Introduction to Queueing Theory: A Modeling Perspective

Glenn Ledder

July 2019, revised October 2019

Contents

This document contains an introduction to queueing theory with emphasis on using queueing theory models to make design decisions. It therefore combines probability with optimization. These concepts are contrasted in a statement I once heard in a talk:

> Probability is the study of the typical for issues of chance.
> Optimization is the study of the exceptional for issues of choice.[1]

The first five sections of these notes develop the concepts and results of queueing theory. This topic is about issues of chance, such as the amount of time one must wait in line before reaching the checkout counter in a store. Hence, our goal will be to characterize the typical behavior of such systems, and we must keep in mind that the actual behavior in any one instance will not necessarily be close to the typical behavior. Sections 6 and 7 use the results of queueing theory in the context of optimization problems . Our goal in these sections will be to identify the choices that are available in a given setting and determine which choice produces the exceptional result. We must keep in mind that choices involving design of probabilistic systems are less certain than choices involving design of deterministic systems. In a deterministic system, we can say exactly what will be the result of any design choice. In a probabilistic system, the best we can say is what will be the average result of a design choice. The optimal decision may not yield the best result in a particular instance, but it will be the decision that results from rational analysis of the options and therefore the best decision that can be made with the information at hand.

These notes assume familiarity with basic calculus and probability theory. In particular, queueing theory makes extensive use of probability distributions and expected value. These concepts are reviewed here to some extent, but the reader may need to look elsewhere for more information.

A Note on Notation

The enormous number of quantities in all of mathematics have to be represented with only a handful of symbols–generally the Latin and Greek alphabets with subscripts. Inevitably this means that many symbols get used differently in different contexts. The symbol λ is used to denote Lagrange multipliers in optimization, adjoints in control theory, eigenvalues in linear algebra and partial differential equations, and the mean customer arrival rate in queueing theory. Clearly, symbols do not have fixed meanings; rather, they mean what we define them to mean. This has implications for both the reading and writing of mathematics. When reading mathematics, one has to be careful to look for information about symbol meanings and be aware that one author's W and W_q might be another author's S and W;[2] not only are the symbols for a given quantity different, but the symbol "W" has different meanings in the different systems. This point is particularly important when reading supplementary material on the internet. It is very likely that the material you are reading has some notational differences from your textbook or lecture notes, and you have to be able to translate from one system to the other. When writing mathematics, it is necessary to be clear about terms and notation to spare your reader any confusion. In general, it is a good idea to define every symbol other than π or e when writing about mathematics and modeling. This is the only way to make sure that the reader will understand what you have written.

[1] I regret not knowing the name of the author to whom this memorable statement should be attributed.
[2] This specific example arises in Section 2.

1 Queueing Theory Basics (see Hillier and Lieberman 17.2,7)

Learning Objectives

1. Know the goals of queueing theory.
2. Be able to identify the defining characteristics of a queue system from the standard 5-character identifiers.
3. Be able to calculate the arrival-service ratio γ and the utilization factor ρ from a given narrative and explain their significance.
4. Know the four principal performance measures of a queue system and be able to calculate them from the steady-state probabilities.
5. Be able to calculate the four performance measures for an M/G/1/∞/∞ system using λ, μ, and σ.

A *queue system* is a system characterized by a bank of parallel service channels with a stream of "customers" who enter at distinct times and receive service, possibly waiting in a queue if all servers are busy. The line of customers in a convenience store is a nice example. Our ultimate goals are

1. To identify the features needed in the specification of a queue system;
2. To develop methods for determining the performance of a system; and
3. To develop protocols that allow system design decisions to be made so as to optimize the overall cost associated with the system.

1.1 Defining a Queue System

A number of elements must be prescribed in order to define a queue system, including

1. The type of probability distribution used for the arrival process, with one or more parameters.
2. The type of probability distribution used for the service process, with one or more parameters.
3. The number of servers in the service station (for example, the number of check-out stations in service at a grocery store).
4. The maximum number of customers that can be in the system, if limited.
5. The size of the population of potential customers, if limited.

Some of this information is presented in compact form as an identifier of the form A/S/s/K/N , where

- A designates the type of distribution used for arrival times,
- S designates the type of distribution used for service times,
- s designates the number of servers,
- K is the maximum number of customers that can be in the system at any one time, and
- N is the size of the population of potential customers.

Typical designators for the distributions are

- M for the exponential distribution (Markovian),
- E_k for an Erlang distribution (which we will use later), and
- D for the deterministic distribution (constant times),
- G for a general distribution (unspecified except for mean and standard deviation).

The size of the calling population is important because customers that enter a queue system should be removed from the list of potential customers. This decreases the mean arrival rate, but usually the decrease is too small to worry about. Unless the calling population is small enough for the decrease to matter, it is best to consider it to be infinite. In this case, the fifth designator N is often omitted. Similarly, the number of customers that can be in the system at any one time is usually limited, but most of the time the capacity is large enough that it is never actually reached. In this case, it is best to consider the maximum system size to be infinite. As with infinite calling population size, it is common to omit the fourth designator K when the queue size is unlimited.

The most commonly used systems are of the form M/M/s, meaning that the arrival and service processes are exponentially distributed and the system size and calling population are unlimited. We will study these systems in Section 4.

1.2 Properties of Queue Systems

The number of customers in a queue system changes over time. When the system first begins to operate, there are generally no customers. Those customers who arrive before the servers are all occupied get to begin service immediately, while customers who arrive later only get to begin service when they get to the front of the queue. Thus the probability that there are 4 customers in the system is initially 0, but it rises as the system remains open.

1.2.1 The arrival-service ratio and the utilization factor

Suppose a queue system has the property that the mean arrival rate does not change as the system size changes. This requires that both the maximum system size K and the calling population N are infinite and that there are no unusual features that can lower the arrival rate as the system size increases. Given a mean rate of service completions of μ for each of s servers, the total service capacity is a mean rate of $s\mu$. The ratio of mean arrival rate to mean total service completion rate, given by

$$\rho = \frac{\lambda}{s\mu}, \tag{1.1}$$

then represents the fraction of the service capacity that is used. For this reason, it is called the *utilization factor* for the queue system. The quantity is often used even in cases where λ is not fixed, but the interpretation as utilization factor no longer holds.

