SAMPLE PATH ANALYSIS OF STOCHASTIC PROCESSES: BUSY PERIODS OF AUTO-CORRELATED SINGLE SERVER QUEUES

A DISSERTATION
IN
Computer Networking
and
Telecommunications Networking

Presented to the Faculty of the University
of Missouri–Kansas City in partial fulfilment of
the requirements for the degree

Doctor of Philosophy

by
Chaitanya N. Garikiparthi

M.S., University of Texas at Dallas, Texas, USA, 2001
B.E., CBIT, Osmania University, India, 1999

Kansas City, Missouri
2007

SAMPLE PATH ANALYSIS OF STOCHASTIC PROCESSES:

BUSY PERIODS OF AUTO-CORRELATED SINGLE SERVER QUEUES

Chaitanya N. Garikiparthi, Candidate for the Doctor of Philosophy Degree

University of Missouri–Kansas City, 2007

ABSTRACT

A number of processes that occur in nature as well as those that are a manifestation of human activities are correlated in nature. They can be described by stochastic non-markovian processes, and most known results in theory are in the steady state domain, assuming that the system has been in operation for a long enough time (ideally infinitely long), and that the state in which the system starts has no effect on the current behavior of the system. Nevertheless steady state assumptions do not hold in many applied situations as the system does not operate for infinitely long and perhaps even gets restarted every so often. In this thesis we provide a framework to stochastically track these processes. Application of this theory provide valuable insights into the transient behavior of these stochastic processes and allows us to model and study the effect of auto-correlations in the driving processes on transient probabilistic (performance) metrics of interest. In particular, we study the busy time (both length and number served) of the single server queue.

Applications of the work shown in this thesis are abound. Most processes in

telecommunications and computer networks exhibit a high degree of variance and are known to exhibit serial correlations across multiple time scales. In order to develop accurate models to represent these systems, we allow the arrival and the service processes that characterize the system to be both general and correlated. We specifically study the busy period and other first passages of an auto-correlated *MEP/MEP/1* queueing system to demonstrate the application of tracking these memory-full processes.

Representing the current state of a system using a relevant starting state vector, and by allowing the driving process to carry correlations across state transitions of the underlying quasi-markovian chain enables us to track these paths very accurately. The analysis presented here is in the transient domain and does not require the underlying processes to be in a steady state. The flexibility that is achieved by being able to model extremely variant (general) process which are allowed to be auto-correlated allows us to accurately model many of these real life processes.

In the first part of the thesis we provide solutions to compute the probabilities for exactly 'n' customers being served in a busy period of *MEP/MEP/1* queueing systems, where both the arrival, and the service processes could both be general and correlated Matrix Exponential Processes. We then present matrix exponential representations to characterize the lengths of sample paths during these busy periods and derive expressions to compute moments for length of the busy period as well as for the number of customers served during the busy period. In the second part of the thesis, we study the effect of increase in threshold level and the correlations in the arrival and service processes on the

mean first passage time to go below a given threshold (given that the system just transitioned from the threshold level $n-1$, to level n). Finally we study the busy periods for finite queueing systems, and again study both the length of the busy period and the number of customers served during such a time.

This abstract of 490 words is approved as to form and content.

Appie van de Liefvoort, Ph.D.
Professor
Computer Science Electrical Engineering Department
School of Computing and Engineering

The undersigned, appointed by the Dean of the School of Graduate Studies, have examined a dissertation titled "Sample Path Analysis of Stochastic Processes: Busy Periods of Auto-correlated Single Server Queues," presented by Chaitanya N. Garikiparthi, candidate for the Doctor of Philosophy degree, and hereby certify that in their opinion it is worthy of acceptance.

Appie van de Liefvoort, Ph.D. Computer Science Electrical Engineering	Date
Cory Beard, Ph.D. Computer Science Electrical Engineering	Date
Deep Medhi, Ph.D. Computer Science Electrical Engineering	Date
Jerry Place, Ph.D. Computer Science Electrical Engineering	Date
Kenneth Mitchell, Ph.D. Computer Science Electrical Engineering	Date
Khosrow Sohraby, Ph.D. Computer Science Electrical Engineering	Date
Victor Wallace, Ph.D. Electrical Engineering and Computer Science, University of Kansas	Date

CONTENTS

ABSTRACT ii

LIST OF TABLES viii

LIST OF ILLUSTRATIONS ix

ACKNOWLEDGEMENTS xi

Chapter

1 SAMPLE PATH ANALYSIS OF CORRELATED QUEUES 1

1.1 Motivation 1

1.2 Busy Period Analysis 3

1.3 Background Material 6

1.4 Literature Survey 10

1.5 Dissertation Structure 12

2 MATRIX EXPONENTIAL PROCESS 14

2.1 Matrix Exponentials 14

2.2 Matrix Exponential Process 20

2.3 Concurrent *MEP*'s and Hat Spaces 22

3 PROBABILITY MASS FUNCTION FOR NUMBER OF CUSTOMERS SERVED DURING THE BUSY PERIOD OF A CORRELATED *MEP/MEP/1* SYSTEM 24

3.1 Introduction 24

3.2 Model Description 25

3.3 Conditional Sample Path Analysis of First Passages in an *MEP/MEP/1* System 28

3.4 Number of Customers Served in Busy Periods of an *MEP/MEP/1* System 35
3.5 Numerical Results 37
3.6 Summary 47
4 BUSY PERIOD LENGTH AND HIGHER LEVEL FIRST PASSAGES 49
4.1 Introduction 49
4.2 Conditional Density for the *min(A,S)* Process 49
4.3 ME Representation for The Length of a Sample Path 50
4.4 Conditional Laplace Transform of a Sample Path During a Busy Period . 53
4.5 Mean Length of a Busy Period 55
4.6 Mean First Passage Time for Different Threshold Levels 58
4.7 Paths That Cross a Given Level During a Busy Period 61
5 BUSY PERIOD ANALYSIS OF FINITE *QBD* PROCESSES 64
5.1 Introduction 64
5.2 Busy Period of a Finite $MEP/MEP/1$ Queue 65
5.3 Numerical Examples 75
5.4 Conclusions 80
6 CONCLUSIONS AND FUTURE WORK 82
6.1 Conclusions 82
6.2 Future Work 83
REFERENCE LIST 85
VITA 90

TABLES

Table		Page
1	$\boldsymbol{H}$ operators for different systems	28
2	M/D/1 Comparison, Utilization = 0.8	39
3	MAP/MAP/1 System, Utilization = 0.75	40
4	Simulation vs Analytical for M/MEP/1, Utilization = 0.75	42
5	Bounds for the First Three Normalized Moments of ME(2) Distributions	46
6	Paths of Height greater than h	62
7	MM1 Finite Queue: $d^s_{n,1}$ for a Utilization = 0.70	76
8	MEP/MEP/1 Finite Queue: $d^s_{n,1}$ for a Utilization = 0.73	77
9	Catalan like sequences related to finite queues	80

LIST OF ILLUSTRATIONS

Figure		Page
1	Birth Death Process	1
2	k-Stage Erlang Distribution	16
3	k-Stage Hyper-Exponential Distribution	17
4	k-stage Coxian Distribution	18
5	Paths serving exactly 3 customers during the first passage, $D_{l,l-1}$	29
6	Paths serving exactly n customers during the first passage, $D_{l,l-1}$	30
7	ME Density that touches the x-axis multiple times	39
8	*MEP/M/1*: Effect of increasing c^2 in uncorrelated case	43
9	*MEP/M/1*: Effect of correlation on Prob[$N_b = n$]	44
10	*MEP/MEP/1*: Effect of increasing c^2	45
11	*MEP(r1,r2,r3,γ_a)/MEP/1*: Effect of Third moment, Util:0.55	46
12	*MEP(r1,r2,r3,γ_a)/MEP/1*: Effect of Third moment, Util:0.83	47
13	*MEP(r1,r2,r3,γ_a)/MEP/1*: Effect of Third moment, Util:0.9, 0.99	48
14	Higher Level First Passages	58
15	Mean Lenght of First Passage From Level n to $n-1$	60
16	Paths within a channel of height h	61
17	Paths with exactly three arrivals and three departures	65
18	Paths with exactly three arrivals and three departures within a channel of width two	66

19 *G/G/1*: Effect of γ_a on mean number served 78

20 *G/G/1*: Effect of γ_a on c^2 for number served 79

21 *G/G/1*: Effect of γ_a on mean busy period length 80

ACKNOWLEDGEMENTS

First and foremost I am deeply grateful to my advisor Dr. Appie van de Liefvoort without whose constant support, encouragement and guidance i would not have completed this dissertation, and for kindling in me a love for numbers. Secondly many thanks go to my other half Renu Paruchuri without whose many sacrifices and constant encouragement i would have not made it through either, and to the cute little light in our life Maruti whom i have missed mostly during the past few years but who nevertheless filled our life with immeasurable joy.

Thanks to all my advisors, Dr. Cory Beard, Dr. Deep Medhi, Dr. Jerry Place, Dr. Kenneth Mitchell, Dr. Khosrow Sohraby and Dr. Victor Wallace for the guidance they have provided. Special thanks go to Dr. Kenneth Mitchell for the numerous long chats (sometimes lasting more than a few hours) which were always inspiring and many a times left me feeling guilty for occupying too much of his time, and to Dr. Khosrow Sohraby whose insightful questions helped our research immensely. Many thanks to my parents Seshulatha Bonam and Vittal Garikiparthi without whom none of this was possible and my brother Aditya Garikiparthi for all the good times, and all our family members. Finally many thanks to all my friends (here and afar) and staff here at UMKC who have made my last few years enjoyable. Notable among them, Jayesh Kumaran, Amit Sinha, Manish

Mehta, Gaurav Agrawal, Balaji Krithikaivasan, Shekhar Srivastava, Shi Zhefu, Jiazhen Zhou, Charlie Zhao, Armin Heindl, Muralikrishna Padavala, Aravind Thoram, Anand Pappuri, Pranojit Chandra, Srilakshmi Katragadda, Vasudeva Sai, Jagadish Bose, Ujwal Manuka, Pavan Kaja, Praveen Chekuri, Praveen Soma, Praveen Patlolla, Debby Dilks, Coretta Carter, Sharon Griffith and Rebecca Edmundson.

CHAPTER 1

SAMPLE PATH ANALYSIS OF CORRELATED QUEUES

1.1 Motivation

A number of stochastic processes that occur in nature and those that are a manifestation of human activities exhibit correlations and are hence memory-full. Most of these processes also show high degrees of variances and provide some unique challenges for researchers trying to build models to capture their behavior. We explore the concept of this ”memory-full”-ness and provide a framework to track these processes. Application of this theory provide valuable insights into the transient behavior of these stochastic processes and allows us to model and study the effect of auto-correlations in the driving processes on some of these transient probabilistic (performance) metrics.

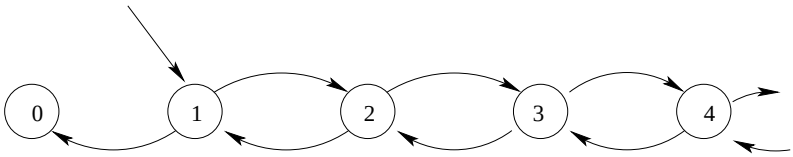

Figure 1: Birth Death Process

Let all possible states the system can assume at any given instant and the transitions from one state to another be represented by a generic birth-death process as shown in Fig. 1. Consider two generic states ‘A’ and ‘B’ representing two possible states the system can assume on such a state space. When the system is in a given state, a new arrival to the system causes a state transition to the right and customers departing (or service completion) cause the system to transition one step to the left. There are multiple

ways of traversing from state A to state B. If the underlying processes driving the chain are serially correlated, then the path that is traversed to reach B (from A) might effect the way in which the system starts once state B is reached; and might effect any other paths of which "A-B" is a sub-part thereof. Most processes occurring in nature are in fact known to show serial correlations at multiple time scales. In this thesis we study the effects of these serial correlations in the driving processes on how these sample paths progress by essentially tracking these correlated memory-full processes.

Applications of the work shown in this thesis are abound. We study busy periods and other first passages of an auto-correlated *MEP/MEP/1* Queueing system to demonstrate the potential and relate the results to applications in computer systems and networks. An application in the financial arena might be to model the progression of price data from a given price A to a price B. Discrete models are commonly used to model the future progression of price series and usually assume the price movement to be uncorrelated and assign equal probabilities to unit up and unit down moves and assume that the next move is independent of the previous. These price moves do show correlations often and we might be able to model the price movement as serially correlated processes. Most of the financial data is also perceived to be non-stationary which adds another dimension of complexity. In biological sciences similar analysis can be used to model the progression of certain growth and/or shrinkage processes. Another application in social sciences is to model the variation in population of a given species of wildlife. The population of a given species in a given region depends on various factors which possibly induce effects that are correlated. For example, a draught or a flood might easily cause the variation in the local population to be neither normal nor independent of previous time instances. The flexibility that is achieved by being able to model extremely variant (general) process

which are allowed to be auto-correlated allows us to accurately model many of these real life processes.

By representing the current state of a system by a relevant starting state vector, and by allowing the driving process to carry correlations across state transitions of the underlying quasi-markovian chain enables us to track these paths very accurately. Any process which can be modeled as a quasi-birth-death (QBD) chain can be studied using the techniques presented here. Note that the analysis presented here is in the transient domain and does not require the underlying processes to be in a steady state.

1.2 Busy Period Analysis

In this section we confine ourselves to the area of computer systems and networks and show how different problems in this area can be reduced to the general problem of probabilistic tracking of memory-full processes. 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.

Consider the operation of a consolidated server. An individual server that forms a part of this consolidation could be perceived to be highly utilized if its queues grow beyond a given threshold, or perhaps its cpu utilizations cross a given threshold level, or perhaps when its disc access times, or some other performance metrics of interest cross certain predefined or dynamically adjusted threshold levels. A question that often arises in such a case is when should a server (or process) be allocated more resources

(processing power, more buffers, etc). The administrator might choose to allocate more resources to this given server so as to improve system performance (response times etc), but if the available resources are limited (which they usually are), he/she needs to make an informed decision on whether or not to allocate additional resources. It is important to know how long the server will be in this state of high utilization. In other words how long will the server be in a state that is above a given threshold? Another interesting question is as follows. Suppose a first threshold level triggers a system to be actively monitored, but no further action (in addition to closely monitoring the system) is required until the system reaches a second threshold level, at which point some action is warranted from the administrator. As soon as the first threshold level is crossed, the solutions proposed in this thesis provide quantitative measures related to crossing the second threshold level based on the current state of the system. For example, probabilities of ever reaching the second level, probabilities for the number of events occurring before we come back to the first threshold level. This would enable the system administrator to pro-actively manage the resources at his disposal. We study this problem and few other related problems by posing them as modifications to the classical Busy Period problem.

Similarly, Queue length fluctuations during a Busy period provide quantitative measures to actively manage a network/system, whereby resources can be allocated in a proactive manner leading to optimal system performance. To study these queue length fluctuations we need to look at the transient Busy Period of the queue from a particular instant in time.

The Busy Period for a system is defined as 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. The systems behavior around some threshold level can be studied by analyzing the system starting immediately after it crosses this threshold level for the given server/process and ending when we reach this threshold for the first time thereafter. This process can hence be represented as a first passage process around this threshold level.

The problems we investigate in this thesis are as follows:

- The probabilities for serving exactly 'n' customers in a busy period of *MEP/MEP/1* queueing systems, where both the arrival, and the service processes could both be general and correlated Matrix Exponential Processes.
- Characterize the lengths of sample paths during these busy periods as an ME process and find the moments for the length of a busy period.
- Study the effect of increase in threshold level and the correlations in the Arrival and Service processes on the mean first passage time to go below a given threshold (given that the system just had a transition from the threshold level $n-1$, to level n).
- Probabilities that sample paths are of heights greater than 'h' during a first passage from level n to level $n-1$ and the effect of starting level and correlations on these probabilities.
- Effect of correlations on the probabilities for n customers being served and on busy periods durations for finite queueing systems.

1.3 Background Material

Most of the known results in queueing theory are related to steady state behavior. When studying such systems it is assumed that the system has been in operation for a long enough time (ideally infinitely long) that the state in which the system starts has no effect on the current behavior of the system. Nevertheless steady state assumptions do not hold in many applied situations as the system does not operate for infinitely long and perhaps even gets restarted every so often.

Busy periods of markovian queues provide insights into the transient nature of the system. Transient system state equations have been solved using a number of techniques. Refer to [16] for an excellent introduction to transient analysis and its historical perspective. We summarize the difference-equation technique in case of an M/M/1 queue and the technique used by Takács for an M/G/1 type queue as presented in [16].

1.3.1 Difference-equation technique

A busy period is defined as the interval of time from the instant a unit arrives at an empty system and its service begins, to the instant when the server becomes free for the first time thereafter. A busy period is a random variable (r.v.), being the first passage time from state 1 to state 0. Denote

T	=	length of the busy period
$b(t)$	=	pdf of T
$N^*(t)$	=	number present at time t during a busy period
		$\{N^*(t), t \geq 0\}$ is a zero-avoiding state process
$q_n(t)$	=	$Pr\{N^*(t) = n\}, n = 1, 2, \ldots$
$\bar{q}_n(s)$	=	Laplace Transform (LT) of $q_n(t)$.

We have $q_1(0) = 1, q_n(0) = 0, n \neq 1$, for $n \geq 2$, $q_n(t)$ satisfying the transient differential-difference equations. Thus, for an *M/M/1* system, using traditional notation,

$$q_n'(t) = -(\lambda + \mu)q_n(t) + \lambda q_{n-1}(t) + \mu q_{n+1}(t), \quad n \geq 2 \tag{1.3.1}$$

As the term $q_0(t)$ will not occur, the equation corresponding to $n = 1$ will be

$$q_1'(t) = -(\lambda + \mu)q_1(t) + \mu q_2(t). \tag{1.3.2}$$

Taking LT of eq. (1.3.1) and eq. (1.3.2),

$$\mu\bar{q}_{n+1}(s) - (s + \lambda + \mu)\bar{q}_n(s) + \lambda\bar{q}_{n-1}(s) = 0, \quad n \geq 2 \tag{1.3.3}$$

This is a difference equation of order 2. Solving the characteristic equation of eq. (1.3.3) and inverting the Laplace transform yields,

$$q_n(t) = \left(\frac{\lambda}{\mu}\right)^{n/2} \frac{n}{\lambda t} e^{-(\lambda+\mu)t} \boldsymbol{I}_n(2t\sqrt{\lambda\mu}), \quad n = 1, 2, \ldots \tag{1.3.4}$$

where $\boldsymbol{I}_n$ is the Bessel function of first kind [16], [15].

