Sample Path Analysis of Busy Periods and Related First Passages of a Correlated *MEP/MEP/1* System

Chaitanya Garikiparthi, Appie van de Liefvoort, Kenneth Mitchell
University of Missouri - Kansas City
School of Computing and Engineering
5100 Rockhill Rd
Kansas City, MO 64110
{cng8t7, appie, mitchellke}@umkc.edu

Abstract

In this paper we study the busy period of an MEP/MEP/1 system, where both the arrival and the service processes can be serially correlated Matrix Exponential Processes. A dynamic programming algorithm is given to compute the probabilities for serving n customers in a busy period and expressions for the first two moments are derived. We study both the effect of correlation in the arrival and service processes and the squared coefficient of variation on these probabilities. The solutions give us qualitative insights into the nature of the busy period of the MEP/MEP/1 system. The resulting algorithms are easily programmable and efficient.

1. Introduction

A study of transient queue length fluctuations during a busy period provides quantitative measures that enable proactive resource management for optimal system performance and capacity utilization. For example, in server consolidation models, an individual server that forms a part of this consolidation is perceived to be highly utilized if its queue grows beyond a given threshold. Dynamic resource allocation remains a challenge and one question that usually arises is whether or not to allocate additional resources (such as processing power, additional buffers, etc) to a given server. A system parameter that can facilitate in making informed decisions in this regard is the number of customers that will be served in this state of high utilization. This problem can be posed as a modification of the classical busy period problem and requires a solution based on a transient system analysis.

Most processes in telecommunications and computer networks exhibit a high degree of variance and are known to be serially correlated. Therefore, in order to develop accurate models to represent these systems, we need to allow for the arrival and the service processes that characterize the system to be both general and correlated.

The busy period for a system is the time interval between any two successive idle periods. It starts when a customer arrives to an empty system and ends when the departing customer leaves the system idle for the first time thereafter. In effect, a simple busy period is equivalent to a first passage from level 1 to level 0. Furthermore, the first passage from a higher level say 'l' to '$(l-1)$' is also of interest. Here, if we let $l-1$ denote a threshold, we are interested in the transient behavior around this threshold.

For an *M/M/1* system, the probabilities for n customers being served during a normal busy period are known, see for example Takács [20]. Takács also derives the joint density for the number served and the length of the busy period where either the inter-arrival times or the service times have an exponential distribution [21]. More recently, Ny and Sericola [17] study the busy period distribution of the *BMAP/PH/1* queue based on an approximation of the exponential of an infinite sized $\boldsymbol{Q}$ matrix using uniformization and truncation. Lucantoni et al. consider the transient BMAP/G/1 queue [13], [14] and derive the two dimensional transform for the joint distribution for the number served in a busy period and its length, which they numerically invert [7]. There is extensive literature studying the tail of the busy period, especially for the M/M/1 queue [2]. One of the observations in that paper is that the tail distributions of busy periods are sub-exponential, which are often hard to model. Boxma and Dumas [6] relate the tail behavior of the active periods of the input sources to the tail of the busy period distribution of a GI/G/1 queue. Asmussen and Bladt [5] use the sample path approach to study the mean busy periods for Markov modulated queues. The probabilities for n customers served during a busy period of a *GI/M/1/N* queue is studied by Agarwal [3] by splitting up the sample paths at suitable renewal epochs . Heindl and Telek [9] studied the busy period of a *MAP/PH/1* system. Lipsky extensively studied first passage times in renewal *ME/ME/1* queues and uses recurrence relations for their solutions [12].

Existing literature on busy periods usually requires either the arrival process or the service process (or both) to be renewal and most proceed by studying the embedded Markov chain at the resulting renewal instants. These techniques are not extendible to *MEP/MEP* systems as there are no such renewal points available. Furthermore, most existing work rely heavily on transform solutions and involve numerical inversions. In this paper we allow both the arrival and the service processes to be non-renewal. This allows us to study the effect of correlation in both the arrival and service processes on busy periods and related performance metrics. We use a combinatorial approach that is analytic and the solutions are obtained using closed form recursive expressions that are easily computable.

Define $D_{l,l-1}$ as the first passage process wherein the system transitions from level l to level $l-1$ ending when level $l-1$ is reached for the first time. In this paper we derive recursive solutions to find the probability for serving 'n' customers during this first passage in an *MEP/MEP/1* queueing system, and we derive moments for the number of customers served during this first passage. We then specialize the solutions obtained to the case of a busy period and study the effect of correlation in the arrival and service processes and the squared coefficient of variance on these probabilities.

The rest of the paper is organized as follows. In Section 2 we give a brief introduction to Linear Algebraic Queueing Theory (LAQT) and develop the $\boldsymbol{H}$ operators that we use extensively throughout the paper. In Section 3 we model $D_{l,l-1}$ of an *MEP/MEP/1* system and derive the recursive expressions for the probability of n customers served during this first passage and the moments for the number served during this first passage. In Section 4 we specialize the results derived in section 3 for a busy period in an *MEP/MEP/1* queue. In Section 5 we perform numerous numerical studies characterizing the effect of correlation in the arrival and service processes and the variability in those processes on various performance measures of interest related to busy periods. Section 6 concludes the paper.

2. Model Description

2.1. Matrix Exponential Process

We use Linear Algebraic Queueing Theory (LAQT) to study the path taken by a queueing system during a busy period. Here, we briefly review the needed material. A matrix exponential (ME) distribution [12] is defined as a probability distribution whose density can be written as

$$f(t) = \boldsymbol{p}(0)\exp(-\boldsymbol{B}t)\,\boldsymbol{B}\boldsymbol{e}', \quad t \geq 0, \qquad (1)$$

where $\boldsymbol{p}(0)$ is the starting operator for the process, $\boldsymbol{B}$ is the process rate operator, and $\boldsymbol{e}'$ is a summing operator, a vector usually consisting of all 1's. The n^{th} moment of the matrix exponential distribution is given by $E[X^n] = n!\boldsymbol{p}(0)\boldsymbol{V}^n\boldsymbol{e}'$, where $\boldsymbol{V}$ is the inverse of $\boldsymbol{B}$. The class of matrix exponential distributions is identical to the class of distributions that possess a rational Laplace-Stieltjes transform. As such, it is more general than continuous phase type distributions which have a similar appearance.