For both modeling and computation, it is also helpful to define the arrival-service ratio

$$\gamma = \frac{\lambda}{\mu}. \tag{1.2}$$

This parameter represents the expected number of arrivals during the average amount of time for a service completion, which we might call the "load" of the system. In modeling, the most frequent scenario is one in which the rates λ and μ are fixed and the problem is to choose the optimal number of servers. The parameter γ is much more useful than ρ in this context because it is strictly a property of the scenario while ρ combines elements of the scenario data (λ and μ) with the independent variable of the optimization problem (S). Computationally, we'll find γ more useful that ρ in cases where the arrival rate depends on the system state.

1.2.2 Performance measures

We are usually only interested in queue systems that reach a steady-state, meaning that eventually the probability of any particular system size no longer changes over time. For example, suppose $\lambda = 2$ and $\mu = 1$ with just one server. On the average, two new customers will arrive for every customer who completes service, so the system size will just keep growing. In contrast, if $\rho = 0.9$, the service capacity is larger than the expected rate of customer arrivals. This means that there will be times when the system is empty because random chance has provided a period between arrivals long enough for the servers to empty the system. Given enough time, there will be some probability P_0 that the system is empty at any particular time. Similarly, there will be probabilities P_1, P_2, and so on, that indicate the probabilities of the system having a total of 1, 2, and so on customers. It is possible that a queue system can achieve steady state even when $\rho > 1$; this requires some additional mechanism, such as a limited system size or a limited calling population, so that the arrival rate falls to 0 when the system reaches its capacity.

Given a system that reaches steady-state for whatever reason, the set of values P_n is an emergent property; that is, it is a property that comes about in the running of the system and requires analysis to determine. Queueing theory is largely about how to determine these steady-state probabilities and some important performance measures. Two of these involve the numbers of customers.

1. The mean number of customers in the system over time, including those who are in the queue as well as those being served. This quantity is usually designated as L.

2. The mean number of customers in the queue, usually denoted L_q. This quantity is seldom of special interest in modeling, but it is mathematically important because it is usually the easiest of the four performance measures to determine.

The other two performance measures involve the average amount of time spent by customers. There are two common choices for terminology and notation, so one must be careful to identify which system is being used, both when reading what others have written and when writing for the benefit of others.

3. The mean amount of time that a customer spends in the system. Some authors call this the "sojourn" time and denote it with the symbol S. Others call it the "waiting" time and use the symbol W.

4. The mean amount of time that a customer spends in the queue. Authors who use "sojourn" time for mean time in the system usually call this the "waiting" time and denote it as W, while authors who use "waiting" time for mean time in the system usually call this the "waiting time in the queue" and denote it as W_q.

The choice between the two notation/terminology systems is a matter of taste. I have two reasons for preferring the S and W system. First, the phrase "waiting time in the queue" is unnecessarily complicated. Second, in everyday language we only think of ourselves as "waiting" when we are in the queue, not when we are being served. Given a choice, it is better to have the mathematical meaning of a word match its nonmathematical meaning.

Any of these four performance measures can be used to quantify the functioning of a queue system. Which we emphasize depends on the context. For internal queue systems, where the customers are machines in a factory and the servers are repair crews, the most logical choice is L because customers in the system represent lost productivity. For external systems, think of an auto repair shop as an example. As a customer, you don't really care how many fellow customers are in the shop, and you are as inconvenienced by slow service as by a long wait to begin service. Assuming your car gets repaired properly at a reasonable cost, it is the sojourn time that will determine if you return to that shop.

There are three simple formulas that relate the four performance measures. First off, the mean amount of time spent in service is of course $1/\mu$, so this is the difference between the waiting times with and without the time spent in service:

$$S = W + \frac{1}{\mu}. \tag{1.3}$$

In addition to this obvious relationship, there are the more subtle relationships that go by the name of *Little's formulas*:

$$L = \bar{\lambda} S, \qquad L_q = \bar{\lambda} W. \tag{1.4}$$

The symbol $\bar{\lambda}$ represents the expected value of the arrival rate.[1] One might easily guess these formulas from dimensional consistency (L is customers and S is time, so L/S is customers per time), but they require some effort to prove. Notice that the formulas (2.3) and (2.4) can also be combined to get a less obvious formula relating L and L_q:

$$L = L_q + \gamma. \tag{1.5}$$

Taken together, formulas (2.3) through (2.5) mean that if we can calculate one of the performance measures, then we can calculate all of them. Usually the easiest of the four is L_q, which we can write as a weighted average of the number of customers in the queue for each state of the system.

$$L_q = P_{s+1} + 2P_{s+2} + 3P_{s+3} + \cdots = \sum_{n=s+1}^{\infty} (n-s)P_n, \tag{1.6}$$

[1] In many systems, the arrival rate is always the specific value denoted as λ, but in other systems the arrival rate depends on the system size.

where P_n is the steady-state probability that there are n customers in the system.[2]

Example

An M/M/2/3/3 queue system has mean service rate $\mu = 2$. The arrival rates depend on the system state: $\lambda_0 = 3$, $\lambda_1 = 2$, $\lambda_2 = 1$, and $\lambda_3 = 0$. The resulting steady-state probabilities are $P_0 = 0.29$, $P_1 = 0.44$, $P_2 = 0.22$, and $P_3 = 0.05$.[3] The expected number of customers in the queue is a weighted average of the numbers of customers in the queue for each state, which (given that there are 2 servers) are 0, 0, 0, and 1 for the states 0, 1, 2, and 3. Hence,[4]

$$L_q = (0)(0.29) + (0)(0.44) + (0)(0.22) + (1)(0.05) = 0.05.$$

Of course the length of the queue is never 0.05, since queue length can only be an integer. But if we make frequent counts of the queue length, we should expect the average of those measurements to be approximately 0.05.

