1-1-2008

Non-Preemptive Shunting in M/M/1 and Dynamic Service Queueing Systems

Steven Lacek

NON - PREEMPTIVE SHUNTING IN M/M/1 AND DYNAMIC SERVICE QUEUEING SYSTEMS

Thesis submitted to
the Graduate College of
Marshall University

In partial fulfillment of
the requirements for the degree of
Master of Arts
Department of Mathematics

By

Steven Lacek

Dr. Alfred Akinsete, Advisor and Committee Chairperson
Dr. Yulia Dementieva, Committee Member
Dr. Ariyadasa Aluthge, Committee Member

Marshall University

May 2008

ABSTRACT

NON - PREEMPTIVE SHUNTING IN M/M/1 AND
DYNAMIC SERVICE QUEUEING SYSTEMS

By Steven Lacek

We provide a study of two queueing systems, namely, an M/M/1 queueing system in which an incoming customer shunts, or skips line, and a dynamic server in an infinite capacity system moving among service nodes. In the former, we explore various aspects of the system, including waiting time, and the relationships between shunting and position in queue and rate of service. Through use of global balance equations, we find the probability that an arriving non-priority customer, finding customers waiting in the system, will shunt to a position other than behind the queue. In the latter, we explore a system in which a server with infinite capacity moves among indexed linear service nodes, receives customers at various nodes, and transports the customers to other indexed nodes in the hierarchy. We determine the expected waiting times at the nodes, expected service times, expected number of customers at a given node, expected number in the system, and expected number in service. The probabilities that an arrival finds n customers at a particular node, and in the entire system are obtained.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to Dr. Akinsete, my thesis committee chair and advisor. Dr. Akinsete is, in all likelihood, the most patient person I have ever met. I hope to be the teacher he is. Thank you for piquing my interest in queue theory and making this work possible.

I would also like to thank Dr. Aluthge and Dr. Dementieva for agreeing to serve on my committee.

Of course, I need to thank Jo Lynn, my wife, and Jude, my son. Without your support and understanding, I could not have done it.

Finally, I thank Mom and Dad.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

Whether it is waiting in line at the supermarket, waiting for traffic to clear, or waiting for a bus, waiting is unavoidable and has become pert of unavoidable human activity. Customers do not enjoy waiting and may find other ways of obtaining service when they are forced to wait too long. Because of this, those who offer service have a great interest in minimizing waiting time while maximizing server utilization. This is one of the driving forces behind queueing theory. An extensive bibliography of early work in queueing theory is provided by Doig [3].

Two types of queuing models are discussed in this work. In our first model, customers needing some type of service enter the system one at a time at a mean rate λ. A single server, working at a mean rate μ attends to customers one at a time. The server will attend to customers on a first come – first serve basis. If an arriving customer finds no other customers in the system, he enters into service immediately. Otherwise, he spends time in service, and exits after service completion. When customers enter the system and find a customer already in service, a queue forms behind the service node. In our model, customers are not lost to the system, that is, once a customer enters the system, he will not leave the system without being served. Also, we will not place a limit on the number of customers that the system can hold. This basic model is examined at

length by Gross and Harris [5], Ross [12], and Takács [14] and Kleinrock [10], among others.

However, none of these authors, nor any other references in the literature, had discussed the results of a customer arriving at such a system who does not enter the system at the end of the queue, but instead jumps to a position within the queue. This act is referred to as shunting. Our model will allow for non – preemptive shunting, meaning that a customer may shunt as long as he does not interrupt the service of the customer being served. In other words, he cannot shunt directly into service.

We explore how the allowance of shunting affects the customers in queue at the time of the shunting. We will use some of the general results discussed in Chapter 2 to compare the customers' waiting time after the shunting to their waiting time before the shunting. We will also examine how the position that the shunting customer takes in line affects the waiting time of the customers in line. Using global balance equations, we also find the probability that a customer will be the victim of shunting.

Our second model consists of a set of independent service nodes. At each of these nodes, customers arrive one at a time and wait for service from a single server who is moving among the nodes. When the server arrives at a node, it takes into service all customers who are waiting at that node. The service being offered consists of the transportation to another node in the hierarchy. Once customers reach their destination, they exit the system. When the server reaches the highest indexed node, all remaining customers exit and the server makes an empty return to the first node.

Under these conditions, we will find the expected time in service of a customer who enters the network at a given point, the expected length of a particular queue in the network, and waiting time of a customer at a given queueing point. We also obtain the expected number of customers in the entire network.

Jackson [7] and Jackson [8] provides key findings in the area of queueing networks. Our system here is similar to Jackson's model, in that there is a series of nodes or, as Jackson referred to them, departments. However, in Jackson's model, customers enter at a node, receive service, move to another node, and receive service again. While Jackson's model provides multiple servers in tandem, our system has a movable or dynamic server, who provides services to arriving customers throughout the system. In other words, our system minimizes cost in service and personnel.

Closer to our work is the model proposed by Afanassieva *et al* [1]. In their work, they presented three different models, starting with a most theoretical model that has infinite waiting space and infinite server capacity such as ours. Their systems accommodate feedbacks, where customers are allowed to either move to a higher node, or to a lower node. Also, one of their systems has no queue waiting time, whereas, our system has a waiting line at each node.

Taube-Netto [13] and Iravani *et al.* [6] both offer works that deal with tandem queues, in which a single server moves between two queues. In their work, customers are not actually moving between the two service nodes, but are exiting after being served. Taube-Netto [13] gives a thorough analysis of such a system in which the server stays at a stage until the queue is cleared, and then switches to the other stage. Iravani *et al.* [6]

further develops Taube-Netto's work by offering optimal switching policies for such a system.

CHAPTER 2

Preliminaries

2.1 General Results

A queueing system (herein, referred to as system) is any model in which a service is being offered and customers arrive to receive that service. Customers may be able to enter service directly, or they may have to wait for service. In either case, once the customer is served, he/she will either leave the system, or return to the system to be served again. A customer may become tired of waiting and decide to leave the system without being served. In such a case, the customer is said to be lost to the system. We will characterize queueing systems using the notation widely credited to Kendall [9]: *A/S/n*, where *A* is the probability distribution of the time between arrivals of customers, or interarrival times to the system, *S* is the probability distribution of service times, and *n* is the number of servers in the system. Chapter 3 refers to the M/M/1 system, this indicates a Markovian arrival, memoryless service, and a unit or single server.

Throughout this work, we will identify the rate at which customers enter the system as λ and the average service rate of one server as μ. In a system with *c* servers, the congestion, or traffic intensity of the queuing system is expressed by $\rho = \frac{\lambda}{c\mu}$. Since ρ is a measure of server utilization, we can find fraction of time that the server is idle by $1 - \rho$. This quantity also measures the probability of finding the system empty. Notice

that when the average arrival rate is greater than or equal to the average service rate, then $\rho > 1$. In such a case, the servers cannot keep up with the incoming customers, and such a system is said to be unstable. In the case where $\rho = 1$, the server works exactly as fast as customers enter the system. Therefore, if at the time when the server began working, a queue had already formed, then the queue will always be there, thus we say that when $\rho = 1$, the system is also unstable.

When $\rho < 1$, if the system is operational for a long time, the system tends toward steady state conditions. When in steady state, we denote the number of customers in the entire system as *N*, the number of customers in queue as N_q, and the number of customers in service as N_s. For a system in transient state, the quantities are measured with respect to time, *t*. We then denote the number of customers in the system at time *t* as *N(t)*, the number of customers in queue at time *t* as $N_q(t)$, and the number of customers in service at time *t* as $N_s(t)$.

