On Queues with Markov Modulated Service Rates

Sai Rajesh Mahabhashyam and Natarajan Gautam*
310 Leonhard Building
The Pennsylvania State University
University Park, PA 16802
{sairajesh, ngautam}@psu.edu

Abstract

In this paper, we consider a queue whose service speed changes according to an external environment that is governed by a Markov process. It is possible that the server changes its service speed many times while serving a customer. We derive first and second moments of the service time of customers in system using first step analysis to obtain an insight on the service process. In fact, we obtain an intriguing result in that the moments of service time actually depend on the arrival process! We also show that the mean service rate is not the reciprocal of the mean service time.

Further, since it is not possible to obtain a closed form expression for the queue length distribution, we use matrix geometric methods to compute performance measures such as average queue length and waiting time. We apply the method of large deviations to obtain tail distributions of the workload in the queue using the concept of effective bandwidth. We present two applications in computer systems: 1) Web server with multi-class requests and 2) CPU with multiple processes. We illustrate the analysis and various methods discussed with the help of numerical examples for the above two applications.

Keywords: Markov modulated processes, first step analysis, matrix geometric method, large deviations.

*Corresponding Author; Phone: (814) 865 1239; Fax (814) 863 4745

1 Introduction

There are several articles in the literature that talk about time-varying arrival rates into queues, for example, non-homogeneous Poisson processes, Markov modulated Poisson process, Markovian arrival processes, variable rate fluid arrival processes, state-dependent arrival processes, etc. However, relatively little work has been devoted to time-varying service rates (exceptions include the state-dependent service rate case and the service rate control case). In fact, most papers and books that deal with properties such as stability of time-varying systems, mainly derive results for time-varying arrivals. One of the reasons for this can be attributed to the fact that in most situations the service speed does not arbitrarily change with time, except say in computer processors, web servers, etc. where a single processor works on multiple jobs in parallel.

In this paper, we consider a single server queueing system with infinite waiting space where customers bring a random amount of work. The server processes the work at different speeds which are piecewise constant over random periods of time. These are typical in computer systems where processors serve several applications in parallel, of which we are interested in one application's performance. The amount of processor speed available for this application depends on the number and type of other applications running on the processor. In that light we model the processor as a single server with processing speeds that vary according to a stochastic environment process. For this single server queueing model with time-varying service rates, our aim is to obtain the following performance measures: (i) mean and variance of the service time in steady state using first-step analysis; (ii) average queue length and average waiting time using matrix geometric methods; (iii) tail distribution of the system workload under steady state using large deviations theory.

From a methodological standpoint, all three techniques (first step analysis, matrix geometric methods as well as large deviations) have been used extensively in the literature to solve problems in queueing theory. The key contribution of this paper, besides innovatively using the above three methods, is to open a new avenue of research problems, as the stochastically varying server speed problem has received very little attention. There are a few articles that are related. The first is Zhou and Gans [28], where the authors state that time-varying service rate problems have not been studied in the literature.

The key difference between Zhou and Gans [28] and ours is that, they consider service speeds that change only when a customer completes service. So, unlike in our paper, the server speed cannot change during the middle of a service. In addition, the service rates mainly take only two values in Zhou and Gans [28]. Boxma and Kurkova [5] considers an $M/G/1$ queue where the speed of the server alternates between two values with high speed periods having exponential distribution and low speed periods having a general distribution. Motivated by the transportation system where if an incident occurs on a road segment all the vehicles on the road have to lower their speed until that incident is cleared, Baykal-Gursoy and Xiao [3] considers an $M/M/\infty$ queueing system subject to random interruptions of exponential distributed durations. Another related article is Núñez-Queija [20] where the author considers a special case of what we consider in this paper, namely the environment process is a specific birth and death process (i.e. queue length process of an $M/M/c/c$ queue). For that system the author uses matrix geometric methods to derive the mean waiting time and mean number in the system. In this paper, we generalize the environment process to any Markov process, and also obtain other performance metrics such as the service time moments and tail distribution of the workload in steady state.

Before describing how this paper is organized, we touch upon some of the related work. There are several articles in literature that consider queues with time varying arrival times. There are many articles (for e.g. Takahashi and Wang [26], Ahn and Jeon [2]) that analyze queues with Markov modulated arrivals. Some researchers like Adan and Kulkarni [1], and Cidon et al [6], analyze queues that have inter-arrival times and service times dependent on each other. On the whole, researchers have focussed more on analyzing queues with time varying and Markov modulated arrivals. On the other hand, there have been very few articles on Markov modulated service times and rates (besides Zhou & Gans [28], Núñez-Queija [20], Boxma & Kurkova [5], and Baykal-Gursoy & Xiao [3]). However, there are several articles on service rate control where service rate is time-varying (for example, Sharma [25]). Some papers like Massey [18] talk about queues with deterministic time varying rates in telecommunication models. Also, Collings and Stoneman [7] considers a $M/M/\infty$ queue with deterministic time varying arrivals and service rates. But, the Markov Modulated service rate problem in which the service rates vary according to an environment process is still unexplored to the best of our knowledge.

The rest of the paper is organized as follows. In Section 2, the problem under consideration is described in detail with all the notation (in Table 1). Two applications related to the problem are described in Sections 2.1 and 2.2. In Section 3, we derive the first and second moments of the service time of customers in system using first step analysis. Having obtained some idea on the service process, we look at some performance measures like average waiting time in system and average queue length using matrix geometric method in Section 4. In Section 5, we apply large deviation analysis to obtain tail distribution of the workload in system. Finally, in Section 6, we conclude our paper with ideas on some extensions and future work.

2 Problem Description

The purpose of this paper is to study a queue whose service capacity varies over time. That is, the speed of the server with which it serves a customer, is determined by an external environment process. In particular, we assume that the server speed changes according to a Continuous Time Markov Chain (CTMC) that is independent of the arrival process and service requirements of the customer. Each customer brings a certain random amount of work, however, the rate at which this work is completed is time-varying. For example, the server serves at different rates (bytes per second) over time to serve a request that needs a certain number of bytes of work. Other than that, the queue is a fairly standard one. We assume that the customers in the queue are served in a First Come First Served (FCFS) manner. For this model, we obtain first and second moments of the service times, average steady state number in system, and tail distributions of workload in system.

The system is represented schematically in Figure 1. Customers arrive into the queue according to a Poisson process with mean rate λ customers per unit time. Each arriving customer brings a certain amount of work distributed exponentially with mean $1/\mu$. Let $X(t)$ be the number of customers in queue at time t. Let $Z(t)$ be the state of the environment process which governs the server speed at time t such that $\{Z(t),\ t \geq 0\}$ is an ergodic CTMC. When the state of the environment process $Z(t) = i$, the service speed available is b_i. That is, the server can do b_i amount of work per unit time. Let θ_i be the instantaneous service completion rate when the environment is

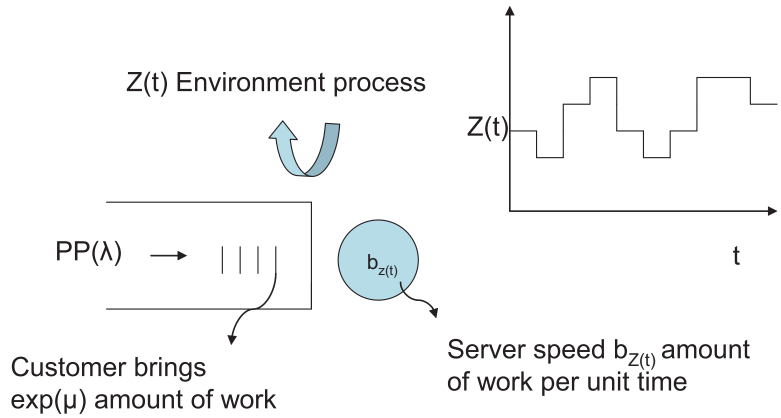

Figure 1: Schematical representation

in state i. Typically, $\theta_{Z(t)} = \mu b_{Z(t)}$. The bivariate stochastic process $\{(Z(t), X(t)), t \geq 0\}$ is also a CTMC. However, closed form solution for the steady state probability distribution is difficult. In steady state, when $t \to \infty$, the processes $X(t) \to X$ and $Z(t) \to Z$ under certain conditions of stability discussed in Section 4. We will discuss and use the process $\{(Z(t), X(t)), t \geq 0\}$ in Sections 3 and 4. All the notation described in this section and the rest of the paper is summarized in Table 1. Before analyzing the system and deriving performance measures, we present two examples below that motivated this research. These examples will be used throughout the paper for numerical results.

