_{i=1}^{M} p_i = 1$. We further simplify the RHS above to $f_i(a,x) = g_i(a)h_i(x)$.

There are many motivations in considering this model. Modeling the $\{A_n, X_n\}$ as above allows us to introduce fairly arbitrary correlations and also model non-linear dependencies between A_n and X_n while specifying marginals for A_n and X_n. This also helps us to consider a more general class of models than that of [3]. The models of [3] are quite restrictive in the cross correlations that can be specified between A_n and X_n and hence, we believe, limited in their applicability. Our model can be used to specify a wide range of positive and negative cross correlations. In fact, the models that we specify arise naturally in the presence of multiclass customers taking service from a single server that maintains a single queue like, for example, request arrivals to web servers and the packet arrival process to a node in a packet communication system. The kind of models that we consider can also be used in randomized control of queues, where, by controlling the p_i, different classes of customers may be given different grades of service. Cidon *et al* [3] describe a spacer controller application which is also applicable for our models. Finally, the mixture distributions model has a well developed theory of identification and parameter estimation [16] which can be applied in the characterization of actual systems. Finally, as we show below, bivariate heavy tail distributions can be well approximated by mixture distributions of the form (1). This assumes significance in light of the recent interest in modeling packet traffic processes by heavy tail distributions [12, 14]. In another application, Ramaswami [15] uses a mixtures of exponentials to model the interarrival and holding time distributions of Internet dial-up calls.

1.1 Mixture Approximations for Bivariate Heavy Tail Distributions

Feldmann and Whitt [7] provide an algorithm to fit a finite mixture of exponentials to a heavy tailed distribution. We extend their result to a class of dependent bivariate heavy tail distributions. We also provide two examples of heavy tailed distributions that can be so approximated.

A bivariate probability density function (pdf) f will be called completely monotone if partial derivatives of f of all orders exist and

$$(-1)^n \frac{\partial^n f(x,y)}{\partial x^k \partial y^{n-k}} \geq 0, \qquad \text{for all } x, y > 0, n \geq 1 \text{ and } k = 0, \ldots, n.$$

The following theorem provides the link between completely monotone pdf's and mixtures of exponential. It is an extension of Bernstein's theorem [8] to the bivariate case.

Theorem 1. Every bivariate completely monotone pdf f can be represented as follows:

$$f(x,y) = \int_0^\infty \int_0^\infty \lambda\mu e^{-\lambda x - \mu y} dG(\lambda, \mu), \qquad x, y \geq 0,$$

for some proper bivariate distribution function G on $(0, \infty)^2$.

The proof of this theorem is a straightforward application of the multivariate Taylor's series expansion and is given in the appendix. Since the distribution functions with finite support are dense in the family of all distribution functions, we have the following result.

Theorem 2. If F is a distribution function with a completely monotone pdf, then there are mixture distributions $F^{(n)}$, $n \geq 1$, of the form

$$F^{(n)}(x,y) = \sum_{i=1}^{k_n} p_{ni} \left(1 - e^{-\lambda_{ni} x}\right) \left(1 - e^{-\mu_{ni} y}\right), \qquad x, y \geq 0,$$

with $\lambda_{ni} \leq \infty$, $p_{ni} > 0$ and $p_{n1} + \cdots + p_{nk_n} = 1$, such that $F^{(n)} \Rightarrow F$ as $n \to \infty$.

We now give two examples of long-tail distributions that are completely monotone. The bivariate Pareto distribution [17] and the bivariate Weibull distribution [10] with respective complementary distribution functions

$$\bar{F}(x,y) = \mathsf{Pr}\,(X > x, Y > y) = (1+x)^{-\lambda_1}(1+y)^{-\lambda_2}(1+\max(x,y))^{-\lambda_0}, x, y \geq 0,$$

and

$$\bar{F}(x,y) = \exp\{-\lambda_1 x^c - \lambda_2 y^c - \lambda_0 (\max\{x,y\})^c\}, \qquad x, y \geq 0,$$

$\lambda_i > 0$, $i = 0, 1, 2$ and $0 < c < 1$ are distributions with completely monotone probability density functions.

From above, the queueing model that we study in this paper is an important step in modeling realistic queueing systems where the interarrival and service times can be from a large class of bivariate distributions and where the variates are dependent.

The rest of the paper is organized as follows. In the next section we describe the model and derive some of its characteristics. Section 3 describes the queueing analysis using Wiener-Hopf factorization technique. We show some numerical examples and conclude with a discussion in Section 4.

2 Preliminaries

As mentioned earlier, we consider the joint distribution of A_n and X_n to be a mixture of bivariate densities of the form $g_i(a)h_i(x)$. In this paper we discuss in detail two models of bivariate mixtures to describe the input $\{A_n, X_n\}$ process – (i) $g_i(\cdot)$ and $h_i(\cdot)$ are both exponential and (ii) $g_i(\cdot)$ are exponential and $h_i(\cdot)$ are deterministic. We obtain the Laplace Stieltjes Transform (LST) of the waiting times for these input process models.