The joint density function for the first *k*-successive events is described by a Matrix Exponential Process (MEP).

$$f_k(x_1,\ldots,x_k) = \boldsymbol{p}(0)\exp(-\boldsymbol{B}x_1)\,\boldsymbol{L}\ldots\exp(-\boldsymbol{B}x_k)\,\boldsymbol{L}\boldsymbol{e}', \qquad (2)$$

where matrix $\boldsymbol{L}$ is the event generator matrix, $\boldsymbol{p}$ is the starting state for the process and $\boldsymbol{e}'$ is a summing operator, a vector usually consisting of all 1's. If the process is stationary, then the starting vector $\boldsymbol{p}$ satisfies $\boldsymbol{p}(0) = \boldsymbol{p}(0)\boldsymbol{V}\boldsymbol{L}$. Examples for such processes are a Poisson process ($\boldsymbol{B}$=[λ], $\boldsymbol{L}$=[λ]), a renewal process ($\boldsymbol{L} = \boldsymbol{B}\boldsymbol{e}'\boldsymbol{p}$), and a Markovian Arrival Process (MAP)($\boldsymbol{B} = -\boldsymbol{D}_0, \boldsymbol{L} = \boldsymbol{D}_1$). Note that $\boldsymbol{B}$ and $\boldsymbol{L}$ are not limited to being Markovian rate matrices. So every MAP is an MEP, but not vice versa (see also [10]). By implication, stationary *MEP's* are dense in the family of all stationary point processes as well, [4]. For additional details, see [12, 11, 22].

2.2. $\boldsymbol{H}$ Operators

Let the arrival and service processes be represented by $< \boldsymbol{B}_a, \boldsymbol{L}_a >$ and $< \boldsymbol{B}_s, \boldsymbol{L}_s >$ respectively. The conditional probability that an arrival event occurs before the service event given that the starting vector is $\boldsymbol{p}$(0) is given by

$$\Pr[\, A < S \;|\; \boldsymbol{p}(0)] = \boldsymbol{p}(0)(\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})^{-1}\widehat{\boldsymbol{L}_a}\boldsymbol{e}'.$$

where $\widehat{\boldsymbol{B}_a} = \boldsymbol{B}_a \otimes \boldsymbol{I}_s$, $\widehat{\boldsymbol{B}_s} = \boldsymbol{I}_a \otimes \boldsymbol{B}_s$, $\widehat{\boldsymbol{L}_a} = \boldsymbol{L}_a \otimes \boldsymbol{I}_s$ and $\widehat{\boldsymbol{L}_s} = \boldsymbol{I}_a \otimes \boldsymbol{L}_s$, and $\otimes$ is the Kronecker product operator which embeds the arrival and service processes into system space. Here, $(\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})^{-1}$ represents the average time that both the arrival and service processes are concurrently active, and $\widehat{\boldsymbol{L}_a}$ represents the arrival event occurring before a service completion. The trailing $\boldsymbol{e}'$ sums up the probabilities distributed in vector form and is usually a column vector of all one's of appropriate dimensions.

In an *MEP/MEP/1* system, the conditional probability that two successive events are both arrivals given the system starts in state $\boldsymbol{p}(0)$ is $\boldsymbol{p}(0)(\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})^{-1}\widehat{\boldsymbol{L}_a} \cdot (\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})^{-1}\widehat{\boldsymbol{L}_a}\boldsymbol{e}'$. Define operators $\boldsymbol{H}_a$ for arrival event happening before the service and $\boldsymbol{H}_s$ for service event happening before the arrival by unconditioning on the initial state of the system, as follows:

$$\boldsymbol{H}_a = (\widehat{\boldsymbol{B}_a}+\widehat{\boldsymbol{B}_s})^{-1}\widehat{\boldsymbol{L}_a} \quad \text{and} \quad \boldsymbol{H}_s = (\widehat{\boldsymbol{B}_a}+\widehat{\boldsymbol{B}_s})^{-1}\widehat{\boldsymbol{L}_s}.$$

Essentially these $\boldsymbol{H}$ operators allow us to track the path evolution by embedding at the event transitions in the continuous time Markov chain. At each observed transition point, the appropriate $\boldsymbol{H}$operator is applied (and normalized if needed) to update the internal state of the discrete time Markov chain, thus allowing both the arrival and service processes involved to be non-renewal. We summarize what $\boldsymbol{H}_a$ and $\boldsymbol{H}_s$ are for different systems in the table 1.

	$\boldsymbol{H}_a$	$\boldsymbol{H}_s$
M/M/1	$\frac{\lambda}{\lambda+\mu}$	$\frac{\mu}{\lambda+\mu}$
M/ME/1	$(\lambda \boldsymbol{I} + \boldsymbol{B}_s)^{-1}\lambda$	$(\lambda \boldsymbol{I} + \boldsymbol{B}_s)^{-1}\boldsymbol{B}_s\boldsymbol{e}_s'\boldsymbol{p}_s$
ME/M/1	$(\boldsymbol{B}_a + \mu\boldsymbol{I})^{-1}\boldsymbol{B}_a\boldsymbol{e}_a'\boldsymbol{p}_a$	$(\boldsymbol{B}_a + \mu\boldsymbol{I})^{-1}\mu$
MEP/MEP/1	$(\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})^{-1}\widehat{\boldsymbol{L}_a}$	$(\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})^{-1}\widehat{\boldsymbol{L}_s}$

Table 1. $\boldsymbol{H}$operators for different systems

Please note that the $\boldsymbol{H}$ operators introduced here differ from the similarly named operators in [12].

3. Conditional sample path analysis of first passages in an *MEP/MEP/1* system

Consider a system that just had a transition from level $(l-1)$ to level l and let $\boldsymbol{p}(0)$ be the current internal state of the system. The events that drive the Markov chain representing this system are either an arrival ($\boldsymbol{H}_a$) or a service completion ($\boldsymbol{H}_s$). As defined earlier, let $D_{l,l-1}$ represent the first passage process wherein the system transitions from level l to level $l-1$ ending when level $l-1$ is reached for the first time. Every sample path that belongs to the process $D_{l,l-1}$ can be represented by a succession of $\boldsymbol{H}_a$'s and $\boldsymbol{H}_s$'s. To compute the probability of occurrence for each of these sample paths we have to pre and post multiply the $\boldsymbol{H}$operator string with $\boldsymbol{p}(0)$ and $\boldsymbol{e}'$ respectively.