2.1 Application 1: Web Server with Multiclass Requests

The above scenario can be applied to a web server processing requests of different classes. We consider multiple classes of requests with varying Quality of Service (QoS) needs. Specifically, there are two main request classes, one is a *streaming* class that has bandwidth requirements, and the other is an *elastic* class that utilizes the processing capacity not used by the streaming requests. The motivation for this comes from the fact that most web pages have both streaming as well as elastic requests that need to be delivered to the users. In the communication networks community, researchers have considered both these classes separately for analysis. Research focusing on streaming traffic can be categorized as 'loss networks' (see Ross [24], Roberts [23], Kaufman [12]) while research that caters to the needs of elastic traffic can be categorized as 'delay networks'. Bonald and Proutiere [4] considered a queueing network with varying flow-rate (elastic traffic as the primary flow) and studied flow-level dynamics. However, very few articles (such as Quieja [20]) have considered both traffic simultaneously for analysis.

In this system, there are N types of requests within the streaming class. Type i (for $i = 1, 2, ..., N$) streaming requests arrive according to a Poisson process with mean arrival rate λ_i. The requests (if accepted) are served in parallel as soon as they arrive. Let the server capacity be S bytes per second. Each type i streaming request consumes a fraction (say r_i bytes/sec) of the server capacity. Each type i streaming request stays for a time exponentially distributed with mean $1/\mu_i$. The remaining capacity unused by the streaming requests is offered to elastic requests which arrive according to a Poisson process with mean rate λ and request files exponentially distributed with mean $1/\mu$. The speed at which the server works on the elastic traffic varies over time, depending on the number and type of streaming requests present.

First, consider a case where $N = 1$. Consider a web server with the following "static" admission control policy. The maximum number of streaming requests allowed simultaneously is 'n'. That is, an arriving streaming request is rejected if there are 'n' ongoing streaming requests at that time, else the arriving request is admitted and allocated a fixed bandwidth 'r_1'. All arriving streaming requests are admitted into the system. That is accomplished by choosing $c > 0$ such that $S = nr_1 + c$, where S is the processing capacity of the web server. The process of serving the streaming requests does not depend on the elastic requests. Also, if there are n streaming requests being served upon an arrival of a new streaming request, the new request is blocked. Thus, in this example, the $Z(t)$ (environment) process is the queue length process of an $M/M/n/n$ queue. This special case is solved by Núñez-Queija [20] using matrix geometric method. In this paper, we have a more general case, where the generator matrix, Q is not just that of an $M/M/n/n$ queue.

For our example, we consider $N = 2$ where there are two bandwidths for the streaming traffic. This is very common in websites that broadcast sports (for example, World cup soccer and cricket) over the internet. The users are given an option to select one of the two bandwidths offered depending on their connection speed. Let us denote the two bandwidths by $r_1 = 0.265$ (low bandwidth) and $r_2 = 0.350$ (high bandwidth). Let the processing capacity of the web server be $S = 0.650$. The arrival rates of requests for the two bandwidths (low and high respectively) are exponentially distributed with parameters $\lambda_1 = 1$, $\lambda_2 = 2$. The holding times are exponentially distributed with parameters $\mu_1 = 2$, $\mu_2 = 3$ for the two bandwidths respectively. The arrival rate and file size of elastic traffic are exponentially distributed with respective parameters $\lambda = 3$ and $\mu = 8$. The possible states of the environment process (i.e. state of streaming traffic) are $(0, 0)$, $(1, 0)$, $(2, 0)$, $(0, 1)$, $(1, 1)$, where the first tuple represents the number of ongoing low bandwidth requests, and the second one represents the number of ongoing high bandwidth requests. The corresponding available bandwidths for the elastic traffic are $b_{(0,0)} = 0.650$, $b_{(1,0)} = 0.385$, $b_{(2,0)} = 0.120$, $b_{(0,1)} = 0.300$, and $b_{(1,1)} = 0.035$. So, the infinitesimal generator matrix Q is given by

$$Q = \begin{bmatrix} -\lambda_1 - \lambda_2 & \lambda_1 & 0 & \lambda_2 & 0 \\ \mu_1 & -\mu_1 - \lambda_1 - \lambda_2 & \lambda_1 & 0 & \lambda_2 \\ 0 & 2\mu_1 & -2\mu_1 & 0 & 0 \\ \mu_2 & 0 & 0 & -\mu_2 - \lambda_1 & \lambda_1 \\ 0 & \mu_2 & 0 & \mu_1 & -\mu_1 - \mu_2 \end{bmatrix} = \begin{bmatrix} -3 & 1 & 0 & 2 & 0 \\ 2 & -5 & 1 & 0 & 2 \\ 0 & 4 & -4 & 0 & 0 \\ 3 & 0 & 0 & -4 & 1 \\ 0 & 3 & 0 & 2 & -5 \end{bmatrix}$$

We will look at this example later again in Sections 3, 4 and 5 to obtain performance measures for the elastic traffic queue.

2.2 Application 2: CPU with Parallel Processes

Another application where the problem in Section 2 can be immediately applied is a Central Processor Unit (CPU) of a computer which runs multiple processes in parallel. The process that is of our interest is a software agent that submits tasks to the CPU continuously throughout the day. Assume that the software agent submits tasks according to a Poisson process and each task has $\exp(\mu)$ work in it that the CPU has to perform. If the only process running on the CPU is that of the agent, it receives all the CPU speed. However if there are few other processes running on the CPU, only a fraction of the CPU speed is available. Hence due to various processes running at the same time, the processing speed for the agent tasks varies over time. This system can be modeled as a queue with time varying service rates. The service rates vary according to an external environment process, which is due to the other processes that run on the CPU. Note that it is not necessary that the CPU is shared equally among all processes. We use a very generic model for the available processing capacity of the CPU. Let the available capacity vary according to a CTMC $\{Z(t),\, t \geq 0\}$ with generator matrix Q such that at time t the available processing speed for the agent tasks is $b_{Z(t)}$. Thus, this can also be modeled as a queueing system described earlier in Section 2.

Throughout the paper, we will consider the following numerical example for illustration purposes. For numerical examples, we will consider various forms of Q and $b_{Z(t)}$. We present an example to illustrate how this application boils down to the considered problem. There are 5 possible server speeds, i.e., $Z(t)$ takes values 1 to 5. They are $b_1 = 1$, $b_2 = 2$, $b_3 = 3$, $b_4 = 4$, and $b_5 = 5$. The infinitesimal generator matrix Q is a 5×5 matrix given by

$$Q = \begin{bmatrix} -6 & 2 & 1 & 2 & 1 \\ 1 & -7 & 3 & 2 & 1 \\ 3 & 2 & -8 & 2 & 1 \\ 2 & 1 & 1 & -5 & 1 \\ 3 & 4 & 1 & 2 & -10 \end{bmatrix}.$$

The mean arrival rate is $\lambda = 2.5$ and the mean task size $1/\mu = 1$.

3 Finding First and Second Moments of Service Time

Let us consider the problem described in Section 2. Since the service rate keeps changing according to an external environment, it is not clear how long each request is served. We will hence derive the first and second moments of service time experienced by an arbitrary customer in steady state. Finally, we will apply the results developed in Section 3.1 to numerical examples.

3.1 Conditional moments

Consider an arbitrary customer in the queue in steady state. In order to obtain the moments of the service time experienced by this customer, we need to know the state of the environment process when the service starts. Hence, we begin by deriving results for the first and second moments of service time and when the service begins with environment in state i, denoted by $E(T_i)$ and $E(T_i^2)$ respectively. We now state and prove a theorem to obtain the first and second moments of this conditional service time, $E(T_i)$ and $E(T_i^2)$. Let q_{ij} be the element of the generator matrix Q of the environment process that corresponds to ith row and jth column and let $q_i = -q_{ii}$. Also, let θ_i be the rate of completion when there are i requests.