Conditioning upon the number of units present at any instant t which complete their service in (t, t+dt), and taking the limit as $dt \to 0$, gives the density for busy period length as

$$b(t) = \frac{1}{t}\rho^{-1/2} \exp^{-(\lambda+\mu)t} \boldsymbol{I}_1(2t\sqrt{\lambda\mu}). \tag{1.3.5}$$

The LST of T is given by

$$b^*(s) = \frac{(\lambda + \mu + s) - \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu}}{2\lambda}. \tag{1.3.6}$$

1.3.2 M/M/1 Busy Period - Matrix Exponential Form

We briefly introduce matrix exponential distributions and show an ME representation for the busy period of the *M/M/1* system; more precisely, we show a finite ME approximation to the *M/M/1* busy period. A matrix exponential (ME) distribution is defined as a probability distribution whose density can be written as

$$f(t) \;=\; \boldsymbol{p}\exp\left(-\boldsymbol{B}t\right)\boldsymbol{B}\boldsymbol{e}', \qquad t \geq 0, \tag{1.3.7}$$

where $\boldsymbol{p}$ 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, but not necessarily so. We will give a detailed introduction to matrix exponential distributions in the next chapter.

To obtain a matrix exponential form for the busy period of the M/M/1 system, we use its moments (coefficients from Taylor series expansion of eq. 1.3.6) as input to the moment matching algorithm [43]. Since the LST is not rational, we do not have a finite matrix exponential representation that matches all the moments of the distribution. We approximate the distribution to an arbitrary precision. The following finite approximation is obtained by matching the first nine moments,

$$\begin{aligned}
\boldsymbol{p} &= \begin{bmatrix} \frac{1}{\mu-\lambda} & 0 & 0 & 0 & 0 \end{bmatrix} \\
\boldsymbol{V} &= \begin{bmatrix}
\frac{\mu}{(\mu-\lambda)^2} & \frac{\mu}{(\mu-\lambda)^3} & 0 & 0 & 0 \\
\frac{\lambda}{\mu-\lambda} & \frac{\lambda+\mu}{(\mu-\lambda)^2} & \frac{\mu}{(\mu-\lambda)^3} & 0 & 0 \\
0 & \frac{\lambda}{\mu-\lambda} & \frac{\lambda+\mu}{(\mu-\lambda)^2} & \frac{\mu}{(\mu-\lambda)^3} & 0 \\
0 & 0 & \frac{\lambda}{\mu-\lambda} & \frac{\lambda+\mu}{(\mu-\lambda)^2} & \frac{\mu}{(\mu-\lambda)^3} \\
0 & 0 & 0 & \frac{\lambda}{\mu-\lambda} & \frac{\lambda+\mu}{(\mu-\lambda)^2}
\end{bmatrix} \\
\boldsymbol{e} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix},
\end{aligned}$$

where $\boldsymbol{V} = \boldsymbol{B}^{-1}$. Substituting $\lambda = \rho\mu$ and using similarity transformations we get the

following matrix exponential representation:

$$
\begin{aligned}
\boldsymbol{p} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \\
\boldsymbol{V} &= \frac{1}{(1-\rho)^2\mu} \begin{bmatrix} (1-\rho) & 1 & 0 & 0 & 0 \\ \rho(1-\rho) & (1+\rho) & 1 & 0 & 0 \\ 0 & \rho & (1+\rho) & 1 & 0 \\ 0 & 0 & \rho & (1+\rho) & 1 \\ 0 & 0 & 0 & \rho & (1+\rho) \end{bmatrix} \\
\boldsymbol{e} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}.
\end{aligned}
$$

1.3.3 Takács Integral equation for M/G/1 System

In this section we show the general methodology used to analyze the busy periods when one of the constituent processes is non-markovian, by taking the case of an $M/G/1$ system. Assume that a busy period is initiated by a single customer. As this initiating customer is in service, the i^{th} customer who arrives during this time period will be called the i^{th} descendant. Let

T = length of the busy period
$G(t)$ = Pr$\{T \leq t\}$
$B(t)$ = Pr$\{v \leq t\}$, where v is the service time
$G^*(s)$ = LST of T
$B^*(s)$ = LST of v

To obtain the busy period distribution, condition on two events - namely, on the duration of the service time v of the initiating customer (customer who starts the new busy period) and on the number A of arrivals during the service time of the initiating customer. Given that $v = x$ and $A = n$, then n sub-busy periods $T_1, \ldots, T_n$ are generated by the n

descendants. Assuming that the $T_i's$ are IID and are independent of x, we have

$$E\{e^{-sT}|v = x, A = n\} = e^{-sx}\left[G^*(s)\right]^n \tag{1.3.8}$$

Finally the LST of T is obtained by un-conditioning on v and A. Hence

$$E\{e^{-sT}\} = \int_{x=0}^{\infty}\sum_{n=0}^{\infty} E\{e^{-sT}|v = x, A = n\}Pr\{A = n\}dB(x) \tag{1.3.9}$$

Simplifying this equation results in the well known functional equation for the LST of busy period as

$$G^*(s) = B^*\left[s + \lambda - \lambda G^*(s)\right]. \tag{1.3.10}$$

Known solution techniques rely on finding either the Laplace transform of the busy period or its derivatives by iteration and then invert them back into the time domain. Specifically note the recursive definition for the Laplace transform of the busy period, $G^*(s)$. This structure is preserved even when one of the processes involved is quasi-birth-death [33] in nature and most known solution approaches revolve around solving for this transform (or a matrix transform thereof) using numerical techniques.

1.4 Literature Survey

For an *M/M/1* system, the probabilities for n customers being served during a normal busy period are known, see for example Takács [40]. 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 [41]. More recently, Ny and Sericola [37] 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 [32], [33] and derive the two dimensional transform for the joint distribution for the number served in a busy period and its length, which are numerically inverted [9].

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 [8] 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 [7] 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 [5] by splitting up the sample paths at suitable renewal epochs. Heindl and Telek [11] studied the busy period of a *MAP/PH/1* system. Osogami and Harchol - Balter [29] study the necessary and sufficient conditions to represent a general process as a Coxian distribution and as an application show that the number of stages which suffice for a busy period duration to be well-represented by a Coxian are solely determined by the service distribution of the first job in the busy period. In [30], Akar and Sohraby present a novel algorithmic approach to compute the stationary probability distribution of finite QDB chains using a hybrid of matrix geometric and invariant subspace methods. Lipsky studied first passage times in renewal *ME/ME/1* queues extensively and uses recurrence relations for their solutions [14].

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 and serially correlated. 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 computed using dynamic programming approach.

1.5 Dissertation Structure

In Chapter 2, we give an overview of matrix exponential distributions, including a list of common distributions with their matrix exponential representations. We also give a review of matrix exponential processes that allows us to represent processes which can be serially correlated and present some examples of matrix exponential processes and present a brief overview of product (hat) spaces.

In Chapter 3, we study the probability that n customers are served during 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 using dynamic programming techniques.

In Chapter 4, we first characterize the conditional min of two matrix exponential processes as a matrix exponential process and use that representation to construct the distribution functions and Laplace transforms for the time it takes to traverse any given sample path. We then use these individual sample path length representations to derive the Laplace transform for the entire busy period length and derive expressions to compute the mean busy period length. In the later half of this chapter, we study how the correlations in arrival and service processes effect the mean first passage time when we now consider a generic first passage from various starting levels. We then compute some busy period related performance metrics for various arrival and service processes.

In Chapter 5, we present an analysis of busy periods of finite *MEP/MEP/1* queues. We study how the moments and auto correlations in the arrival and service processes affect the busy period for these finite queues. Due to the restrictions presented by the finite queue boundaries and the effect of the boundary on the state transitions leading to the boundary, certain queueing studies, including the busy period analysis, are more intricate for the finite system as compared to their infinite counterparts. We derive the corresponding matrix quadratic equations for finite case and draw attention to the (dis)similarities to the matrix quadratic equation in relation to the infinite queueing situation and provide numerical examples.

In Chapter 6, we offer some concluding remarks as well as some directions for future work.

CHAPTER 2

MATRIX EXPONENTIAL PROCESS

2.1 Matrix Exponentials

A matrix exponential (ME) distribution [14] is a probability distribution function represented by the tuple $< \boldsymbol{p}, \boldsymbol{B}, \boldsymbol{e}' >$ where $\boldsymbol{p}$ is the starting operator for the process, $\boldsymbol{B}$ is the process rate operator, and the vector $\boldsymbol{e}'$ is a summing operator usually consisting of all ones. The density and the cumulative distribution functions are given by

$$f(t) = \boldsymbol{p} \exp(-\boldsymbol{B}t)\, \boldsymbol{B}\boldsymbol{e}', \quad t \geq 0, \tag{2.1.1}$$

$$F(t) = 1 - \boldsymbol{p} \exp(-\boldsymbol{B}t)\, \boldsymbol{e}', \quad t \geq 0. \tag{2.1.2}$$

The power moments of a matrix exponential distribution are given by

$$E[X^n] = \int_0^\infty t^i \, dF(t) = n!\, \boldsymbol{p}\boldsymbol{V}^n\boldsymbol{e}', \tag{2.1.3}$$

where $\boldsymbol{V} = \boldsymbol{B}^{-1}$. The matrix $\boldsymbol{V}$ is also known as the process time operator.

The Laplace-Stieltjes transform of a matrix exponential distribution is given by

$$F^*(s) = \int_0^\infty \exp(-s\,t)\, \boldsymbol{p} \exp(-\boldsymbol{B}\,t)\, \boldsymbol{B}\boldsymbol{e}' dt = \boldsymbol{p}\,(s\boldsymbol{I} + \boldsymbol{B})^{-1}\, \boldsymbol{B}\boldsymbol{e}'. \tag{2.1.4}$$

The class of matrix exponential distributions is identical to the class of distributions that possess a rational Laplace-Stieltjes transform, i.e., all distributions that have a rational Laplace-Stieltjes transform can be represented as a matrix exponential distribution. Distributions that do not have a rational Laplace-Stieltjes transform can be closely approximated by distributions having rational Laplace-Stieltjes transform (see [31]). These

representations need not be unique. If $< \boldsymbol{p}, \boldsymbol{B}, \boldsymbol{e} >$ is a matrix exponential representation of a distribution F, then $< \boldsymbol{p}\boldsymbol{X}^{-1}, \boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^{-1}, \boldsymbol{X}\boldsymbol{e} >$ is also a matrix exponential representation of F, where $\boldsymbol{X}$ is a non-singular matrix.

For any given rational Laplace-Stieltjes transform the problem of when there exists a corresponding matrix exponential distribution was addressed by Fackrell [17]. Given a sequence of moments of a distribution the problem of when is it a matrix exponential distribution of finite degree was addressed by Van de Liefvoort [27], who also proposed an algorithm for constructing a minimal matrix exponential representation.

The class of matrix exponential distributions have representations that closely resemble the phase-type distributions, which have an additional requirement that $\boldsymbol{p}$ and $\boldsymbol{B}$ are probabilistically interpretable. Below are some examples of matrix exponential distributions

- Exponential Distribution

  The density function $f(t)$ and the Laplace-Stieltjes transform $F^*(s)$ of an exponential distribution

  $$f(t) = \exp(-\lambda\, t)\, \lambda, \quad F^*(s) = \frac{\lambda}{\lambda + s}. \tag{2.1.5}$$

  A matrix exponential representation is

  $$\boldsymbol{p} = [1]\,, \ \boldsymbol{B} = [\lambda]\,, \ \boldsymbol{e}' = [1]\,. \tag{2.1.6}$$

- Erlang Distribution

  A k-stage Erlang distribution where the time spent in each stage is exponentially distributed with the rate λ is presented in Fig. 2.1 whose density function $f(t)$ and

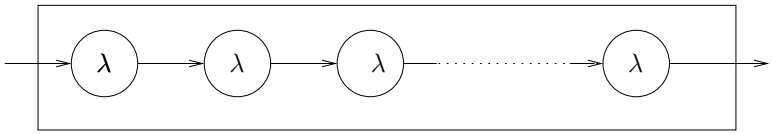

Figure 2: k-Stage Erlang Distribution

its corresponding Laplace-Stieltjes transform $F^*(s)$ are given by

$$f(t) = \frac{\exp(-\lambda\, t)\, \lambda^k\, t^{k-1}}{k!}, \qquad F^*(s) = \left(\tfrac{\lambda}{\lambda+s}\right)^k. \tag{2.1.7}$$

A matrix exponential representation is

$$\boldsymbol{p} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix},\ \boldsymbol{B} = \begin{bmatrix} \lambda & -\lambda & 0 & \cdots & 0 \\ 0 & \lambda & -\lambda & \cdots & 0 \\ 0 & 0 & \lambda & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{bmatrix},\ \boldsymbol{e}' = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

- Hyper-Exponential Distribution

  A k-stage hyper-exponential distribution where the time spent in stage i is exponentially distributed with the rate λ_i and the probability of starting in each stage is given by α_i is presented in Fig. 3. The density function of a k-stage hyper-exponential process $f(t)$ and its corresponding Laplace-Stieltjes transform $F^*(s)$ are given by

$$f(t) = \sum \alpha_i \exp(-\lambda_i\, t)\, \lambda_i, \qquad F^*(s) = \textstyle\sum_{i=1}^{k} \alpha_i\, \frac{\lambda_i}{(\lambda_i+s)}. \tag{2.1.8}$$

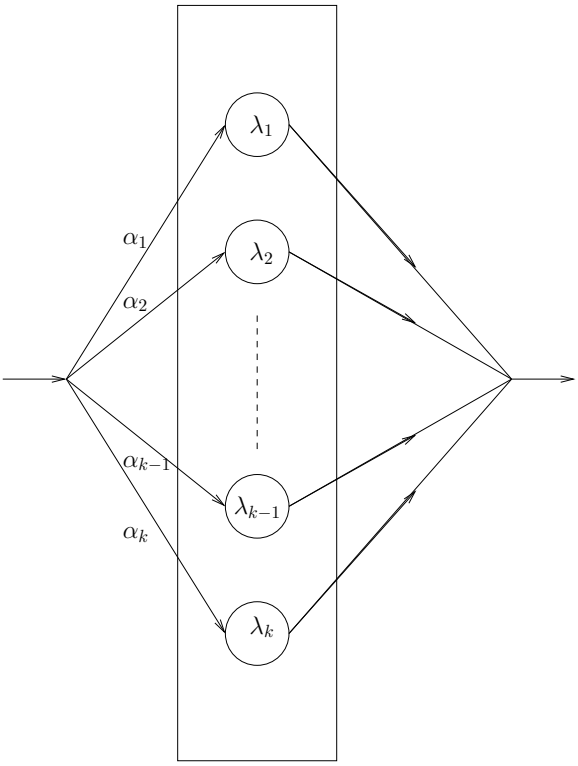

Figure 3: k-Stage Hyper-Exponential Distribution

An matrix exponential representation is

$$\boldsymbol{p} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_k \end{bmatrix}, \ \boldsymbol{B} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix}, \ \boldsymbol{e}' = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

- Coxian Distributions

  This class of distributions was introduced by Cox [31], who showed that any non-exponential probability distribution with rational Laplace-Stieltjes transform can be

represented as a series of exponential stages with possibly complex valued transition rates, and where with some probability the next stage is entered or with complementary probability the process stops. A k-stage Coxian distribution is presented in Fig. 4 where the time spent in stage i is exponentially distributed with the rate λ_i and the probability of entering that stage is α_i, both λ_i and α_i are possibly complex valued. The Laplace-Stieltjes transform $F^*(s)$ of a k-stage Coxian distribution is given in partial fraction form by

$$F^*(s) = \sum_{i=1}^{k} \alpha_i \prod_{j=1}^{i-1}(1-\alpha_i) \prod_{j=1}^{i} \frac{\lambda_j}{(s+\lambda_j)}. \tag{2.1.9}$$

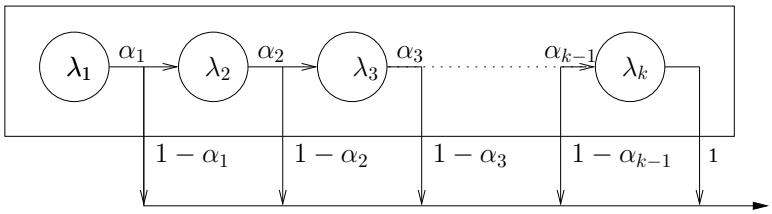

Figure 4: k-stage Coxian Distribution

An matrix exponential representation is

$$\boldsymbol{p} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix}, \ \boldsymbol{B} = \begin{bmatrix} \lambda_1 & -\alpha_1\,\lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & -\alpha_2\,\lambda_2 & \ddots & 0 \\ 0 & 0 & \lambda_3 & \vdots & 0 \\ \vdots & \ddots & \ddots & \vdots & \\ 0 & 0 & \cdots & 0 & \lambda_k \end{bmatrix}, \ \boldsymbol{e}' = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

- General Canonical Form from RLT

For any distribution that has a rational Laplace-Stieltjes transform, $F^*(s)$,

$$F^*(s) = \frac{b_0 + b_1\,s + \cdots + b_{m-1}\,s^{m-1}}{a_0 + a_1\,s + \cdots + a_{m-1}\,s^{m-1} + s^m}. \tag{2.1.10}$$

A matrix exponential representation $< \boldsymbol{p}, \boldsymbol{B}, \boldsymbol{e}' >$ for this distribution is given in companion canonical form as

$$\boldsymbol{p} = \begin{bmatrix} \frac{b_0}{a_0} & (-1)^1 \frac{b_1}{a_0} & \cdots & (-1)^{m-2} \frac{b_{m-2}}{a_0} & (-1)^{m-1} \frac{b_{m-1}}{a_0} \end{bmatrix},$$

$$\boldsymbol{B} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & 0 \\ (-1)^{m-1} a_0 & (-1)^{m-2} a_1 & \cdots & (-1)^1 a_{m-2} & a_{m-1} \end{bmatrix}, \ \boldsymbol{e}' = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

It is to be noted here that even if a given density function can be represented as a phase type, the equivalent canonical form shown above is not and does not have any phase type interpretations associated. Conversely, some density functions which cannot be expressed as phase types can still be represented using this canonical form as a matrix exponential .