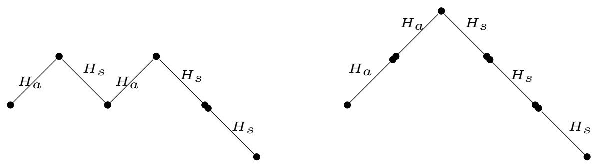

Figure 1. Paths serving exactly 3 customers during the first passage, $D_{l,l-1}$

The number of possibilities to serve exactly n customers during this first passage is given by C_{n-1}, the $(n-1)^{st}$ Catalan number [19]. The n^{th} Catalan number C_n is computed either as $\frac{1}{n+1}\binom{2n}{n}$, n $\geq$ 0, or from the recursive definition for Catalan numbers as $C_n = \sum_{i=0}^{n-1} C_i C_{n-i-1},\ C_0 = C_1 = 1$. For example, exactly three customers can be served during a first passage from level l to level $l-1$ by following one of the two paths shown in Fig. 1 and the probabilities associated with each of those paths are $\boldsymbol{p}(0)\boldsymbol{H}_a\boldsymbol{H}_s\boldsymbol{H}_a\boldsymbol{H}_s\boldsymbol{H}_s\boldsymbol{e}'$ and $\boldsymbol{p}(0)\boldsymbol{H}_a\boldsymbol{H}_a\boldsymbol{H}_s\boldsymbol{H}_s\boldsymbol{H}_s\boldsymbol{e}'$ respectively. In the *M/M/1* case these two paths would be equi-probable with a probability of $\frac{\lambda^2\mu^3}{(\lambda+\mu)^5}$ and hence the probability for exactly three customers being served during $D_{l,l-1}$ is given by $\frac{2\lambda^2\mu^3}{(\lambda+\mu)^5}$.

A busy period is a special case of this first passage when $l = 1$. Let $N_{l,l-1}$ be the discrete random variable for the number of customers served during the first passage $D_{l,l-1}$. Hence, in an *M/M/1* system,

$$d_{n,1} \triangleq \text{Prob}[N_{1,0} = \text{n}\ ] = C_{n-1}\frac{\lambda^{n-1}\mu^n}{(\lambda+\mu)^{2n-1}}, \quad n \geq 1.$$

In the case of a *MEP/MEP/1* system, the matrices involved are generally non-commutative ($\boldsymbol{H}_a\boldsymbol{H}_s \neq \boldsymbol{H}_s\boldsymbol{H}_a$) and the paths have different probabilities associated with them. The relationship among these different paths that serve a given number of customers during this first passage leads us to define a set of recurrence relations for these probability matrices, resulting in a direct generalization of the recursive definition for scalar Catalan numbers to matrices.

If $N_{l,l-1} = 1$, then the first arrival that started the process is followed by a departure; the probability of this occurring is $\boldsymbol{p}(0)\boldsymbol{H}_s\boldsymbol{e}'$. In all the other cases, at

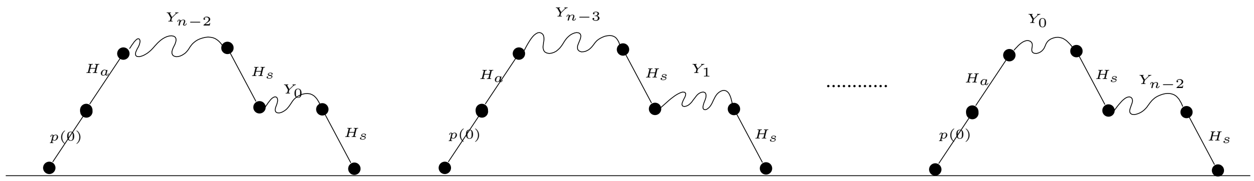

Figure 2. Paths serving exactly n customers during the first passage, $D_{l,l-1}$

least one more arrival ($\boldsymbol{H}_a$) occurs before the first departure ($\boldsymbol{H}_s$). We consider the remaining process (after the second arrival), as two sub processes, where a certain number of customers are serviced before returning to level l for the first time followed by a certain number of customers serviced before finally returning to level $l-1$ (See Fig. 2). In this respect, each of these sub-paths is similar to a Dyck path [18] starting from the starting point of the sub-path. Thus exactly n customers can be served during this first passage ($D_{l,l-1}$) by serving $n-i$ customers before returning to level l for the first time, followed by serving $i-1$ customers before the last customer departs the system, followed by the final departure event returning the system to level $l-1$ for the first time.

The above insight and explicit enumeration of all the possible paths for a few cases allows us to define the following set of recurrence relations. Please note that theses derivations are independent of the current state of the system (as long as the server is active). Let,

$$\begin{aligned}
\boldsymbol{Y}_0 &= \boldsymbol{I}, \\
\boldsymbol{Y}_1 &= \boldsymbol{H}_a \boldsymbol{Y}_0 \boldsymbol{H}_s \boldsymbol{Y}_0, \\
\boldsymbol{Y}_2 &= \boldsymbol{H}_a \boldsymbol{Y}_1 \boldsymbol{H}_s \boldsymbol{Y}_0 + \boldsymbol{H}_a \boldsymbol{Y}_0 \boldsymbol{H}_s \boldsymbol{Y}_1, \\
&\vdots \\
\boldsymbol{Y}_{n-1} &= \boldsymbol{H}_a \left[\boldsymbol{Y}_{n-2} \boldsymbol{H}_s \boldsymbol{Y}_0 + \boldsymbol{Y}_{n-3} \boldsymbol{H}_s \boldsymbol{Y}_1 + \ldots \right. \\
&\qquad \left. + \ldots + \boldsymbol{Y}_0 \boldsymbol{H}_s \boldsymbol{Y}_{n-2} \right], \\
\boldsymbol{Y}_n &= \boldsymbol{H}_a \left[\boldsymbol{Y}_{n-1} \boldsymbol{H}_s \boldsymbol{Y}_0 + \boldsymbol{Y}_{n-2} \boldsymbol{H}_s \boldsymbol{Y}_1 + \ldots + \right. \\
&\qquad \left. + \ldots + \boldsymbol{Y}_0 \boldsymbol{H}_s \boldsymbol{Y}_{n-1} \right].
\end{aligned}$$

where $\boldsymbol{I}$ is an identity matrix of the dimensions of either the service process or the arrival process whichever is an *MEP* and it would be in the product space if both of these are *MEPs*. $\boldsymbol{Y}_i$ is the operator that transfers the internal state of the system as the system transitions from level l back to level l while traversing only states $l, l+1, l+2, \ldots$ and after having served exactly i customers. Here, $\boldsymbol{Y}_i$ is independent of the level l, as all the information that differentiates transitions for different levels is present in the system starting vector on which $\boldsymbol{Y}_i$ operates, and $\boldsymbol{Y}_i$ depends only on the number of arrivals and departures. Furthermore, the operator $\boldsymbol{Y}_i \boldsymbol{H}_s$ represents serving exactly $(i+1)$ customers while transitioning down by one level. In short

$$\begin{aligned}
\boldsymbol{Y}_0 &= \boldsymbol{I}, \\
\boldsymbol{Y}_{n-1} &= \sum_{i=0}^{n-2} \boldsymbol{H}_a \boldsymbol{Y}_{n-i-2} \boldsymbol{H}_s \boldsymbol{Y}_i, \qquad n > 1.
\end{aligned}$$

Please note the similarity between the above definition for $\boldsymbol{Y}_{n-1}$ and the recursive definition for Catalan numbers. Indeed, if one would unravel the recurrence relation, there would be C_{n-1} terms in the expression for $\boldsymbol{Y}_{n-1}$. Also note that the definition for $\boldsymbol{Y}_{n-1}$ is order preserving and hence the correlation that is present in the arrival and service events are effectively captured therein.