Table 1: Notation

Variables	Explanation
λ	Arrival rate of requests
$1/\mu$	Average work requested per customer
$Z(t)$	State of the environment process at time t
Z	When $t \to \infty$ $Z(t) \to Z$
$X(t)$	Number of requests in system at time t
X	When $t \to \infty$ $X(t) \to X$
b_i	Service capacity when $Z(t) = i$
θ_i	Rate of completion of requests when there are i requests
Q	Infinitisemal generator matrix of the environment process
W	Average delay of requests
$S^{(i)}$	State space with i requests
$\pi^{(i)}$	Steady state probability vector for state space $S^{(i)}$
p_i	Steady state probability that $Z = i$
$\hat{\pi}_i$	Steady state probability that a customer sees $Z = i$at the beginning of service
T_i	Service time of a customer whose service begins with $Z = i$
S	Total processing capacity of the web server
n	Maximum no. of streaming requests allowed simultaneously
r_i	Bandwidth allocated to each streaming request of type i when admitted in example 1
c	Minimum bandwidth for elastic requests in example 1
λ_i	Arrival rate of type i of streaming requests in example 1
R	Auxilliary matrix used in calculating average delay
$A(t)$	Total amount of traffic generated by a source over time $(0, t]$
$eb(v)$	Effective bandwidth with parameter v
$W(t)$	Amount of workload at time t
$\bar{R}$	Diagonal rate matrix
$e(M)$	Largest eigen value of square matrix M
W_i	Amount of work brought in by ith customer

Theorem 1 *The first and second moments of the service time conditional on the state of the environment process being "i" at beginning of service, are obtained by solving the set of equations*

$$f_i E(T_i) - \sum_{j \neq i} q_{ij} E(T_j) \;\; = \;\; 1, \tag{1}$$

and $E(T_i^2)$ is obtained by solving the set of equations

$$f_i^2 E(T_i^2) - f_i \sum_{j \neq i} q_{ij} E(T_j^2) \;\; = \;\; 2(1 + \sum_{j \neq i} q_{ij} E(T_j)), \tag{2}$$

where $f_i = q_i + \theta_i$.

Proof.

The proof applies first step analysis (Dorman [9], Kulkarni [15]) based on the Laplace Stieltjes transform (LST) of the conditional service time. Let T_i be the random variable denoting total service time that begins in state i for an arbitrary customer in steady state. Let V_i be the service time if the CTMC were always in state i. Clearly, V_i is exponentially distributed with parameter θ_i. Let f_i= $\sum_{j \neq i} q_{ij} + \theta_i = q_i + \theta_i$. Let R_{ij} be the random variable denoting the time between state change from i to j. Hence, we have

$$T_i = \min(V_i, R_{ij(j \neq i)}) \;\; + \;\; \begin{cases} 0 & w.p \quad \theta_i/f_i \quad j = i \\ T_j & w.p \quad q_{ij}/f_i \quad \forall j \neq i. \end{cases}$$

Taking LSTs on both sides, we get

$$E(e^{-sT_i}) \;\; = \;\; \frac{f_i}{s + f_i} * [\frac{\theta_i}{f_i} + \frac{\sum_{j \neq i} q_{ij} E(e^{-sT_j})}{f_i}].$$

Arranging terms, we have

$$E(e^{-sT_i}) \;\; = \;\; \frac{\theta_i}{s + f_i} + \frac{\sum_{j \neq i} q_{ij} E(e^{-sT_j})}{s + f_i}. \tag{3}$$

Taking derivative of Equation (3) with respect to s, and substituting $s = 0$, we have

$$f_i E(T_i) - \sum_{j \neq i} q_{ij} E(T_j) \;\; = \;\; 1. \tag{4}$$

This is identical to Equation (1). By solving Equation (4), we get $E(T_i)$ $\forall$ $i = 1, \; 2, \; ..., \; n$, where n is the number of states of the environment process. Taking the second derivative of the LST in Equation (3) w.r.t s, and substituting $s = 0$, we get

$$f_i^2 E(T_i^2) - f_i \sum_{j \neq i} q_{ij} E(T_j^2) \;\; = \;\; 2(1 + \sum_{j \neq i} q_{ij} E(T_j)). \tag{5}$$

Solving Equation (5), we get $E(T_i^2)\ \forall\ i = 1,\ 2,\ ...,\ n$.

∎

Now, in order to derive the unconditional first and second service time moments, $E(T)$ and $E(T^2)$ respectively, we use

$$E(T) = \sum_{i=1}^{n} \hat{\pi}_i E(T_i), \tag{6}$$

$$E(T^2) = \sum_{i=1}^{n} \hat{\pi}_i E(T_i^2) \tag{7}$$

where $\hat{\pi}_i$ is the probability that a customer sees the environment process in state i at the beginning of service.

For Equations 6 and 7 we need to obtain an expression for $\hat{\pi}_i$. Our conjecture is that $\hat{\pi}_i$ is related to p_i, which is the steady state probability that the environment is in state i. The p_i's can be obtained by solving $[p_i]Q = 0$ and $\sum_i p_i = 1$. We devote the next section to obtaining $\hat{\pi}_i$, the probability that a customer's service begins when environment process in state i.

3.2 Computing $\hat{\pi}_i$

Upon running extensive simulations (as shown in Section 3.3.1), we observe that $\hat{\pi}_i$ is not only a function of p_i but it also depends on λ, the arrival rate. This is indeed an intriguing result, as the service time is in fact a function of the arrival rate! Regrettably, we find that it is intractable to obtain a closed-form expressions of $\hat{\pi}_i$ (except for some special cases, which we will show next) in terms of λ. We now analytically illustrate how $\hat{\pi}_i$ is different for two special cases $\lambda \to 0$ and $\lambda \to \infty$. Later in Section 3.3.1, we suggest that we could use one of the special cases as an approximation.

3.2.1 Special Case 1: Arrival rate approaching zero

For this special case of $\lambda \to 0$, we use the notation ${\hat{\pi}_i}^0$ for $\hat{\pi}_i$, i.e.

$${\hat{\pi}_i}^0 = \lim_{\lambda \to 0} \hat{\pi}_i.$$

The following theorem derives an expression for ${\hat{\pi}_i}^0$.

Theorem 2 *In the asymptotic case of $\hat{\pi}_i$ when $\lambda \to 0$, ${\hat{\pi}_i}^0 = p_i$, where p_i steady state probability that the environment is in state i.*

Proof.

When $\lambda \to 0$, we can say that each arrival will see the queue empty w.p. 1 as the inter-arrival times are very large. Therefore for the customers, service will begin as soon as an arrival takes place. In addition, due to PASTA, an arriving customer will see the environment process $\{Z(t),\ t \geq 0\}$ in state i w.p. p_i. Hence, service will begin when environment process is in state i w.p. p_i. Therefore, ${\hat{\pi}_i}^0 = p_i$.

∎

Remark 1: From Theorem 2, when $\lambda \to 0$, $E(T) = \sum_{i=1}^{n} p_i E(T_i)$.

3.2.2 Special case 2: Arrival rate approaching infinity

Let us consider another asymptotic case in which the rate of arrivals approaches infinity. Since $\lambda \to \infty$, we can also say that the number in queue waiting for service is infinite, or $X(t) = \infty$. This means that the server would serve non-stop. For this special case of $\lambda \to \infty$, we use the notation $\hat{\pi}_i{}^\infty$ for $\hat{\pi}_i$, i.e.

$$\hat{\pi}_i{}^\infty = \lim_{\lambda \to \infty} \hat{\pi}_i.$$

We now present a theorem that relates $\hat{\pi}_i{}^\infty$ to p_i.

Theorem 3 *The relationship between the service start probabilities $\hat{\pi}_i{}^\infty$ and stationary probabilities p_i is given by $\hat{\pi}_i{}^\infty = p_i\theta_i / \sum_{j=1}^n p_j\theta_j$, where θ_i is the rate of service completion when the environment process is in state i.*

Proof.