When the system is in steady state, we will denote the probability that $N = n$ as p_n, and the probability that $N(t) = n$ as $p_n(t)$ when the system is not in steady state.

The expected number of customers in the system at steady state is denoted by $L = E[N]$, where,

$$L = E[N] = \sum_{n=0}^{\infty} n \cdot p_n \tag{2.1}$$

Similarly, the expected number of customers in queue is denoted as $L_q = E[N_q]$. The total number of customers in queue is equal to the total number of customers in the

system minus the number of customers in service. So when the system has *c* servers, we have that,

$$L_q = \sum_{n=c+1}^{\infty} (n-c) \cdot p_n \tag{2.2}$$

Other measures of great interest are the amount of time spent in the system, *T*, the amount of time spent in queue, T_q, and the time in service, *S*. The quantities *T*, T_q, and *S* are random variables, and we can readily see that $T = T_q + S$. Using these random variables, we now introduce the expected, or mean waiting time, $W = E[T]$, and the expected, queue waiting time, $W_q = E[T_q]$. Thus we have that,

$$\begin{aligned} T &= T_q + S \\ E[T] &= E[T_q] + E[S] \\ W &= W_q + E[S] \end{aligned}$$

If the mean service rate is μ, then the mean service time is $\frac{1}{\mu}$, so that,

$$W = W_q + \frac{1}{\mu} \tag{2.3}$$

2.2 Poisson Processes

For the queuing systems discussed in this work, we will assume that customers arrive one at a time and the rate of arrival is independent of the length of the queue and of the rate of service. In order to better understand this process, the following definitions are necessary.

Definition 1 (Stochastic Process) A *stochastic process* $\{X(t),\ t \in T\}$ is a collection of random variables, $X(t)$, indexed by $t \in T$. $X(t)$ is referred to as the *state* of the stochastic process at time *t*.

Definition 2 (Counting Process) A *counting process* $\{X(t),\ t \in T\}$, is a stochastic process wherein the state represents the number of times an event has occurred by time *t*.

In order for a stochastic process to be considered a counting process, the following must hold:

i. $X(t)$ is an integer greater than or equal to zero.

ii. If $s < t$, then $X(s) \le X(t)$.

iii. When $s < t$, $X(t) - X(s)$ is the number of events in the interval $(s,t]$.

Condition (iii) allows us to view time as a set of disjoint intervals. If the number of events that occur in disjoint intervals are independent of each other, then the counting process has *independent increments*. Further, if the distribution of events occurring in any interval depends only on the length of the interval, then the counting process has

stationary increments. Therefore, the arrival process described at the beginning of this section is a counting process with independent, stationary increments.

Definition 3 (Poisson Process, Type 1) A counting process $\{X(t),\ t \in T\}$ is a Poisson process with rate λ, λ > 0, if the following conditions hold:

i. $X(0) = 0$

ii. The process has independent increments.

iii. The number of events in any interval of length $t > 0$ is Poisson distributed with mean $\lambda \cdot t$.

Condition (i) implies that our "time" begins when the first customer arrives. Condition (ii) states that the number of customers who arrive within any given interval of time is independent of the number of customers who arrive in any other interval of time. Lastly, using the definition of a Poisson distribution, condition (iii) shows that the probability there are *n* arrivals in an interval of length *t*, beginning at time *s* is

$$\Pr\{X(s+t) - X(s) = n\} = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \text{ where } n = 0,\ 1,\ 2,\ \ldots \qquad (2.4)$$

Note here that the distribution of events occurring in any interval does not depend on the time at which the interval starts, but depends instead only on the length of the interval.

Our desired queueing system arrival process satisfies the first two conditions, but we cannot say that the arrival process of our queueing system is a Poisson process. In order to do so, we now consider a second definition of a Poisson process:

Definition 4 (Poisson Process, Type 2) A counting process $\{X(t),\ t \in T\}$ is a Poisson process with rate λ, λ > 0, if the following conditions hold:

i. $X(0) = 0$

ii. The process has stationary and independent increments.

iii. $\Pr\{X(h) = 1\} = \lambda h + o(h)$

iv. $\Pr\{X(h) \geq 2\} = o(h)$

Where any function f is said to be $o(h)$ if $\lim_{h \to 0} \frac{f(h)}{h} = 0$

Ross [12] provides proof that Definition 4 implies Definition 3. Here we prove that Definition 3 implies Definition 4, and thus showing Definition 3 and Definition 4 are equivalent.

Proof: Suppose that condition iii. of Definition 3 holds. Then, for any time $h > 0$, we have that,

$$\Pr\{X(h) = 1\} = \Pr\{X(0+h) - X(0) = 1\} = e^{-\lambda h} \frac{(\lambda h)^1}{1!} = e^{-\lambda h} \cdot \lambda h$$

but since $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$, we have,

$$\Pr\{X(h)=1\}=\left[\sum_{n=0}^{\infty}\frac{(-\lambda h)^n}{n!}\right]\cdot\lambda h$$

$$=\left[\frac{(-\lambda h)^0}{0!}\right]\lambda h+\left[\sum_{n=1}^{\infty}\frac{(-\lambda h)^n}{n!}\right]\cdot\lambda h$$

$$=\lambda h+\left[\sum_{n=1}^{\infty}\frac{(-\lambda h)^n}{n!}\right]\cdot\lambda h$$

Now we let $f(h)=\left[\sum_{n=1}^{\infty}\frac{(-\lambda h)^n}{n!}\right]\cdot\lambda h$ and consider,

$$\lim_{h\to 0}\frac{f(h)}{h}=\lim_{h\to 0}\frac{\left[\sum_{n=1}^{\infty}\frac{(-\lambda h)^n}{n!}\right]\cdot\lambda h}{h}=\lim_{h\to 0}\left[\sum_{n=1}^{\infty}\frac{(-\lambda h)^n}{n!}\right]\cdot\lambda$$

$$=\lambda\cdot\lim_{h\to 0}\left[\sum_{n=1}^{\infty}\frac{(-\lambda h)^n}{n!}\right]$$

$$=\lambda\cdot\sum_{n=1}^{\infty}\left[\lim_{h\to 0}\frac{(-\lambda h)^n}{n!}\right]$$

$$=\lambda\cdot\sum_{n=1}^{\infty}\left[\frac{0}{n!}\right]=0$$

Since $\lim_{h\to 0}\frac{f(h)}{h}=0$, we have that $\Pr\{X(h)=1\}=\lambda h+o(h)$, thus showing Definition 3 implies condition (iii) of Definition 4. Now, since we are given that the process is a counting process, we know that $X(h)$ is some integer greater than or equal to zero, thus

we only now need consider $\Pr\{X(h) \geq 2\}$. Again, in conditions of Definition 3, we have that,

$$\Pr\{X(s+t) - X(s) = n\} = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \quad n = 0,\ 1,\ 2,\ \ldots$$

let $f(h) = e^{-\lambda h} \frac{(\lambda h)^n}{n!}$ and we consider,

$$\lim_{h \to 0} \frac{f(h)}{h} = \lim_{h \to 0} \frac{e^{-\lambda h} \frac{(\lambda h)^n}{n!}}{h}$$

$$= \lim_{h \to 0} e^{-\lambda h} \cdot \frac{\lambda^n h^n}{h \cdot n!}$$

$$= \lim_{h \to 0} e^{-\lambda h} \cdot \frac{\lambda^n h^{n-1}}{n!}$$

$$= \lim_{h \to 0} e^{-\lambda h} \cdot \lim_{h \to 0} \frac{\lambda^n h^{n-1}}{n!}$$

When $n \geq 2$, we see that

$$\lim_{h \to 0} e^{-\lambda h} \cdot \lim_{h \to 0} \frac{\lambda^n h^{n-1}}{n!} = 1 \cdot 0 = 0$$

Therefore, we have $f(h) = e^{-\lambda h} \frac{(\lambda h)^n}{n!}$ is $o(h)$ for $n \geq 2$, and thus,

$\Pr\{X(h) \geq 2\} = o(h)$. This shows that Definition 3 implies Definition 4, and along with Ross's proof that Definition 4 implies Definition 3, the two are equivalent. □

Note that Definition 4 allows us to shorten the length of each time increment until there is only one arrival for each increment. Through equivalence of the two definitions,

we can say that our arrival process is a Poisson process. Note that the lengths of the increments need not be uniform since they are independent of each other.