Now the probability that exactly n customers are served during $D_{l,l-1}$ conditioned on the internal system state being in $p(0)$ at the transition from level $l-1$ to level l is given by,

$$d_{n,l} \triangleq \text{Prob}[N_{l,l-1} = \text{n}\,] = \boldsymbol{p}(0) \boldsymbol{Y}_{n-1} \boldsymbol{H}_s \boldsymbol{e}', \qquad n \geq 1.$$

where $\boldsymbol{e}'$ is a column vector of all 1's whose dimensions depend on whether the system is an *M/M/1*, *M/MEP/1*, *MEP/M/1* or an *MEP/MEP/1*. For the *M/MEP/1* and *MEP/M/1*, its dimension corresponds to either the service processes or the arrival processes dimension respectively, and for an *MEP/MEP/1* system $\boldsymbol{e}'$ is in the product space given by $\boldsymbol{e}' = \widehat{\boldsymbol{e}'_a} \boldsymbol{e}'_s$, where $\widehat{\boldsymbol{e}'_a} = \boldsymbol{e}'_a \otimes \boldsymbol{I}_s$. For the scalar *M/M/1* case it is trivial.

It is to be noted here that this derivation for the probability of n customers being served in this first passage is a general result and does not require the queueing process involved to be recurrent, i.e., this derivation holds even for an unstable system (utilization $\rho > 1$). In particular, this setup will allow the computation of the probability that the busy period is finite.

3.1. Moments for the number of customers served during $D_{l,l-1}$

The z-transform for the number of customers served during this first passage $D_{l,l-1}$ is

$$y(z) = \sum_{n=1}^{\infty} Prob[N_{l,l-1} = n].z^n = b_1 z + b_2 z^2 + \ldots$$

Since $\boldsymbol{Y}_{n-1}$ forms the core of $d_{n,l}$, one can now define the matrix z-transform $\boldsymbol{Y}(z) = \boldsymbol{Y}_0 z^1 + \boldsymbol{Y}_1 z^2 + \boldsymbol{Y}_2 z^3 + \ldots$.

From the definition of $\boldsymbol{Y}_n$ one arrives at the matrix quadratic form for $\boldsymbol{Y}(z)$ as follows:

$$\begin{array}{rcl}
z^1 \boldsymbol{Y}_0 & = & \boldsymbol{I} z^1 \\
z^2 \boldsymbol{Y}_1 & = & (\boldsymbol{H}_a \boldsymbol{Y}_0 z^1 \boldsymbol{H}_s \boldsymbol{Y}_0 z^1) \\
z^3 \boldsymbol{Y}_2 & = & (\boldsymbol{H}_a \boldsymbol{Y}_1 z^2 \boldsymbol{H}_s \boldsymbol{Y}_0 z^1 + \boldsymbol{H}_a \boldsymbol{Y}_0 z^1 \boldsymbol{H}_s \boldsymbol{Y}_1 z^2) \\
\vdots & = & \vdots \\
z^{n+1} \boldsymbol{Y}_n & = & (\boldsymbol{H}_a [\boldsymbol{Y}_{n-1} z^n \boldsymbol{H}_s \boldsymbol{Y}_0 z^1 + \boldsymbol{Y}_{n-2} z^{n-1} \boldsymbol{H}_s . \\
 & & \quad \boldsymbol{Y}_1 z^2 + \ldots + \boldsymbol{Y}_0 z^1 \boldsymbol{H}_s \boldsymbol{Y}_{n-1} z^n]) \\
\hline
\boldsymbol{Y}(z) & = & z\boldsymbol{I} + \boldsymbol{H}_a (\boldsymbol{Y}_0 z^1 + \boldsymbol{Y}_1 z^2 + \boldsymbol{Y}_2 z^3 + \ldots) \\
 & & \quad \boldsymbol{H}_s (\boldsymbol{Y}_0 z^1 + \boldsymbol{Y}_1 z^2 + \boldsymbol{Y}_2 z^3 + \ldots)
\end{array}$$

Thus, $\boldsymbol{Y}(z)$ satisfies the matrix quadratic equation

$$\boldsymbol{Y}(z) = z\boldsymbol{I} + \boldsymbol{H}_a \boldsymbol{Y}(z) \boldsymbol{H}_s \boldsymbol{Y}(z). \tag{3}$$

Notice that this matrix quadratic form for $\boldsymbol{Y}(z)$ (equation (3)) is closely related to the common matrix quadratic equation for the matrix $\boldsymbol{G}$ that occurs in literature [14], [16]. In fact, "$\boldsymbol{Y}(1)\boldsymbol{H}_s$" is equivalent to the matrix $\boldsymbol{G}$ if the system under consideration has *MAP* processes, and $\boldsymbol{Y}(1)\boldsymbol{H}_s$ extends the functionality of $\boldsymbol{G}$ to our current more general situation. The current derivation is a combinatorial approach and implemented with dynamic programming techniques to keep the computational costs in control. Also, the matrix $\boldsymbol{Y}$ is constructed from the individual components as a limiting process which gives us qualitative insights into the recursive structure of the busy period.

Taking the derivative of $\boldsymbol{Y}(z)$ in equation (3),

$$\boldsymbol{Y}'(z) = \boldsymbol{I} + \boldsymbol{H}_a \boldsymbol{Y}'(z) \boldsymbol{H}_s \boldsymbol{Y}(z) + \boldsymbol{H}_a \boldsymbol{Y}(z) \boldsymbol{H}_s \boldsymbol{Y}'(z),$$

and evaluating at z=1, gives

$$\boldsymbol{Y}'(1) = \boldsymbol{I} + \boldsymbol{H}_a \boldsymbol{Y}'(1) \boldsymbol{H}_s \boldsymbol{Y} + \boldsymbol{H}_a \boldsymbol{Y} \boldsymbol{H}_s \boldsymbol{Y}'(1), \tag{4}$$

where,

$$\boldsymbol{Y} = \sum_{n=0}^{n=\infty} \boldsymbol{Y}_n. \tag{5}$$

Here $\boldsymbol{Y}$ should be directly computed from its individual components as a limiting process. Alternatively, if the busy period is known to be recurrent ($\rho < 1$), then $\boldsymbol{Y}$ can be computed by a fixed point iteration on the z-transform equation for $\boldsymbol{Y}(z)$ at $z = 1$. Empirical studies show that this fixed point iteration does converge when the busy period is recurrent, and a proof will be shown in future work.