We can use formula (2.4) to get W, but first we need to calculate $\bar{\lambda}$ as a weighted average of the λ_n:

$$\bar{\lambda} = (3)(0.29) + (2)(0.44) + (1)(0.22) + (0)(0.05) = 1.97.$$

Thus,

$$W = \frac{0.05}{1.97} = 0.0254, \quad \underline{W} = 0.0254 + 0.5 = 0.5254, \quad L = (1.97)(0.5254) = 1.035.$$

1.3 M/G/1/∞/∞ Results

Usually, the performance measures of a system can only be determined after the steady-state probabilities are known, and this in turn can be done only for an M/M system. However, there is a simple formula for the queue length L_q that works for any M/G/1/∞/∞ system;[5] that is, systems for which

1. There are no limits to the number of potential customers or the number of customers who can be in the system at any one time;
2. Arrival times are exponentially distributed with mean rate λ;
3. Service times have a mean of $\mu_T = 1/\mu$ and a standard deviation σ, but no specific service distribution.[6]
4. There is one server.

We present this formula here without derivation, as the derivation is beyond the scope of this presentation:

$$L_q = \frac{\rho^2(1 + \mu^2\sigma^2)}{2(1 - \rho)}, \qquad \rho = \frac{\lambda}{\mu}, \tag{1.7}$$

[2]Generally it is easier to use the expanded version than the compact summation formula.

[3]We'll work out these probabilities in Section 3.

[4]We can also obtain the result formally from (2.4), but it is just as easy and a better learning experience to do it from first principles.

[5]In the rest of this section we omit the fourth and fifth designators, as is typically done when they are infinite.

[6]Note that μ refers to a rate and σ to a time.

leading to

$$L = \frac{2\rho - \rho^2(1 - \mu^2\sigma^2)}{2(1 - \rho)}. \tag{1.8}$$

Formula (2.7) is called the *Pollaczek-Khintchine formula*.

For the specific case where the service times are exponentially distributed (M/M/1), the standard deviation is $\sigma = 1/\mu$ and the result reduces to

$$L = \frac{\rho}{1 - \rho}, \tag{1.9}$$

while the assumption of uniform service time $1/\mu$ (M/D/1) yields

$$L = \frac{\rho(2 - \rho)}{2(1 - \rho)}. \tag{1.10}$$

Note that formula (2.10) has an extra factor $(1 - \rho/2)$ compared to formula (2.7). Given that this factor is less than one, we see that greater uniformity in service times improves system performance by as much as 50%.[7] See Figure 2.1. This is a general characteristic of queue systems (although the maximum amount of improvement might be different from 50%). Less variability with the same mean service rate is always better. There are no obvious design implications of this result, since we don't get to choose the characteristics of service jobs. It may influence the decision to add a server, as additional servers have more benefit when service times have a higher variability.

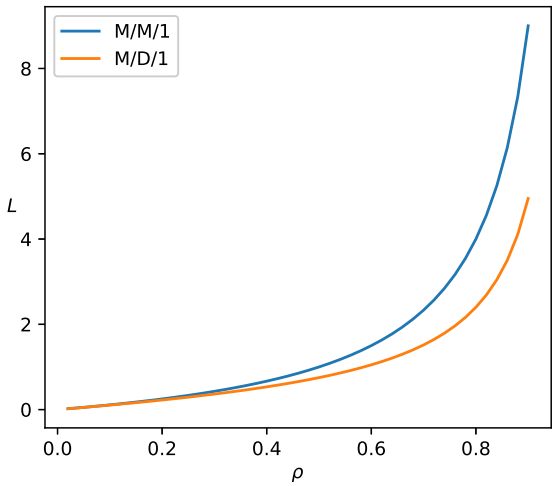

Figure 1.1: Dependence of expected system size on arrival-service ratio for exponential and deterministic service distributions.

[7] Other service distributions generally produce results that fall between M/M/1 and M/D/1.

2 Stochastic Processes (see Hillier and Lieberman 17.4)

Learning Objectives

1. Understand what is meant by the term *stochastic process*.
2. Be able to explain why the arrival process for a queue system is a stochastic process.
3. Be able to discuss the lack of history property that we expect to be valid for arrival processes.
4. Be able to calculate the probability of the next event occuring within some specified time interval for the exponential distribution.
5. Be able to show that the exponential distribution has the lack of history property.
6. Be able to explain why the exponential distribution is probably not a good model for queue service processes.

Mathematicians generally introduce concepts by presenting mathematical definitions. This is unfortunate for beginning courses in any subject, because these definitions do not just come from nothing. Consider the derivative, which mathematicians define as the limit of a particular quotient of differences. Why bother? Because that limit corresponds to something we want to study with mathematics. Without understanding why the derivative is important, we can't properly appreciate how it gets defined. We can't understand the derivative without understanding what it represents. So in addition to the *mathematical* definition, there is also a *semantic* definition, which is: "the instantaneous rate of change of a function." To the user of mathematics, mathematical definitions may or may not be important, but semantic definitions are always critical. In these notes, we'll focus first on the semantic definitions of concepts and only add mathematical definitions when they are needed.

2.1 Definition of a Stochastic Process

We define a *stochastic process* as "a time sequence represented by a variable whose values are subject to random variation." If you stand outside a convenience store, you will see customers entering and leaving. The arrival and departure times are time sequences that can be represented by a variable indicating the elapsed time between successive arrivals or departures. These times are subject to random variation, so the arrival and departure streams are stochastic processes.

Even though there is no way to predict the next number in the sequence, a mathematically defined stochastic process is not completely unpredictable because it is associated with a given probability distribution. This means that the average behavior of the stochastic process is predictable even though individual values are not. The probability distribution defines a theoretical mean value and the probability that the next randomly chosen number will be in any particular interval. Given a long enough sequence, the expectations provided by the underlying probability distribution will be met.

Real stochastic processes seldom conform exactly to a mathematical abstraction. Nevertheless, we can use mathematical models to approximate a real system. Observation can help in

the choice of a model; for example, we can estimate the average arrival rate by counting the number of arrivals that occur in a 6-hour period.