2.3 Markov Chains

We will now turn our attention to a class of stochastic processes known as Markov chain. Some of the properties of Markov chains are outlined in the following definitions.

Definition 5: (Markov process) A discrete stochastic process is said to be a Markov process if the present state of the process depends only on the immediately proceeding state. That is, $\Pr\{X_t = j | X_{t-1} = i, X_{t-2} = i_2, X_{t-3} = i_3, \ldots, X_0 = i_0\} = \Pr\{X_t = j | X_{t-1} = i\} = P_{ij}$.

In other words, given the Markov process is in state *i*, it will next move to state *j* with probability P_{ij}.

Definition 6: (Markov chain) A Markov process with discrete state space and parameter space is a Markov chain.

We assume that if the Markov chain is in state *i*, and transition to state *j* is possible, then $P_{ij} \geq 0$. Also, since *i* and *j* are indexes, we assume that $i, j \geq 0$. Also, we will assume that if the Markov chain is in state *i*. then it must move to some state *j*, and so we have that $\sum_{j=0}^{\infty} P_{ij} = 1$. It is often helpful to represent a Markov chain in matrix form as follows:

$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ P_{20} & P_{21} & P_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Markov chains, like queueing systems possess steady states. That is, if the Markov chain is allowed to operate for a long period of time, then we may begin to see a limiting value for P_{ij}. In the event that a Markov chain has made *m* state changes, we say that,

$$\lim_{m\to\infty} P_{ij}^{(m)} = \pi_j \text{ for all } i$$

This shows that after a "long" period of time, there is a steady probability that the Markov chain will settle in state *j*. It is this thinking that we will use when we talk about queueing systems that have obtained a steady state. For more in depth work with Markov chains, the reader is referred to Ross [12], Foster [4], and Gross and Harris [5], among others.

CHAPTER 3

Non-Preemptive Shunting in an M/M/1 System

3.1 General Results

We begin this chapter by clarifying the meaning of shunting. In a first in-first out queueing system, we say that a customer has shunted if he enters the system in any position other than at the end of the queue. It is assumed that the arrival process of the shunting customer does not differ from that of the customers in the system, other than the point of entry to the system. This process is also being referred to as "skipping line" or "jumping line." Being the victim of a line jumper can be very frustrating, and so we will look at the effects this action has on the expected waiting time of customers already in queue. An example of this might be seen at a toll booth, where a long line of cars is proceeding to a vacant toll taker. Immediately before the first car in line arrives at the vacant window, a car from a different lane decides to shunt in front of the line of cars. The driver who was cut off, and other drivers behind him in line, are forced to wait through the impatient driver's service.

This process is illustrated in Figure 3.1 below.

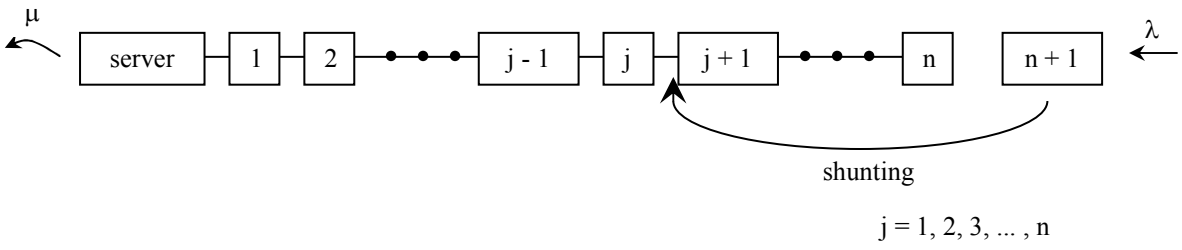

Figure 3.1

Suppose we have a first in-first out, M/M/1 system with arrival rate λ, and service rate μ. Recalling our notation from Chapter 1, we denote the time spent in the system, time spent in the queue, and time spent in service by the n^{th} customer as T, T_q, and S respectively, and we had the following:

$$T = T_q + S$$

and

$$W = W_q + \frac{1}{\mu}$$

for expected time.

Now, suppose that a customer on arrival, meets *n* customers in the system, including the customer being served, and enters the queue somewhere ahead of customer *n*, say at epoch *j*. When this occurs, we will refer to the time spent in the system by customer *n* as T'. We will denote the service time of the shunting customer as $S^{(j)}$. Thus, when shunting occurs, we have,

$$T' = T_q + S + S^{(j)}$$

Using the fact that $T = T_q + S$, we have:

$$T' = T + S^{(j)} \tag{3.1}$$

This is not surprising, since it simply states that when shunting occurs, the waiting time of a customer is increased by the amount of time it takes to serve the customer who has jumped in front of him.

Now, let $W' = E\left[T'\right]$. Because this system is M/M/1, service rate is independent of where the customer entered the line, so the average service time for each customer is $\frac{1}{\mu}$, therefore,

$$E\left[T'\right] = E[T] + E\left[S^{(j)}\right]$$

$$E\left[T'\right] = W + E\left[S^{(j)}\right]$$

$$W' = W + \frac{1}{\mu} \tag{3.2}$$

Equation 2.3 gives

$$W' = W_q + \frac{1}{\mu} + \frac{1}{\mu}$$

or

$$W' = W_q + \frac{2}{\mu} \tag{3.3}$$

Under the assumption that our system is in steady state, if customer *j* entered the system in accordance with the Poisson arrival process and is to be served at the same rate as other customers, it holds that

$$L = E[N] = \sum_{n=0}^{\infty} n \cdot p_n = \sum_{n=1}^{\infty} n \cdot p_n$$

Let L' be the expected queue length of the system after shunting has occurred. $L' = E[N+1]$, since the system has a total of $n+1$ customers, including the one that shunted. So we have that, $L' = E[N+1] = \sum_{n=0}^{\infty} (n+1) p_n$. Since $\sum_{n=0}^{\infty} p_n = 1$, we have that,

$L' = 1 + \sum_{n=0}^{\infty} n \cdot p_n$. We know from the classical M/M/1 results that $p_n = (1-\rho) \cdot \rho^n$,

where $\rho = \frac{\lambda}{\mu}$. Hence,

$$L' = 1 + \sum_{n=0}^{\infty} n \cdot \rho^n (1-\rho)$$

$$= 1 + (1-\rho) \cdot \sum_{n=1}^{\infty} n \cdot \rho^n$$

$$= 1 + \rho(1-\rho) \cdot \sum_{n=1}^{\infty} n \cdot \rho^{n-1}$$

$$= \frac{1}{1-\rho}$$

Under the previously stated conditions, it also holds that

$$L_q = E[N_q] = \sum_{n=1}^{\infty} (n-1) p_n$$

Let L_q' be the expected number of customers in queue after a shunting occurs. Thus we have

$$\begin{aligned} L_q' &= \sum_{n=1}^{\infty} ((n+1)-1) p_n \\ &= \sum_{n=1}^{\infty} n \cdot p_n \\ &= L \end{aligned}$$

3.2 Shunting and Position in Queue

We now consider how a customer's position in queue and being the victim of shunting are related. Since expected waiting time after shunting is given in terms of a customer's expected time in queue, Equation 3.2 and Equation 3.3 hold for any customer in queue, regardless of his/her position behind the position that the line jumper takes.