Let S_m denote the time when the mth event occurs where the mth event can be either a state change in the environment process or a service completion. Let $Y_m = Z(S_m+)$ be the state of the environment process just after the mth event took place. Consider a Markov regenerative sequence $\{(Y_m,\, S_m),\, m \geq 0\}$ from which we can build a semi Markov process (SMP). Therefore, we have

$$P\{S_n > t | Y_n = i\} = e^{-(\theta_i + q_i)t},$$

where $q_i = \sum_{j \neq i} q_{ij} = -q_{ii}$. The kernel of the SMP is

$$G_{ij}(t) = P\{Y_{n+1} = j, S_n \leq t | Y_n = i\} \quad = \quad \begin{cases} \frac{q_{ij}}{q_i + \theta_i}(1 - e^{-(q_i + \theta_i)t}) & if \quad j \neq i, \\ \frac{\theta_i}{q_i + \theta_i}(1 - e^{-(q_i + \theta_i)t}) & if \quad j = i. \end{cases}$$

The steady state probability that the SMP defined above is in state i is same as that of the environment process. Hence we have p_i as the probability that the SMP is in state i in steady state. Let $G(\infty) = [G_{ij}(\infty)]$. What is unknown at this time is $\triangle_i$, the probabilities that the SMP is in state i at the beginning of an epoch. It can be obtained by solving $\triangle = \triangle G(\infty)$ and $\sum \triangle_i = 1$. However, what we are interested is in the relation between $\triangle_i$ and p_i. Using SMP results (see Kulkarni [15]), we have

$$p_i = \frac{\triangle_i / (\theta_i + q_i)}{\sum_j \triangle_j / (\theta_j + q_j)}.$$

Notice that a fraction $\frac{\theta_i}{\theta_i + q_i}$ of state changes in the SMP correspond to a service completion. Since there are infinite customers in the queue, beginning of service of a new customer starts immediately after a service completion. So, the fraction $\frac{\theta_i}{\theta_i + q_i}$ of state changes in the SMP correspond to a new customer beginning service. Therefore, we have $\hat{\pi_i^\infty} = \triangle_i(\frac{\theta_i}{\theta_i + q_i})$. Hence, $p_i = \frac{\hat{\pi_i^\infty}/\theta_i}{\sum_j \hat{\pi_j^\infty}/\theta_j}$, i.e. $p_i \propto \hat{\pi_i^\infty}/\theta_i$ or $\hat{\pi_i^\infty} \propto p_i\theta_i$. Therefore, by normalizing we have

$$\hat{\pi_i^\infty} = p_i\theta_i / \sum_{j=1}^n p_j\theta_j.$$

■

Remark 2: From Theorem 3, when $\lambda \to \infty$, $E(T) = \sum_{i=1}^{n} p_i\theta_i E(T_i) / \sum_{j=1}^{n} p_j\theta_j$. Using result from point processes it is possible to show that $E(T)$ will converge to the reciprocal of the time-average service rate. This will be addressed in Section 4.3 when we discuss time-average service rate.

Remark 3: Clearly from Remarks 1 and 2, $E(T)$ depends on the choice of λ.

3.3 Simulation experiments, approximations and numerical examples

In Section 3.2, we derived $\hat{\pi}_i$ for special cases of arrival rates (namely $\lambda \to 0$ and $\lambda \to \infty$). We ran simulation experiments to study the effect of arrival rate λ on the average service time. Further, we investigate good approximations for $\hat{\pi}_i$ for any general arrival rate. In Section 3.3.1, we discuss the simulation results and in Section 3.3.2, we provide some more interesting numerical results for the examples given in Sections 2.1 and 2.2.

3.3.1 Simulation results

Simulation runs were conducted to observe the trend of the first and second moments of service time with respect to variations in the inter-arrival time of customers in to the system. A plot with interarrival time (in appropriate units of time) on X-axis and average service time on Y-axis is shown in Figure 2. The observations from the simulation are as follows. For the run when interarrival time is just more than the average service time obtained for case $\lambda \to \infty$, it was observed that the queue is stable, and the average service time is approximately the same as that of $\lambda \to \infty$. When inter-arrival times are increased, it was observed that the average service time also increases but not as much as inter-arrival times. When inter-arrival times are very large, the average service times almost remain unchanged and asymptotically converge to the average service time obtained for the case $\lambda \to 0$ as seen in Figure 2. For Figure 2, $E(T)$ when $\lambda \to \infty$ is 0.307 sec and $E(T)$ when $\lambda \to 0$ is 0.360 sec. For large λ, the $E(T)$ corresponding to $\lambda \to \infty$ would be a good approximation. We will revisit this again while discussing stability condition in Section 4.3.

3.3.2 Numerical results

Based on the findings in Section 3.3.1, we know that for low inter-arrival times, $\hat{\pi}_i^{\infty}$ can be used as an approximation for $\hat{\pi}_i$. So, in this Section, we make use of that approximation to have some insights on the service process. Using Equations 6 and 7, we obtain first and second moments of the service time as 0.3069 and 0.3081 respectively for application 1 in Section 2.1 as well as 0.3640 and 0.3298 respectively for application 2 in Section 2.2. Also, the mean and variance of service time increase with $1/\mu$, the mean amount of work brought in by a customer. However, it is not clear how the service process is affected by the ratio of Q versus λ. To address this, we consider cases when the $Z(t)$ process changes quicker and slower than the $X(t)$ process. This is done when the magnitude of all the values in matrix Q are increased and reduced respectively. Let us say that the factor of magnification of Q matrix be m, i.e. the new Q matrix is mQ. In Table 2, we vary m from 0.001 to 1000 and illustrate this. As $m \to 0$ and $m \to \infty$, the second moment stabilizes to two values. The first moments (mean) of service time for examples in Sections 2.1 and 2.2 are

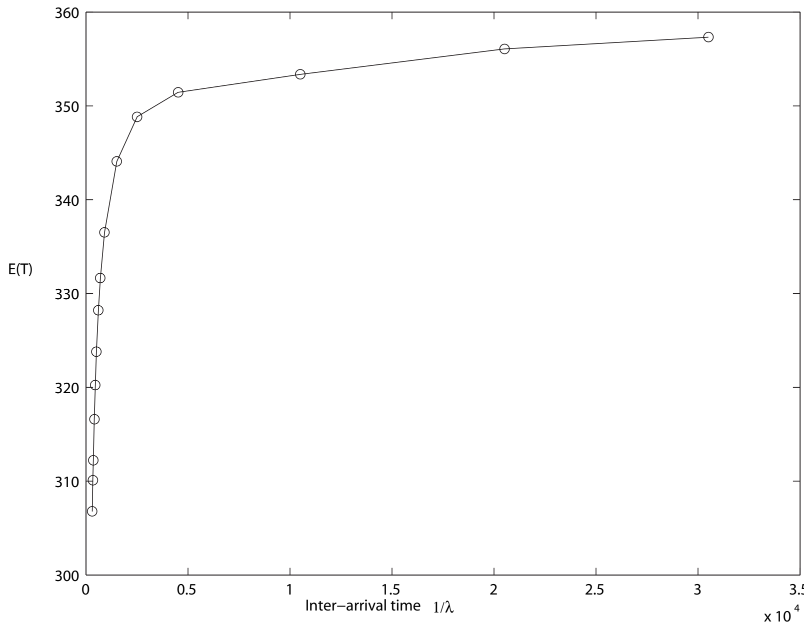

Figure 2: Interarrival time vs Average service time

respectively 0.3069 and 0.3640 for all the cases. It is seen that as m increases, the second moment (and thereby variance) of the service time decreases.

Table 2: Second moment of service time

m	0.001	0.01	0.1	1	10	100	1000
Second moment (Application 1)	0.4527	0.4219	0.3081	0.2193	0.1928	0.1888	0.1884
Second moment (Application 2)	0.3651	0.3603	0.3298	0.2813	0.2669	0.2652	0.2650

4 Performance Measures: Average Delay and Queue Length

Having derived the moments of service time, we now concentrate on obtaining system performance measures such as mean delay in this section (average queue length can be easily computed using Little's law). For this, we need to analyze the stochastic process $\{(Z(t), X(t)), t \geq 0\}$, a two dimensional CTMC, which is also a Quasi-birth-death (QBD) process. It is not possible to use generating function or other such techniques to obtain closed form expression for steady state probabilities. Additionally, we cannot model the $X(t)$ process as an $M/G/1$ queue and use the first and second moments of service time, as the service times are not iid. Hence we resort to Matrix Geometric method (MGM), which we will explain in detail. Since QBD has a special structure, MGM can be used to obtain the desired performance measure. To understand the QBD process, we review the concept of birth-death process first and then QBD process (Section 4.1), which will be useful while applying MGM. The general theory of MGM and how it is applied to our problem is explained in Section 4.2. Although we understand that MGM is a mature tool, we feel that a brief explanation would ensure smoother reading.