First consider the case when $g_i(a)$ and $h_i(x)$ are all exponential densities with rate λ_i and μ_i respectively. We will call this the MEE model. The dependency between the variables X_n and A_n is introduced via the parameters of their distributions. The LST of the distributions of X_n and A_n, $\mathcal{X}(s)$ and $\mathcal{A}(s)$, respectively, and the covariance between A_n and X_n, $\mathsf{cov}(X, A)$, is given below. Since the distributions are independent of n, we drop the subscripts and write the following:

$$\mathcal{X}(s) = \sum_{i=1}^{M} p_i \frac{\mu_i}{\mu_i + s}, \qquad \mathcal{A}(s) = \sum_{i=1}^{M} p_i \frac{\lambda_i}{\lambda_i + s}, \tag{2}$$

$$\mathsf{E}\,(X) = \sum_{i=1}^{M} \frac{p_i}{\mu_i}, \qquad \mathsf{E}\,(A) = \sum_{i=1}^{M} \frac{p_i}{\lambda_i}, \qquad \mathsf{E}\,(AX) = \sum_{i=1}^{M} \frac{p_i}{\lambda_i \mu_i}, \tag{3}$$

$$\mathsf{cov}(A, X) \;=\; \sum_{i=1}^{M} \frac{p_i}{\lambda_i \mu_i} - \left(\sum_{i=1}^{M} \frac{p_i}{\mu_i}\right)\left(\sum_{i=1}^{M} \frac{p_i}{\lambda_i}\right) \;=\; \sum_{i=1}^{M} \frac{p_i}{\mu_i}\left(\frac{1}{\lambda_i} - \sum_{j=1}^{M} \frac{p_j}{\lambda_j}\right). \tag{4}$$

The covariance will be positive if larger values of μ_i are associated with larger values of λ_i and negative if it is otherwise. The LST of the joint distribution X_n and A_n, $\mathcal{XA}(s_1, s_2)$ is given by

$$\mathcal{XA}(s_1, s_2) \;=\; \mathsf{E}\left(e^{-s_1 X - s_2 A}\right) \;=\; \sum_{i=1}^{M} p_i \frac{\lambda_i \mu_i}{(\mu_i + s_1)(\lambda_i + s_2)} \tag{5}$$

In the second model, that we call the MED model, the $h_i(\cdot)$ are deterministic. Another way to look at this model would be to say that X_n has a discrete distribution with $\Pr(X_n = 1/\mu_i) = p_i$, $i = 1, \ldots, M$ and given $X_n = 1/\mu_i$, the distribution of A_n is exponentially distributed with rate λ_i. One motivation for the choice of a discrete distribution for the service times is from packet communication networks where the probability mass function of the packet lengths is seldom smooth with a large fraction of the packet lengths taking only a small number of distinct values. Estimation of parameters for this model in the context of packet queues is quite straightforward. In this case we have $\mathcal{A}(s)$, $\mathsf{E}\,(X)$, $\mathsf{E}\,(A)$, $\mathsf{E}\,(XA)$ and $\mathsf{cov}(X, A)$ as before while

$$\mathcal{X}(s) = \sum_{i=1}^{M} p_i e^{-s/\mu_i}, \qquad \mathcal{XA}(s_1, s_2) \;=\; \sum_{i=1}^{M} p_i e^{-s_1/\mu_i} \frac{\lambda_i}{\lambda_i + s_2}. \tag{6}$$

3 Queueing Analysis

3.1 Background

Queueing analysis in the presence of cross correlation between A_n and X_n follows the spectral factorization technique that is used to analyze GI/GI/1 queues. (See, for example, Chapter 8 of [11]). Define $U_n = X_n - A_n$. If W_n is the waiting time of the n^{th} customer, the following recursive equation for W_{n+1} is well known:

$$W_{n+1} = \max(0, W_n + U_n). \tag{7}$$

Let $\mathsf{A}(a)$, $\mathsf{X}(x)$, $\mathsf{W}(w)$ and $\mathsf{U}(u)$ denote the distributions of A, X, W and U respectively. Also, $\mathcal{W}(s)$ will be Laplace Transform of $\mathsf{W}(w)$ and $\mathcal{U}(s)$, the LST of $\mathsf{U}(u)$. Note that $\mathcal{U}(s) = \mathcal{XA}(s, -s)$.

Since the bivariate sequence (X_n, A_n) is i.i.d., the sequence U_n is i.i.d. and the waiting time distribution $\mathsf{W}(w)$, satisfies the Lindley's integral,

$$\mathsf{W}(w) = \begin{cases} \int_{-\infty}^{w} \mathsf{W}(w-u)\, d\mathsf{U}(u) & \text{for } w \geq 0 \\ 0 & \text{for } w < 0 \end{cases} \tag{8}$$

which is an integral equation of the Wiener-Hopf type. If

$$\lim_{u \to -\infty} \frac{\mathsf{U}(u)}{e^{Du}} < \infty, \tag{9}$$

then following [11], to obtain $\mathcal{W}(s)$ we first factorize $\mathcal{U}(s) - 1$ as

$$\mathcal{U}(s) - 1 = \frac{\Psi_+(s)}{\Psi_-(s)}, \tag{10}$$

where $\Psi_+(s)$ satisfies the following conditions.

- For $\mathrm{Re}(s) > 0$, $\Psi_+(s)$ is an analytic function of s and has no zeroes in that half of the s-plane,
- For $\mathrm{Re}(s) > 0$, $\lim_{|s| \to \infty} \frac{\Psi_+(s)}{s} = 1$,

and $\Psi_-(s)$ satisfies

- For $\mathrm{Re}(s) < D$, $\Psi_-(s)$ is an analytic function of s and has no zeroes in that half of the s-plane,
- For $\mathrm{Re}(s) < D$, $\lim_{|s| \to \infty} \frac{\Psi_-(s)}{s} = -1$.

Then, $\mathcal{W}(s)$ is given by

$$\mathcal{W}(s) = \frac{1}{\Psi_+(s)} \lim_{s \to 0} \frac{\Psi_+(s)}{s}. \tag{11}$$

Note that $\mathcal{W}(s)$ is the Laplace Transform of the distribution of the waiting time not of its density. In the following, we show that MEE and the MED models are amenable to this solution method.