However, if a customer is near the front of the queue, then T_q will obviously be less than if the customer is near the end of the queue. We consider the limit:

$$\lim_{T_q \to 0} \frac{T'}{T} = \lim_{T_q \to 0} \frac{T_q + S + S^{(j)}}{T_q + S} = \frac{S + S^{(j)}}{S}.$$

We can interpret this to mean that if, at time *t*, a shunting occurs in front of customer *n* whose time left in queue at time *t* is near zero, then customer *n*'s expected service time is effectively doubled. In more general terms, the closer a customer is to the point of service, the more impact shunting will have on his waiting time.

We now consider the limit:

$$\lim_{T_q \to \infty} \frac{T'}{T} = \lim_{T_q \to \infty} \frac{T_q + S + S^{(j)}}{T_q + S} = 1$$

This shows that if customer *n* is significantly far from the point of service, then shunting has little or no effect on him.

3.3 Shunting and Service Rates

It is also interesting to consider the connection between service time and shunting. From our own experiences, we can relate to the situation when being served by a very efficient server, we are more likely to allow someone to skip ahead of us in queue. At the same time, if we are in a very long queue that is barely creeping along, we may allow someone to jump in front of us. To illustrate this, we consider the following:

$$\lim_{\mu\to\infty} W' = \lim_{\mu\to\infty}\left[W+\frac{1}{\mu}\right]$$

$$= \lim_{\mu\to\infty}\left[W_q+\frac{1}{\mu}+\frac{1}{\mu}\right]$$

$$= W_q$$

And since

$$\lim_{\mu\to\infty} W = \lim_{\mu\to\infty}\left[W_q+\frac{1}{\mu}\right]$$

$$= W_q$$

we can see that as the rate of service goes to zero, or equivalently, the service mean grows rapidly, then shunting does not affect a customers expected wait time significantly.

In the same vein, we see that,

$$\lim_{\mu \to 0} W' = \lim_{\mu \to 0} \left[W + \frac{1}{\mu} \right]$$

$$= \lim_{\mu \to 0} \left[W_q + \frac{2}{\mu} \right]$$

$$= \infty$$

and

$$\lim_{\mu \to 0} W = \lim_{\mu \to 0} \left[W_q + \frac{1}{\mu} \right]$$

$$= \infty$$

which tells us that as the rate of service goes to infinity, or equivalently, the service mean decreases rapidly, then shunting likewise has no significant effect on the expected wait time of a customer.

3.4 Probability of Shunting

To conclude this chapter, we explore the probabilities that are associated with shunting in an M/M/1, FCFS queueing system. We begin by establishing the global balance equations of an M/M/1 system. Gross and Harris [5] provide outstanding details on the subject of global balance equations.

First, we note that since arrivals follow a Poisson process, the number of arrivals over some interval, $(t, t+\Delta t)$ is equal to 1. Also, departures from the system occur one at a time. Let $p_n(t, t+\Delta t)$ be the probability that there are *n* customers in the system in the time interval $(t, t+\Delta t)$. Thus we have

$$\begin{aligned} p_n(t,t+\Delta t) &= p_n(t)(1-\lambda\Delta t)(1-\mu\Delta t) + p_{n-1}(t)(\lambda\Delta t)(1-\mu\Delta t) \\ &\quad + p_{n+1}(t)(1-\lambda\Delta t)(\mu\Delta t) \end{aligned}$$

$$\begin{aligned} p_n(t,t+\Delta t) &= p_n(t) - \lambda p_n(t)\Delta t - \mu p_n(t)\Delta t + \mu\lambda p_n(t)(\Delta t)^2 + \lambda p_{n-1}(t)\Delta t \\ &\quad - \mu\lambda p_{n-1}(t)(\Delta t)^2 + \mu p_{n+1}(t)\Delta t - \mu\lambda p_{n+1}(t)(\Delta t)^2 \end{aligned}$$

$$\begin{aligned} p_n(t,t+\Delta t) - p_n(t) &= -(\lambda+\mu)p_n(t)\Delta t + \mu\lambda p_n(t)(\Delta t)^2 + \lambda p_{n-1}(t)\Delta t - \mu\lambda p_{n-1}(t)(\Delta t)^2 \\ &\quad + \mu p_{n+1}(t)\Delta t - \mu\lambda p_{n+1}(t)(\Delta t)^2 \end{aligned}$$

$$\lim_{\Delta t\to 0} \frac{p_n(t,t+\Delta t) - p_n(t)}{\Delta t} = -(\lambda+\mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t)$$

Since $\lim_{\Delta t\to 0} \frac{p_n(t,t+\Delta t) - p_n(t)}{\Delta t} = p_n'(t)$, and when the system is in steady state, $p_n'(t) = 0$,

we then have that $0 = -(\lambda+\mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t)$

Also, when in steady state, $p_n(t) = p_n$, thus we have

$$0 = -(\lambda + \mu)p_n + \lambda \cdot p_{n-1} + \mu \cdot p_{n+1} \quad \text{when } n \geq 1$$
$$\lambda \cdot p_n + \mu \cdot p_n = \lambda \cdot p_{n-1} + \mu \cdot p_{n+1} \tag{3.4}$$

and
$$\lambda \cdot p_0 = \mu \cdot p_1 \quad \text{when } n = 0 \tag{3.5}$$

Using the fact that $1 - \rho = p_0$ we have:

$$\lambda \cdot p_0 = \mu \cdot p_1$$
$$p_1 = \rho(1 - \rho)$$

Recursively, we find that when $n = 1$

$$\lambda \cdot p_1 + \mu \cdot p_1 = \lambda \cdot p_0 + \mu \cdot p_2$$

$$\mu \cdot p_2 = \lambda \cdot p_1 + \mu \cdot p_1 - \lambda \cdot p_0$$

$$p_2 = \frac{\lambda \cdot \rho \cdot p_0 + \mu \cdot \rho \cdot p_0 - \lambda \cdot p_0}{\mu}$$

$$p_2 = p_0 \cdot \rho^2$$

$$p_2 = (1 - \rho) \cdot \rho^2$$

If we continue this pattern we see that

$$p_n = (1 - \rho) \cdot \rho^n . \tag{3.6}$$

Again, since the shunting customer follows the same Poisson arrival process and is served in the same manner as all other customer, this hold for our system.

We now let $p_{j|n}(t)$ be the probability that given *n* customers in the system at time *t*, an arrival to the system will take the position immediately behind customer *j*, for all $1 \le j \le n$. We say that the new customer takes the position behind customer *j* to account for the possibility that the new customer does not shunt, and takes his position behind customer *n*. Also, this ensures that the shunting remains non-preemptive. $p_{j|n}(t)$ can be represented as a row vector of *n* elements.

$$\hat{p}_{j|n}(t) = \begin{pmatrix} p_{1|n}(t) & p_{2|n}(t) & \cdots & p_{n|n}(t) \end{pmatrix}$$

Since we are dealing with a system that is in steady state, we now have that

$$\hat{p}_{j|n} = \begin{pmatrix} p_{1|n} & p_{2|n} & \cdots & p_{n|n} \end{pmatrix}$$

Such a vector exists for every *n*, giving the matrix,

$$P_{j|n} = \begin{pmatrix} p_{1|1} & 0 & 0 & \cdots & 0 \\ p_{1|2} & p_{2|2} & 0 & \cdots & 0 \\ p_{1|3} & p_{2|3} & p_{3|3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{1|n} & p_{2|n} & p_{3|n} & \cdots & p_{n|n} \end{pmatrix}$$

We will assume that no customers are lost to the system. Since it is given that there are *n* customers in the system, then $\sum_{j=1}^{n} p_{j|n} = 1$. We wish to determine the limiting probability of this stochastic matrix. In order to do so, we will follow a procedure detailed by Ross [12] among others. We will consider the $n = 2$ and $n = 3$ cases, and then generalize our findings for all finite *n*.