4.1 Quasi-birth and death process

In order to understand Quasi-birth-death process, we will start with the definition of birth-death process as per Thorne [27].

Definition: The continuous-time Markov process $\{Y(t) : t \geq 0\}$ *is a birth-death process if the only two possible transitions are* $n \to n+1$ *with birth rates* $q(n, n+1), n \geq 0$*, and* $n \to n-1$ *with death rates* $q(n, n-1), n \geq 1$*. If* S *is the state space, the infinitesimal generator matrix for a birth-death process is as follows:*

$$\begin{pmatrix} -q(0,1) & q(0,1) & 0 & 0 & 0 & \dots \\ q(1,0) & -q(1) & q(1,2) & 0 & 0 & \dots \\ 0 & q(2,1) & -q(2) & q(2,3) & 0 & \dots \\ 0 & 0 & q(3,2) & -q(3) & q(3,4) & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where $q(j) = q(j, j-1) + q(j, j+1)\ \forall j \,\epsilon\, S$ *and j>0.*

The above process cannot skip adjacent states and hence it is called “skip free” in the states. Birth-death process is a special case of more general class called Quasi-birth-death process. As per Latouche et al [17], the definition for quasi birth-death process is as follows.

Definition: A continuous time Quasi-Birth-Death (QBD) process is a continuous time Markov process whose infinitesimal generator matrix is of the block partitioned form

$$Q_{QBD} = \begin{pmatrix} B_1 & A_0 & 0 & 0 & 0 & \dots \\ A_2 & A_1 & A_0 & 0 & 0 & \dots \\ 0 & A_2 & A_1 & A_0 & 0 & \dots \\ 0 & 0 & A_2 & A_1 & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where A_0, A_1, A_2 *and* B_1 *are* $n \times n$ *matrices.*

After partitioning the states into subsets called levels, such that a position within the level is termed as phase. The process can jump down one level, stay in the same level or jump up one level, and the rate that these transitions occur are given by A_2, A_1 and A_0 respectively. The process is said to be skip free between levels.

4.2 Matrix Geometric Method for QBD Analysis

There exists a matrix geometric relation among the stationary probabilities of the $\{(Z(t), X(t)), t \geq 0\}$ process. In matrix geometric method, an auxiliary matrix called R is used in the calculation of stationary probabilities and other measures of interest like waiting time and mean queue length. The main computational effort is in obtaining R, which is frequently done numerically. For a level independent infinite level CTMC, R is obtained as follows.

Computation of matrix R: The matrix R has a following quadratic relation (Ramaswami [21], Riska [22]):

$$A_0 + RA_1 + R^2 A_2 = 0. \tag{8}$$

Equation (8) has to be used recursively to solve for R. There is a lot of research going on to find R efficiently but we will not go into those details in this paper. Neuts [19] defines infinite-state Markov chains with a repetitive structure with state space partitioned into the boundary states $S^{(0)} = \{s_0^1, ..., s_0^n\}$ and a set of states $S^{(i)} = \{s_i^1, ..., s_i^n\}$ $\forall i \geq 1$, that correspond to the repetitive portion of the chain. Let $\pi^{(i)}$ be the steady state probability vector of states $S^{(i)}$. Then

$$\pi^{(i)} = \pi^{(1)} R^{i-1} \;\; \forall \; i \geq 1.$$

Solving $\pi Q_{QBD} = 0$ will give both $\pi^{(0)}$ and $\pi^{(1)}$. The following set of equations are obtained:

$$\begin{aligned} \pi^{(0)} B_0 + \pi^{(1)} A_2 &= 0, \\ \pi^{(0)} A_0 + \pi^{(1)} (A_1 + RA_2) &= 0, \\ \pi^{(0)} e + \pi^{(1)} (I - R)^{-1} e &= 1. \end{aligned} \tag{9}$$

Once $\pi^{(0)}$ and $\pi^{(1)}$ are obtained, the expected waiting (including service) time of a job in the system can be calculated as follows:

$$W = \lambda^{-1}(\pi^{(1)}(I - R)^{-1} e + \pi^{(1)} R (I - R)^{-2} e) \tag{10}$$

where W is the average delay or holding time and e is the column vector of ones. This is how average delay can be computed.

Now, we resort to our QBD process $\{(Z(t),\, X(t)),\, t \geq 0\}$, which indeed is a CTMC. Recall that the environment process $\{(Z(t), X(t)), t \geq 0\}$ is a CTMC with q_{ij}, the rate of transitioning from state i to state j. Also λ is the arrival rate of requests and θ_i is the rate of completion of requests when there are i requests. The following theorem derives A_0, A_1, A_2, B_1 matrices for our QBD process in terms of λ, θ_i and q_{ij}.

Theorem 4 *The structure of the infinitesimal generator for this QBD problem is shown below:*

$$\mathbf{A_0} = \begin{pmatrix} \lambda & 0 & 0 & \dots & 0 \\ 0 & \lambda & 0 & \dots & 0 \\ 0 & 0 & \lambda & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \lambda \end{pmatrix} \qquad \mathbf{A_1} = \begin{pmatrix} s(1) & q_{12} & q_{13} & \dots & q_{1n} \\ q_{21} & s(2) & q_{23} & \dots & q_{2n} \\ q_{31} & q_{32} & s(3) & q_{34} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ q_{n,1} & q_{n,2} & \dots & q_{n,n-1} & s(n) \end{pmatrix}$$

where $s(i) = -\sum_{j\neq i} q_{ij} - \lambda - \theta_i$ *and*

$$\mathbf{A_2} = \begin{pmatrix} \theta_1 & 0 & 0 & \dots & 0 \\ 0 & \theta_2 & 0 & \dots & 0 \\ 0 & 0 & \theta_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \theta_n \end{pmatrix} \qquad \mathbf{B_1} = \begin{pmatrix} u(1) & q_{12} & q_{13} & \dots & q_{1n} \\ q_{21} & u(2) & q_{23} & \dots & q_{2n} \\ q_{31} & q_{32} & u(3) & q_{34} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ q_{n,1} & q_{n,2} & \dots & q_{n,n} & u(n) \end{pmatrix}$$

where $u(i) = -\sum_{j\neq i} q_{ij} - \lambda$.

Proof.

Consider the CTMC $\{(Z(t),\, X(t)),\, t \geq 0\}$. Rearranging the states of the CTMC suitably, we can write down its generator matrix in QBD form. From that matrix, we can obtain A_0, A_1, A_2, B_1 matrices as given in the theorem.

■

Note that $\{(Z(t),\, X(t)),\, t \geq 0\}$ is a level independent infinite-level QBD process. Hence, substituting the above matrices A_0, A_1, A_2, B_1 in the set of Equations (9), we get $\pi^{(0)}$ and $\pi^{(1)}$. Then, from Equation (10), we obtain the average delay. The average delay assumes that the QBD is ergodic. We now state the condition for stability for the $\{(Z(t),\, X(t)),\, t \geq 0\}$ process with the understanding that $\{Z(t),\, t \geq 0\}$ is ergodic.

4.3 Stability condition for the system

Note that the average service rate offered by the queue is $\mu E(b_{Z(t)})$. The following theorem states the stability condition.

Theorem 5 *The necessary and sufficient condition for the stability of the queue with queue length process* $\{X(t),\, t \geq 0\}$ *is*

$$\frac{\lambda}{\mu E[b_{Z(t)}]} < 1,$$

where $E[b_{Z(t)}]$ *is the average service speed of the server.*

Proof.