3.2 The MEE Model

From Section 2, writing $1 = \sum_{i=1}^{M} p_i$, we have, after some algebra,

$$\mathcal{U}(s) - 1 = \mathcal{X}\mathcal{A}(s, -s) - 1 = \sum_{i=1}^{M} p_i s \left(\frac{s + (\mu_i - \lambda_i)}{(\mu_i + s)(\lambda_i - s)} \right). \tag{12}$$

To illustrate the method, we first take $M = 2$. We then extend the procedure to the case of general M. In fact, what we describe below is an algorithm for determining the functions $\Psi_+(s)$ and $\Psi_-(s)$. As shown below, these depend on writing $\mathcal{U}(s) - 1$ above as a ratio of two polynomials and determining the zeroes of these two polynomials. We write $\mathcal{U}(s) - 1 = \alpha(s)/\beta(s)$, where

$$\begin{aligned}\alpha(s) &= s(p_1(s + (\mu_1 - \lambda_1))(\lambda_2 - s)(\mu_2 + s) + p_2(s + (\mu_2 - \lambda_2))(\lambda_1 - s)(\mu_1 + s)) \\ \beta(s) &= (\lambda_1 - s)(\mu_1 + s)(\lambda_2 - s)(\mu_2 + s).\end{aligned}$$

The queue stability condition will require that the mean interarrival time is strictly greater than the mean service time, i.e.,

$$\left(\frac{p_1}{\lambda_1} + \frac{p_2}{\lambda_2} \right) - \left(\frac{p_1}{\mu_1} + \frac{p_2}{\mu_2} \right) > 0, \tag{13}$$

Let $\gamma(s) = \alpha(s)/s$. We see that

$$\gamma(0) = p_1(\mu_1 - \lambda_1)\lambda_2\mu_2 + p_2(\mu_2 - \lambda_2)\lambda_1\mu_1, \tag{14}$$

and from the stability condition above, $\gamma(0) > 0$. This condition also turns out to be necessary for the existence of spectral decomposition.

The function $\alpha(s)$ has 4 zeroes which are located as follows. One is at $s = 0$ and the other three, denoted by s_0, s_1 and s_2, are located depending on the relation between λ_i and μ_i. Without loss of generality, we will assume $\mu_1 < \mu_2$. For illustration we will consider the condition $\lambda_1 < \lambda_2$. In this case, we will show that

$$-\mu_2 < s_0 < -\mu_1 < s_1 < 0 < \lambda_1 < s_2 < \lambda_2. \tag{15}$$

If $\lambda_1 > \lambda_2$, then we will have $0 < \lambda_2 < s_2 < \lambda_1$. Moreover, $\alpha(s)$ is positive for $s = -\mu_1$ and negative for $s = -\mu_2$, which implies that there is a zero between these two points. Also, $\alpha(s) > 0$ for $s = \lambda_1$ and $\alpha(s) < 0$ for $s = \lambda_2$. This means that $\alpha(s)$ has one zero between λ_1 and λ_2. Since $\gamma(0) > 0$, $\alpha(s)$ has a zero at a point $s_1 \in (-\mu_1, 0)$. Note that the coefficient of s^3 in $\gamma(s)$ is -1. If we factor this out we can write the spectral factorization as

$$\begin{aligned}\mathcal{U}(s) - 1 &= \frac{-s(s - s_0)(s - s_1)(s - s_2)}{(\lambda_1 - s)(\mu_1 + s)(\lambda_2 - s)(\mu_2 + s)} \\ &= \frac{\Psi_+(s)}{\Psi_-(s)} = \frac{\frac{s(s-s_0)(s-s_1)}{(\mu_1+s)(\mu_2+s)}}{-\frac{(\lambda_1-s)(\lambda_2-s)}{(s-s_2)}}.\end{aligned} \tag{16}$$

Since $\Psi_+(s)$ cannot have any zeroes for $s > 0$, the factor $(s - s_2)$ in (16) gets associated with $\Psi_-(s)$. Similarly, $\Psi_+(s)$ must be analytic for $Re(s) > 0$, and so the factor $(\lambda_1 - s)(\lambda_2 - s)$ is associated with $\Psi_-(s)$. Since $0 < \lambda_1 < s_2$, $\Psi_-(s)$ is analytic for $Re(s) < \lambda_1$ and does not have any zeroes in this half plane. Now it is easy to check that the limiting conditions imposed on $\Psi_+(s)$ and $\Psi_-(s)$ in Section 3.1 are satisfied. Further, it is easy to verify that these conditions cannot be satisfied for any choice of functions $\Psi_+(s)$ and $\Psi_-(s)$ when $\gamma(0) < 0$.

Using (11),

$$\mathcal{W}(s) = \frac{s_0 s_1}{\mu_1 \mu_2} \frac{(\mu_1 + s)(\mu_2 + s)}{s(s - s_0)(s - s_1)} \tag{17}$$

which is the Laplace transform of the waiting time distribution. So the LST of the pdf of the waiting time is

$$w(s) = s\mathcal{W}(s) = \frac{s_0 s_1}{\mu_1 \mu_2} \frac{(\mu_1 + s)(\mu_2 + s)}{(s - s_0)(s - s_1)}. \tag{18}$$

From the above it is easily seen that the mean waiting time satisfies,

$$\bar{w} \;=\; -\,w'(0) \;=\; -\left(\frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{s_0} + \frac{1}{s_1}\right) \;>\; 0, \tag{19}$$

where the last inequality follows from (15).