2.2 The Arrival Process for a Queue System

Suppose you stand inside a convenience store and watch the checkout counter. Customers[8] enter the store, find their items, and then walk up to the cashier. The queue system is just the cashier (not the whole store), so each customer's arrival occurs at the moment when they reach the counter. Since customers are independent, there is no way to predict when the next customer will arrive and no connection between the arrivals of different customers. Two customers might arrive moments apart or there might be several minutes between two consecutive customers. The arrival process can be thought of as a stochastic process that generates a sequence of inter-arrival times.

Because stochastic processes are predictable in the aggregate, we can assume that there is an average arrival rate, which we will call λ. Note that the dimension of λ is "customers/time."[9] When we apply a stochastic process model to a real situation, we'll need to make observations to help us decide what value to use for λ in the model.[10]

There is one fundamental qualitative feature of stochastic arrival processes when there is a large number of potential customers and those customers are independent of each other:

- The time at which the next arrival occurs, as measured from the current time, does not depend on the time at which the previous arrival occurred.

Suppose the mean arrival rate is 15 customers per hour, so that the average time between arrivals is 4 minutes. Now suppose you start your stopwatch at exactly 1:00. The expected time for the next arrival is 1:04. Suppose no customers arrive in the first 3 minutes. When do we now expect the next arrival? It might seem logical that the answer is still 1:04, but that is not correct. It is now 1:03, and our best guess for the next arrival is that it will happen in 4 minutes, at 1:07. If we get to 1:16 without an arrival, the expected time for the next arrival is still 4 minutes away. To summarize: additional elapsed time does not change the expectation for the next arrival time, because the arrival time of customer n is unrelated to the arrival time of customer $n-1$. This is hard to conceptualize because it is so different from everyday experience, when events become more likely as the wait for them lengthens.[11]

[8] In the queueing theory context, think of a customer as encompassing groups of individuals who are shopping together as well as single individuals.

[9] It is important to note the dimensions of quantities, as dimensional reasoning is very helpful in mathematical modeling.

[10] In a real situation, the average arrival rate is probably not a well-defined quantity, since the arrival rate of customers probably depends on the time of day and the weather. But remember that we are talking here about a stochastic process as a *model* for a real situation.

[11] I know of only one example other than stochastic processes that is like this. When I was in college in the 1970's, scientists thought that we would see a practical fusion reactor in about 20 years. When I started teaching in 1989, scientists at that time still estimated 20 years for a practical fusion reactor. Scientists in 2019 still estimate that a practical fusion reactor is unlikely within the next 20 years. Of course the story here is not about a history-less property, but about how hard it is to predict the difficulty of something we don't yet know how to do.

2.3 The Exponential Distribution

There are three ways to define a probability distribution, two that are easier to understand and one that is more broadly useful. One of the two easier ones is the *survival function*, which we define as

$$S(t) = P\{T > t\}; \tag{2.1}$$

that is, $S(t)$ is the probability for any time t that the next event has not yet occured at that time.[12] There are some simple properties that are common to the survival functions in queueing theory:

$$S(0) = 1, \qquad S'(t) \leq 0, \qquad S(\infty) = 0; \tag{2.2}$$

these say that the next event always occurs after time 0, that the probability the event has not yet occurred decreases over time, and that the next event always occurs at some finite time.[13]

Given a survival function $S(t)$, the *probability density function* is defined as

$$f(t) = -S'(t). \tag{2.3}$$

Note that

$$\int_t^\infty f(t)\,dt = -\int_t^\infty S'(t)\,dt = \int_\infty^t S'(t)\,dt = S(t) - S(\infty) = S(t) = P\{T > t\}; \tag{2.4}$$

Thus, a probability distribution can also be defined through a probability density function, using a definite integral to determine probabilities. Similarly,

$$\int_{t_1}^{t_2} f(t)\,dt = S(t_2) - S(t_1) = P\{t_1 < T < t_2\}.^{14} \tag{2.5}$$

The exponential distribution is so named because of its defining functions:

$$S(t) = e^{-\alpha t}, \qquad f(t) = \alpha e^{-\alpha t}. \tag{2.6}$$

The parameter α allows us to "tune" the distribution so that it has whatever mean rate λ we desire. To connect the distribution parameter α with the rate parameter λ, we compute the expected value of the random variable T. This is done by taking a weighted average of the possible times t using the probability density function to provide the weighting.[15] Specifically, we get

$$E(T) = \int_0^\infty t \cdot \alpha e^{-\alpha t}\,dt = \cdots = \frac{1}{\alpha}.^{16} \tag{2.7}$$

This means that the mean value of the time T is $1/\alpha$, and we can interpret α as the mean rate of arrivals. For an arrival distribution, we'll simply use the exponential distribution with parameter λ.

[12] The cumulative density function $F(t) = 1 - S(t)$ is more commonly used, but the survival function seems more natural for queueing theory, which is about waiting for things that haven't happened yet.

[13] Outside of queueing theory, some probability distributions have random variables that can be negative. Survival functions can be used in this case, with $S(\infty) = 1$ instead of $(S(0) = 1$, but it is more natural to use the cumulative distribution function in this context.

[14] It is customary in probability to make one of the inequalities be $\leq$ rather than $<$. This is a distinction without a differrence, since the probability of $T = t$ is 0 for any t.

[15] These notes assume some familiarity with the basic mathematics of probability distributions. Some readers may need to consult other sources to review this mathematics.

[16] I am using the symbol $\cdots$ to indicate some calculation details that the reader should be able to fill in. Here you can integrate by parts, or you can use symbolic math software such as Wolfram Alpha.

2.4 Key Properties of the Exponential Distribution

Why are we studying the exponential distribution? Because it has the history-free property that we identified as characteristic of arrival times for independent customers. To see this, note first that the probability of $T > t$ is given by $S(t)$ from formula (1.6) . Now suppose a time t_1 passes without an arrival. Then the probability that the next arrival occurs between times t_1 and $t_1 + t$ is given by the conditional probability formula

$$P\{T > t_1{+}t \mid T > t_1\} = \frac{P\{T > t_1{+}t\}}{P\{T > t_1\}} = \frac{e^{-\alpha(t_1+t)}}{e^{-\alpha t_1}} = e^{-\alpha t} = P\{T > t\}.$$

This says that the probability that the next arrival will require a time greater than t, given that the interval t_1 has already passed, is the same as the initial probability that the next arrival requires a time greater than t from the start.