The proof follows from Núñez-Queija [20], where $\{Z(t),\, t \geq 0\}$ is a birth and death process. To extend the analysis in Núñez-Queija [20] to the general case of $\{Z(t),\, t \geq 0\}$ is straightforward.

■

Remark 4: Numerical investigations reveal that the MGM stability condition given in [22] is same as given in the above theorem. The MGM stability condition can only be checked numerically. This provides a closed form expression that can be used for the analysis of stability.

Recall that $E(T)$ is the average service time and it is a function of λ. One of the curious questions that comes to mind is the relationship between $E(T)$ and $\mu E(b_{Z(t)})$.

As discussed in Remark 2, as $\lambda \to \infty$

$$E(T) \to \frac{1}{\mu E(b_{Z(t)})}.$$

Therefore, we have,

$$\frac{1}{\mu E(b_{Z(t)})} = \sum_{i=1}^{n} p_i \theta_i E(T_i) / \sum_{j=1}^{n} p_j \theta_j.$$

Remark 5: However, based on simulations,

$$E(T) < \frac{1}{\mu E(b_{Z(t)})}.$$

So, if $\lambda < \mu E(b_{Z(t)})$ (i.e. stability condition), then automatically $\lambda < \frac{1}{E(T)}$. Therefore stability condition is indeed accurate. The reason the mean service rate is not reciprocal of the mean service time is because the former is averaged over time (and includes server idle periods when the environment process is not stochastically identical to when the service is busy) whereas the mean service time is averaged over customers, i.e. when the server is busy.

4.4 Numerical Examples

We consider the numerical examples described in Section 2.1 and 2.2. Some of the obvious results we obtain are: the average waiting time increases with λ, decreases with μ and decreases with $E[b_{Z(t)}]$. However, it is not clear how the waiting times are affected based on the relative frequency of change in the $X(t)$ and $Z(t)$ processes. Similar to Section 3.3, we multiply Q by a factor m and change m to reflect the relative frequency of change in $X(t)$ and $Z(t)$. If the matrix is scaled up i.e. large m, it means that the $Z(t)$ process changes states faster and if Q is scaled down, $Z(t)$ process changes occasionally both in comparison to $X(t)$. In table 3, we vary m from 0.001 to 1000 for both the examples. It can be seen that as m decreases the waiting increases drastically. Higher m implies higher rate of change in $Z(t)$ process. However, as $m \to \infty$, the average waiting time seems to converge to the M/M/1 waiting time (with average service time $1/\mu E[b_{Z(t)}]$).

Table 3: Average waiting time

m	0.001	0.01	0.1	1	10	100	1000
Average waiting time (e.g. 1)	342.50	42.03	11.72	4.82	3.96	3.88	3.87
Average waiting time (e.g. 2)	76.99	22.51	6.90	4.35	4.07	4.05	4.04

5 Large Deviations Analysis for Tail Distributions

In the previous section, we used MGM method to obtain the mean delay W and thereby mean number in the system (via Little's law, λW). The next question is whether it would be possible to obtain the distribution of any of the performance measures. As it turns out, in the 1990's several researchers (Elwalid and Mitra [10], Kelly [13], Kesidis et al [14], Courcoubetis and Weber [8]) used the method of large deviations to obtain tail distributions of the system workload. In that spirit,

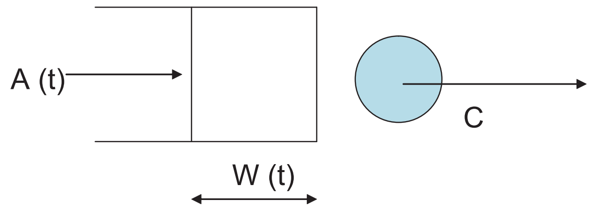

Figure 3: Single Buffer Model

we will take advantage of the theory of large deviations, specifically using effective bandwidth of the traffic, for our analysis to obtain the tail distribution of the workload in the queue under steady state. One of the advantages of this methodology is that it can be easily extended to the non-exponential case, which we will explain later. We first state some preliminary results that would be useful for the reader in terms of the large-deviations principle.

5.1 Preliminaries

Consider a queueing system where a stream of customers arrive into a queue according to some random process and each customer brings some random amount of work. Let $A(t)$ be the total amount of work that arrives into the queue over time $(0, t]$. Note that $A(t)$ is a random variable. In fact, the arrivals need not be discrete, but also fluids so that $A(t)$ is the amount of fluid that arrives in time $(0, t]$. The server works at a constant speed c, so that whenever there is work queued up, it exits at rate c. The single buffer fluid model is shown in figure 3. Care must be taken in fluid arrivals, work load does not jump whenever an arrival occurs, but increases steadily. This is not an issue for discrete arrivals. This fact is shown in figure 4. Let $W(t)$ be the amount of workload in the system at time t. The system is stable (i.e. $W(t) \to W(\infty)$ as $t \to \infty$) if

$$\lim_{t\to\infty} \frac{A(t)}{t} < c.$$

The theory of large deviations can be used to obtain the tail distribution of the random variable $W(\infty)$. This requires the use of the effective bandwidth concept. The effective bandwidth of the input stream of traffic is defined in terms of $A(t)$ as

$$eb(v) = \lim_{t\to 0} \frac{1}{vt} \log E[e^{vA(t)}].$$

According to the theory of large deviations, for large values of u,

$$P\{W(\infty) > u\} \approx e^{-\eta u}, \tag{11}$$

where η is the unique solution (if the system is stable) to

$$eb(\eta) = c. \tag{12}$$

Note that in order to use the above result, the server speed c must be a constant at all times. However, in our problem the server speed changes. We will see in the next section how to use a

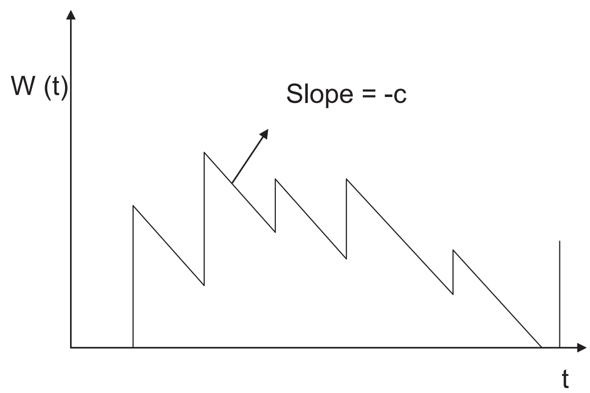

Figure 4: Workload Sample path for discrete queues

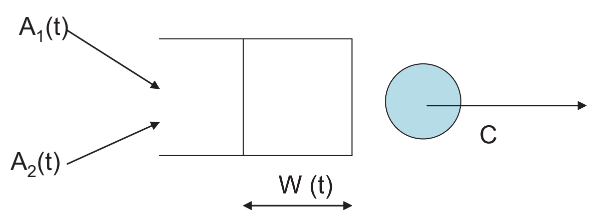

Figure 5: Single buffer with two sources

compensating source so that we create an equivalent system where the server speed can be kept a constant. To use the large deviations result for this system we need another result. Consider two independent streams of traffic that enter a queue with server speed c. Let $A_1(t)$ and $A_2(t)$ be the total amount of work that arrives into the queue over time $(0, t]$ from the two streams as shown in figure 5. For $i = 1,\ 2$, let the effective bandwidth of traffic stream i be

$$eb_i(v) = \lim_{t \to 0} \frac{1}{vt} \log E[e^{vA_i(t)}].$$

Then, according to the theory of large deviations, for large values of u,

$$P\{W(\infty) > u\} \approx e^{-\eta u}, \tag{13}$$

where η is the unique solution (if the system is stable) to

$$eb_1(\eta) + eb_2(\eta) = c. \tag{14}$$

In the next subsection we will use all the results here for our problem defined in Section 2.