To compare the waiting times in the presence and absence of dependency, we now derive an expression for the mean waiting time in a queue when A_n and X_n have the same marginals as in the MEE model, but are independent. Under this assumption the LST of (X_n, A_n) will be

$$\mathcal{X\!A}(s_1, s_2) \;=\; \mathsf{E}\left(e^{-s_1 X - s_2 A}\right) \;=\; \sum_{i=1}^{M} p_i \frac{\lambda_i}{\lambda_i + s_2} \sum_{i=1}^{M} p_i \frac{\mu_i}{\mu_i + s_1}. \tag{20}$$

We carry out the queueing analysis as above for $M = 2$.

$$\mathcal{U}(s) - 1 = \mathcal{X\!A}(s, -s) - (p_1 + p_2)(p_1 + p_2) = s \sum_{i,j=1}^{2} \left(p_i p_j \frac{s + (\mu_i - \lambda_j)}{(\mu_i + s)(\lambda_j - s)}\right) \tag{21}$$

Writing as before

$$\mathcal{U}(s) - 1 \;=\; \frac{\alpha(s)}{\beta(s)} \;=\; \frac{s\gamma(s)}{\beta(s)}, \tag{22}$$

where $\beta(s)$ is the same as in the MEE Model above while

$$\gamma(s) = \sum_{i,j=1}^{2} \left(p_i p_j (s + (\mu_i - \lambda_j))(\mu_{i'} + s)(\lambda_{j'} - s)\right), \tag{23}$$

where $i' = 1$ if $i = 2$ and vice-versa. A little algebra shows that $\gamma(0)$ is the same as in the MEE Model and again, spectral decomposition is possible when $\gamma(0) > 0$. This together with the fact that the coefficient of s^3 in $\gamma(s)$ is negative, the function $\gamma(s)$ has two negative and one positive zero as in the MEE Model. Let the zeroes of (23) be $q_0 < q_1 < 0 < q_2$. The spectral factorization, the Laplace transform of the waiting time distribution and the mean waiting time will satisfy the same formulas as in the MEE Model with s_i replaced by q_i, $i = 0, 1, 2$. Thus the difference between the mean waiting times of the MEE Model and the model with independence will be $-\sum_{i=0}^{1}(s_i^{-1} - q_i^{-1})$.

The procedure used to analyze the MEE Model with $M = 2$ easily extends to the general $M > 1$ case. Let $\mu_{(1)} < \mu_{(2)} < \cdots < \mu_{(M)}$ be the ordered values of the μ_i's and let $\lambda_{(1)} < \lambda_{(2)} < \cdots < \lambda_{(M)}$ denote the ordered λ_i values. Then, $\mathcal{U}(s) - 1 = \alpha(s)/\beta(s)$, where

$$\begin{aligned}\alpha(s) &= s\sum_{i=1}^{M}(p_i(s + (\mu_i - \lambda_i))\prod_{j\neq i}(\lambda_j - s)(\mu_j + s)) \\ \beta(s) &= \prod_{i=1}^{M}(\lambda_i - s)(\mu_i + s).\end{aligned}$$

The zeroes s_i, $i = 0, \ldots, (2M-2)$, of $\alpha(s)$ satisfy at $-\mu_{(M)} < s_0 < -\mu_{(M-1)} < \cdots < -\mu_{(1)} < s_{M-1} < 0 < \lambda_{(1)} < s_M < \lambda_{(2)} < \cdots < \lambda_{(M-1)} < s_{2M-2} < \lambda_{(M)}$. The spectral factorization in this case will be

$$\mathcal{U}(s) - 1 \;=\; \frac{\Psi_+(s)}{\Psi_-(s)} \;=\; \frac{s\frac{\prod_{i=0}^{M-1}(s-s_i)}{\prod_{i=1}^{M}(\mu_i+s)}}{(-1)^{M-1}\frac{\prod_{i=1}^{M}(\lambda_i-s)}{\prod_{i=M}^{2M-2}(s-s_i)}}$$

The Laplace transform of the pdf of the waiting time and the mean waiting time are given by

$$w(s) = \prod_{i=1}^{M}\frac{(-s_{i-1})(\mu_i + s)}{\mu_i(s - s_{i-1})}.$$

Taking logarithms on both sides and differentiating, we get

$$\bar{w} = -\sum_{i=1}^{M}\left(\frac{1}{\mu_i} + \frac{1}{s_{i-1}}\right).$$

Note that the waiting time distribution depends on the λ_i, p_i, $i = 1, \ldots, M$ through the s_i, $i = 0, \ldots, M-1$.

3.3 The MED Model

From (6) and using $\sum_{i=1}^{M} p_i = 1$, we obtain

$$\mathcal{U}(s) - 1 = \sum_{i=1}^{M} p_i\left(\frac{\lambda_i(e^{-s/\mu_i} - 1) + s}{\lambda_i - s}\right). \tag{24}$$

Again, for ease in illustration we first take $M = 2$. Then the above equation reduces to

$$\mathcal{U}(s) - 1 = \frac{p_1(\lambda_1(e^{-s/\mu_1} - 1) + s)(\lambda_2 - s) + p_2(\lambda_2(e^{-s/\mu_2} - 1) + s)(\lambda_1 - s)}{(\lambda_1 - s)(\lambda_2 - s)} \tag{25}$$

Let $f(s) = p_1(\lambda_1(e^{-s/\mu_1} - 1) + s)(\lambda_2 - s) + p_2(\lambda_2(e^{-s/\mu_2} - 1) + s)(\lambda_1 - s)$ be the numerator in the above expression. If $\lambda_1 > \lambda_2$ (resp. $\lambda_1 < \lambda_2$) then $f(\lambda_1) > 0$ and $f(\lambda_2) < 0$ (resp. $f(\lambda_1) < 0$ and $f(\lambda_2) > 0$) implying that the function $f(s)$ has a zero between λ_1 and λ_2. Denote this zero by s_0. We write

$$\mathcal{U}(s) - 1 = \frac{\Psi_+(s)}{\Psi_-(s)}, \tag{26}$$

where,

$$\Psi_+(s) = -f(s)/(s - s_0), \qquad \Psi_-(s) = -(\lambda_1 - s)(\lambda_2 - s)/(s - s_0) \tag{27}$$