This lack of history property is what will allow us to analyze queue models theoretically. Because of this convenience, we'll want to use the exponential distribution for service times as well as arrival times. There are some difficulties in using an exponential distribution service model, however. Note that the function f is always decreasing. This means that the probability that the next arrival will occur between times 0 and t_1 is larger than the probability that the next arrival will occur between times t_1 and $2t_1$, for any t_1. See Figure 1.1. This is a reasonable property for arrival processes, but it is questionable for service processes. Think about the amount of time it takes to make your purchase in the convenience store. It is certainly more likely to be in the range 0 to 5 minutes than 5 to 10 minutes. But is it more likely that the time will be between 0 and 10 seconds than between 10 and 20 seconds? Almost certainly not. This is an issue we'll address in Section 5.

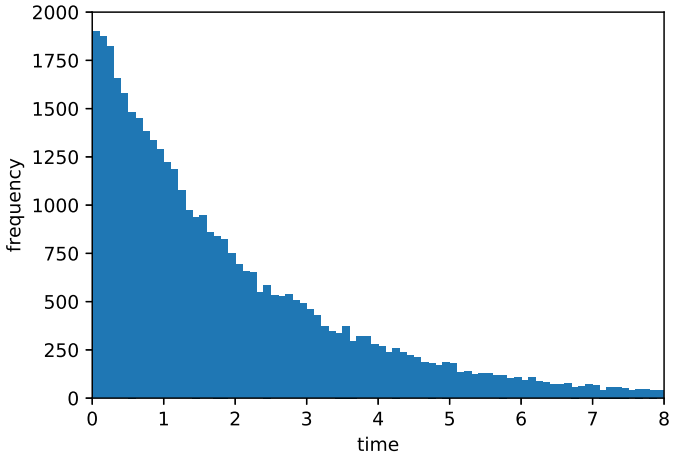

Figure 2.2: Histogram of 40000 values drawn from the exponential distribution with mean rate $\mu = 0.5$. Note that the mean time is $\mu_T = 1/\mu = 2$.

2.5 Building Intuition

It is easy to have too much faith in theoretical results and too little understanding of the unpredictability of individual events. One should experiment with computer simulations to see how much the actual mean rate can deviate from the theoretical mean rate when the duration of the simulation is short. Figure 1.2 shows the means for samples of size 100 drawn from the exponential distribution.[17] Note that it is not especially unusual for the actual mean of 100 trials to be less than 80% or more than 120% of the theoretical value.

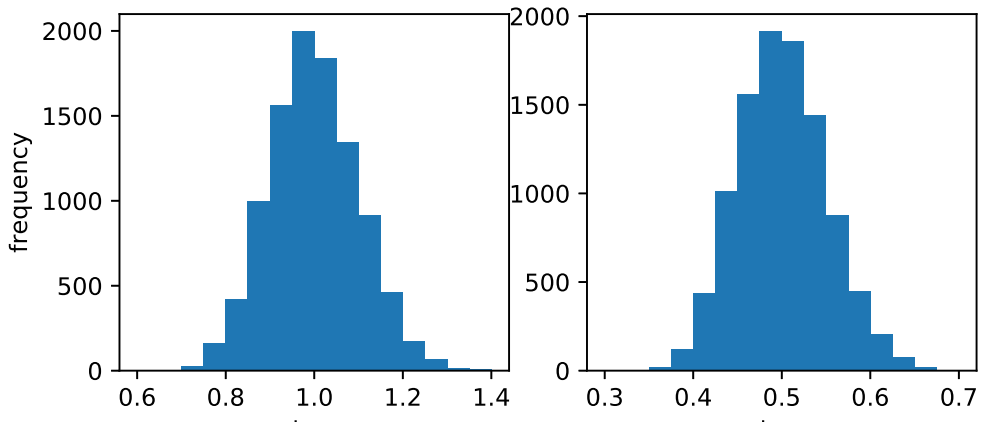

Figure 2.3: Histograms of means of 10000 samples of size 100, with rates $\alpha = 1$ (left) and $\alpha = 2$ (right).

3 Analysis of Finite Queues (see Hillier and Lieberman 17.5,6)

Learning Objectives

1. Understand the assumptions necessary for rate diagrams to be valid.
2. Be able to sketch rate diagrams and use them to determine steady-state probabilities for systems with limited queue size.

We've seen that the four steady-state performance measures are related by three simple equations,

$$L = \bar{\lambda} S, \qquad L_q = \bar{\lambda} W, \qquad S = W + \frac{1}{\mu}, \tag{3.1}$$

where $\bar{\lambda}$ is the mean arrival rate for the system, leaving us one equation short. The usual way to complete the set is to compute the queue length in terms of the steady-state probabilities

[17] See the R program ExpDistTest.R or the python program ExpDistTest.py.

for the system states:

$$L_q = P_{s+1} + 2P_{s+2} + 3P_{s+3} + \cdots = \sum_{n=s+1}^{\infty} (n-s)P_n. \tag{3.2}$$

In general, there is no way to compute these steady-state probabilities other than to estimate them with simulations. However, in the case where both arrival and service times are exponentially distributed, we can compute the probabilities as the solutions of the steady-state versions of a set of differential equations. We'll do this in two stages. In this section we'll consider the relatively straightforward case that occurs when the queue size is limited. Then in Section 4, we'll deal with the additional mathematics needed for queues that are unlimited in size. Keep in mind that the formulas require careful interpretation for systems with limited queue sizes because customers can be turned away. This means that the mean arrival rate $\bar{\lambda}$ is actually less than the usual arrival rate λ.