5.2 Tail Distributions Using a Compensating Source

We now use the large deviation results based on effective bandwidths from the previous subsection. First consider the queueing system in Section 2. Customers arrive according to a Poisson process (with mean rate λ). We first consider the general case where each customer brings a random amount of work, independent and identically distributed as other customers such that the CDF of the amount of work is $G(\cdot)$. Subsequently we will consider the case on Section 2, namely each customer bringing $\exp(\mu)$ work with them. Let W_i be the amount of work brought by the i^{th} customer, $G(u) = P\{W_i \leq u\}$ with LST $\tilde{G}(s) = E[e^{-sW_i}]$. Let $N(t)$ be the number of arrivals in time $(0, t]$. In addition, $A(t) = W_1 + W_2 + ... + W_{N(t)}$. In order to compute the effective bandwidth, we first derive $E[e^{vA(t)}]$. We have,

$$\begin{aligned} E[e^{vA(t)}] &= \sum_{i=0}^{\infty} E[e^{vA(t)}|N(t) = i]P(N(t) = i) \\ &= e^{-\lambda t} + \sum_{i=1}^{\infty} [\tilde{G}(-v)]^i e^{-\lambda t} \frac{(\lambda t)^i}{i!} \\ &= e^{-\lambda t} + [e^{\lambda t \tilde{G}(-v)} - 1]e^{-\lambda t} \\ &= e^{\lambda t(\tilde{G}(-v)-1)}. \end{aligned}$$

The first equation above is by conditioning on the number of arrivals in time $(0, t]$ and the second equation is due to the definition of LST. Therefore, the effective bandwidth of this stream of traffic is $eb(v) = \frac{\lambda}{v}[\tilde{G}(-v) - 1]$.

Now, for the special case when the amount of work is exponentially distributed with mean $1/\mu$, we have $G(u) = 1 - e^{-\mu u}$ and $\tilde{G}(-v) = \frac{\mu}{\mu - v}$. So, $eb(v) = \frac{\lambda}{v}[\frac{\mu}{\mu - v} - 1] = \frac{\lambda}{\mu - v}$. If the server speed is a constant speed, in particular for the standard $M/M/1$ queue, $c = 1$. Thus $eb(\eta) = c = 1$ results in $\eta = \mu - \lambda$. Hence for large u, $P\{W(\infty) > u\} = e^{-(\mu-\lambda)u}$, which in fact, is true for all u from standard $M/M/1$ results.

In our system considered in Section 2, the server speed c changes with time. However in order to use the large deviations results we need c to be a constant. A technique that can be used when the service speed is not constant is to have a compensating stream of traffic such that the $W(t)$ process is unchanged and c is a constant. Using that, we can state the following theorem.

Theorem 6 *The tail distribution (for large u) for the amount of work in the queue in Section 2 is:*

$$P\{W(\infty) > u\} = e^{-\eta u}.$$

where η is obtained by solving

$$e(\bar{R} + Q/\eta) + \frac{\lambda}{\mu - \eta} = \max_i b_i,$$

where $e(A)$ is the largest eigen value of matrix A and $\bar{R}$ is $diag(c - b_i)$.

Proof.

Consider the following two systems (the first is what we described in Section 2 and the second is a fictitious one):

- System 1: A single stream of traffic where customers arrive according to a Poisson process with mean rate λ and each customer brings an $\exp(\mu)$ amount of work. The server serves at different rates according to an environment process $\{Z(t), t \geq 0\}$ which is a CTMC with generator matrix Q. At time t the service speed c is $b_{Z(t)}$.

- System 2: A fictitious queueing system where there are two input streams and a server with a constant speed $c = \max_i b_i$. The first input stream is one where fluid enters the queue at rate $c - b_{Z(t)}$ at time t. The second input stream is the usual one where customers arrive according to a Poisson process with mean rate λ and each customer brings an $\exp(\mu)$ amount of work.

The amount of work remaining at time t in both systems are identical for all t. This is because at time t when the server speed is $b_{Z(t)}$ in System 1, in System 2 the server speed is constant at c and an additional $c - b_{Z(t)}$ work flows into the system nullifying the extra $c - b_{Z(t)}$ capacity that is available. In order to do this we require $c = \max_i b_i$ so that $c - b_{Z(t)} \geq 0$ for all t. Since both System 1 and System 2 have the same workload at all times, we study System 2 using large deviations with the understanding that the tail distributions of the work load would be identical to that of System 1.

In order to use the large deviations result for System 2, we need effective bandwidths of both streams. For the first stream, let $\bar{R}$ be a diagonal rate matrix such that, $\bar{R}_i = diag(c - b_i)$. Note that the rows of the R matrix need to correspond to the rows of the Q matrix. Let $e(M)$ denote the largest real-eigen value of a square matrix M. Then the effective bandwidth of the first stream is $eb(v)$=$e(\bar{R} + Q/v)$ (due to a result in Elwalid and Mitra [10]). From the previous subsection, the effective bandwidth of the second stream is $\lambda/(\mu - v)$. Using Equation (14), we obtain η by solving

$$e(\bar{R} + Q/\eta) + \frac{\lambda}{\mu - \eta} = \max_i b_i. \tag{15}$$

Therefore the tail distribution (for large u) for the amount of work in the queue of System 1 is:

$$P\{W(\infty) > u\} = e^{-\eta u}. \tag{16}$$

■

Recall that we had earlier mentioned that the amount of work a customer brings need not be exponential, we can still obtain the effective bandwidth. In fact the $\{Z(t), t \geq 0\}$ process need not even be a CTMC. If the input process is a semi-Markov process (See Gautam et al [11]) or a Markov regenerative process (See Kulkarni [16]), computing effective bandwidth is possible. Therefore these tail probabilities can be applied to a much more general setting than what is considered here. One of the main applications of tail distributions is to obtain overflow probabilities in buffers. These are illustrated in the following subsection.

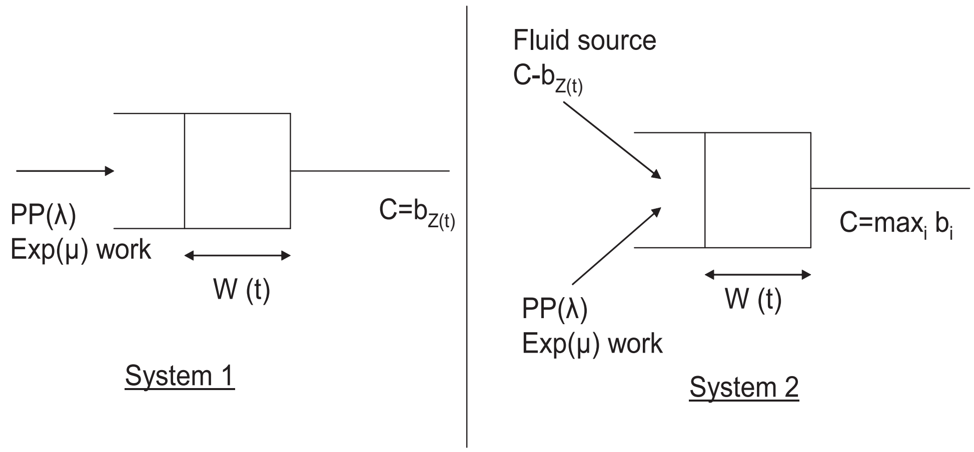

Figure 6: Representation of systems 1 and 2

5.3 Numerical Example

Consider the example in Section 2.1 where mean arrival rate is $\lambda = 3$ and work to be done is exponentially distributed with parameter $\mu = 8$. The total bandwidth $S = 0.650$. The two available bandwidths are $r_1 = 0.265$ and $r_2 = 0.350$. We show the steps involved only for the one in Section 2.1 for illustration purposes (as a very similar process is to be followed for the example in Section 2.2).

Using the b values in Section 2.1, $c = \max(b_{(0,0)}, b_{(1,0)}, b_{(2,0)}, b_{(0,1)}, b_{(1,1)}) = 0.650$ and

$$\bar{R} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0.265 & 0 & 0 & 0 \\ 0 & 0 & 0.530 & 0 & 0 \\ 0 & 0 & 0 & 0.350 & 0 \\ 0 & 0 & 0 & 0 & 0.615 \end{bmatrix}$$

Using Q matrix in Section 2.1 and $\bar{R}$, solving for η in Equation (15), we get $\eta = 6.3081$. For large u, we use Equation 16 to obtain $P(W(\infty) > u) = e^{-6.3081u}$. Thus, we can find tail distribution of workload in system in steady state. For example, the probability that the work load exceeds 2 (i.e. $u = 2$), is approximately 3.32×10^{-6}.