- Phase Type Distributions

  Neuts [19] introduced the phase type distributions by defining a continuous time Markov chain with an absorbing state. A phase-type distribution defined by $(\widehat{\boldsymbol{\alpha}}, \boldsymbol{T})$ is the distribution of time until absorption in a finite-state, continuous-time Markov process with one absorbing state. The matrix $\boldsymbol{T}$ represents the transitions among the non absorbing states and the vectors $\widehat{\boldsymbol{\alpha}}$ is the entry vector giving the probability distribution of the initial state. A matrix exponential representation is given by

$$\boldsymbol{p} = \widehat{\boldsymbol{\alpha}}; \boldsymbol{B} = -\boldsymbol{T}, \ \boldsymbol{e}' = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

- The following is an example of a distribution that does not have a phase type representation. This distribution was introduced in [31] and the density function and corresponding Laplace-Stieltjes transform are given by

$$f(t) = 8\,\mu\,[sin(\mu\,t)]^2 \exp(-2\mu\,t), \quad t \geq 0$$
$$F^*(s) = \frac{16\,\mu^3}{(s+2\,\mu)\,(s^2+4\,s\mu+8\,\mu^2)}. \tag{2.1.11}$$

A matrix exponential representation using only real numbers,

$$\boldsymbol{p} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix},\ \boldsymbol{B} = \mu \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 2 \\ 15 & -15 & 8 \end{bmatrix},\ \boldsymbol{e'} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

The companion canonical representation is,

$$\boldsymbol{p} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix},\ \boldsymbol{B} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 16\,\mu^3 & -16\,\mu^2 & 6\,\mu \end{bmatrix}\ \boldsymbol{e'} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

For other properties of the matrix exponential distributions, see [14] and on methods to compute $f(t)$ see [20].

2.2 Matrix Exponential Process

The matrix exponential process (MEP) is defined by the joint density function of first k-successive intervals between events where the inter event times are matrix exponentially distributed

$$f_{1,2,\ldots,k}(x_1,\ldots,x_k) = \boldsymbol{p}(0)\,\exp(-\boldsymbol{B}x_1)\boldsymbol{L}\ldots\exp(-\boldsymbol{B}x_k)\boldsymbol{L}\boldsymbol{e'}, \tag{2.2.1}$$

then this describes an matrix exponential process, where $\boldsymbol{p}(0)$ is the state of the process at time zero (also referred to as the starting operator), and $\boldsymbol{L}$ is the instantaneous event generator matrix. The matrix $\boldsymbol{L}$ reflects the rate of transitions between the internal state

of the process immediately before the event and immediately after the event. The induced process $\boldsymbol{p}(0)\boldsymbol{Y}^k,\ (k = 0, 1, \dots)$ describes the sequence of states immediately after the start of a new interval at event times, where $\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{L}$. If the process is assumed to be covariance stationary, then the $\boldsymbol{p}(0)$ is the stationary vector for the process at embedded event points. Below are examples of some matrix exponential processes.

- Poisson process

  In a Poisson process, the intervals between consecutive events are independent and identically distributed exponential random variables. A Poisson process with the rate λ has a MEP representation given by

  $$\boldsymbol{p}(0) = [1],\ \boldsymbol{B} = [\lambda],\ \boldsymbol{L} = [\lambda]. \tag{2.2.2}$$

- Matrix Exponential Renewal Process

  Renewal process defines processes whose inter-event times are independent of each other. The event generator matrix for renewal process whose inter-event times are characterized by matrix exponential distributions is given by

  $$\boldsymbol{L} = \boldsymbol{B}\boldsymbol{e}'\boldsymbol{p}. \tag{2.2.3}$$

  Note that the rank of the matrix $\boldsymbol{L}$ is one for a renewal process.

- Markov Arrival Processes (MAP)

  The Markovian Arrival Process (MAP), which is a generalization of the PH distribution was introduced by Neuts [21] to model non Markovian point processes. A MAP is a non renewal process represented by two matrices $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ rather than a matrix and a vector as in the phase type distribution. The matrix $\boldsymbol{D}_0$ is similar to the $\boldsymbol{T}$ matrix for a PH distribution, which contains the transitions between the transient

states of the underlying Markov chain. The rows of the matrix $\boldsymbol{D}_1$ describe how the transient states of the underlying Markov chain are reentered after an absorption event.The equivalent MEP representation is given by

$$\boldsymbol{B} = -\boldsymbol{D}_0, \boldsymbol{L} = \boldsymbol{D}_1. \tag{2.2.4}$$

The Markov modulated Poisson process (MMPP) is a special case of MAP where the matrix $\boldsymbol{D}_1$ is diagonal i.e., event transitions do not result in change of state.

The expression for the lag-l covariance, the covariance between the first interval and the $(l+1)$th is

$$\text{cov}[X_1, X_{l+1}] = \boldsymbol{pV}(\boldsymbol{Y})^l\boldsymbol{Ve'} - (\boldsymbol{pVe'})^2, \;\; l \geq 0.$$

The auto-correlation at lag-l, r[l], can be found by dividing cov[X_1, X_l] by the variance

$$\text{var[X]} = 2\boldsymbol{pV}^2\boldsymbol{e'} - (\boldsymbol{pVe'})^2.$$

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 [22]). By implication, stationary MEPs are dense in the family of all stationary point processes as well, [23]. For additional details see [14, 24–26].

2.3 Concurrent *MEP*'s and Hat Spaces

It is not unusual that multiple processes each acting on their own operator spaces act concurrently on a given state. Kronecker product is one way of representing the embedding (or combining) of these two disjoint operator spaces, into a bigger

product space. In general, if $\boldsymbol{K}_1$ is an $m_1 \times n_1$ matrix operating on objects in space 1, and $\boldsymbol{K}_2$ is an $m_2 \times n_2$ matrix of space 2, the Kronecker product of $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$, denoted by $\boldsymbol{K}_1 \bigotimes \boldsymbol{K}_2$, is the matrix of size $(m_1 m_2) \times (n_1 n_2)$ that is obtained by multiplying each element of $\boldsymbol{K}_1$ by the full matrix, $\boldsymbol{K}_2$.

As a particular example, let an arrival process represented by $< \boldsymbol{p}_a, \boldsymbol{B}_a, \boldsymbol{L}_a, \boldsymbol{e}_a >$ and a service processes represented by $< \boldsymbol{p}_s, \boldsymbol{B}_s, \boldsymbol{L}_s, \boldsymbol{e}_s >$ act concurrently on the internal state of the system. Using Kronecker products we can construct a product space which represents the concurrency of these two process by embedding both the arrivals and services into the product space as follows [14].

$$\begin{aligned}
\widehat{\boldsymbol{e}_a} &= \boldsymbol{e}_a \otimes \boldsymbol{I}_s, & \widehat{\boldsymbol{e}_s} &= \boldsymbol{I}_a \otimes \boldsymbol{e}_s \\
\widehat{\boldsymbol{B}_a} &= \boldsymbol{B}_a \otimes \boldsymbol{I}_s, & \widehat{\boldsymbol{B}_s} &= \boldsymbol{I}_a \otimes \boldsymbol{B}_s \\
\widehat{\boldsymbol{L}_a} &= \boldsymbol{L}_a \otimes \boldsymbol{I}_s, & \widehat{\boldsymbol{L}_s} &= \boldsymbol{I}_a \otimes \boldsymbol{L}_s \\
\widehat{\boldsymbol{e}_a} &= \boldsymbol{e}_a \otimes \boldsymbol{I}_s, & \widehat{\boldsymbol{e}_s} &= \boldsymbol{I}_a \otimes \boldsymbol{e}_s
\end{aligned}$$

where, $\boldsymbol{I}_a$ and $\boldsymbol{I}_s$ are identity matrices in the arrival and service spaces respectively and the symbol $\hat{\cdot}$ (called caret or hat) represents a process in the embedded space. Once the matrices are embedded into the product space, the concurrent process rate matrix for example is given by $\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_a}$.

CHAPTER 3

PROBABILITY MASS FUNCTION FOR NUMBER OF CUSTOMERS SERVED DURING THE BUSY PERIOD OF A CORRELATED *MEP/MEP/1* SYSTEM

3.1 Introduction

In this chapter we study the probability that n customers are served during 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 using dynamic programming techniques.

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.

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 chapter 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.

3.2 Model Description

3.2.1 *QBD* Processes

A finite QBD process is a Markov process with infinitesimal generator $\tilde{\boldsymbol{Q}}$ [36], given by

$$\tilde{\boldsymbol{Q}} = \begin{bmatrix} \boldsymbol{B_0} & \boldsymbol{A}_0 & & & & \\ \boldsymbol{B_1} & \boldsymbol{A}_1 & \boldsymbol{A}_0 & & & \\ & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{C_0} \\ & & & & \boldsymbol{A}_2 & \boldsymbol{C_1} \end{bmatrix}. \tag{3.2.1}$$

Define the embedding operators $\boldsymbol{H}_a$ reflecting an arrival event occurring before the service and $\boldsymbol{H}_s$ representing a service event occurring before the arrival map into the *QBD* space as follows:

$$\boldsymbol{H}_a = (\boldsymbol{A}_1)^{-1}\boldsymbol{A}_0, \tag{3.2.2}$$

$$\boldsymbol{H}_s = (\boldsymbol{A}_1)^{-1}\boldsymbol{A}_2. \tag{3.2.3}$$

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.

The conditional probability that an arrival event (r.v. A) occurs before the service event (r.v. S), given that the starting vector is $\boldsymbol{p}(0)$ is given by

$$\Pr[\, A < S \;\mid\; \boldsymbol{p}(0)] = \boldsymbol{p}(0)\boldsymbol{H}_a\boldsymbol{e}'. \tag{3.2.4}$$

where 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 a QBD system, the conditional probability that two successive events are both arrivals, given that the process starts in $\boldsymbol{p}(0)$ is

$$\boldsymbol{p}(0)(\boldsymbol{H}_a)^2\boldsymbol{e}'. \tag{3.2.5}$$

The corresponding discrete-time QBD process is governed by

$$\tilde{\boldsymbol{P}} = \begin{bmatrix} \mathbf{0} & \boldsymbol{H}_1 & & & & \\ \boldsymbol{H}_3 & 0 & \boldsymbol{H}_a & & & \\ & \boldsymbol{H}_s & \mathbf{0} & \boldsymbol{H}_a & & \\ & & \ddots & \ddots & \ddots & \\ & & & \boldsymbol{H}_s & \mathbf{0} & \boldsymbol{H}_2 \\ & & & & \boldsymbol{H}_4 & \mathbf{0} \end{bmatrix}, \tag{3.2.6}$$

where $\boldsymbol{H}_1 = (\boldsymbol{B}_0)^{-1}\boldsymbol{A}_0$,$\boldsymbol{H}_2 = (\boldsymbol{A}_1)^{-1}\boldsymbol{C}_0$,$\boldsymbol{H}_3 = (\boldsymbol{A}_1)^{-1}\boldsymbol{B}_1$, and $\boldsymbol{H}_4 = (\boldsymbol{C}_1)^{-1}\boldsymbol{A}_2$.

In the particular systems that we study here will have $\boldsymbol{H}_3 = \boldsymbol{H}_s$ reflecting that the service process is suspended if no customers are present, without effecting the internal state. Also, in our case, $\boldsymbol{H}_4 = (\boldsymbol{I} - \boldsymbol{H}_a)^{-1}\boldsymbol{H}_s$.

3.2.2 $\boldsymbol{H}$ Operators

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

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

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. 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 $\boldsymbol{p}(0)(\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})^{-1}\widehat{\boldsymbol{L}_a}\ \cdot (\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})^{-1}\widehat{\boldsymbol{L}_a}\boldsymbol{e}'$. The operators $\boldsymbol{H}_a$ for arrival event happening before the service and $\boldsymbol{H}_s$ for service event happening before the arrival are given by unconditioning on the initial state of the system.

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

Essentially these $\boldsymbol{H}$ operators allow us to track the path evolution by considering event transitions embedded 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 [14].

3.3 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 $\boldsymbol{p}(l)$ 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 $\boldsymbol{p}(l)$ and $\boldsymbol{e}'$ respectively.

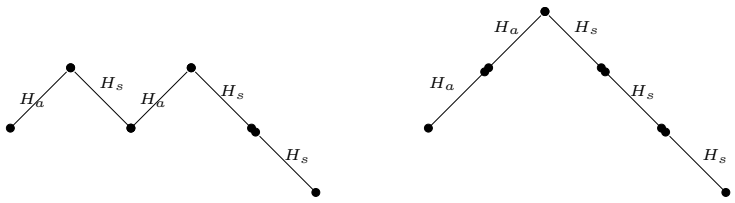

Figure 5: Paths serving exactly 3 customers during the first passage, $D_{l,l-1}$

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 [39]. 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 $C_n = \sum_{i=0}^{n-1} C_i C_{n-i-1},\ C_0 = C_1 = 1,$ by using dynamic programming techniques. 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. 5 and the probabilities associated with each of those paths are $\boldsymbol{p}(l)\boldsymbol{H}_a\boldsymbol{H}_s\boldsymbol{H}_a\boldsymbol{H}_s\boldsymbol{H}_s\boldsymbol{e}'$ and $\boldsymbol{p}(l)\boldsymbol{H}_a\boldsymbol{H}_a\boldsymbol{H}_s\boldsymbol{H}_s\boldsymbol{H}_s\boldsymbol{e}'$ 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.

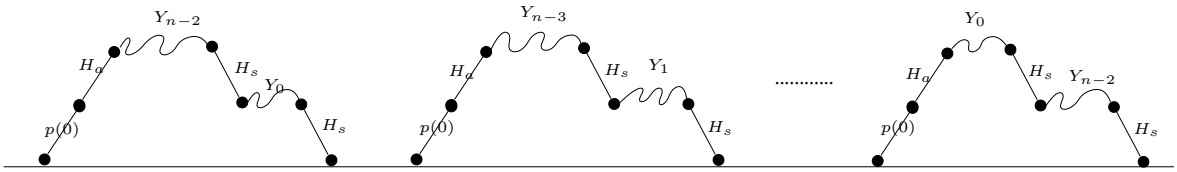

Figure 6: Paths serving exactly n customers during the first passage, $D_{l,l-1}$

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 $\boldsymbol{p}(l)\boldsymbol{H}_s\boldsymbol{e}'$. In all the other cases, at 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 $k_1, k_1 \geq 0$ customers are serviced before returning to level l for the first time followed by k_2 customers served before finally returning to level $l-1$ (See Fig. 6). In this respect, each of these sub-paths is similar to a Dyck path [38]. Thus exactly n customers can be served during this first passage ($D_{l,l-1}$) by serving $n-i$ ($k1 = n-i$) customers before returning to level l for the first time, followed by serving $i-1$ ($k_2 = i-1$) 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 these derivations are independent of the current state of the system (as long as the server is active). Let,

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

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 $p(l)$ at the transition from level $l-1$ to level l is given by,

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

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 $\widehat{\boldsymbol{e}'_a} = \boldsymbol{e}'_a \otimes \boldsymbol{I}_s$. We show the computation of the starting vector for a normal busy period in section (3.4) and the starting vector for a higher level first passage in Chapter(4).

3.3.1 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 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:

$$\begin{aligned}
z^1\boldsymbol{Y}_0 &= \boldsymbol{I}z^1 \\
z^2\boldsymbol{Y}_1 &= (\boldsymbol{H}_a\boldsymbol{Y}_0z^1\boldsymbol{H}_s\boldsymbol{Y}_0z^1) \\
z^3\boldsymbol{Y}_2 &= (\boldsymbol{H}_a\boldsymbol{Y}_1z^2\boldsymbol{H}_s\boldsymbol{Y}_0z^1 + \boldsymbol{H}_a\boldsymbol{Y}_0z^1\boldsymbol{H}_s\boldsymbol{Y}_1z^2) \\
\vdots &= \quad \vdots \\
z^{n+1}\boldsymbol{Y}_n &= (\boldsymbol{H}_a[\boldsymbol{Y}_{n-1}z^n\boldsymbol{H}_s\boldsymbol{Y}_0z^1 + \boldsymbol{Y}_{n-2}z^{n-1}\boldsymbol{H}_s\boldsymbol{Y}_1z^2 + \ldots + \boldsymbol{Y}_0z^1\boldsymbol{H}_s\boldsymbol{Y}_{n-1}z^n]) \\
\hline
\boldsymbol{Y}(z) &= z\boldsymbol{I} + \boldsymbol{H}_a(\boldsymbol{Y}_0z^1 + \boldsymbol{Y}_1z^2 + \boldsymbol{Y}_2z^3 + \ldots)\boldsymbol{H}_s(\boldsymbol{Y}_0z^1 + \boldsymbol{Y}_1z^2 + \boldsymbol{Y}_2z^3 + \ldots)
\end{aligned}$$

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

$$\boldsymbol{Y}(z) = z\boldsymbol{I} + \boldsymbol{H}_a\boldsymbol{Y}(z)\boldsymbol{H}_s\boldsymbol{Y}(z). \tag{3.3.1}$$

This matrix quadratic form for $\boldsymbol{Y}(z)$ (equation (3.3.1)) is closely related to the common matrix quadratic equation for the matrix $\boldsymbol{G}$ that occurs in literature [33], [35]. 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 (3.3.1),

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

and evaluating at z=1, gives

$$\boldsymbol{Y}'(1) = \boldsymbol{I} + \boldsymbol{H}_a\boldsymbol{Y}'(1)\boldsymbol{H}_s\boldsymbol{Y}(1) + \boldsymbol{H}_a\boldsymbol{Y}(1)\boldsymbol{H}_s\boldsymbol{Y}'(1), \tag{3.3.2}$$

Here $\boldsymbol{Y}(1)$ 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 (3.3.2) or as a limiting process. The mean number served during this conditional first passage is given by

$$\text{E}[N_{l,l-1}] = \boldsymbol{p}(l)\boldsymbol{Y}'(1)\boldsymbol{H}_s\boldsymbol{e}'. \tag{3.3.3}$$

Similarly, the second moment is computed as,

$$\text{E}[N_{l,l-1}^2] = \boldsymbol{p}(l)\boldsymbol{Y}''(1)\boldsymbol{H}_s\boldsymbol{e}' + \boldsymbol{p}(l)\boldsymbol{Y}'(1)\boldsymbol{H}_s\boldsymbol{e}'. \tag{3.3.4}$$

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

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

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. The matrix $\boldsymbol{Y}$ can be obtained by iteration on the z-transform equations

using $O(m^3)$ computations per iteration. Also the space complexity for computing $\boldsymbol{Y}$ is of order $O(m^3 n)$.