3.1 An Example: the M/M/2/3/∞ Queue

We begin with the simplest example that shows all of the features of the method: the M/M/2/3/∞ system. This system has just four possible states, $n = 0$ to $n = 3$. Each arrival and service completion event changes the state of the system. The probabilities of the different states change according to rates based on the current state and the mean arrival and service rates. These are given in Figure 3.1 in terms of λ and μ.

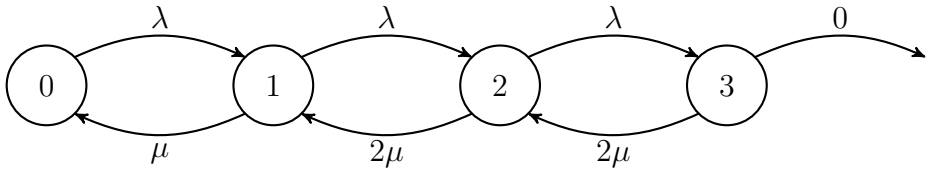

Figure 3.1: A schematic diagram of the M/M/2/3 queue

Each arrow in the figure is labeled with a rate parameter corresponding to the corresponding transition. On the upper arrows, we see that customers enter at rate λ when the state is 0, 1, or 2, but they do not enter at all when the state is 3. The model assumes that any customer who tries will leave, never to return.[1] On the lower arrows, the service rate parameter is μ when the state is 1, but 2μ when the state is either 2 or 3. This is because both servers are busy at these states, so the overall service rate when the system is full is twice as large as the rate for one server.

3.1.1 Steady-State Equations for the M/M/2/3/∞ Queue

Exponentially distributed times correspond to rates that are proportional to the probability of the appropriate state. For example, we have the rate parameter λ for arrivals at state 0. This corresponds to a rate of λP_0, where P_0 is the probability of state 0. Similarly, we have the rate

[1] An example would be a gasoline station that has 2 pumps and space for one car to wait. Any customer who arrives when the system is full will go to a different station.

parameter μ for service completions at state 1, corresponding to the rate μP_1. These two rates change the probability of state 0 according to the differential equation

$$\frac{dP_0}{dt} = \mu P_1 - \lambda P_0.$$

If the queue system opens at time 0 with no customers, then $P_0(0) = 1$ and $P_1(0) = 0$, so the probability P_0 will initially decrease at rate $-\lambda$. Over time, P_0 will decrease and P_1 will increase until they reach steady-state values. Since they are no longer changing at that point, the differential equation reduces to the steady state equation $\mu P_1 = \lambda P_0$, which we write in terms of the single parameter $\gamma = \lambda/\mu$ as

$$P_1 = \frac{\lambda}{\mu} P_0 = \gamma P_0. \tag{3.3}$$

The two rates that affect P_0 also affect P_1, but in the opposite way. Two other rates affect P_1 as well. In sum,

$$\frac{dP_1}{dt} = 2\mu P_2 - \lambda P_1 - \mu P_1 + \lambda P_0,$$

or

$$\frac{dP_1}{dt} = 2\mu P_2 - \lambda P_1 - \frac{dP_0}{dt}.$$

At steady state, both P_1 and P_0 are unchanging, so the last equation reduces to $2\mu P_2 = \lambda P_1$, which we write as

$$P_2 = \frac{\lambda}{2\mu} P_1 = \frac{\gamma}{2} P_1 = \frac{\gamma^2}{2} P_0. \tag{3.4}$$

A similar study of the changes in P_2 yields

$$P_3 = \frac{\gamma}{2} P_2 = \frac{\gamma^3}{4} P_0. \tag{3.5}$$

The overall change in P_3 is already 0 from equation (3.5), so we do not get a fourth equation.

3.1.2 Steady-State Probabilities for the M/M/2/3/∞ Queue

In addition to the three equations we've found from the differential equations, we also know that the four probabilities must add up to 1. Having written each of the probabilities with $n > 0$ in terms of P_0 , we can now use this fact to calculate P_0. We have

$$1 = P_0 + P_1 + P_2 + P_3 = \left(1 + \gamma + \frac{\gamma^2}{2} + \frac{\gamma^3}{4}\right) P_0; \tag{3.6}$$

hence,

$$P_0 = \frac{1}{1 + \gamma + \frac{\gamma^2}{2} + \frac{\gamma^3}{4}}, \tag{3.7}$$

and the rest of the probabilities are easily found from formulas (3.3)–(3.5).

For example, if the mean arrival and service times are both 2, then $\gamma = 1$ and the probabilities are

$$P_0 = \frac{4}{11}, \qquad P_1 = \frac{4}{11}, \qquad P_2 = \frac{2}{11}, \qquad P_3 = \frac{1}{11}. \tag{3.8}$$

3.1.3 The Performance Measures for the M/M/2/3/∞ Queue

The mean system state is an average of the possible states 0, 1, 2, and 3, weighted by the probabilities of those states. In our specific example with $\gamma = 1$ and probabilities from fomulas (3.8), we have

$$L = 0P_0 + 1P_1 + 2P_2 + 3P_3 = 1. \tag{3.9}$$

Similarly, the mean number of customers in the queue is

$$L_q = 0P_0 + 0P_1 + 0P_2 + 1P_3 = \frac{1}{11}, \tag{3.10}$$

since there is 1 customer in the queue at state 3 and no customers in the queue at states 0 through 2.

Now is when it gets tricky. If we naively use the formulas

$$L = \lambda S, \qquad L_q = \lambda W,$$

with $\lambda = 2$, we get

$$S = \frac{L}{2} = \frac{1}{2}, \qquad W = \frac{L_q}{2} = \frac{1}{22}.$$

But these results cannot be correct, because they do not satisfy $S - W = 1/\mu = 1/2$. The problem here is that the $\bar{\lambda}$ in Little's formulas (3.1ab) is the mean arrival rate *for the system* whereas the λ weve been using in Figure 3.1 and our calculation of the probabilities is the mean arrival rate *when the system is not full.*[2] Instead we can compute $\bar{\lambda}$ as a weighted average of the arrival rates for the different states:

$$\bar{\lambda} = \lambda P_0 + \lambda P_1 + \lambda P_2 + 0P_3 = \lambda(1 - P_3). \tag{3.11}$$

The fraction $\bar{\lambda}/\lambda = 1 - P_3$ is another critical performance measure of the system, representing the fraction of customers who are lost because of limited system capacity. In the present example, 1/11 of potential customers are lost, so the actual mean arrival rate is (10/11)*2, or 20/11. Little's formulas then give

$$W = \frac{L_q}{20/11} = \frac{1}{20}, \qquad S = \frac{L}{20/11} = \frac{11}{20}.$$

As required, these satisfy $S - W = 1/\mu$.