6 Concluding Remarks

6.1 Conclusions and Extensions

We consider a queue where the service rate changes over time according to a changing environment process that is governed by a Markov process. To get a grasp on the service process, we derive equations to find the first and second moment of the service time using Laplace-Stieltjes transforms. We obtain a curious result in that the mean service time is not equal to the reciprocal of the mean service rate. In fact, the mean and variance of service time depend on the arrival rate.

The performance measures, mean waiting time and average queue length, are calculated using matrix geometric method. In the analysis, we use a combination of fluid traffic as well as a discrete traffic to obtain work load distribution, thereby overflow probabilities. The analysis is illustrated using numerical examples wherever appropriate. We used simulations to cross check our analytical results everywhere. It was observed that the second moment of service time is higher and in turn, the average waiting time is higher if the environment process varies slowly with respect to queue length keeping the mean service time constant.

In this paper, we have assumed FCFS service discipline throughout. This can be extended to other service disciplines like processor sharing directly. The transition diagram of the bivariate stochastic process remains the same and so most of the analysis is still valid for other service disciplines. The performance measures like average queue length and average time in system (delay) obtained from matrix geometric method will be exactly the same. Even if the elastic requests are served in parallel, the infinitesimal generator matrix Q is identical to the one if the elastic requests are served in FCFS fashion. That is, the transition diagram of the FCFS is exactly the same for both cases. Likewise, the large deviations results would be unaffected as the results only calculate the total workload. All work-conserving schemes would yield identical results.

6.2 Future Work

In section 4.2, we compute average waiting time of the customers in system. In section 3, we derive equations to obtain first and second moments of service time of the customers. Obtaining the distribution or at least second moment (thereby variance) of the waiting time in system, in addition to average waiting time, would be much more helpful in understanding the system. Thus, one of the future tasks would be to compute the second moment of the waiting time. The eventual goal is to obtain the distribution of waiting time.

In future, we will consider extending the results in this paper to the case of non-exponential systems. In particular, the arrival process would be renewal, work brought by each arrival is from a general distribution, and/or the environment process $\{Z(t), t \geq 0\}$ is a more general process such as a Markov regenerative process. Although we indicated in Section 6.1 that these extensions are straight forward for large deviations case, the MGM analysis can be done only under special circumstances such as Phase type distributions.

Other applications in the future include solving design and control problems. For example, consider the application in Section 2.1. For the web server, it is important to design the number of each type of streaming request to admit (i.e. $n_1, n_2, ...$). To solve such a problem, a mathematical optimization model can be formulated such that by solving it, we can design a web server system. In addition, we can study adaptive control systems in CPU application in Section 2.2. In particular, the agent can decide (based on CPU load) when to accept tasks and when to send tasks to the CPU. The authors are in the process of investigating some of the above aspects.

Acknowledgements

The research is partially supported by NSF Grants ANI-0219747 and ACI-0325056. The authors thank the editor and anonymous reviewers for their comments and suggestions that led to considerable improvements in the content and presentation of this paper. The authors are grateful to Prof. George Kesidis, Prof. Raj Acharya, Prof. Anand Sivasubramaniam, Dr. Donna Ghosh, Prof. Venkatesh Sarangan for initial discussions that led to the formulation of this research. The authors deeply indebted to Prof. Takis Konstantopoulos and Prof. Karl Sigman for their valuable input in the analysis of the paper.

References

[1] Adan,I. J. B. F., Kulkarni, V. G., 2003, "Single server queue with Markov dependent inter-arrival and service Times", Queueing Systems, 45(2), pp. 113-134.

[2] Ahn, S., and Jeon, J., 2000, "Large deviation results for queues with Continuous time Markov modulated Arrivals", Engineering Simulation, vol. 18, no. 1, pp 67-74.

[3] Baykal-Gursoy, M., and Xiao, W., 2004, "Stochastic decomposition in $M/M/\infty$ queues with Markov modulated service rates", Queueing Systems, 48, pp. 75-88.

[4] Bonald, T., and Proutiere, A., 2003, "Insensitive bandwidth sharing in data networks", Queueing Systems, vol. 44, pp. 69-100.

[5] Boxma, O. J., and Kurkova, I. A., 2001, "The $M/G/1$ queue with two service speeds", Adv. Applied Probability, 33, pp. 520-540.

[6] Cidon, I., Gu'erin, R., Khamisy, A., and Sidi, M., 1991, "On Queues with inter-arrival times proportional to service times", Technion, EE PUB, (811).

[7] Collings, T., and Stoneman, C., 1976,"The $M/M/\infty$ queue with varying arrival and service rates", Operations Research 24, 760-773.

[8] Courcoubetis, C., and Weber, R., 1995, "Effective bandwidths for stationary sources", Prob. Eng. Inf. Sci., vol. 9, pp. 285-296.

[9] Dorman, K., karindorman.stat.iastate.edu/stat432/2003-09-12.pdf.

[10] Elwalid, A. I., and Mitra, D., 1993, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks", IEEE/ACM Transactions on Networking, vol. 1, no. 3, pp. 329-343.

[11] Gautam, N., Kulkarni, V. G., Rolski, T., and Palmowski, Z., 1999, "Bounds for fluid models driven by semi-Markov inputs", Probability in Engineering and Informational Sciences, Vol. 13, No. 4, pp. 429-475.

[12] Kaufman, J. S., 1981, "Blocking in shared resource environment", IEEE Transactions on Communications, vol. 29, no. 10, pp. 1494-1481.

[13] Kelly, F. P., 1996, Notes on Effective bandwidth, Stochastic Networks: Theory and Applications, Oxford University Press.

[14] Kesidis, G., Walrand, J., and Chang, C-S., 1993, "Effective bandwidths for multiclass Markov fluids and other ATM sources", IEEE/ACM Transactions on Networking, vol. 1, no. 3, pp. 424-428.

[15] Kulkarni, V. G., 1995, Modeling and analysis of stochastic systems. Chapman and Hall texts in Statistical series, CRC press.

[16] Kulkarni, V. G., 1997, "Effective bandwidths for Markov regenerative sources", Queueing Systems, vol. 24, pp. 137-153.

[17] Latouche, G., Pierce, C., and Taylor, P., 1997, "Invariant measures for quasi-birth and death processes", Stochastic Models.

[18] Massey, W., 2002, "The analysis of queues with time-varying rates for telecommunication models", Telecommunication Systems, vol. 21:2-4, pp. 173-204.

[19] Neuts, M. F., 1981, Matrix geometric Solutions in Stochastic models: An algorithmic Approach. John Hopkins University, University Press.

[20] Núñez-Queija, R., 1997, "Steady state analysis of a queue with varying service rate", Technical Report PNA-R9712, CWI.

[21] Ramaswami, V., 2000, Algorithmic analysis of stochastic models: The changing face of mathematics, Ramanujam Endowment Lecture at Anna University, Chennai, India.

[22] Riska, A., and Smirni, E., 2002, "Mamsolver: A matix analytical method tool", in the Proceedings of the 12-th International Conference on Modeling Techniques and Tools for Computer and Communication Systems Performance Evaluation, vol. LNCS 2324, pp.205-211.

[23] Roberts, J. W., 1981, "A Service system with heterogeneous user requirements", Performance of Data Communication Systems and their Applications, pp. 423-431.

[24] Ross, K. W., 1995, Multiservice loss models for broadband telecommunication networks. New York: Springer-Verlag.

[25] Sharma, V., 2001, "Queues with service rate controlled by a delayed feedback", Queueing Systems, vol 39, No. 4, pp. 303-315.

[26] Takahashi, H., and Wang, L., 1990, "Approximate analysis of a queueing system with Markov modulated arrivals", Electronics and Communications in Japan, Part 1, vol. 73, no. 11, pp 12-21.

[27] Thorne, J. R., Modeling Complex Queuing Situations with Markov Processes, *URL: charybdis.mit.csu.edu.au/mantolov/CD/ICITA2002/ papers/077-1.pdf*.

[28] Yong-Pin Zhou and Noah Gans, 1999 "A single-server queue with Markov modulated service times", October 1999, URL: http://fic.wharton.upenn.edu/fic/papers/99/9940.pdf.