3.4 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

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

i.e., $\boldsymbol{p}_{bp}$ is the normalized left eigenvector corresponding to an eigenvalue of 1 for the matrix $\boldsymbol{Y}\boldsymbol{H}_s\boldsymbol{V}_a\boldsymbol{L}_a$. 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 $\boldsymbol{V}_a\boldsymbol{L}_a$, 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,

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

and mean number of customers served during a busy period is

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

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

Algorithm 1 To compute 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, $\boldsymbol{Y}^{(0)} = \boldsymbol{I}$

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

Alternately, $\boldsymbol{Y}$ can be computed as a limiting process from the summation of individual $\boldsymbol{Y}'_n s$.

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,

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

3.5 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.

3.5.1 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 [40]

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

where the combinatorial multiplier is the n^{th} Catalan number.

The mean number served and the variance for number served in this *M/M/1* system busy period are given by

$$E(N) = \frac{1}{1-\rho}$$
$$\text{and } Var(N) = \frac{\rho(1+\rho)}{(1-\rho)^3}.$$

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.

A closed form explicit result is known when the service distribution is deterministic, an *M/D/1* system [10]. 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

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

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.7. 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 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. Please note that even with the approximation to the deterministic distribution, the results are very

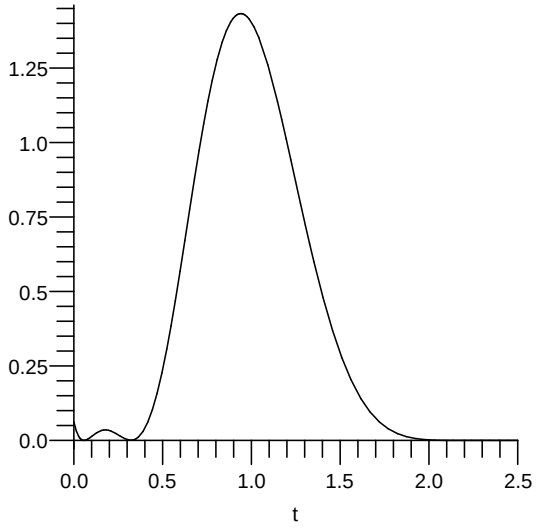

Figure 7: ME Density that touches the x-axis multiple times

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

close to the known Borel distribution.

3.5.2 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

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

and the service process is represented by the rate matrices

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

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

3.5.3 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

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

where $p_1 = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{c^2-1}{c^2+1}}$. This process is uncorrelated. In order to construct correlated processes with geometrically decaying covariances that share the same marginals, we use the approach presented in [34]. Define $\boldsymbol{L}^{(\gamma)}$ for $-1 < \gamma < 1$ as

$$\boldsymbol{L}^{\gamma} = (1-\gamma)(\boldsymbol{B}\boldsymbol{e}'\boldsymbol{p} - \boldsymbol{B}) + \boldsymbol{B}. \tag{3.5.1}$$

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

3.5.3.1 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.

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

3.5.3.2 MEP/M/1 System

As can be seen in Fig. 8, 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 number of busy periods serving one customer is decreasing.

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. 9). When the c^2 is increasing for the non-correlated case, the number of busy periods where fewer than five customers are 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

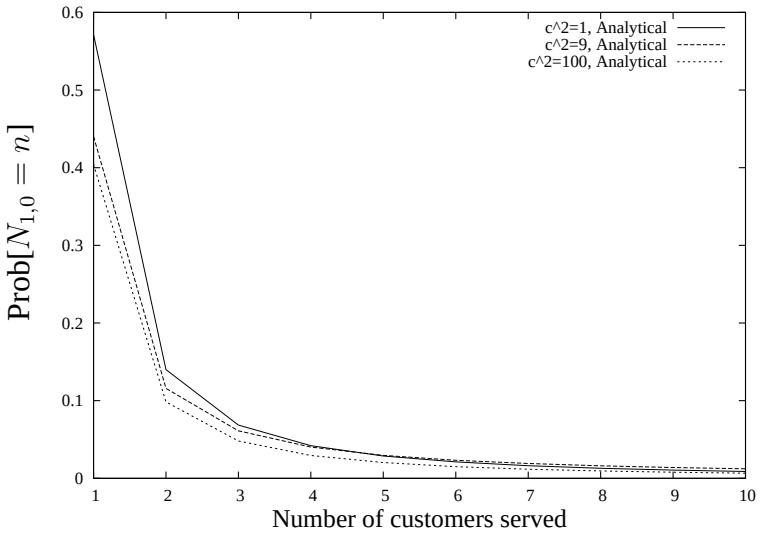

Figure 8: *MEP/M/1*: Effect of increasing c^2 in uncorrelated case

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. Also notice from Fig. 9 that in all the three cases for various c^2 values, the probabilities for one customer served in correlated case tend to be segregated and very different than the probabilities in the uncorrelated case.

3.5.4 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

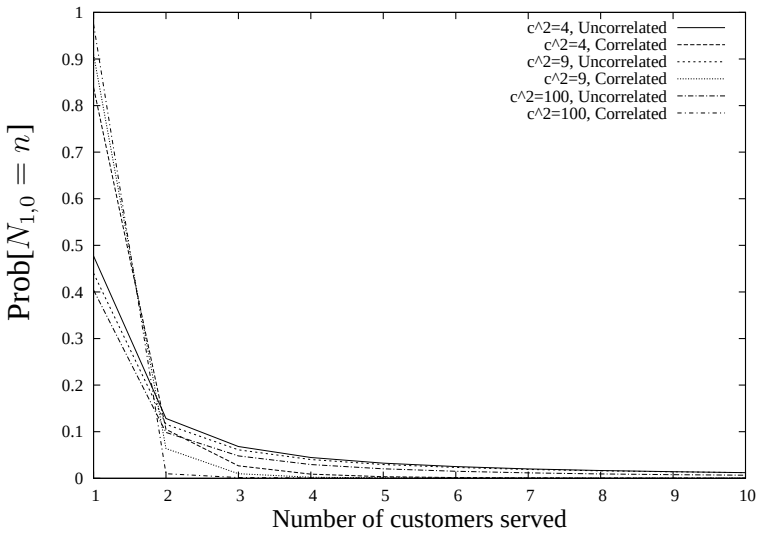

Figure 9: *MEP/M/1*: Effect of correlation on Prob[$N_b = n$]

and γ of 0.99 represents a system where the arrivals and service demands are both very erratic and correlated (bursty). Fig. 10 represents this effect.

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

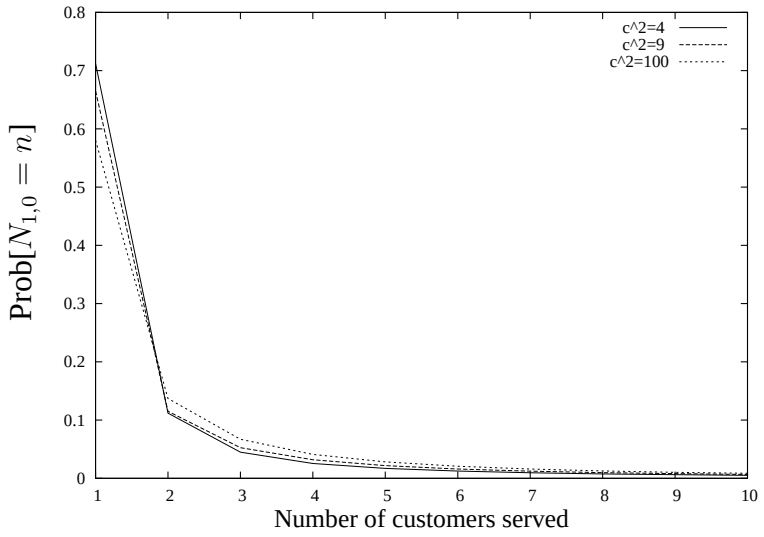

Figure 10: *MEP/MEP/1*: Effect of increasing c^2

3.5.4.1 Effect of Third Moment on MEP(r_1, r_2, r_3, γ)/M/1 Queue Busy Period

Consider a queue where the marginals of the arrival process are characterized by the first three reduced moments and the correlation decay parameter γ, of the arrival processes. We use this characterization so that the impact of the third-moment on the expected length of the busy periods can be studied. Such an arrival distribution can be represented in LAQT with the moment canonical form [44]

$$\boldsymbol{p}_a = \begin{bmatrix} 1 & 0 \end{bmatrix}, \ \boldsymbol{B}_a^{-1} = \begin{bmatrix} r_1 & r_1 \\ \frac{r_2 - r_1^2}{r_1} & \frac{r_3 - 2r_1 r_2 + r_1^3}{r_2 - r_1^2} \end{bmatrix}, \ \boldsymbol{e}'_a = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{3.5.2}$$

The bounds on the first 3 moments are given by Table 5 which is reproduced from [20].

The bounds on the value of the correlation parameter γ for fixed first three moments can be found in [20]. Let the mean service rate be defined by μ. In this case we assume the first reduced moment of the arrival process is normalized to 1 and $r2$ is set to 3. Plots. 11, 12 and 13 show the effect of the third moment on the mean length of busy

Table 5: Bounds for the First Three Normalized Moments of ME(2) Distributions

$r_1 > 0$	
hypoexponential	hyperexponential
$\frac{3}{4}r_1^2 \leq r_2 < r_1^2$ $(\Rightarrow h_2 < 0)$	$r_1^2 < r_2$ $(\Leftrightarrow 0 < h_2)$
$r_1(2h_2 + r_2) + 2(-h_2)^{\frac{3}{2}} \leq r_3$ $r_3 \leq r_1(h_2 + r_2)$	$\frac{r_2^2}{r_1} < r_3$ $(\Leftrightarrow 0 < h_3)$

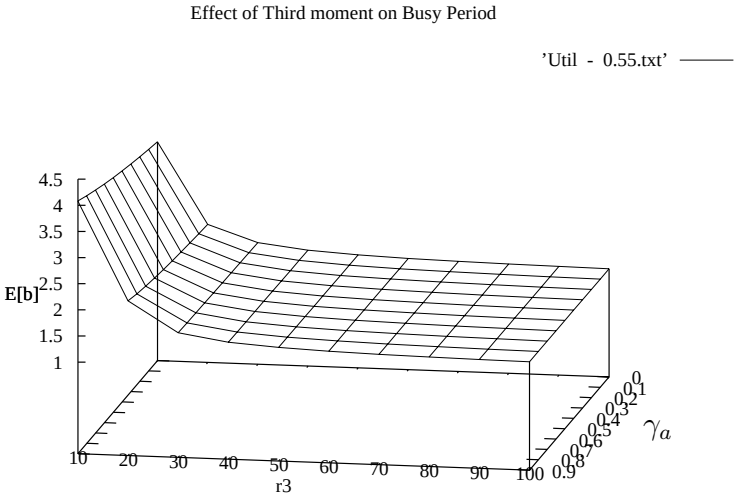

Figure 11: *MEP(r1,r2,r3,γ_a)/MEP/1*: Effect of Third moment, Util:0.55

periods as the utilizations and the correlation decay parameter γ are varied. A noticeable observation is the effect of the third moment at γ_a above 0.7. As the third moment decreases from 100 to 10, the mean busy period length increases until $r3$ reaches a certain critical value and decreases for further decrease in $r3$. This effect tends to be present at all higher utilizations and the critical $r3$ value tends to move higher as utilization increases.

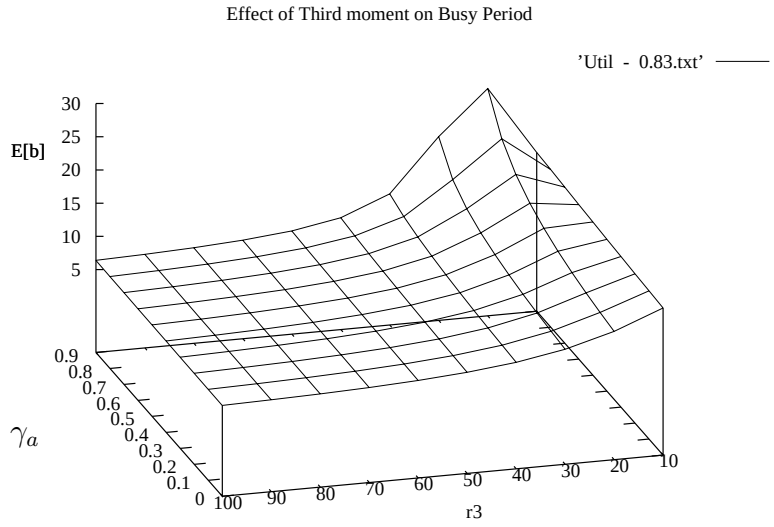

Figure 12: *MEP(r1,r2,r3,γ_a)/MEP/1*: Effect of Third moment, Util:0.83

3.6 Summary

In this chapter 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.

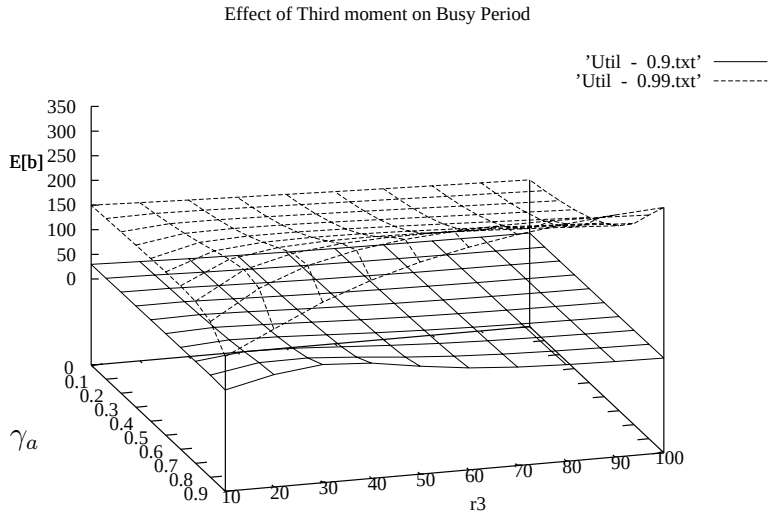

Figure 13: *MEP(r1,r2,r3,γ_a)/MEP/1*: Effect of Third moment, Util:0.9, 0.99

CHAPTER 4

BUSY PERIOD LENGTH AND HIGHER LEVEL FIRST PASSAGES

4.1 Introduction

In this chapter we first characterize the conditional min of two matrix exponential processes as a matrix exponential process and use that representation to construct the distribution functions and Laplace transforms for the time it takes to traverse any given sample path. We use these individual sample path length representations to derive the Laplace transform for the entire busy period length and derive expressions to compute the mean busy period length. In the later half of this chapter, we study how the correlations in arrival and service processes effect the mean first passage time when we now consider a generic first passage from level l to a level $l-1$ for various values of l (as opposed to the transition from level 1 to level 0). We then compute the probabilities that the sample paths are of height greater than a given threshold during a given first passage and also compute and compare the moments for number served and the mean time for the first passage for various levels against the same performance metrics for a normal busy period for various arrival and service processes.

4.2 Conditional Density for the *min(A,S)* Process

Consider two contesting processes, A and S (representing Arrivals and Service completions), both represented by the corrosponding matrix exponential notations $< \boldsymbol{p}_a, \boldsymbol{B}_a, \boldsymbol{e}_a >$

and $< \boldsymbol{p}_s, \boldsymbol{B}_s, \boldsymbol{e}_s >$. Then the conditional density for the min process given that the arrival process occurs before the service process is,

$$\begin{aligned}
\text{Pr}[\ min(A,S) = t \mid A < S\ ] &= \frac{\text{Pr}[\ min(A,S) = t \text{ and } A < S\ ]}{\text{Pr}[\ A < S\ ]} \\
&= \frac{\boldsymbol{p}_a \exp(-\boldsymbol{B}_a t)\boldsymbol{L}_a \boldsymbol{e}_a \boldsymbol{p}_s \exp(-\boldsymbol{B}_s t)\boldsymbol{e}_s}{\boldsymbol{p}_a \widehat{\boldsymbol{p}_s}(\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})^{-1}\widehat{\boldsymbol{L}_a}\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s} \\
&= \boldsymbol{p}_a \widehat{\boldsymbol{p}_s} \exp(-(\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})t)\frac{\widehat{\boldsymbol{L}_a}}{\boldsymbol{p}_a \widehat{\boldsymbol{p}_s}(\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})^{-1}\widehat{\boldsymbol{L}_a}\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s}\widehat{\boldsymbol{e}_a}e_s
\end{aligned}$$

The expression in the denominator is the probability that an Arrival event occurs before a Service completion and hence is a scalar (less than 1), say α. The effect of conditioning on the fact that arrival occurs before service event, is that the Arrival processes gets effectively accelerated (from $\boldsymbol{L}_a$ to $\frac{\boldsymbol{L}_a}{\alpha}$). This in essence is the effect of knowing that additional piece of information. If we consider this as a new matrix exponential process, we no longer have the usual equality $\boldsymbol{Be} = \boldsymbol{Le}$ since $(\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s \neq \frac{\boldsymbol{L}_a}{\alpha}\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s$. But nonetheless this is a valid matrix exponential density. It can easily be seen that the integral of the above conditional density from 0 to ∞ equals 1.