We begin by finding the eigenvalues of $P_{j|2} = \begin{pmatrix} p_{1|1} & 0 \\ p_{1|2} & p_{2|2} \end{pmatrix}$. Note that we are using λ for our eigenvalues as is customary. The reader should not confuse λ here for arrival rate.

$$P_{j|2} = \begin{pmatrix} p_{1|1} & 0 \\ p_{1|2} & p_{2|2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ p_{1|2} & p_{2|2} \end{pmatrix}$$

$$\det\left[P_{j|2} - I \cdot \lambda\right] = 0$$

$$\det\begin{pmatrix} 1-\lambda & 0 \\ p_{1|2} & p_{2|2} - \lambda \end{pmatrix} = 0$$

$$(1-\lambda)(p_{2|2} - \lambda) = 0$$

Thus $\lambda_1 = 1$ and $\lambda_2 = p_{2|2}$. From this, we have the matrix $V = \begin{pmatrix} 1 & 0 \\ 0 & p_{2|2} \end{pmatrix}$. Now, we find the eigenvectors:

$$\left(P_{j|2} - I\lambda_t\right) \cdot H_t = 0$$

First we use $\lambda_1 = 1$.

$$\begin{pmatrix} 0 & 0 \\ p_{1|2} & p_{2|2}-1 \end{pmatrix} \cdot \begin{pmatrix} h_{11} \\ h_{12} \end{pmatrix} = 0$$

$$0 \cdot h_{11} + 0 \cdot h_{12} = 0$$

$$p_{1|2} \cdot h_{11} + \left(p_{2|2} - 1\right) \cdot h_{12} = 0$$

$$p_{1|2} \cdot h_{11} - p_{1|2} \cdot h_{11} = 0$$

This system of equations yields $h_{11} = h_{12}$, thus we let $\begin{pmatrix} h_{11} \\ h_{12} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Next, we use $\lambda_2 = p_{2|2}$.

$$\begin{pmatrix} p_{1|2} & 0 \\ p_{1|2} & 0 \end{pmatrix} \cdot \begin{pmatrix} h_{21} \\ h_{22} \end{pmatrix} = 0$$

$$p_{1|2} \cdot h_{21} + 0 \cdot h_{22} = 0$$

$$p_{1|2} \cdot h_{21} + 0 \cdot h_{22} = 0$$

This gives us that $h_{21} = 0$ and $h_{22} = 1$, thus $\begin{pmatrix} h_{21} \\ h_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Now, we have that $H = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ and that $H = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, where $H^{-1} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$. Now, we apply the known result,

$$P_{j|2}^{(K)} = H \cdot V^K \cdot H^{-1}$$

$$P_{j|2}^{(K)} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1^K & 0 \\ 0 & p_{2|2}^K \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$$

$$P_{j|2}^{(K)} = \begin{pmatrix} 1 & 0 \\ \left(1 - p_{2|2}^K\right) & p_{2|2}^K \end{pmatrix}$$

If we now take the limit, we have

$$\lim_{K \to \infty} P_{j|2}^{(K)} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

After applying the same technique to

$$P_{j|3} = \begin{pmatrix} p_{1|1} & 0 & 0 \\ p_{1|2} & p_{2|2} & 0 \\ p_{1|3} & p_{2|3} & p_{3|3} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ p_{1|2} & p_{2|2} & 0 \\ p_{1|3} & p_{2|3} & p_{3|3} \end{pmatrix}$$

we find $H = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & \dfrac{p_{2|3}}{p_{2|2} - p_{3|3}} & 1 \end{pmatrix}$, $V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & p_{2|2} & 0 \\ 0 & 0 & p_{3|3} \end{pmatrix}$,

and $H^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ \dfrac{p_{2|3} - p_{2|2} + p_{3|3}}{p_{2|2} - p_{3|3}} & -\dfrac{p_{2|3}}{p_{2|2} - p_{3|3}} & 1 \end{pmatrix}$

From these, we have

$$P_{j|3}^{(K)} = \begin{pmatrix} 1 & 0 & 0 \\ 1-p_{2|2}^{K} & p_{2|2}^{K} & 0 \\ \left(\dfrac{p_{3|3}^{K}\left(p_{3|3}-p_{2|2}+p_{2|3}\right)-p_{2|2}^{K}\cdot p_{2|3}+p_{2|2}-p_{3|3}}{p_{2|2}-p_{3|3}} \right) & \left(p_{2|3}\cdot\dfrac{p_{2|2}^{K}-p_{3|3}^{K}}{p_{2|2}-p_{3|3}} \right) & p_{3|3}^{K} \end{pmatrix}$$

$$\lim_{K\to\infty} P_{j|3}^{(K)} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Thus, if we generalize, we have $\lim_{K\to\infty} P_{j|n}^{(K)} = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$

Applying the definition of conditional probability,

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

we have

$$p_{j|n} = \frac{p_{nj}}{p_n},$$

where $j = 1,2,\ldots,n$, and p_{nj} is the probability that a customer new to the system finds n customers in queue and takes the position immediately behind the j^{th} customer.

Equivalently, we have $$p_{nj} = p_{j|n} \cdot p_n . \qquad (3.7)$$

So, we have the matrix

$$P_{nj} = \begin{pmatrix} p_1 \cdot p_{1|1} & 0 & 0 & \cdots & 0 \\ p_2 \cdot p_{1|2} & p_2 \cdot p_{2|2} & 0 & \cdots & 0 \\ p_3 \cdot p_{1|3} & p_3 \cdot p_{2|3} & p_3 \cdot p_{3|3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_n \cdot p_{1|n} & p_n \cdot p_{2|n} & p_n \cdot p_{3|n} & \cdots & p_n \cdot p_{n|n} \end{pmatrix}$$

Based on our results for $p_{j|n}$, we have that $P_{nj} = p_n$ when $j = 1$ and $P_{nj} = 0$ elsewhere.

Let us consider the case when the probability of shunting decreases geometrically as the position of shunting moves away from the service point. Call this case the geometric shunting. In the case of a network with *n* number of customers, this gives,

$$p_{1|n} = \alpha,\ p_{2|n} = \alpha^2,\ \ldots,\ p_{n|n} = \alpha^n, \text{ for } 0 < \alpha \le 1$$

When $n = 1$, obviously, $p_{1|1} = 1$.