Similarly, we can compute $\boldsymbol{Y}'(1)$ either by iteration on equation (4) or as a limiting process. The mean for the number served during this conditional first passage is given by

$$\mathrm{E}[N_{l,l-1}] = \boldsymbol{p}(0) \boldsymbol{Y}'(1) \boldsymbol{H}_s \boldsymbol{e}'. \tag{6}$$

Similarly the second moment is

$$\mathrm{E}[N_{l,l-1}^2] = \boldsymbol{p}(0) \boldsymbol{Y}''(1) \boldsymbol{H}_s \boldsymbol{e}'. \tag{7}$$

where $\boldsymbol{Y}''(1)$ is computed either as a limiting process or by iteration on

$$\begin{array}{rcl}
\boldsymbol{Y}''(1) & = & \boldsymbol{H}_a \boldsymbol{Y}''(1) \boldsymbol{H}_s \boldsymbol{Y} + 2\boldsymbol{H}_a \boldsymbol{Y}'(1) \boldsymbol{H}_s \boldsymbol{Y}'(1) \\
 & & + \boldsymbol{H}_a \boldsymbol{Y} \boldsymbol{H}_s \boldsymbol{Y}''(1).
\end{array}$$

If the $\boldsymbol{H}$'s are of size *m by m* then the computation of $\boldsymbol{Y}_n$ would take $3n$ matrix multiplications and n matrix summations. Hence the time complexity is of order $O(m^3 n)$, which is computationally manageable, especially since the matrix dimensions do not grow with path lengths. Also the matrix $\boldsymbol{Y}$ can be obtained by iteration on the z-transform equations using $O(m^3)$ computations per iteration.

4. Number of customers served in busy periods of an *MEP/MEP/1* system

As mentioned in the previous section, the busy period is a special case of the first passage process $D_{l,l-1}$ when $l = 1$. Let the internal state of the system at the start of a busy period be represented by $\boldsymbol{p}_{bp}$. Assuming that the utilization of the system is less than one ($\rho < 1$) and hence that a busy period always ends, this starting vector ($\boldsymbol{p}_{bp}$) is the normalized invariance vector for the start of a random busy period and is the solution to the following equation

$$\boldsymbol{p}_{bp} \boldsymbol{Y} \boldsymbol{H}_s \boldsymbol{Y}_a = \boldsymbol{p}_{bp}.$$

i.e., $\boldsymbol{p}_{bp}$ is the normalized left eigenvector corresponding to an eigenvalue of 1 for the matrix $\boldsymbol{Y} \boldsymbol{H}_s \boldsymbol{Y}_a$. The intuition is that if the process starts in $\boldsymbol{p}_{bp}$ at the start of a random busy period, its value at the start of the next busy period is given by traversing one of the possible paths $\boldsymbol{p}_{bp} \boldsymbol{Y}$, followed by the final departure $\boldsymbol{H}_s$

(back to state zero), after which only the arrival process is active until the next arrival event $\boldsymbol{Y}_a$ (computed as $\boldsymbol{V}_a\boldsymbol{L}_a$), thus starting the next busy period.

Once the starting vector for a busy period is known, the expressions for Prob$[N_{1,0} = n]$ and E$[N_{1,0}]$ follow directly from the results in the previous section. Hence, the probability that exactly n customers are served during a busy period is given by,

$$d_{n,1} = \text{Prob}[N_{1,0} = \text{n}\,] = \boldsymbol{p}_{bp}\boldsymbol{Y}_{n-1}\boldsymbol{H}_s\boldsymbol{e}', \quad n \geq 1,$$

and mean number of customers served during a busy period is

$$\text{E}[N_{1,0}] = \boldsymbol{p}_{bp}\boldsymbol{Y}'(1)\boldsymbol{H}_s\boldsymbol{e}'.$$

We summarize the procedure to compute these metrics in Algorithm 1.

Algorithm 1 To compute the Prob$[N_{1,0} = n]$ and mean for the number served during busy period of a *MEP/MEP/1* system

---

1: Setup $\boldsymbol{H}_a$ and $\boldsymbol{H}_s$ from the arrival and service process representations.

2: Compute $\boldsymbol{Y}$ by a fixed point iteration on

$$\boldsymbol{Y} = \boldsymbol{I} + \boldsymbol{H}_a\boldsymbol{Y}\boldsymbol{H}_s\boldsymbol{Y},$$

using, $\boldsymbol{Y}^{(0)} = \boldsymbol{I}$

$$\boldsymbol{Y}^{(i)} = (\boldsymbol{I} + \boldsymbol{H}_a\boldsymbol{Y}^{(i-1)}\boldsymbol{H}_s\boldsymbol{Y}^{(i-1)}), \quad i > 0.$$

Alternately, $\boldsymbol{Y}$ can be computed as a limiting process. See equation 5.

3: Find $\boldsymbol{p}_{bp}$, the left eigenvector corresponding to an eigenvalue of 1 for $\boldsymbol{Y}\boldsymbol{H}_s\boldsymbol{V}_a\boldsymbol{L}_a$.

4: To compute Prob$[N_{1,0} = n]$:

- Compute $\boldsymbol{Y}_{n-1}$ using, $\boldsymbol{Y}_0 = \boldsymbol{I}$,

$$\boldsymbol{Y}_{n-1} = \sum_{i=0}^{n-2}\boldsymbol{H}_a\boldsymbol{Y}_{n-i-2}\boldsymbol{H}_s\boldsymbol{Y}_i \qquad n > 1$$

- Probability that exactly n customers are served in a busy period is

$$\text{Prob}[\,N_{1,0} = \text{n}\,] = \boldsymbol{p}_{bp}\boldsymbol{Y}_{n-1}\boldsymbol{H}_s\boldsymbol{e}', \qquad n \geq 1.$$

5: To compute the mean number served in a busy period:

- Find $\boldsymbol{Y}'(1)$ using fixed point iteration on

$$\boldsymbol{Y}'(1) = \boldsymbol{I} + \boldsymbol{H}_a\boldsymbol{Y}'(1)\boldsymbol{H}_s\boldsymbol{Y} + \boldsymbol{H}_a\boldsymbol{Y}\boldsymbol{H}_s\boldsymbol{Y}'(1).$$

- Mean for number served is given by,

$$E[N_{1,0}] = \boldsymbol{p}_{bp}\boldsymbol{Y}'(1)\boldsymbol{H}_s\boldsymbol{e}'.$$

---

5. Numerical Results

Using the general derivation for the *MEP/MEP/1* system presented above, we compare our results to existing solutions for the number served during the busy period for an *M/M/1* and an *M/D/1* system. We then compare and validate our analytical results with trace driven simulations for *M/MEP/1*, *MEP/M/1* and *MEP/MEP/1* systems. Finally, we perform parametric studies on an *MEP/MEP/1* system using our analytical solutions.