4.3 ME Representation for The Length of a Sample Path

Consider a sample path during a busy period where immediately after the start of a busy period, we have an arrival followed by a departure event. The length of this sample path is the convolution of two stochastic processes, representing the occurence of

an arrival event followed by a departure event (represented as "AD").

$$\Pr[\text{“}AD\text{”} = t_1 dt_1] = \int_{t_1=0}^{t} \boldsymbol{p}_{bp} \exp(-(\widehat{\boldsymbol{B}_a + \boldsymbol{B}_s})t_1)\frac{\boldsymbol{L}_a}{\alpha_1} \exp(-(\widehat{\boldsymbol{B}_a + \boldsymbol{B}_s})(t-t_1))\frac{\boldsymbol{L}_s}{\alpha_2} dt_1 \tag{4.3.1}$$

where $\alpha_1 = \boldsymbol{p}_{bp}(\widehat{\boldsymbol{B}_a + \boldsymbol{B}_s})^{-1}\widehat{\boldsymbol{L}_a}\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s = \boldsymbol{p}_{bp}\boldsymbol{H}_a\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s$ and $\alpha_2 = \dfrac{\boldsymbol{p}_{bp}\boldsymbol{H}_a}{\boldsymbol{p}_{bp}\boldsymbol{H}_a\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s}\boldsymbol{H}_s\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s$

The above density of sample path ("AD") can be written in an matrix exponential form using the following $< \boldsymbol{p}_{pp}, \boldsymbol{B}_{pp}, \boldsymbol{L}_{pp}, \boldsymbol{e}_{pp} >$ where,

$< \boldsymbol{p}_{pp}, \boldsymbol{B}_{pp}, \boldsymbol{L}_{pp}, \boldsymbol{e}_{pp} >$ where,

$$\boldsymbol{p}_{pp} = \begin{bmatrix} \boldsymbol{p}_{bp} & 0 \end{bmatrix}$$

$$\boldsymbol{B}_{pp} = \begin{bmatrix} (\widehat{\boldsymbol{B}_a + \boldsymbol{B}_s}) & \frac{-\boldsymbol{L}_a}{\boldsymbol{p}_{bp}\boldsymbol{H}_a\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s} \\ 0 & (\widehat{\boldsymbol{B}_a + \boldsymbol{B}_s}) \end{bmatrix}, \boldsymbol{L}_{pp} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{\boldsymbol{L}_s}{\frac{\boldsymbol{p}_{bp}\boldsymbol{H}_a}{\boldsymbol{p}_{bp}\boldsymbol{H}_a\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s}\boldsymbol{H}_s\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s} \end{bmatrix}, \boldsymbol{e}_{pp} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \widehat{\boldsymbol{e}_a}\boldsymbol{e}_s \end{bmatrix}.$$

The equivalence of the above two forms can be verified by computing the Laplace transforms of the above two representations. Let $F_1^*(s)$ and $F_2^*(s)$ represent the Laplace transforms of the convolution form and the matrix exponential form respectively. The Laplace transform of the matrix exponential representation, is given by

$$F_2^*(s) = \boldsymbol{p}_{pp}(\boldsymbol{B}_{pp} + s\boldsymbol{I})^{-1}\boldsymbol{L}_{pp}\boldsymbol{e}'_{pp}. \tag{4.3.2}$$

Inverse of a block matrix can be written as,

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & S_A^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix}$$

where S_A, the $Schur$ complement of A is given by $S_A = D - CA^{-1}B$

$$\begin{aligned}
(\boldsymbol{B}_{pp} + s\boldsymbol{I})^{-1} &= \begin{bmatrix} I & (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} \frac{\boldsymbol{L}_a}{\alpha_1} \\ 0 & I \end{bmatrix} \cdot \\
&\quad \cdot \begin{bmatrix} (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} & 0 \\ 0 & (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \\
&= \begin{bmatrix} (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} & (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} \frac{\boldsymbol{L}_a}{\alpha_1} (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} \\ 0 & (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} \end{bmatrix}.
\end{aligned}$$

Hence,

$$\begin{aligned}
F_2^*(s) &= \begin{bmatrix} \boldsymbol{p}_{bp} (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} & \boldsymbol{p}_{bp} (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} \frac{\boldsymbol{L}_a}{\alpha_1} (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} \end{bmatrix} \cdot \\
&\quad \cdot \begin{bmatrix} 0 & 0 \\ 0 & \frac{\boldsymbol{L}_s}{\alpha_2} \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \widehat{\boldsymbol{e}_a \boldsymbol{e}_s} \end{bmatrix} \\
&= \begin{bmatrix} 0 & \boldsymbol{p}_{bp} (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} \frac{\boldsymbol{L}_a}{\alpha_1} (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} \frac{\boldsymbol{L}_s}{\alpha_2} \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \widehat{\boldsymbol{e}_a \boldsymbol{e}_s} \end{bmatrix} \\
&= \boldsymbol{p}_{bp} (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} \frac{\boldsymbol{L}_a}{\alpha_1} (\boldsymbol{B}_a \widehat{+ \boldsymbol{B}_s} + s\boldsymbol{I})^{-1} \frac{\boldsymbol{L}_s}{\alpha_2} \widehat{\boldsymbol{e}_a \boldsymbol{e}_s} = F_1^*(s)
\end{aligned}$$

Hence both the representations are equivalent, and the shown matrix exponential representation corresponds to the convolved sample path density. In the next section we show the matrix exponential form for the sample path "ADD" (Arrival followed by two consecutive departures) and compute the conditional Laplace transform which will then be used to derive the moments of the busy period.

4.4 Conditional Laplace Transform of a Sample Path During a Busy Period

Let us consider all possible paths during which exactly two customers are served during a busy period. As the busy period starts with the first arriving customer, there is only one such path possible, another arrival followed by two consecutive departures, i.e., "A-D-D". The probability of this path being taken is $\boldsymbol{p}_{bp}\boldsymbol{H}_a\boldsymbol{H}_s\boldsymbol{H}_s\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s$. The matrix exponential representation for the length of this busy period is given by $< \boldsymbol{p}_{pp}, \boldsymbol{B}_{pp}, \boldsymbol{L}_{pp}, \boldsymbol{e}_{pp} >$ where,

$$\boldsymbol{p}_{pp} = \begin{bmatrix} \boldsymbol{p}_{bp} & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{B}_{pp} = \begin{bmatrix} (\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s}) & \dfrac{-\widehat{\boldsymbol{L}_a}}{\boldsymbol{p}_{bp}\boldsymbol{H}_a\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s} & 0 \\ 0 & (\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s}) & \dfrac{-\widehat{\boldsymbol{L}_s}}{\dfrac{\boldsymbol{p}_{bp}\boldsymbol{H}_a}{\boldsymbol{p}_{bp}\boldsymbol{H}_a\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s}\boldsymbol{H}_s\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s} \\ 0 & 0 & (\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s}) \end{bmatrix}$$

$$\boldsymbol{L}_{pp} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \dfrac{\widehat{\boldsymbol{L}_s}}{\dfrac{\boldsymbol{p}_{bp}\boldsymbol{H}_a\boldsymbol{H}_s}{\boldsymbol{p}_{bp}\boldsymbol{H}_a\boldsymbol{H}_s\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s}\boldsymbol{H}_s\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s} \end{bmatrix}, \boldsymbol{e}_{pp} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \widehat{\boldsymbol{e}_a}\boldsymbol{e}_s \end{bmatrix}$$

The Laplace transform for this path is given by

$$F^*(s) = \boldsymbol{p}_{bp}\frac{\boldsymbol{F}_a^*(s)}{\boldsymbol{p}_{bp}\boldsymbol{H}_a\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s}\frac{\boldsymbol{F}_s^*(s)}{\dfrac{\boldsymbol{p}_{bp}\boldsymbol{H}_a\boldsymbol{H}_s\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s}{\boldsymbol{p}_{bp}\boldsymbol{H}_a\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s}}\frac{\boldsymbol{F}_s^*(s)}{\dfrac{\boldsymbol{p}_{bp}\boldsymbol{H}_a\boldsymbol{H}_s\boldsymbol{H}_s\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s}{\boldsymbol{p}_{bp}\boldsymbol{H}_a\boldsymbol{H}_s\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s}}\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s$$

where $\boldsymbol{F}_a^*(s) = (\boldsymbol{B}_a + \widehat{\boldsymbol{B}_s} + s\boldsymbol{I})^{-1}\boldsymbol{L}_a$ and $\boldsymbol{F}_s^*(s) = (\boldsymbol{B}_a + \widehat{\boldsymbol{B}_s} + s\boldsymbol{I})^{-1}\boldsymbol{L}_s$ are matrices.

The conditional Laplace transform for this path, conditioned by the path "ADD" being taken (which occurs with a probability $\boldsymbol{p}_{bp}\boldsymbol{H}_a\boldsymbol{H}_s\boldsymbol{H}_s\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s$), is given by,

$$F^*(s \mid \text{“}ADD\text{”}) \;=\; \boldsymbol{p}_{bp}\boldsymbol{F}_a^*(s)\boldsymbol{F}_s^*(s)\boldsymbol{F}_s^*(s)\widehat{\boldsymbol{e}_a}\boldsymbol{e}_s$$

4.5 Mean Length of a Busy Period

Noting the similarity between the above formulation for the conditional Laplace transform and the probability of a certain path being taken during a busy period, we can derive the joint transform equation for the number of customers served during a busy period and its length. We have,

$$\begin{aligned}
\boldsymbol{F}_0^*(s) &= \boldsymbol{I} \\
\boldsymbol{F}_1^*(s) &= \boldsymbol{F}_a^*(s)\boldsymbol{F}_0^*(s)\boldsymbol{F}_s^*(s)\boldsymbol{F}_0^*(s) \\
\boldsymbol{F}_2^*(s) &= \boldsymbol{F}_a^*(s)\boldsymbol{F}_1^*(s)\boldsymbol{F}_s^*(s)\boldsymbol{F}_0^*(s) + \boldsymbol{F}_a^*(s)\boldsymbol{F}_0^*(s)\boldsymbol{F}_s^*(s)\boldsymbol{F}_1^*(s) \\
\vdots \quad &\quad \vdots \\
\boldsymbol{F}_n^*(s) &= \boldsymbol{F}_a^*(s).\left[\boldsymbol{F}_{n-1}^*(s)\boldsymbol{F}_s^*(s)\boldsymbol{F}_0^*(s) + \boldsymbol{F}_{n-2}^*(s)\boldsymbol{F}_s^*(s)\boldsymbol{F}_1^*(s) + \ldots + \boldsymbol{F}_0^*(s)\boldsymbol{F}_s^*(s)\boldsymbol{F}_{n-1}^*(s)\right]
\end{aligned}$$

The Z-transform of the above set of equations gives the two-dimensional transform for number served during the busy period and the length of the busy period and is given by the matrix functional equation

$$\boldsymbol{F}^*(s,z) = z\,\boldsymbol{I} + \boldsymbol{F}_a^*(s)\boldsymbol{F}^*(s,z)\boldsymbol{F}_s^*(s)\boldsymbol{F}^*(s,z). \tag{4.5.1}$$

Evaluating the joint transform at $z = 1$ and including the final departure gives the matrix required to compute the Laplace transform for the busy period duration as

$$\boldsymbol{F}^*(s)\boldsymbol{F}_s^*(s) = \boldsymbol{F}_s^*(s) + \boldsymbol{F}_a^*(s)\boldsymbol{F}^*(s)\boldsymbol{F}_s^*(s)\boldsymbol{F}^*(s)\boldsymbol{F}_s^*(s). \tag{4.5.2}$$

Substituting $\boldsymbol{F}^*(s)\boldsymbol{F}_s^*(s) = \boldsymbol{F}_T^*(s)$ and $\boldsymbol{F}^*(s)\boldsymbol{F}_s^*(0) = \boldsymbol{F}_T^*$, taking the derivative with respect to s and evaluating at $s = 0$, we get,

$$(\boldsymbol{F}_T^*(0))' = \left(\boldsymbol{F}_s^*(s)' + \boldsymbol{F}_a^*(s)'\boldsymbol{F}_T^*(s)^2 + \boldsymbol{F}_a^*(s)\boldsymbol{F}_T^*(s)'\boldsymbol{F}_T^*(s) \; + \boldsymbol{F}_a^*(s)\boldsymbol{F}_T^*(s)\boldsymbol{F}_T^*(s)'\right)\Big|_{s=0}. \tag{4.5.3}$$

Using $\boldsymbol{F}_a^*(0) = \boldsymbol{H}_a$, $\boldsymbol{F}_s^*(0) = \boldsymbol{H}_s$, $\boldsymbol{F}_a^*(0)' = -\boldsymbol{D}\boldsymbol{H}_a$ and $\boldsymbol{F}_s^*(0)' = -\boldsymbol{D}\boldsymbol{H}_s$, $\boldsymbol{F}_T^*(0)'$ is obtained by iteration on

$$\boldsymbol{F}_T^*(0)' = -\boldsymbol{D}\boldsymbol{H}_s - \boldsymbol{D}\boldsymbol{H}_a\, \boldsymbol{F}_T^*(0)^2 + \boldsymbol{H}_a\, \boldsymbol{F}_T^*(s)'\, \boldsymbol{F}_T^*(0) + \boldsymbol{H}_a\boldsymbol{F}_T^*(0)\, \boldsymbol{F}_T^*(0)'$$

where, $\boldsymbol{D} = (\widehat{\boldsymbol{B}_a} + \widehat{\boldsymbol{B}_s})^{-1}$ and $\boldsymbol{F}_T^*(0) = \boldsymbol{Y}\boldsymbol{H}_s$

Let τ_b represent the r.v. for the length of a busy period, then

$$\begin{aligned} \mathrm{E}[\tau_b] &= -\frac{d}{ds}\left(\boldsymbol{p}_{bp}\boldsymbol{F}^*(s)\boldsymbol{F}_s^*(s)\boldsymbol{e}'\right)\Big|_{s=0} \\ &= -\frac{d}{ds}\left(\boldsymbol{p}_{bp}\boldsymbol{F}_T^*(s)\boldsymbol{e}'\right)\Big|_{s=0} = -\boldsymbol{p}_{bp}\, \boldsymbol{F}_T^*(0)'\, \boldsymbol{e}' \end{aligned}$$

4.5.1 Simplifications in an M/M/1 Case

In this case

$$\boldsymbol{F}_a^*(s) = \frac{\lambda}{\lambda+\mu+s} \quad \text{and} \quad \boldsymbol{F}_s^*(s) = \frac{\mu}{\lambda+\mu+s}$$

Hence the Laplace transform for the length of busy period Eq. (4.5.2), simplifies to

$$\boldsymbol{F}^*(s) = 1 + \frac{\lambda}{\lambda+\mu+s}\boldsymbol{F}^*(s)\frac{\mu}{\lambda+\mu+s}\boldsymbol{F}^*(s)$$

Therefore,

$$\lambda\mu\boldsymbol{F}^*(s)^2 - (\lambda+\mu+s)^2\boldsymbol{F}^*(s) - (\lambda+\mu+s)^2 = 0$$

Solving for $\boldsymbol{F}^{\,*}(s)$ and selecting the appropriate root using the condition that $\boldsymbol{F}^{\,*}(s)\Big|_{s=0} = 1$ and post-multiplying with $\boldsymbol{F}_s^*(s)$ gives the well know transform for the $M/M/1$ busy period,

$$\boldsymbol{F}^*(s) = \frac{(\lambda+\mu+s) - \sqrt{(\lambda+\mu+s)^2 - 4\lambda\mu}}{2\lambda}. \tag{4.5.4}$$

Also note that the Laplace transform for an *M/M/1* case can be written directly as a summation of conditional Laplace transforms, conditioned on the number of customers served during a busy period, i.e.,

$$\boldsymbol{F}^*(s) = \sum_{n=0}^{\infty} \frac{1}{n+1}\binom{2n}{n}\frac{\lambda^{2n}\mu^{2n+2}}{(\lambda+\mu)^{4n+2}}\frac{(\lambda+\mu)^{2n+1}}{(\lambda+\mu+s)^{2n+1}}$$

4.6 Mean First Passage Time for Different Threshold Levels

Consider a system where the system just transitioned from level $n-1$ to level n and we are interested in the mean first passage time to reach back to level $n-1$. In this section we show the effect of correlations in arrival and service processes on the mean first passage time to go from back to this threshold level $n-1$ for different threshold levels.

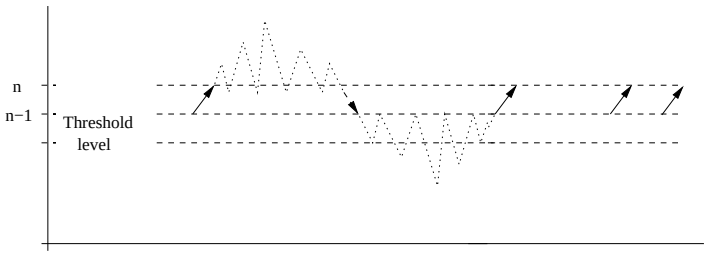

Figure 14: Higher Level First Passages

This first passage time differs from a normal busy period only in the way the process starts. Once that starting vector for this "Elevated Busy Period" is known, then the rest of the analysis is similar to a normal busy period. To compute this starting vector, we consider all possible paths that result in such a transition and then compute the invariance vector. If we let $P_{n-1,n}$ denote the probability matrix that represents all the possible paths that lead the queue from level $n-1$ to level n for the first time, using a common first passage argument [14], we can compute $P_{n-1,n}$ using the following set of recurrence relations.

$$\begin{aligned} \boldsymbol{P}_{0,1} &= \widehat{\boldsymbol{V}}_a \widehat{\boldsymbol{L}}_a \\ \boldsymbol{P}_{1,2} &= (\boldsymbol{I} - \boldsymbol{H}_s \boldsymbol{P}_{0,1})^{-1} \boldsymbol{H}_a \end{aligned}$$

$$\vdots \qquad\qquad \vdots$$

$$\boldsymbol{P}_{n-1,n} = (\boldsymbol{I} - \boldsymbol{H}_s \boldsymbol{P}_{n-2,n-1})^{-1} \boldsymbol{H}_a$$

If we cross a given threshold $(n-1)$ and start the process in the n^{th} state with a starting vector $p_{n-1,n}$, then $\boldsymbol{Y}\boldsymbol{H}_s\boldsymbol{P}_{n-1,n}$ represents all the possible ways in which we will cross the same threshold for the first time, after dropping below the threshold. Hence the required starting vector for crossing a threshold of $n-1$ reaching n is computed as the invariance vector for the matrix $\boldsymbol{Y}\boldsymbol{H}_s\boldsymbol{P}_{n-1,n}$.