When $n = 2$, $0 < \beta \le 1$,

$$p_{j|2} = \begin{pmatrix} \alpha & 0 \\ \beta & \beta^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0.6180 & 0.3820 \end{pmatrix}$$

And when $n = 3$, $0 < \gamma \le 1$,

$$p_{j|3} = \begin{pmatrix} \alpha & 0 & 0 \\ \beta & \beta^2 & 0 \\ \gamma & \gamma^2 & \gamma^3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0.6180 & 0.3820 & 0 \\ 0.5437 & 0.2956 & 0.1607 \end{pmatrix}$$

For $n = 4$, we have

$$p_{j|4} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.6180 & 0.3820 & 0 & 0 \\ 0.5437 & 0.2956 & 0.1607 & 0 \\ 0.5188 & 0.2691 & 0.1396 & 0.0725 \end{pmatrix}$$

We see here that for a given *n*, the tendency to shunt at a point closer to the server is higher than shunting elsewhere. Also, in a geometric shunting, an arrival is less likely to shunt when the queue is long.

Also, we see that as the length of the queue increases, the probability that a shunting will occur immediately behind the first customer decreases. This pattern holds when we consider the probability of shunting immediately behind the second customer, and so on.

We also consider the case where the probability of shunting at a point decreases linearly on the number of customers to the end of the queue. Call this a linearly dependent shunting. That is, $p_{j|n} = (n - j + 1) \cdot \alpha$, where $0 < \alpha \leq 1$

For this case, $p_{1|n} = n \cdot \alpha$, $p_{2|n} = (n-1)\alpha$, …, $p_{n|n} = \alpha$.

When $n = 1$, $p_{1|1} = 1$

when $n = 2$, $0 < \beta \leq 1$,

$$p_{j|2} = \begin{pmatrix} \alpha & 0 \\ 2\beta & \beta \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2/3 & 1/3 \end{pmatrix}$$

when $n = 3$, $0 < \gamma \leq 1$,

$$p_{j|3} = \begin{pmatrix} \alpha & 0 & 0 \\ 2\beta & \beta & 0 \\ 3\gamma & 2\gamma & \gamma \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2/3 & 1/3 & 0 \\ 1/2 & 1/3 & 1/6 \end{pmatrix}$$

And generally,

$$p_{j|n} = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 2/3 & 1/3 & 0 & \cdots \\ 1/2 & 1/3 & 1/6 & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ n\cdot\alpha & (n-1)\cdot\alpha & \cdots & \alpha \end{pmatrix}$$

where $\alpha = \dfrac{2}{n(n+1)}$.

Similar interpretation can be obtained for the linearly dependent shunting, as we did in the case of the geometric shunting.

CHAPTER 4

Indexed Stations Served by a Single Dynamic Server

4.1 Overview of the System

In this chapter, we will explore a queueing system which is served by a single server who transports customers to and from *K* indexed points of service, or nodes. The server begins at node 1, takes into service any customers who are waiting at node 1, and moves to node 2 where any customer whose final destination was node 2 departs from the system. At the same time, any customers waiting for service at node 2 enter into bulk service. The process continues until the server reaches node *K*, where he/she returns, without serving anyone, to node 1, where the process starts over. We will refer to this as a cyclic service. The system is illustrated in Figure 4.1 below.

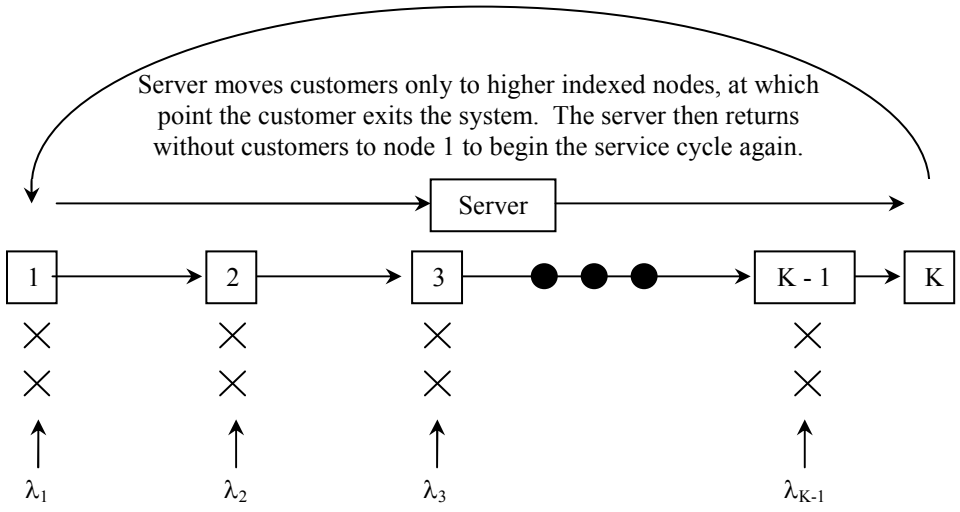

Figure 4.1

4.2 General Results

We will assume that there is no limit to the size of the queue at any service point, and that there is no limit to the number of customers who can be in service at any given time. Customers arrives at node *i*, $1 \le i < K$, following a Poisson arrival process with rate λ_i. We assume that all λ_i are independent of each other for all $1 \le i < K$, and we let $\lambda = \sum_{i=1}^{K-1} \lambda_i$. Customers move from node *i* to node *j*, $i < j \le K$ with probability p_{ij}, such that for every $1 \le i < K$, $\sum_{j=i+1}^{K} p_{ij} = 1$. Let the time of the server's return from node *K* to node 1 be r_{K1}, and let the time it takes the server to move from node i to node j, or routing time from *i* to *j*, be the random variable r_{ij} for all $i, j = 1,2,\ldots,K$, $i < j \le K$, except that when $i = K$, then $j = 1$. We assume routing times are independent of each other and independent of the number of customers waiting in queue at each node. Further, we assume that r_{ij} includes the loading time of node *i* and unloading time of node *j*.

We can see, for example, that a customer who enters at node 2, and whose destination is node 7, will need to wait through the loading/unloading of node 3 through node 6. In this case, the routing time will be $r_{27} = r_{23} + r_{34} + r_{45} + r_{56} + r_{67}$. From this, we also see that the routing time from node 1 to node *K* is

$$r_{12} + r_{23} + r_{34} + \cdots + r_{K-1,K} = \sum_{k=1}^{K-1} r_{k,k+1}$$

and thus the time of a service cycle is

$$T = r_{K1} + \sum_{k=1}^{K-1} r_{k,k+1}$$

Note that the routing time from node K to node 1 is not necessarily equal to the routing time from node 1 to node K. That is, $r_{K1} \neq \sum_{k=1}^{K-1} r_{k,k+1}$.

We will say that $0 < T \leq \Theta$, where Θ is some finite, positive value. Since r_{ij} is a random variable, its expected value is $E[r_{ij}] = \tau_{ij}$.

So, we have that the expected service time of a customer who arrives at node i is:

$$\frac{1}{\mu_i} = E\left[S^{(i)}\right] = \sum_{j=i+1}^{K} p_{ij} \cdot \tau_{ij}, \quad 1 \leq i < j \leq K \tag{4.1}$$

with equivalent expected service rate expressed by

$$\mu_i = \frac{1}{\sum_{j=i+1}^{K} p_{ij} \cdot \tau_{ij}}$$

The expected service rate of the entire system, μ, is of obvious interest. Since μ is the mean of all μ_i, $1 \leq i \leq K-1$, we have,

$$\mu = \frac{1}{K-1} \cdot \sum_i \mu_i$$

$$= \frac{1}{K-1} \cdot \sum_i \left(\frac{1}{\sum_{j=i+1}^{K} p_{ij} \cdot \tau_{ij}} \right)$$

Hence,

$$\frac{1}{\mu} = \left[\frac{1}{K-1} \cdot \sum_{i} \left(\frac{1}{\sum_{j=i+1}^{K} p_{ij} \cdot \tau_{ij}} \right) \right]^{-1}, \quad 1 \le i < j \le K \tag{4.2}$$