3.2 The General Case for M/M/s/K/∞

In general, the arrival and service rates could vary with the system state for a number of reasons. Perhaps server 1 is faster than server 2, or perhaps some of the potential customers balk (leave the system without service) before the queue is full. So a more general rate diagram (see Figure 3.2) would use λ_n and μ_n as the rate parameters for arrivals and service, respectively, when there are n customers in the system. With a maximum state of K, we have $\lambda_K = 0$ and we don't need μ_{K+1}.

[2] Hillier and Lieberman use "λ" in Little's formulas, which is a misleading use of the notation.

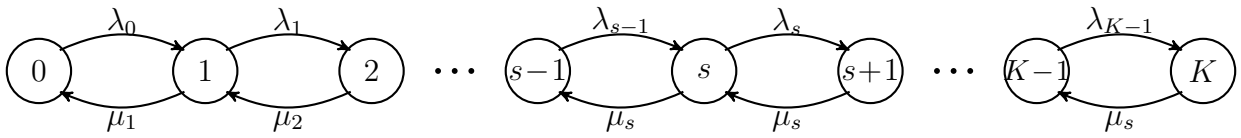

Figure 3.2: A schematic diagram of the most general form of M/M/s/K queue

The differential equations are constructed in the same way as in the previous example. There are $K+1$ unknown probabilities related by K steady-state equations, giving us

$$\mu_1 P_1 = \lambda_0 P_0, \qquad \mu_2 P_2 = \lambda_1 P_1, \qquad \ldots, \qquad \mu_s P_K = \lambda_{K-1} P_{K-1}. \tag{3.12}$$

These equations allow us to write all of the probabilities in terms of P_0:

$$P_1 = \frac{\lambda_0}{\mu_1} P_0, \qquad P_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} P_0, \qquad \ldots, \qquad P_K = \frac{\lambda_0 \lambda_1 \cdots \lambda_{K-1}}{\mu_1 \mu_2 \cdots \mu_{s-1} \mu_s^{K-s+1}} P_0. \tag{3.13}$$

Then we get P_0 by setting the sum of the probabilities P_0 through P_K equal to 1. Once the probabilities are known, we get L and L_q from

$$L = \sum_{n=0}^{K} n P_n, \qquad L_q = \sum_{n=s}^{K} (n-s) P_n, \tag{3.14}$$

and then

$$\bar{\lambda} = \sum_{n=0}^{K-1} \lambda_n P_n, \qquad W = \frac{L_q}{\bar{\lambda}}, \qquad S = \frac{L}{\bar{\lambda}}. \tag{3.15}$$

We can confirm our results by checking $S - W = 1/\mu$. Note that in the usual case where $\lambda_n = \lambda$ for $n < K$, we have

$$\bar{\lambda} = (1 - P_K)\lambda \tag{3.16}$$

3.3 An Example with Finite Calling Population

Suppose we have an M/M/2 system in which the system size is limited not by its own capacity but by the number of potential customers. An example would be a system where the customers are the three equivalent machines in a factory and there are two repair crews. This would be an M/M/2/3/3 queue system.

Note that the rate μ is usually taken to be the service rate for each server, while the rate λ is usually taken to be the arrival rate for the whole population. When the calling population is limited, it makes more sense to take λ to be the arrival rate for each potential customer.[3] Thus, the arrival rates are 3λ, 2λ, and λ when the system size is 0, 1, and 2, respectively, as shown in Figure 3.3.

[3]Note here the important point that symbols do not have fixed meanings; rather, they mean what we define them to mean. There are only so many symbols available, so all of them have to have different meanings in different contexts. The choice of what meaning a symbol has in a given context is part of the process of developing the mathematics. Mathematics is not just about knowing formulas; it is also about understanding what the formulas mean in the context in which they occur.

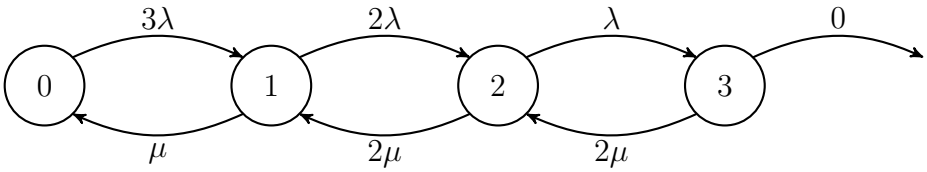

Figure 3.3: A schematic diagram of the M/M/2/3/3 queue

The reader should go through the calculations to confirm the result

$$P_0 = \frac{1}{1 + 3\gamma + 3\gamma^2 + \frac{3}{2}\gamma^3} \tag{3.17}$$

and also confirm the specific probabilities given in the example of Section 2 for the case $\lambda = 1$, $\mu = 2$.

4 M/M/s Queue Theory (see Hillier and Lieberman 17.5,6)

Learning Objectives

1. Be able to derive the formulas for L_q and L for M/M/1 and M/M/2 queues.
2. Use a computer program that calculates L_q and L for given λ, μ, and s for an M/M/s queue system to obtain results and plot them as graphs.