To show the effect of increase in threshold level on this first passage time, and to study the effect of correlation in the arrival process and service process on the mean of this first passage time, we use the general setup for the *MEP/MEP/1* system. For a c^2 of 25 for the arrival process and 9 for the service process, at a Utilization of 0.75 and correlation decay parameter of 0.7 (where applicable), we plot this mean first passage time as a function of the threshold level in Fig.15.

When both the arrival and service processes have a correlation decay parameter of 0.7, the starting vector for the correlated G/G/1 case for a transition from state 1 to state 0 is $\boldsymbol{p}_{0,1} = (0.853, 0.016, 0.125, 0.005)$ and the mean length of the first passage time from level 1 to level 0 is 3.92. As we increase the threshold level, the mean first passage time from level n to $n-1$ increases and converges. In this example, the starting vector converges to $\boldsymbol{p}_{n,n-1} = (0.45, 0.535, 0.006, 0.008)$ and the mean first passage time converges to 36.46 which is quite higher than the mean first passage time from level 1 to level 0. This increase in mean busy period as threshold level increases can be understood

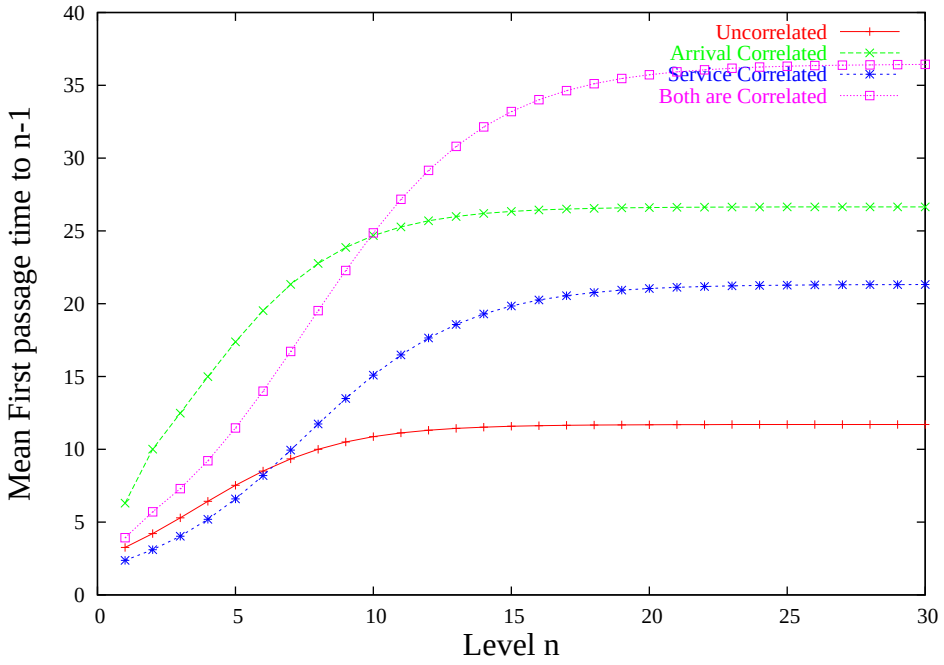

Figure 15: Mean Lenght of First Passage From Level n to $n-1$

by noticing that for the queue to cross the higher threshold the internal states of the Arrival and Service processes should already be such that either the Arrival process is in its faster state or the Service process is in it slower state or both; and due to the correlation in these processes, the arrival and/or the service processes tend to remain in those same states for a while, which means that the transient queue length at a higher threshold is bound to increase more than in the case when the the threshold was lower. After a certain height the mean busy period converges because once the queue reaches a certain height, the probability of the Arrival process being in the slower state or the Service process being in the faster state is so low that a further increase in threshold does not effect the starting phases for the arrival and service processes.

4.7 Paths That Cross a Given Level During a Busy Period

In this section we compute the probabilities of going above a height h during the first passage from level n to level $n-1$. We then show the effect of correlations in the arrival and service processes and the effect of the starting level (n), on these probabilities.

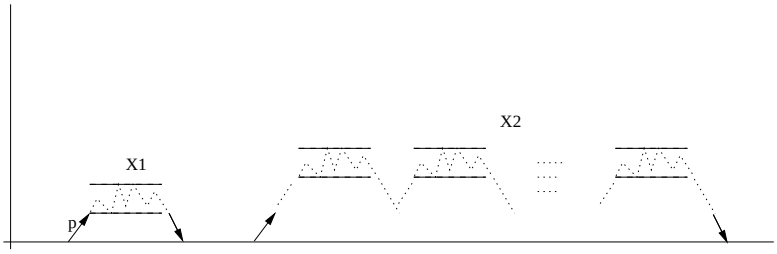

Figure 16: Paths within a channel of height h

Let $\boldsymbol{X}_h$ represent all the possible paths that start at a given level and end at the same level, never going below that level and are of height atmost h, i.e., all paths within a channel of height h. We now have the following set of recurrence relations for $\boldsymbol{X}'_h s$

$$\begin{aligned}
\boldsymbol{X}_0 &= \boldsymbol{I} \\
\boldsymbol{X}_1 &= (\boldsymbol{I} - \boldsymbol{H}_a \boldsymbol{X}_0 \boldsymbol{H}_s)^{-1} \\
\boldsymbol{X}_2 &= (\boldsymbol{I} - \boldsymbol{H}_a \boldsymbol{X}_1 \boldsymbol{H}_s)^{-1} \\
\vdots \quad &= \quad \vdots
\end{aligned}$$

Let M_n be the r.v for the density for Maximum height during a first passage from level n to level $n-1$. Hence,

$$\text{Prob}[M_n \leq h\ ] = \boldsymbol{p}\boldsymbol{X}_{h-1}\boldsymbol{H}_s\boldsymbol{e} \qquad\qquad h \geq 1. \tag{4.7.1}$$

The probability that during a busy period a given level is crossed is

$$\text{Prob}[M_n > h\ ] = 1 - \boldsymbol{p}\boldsymbol{X}_{h-1}\boldsymbol{H}_s\boldsymbol{e} \qquad\qquad h \geq 1 \tag{4.7.2}$$

The vector $\boldsymbol{p}$ for starting at different levels is computed using the approach presented in the previous section. The effect of both the correlations in the arrival and service processes as well as the effect of changing the base level on these probabilities is shown in Table. 6. The correlation decay parameter gamma is set to 0.7 and a squared coefficient of variation of 9 is used for both the Arrival and Service processes. Utilization is set to 0.75.

	$M/M/1$	$M/M/1$	$G/G/1$	$G/G/1$
	From 1-0	From 10-9	From 1-0	From 10-9
h	Prob. that Height greater than h			
1	0.428571	0.428571	0.369322	0.620431799
2	0.243243	0.243243	0.212039	0.501704526
3	0.154286	0.154286	0.140600	0.442551601
4	0.103713	0.103713	0.101558	0.406558265
⋮	⋮	⋮	⋮	⋮
10	0.014699	0.014699	0.039565	0.316721063
Mean length of the busy period?				
Mean length	0.799	0.799	2.07	16.32
Mean nos. served in a busy period?				
Mean nos.	4	4	10.38	70.81
csquare	5.25	5.25	61.2	10.8

Table 6: Paths of Height greater than h

Some very interesting numbers can be seen in the table above. For example, the probability that the queue grows above a height of 10 during a busy period for a *G/G/1* system for a transition from level 1 to level 0 is 0.03956 where as the same probability for

a transition from level 10 to level 9 is 0.3167 which is orders of magnitude higher. The effect of this can be clearly seen in both the mean length of the busy period which increases from 2.07 to 16.32 and the mean number served during a busy period which increases from 10.38 to 70.81. This is purely the effect of the increase in base level. Also note the increase in csquare for the the mean number served in a busy period increases from 5.25 in an *M/M/1* case to 61.2 in a *G/G/1* case; this is the sum effect of the correlations and variances in the arrival and service processes. Also there are some not so obvious numbers such as the decrease in csquare of a *G/G/1* system as the base level changes from 1 to 10; but perhaps the reason for this is that at a higher level there is a lower chance of having fewer number of customers served, i.e., the mean number served is high (70.81) with relatively small variance of 10.8 compared to a high variance of 61.2 in the case where the base level is 1.

CHAPTER 5

BUSY PERIOD ANALYSIS OF FINITE *QBD* PROCESSES

5.1 Introduction

In this section we present solutions for the number of customers served during a busy period and the length of a busy period for finite *MEP/MEP/1* system where either one or both of the arrival or service processes can be serially auto-correlated. We present numerical results and study how the moments and auto correlations in the arrival and service processes affect the busy period. This includes the probabilities of serving exactly n customers during a busy period and the moments of the length of the busy period for different allowable system (queue) sizes. The resulting algorithms use dynamic programming techniques and are easily implemented.

Consider a server in a certain facility that has finite resources (memory, disk space etc.) Most performance measure studies of interest like latencies, system times, waiting times etc. study the system from the perspective of an incoming customer with an objective of reducing the delays experienced by the customer as he progresses through the system. If one needs to take certain proactive measures, for example to avoid certain breakdowns, it is equally important to study the system from the service providers perspective. An understanding of how many customers are being served by the server on a continuous basis, i.e., between the servers idle times, is instrumental in devising proactive schemes to achieve optimal performance.

Due to the restrictions presented by the finite boundaries and the effect of the

boundary on the state transitions leading to the boundary, certain queueing studies, including the busy period analysis, are more intricate for the finite system as compared to their infinite counterparts.

5.2 Busy Period of a Finite $MEP/MEP/1$ Queue

Now consider a finite queueing system where the maximum system size (and hence the the highest level a sample path can take) is limited to q. Fig. 17 shows all

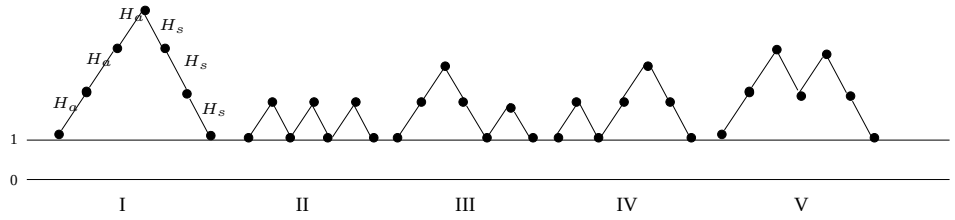

Figure 17: Paths with exactly three arrivals and three departures

possible paths wherein exactly three arrivals and three departures occur during a sample path, such that the sample path always stays above the starting level (level 1, in this case) and ends exactly at this level; i.e., all sample paths that begin at some level and end at the same level never taking any excursions below that level. In the case of an infinite queue there are five such possible paths (the count given by Catalan numbers).

We now require that the sample paths do not cross a given height representing a finite queue of size q. We allow the arrival process to be active when the queue is full, and arrivals to a full queue are lost and cleared. All possible sample paths that represent exactly three arrivals and three departures within a channel of width two are shown in Fig. 18. The loops at the tops of the sample paths represent arrivals that are being dropped when the queue is full. The number of paths is no longer given by the Catalan numbers.

Let $N^s_{1,0}$ represent the discrete random variable for the number of customers served during a busy period when the maximum allowable system size is s, and let $d^s_{n,1}$ represent the probability that $N^s_{1,0} = n$.

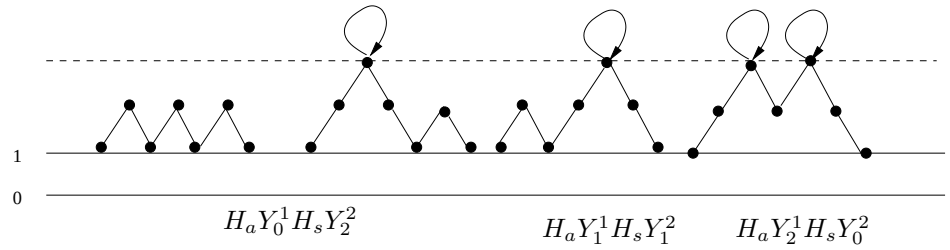

Figure 18: Paths with exactly three arrivals and three departures within a channel of width two

When computing $d^s_{n,1}$, all possible paths must be considered that fit within a channel of height q and that have exactly $n-1$ service completions. Note that the channel being considered starts after the arrival of the first customer, hence the $n-1$ arrivals and service completions in the channel plus one additional service that completes the transition of the sample paths from level one to level zero at the end, completes the corresponding busy period, resulting in n customers being served.

Let $\boldsymbol{Y}^j_i$ correspond to all possible paths of height less than or equal to j with exactly i customers being served and thus exactly i down transitions. Now we have,

$$d^s_{n,1} = \boldsymbol{p}(0)\boldsymbol{Y}^{s-1}_{n-1}\boldsymbol{H}_s\boldsymbol{e}$$

conditioned upon the process starting in the vector $\boldsymbol{p}(0)$. The actual starting vector for the busy period will be determined later.

We will now concentrate on deriving the recursive definitions for various sample paths that serve i customers within a channel of width j, $\boldsymbol{Y}^j_i$.

If the first arrival is immediately followed by a service completion we have exactly one customer served, the probability of this occurring is $\boldsymbol{p}(0)\boldsymbol{H}_s\boldsymbol{e}$. In all other cases, the second arrival occurring before the first departure. This also means that a transition from two customers in the system down to one customer in the system must occur at least once before the busy period ends. Let there be k service completions in the transition from level two, to level 1 for the first time without exceeding a height of $j-1$. Then there are $i-1-k$ services in the remainder of the busy period without exceeding a ceiling of j.

$\boldsymbol{Y}_i^j$'s are recursively defined as follows:

$$\begin{aligned}
&\boldsymbol{Y}_0^0 = (\boldsymbol{I}-\boldsymbol{H}_a)^{-1}, \quad \boldsymbol{Y}_0^1 = \boldsymbol{Y}_0^2 = \ldots = \boldsymbol{I}, \\
&\boldsymbol{Y}_1^1 = \boldsymbol{H}_a(\boldsymbol{I}-\boldsymbol{H}_a)^{-1}\boldsymbol{H}_s, \quad \boldsymbol{Y}_n^1 = (\boldsymbol{Y}_1^1)^n, \qquad n \geq 2 \\
&\qquad\qquad \boldsymbol{Y}_i^j = \sum_{k=0}^{i-1} \boldsymbol{H}_a \boldsymbol{Y}_k^{j-1} \boldsymbol{H}_s \boldsymbol{Y}_{i-1-k}^j, \qquad i \geq 2, 2 \leq j \leq i, \\
&\qquad\qquad\qquad\qquad \boldsymbol{Y}_i^j = \boldsymbol{Y}_i, \qquad j > i.
\end{aligned} \tag{5.2.1}$$

Notice that the general structure of the definition of $\boldsymbol{Y}_i^j$ still resembles the general Catalan recursion, $C_n = \sum_{i=0}^{n=1} C_i C_{n-i-1}$. In this respect, each of these sub-paths is similar to a Dyck path [38] starting from the starting point of the sub-path.

For different allowable heights, we have a complete set of $\boldsymbol{Y}_i^j$'s. The following matrix gives a better understanding of the relationship between all the different $\boldsymbol{Y}_i^j$'s and the $\boldsymbol{Y}_i$'s that we see in the case of infinite queues.

$$\begin{bmatrix}
\boldsymbol{Y}_0^0 & \boldsymbol{Y}_0 & \boldsymbol{Y}_0 & . & . & . & . & . \\
\boldsymbol{Y}_1^1 & \boldsymbol{Y}_1 & \boldsymbol{Y}_1 & . & . & . & . & . \\
\boldsymbol{Y}_2^1 & \boldsymbol{Y}_2^2 & \boldsymbol{Y}_2 & \boldsymbol{Y}_2 & . & . & . & . \\
\boldsymbol{Y}_3^1 & \boldsymbol{Y}_3^2 & \boldsymbol{Y}_3^3 & \boldsymbol{Y}_3 & \boldsymbol{Y}_3 & . & . & . \\
\boldsymbol{Y}_4^1 & \boldsymbol{Y}_4^2 & \boldsymbol{Y}_4^3 & \boldsymbol{Y}_4^4 & \boldsymbol{Y}_4 & \boldsymbol{Y}_4 & . & . \\
\boldsymbol{Y}_5^1 & \boldsymbol{Y}_5^2 & \boldsymbol{Y}_5^3 & \boldsymbol{Y}_5^4 & \boldsymbol{Y}_5^5 & \boldsymbol{Y}_5 & \boldsymbol{Y}_5 & . \\
. & . & . & . & . & . & . & .
\end{bmatrix}$$

Consider the fourth row from the matrix shown. The individual elements, $\boldsymbol{Y}_3^1, \boldsymbol{Y}_3^2, \boldsymbol{Y}_3^3, \ldots$ are used to compute the probabilities of serving exactly three customers during a busy period and a finite queue of size 1, 2, 3 or higher respectively. Note that if only three customers are served, the finiteness of the queue does not have any impact of queues of size four or more.

It should be noted from the recursive definitions for $\boldsymbol{Y}_i^j$'s and the general matrix structure for different $\boldsymbol{Y}_i^j$'s, that once the boundary equations for $\boldsymbol{Y}_0^j$'s have been defined and the first column of the matrix corresponding to $\boldsymbol{Y}_n^1$'s are defined, every other element of the matrix can be computed using a dynamic programming approach.

In the case of infinite queues, the number of possible paths in which n customers can be served is given by the $(n-1)^{st}$ Catalan number; we are dealing with $\boldsymbol{Y}_i$ without much regard for the ceiling, as the effective ceiling was at infinity, thus $\boldsymbol{Y}_i = \boldsymbol{Y}_i^\infty$. For finite queues, we now have a full gamut of $\boldsymbol{Y}_i^j$'s, and the number of possible paths is not given by the Catalan numbers, but the general structure of Catalan recursion is still preserved.