Further, the point s_0 is not a zero of the derivative of $f(s)$, that is, $f'(s_0) \neq 0$. This ensures that all conditions of Section 3.1 are satisfied and the LST of the waiting time distribution will be

$$\mathcal{W}(s) = \frac{K}{\Psi_+(s)} \tag{28}$$

where

$$K = \lim_{s \to 0} \Psi_+(s)/s = \frac{p_1\lambda_2(1 - \lambda_1/\mu_1) + p_2\lambda_1(1 - \lambda_2/\mu_2)}{s_0} \tag{29}$$

The LST of the density of the waiting time is

$$w(s) = -\frac{Ks(s - s_0)}{f(s)} = -\frac{K(s - s_0)}{g(s)}, \tag{30}$$

where $g(s) = f(s)/s$. $w'(s)$, the derivative of $w(s)$, is obtained as

$$w'(s) = -K\left(\frac{g(s) - g'(s)(s - s_0)}{g^2(s)}\right) \tag{31}$$

The mean of the waiting time $\bar{w} = -w'(0)$ will simplify to

$$\bar{w} \;=\; K\left(\frac{Ks_0 + g'(0)s_0}{K^2 s_0^2}\right) \;=\; \frac{K + g'(0)}{Ks_0} \tag{32}$$

Here K is obtained from (29), s_0 is the solution of $f(\cdot)$ as defined before and $g'(0)$ can be shown to be given by

$$g'(0) = p_1\lambda_1\lambda_2/(2\mu_1^2) - p_1(1 - \lambda_1/\mu_1) + p_2\lambda_1\lambda_2/(2\mu_2^2) - p_2(1 - \lambda_2/\mu_2) \tag{33}$$

We now consider the case of a general $M \geq 2$. Following the analysis for the case $M = 2$ we write

$$\mathcal{U}(s) - 1 = \frac{f(s)}{\prod_{i=1}^{M}(\lambda_i - s)},$$

where

$$f(s) = \sum_{i=1}^{M} p_i(\lambda_i(e^{-s/\mu_i} - 1) + s)\prod_{j \neq i}(\lambda_j - s)$$

As before $f(0) = 0$, and if $\lambda_{(i)}$, $i = 1, \ldots, M$ are the ordered values of the λ_i's, then the function f changes sign at each successive $\lambda_{(i)}$ starting from $f(\lambda_{(1)}) > 0$. The function f has $M - 1$ zeroes for $s > 0$, $s_1, \ldots, s_{M-1}$ satisfying $\lambda_{(i)} < s_i < \lambda_{(i+1)}$, $i = 1, \ldots, M - 1$. The functions Ψ_+ and Ψ_- in the spectral factorization are given by

$$\Psi_+(s) = (-1)^{M-1}\frac{f(s)}{\prod_{i=1}^{M-1}(s - s_i)}$$

and

$$\Psi_-(s) = (-1)^{M-1}\frac{\prod_{i=1}^{M}(\lambda_i - s)}{\prod_{i=1}^{M-1}(s - s_i)}.$$

Let $K = \lim_{s\to 0} \Psi_+(s)/s$. Then

$$K = \frac{\sum_{i=1}^{M} p_i(1-\frac{\lambda_i}{\mu_i})\prod_{j\neq i}\lambda_j}{\prod_{i=1}^{M-1} s_i}.$$

The Laplace transform of the pdf of the waiting time is then given by

$$w(s) = \frac{K}{\Psi_+(s)} = (-1)^{M-1}K\frac{\prod_{i=1}^{M-1}(s-s_i)}{g(s)},$$

where $g(s) = f(s)/s$. Observing that $g(0) = K\prod_{i=1}^{M-1} s_i$, it is easy to see that the mean waiting time satisfies

$$\bar{w} = -w'(0) = \frac{K\sum_{i=1}^{M-1}\prod_{j\neq i} s_j + g'(0)}{K\prod_{i=1}^{M-1} s_i}$$

with

$$g'(0) = \sum_{i=1}^{M} p_i\left(\frac{\prod_{j=1}^{M}\lambda_j}{2\mu_i^2} - \left(1-\frac{\lambda_i}{\mu_i}\right)\left(\sum_{j\neq i}\prod_{k\neq i,j}\lambda_k\right)\right).$$

4 Numerical Results and Discussion

We will consider a few numerical examples for the case of $M = 2$ to consider the effect of the dependency between the interarrival and corresponding service times. For the purpose of illustration we first show the variation of the mean waiting time as a function of the cross correlation coefficient, r, which can be obtained from (4).

$$r = \frac{\sum_{i=1}^{2}\frac{p_i}{\mu_i\lambda_i} - \sum_{i=1}^{2}\frac{p_i}{\mu_i}\sum_{i=1}^{2}\frac{p_i}{\lambda_i}}{\sqrt{(\sum_{i=1}^{2}\frac{p_i}{\mu_i^2} - \frac{2p_1p_2}{\mu_1\mu_2})(\sum_{i=1}^{2}\frac{p_i}{\lambda_i^2} - \frac{2p_1p_2}{\lambda_1\lambda_2})}}$$

Note that the same value of r can be obtained from multiple choice of the model parameters values. The mean waiting time being a function of the negative zeroes of the polynomial $\alpha(s)$ does not admit a representation in terms of the correlation r. The dependency between the interarrival and service times being non-linear, cannot entirely be captured by the correlation coefficient. Hence the mean waiting time will depend not only on r but also on the individual parameter values. This is illustrated next by fixing four different values r, for each of which we obtain four different values of the model parameters and plot the variation of the mean waiting time as a function of the utilization.