5.1. Comparison to M/M/1 and M/D/1:

For the *M/M/1* case, the probabilities that $n+1$ customers are served in a busy period is given by [20]

$$\text{Prob}[N_b = n+1] = \frac{1}{n+1}\binom{2n}{n}\frac{\lambda^n\mu^{n+1}}{(\lambda+\mu)^{2n+1}}, \ \ n \geq 0,$$

where the combinatorial multiplier is the n^{th} Catalan number. Our results match exactly with this closed form solution and, as we mentioned in Section 3, we consider our derivation as a generalization of Catalan numbers for matrices.

For the M/G/1 case, a closed form explicit result is known when the service distribution is deterministic [8]. In this case, the probability of n number of customers served in a busy period (f_n) is given by the Borel distribution

$$f_n = \frac{1}{n}\frac{(\lambda\tau n)^{n-1}}{(n-1)!}e^{-\lambda\tau n}, \quad n \geq 1.$$

Consider now the ME density with representation $< \boldsymbol{p}_5, \boldsymbol{B}_5, \boldsymbol{e}_5 >$, where

$$\begin{aligned}
\boldsymbol{p}_5 &= \begin{bmatrix} 1 & \frac{3}{10} & \frac{7}{160} & \frac{1}{400} & \frac{1}{7680} \end{bmatrix} \\
\boldsymbol{B}_5 &= \begin{bmatrix} 0 & \frac{3}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{3}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{3}{2} \\ 480 & -576 & 300 & -90 & 15 \end{bmatrix} \\
\boldsymbol{e}_5 &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}.
\end{aligned}$$

This ME form represents the function

$$\begin{aligned}
f(t) &= \frac{1}{960}\,(12939 - 14896\,\cos{(3\,t)} \\
&\quad -9504\,\sin{(3\,t)} + 2017\,\cos{(6\,t)} \\
&\quad +4344\,\sin{(6\,t)})e^{-3\,t}
\end{aligned}$$

The above *ME* is an example of a distribution that is not also of a Phase type because the density is equal to zero for various values of t as can be seen in fig.3. This distribution has a mean of 1 and c^2 of $\frac{1}{12}$. The ten fold convolution of this density has a mean of 1.0 and a squared coefficient of variation (c^2) of 0.004, and is used in this paper to approximate a deterministic distribution. With this *ME* as the service process representation and with a Poisson arrival stream with a mean rate $\lambda = 0.8$, we get the probabilities shown in Table 2.

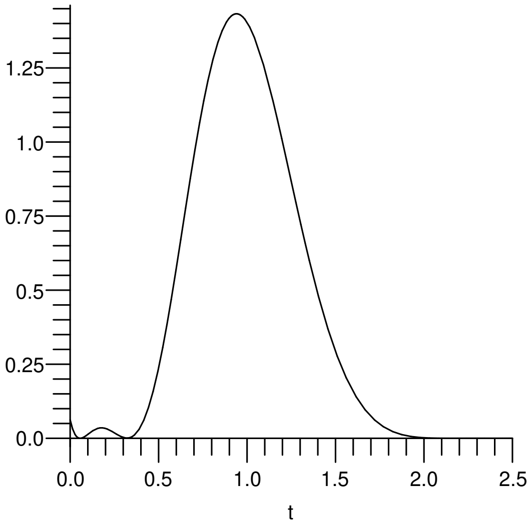

Figure 3. ME Density that touches the x-axis multiple times

Table 2. M/D/1 Comparison, Utilization = 0.8

n	Borel distribution Prob$[N_{1,0} = n]$	Our Result Prob$[N_{1,0} = n]$
1	0.4493289641	0.449926477600
2	0.1615172144	0.161409917100
3	0.08708923515	0.086978314120
4	0.05565399583	0.055568460030
5	0.03907336297	0.03900843250

Please note that even with the approximation to the deterministic distribution, the results are very close to the known Borel distribution.

5.2 MAP/MAP/1 System

Since *MAP*'s form a subset of the *MEP*'s, we can compute these probabilities (Prob$[N_{1,0} = n]$) for a *MAP/MAP/1* system. Consider a *MAP/MAP/1* system where the arrival is represented by

$$\boldsymbol{D}_0 = \begin{bmatrix} -7.1041 & 0 \\ 0 & -0.3959 \end{bmatrix},$$
$$\boldsymbol{D}_1 = \begin{bmatrix} 6.9916 & 0.1125 \\ 0.1125 & 0.2834 \end{bmatrix},$$

and the service process is represented by the rate matrices

$$\boldsymbol{D}_0 = \begin{bmatrix} -9.4721 & 0 \\ 0 & -0.5279 \end{bmatrix},$$
$$\boldsymbol{D}_1 = \begin{bmatrix} 9.3221 & 0.15 \\ 0.15 & 0.3778 \end{bmatrix}.$$

This is equivalent to an *MEP/MEP/1* system where for both the arrival and service processes the $\boldsymbol{B}$'s and $\boldsymbol{L}$'s can be derived from the corresponding $\boldsymbol{D}_0$'s and $\boldsymbol{D}_1$'s, i.e., from the arrival process $\boldsymbol{D}$'s we can get, $\boldsymbol{B}_a = -\boldsymbol{D}_0$, $\boldsymbol{L}_a = \boldsymbol{D}_1$, and from the service process $\boldsymbol{D}$'s we get, $\boldsymbol{B}_s = -\boldsymbol{D}_0$, $\boldsymbol{L}_s = \boldsymbol{D}_1$ respectively. This system has a utilization of 0.75 with a correlation decay parameter of 0.7 and c^2 of 9.0 for both the arrival and service processes. The corresponding probabilities are shown in Table 3.