Also note that by closing the individual sub-matrices ($\boldsymbol{Y}_i^j$'s) using the relevant

starting and ending vectors, we can directly read the corresponding probabilities.

$$\begin{bmatrix}
\boldsymbol{p}\boldsymbol{Y}_0^0\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_0\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_0\boldsymbol{e}' & . & . & . & . & . \\
\boldsymbol{p}\boldsymbol{Y}_1^1\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_1\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_1\boldsymbol{e}' & . & . & . & . & . \\
\boldsymbol{p}\boldsymbol{Y}_2^1\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_2^2\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_2\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_2\boldsymbol{e}' & . & . & . & . \\
\boldsymbol{p}\boldsymbol{Y}_3^1\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_3^2\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_3^3\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_3\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_3\boldsymbol{e}' & . & . & . \\
\boldsymbol{p}\boldsymbol{Y}_4^1\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_4^2\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_4^3\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_4^4\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_4\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_4\boldsymbol{e}' & . & . \\
\boldsymbol{p}\boldsymbol{Y}_5^1\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_5^2\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_5^3\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_5^4\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_5^5\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_5\boldsymbol{e}' & \boldsymbol{p}\boldsymbol{Y}_5\boldsymbol{e}' & . \\
. & . & . & . & . & . & . & .
\end{bmatrix}$$

For example, $\boldsymbol{p}\boldsymbol{Y}_4^3\boldsymbol{e}'$ gives the probability that exactly five customers are served during a busy period when the maximum system size is limited to four.

5.2.1 Computing The Starting Vector

The starting vector $\boldsymbol{p}_{bp}$, for the busy period in this finite queueing case is computed using

$$\boldsymbol{p}_{bp}\boldsymbol{Y}^q\boldsymbol{H}_s\boldsymbol{H}_1 = \boldsymbol{p}_{bp}.$$

where, $\boldsymbol{Y}^q = \sum_{i=0}^{\infty} \boldsymbol{Y}_i^q$ represents all possible paths that lie within a channel of width q. The equation for computing the starting vector of a random busy period symbolizes the invariance for the system state between the starts of two successive busy periods. The invariance equation for $\boldsymbol{p}_{bp}$ is similar to that of the infinite queue case except that now the paths that comprise the $\boldsymbol{Y}$ are limited by the size of the queue, hence replaced by $\boldsymbol{Y}^q$. The intuition is still valid, that at the start of a random busy period if the starting vector is $\boldsymbol{p}_{bp}$ then following one of the possible paths $\boldsymbol{Y}^q$ the busy period ends ($\boldsymbol{H}_s$), followed by an arrival event occuring causing the start of the next busy period.

An alternate method [14], to compute the starting vector is by introducing $\boldsymbol{X}_h$,

representing all possible paths within a band of height h with possible loops at the top, then for a given maximum system size s, $\boldsymbol{X}_{s-1}$ can be computed using,

$$\begin{aligned}
\boldsymbol{X}_0 &= (\boldsymbol{I} - \boldsymbol{H}_a)^{-1}, \\
\boldsymbol{X}_1 &= (\boldsymbol{I} - \boldsymbol{H}_a \boldsymbol{X}_0 \boldsymbol{H}_s)^{-1}, \\
\boldsymbol{X}_2 &= (\boldsymbol{I} - \boldsymbol{H}_a \boldsymbol{X}_1 \boldsymbol{H}_s)^{-1}, \\
\vdots &= \quad \vdots \\
\boldsymbol{X}_{s-1} &= (\boldsymbol{I} - \boldsymbol{H}_a \boldsymbol{X}_{s-2} \boldsymbol{H}_s)^{-1},
\end{aligned}$$

and the starting vector $\boldsymbol{p}_{bp}$ is computed using

$$\boldsymbol{p}_{bp} \boldsymbol{X}_{s-1} \boldsymbol{H}_s \boldsymbol{H}_1 = \boldsymbol{p}_{bp}.$$

Now, the probability that exactly n customers are served during a busy period of a finite queue where the maximum height of a sample path or the maximum system size is restricted to s (corresponding to a maximum channel width of size $q = s - 1$ followed by the first arrival), is given by

$$d^s_{n,1} = Prob[N^s_{1,0} = n] = \boldsymbol{p}_{bp} \boldsymbol{Y}^{s-1}_{n-1} \boldsymbol{H}_s \boldsymbol{e}', \quad n \geq 1.$$

5.2.2 Mean Number Served During a Finite Queue Busy Period

For a given channel width q, define the matrix z-transform $\boldsymbol{Y}^q(z) = \boldsymbol{Y}^q_0 z^1 + \boldsymbol{Y}^q_1 z^2 + \boldsymbol{Y}^q_2 z^3 + \ldots$ for $q > 1$. We can now derive the following recurrence relation for $\boldsymbol{Y}^q(z)$.

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

$$
\begin{aligned}
\boldsymbol{Y}(z)^q &= z\boldsymbol{I} + \boldsymbol{H}_a(\boldsymbol{Y}_0^{q-1}z^1 + \boldsymbol{Y}_1^{q-1}z^2 + \boldsymbol{Y}_2^{q-1}z^3 + \ldots) \\
&\qquad \boldsymbol{H}_s(\boldsymbol{Y}_0^q z^1 + \boldsymbol{Y}_1^q z^2 + \boldsymbol{Y}_2^q z^3 + \ldots).
\end{aligned}
$$

Thus, $\boldsymbol{Y}^q(z)$ satisfies the matrix recurrence equation

$$
\boldsymbol{Y}^q(z) = z\boldsymbol{I} + \boldsymbol{H}_a\boldsymbol{Y}^{q-1}(z)\boldsymbol{H}_s\boldsymbol{Y}^q(z), \qquad q > 1. \tag{5.2.2}
$$

Notice the similarity to the matrix quadratic equation in relation to the infinite queueing situation. However we now have different $\boldsymbol{Y}^q(z)$'s for different allowable queue sizes, q. The boundary equation in the case where $q = 1$ is as follows

$$
\begin{aligned}
\boldsymbol{Y}^1(z) &= \boldsymbol{Y}_0^1 z^1 + \boldsymbol{Y}_1^1 z^2 + \boldsymbol{Y}_2^1 z^3 + \ldots, \\
&= z\boldsymbol{I} + Y_1^1 z^2(\boldsymbol{I} + (\boldsymbol{Y}_1^1)z + (\boldsymbol{Y}_1^1)^2 z^2 + \ldots, \\
&= z\boldsymbol{I} + \boldsymbol{Y}_1^1 z^2(\boldsymbol{I} - \boldsymbol{Y}_1^1 z)^{-1}.
\end{aligned}
$$

At $z = 1$,

$$\begin{aligned}\boldsymbol{Y}^1(1) &= \boldsymbol{I} + \boldsymbol{Y}_1^1(\boldsymbol{I} - \boldsymbol{Y}_1^1)^{-1}, \\ &= (\boldsymbol{I} - \boldsymbol{H}_a(\boldsymbol{I} - \boldsymbol{H}_a)^{-1}\boldsymbol{H}_s)^{-1}.\end{aligned}$$

Hence for any given allowable height q, $\boldsymbol{Y}^q(1)$ can be computed using

$$\boldsymbol{Y}^q(1) = \left(\boldsymbol{I} - \boldsymbol{H}_a\boldsymbol{Y}^{q-1}(1)\boldsymbol{H}_s\right)^{-1}, \qquad q > 1.$$

Taking the derivative of Eq. (5.2.2) w.r.t z and evaluating at $z = 1$ gives,

$$\boldsymbol{Y}^{q\,\prime}(1) = \boldsymbol{I} + \boldsymbol{H}_a\,\boldsymbol{Y}^{q-1\,\prime}(z)\boldsymbol{H}_s\boldsymbol{Y}^q(1) + \boldsymbol{H}_a\boldsymbol{Y}^{q-1}(1)\boldsymbol{H}_s\,\boldsymbol{Y}^{q\,\prime}(1). \tag{5.2.3}$$

The base case when $q = 1$, is

$$\boldsymbol{Y}^1(z) = \boldsymbol{I}z + \boldsymbol{Y}_1^1 z^2 + \boldsymbol{Y}_2^1 z^3 + \ldots$$

Taking the derivative at evaluating at $z = 1$,

$$\begin{aligned}\boldsymbol{Y}^{1\,\prime}(1) &= \boldsymbol{I} + 2\left(\boldsymbol{Y}_1^1\right)^2 + 3\left(\boldsymbol{Y}_1^1\right)^3 + 4\left(\boldsymbol{Y}_1^1\right)^4 + \ldots \\ &= (\boldsymbol{I} + \boldsymbol{Y}_1^1 + {\boldsymbol{Y}_1^1}^2 + {\boldsymbol{Y}_1^1}^3 + \ldots) + (\boldsymbol{Y}_1^1 + {\boldsymbol{Y}_1^1}^2 + {\boldsymbol{Y}_1^1}^3 + \ldots) + \ldots \\ &= (\boldsymbol{I} - \boldsymbol{Y}_1^1)^{-1} + \boldsymbol{Y}_1^1(\boldsymbol{I} - \boldsymbol{Y}_1^1)^{-1} + \boldsymbol{Y}_1^2(\boldsymbol{I} - \boldsymbol{Y}_1^1)^{-1} + \ldots \\ &= (\boldsymbol{I} - \boldsymbol{Y}_1^1)^{-1}.(\boldsymbol{I} - \boldsymbol{Y}_1^1)^{-1} \\ &= (\boldsymbol{I} - H_a(\boldsymbol{I} - \boldsymbol{H}_a)^{-1}\boldsymbol{H}_s)^{-2}.\end{aligned}$$

Hence we can compute $\boldsymbol{Y}^{q\,\prime}(1)$ for a given q using

$$\begin{aligned}\boldsymbol{Y}^{q\,\prime}(1) &= \left(\boldsymbol{I} - \boldsymbol{H}_a\boldsymbol{Y}^{q-1}(1)\boldsymbol{H}_s\right)^{-1} \\ &\quad \cdot \left(\boldsymbol{I} + \boldsymbol{H}_a\,\boldsymbol{Y}^{q-1\,\prime}(1)\boldsymbol{H}_s\boldsymbol{Y}^q(1)\right),\ q > 1.\end{aligned} \tag{5.2.4}$$

Now the mean number of customers served during a busy period when the allowable system size is s, is given by

$$\mathrm{E}[N^s_{1,0}] = \boldsymbol{p}(0)\ \boldsymbol{Y}^{s-1\,\prime}(1)\boldsymbol{H}_s\boldsymbol{e}'. \tag{5.2.5}$$

Similarly the second moment is computed using,

$$\boldsymbol{Y}^{1\,\prime\prime}(1) = 2(\boldsymbol{I} - \boldsymbol{Y}^1_1)^{-3}\boldsymbol{Y}^1_1,$$

and

$$\begin{aligned}\boldsymbol{Y}^{q\,\prime\prime}(1) \;\; &= \;\; \left(\boldsymbol{I} - \boldsymbol{H}_a\boldsymbol{Y}^{q-1}(1)\boldsymbol{H}_s\right)^{-1} \cdot \Big(\boldsymbol{H}_a\ \boldsymbol{Y}^{q-1\,\prime\prime}(1)\boldsymbol{H}_s\boldsymbol{Y}^q(1) \\ &\qquad +2\boldsymbol{H}_a\ \boldsymbol{Y}^{q-1\,\prime}(1)\boldsymbol{H}_s\ \boldsymbol{Y}^q(1)'\Big),\ q>1.\end{aligned} \tag{5.2.6}$$

It is to be noted that though the busy period analysis for a finite queue is similar to that of an infinite queue as presented in [28], it is considerably more intricate due to the fact that we now have an entire set of $\boldsymbol{Y}^j_i$'s instead of simply $\boldsymbol{Y}_i$'s. The constructive mechanism presented in the case of the infinite queue however does provide a basic mechanism to study the finite queue.

5.2.3 Mean Length of a Finite Queue Busy Period

Using the conditional Laplace transform derived in the previous section we construct recursive equations similar to Eq.(8) representing the Laplace transforms for the length of the sample paths during a busy period. The joint transform is given by

$$\boldsymbol{F}^*(s,z)^q = z\,\boldsymbol{I} + \boldsymbol{F}^*_a(s)\boldsymbol{F}^*(s,z)^{q-1}\boldsymbol{F}^*_s(s)\boldsymbol{F}^*(s,z)^q. \tag{5.2.7}$$

where $\boldsymbol{F}^*(s,z)^q$ represents the joint transform for the length of the sample paths and the number served when the allowable height of the channel is q. The boundary condition at $z=1$ and $q=0$ is given by $\boldsymbol{F}^{0\,*}(s,1)=(\boldsymbol{I}-\boldsymbol{F}_a^*(s))^{-1}$, and for any given finite queue size, the transform for the busy period for the system size is given by

$$\boldsymbol{F}_T^{q\,*}(s)=\boldsymbol{p}_{bp}\,\boldsymbol{F}^{q-1\,*}(s)\boldsymbol{F}_s^*(s)\boldsymbol{e}', \qquad q\geq 1. \tag{5.2.8}$$

The mean length of the busy period in this finite queueing case is computed using the following.

$$\begin{aligned}\boldsymbol{F}^*(s)^{q\prime}\Big|_{s=0} &= \left(\boldsymbol{I}-\boldsymbol{F}_a^*(0)\boldsymbol{F}^*(0)^{q-1}\boldsymbol{F}_s^*(0)\right)^{-1}\cdot\left(\boldsymbol{F}_a^*(0)'\,\boldsymbol{F}^*(0)^{q-1}\boldsymbol{F}_s^*(0)\boldsymbol{F}^*(0)^q+\right.\\ &\quad\left.+\boldsymbol{F}_a^*(0)\;\boldsymbol{F}^*(0)^{q-1\,\prime}\,\boldsymbol{F}_s^*(0)\boldsymbol{F}^*(0)^q+\;\boldsymbol{F}_a^*(0)\boldsymbol{F}^*(0)^{q-1}\;\boldsymbol{F}_s^*(0)'\,\boldsymbol{F}^*(0)^q\right),\end{aligned}$$

where $\boldsymbol{F}_a^*(0)=\boldsymbol{H}_a$, $\boldsymbol{F}_s^*(0)=\boldsymbol{H}_s$, $\boldsymbol{F}^*(0)^q=\boldsymbol{Y}^q$, $\boldsymbol{F}^*(0)^{q-1}=\boldsymbol{Y}^{q-1}$, $\boldsymbol{F}_a^*(0)'=-\boldsymbol{D}\boldsymbol{H}_a$, $\boldsymbol{F}_s^*(0)'=-\boldsymbol{D}\boldsymbol{H}_s$, and

$$\begin{aligned}\boldsymbol{F}^*(0)^{1\,\prime} &= -\left(\boldsymbol{I}-\boldsymbol{H}_a(\boldsymbol{I}-\boldsymbol{H}_a)^{-1}\boldsymbol{H}_s\right)^{-2}\cdot\left[\boldsymbol{D}\boldsymbol{H}_a(\boldsymbol{I}-\boldsymbol{H}_a)^{-1}\boldsymbol{H}_s+\right.\\ &\quad\left.+\boldsymbol{H}_a(\boldsymbol{I}-\boldsymbol{H}_a)^{-2}\boldsymbol{D}\boldsymbol{H}_a\boldsymbol{H}_s+\boldsymbol{H}_a(\boldsymbol{I}-\boldsymbol{H}_a)^{-1}\boldsymbol{D}\boldsymbol{H}_s\right].\end{aligned}$$

Let τ_b^m represent the r.v for busy period duration of a finite queue where the maximum allowable system size is m, then,

$$\begin{aligned}\mathrm{E}[\tau_b^m] &= -\frac{d}{ds}\left(\boldsymbol{p}_{bp}\boldsymbol{F}^*(s)^{m-1}\boldsymbol{F}_s^*(s)\boldsymbol{e}'\right)\Big|_{s=0}\\ &= -\boldsymbol{p}_{bp}\left(\boldsymbol{F}^*(0)^{m-1\,\prime}\,\boldsymbol{H}_s-\boldsymbol{Y}^{m-1}\boldsymbol{D}\boldsymbol{H}_s\right)\boldsymbol{e}'.\end{aligned} \tag{5.2.9}$$

5.3 Numerical Examples

We consider here various systems where the arrival and/or service process are markovian, renewal matrix exponential and correlated matrix exponential.

5.3.1 M/M/1 Case

Consider a simple *M/M/1* case where the $\boldsymbol{H}_a$ and $\boldsymbol{H}_s$ are scalars, some simplifications are evident. For example, since $(\boldsymbol{I} - \boldsymbol{H}_a)^{-1}\boldsymbol{H}_s$ is now equal to 1, we have

$$\boldsymbol{Y}_1^1 = \boldsymbol{H}_a, \quad \boldsymbol{Y}_n^1 = \boldsymbol{H}_a^n \;\; n \geq 2.$$

But no better structure is evident yet for higher level $\boldsymbol{Y}_i^j$'s than as defined by Eq. (5.2.2). The count process for the number of possible paths that serve n customers during a busy period of this system is not given by the catalan numbers. In-fact, there can be an infinite number of possible sample paths due to the arrivals that get dropped represented by $(\boldsymbol{I} - \boldsymbol{H}_a)^{-1}$. For any given fixed queue size we can however compute the probabilities for n customers served during a busy period using Eq. (5.2.2). For an $M/M/1$ queue with a utilization of 0.7, the probabilities $d_{n,1}^s$ are shown in Table 7. Notice that for a given system size, the probabilities for n customers served differ from the infinite queueing situation starting when n is equal to the maximum allowable system size. For example $d_{3,1}^3$=0.093178 where as for the rest of the system sizes, $d_{3,1}^5$ = $d_{3,1}^{10}$ = $d_{3,1}^{100}$ = 0.069021; this is as expected.