Without loss of generality all rate parameters will be normalized with respect to μ_1 and hence we will take $\mu_1 = 1$. Also, $p_2 = 1 - p_1$. Thus even for this simple model, we see that there are four parameters—μ_2, λ_1, λ_2 and p_1 and an exhaustive analysis of the system behavior for all parametric variations is probably an overkill. Further, if one were to study the effect of cross correlation on the waiting times, we would have to fix the moments of the interarrival and service times and solve for the parameters. We obtain a set of nonlinear simultaneous equations that are difficult to solve. Hence we do the following. We first fix the server utilization, ρ

($\rho = \frac{p_1/\mu_1+p_2/\mu_2}{p_1/\lambda_1+p_2/\lambda_2}$) and μ_2. We vary λ_1 from a $\lambda_{\max}$ towards 0 while keeping $\lambda_1 + \lambda_2$ a constant. For every combination of λ_1, λ_2, μ_1 and μ_2 we solve for p_1 to obtain the desired utilization. The idea is to obtain a range of cross correlations while keeping the utilization constant. Some combinations of λ_i and μ_i will not be able to support the utilization and hence such combinations will be discarded.

Fig. 1 shows the correlation coefficient and the corresponding mean waiting time for each combination of (λ_1, λ_2). Note that while for each combination of (λ_1, λ_2) we get a distinct value of r, it is not monotonic in λ_2. For the plots in the figure, the utilization is 0.85, $\lambda_{\max} = 10$, $\mu_1 = 1$ and $\mu_2 = 10$. Fig. 1 also shows the mean waiting times for the $H_2/H_2/1$ queue with identical parameters to that of the MEE queue except that here the A_n and X_n are independent. Observe that the results from the $H_2/H_2/1$ closely match that of the MEE queue with a maximum error of -15% for high negative cross correlation and about +10% for high positive cross correlation. Our extensive numerical study indicates that the results are similar for different values of the utilization and also for the MED model.

Next, we present the numerical results for the waiting time analysis as a function of the utilization while keeping the correlation coefficient constant. Choosing $\lambda_2 = b\lambda_1$ and $\mu_2 = c\mu_1$, where $b, c > 0$ are constants, we see that r depends only on p_1 when b and c are fixed. In fact, it can be seen that fixing b and c, to obtain a given r there will be two solutions for p_1.

For the MEE model, a cross correlation coefficient in the interval $[-0.25, 0.45]$ can be obtained using values of the parameters $b, c \in (0.1, 10.0)$ and $0.2 < p_1 < 0.8$. Values of r outside the above range can also be supported but we believe that the values of b, c and p_1 may not be representative a realistic system. In that case higher values of M may be used. For the MED model, we find that the r coefficient in the range $[-0.5, 0.5]$ can be obtained using parameters in the above ranges. In the following we present the queue performance measures for two extreme values of r. Note that the same r can be obtained for different combinations of b, c and p_1. Table 1 lists the b, c and the p_1 values for the r for which we obtain the numerical results. Numerical results are obtained by finding λ_1 for the specified utilization.

Fig. 2 shows the mean waiting time in the MEE queue as a function of the utilization for different sets of parameters for the two cross correlation values that we consider. Also shown are the mean waiting times in the $H_2/H_2/1$ queue with the same parameters. Observe that in the case of $r = -0.2$, except for a 15–20% higher mean waiting time in the MEE queue, the 'behavior' is pretty much the same in both queues for the same parameters. With positive correlations, the advantage is in favor of the MEE queue although the percentage difference is significantly higher at 50-60%. But that is because of the magnitude of the correlation. With a positive correlation of $r = 0.2$, the difference was similar to that in the case of the negative correlation. Also, observe that for similar values of b and c and cross correlations, the MED has a lower mean waiting time than the MEE model. This can be explained by the lower variability

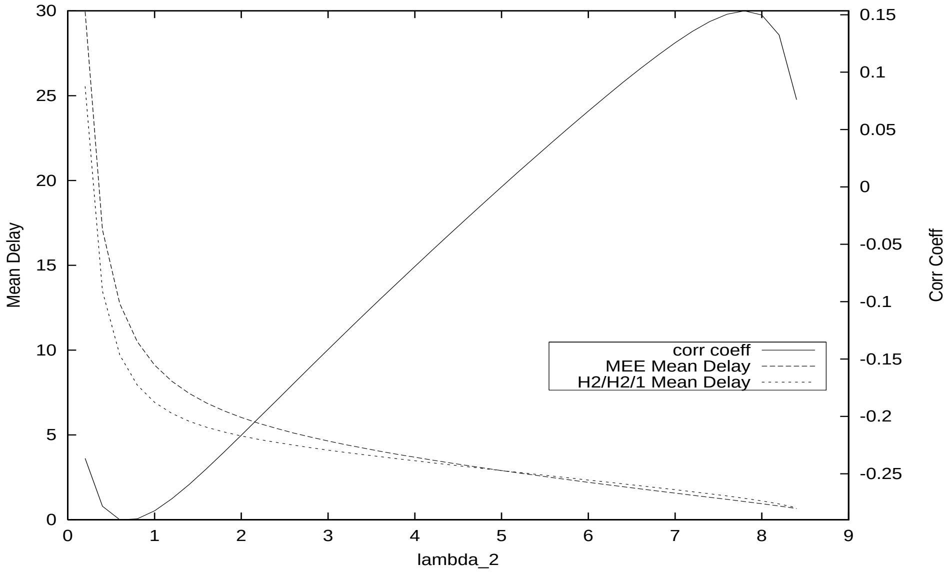
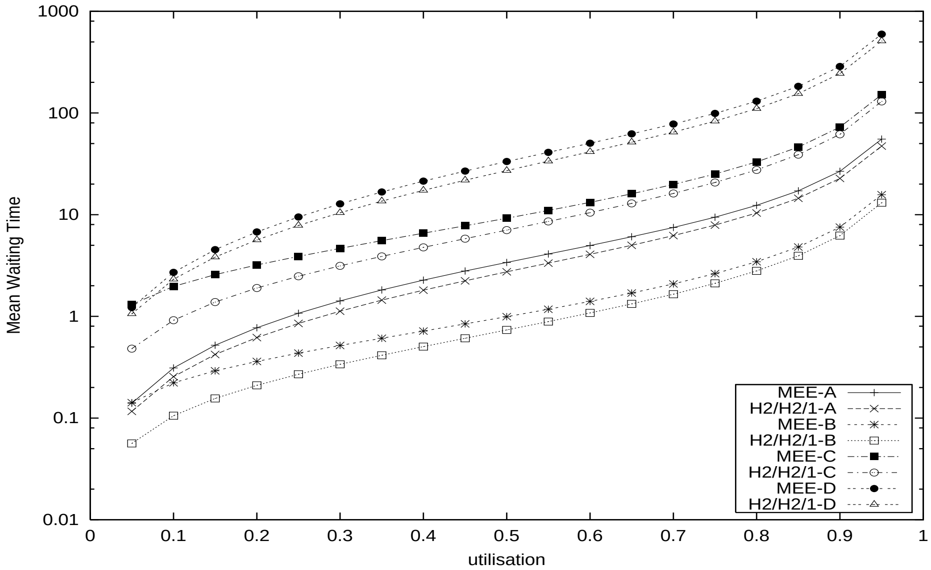