The customer's time spent waiting in queue will depend on the location of the server at the time of the customer's arrival. If the customer and the server arrive at a node at exactly the same time, then we say that the customer's time spent waiting in queue is 0. However, if the customer arrives at a node the moment the server is leaving the node, then the customer's time spent waiting in queue is T. The time spent waiting in queue is a continuous random variable that has a uniform distribution. Thus, a customer arriving at node *i* has an expected queue waiting time of

$$W_q^{(i)} = \frac{0 + \Theta}{2} = \frac{\Theta}{2}$$

This holds for all $i = 1,2,3,\ldots,K-1$, thus we can say

$$W_q = W_q^{(i)} = \frac{\Theta}{2} \tag{4.3}$$

Using Little's formula, which is given without proof in Gross and Harris [5], among others, and originally proven by Little [11], we find the expected length of the queue at node *i*, $L_q^{(i)}$ to be,

$$\begin{aligned} L_q^{(i)} &= \lambda_i \cdot W_q^{(i)} \\ L_q^{(i)} &= \lambda_i \cdot \frac{\Theta}{2} \end{aligned} \tag{4.4}$$

And we see that the expected number of customers waiting in the entire system is

$$L_q = \sum_{i=1}^{K-1} L_q^{(i)}$$

$$L_q = \frac{\Theta}{2} \cdot \sum_{i=1}^{K-1} \lambda_i$$

$$L_q = \lambda \cdot \frac{\Theta}{2} \tag{4.5}$$

Note that since $W_q = W_q^{(i)}$, this confirms that $L_q = \lambda \cdot W_q$ holds for this system.

We can now determine the expected time in the system for a customer who arrives at node *i*. The time in system for a customer who enters at node *i* is given by

$$W^{(i)} = W_q^{(i)} + E\left[S^{(i)}\right]$$

or (4.6)

$$W^{(i)} = \frac{\Theta}{2} + \sum_{j=i+1}^{K} p_{ij} \cdot \tau_{ij}$$

Hence,

$$W = \frac{\Theta}{2} + \frac{1}{\mu}$$

where, $\frac{1}{\mu}$ is defined as in (4.2).

And finally, by Little's formula, the expected number of customers in the entire system now becomes,

$$L = \lambda\left(\frac{\Theta}{2} + \frac{1}{\mu}\right)$$

4.3 State Probabilities for the Model

In this section, we discuss the probability $p_n^{(i)}$ that an arriving customer at node *i* finds *n* customers waiting for service. We begin by pointing out that at each node, the queue discipline is Poisson arrival and bulk service, Gross and Harris [5].

The global balance equations for node *i* is obtained as follows: We assume that there is a limit to the number of customers that can be taken into service, namely *N*. In other words, there is a constraint on the buffer space at the service node. We know from the Poisson postulates, that

$$p_n^{(i)}(t+\Delta t) = p_n^{(i)}(t)(1-\lambda_i\Delta t)(1-\mu_i\Delta t) + p_{n-1}^{(i)}(t)(\lambda_i\Delta t)(1-\mu_i\Delta t) + p_{n+N}^{(i)}(t)(\mu_i\Delta t)(1-\lambda_i\Delta t)$$

$$p_n^{(i)}(t+\Delta t) = p_n^{(i)}(t) - \lambda_i p_n^{(i)}(t)(\Delta t) - \mu_i p_n^{(i)}(t)(\Delta t) + \lambda_i\mu_i p_n^{(i)}(t)(\Delta t)^2 + \lambda_i p_{n-1}^{(i)}(t)(\Delta t) - \lambda_i\mu_i p_{n-1}^{(i)}(\Delta t)^2 + \mu_i p_{n+N}^{(i)}(t)(\Delta t) - \lambda_i\mu_i p_{n+N}^{(i)}(\Delta t)^2$$

$$\frac{p_n^{(i)}(t+\Delta t) - p_n^{(i)}(t)}{\Delta t} = -\lambda_i p_n^{(i)}(t) - \mu_i p_n^{(i)}(t) + \lambda_i\mu_i p_n^{(i)}(t)(\Delta t) + \lambda_i p_{n-1}^{(i)}(t) - \lambda_i\mu_i p_{n-1}^{(i)}(\Delta t) + \mu_i p_{n+N}^{(i)}(t) - \lambda_i\mu_i p_{n+N}^{(i)}(\Delta t)$$

$$\lim_{\Delta t\to 0}\frac{p_n^{(i)}(t+\Delta t) - p_n^{(i)}(t)}{\Delta t} = -\lambda_i p_n^{(i)}(t) - \mu_i p_n^{(i)}(t) + \lambda_i p_{n-1}^{(i)}(t) + \mu_i p_{n+N}^{(i)}(t)$$

$$p_n^{(i)\prime}(t) = -(\lambda_i + \mu_i)p_n^{(i)}(t) + \lambda_i p_{n-1}^{(i)}(t) + \mu_i \cdot p_{n+N}^{(i)}(t)$$

If steady state exists, $p_n^{(i)\prime}(t) = 0$, so that

$$0 = -(\lambda_i + \mu_i)p_n^{(i)} + \lambda_i p_{n-1}^{(i)} + \mu_i \cdot p_{n+N}^{(i)}, \qquad n \geq 1 \tag{4.7}$$

with

$$0 = -\lambda_i p_0^{(i)} + \mu_i \cdot p_1^{(i)} + \mu_i \cdot p_2^{(i)} + \cdots + \mu_i \cdot p_N^{(i)} \qquad \text{when } n = 0$$

The following digression is needed for our discussion: A *linear operator,* $D = \{D^0, D^1, \ldots, D^n\}$, for the sequence $\{a_1, a_2, a_3, \ldots, a_N\}$ is defined as,

$$\begin{aligned} D^0 a_n &= a_n \\ D^1 a_n &= a_{n+1} \qquad \text{for all } m, n. \\ D^m a_n &= a_{n+m} \end{aligned}$$

With this, we can write the equation

$$C_n a_n + C_{n+1} a_{n+1} + \cdots + C_{n+N} a_{n+N} = 0,$$

as,

$$C_n D^0 a_n + C_{n+1} D^1 a_n + \cdots + C_{n+N} D^N a_n = 0$$

$$\left[\sum_{j=n}^{n+N} C_j \cdot D^{j-n} \right] a_n = 0$$

Letting $n - 1 = m$, equation (4.7) may now be expressed as

$$0 = \lambda_i p_m^{(i)} - (\lambda_i + \mu_i)p_{m+1}^{(i)} + \mu_i \cdot p_{m+1+N}^{(i)}$$

So that

$$\left[\lambda_i - (\lambda_i + \mu_i)D + \mu_i \cdot D^{N+1}\right] p_m^{(i)} = 0, \quad m \geq 0$$

or equivalently,

$$\left[\lambda_i - (\lambda_i + \mu_i)D + \mu_i \cdot D^{N+1}\right] = 0$$

This is referred to as the operator equation, and its roots are given as $\left(r_1^{(i)}, r_2^{(i)}, \ldots, r_{N+1}^{(i)}\right)$.