The Laplace transform for the length of the busy period in finite queueing situation

Table 7: MM1 Finite Queue: $d^s_{n,1}$ for a Utilization = 0.70

n	$s=3$	$s=5$	$s=10$	$s=100$
	$d^s_{n,1}$			
1	0.588235	0.588235	0.588235	0.588235
2	0.142479	0.142479	0.142479	0.142479
3	0.093178	0.069021	0.069021	0.069021
4	0.060937	0.041795	0.041795	0.041795
5	0.039851	0.029762	0.028345	0.028345
6	0.026062	0.023093	0.020597	0.020597
7	0.017044	0.018606	0.015679	0.015679
8	0.011146	0.015221	0.012343	0.012343
9	0.00729	0.012524	0.009965	0.009965
10	0.004767	0.010329	0.008208	0.008207

is of the general form

$$f^{q\,*}(s) = \cfrac{a}{1 - \cfrac{b}{1 - \cfrac{b}{1 - \cfrac{b}{\ddots\, 1 - \cfrac{b}{1 - \frac{b}{c}}}}}}, \tag{5.3.1}$$

where $\rho = \frac{\lambda}{\mu}$, $a = \frac{1}{1+\rho+s}$, $b = \frac{\rho}{(1+\rho+s)^2}$, $c = 1 - \frac{\rho}{1+\rho+s}$, and the depth of the continued fraction depends on the maximum allowable system size. The problem of finding a closed form expression for the above finite continued partial fraction seems to be an open problem as of yet. However, for any given finite system size we show below a general form for the first two moments for the length of the busy period. Let τ^s be the random variable representing the length of the busy period where the finite system size is limited by s,

$$E[\tau^s] = \frac{1-\rho^s}{1-\rho}, \qquad s > 1 \tag{5.3.2}$$

$$E[{\tau^s}^2] = 2! \left(\sum_{n=1}^{s} T_n \rho^{n-1} + \sum_{n=s-1}^{1} T_n \rho^{2s-1-n} \right), \qquad s > 1, \tag{5.3.3}$$

where, T_n are the Triangular numbers given by $T_n = \frac{n(n+1)}{2}, n \geq 1$.

5.3.2 MEP/MEP/1 Case

Consider an *MEP/MEP/1* system where the arrival process is represented by

$$\boldsymbol{p}_a = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \quad \boldsymbol{B}_a = 1.4 * \begin{bmatrix} 1 & 0 \\ 0 & 25 \end{bmatrix},$$

$$\boldsymbol{L}_a = 1.4 * \begin{bmatrix} (1+\gamma_a) & (1-\gamma_a) \\ (1-\gamma_a) & (1+\gamma_a) \end{bmatrix},$$

where $\gamma_a = 0.9$, is the parameter that controls the correlation decay of the process. Note that the marginal distribution is independent of γ_a and has a mean of 0.3714 and a squared coefficient of variation $c^2 = 2.7$. Similarly, consider the service process represented by

$$\boldsymbol{p}_s = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \end{bmatrix}, \quad \boldsymbol{B}_s = \begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix}, \quad \boldsymbol{L}_s = \begin{bmatrix} 2.775 & 0.225 \\ 0.1 & 3.9 \end{bmatrix}.$$

This system has a utilization of 0.73 and the probabilities of serving n customers in a busy period for different allowable system sizes ($d^s_{n,1}$) are shown in Table 8.

Table 8: MEP/MEP/1 Finite Queue: $d^s_{n,1}$ for a Utilization = 0.73

n	$s=3$	$s=5$	$s=10$	$s=100$
	$d^s_{n,1}$			
1	0.692976	0.692976	0.692976	0.692976
2	0.130622	0.130622	0.130622	0.130622
3	0.068845	0.050156	0.050156	0.050156
4	0.038867	0.024632	0.024632	0.024632
5	0.023346	0.018133	0.013933	0.013933
6	0.014738	0.015667	0.008724	0.008724
7	0.009649	0.013558	0.005931	0.005931
8	0.006477	0.011368	0.004325	0.004325
9	0.004418	0.009268	0.003348	0.003348
10	0.003046	0.007407	0.003097	0.002724

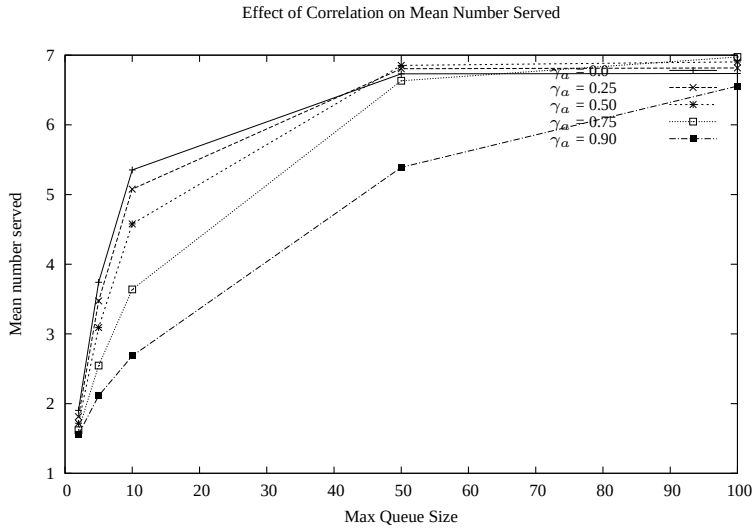

Figure 19: *G/G/1*: Effect of γ_a on mean number served

We show the effect of the correlation parameter γ_a on the mean and squared coefficient of variation of number of customers served in the finite case in Fig. 19 and Fig. 20 respectively. An interesting observation is that for a given utilization, as we allow the maximum queue size attainable to grow, the mean number of customers served during a busy period tend to converge. Though this gives an impression that at higher allowable system sizes, the correlation parameter does not have a considerable impact on the number served, the impact of this increase in correlation parameter can be clearly seen as it effects the variance and hence the squared coefficient of variance for the number served.

We show the effect of γ_a on the mean length of the busy period in Fig. 21. Comparing the mean busy period lengths for a maximum allowable size of 50 and 100, it can be seen that except for when γ_a is 0.9, the mean busy period lengths for a maximum height

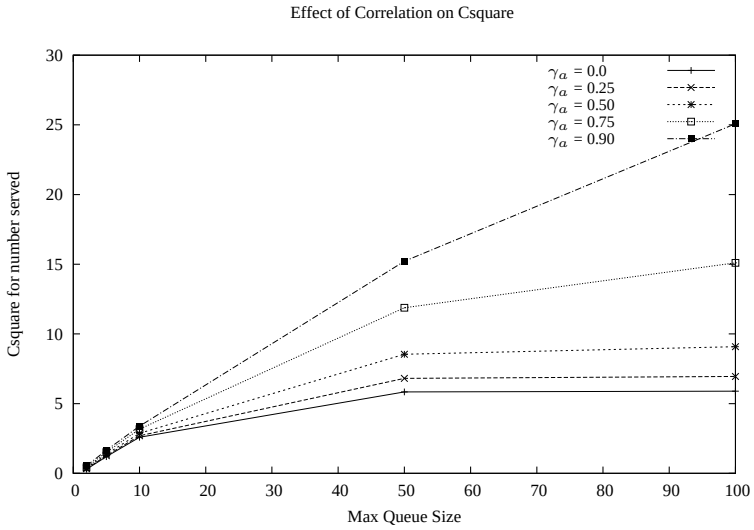

Figure 20: *G/G/1*: Effect of γ_a on c^2 for number served

of 50 and 100 match to atleast one significant digit after the decimal. Again, this is as expected. As the correlation decay parameter increases, we would expect the finite busy period means to approach the infinite queue's busy period mean at increasing maximum allowable queue sizes and hence they tend to converge slower.

For different allowable system sizes, by ignoring the loops that form at the top, a set of count processes are generated from the the set of equations. 5.2.2. Some of these number series are known to be related to the number of possible paths in finite spaces. However, the set of equations as defined in 5.2.2 allows us to unify them all. The different number series are readily obtained by setting $\boldsymbol{Y}_0^0 = 1$, $\boldsymbol{H}_a = 1$ and $\boldsymbol{H}_s - 1$ and are shown in Table 9.

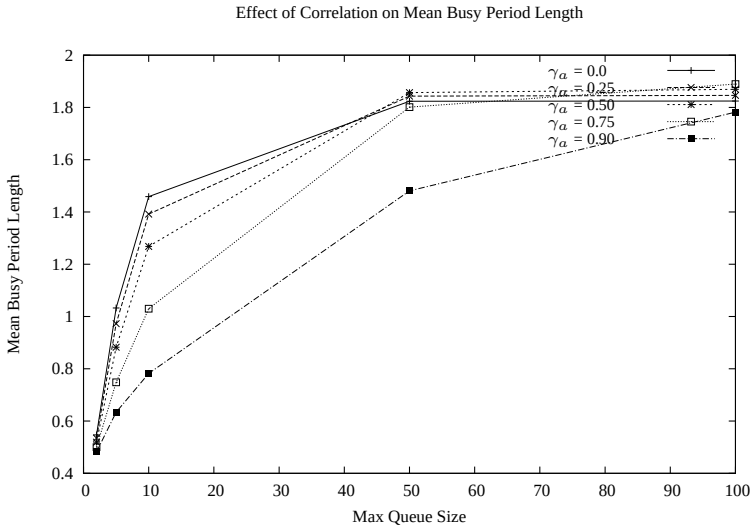

Figure 21: *G/G/1*: Effect of γ_a on mean busy period length

5.4 Conclusions

In this chapter we derived closed form recursive solutions to compute the probabilities for n customers served during the busy period of a finite *MEP/MEP/1* system wherein both the arrivals and services can be auto-correlated. We also derived expressions to compute the first two moments for the number of customers served and expression to compute the mean length of a busy period in finite queues. This framework provides us with tools

Table 9: Catalan like sequences related to finite queues

Max height	Number Sequence
3	1, 1, 2, 5, 13, 34, 89, 233, 610, 1597, 4181, 10946, 28657, 75025
4	1, 1, 2, 5, 14, 41, 122, 365, 1094, 3281, 9842, 29525, 88574, 265721
5	1, 1, 2, 5, 14, 42, 131, 417, 1341, 4334, 14041, 45542, 147798, 479779
6	1, 1, 2, 5, 14, 42, 132, 428, 1416, 4744, 16016, 54320, 184736, 629280
7	1, 1, 2, 5, 14, 42, 132, 429, 1429, 4846, 16645, 57686, 201158, 704420
8	1, 1, 2, 5, 14, 42, 132, 429, 1430, 4861, 16778, 58598, 206516, 732825
9	1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16795, 58766, 207783, 740924

to study similar first passages starting from a random or an environment-defined starting vectors. 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.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

In this thesis we developed a framework which can be used to study and better understand many stochastic processes occurring in nature. The framework presented allows the constituting processes to be both general and correlated and hence lead to more realistic models. Applications in finance, biological and social sciences are noted and we study the transient busy period as it relates to computer networks and queues in detail to demonstrate the application of tracking these memory-full processes.

Essentially by representing the current state of a system using a relevant starting state vector, and by allowing the driving process to carry correlations across state transitions of the underlying quasi-markovian chain enables us to track these paths very accurately.

In the first part of the thesis we provided solutions to compute the probabilities for exactly 'n' customers being served in a busy period of *MEP/MEP/1* queueing systems and provided some numeric results both using this analytic approach as well as by simulation. We presented the results in this part in an algorithm hence making it a straight forward task to compute the performance metrics of interest. We then presented new matrix exponential representations to characterize the lengths of sample paths during these busy periods and derived expressions to compute the moments for the length of a busy

period as well as for the number of customers served during the busy period. We studied the effect of various parameters effecting the constituting processes by demonstrating the effects of the the first three moments and the auto-correlations in the arrival and service processes on busy periods.

In the second part of the thesis, we study the effect of increase in threshold level and the correlations in the arrival and service processes on the mean first passage time to go below a given threshold. Finally we studied the busy periods for general finite queueing systems and derived recursive matrix quadratic equations which have a similar structure to the matrix quadratic equations in an infinite case but are relatively more intricate.

6.2 Future Work

As noted earlier, the matrix quadratic form for $\boldsymbol{Y}(z)$ derived in Chapter.3 is closely related to the well known matrix quadratic equation for the matrix $\boldsymbol{G}$ that occurs frequently in matrix geometric literature [33], [35]. In fact, "$\boldsymbol{Y}(1)\boldsymbol{H}_s$" is equivalent to the matrix $\boldsymbol{G}$ if the system under consideration has *MAP* process as one of the driving processes.

In [3] and [4], Kumaran et al. propose a spectral decomposition based approach to compute various performance metrics, including the waiting time, system size etc., for a single server *MEP/MEP/1* queue. The method essentially involves constructing a Coupling matrix $\boldsymbol{C}$ (introduced by Van de Liefvoort [27]) from the arrival and service processes, which is then spectrally decomposed to form the matrix $\boldsymbol{R}$, from which explicit solutions are derived for waiting times etc. This matrix $\boldsymbol{R}$ thus constructed is found to be similar to the similarly named matrix $\boldsymbol{R}$ that occurs in matrix geometric literature, in

the sense that they have the same eigen values. Let the $\boldsymbol{R}$ matrix that occurs in matrix geometric literature be denoted by $\boldsymbol{R}_n$ and the $\boldsymbol{R}$ matrix computed from the Coupling matrix be denoted by $\boldsymbol{R}_a$. We suspect that there exists a matrix transformation that relates these two different $\boldsymbol{R}$ matrices. If such a transformation were found, then the computation of the matrix $\boldsymbol{Y}\boldsymbol{H}_s$ or alternatively $\boldsymbol{G}$ can be performed very efficiently as the matrix $\boldsymbol{G}$ and $\boldsymbol{R}_n$ are related by $\boldsymbol{G} = (\boldsymbol{I} - \boldsymbol{R}_n\boldsymbol{H}_s)^{-1}\boldsymbol{H}_s$. Such a transformation remains elusive yet.

REFERENCE LIST

[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] J. Kumaran, K. Mitchell and A. Van de Liefvoort, "A spectral approach to compute performance measures in a correlated single server queue". *SIGMETRICS, Performance Evaluation Review 33*, 2 (2005), 12-14.

[4] J. Kumaran, K. Mitchell and A. Van de Liefvoort, "The waiting time distribution of an $MEP/MEP/1$ queue.". *Proceedings of the 19th International Teletrafic Congress (ITC19)*, (2005), pp. 2327-2336.

[5] 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.

[6] 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.

[7] 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).

[8] O. J. Boxma and V. Dumas, " The busy period in the fluid queue". *Centrum voor Wiskunde*

en Informatica (CWI), Amsterdam, Netherlands. PNA-R9718, 1997.

[9] 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.

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

[11] 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.

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

[13] 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.

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

[15] G. Watson, "A Treatise on the Theory of Bessel Functions, 2nd ed." Cambridge, England: Cambridge University Press, 1966.

[16] J. Medhi, " Stochastic Models in Queueing Theory", Second Edition. Academic Press, 2003. ISBN:0-12-487462-2.

[17] Fackrell, M. *Characterization of matrix-exponential distributions*. PhD thesis, Adelaide University, 2003.

[18] Van de Liefvoort, A. The waiting time distribution and its moments of the *PH/PH/1* queue. *Operations Research Letters 9*, 4 (1990), 261–269.

[19] Neuts, M. *Matrix-Geometric Solutions in Stochastic Models*. Johns Hopkins University

Press, Baltimore, MD, 1981.

[20] Heindl, A., Mitchell, K., and Van de Liefvoort, A. The correlation region of second-order MAPs with application to queueing network decomposition. In *Lecture Notes in Computer Science*, vol. 2794. Sep 2003, pp. 237 – 254.

[21] Neuts, M. A versatile Markovian point process. *Journal of Applied Probability 16* (1979), 764–779.

[22] Lee, Y., Van de Liefvoort, A., and Wallace, V. Modeling correlated traffic with a generalized IPP. *Performance Evaluation 40* (2000), 99–114.

[23] Asmussen, S., and Koole, G. Marked point processes as limits of Markovian arrival streams. *Journal of Applied Probability 30* (1993), 365–372.

[24] Lipsky, L., Fiorini, P., Hsin, W., and Van de Liefvoort, A. Auto-correlation of lag-k for customers departing from semi-Markov processes. Tech. Rep. TUM-19506, Technical University, Munich, 1995.

[25] Lee, Y., Van de Liefvoort, A., and Wallace, V. Generalized Bernoulli process as a discrete-time model for network traffic. Tech. rep., University of Kansas, EECS Department, University of Kansas, Lawrence, KS, 1998. DesignLab Technical Report No. DL-1998-04.

[26] Van de Liefvoort, A., and Heindl, A. Approximating matrix-exponential distributions by global randomization. *Stochastic Models 21* (2005), 669–693.

[27] Van de Liefvoort, A. The waiting time distribution and its moments of the *PH/PH/1* queue. *Operations Research Letters 9*, 4 (1990), 261–269.

[28] C. Garikiparthi, A. van de Liefvoort and K. Mitchell, *Sample Path Analysis of Busy Periods and Related First Passages of a Correlated MEP/MEP/1 System*. Fourth International Conference on Quantitative Evaluation of Systems, QEST 2007, Scotland, UK. Pg. 277-286.

[29] T. Osogami, M. Harchol-Balter. “Necessary and sufficient conditions for representing general distributions by Coxians”. *Technical Report: CMU-CS-02-178*, Carnegie Mellon University, PA, Sept 2002.

[30] N. Akar, K. Sohraby. “Finite and Infinite QBD Chains: A Simple and Unifying Algorithmic Approach”. *Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution*, 1997 INFOCOM. Pg. 1105.

[31] Cox, D. Use of complex probabilities in the theory of stochastic processes. In *Proc. Cambridge Philosophical Society* (1955), vol. 51, pp. 313–319.

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

[33] 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.

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

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

[36] M. Neuts, “ Structured Stochastic Matrices of M/G/1 tyep and Their Applications”. Marcel Dekker, Inc., New York, 1989.

[37] 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.

[38] 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.

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

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

[41] 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.

[42] 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.

[43] A. Van de Liefvoort, "The moment problem for continuous distributions". *Technical Report WP-CM-02*. School of Interdisciplinary Computing and Engineering, University of Missouri Kansas City, USA, 1990.

[44] Mitchell, K. Constructing correlated sequence of matrix exponentials with invariant first-order properties. *Operations Research Letters 28*, 1 (2001), 27–34.

VITA

Chaitanya Garikiparthi is a doctoral candidate in the School of Computing and Engineering at the University of Missouri - Kansas City. His main research interests are performance analysis and modeling (using linear algebraic queueing theory), applied probability and stochastic processes. He received a Bachelors in Mechanical Engieering from CBIT, Osmania University, India in 1999 and a Masters in Computer Sciences from the University of Texas at Dallas in 2001. He did brief internships at Nortel Networks in 2001, and at Motorola in 2006.