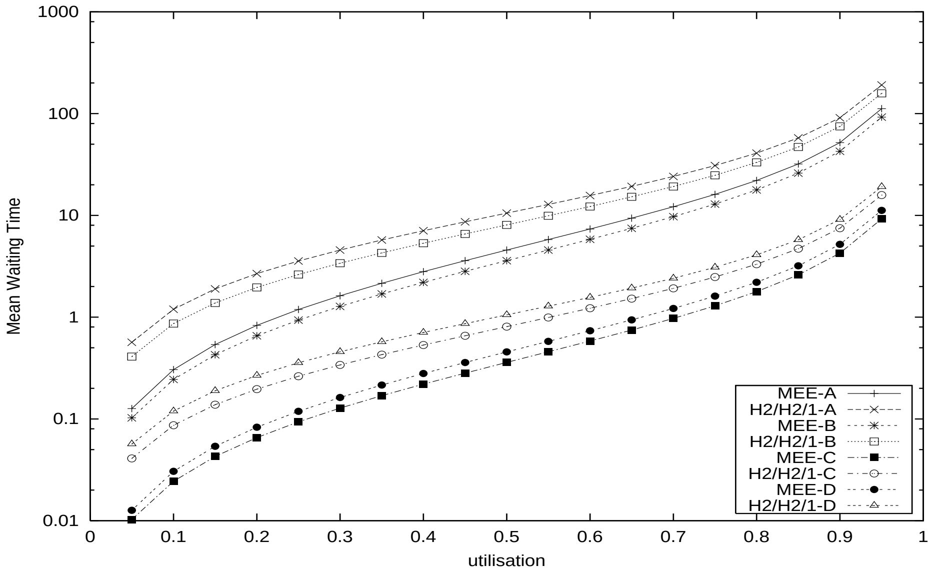
Figure 1: Mean waiting time and r the MEE model. Also shown is the mean waiting time for the $H_2/H_2/1$ queue with identical parameters as that of the MEE queue except that A_n and X_n are independent.

Negative Cross Correlation								
	MEE				MED			
Name	b	c	p_1	r	b	c	p_1	r
A	0.10	0.50	0.819	-0.2	0.10	10.0	0.414	-0.5
B	0.10	0.50	0.185	-0.2	0.10	10.0	0.993	-0.5
C	10.0	0.10	0.862	-0.2	10.0	0.10	0.007	-0.5
D	10.0	0.10	0.139	-0.2	10.0	0.10	0.586	-0.5
Positive Cross Correlation								
	MEE				MED			
Name	b	c	p_1	r	b	c	p_1	r
A	0.10	0.10	0.889	+0.4	0.10	0.10	0.414	+0.5
B	0.10	0.10	0.926	+0.4	0.10	0.10	0.993	+0.5
C	10.0	10.0	0.074	+0.4	10.0	10.0	0.586	+0.5
D	10.0	10.9	0.111	+0.4	10.0	10.0	0.007	+0.5

Table 1: b, c and p_1 for different r for which we obtain the numerical results.



Negative Cross Correlations



Positive Cross Correlations

Figure 2: Mean waiting time as a function of the utilization in the MEE queue. Also shown are the waiting times for the $H_2/H_2/1$ queue with identical parameters as that of the MEE queue except that A_n and X_n are independent.

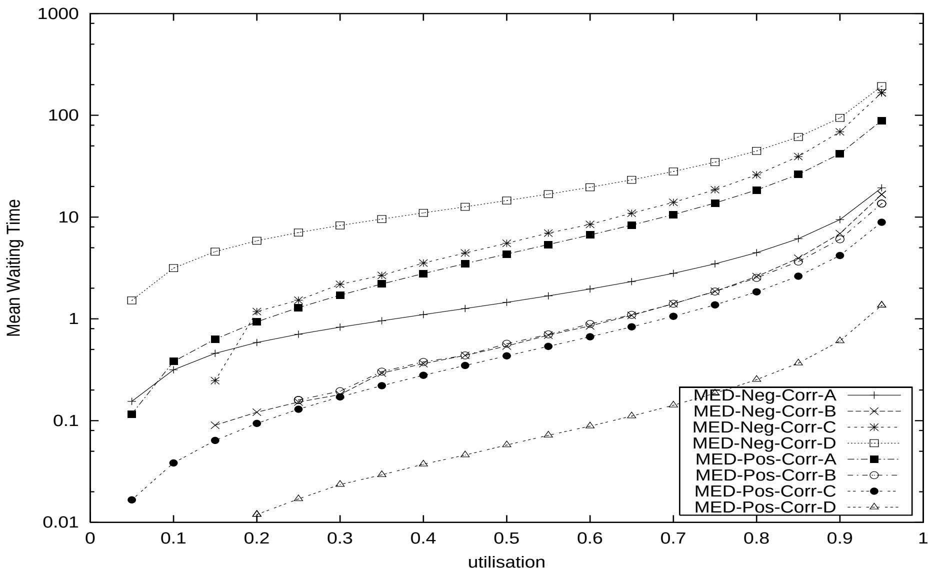


Figure 3: Mean waiting time as a function of the utilization in the MED queue. Figure shows the results for both negative and positive cross correlations.

of the components of the mixture in the MED model.