Table 3. MAP/MAP/1 System, Utilization = 0.75

n	Prob$[N_{1,0} = n]$
1	0.63060456
2	0.12076481
3	0.05671077
4	0.03417086
5	0.02313089
$n > 5$	0.13461808

5.3 Simulation Results

For simulations, we generate traces using an ME process that is correlated. For this purpose, we use a hyper-exponential distribution with starting vector ($\boldsymbol{p}$), where the rate matrix ($\boldsymbol{B}$) is adjusted for the required c^2 (squared coefficient of variation) and the event transition matrix $\boldsymbol{L}$ is adjusted to control the correlation decay. It has the *ME* representation

$$\begin{aligned} \boldsymbol{p} &= \begin{bmatrix} p_1 & 1-p_1 \end{bmatrix}, \\ \boldsymbol{B} &= \lambda \begin{bmatrix} 2p_1 & 0 \\ 0 & 2(1-p_1) \end{bmatrix}, \\ \boldsymbol{L} &= \boldsymbol{B}\boldsymbol{e}'\boldsymbol{p}, \end{aligned}$$

where $p_1 = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{c^2-1}{c^2+1}}$. This process is an uncorrelated sequence. In order to construct correlated processes with geometrically decaying covariances that share the same marginals, we use the approach presented in [15]. Define $\boldsymbol{L}^{(\gamma)}$ for $-1 < \gamma < 1$ as

$$\boldsymbol{L}^{\gamma} = (1-\gamma)(\boldsymbol{B}\boldsymbol{e}'\boldsymbol{p} - \boldsymbol{B}) + \boldsymbol{B}. \tag{8}$$

The $\boldsymbol{L}^{(\gamma)}$ thus constructed introduces geometrically decaying correlations in the process, while leaving the marginals (and therefore the c^2) invariant.

5.3.1 M/MEP/1 System

For an *M/MEP/1* system, the effect of increasing the c^2 on the probabilities for n customers being served during a busy period while keeping γ (correlation decay parameter) at 0.99 is shown in Table 4. As can be seen from the table, the simulation results follow the analytic results closely. As the c^2 of the service process increases, there will be many requests with short service demands (compared to interarrival times), hence increasing the count of busy periods in which fewer customers are served. However, there will also be arrivals that have longer service demands, but since they are correlated, they tend to cause fewer very long busy periods, hence not contributing significantly to the count of busy periods.

5.3.2 MEP/M/1 System

As can be seen in Fig. 4, as the c^2 of the arrival process increases, the probability for only one customer served in a busy period decreases. In other words, the relative number of busy periods serving one customer is decreasing.

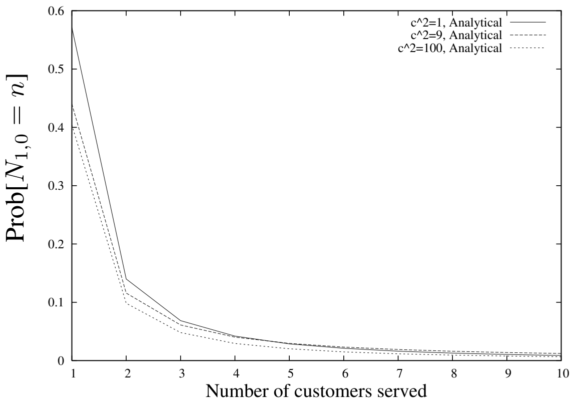

Figure 4. *MEP/M/1*: **Effect of increasing** c^2 **in uncorrelated case**

The effect of increasing the c^2 of the arrival process in a correlated vs non-correlated *(MEP/M/1 vs ME/M/1)* system is interesting to note (See Fig. 5). When the c^2 is increasing for the non-correlated case, the number of busy periods with fewer customers served decreases and the busy periods with more number of customers served gradually increases, hence the probability for one customer served in a busy period decreases (Prob[N_b = 1] = 0.404 for a $c^2 = 100$ and $\gamma = 0$). On the other hand, when the arrival process is highly correlated, there are a few busy periods that are extremely long and there are fewer busy periods where only one customer is served (as compared to the normal *M/M/1* case). Because of these extremely long busy periods and a decrease in total busy period count, the probability that only one customer is served increases (Prob[N_b = 1] = 0.974 for a $c^2 = 100$ and $\gamma = 0.99$), even though the absolute count of busy periods where exactly one customer is served decreases.

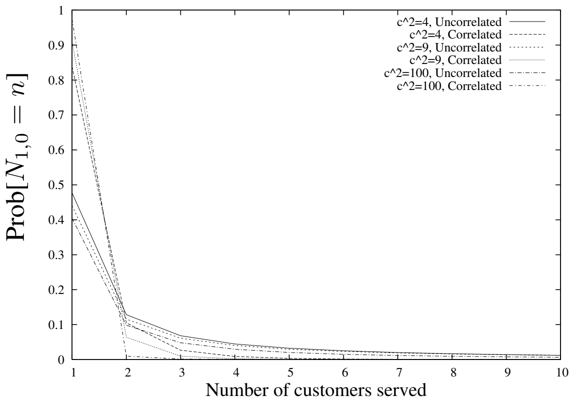

Figure 5. *MEP/M/1*: **Effect of correlation on Prob[$N_b = n$]**

5.3.3 Parametric studies using the *MEP/MEP/1* model

In this section we show how the values of c^2 and γ affect the system under study. For this purpose we use the general derivation used for the *MEP/MEP/1* system. With γ fixed at 0.99 for both the arrival and service processes, we increase the value of c^2 for both the processes from 4 to 100 while keeping the system utilization at 0.75. A c^2 of 100 and γ of 0.99 represents a system where the arrivals and service demands are both very erratic and correlated (bursty). Fig. 6 represents this effect.

It should be noted that the probability density for the number served for a highly correlated and variant

Table 4. Simulation vs Analytical for M/MEP/1, Utilization = 0.75

	$c^2 = 1$		$c^2 = 9$		$c^2 = 100$	
	Analytical	Simulation	Analytical	Simulation	Analytical	Simulation
n	Prob$[N_b = n]$					
1	0.571428571	0.57117927	0.715999851	0.71601889	0.726246184	0.72640603
2	0.139941691	0.139843148	0.145404403	0.14564553	0.144367891	0.144379278
3	0.068542869	0.068994283	0.059055301	0.059103117	0.057396756	0.057232489
4	0.041965022	0.042159178	0.029981189	0.029880241	0.028524242	0.028464697
5	0.028776015	0.028677701	0.017047444	0.016780497	0.015876653	0.015988375
6	0.021141562	0.021184264	0.010385734	0.010434037	0.009468192	0.009456229
7	0.016272223	0.016016045	0.006628639	0.006665474	0.005915325	0.005884958
8	0.012951361	0.013016432	0.004374999	0.004417871	0.003821632	0.003816901
9	0.01057254	0.01059373	0.002961674	0.002941219	0.002532298	0.002588751
10	0.008803257	0.008900555	0.00204508	0.002003509	0.001711516	0.0017103