From this, we have

$$p_n^{(i)} = \sum_{j=1}^{N+1} C_j^{(i)} \cdot \left(r_j^{(i)}\right)^n, \quad n \geq 0$$

In order to maintain $\sum_{n=0}^{\infty} p_n^{(i)} = 1$, all $C_j \cdot r_j^n$ must be less than 1, therefore, either our roots are less than 1, or $C_j = 0$. To determine the number of roots that are on the interval (0,1) we use **Rouché's Theorem**, which states:

Theorem: *If f and g are functions analytic inside and on a closed contour C and if* $|g| < |f|$ *on C, then f and* $f + g$ *have the same number of roots inside C.*

To employ this theorem, for any *i*, we let $g(D) = D^{N+1}$ and we let $f(D) = \frac{\lambda_i}{\mu_i} - \left(\frac{\lambda_i}{\mu_i} + 1\right) \cdot D$. We have that both *f* and *g* are polynomials and are therefore analytic. Now let *C* be a circle of radius $1 - \xi$, with ξ chosen to be sufficiently small. That is, $|D| \leq 1 - \xi$. We then have

$$|f(D)| = \left| D\left(\frac{\lambda_i}{\mu_i} + 1\right) - \frac{\lambda_i}{\mu_i} \right|$$

$$\geq \left| |D| \cdot \left(\frac{\lambda_i}{\mu_i} + 1\right) - \frac{\lambda_i}{\mu_i} \right| = \left| (1 - \xi) \cdot \left(\frac{\lambda_i}{\mu_i} + 1\right) - \frac{\lambda_i}{\mu_i} \right|$$

$$> \left| 1 - \xi \cdot \left(\frac{\lambda_i}{\mu_i} + 1\right) \right|$$

$$> 1 - \xi \cdot \left(\frac{\lambda_i}{\mu_i} + 1 \right)$$

since $1 - \xi \cdot \left(\frac{\lambda_i}{\mu_i} + 1 \right) < 1 - \xi$,

then,

$$|f(D)| > 1 - \xi \left(\frac{\lambda_i}{\mu_i} + 1 \right)$$

$$> (1 - \xi)^{N+1}$$

Thus, $|f(D)| > |g(D)|$

Over the interval (0, 1), *f(D)* is a decreasing function, which maps to the interval $\left(\frac{\lambda_i}{\mu_i}, -1 \right)$. Thus, *f(D)* has one root in C, namely

$$D = \frac{\lambda_i / \mu_i}{\lambda_i / \mu_i + 1} = \frac{\lambda_i}{\lambda_i + \mu}.$$

Therefore, by Rouché's theorem, we have that $\left[\lambda_i - (\lambda_i + \mu_i) D + \mu_i \cdot D^{N+1} \right] = 0$ has one root in $(0,1)$, namely r_i^n.

Thus, when

$$\left[\lambda_i - (\lambda_i + \mu_i) D + \mu_i \cdot D^{N+1} \right] p_m^{(i)} = 0, \qquad \begin{matrix} m = n+1 \\ m \geq 0 \end{matrix}$$

we have

$$p_n^{(i)} = C \cdot r_i^n \qquad \text{for } n \geq 0 \text{ and } 0 < r < 1$$

Realizing again that $\sum_{n=0}^{\infty} p_n^{(i)} = 1$, we can show that $p_n^{(i)} = (1 - r_i) \cdot r_i^n$. Thus we have,

$$p_n^{(i)} = p_0^{(i)} \cdot r_i^n$$

Where, $$p_0^{(i)} = 1 - r_i$$

Jackson [7] provides that if P_n^i is the probability that there are *n* customers waiting at node *i*, then the state of the system can be described by the product

$$P(n_1, n_2, \ldots, n_{K-1}) = P_{n_1}^1 \cdot P_{n_2}^2 \cdot \ldots \cdot P_{n_{K-1}}^{K-1}$$

CHAPTER 5

Conclusion

We have used classical queueing theory techniques to obtain important measures of system performance in an M/M/1 system that allows non – preemptive shunting. We have determined the expected queue length, expected time in queue, and expected time in system. From our results, we determined that the further a customer is from the point of service, the less shunting will affect his expected waiting time. Also, we have shown that as the service rate reaches either zero or infinity, shunting has no perceived effect on a customers' expected waiting time. Using global balance equations and stochastic matrices, we determined that in an M/M/1 system that allows shunting, the limiting probability of shunting immediately behind the j^{th} customer given n customers in the system is $p_{1|n} = 1$ for and $p_{j|n} = 0$ where $j \neq 1$.

Further study is needed in the preemptive shunting case, where a customer could interrupt the customer currently in service. Much study has already been given to the topic of preemptive priority queues. This could provide the basis by which a preemptive shunting model could can be developed. Using global balance equations for preemptive priority queues that are already known and this work, a preemptive shunting model could be obtained, and consequently the measures of system performance.

Also, in a preemptive shunting queue, an admission policy might prove to be necessary, since there is the potential for the original customer in service to be interrupted by a shunter. This provides a large source of topics, including defining admission

policies and from there, finding measures of system performance. Beyond that, work could be done to find optimal admission policies based on desired system performance. Chang and Chen [2] have worked with admission policies regarding tandem queues. Some of their work may translate to our model here.

Also in this work, we have defined a system in which a single server transports customers through a hierarchy of independent, indexed service nodes. We again used classical methods to determine measures of system performance including expected waiting time in queue, expected service time, and expected queue length for each service node.

Since our model operates with highly theoretical boundaries including infinite queue capacity and infinite service capacity, it would be of practical interest to place limiting values on these fields. This provides a topic for further study of this model.

Also, our server moves in only one direction, this is clearly not the most efficient use of the server. More work is needed to show similar results when the server is allowed to transport customers in two directions. Along the same vein, results should be found for when the server is allowed to rest at either node 1 or node K until called upon by a customer at a node, particularly server idleness.

BIBLIOGRAPHY

[1] Afanassieva, L. G., Fayolle, G., and Nazarov, L. V., "Preliminary models for Moving Server Networks," Rapport de recherché, INRIA, No. 2469, Janvier 1995.

[2] Chang, K. H., and Chen, W. F., "Admission Control Policies for Two-Stage Tandem Queues with no waiting spaces," Computer & Operations Research, Vol. 30 Issue 4, (2003), 589 – 601.

[3] Doig, A., "A Bibliography on the Theory of Queues," Biometrika, Vol. 44, No. 3/4, (1957), 490 – 514.

[4] Foster, F. G., "On the Stochastic Matrices Associated with Certain Queuing Processes," The Annals of Mathematical Statistics, Vol. 24, No. 3, (1953), 350 – 366.

[5] Gross, D., and Harris, C. M., *Fundamentals of Queueing Theory*, 3rd ed., John Wiley & Sons, New York, 1998.

[6] Iravani, S. M. R., Posner, M. J. M., and Buzacott, J. A., "A two-stage tandem queue attended by a moving server with holding and switching costs," Queueing Systems, Vol. 26, No. 3 – 4. (1997), 203 – 228.

[7] Jackson, J. R., "Networks of Waiting Lines," Operations Research, Vol. 5, Issue 4, (1957), 518 – 521.

[8] Jackson, J. R., "Jobshop-Like Queueing Systems," Management Science, Vol. 10, No. 1, (1963), 131 – 142.

[9] Kendall, D. G., "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of the Imbedded Markov Chain," The Annals of Mathematical Statistics, Vol. 24, No. 3, (1953), 338 – 354.

[10] Kleinrock, L., *Queueing Systems, Volume I: Theory*, John Wiley & Sons, New York, 1975.

[11] Little, J. D. C., "A Proof for the Queueing Formula: $L = \lambda W$," Operations Research, Issue 3, (1961), 383 – 387.

[12] Ross, S. M., *Introduction to Probability Models*, 8th ed., Academic Press, California, 2003.

[13] Taube-Netto, M., "Two Queues in Tandem Attended by a Single Server," Operations Research, Vol. 25, No. 1, (1977), 140 – 147.

[14] Takács, L., *Introduction to the Theory of Queues*, Oxford University Press, New York, 1962.