From extensive numerical analysis that we performed with the model that we have introduced in this paper, our general conclusion is that in the presence of cross correlation between the service and interarrival times, the marginal distributions provide an excellent first approximation to the mean waiting time. The effect of the correlation is systematic and is well approximated by a multiplicative constant that is not significantly affected by the load and the parameters. This was also seen in the numerical results of Fig. 1.

In modeling a real system using the models from this paper, the values of M, λ_i, μ_i and p_i are estimated from the data. Alternatively, each component of the mixture corresponds to a component of the system that we are modeling. This determines value of cross correlation coefficient. Thus the models described in this paper should be seen as a mechanism to model the dependency between the A_n and X_n and not just as means to study the effect of cross correlation on the mean waiting times.

Appendix

We now give a proof of Theorem 1. Let $f : \mathcal{R}^2 \to [0, \infty)$ be a completely monotone function. Fix $a > 0$ and let $0 < s_i < 1$, $i = 1, 2$. Let $f^{n,k}(x, y) = \partial^n f(x, y)/\partial x^k \partial y^{n-k}$. Consider the Taylor expansion of the function $f(a(1-s_1), a(1-s_2))$ as a function of (s_1, s_2) around (a, a):

$$f(a(1-s_1), a(1-s_2)) = \sum_{n=0}^{\infty} \frac{(-a)^n}{n!} \sum_{k=0}^{n} \binom{n}{k} f^{n,k}(a, a) s_1^k s_2^{n-k}$$

Define $f_a(x,y) = f(a(1-e^{-x/a}), a(1-e^{y/a}))$. Using the above expansion we get

$$f_a(x,y) = \sum_{n=0}^{\infty} \frac{(-a)^n}{n!} \sum_{k=0}^{n} \binom{n}{k} f^{n,k}(a,a) e^{-xk/a} e^{-y(n-k)/a}$$

The right hand side of the above equation is the Laplace transform of a measure G_a that puts mass $(-a)^n f^{n,k}(a,a)/n!(n-k)!$ at the point $(k/a, (n-k)/a)$, $k = 0, 1, \ldots, n$ $n = 1, 2, \ldots$. Now $f_a \to f$ as $a \to \infty$. By the extended continuity theorem there exists therefore a measure G such that $G_a \to G$ and f is its Laplace transform.

References

[1] Chao, X. Monotone effect of dependency between interarrival and service times in a simple queueing system. Operations Research Letters **1995**, 17, 47–51.

[2] Cidon, I.; Guerin, R.; Khamisy, A.; Sidi, M. Analysis of a correlated queue in communication systems. IEEE Trans on Inform Theory**1993**, IT-39 (2), 456–465.

[3] Cidon, I.; Guerin, R.; Khamisy, A. and Sidi, M. On queues with inter-arrival times proportional to service times. Probability in the Engineering and Informational Sciences **1996**, 10, 87–107.

[4] Conolly, B. W. The waiting time process for a certain correlated queue. Operations Research **1968**, 15, 1006–1015.

[5] Conolly, B. W.; Hadidi, N. A correlated queue. Journal of Applied Probability **1969**, 6, 122–136.

[6] Conolly, B. W.; Choo, Q. H. The waiting time process for a generalized correlated queue with exponential demand and service. SIAM Jour. Appl. Math. **1979**, 37, 263–275.

[7] Feldmann, A., and Whitt, W. Fitting Mixtures of Exponentials to Long-Tail Distributions to Analyze Network Performance Models. Performance Evaluation**1998**, 31(8), 963–976.

[8] Feller, W. An introduction to probability theory and its applications **1971**, *2*, Wiley, New York.

[9] Fendick, K .W.; Saksena, V. R.; Whitt, W. Dependence in packet queues. IEEE Trans. Commun. **1989**, 37, 1173–1183.

[10] Hanagal, D.D. A multivariate Weibull distribution. Economic Quality Control **1996**, 11, 193–200.

[11] Kleinrock, L. Queueing Systems **1975**, *1* John Wiley and Sons.

[12] Leland, W. E.; Taqqu,M. S.; Willinger, W.; Wilson, D. V. On the self similar nature of Ethernet traffic **1994**, IEEE/ACM Transactions on Networking, 2, 1-15.

[13] Müller, A. On the waiting time in queues with dependency between interarrival and service times. Operations Research Letters **2000**, 26, 43–47.

[14] Paxson, V.; Floyd,S. Wide area traffic: The failure of Poisson modeling. IEEE/ACM Transactions on Networking **1995**, 3, 226–244.

[15] Ramaswami, V. Containing the congestion effects of internet dial up calls. **Invited Talk,** Intnl. Conf. on Stoch. Models, Kochi India, Dec. 17–21 2002.

[16] Titterington, D. M.; Smith, A. F. M.; Makov, U. E. Statistical Analysis of Finite Mixture Distributions **1986**, John Wiley and Sons.

[17] Veenus, P.; Nair, K. R. M. Characterization of a bivariate Pareto distribution. Journal of Indian Statistical Association **1994**, 32, 15–20.