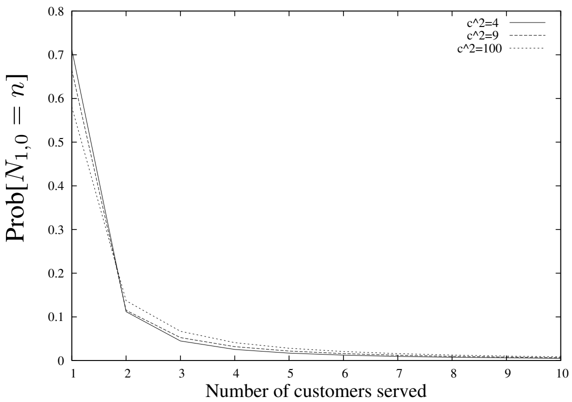

Figure 6. *MEP/MEP/1*: **Effect of increasing** c^2

MEP/MEP/1 system matches very closely with a simple *M/M/1* system. For example, the probability for serving exactly one customer has a value of 0.712 for a c^2 of 4 and goes down to 0.580 when the c^2 is 100, which is very close to that in an *M/M/1* system, 0.571. This result is quite counterintuitive, since we would expect the busy periods of a highly correlated *MEP/MEP/1* system to be somewhat different than that of an *M/M/1* system. Note however that only the relative count of busy periods that serve n customers stays the same. The c^2 for number served during a busy period however changes from 5.25 for an *M/M/1* system to 210 for an *MEP/MEP/1* system. Hence in an *MEP/MEP/1* system, there are some busy periods that are extremely long even though the averages look similar to an *M/M/1* system.

6 Summary and Future work

In this paper we derived closed form recursive solutions to compute the probability density for n customers served during the first passage, $D_{l,l-1}$, in a correlated *MEP/MEP/1* system. These conditional first passages provide us with tools to study similar first passages starting from a random or an environment-defined starting vector. We then analyzed the busy period of a *MEP/MEP/1* queue as a special case of these first passages and studied how these performance metrics are affected by the correlation in arrival and service processes. This approach to the busy period gives us qualitative insight into its structure and lays a general framework to analyze other transient system properties. The algorithms developed are easily programmable using dynamic programming techniques and can be incorporated into real life performance analysis tools. Future work will involve studying first passage time distributions and k-busy periods based on similar analytical techniques as presented in this paper.

Acknowledgments

We are thankful to the anonymous reviewers who have provided very useful input. Some of the additions mentioned are being studied.

References

[1] J. Abate and W. Whitt, (1992). " Numerical Inversion of Probability Generating Functions". *Operations Research Letters*, Vol. 12, 245-251

[2] J. Abate and W. Whitt, "Approximations for the M/M/1 Busy Period". *Queueing Theory and its Applications*,

Liber Amicorum for Professor J. W. Cohen, North-Holland, Amsterdam, 1988, pp. 149–191

[3] M. Agarwal, "Distribution of number served during a busy period of GI/M/1/N queues-lattice path approach ". *Journal of Statistical Planning and Inference*, Vol. 101, 2002, Pg. 7–21.

[4] S. Asmussen and G. Koole, " Marked point processes as limits of Markovian arrival streams". *Journal of Applied Probability*, Vol. 30 (1993). No. 2, 365-372.

[5] S. Asmussen and M. Bladt, " A sample path approach to mean busy periods for Markov-modulated queues and fluids". *Advances in Applied Probability*, Vol. 26, No. 4, 1117-1121(1994).

[6] O. J. Boxma and V. Dumas, " The busy period in the fluid queue". *Centrum voor Wiskunde en Informatica (CWI)*, Amsterdam, Netherlands. PNA-R9718, 1997.

[7] G. L. Choudhury, D. Lucantoni and W. Whitt, (1994). " Multidimensional transform inversion with applications to the transient M/G/1 queue". *Annals of Applied Probability* Vol. 4, 719– 740.

[8] R. Cooper, " Introduction to Queueing Theory". Page 231, ISBN-10: 0444003797.

[9] A. Heindl and M. Telek. " Output models of MAP/PH/1(/K) queues for an efficient network decomposition". *Performance Evaluation*, Vol. 49(1-4):321-339, 2002.

[10] A. Lee, A. Van de Liefvoort and V. Wallace, " Modeling correlated traffic with generalized IPP". *Performance Evaluation*, Vol. 40 (2000), Pg. 99-114.

[11] L. Lipsky, P. Fiorini, W.J. Hsin and A. Van de Liefvoort, " Auto-correlation of lag-k for customers departing from semi-Markov processes". *Tech. report, Institut für Informatik*, Technische Universität München Technical Report, 342/04/95.

[12] L. Lipsky, " Queueing Theory: A Linear Algebraic Approach". New York: MacMillan, 1992. ISBN-10: 0023709529.

[13] D. Lucantoni, " The $BMAP/G/1$ Queue: A Tutorial". *Lecture Notes in Computer Science*, Vol.729, 1993, Pg. 330-358.

[14] D. Lucantoni, " Further transient analysis of the $BMAP/G/1$ Queue". *Special issue in honor of Marcel F. Neuts. Communications in Statististics - Stochastic Models*, Vol. 14 (1998), no. 1-2, 461–478.

[15] K. Mitchell, " Constructing Correlated Sequence of Matrix Exponentials with Invariant First-Order Properties", *Operations Research Letters*, vol. 28 $no.1$, pages 27-34, 2001.

[16] M. Neuts, " Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach". The Johns Hopking Univ. Press, 1981. ISBN: 0486683427.

[17] L.-M. Le Ny, B. Sericola. " Busy Period Distribution of the BMAP/PH/1 Queue". *Proceedings of the 9th International Conference on Analytical and Stochastic Modelling Techniques (ASMT)*, Darmstadt, Germany, June 2002.

[18] P. Peart and W.J. Woan, " Dyck paths with no peaks at height k" , *Journal of Integer Sequences*, vol. 4, 2001, Article 01.1.3A.

[19] R. P. Stanley, " Enumerative Combinatorics" Vol. 1. Cambridge, England: Cambridge University Press, 1999a. ISBN-10: 0521663512.

[20] L. Takacs, " Introduction to the Theory of Queues", New York, Oxford University Press 1962. See pages 31-37.

[21] L. Takacs, " A Generalization of the Ballot Problem and its Applications in the Theory of Queues". *Journal of the American Statistical Association*, Vol.57, No. 298 (June, 1962), Pg. 327-337.

[22] A. Van de Liefvoort and A. Heindl, " Approximating matrix-exponential distribtions by global randomization". *Stochastic Models*, Vol. 21 (2005), No. 2-3, Pg. 669-693.