In Section 3 we learned the basic procedure for analyzing the steady-state probability distribution and performance measures for M/M queues with a limited queue size. The limited queue size means that there are only a finite number of unknowns. Now we consider queue systems with unlimited queue size. We further assume that the arrival rates are the same for all states and each server has the same service rate. We can use the same method as before, but we will have infinitely many unknown probabilities, leading to a formula for P_0 that involves an infinite sum. The needed sum can be computed analytically (although the algebra is tedious). We therefore consider systems that have the rate diagram shown in Figure 4.1. As is commonly done, we will refer to these systems in this section as M/M/s rather than M/M/$s/\infty/\infty$.
Some general infinite sum formulas will be needed. First is the geometric series formula:

$$1 + x + x^2 + x^3 + \cdots = \frac{1}{1 - x}, \qquad 0 < x < 1. \tag{4.1}$$

Note that the formula works for $|x| < 1$, but we will only use it with positive x. The second formula comes from differentiating formula (4.1) term by term:

$$1 + 2x + 3x^2 + \cdots = \frac{1}{(1 - x)^2}, \qquad 0 < x < 1. \tag{4.2}$$

We'll work out the details for M/M/1 and M/M/2 as preludes to the general case.

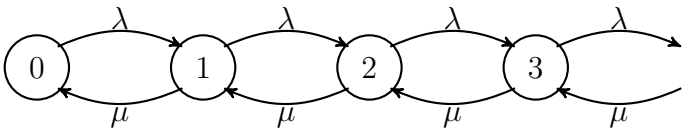

Figure 4.1: A schematic diagram of the M/M/1 queue

4.1 The Steady-State Probabilities for M/M/1

Figure 4.1 shows the M/M/1 queue system. Note that

$$\rho = \gamma = \frac{\lambda}{\mu}, \tag{4.3}$$

so it makes no difference whether we use ρ or γ. From the rate diagram we have

$$P_1 = \rho P_0, \quad P_2 = \rho P_1 = \rho^2 P_0, \quad P_3 = \rho^3 P_0, \quad \ldots, \quad P_n = \rho^n P_0.$$

All the probabilities sum to 1, so

$$1 = P_0 + \rho P_0 + \rho^2 P_0 + \cdots = P_0(1 + \rho + \rho^2 + \cdots) = \frac{P_0}{1-\rho},$$

where the infinite sum follows from formula (4.1). Thus, we have probabilities

$$P_0 = 1 - \rho, \qquad P_n = \rho^n(1-\rho). \tag{4.4}$$

Next we can compute L_q using these results. Note that the queue is empty when the system state is either 0 or 1 and $n-1$ for larger states.

$$L_q = 0P_0 + 0P_1 + P_2 + 2P_3 + 3P_4 + \cdots = P_2 + 2\rho P_2 + 3\rho^2 P_2 + \cdots = P_2(1 + 2\rho + 3\rho^2 + \cdots).$$

We can now use the sum formula (4.2) to obtain the result

$$L_q = \frac{P_2}{(1-\rho)^2} = \frac{\rho^2(1-\rho)}{(1-\rho)^2} = \frac{\rho^2}{1-\rho}. \tag{4.5}$$

The expected system size, using formula (2.5), is then

$$L = L_q + \gamma = \cdots = \frac{\rho}{1-\rho}. \tag{4.6}$$

4.2 The Steady-State Probabilities for M/M/2

The M/M/2 system is illustrated by the rate diagram of Figure 4.2. The key parameters are

$$\gamma = \frac{\lambda}{\mu}, \qquad \rho = \frac{\lambda}{2\mu}; \tag{4.7}$$

at each point in the calculation, we'll try to use the one that is most convenient.

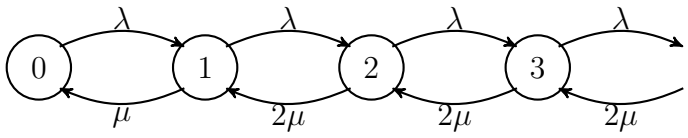

Figure 4.2: A schematic diagram of the M/M/2 queue

From the rate diagram, we have

$$P_1 = \gamma P_0, \quad P_2 = \rho P_1, \quad P_3 = \rho^2 P_1, \quad \ldots. \tag{4.8}$$

The probabilities sum to 1, so

$$1 = P_0 + P_1 + \rho P_1 + \rho^2 P_1 + \cdots = P_0 + P_1(1 + \rho + \rho^2 + \cdots) = P_0 + \frac{P_1}{1-\rho}. \tag{4.9}$$

Using $P_1 = \gamma P_0 = 2\rho P_0$ and some simplification, we obtain

$$P_0 = \frac{1-\rho}{1+\rho}, \tag{4.10}$$

after which the other probabilities can be calculated from formulas (4.8). Following the same method as for $s = 1$ (note that the queue size is 0 until state $n = 3$ and then $n - 2$), the expected queue size is

$$L_q = P_3 + 2P_4 + 3P_5 + \cdots = (1 + 2\rho + 3\rho^2 + \cdots)P_3 = \frac{P_3}{(1-\rho)^2} = \cdots = \frac{2\rho^3}{1-\rho^2}. \tag{4.11}$$

Finally, the expected system size is

$$L = L_q + \gamma = L_q + 2\rho = \cdots = \frac{2\rho}{1-\rho^2}. \tag{4.12}$$

4.3 The Steady-State Probabilities for M/M/s with $s \geq 2$

The same procedure used for M/M/2 works for M/M/s with $s \geq 2$, but parts of the calculation get much messier. There are a lot of different ways to organize the work; the one used here is at least arguably the simplest. The key is based on maintaining the structure of formula (4.9), dividing the sum into a portion consisting of early terms (here just P_0) and a portion that is easily computed using formula (4.1).

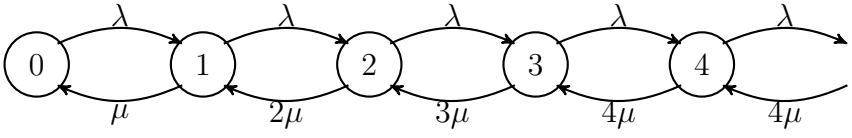

Figure 4.3: A schematic diagram of the M/M/4 queue

In general, we have

$$\rho = \frac{\lambda}{s\mu}, \qquad \gamma = \frac{\lambda}{\mu}. \tag{4.13}$$

Figure 4.3 illustrates the M/M/4 queue system as an example for the general case. We begin by looking at just the portion of the rate diagram that starts at $n = 3$. In terms of P_3, we have

$$P_4 = \rho P_3, \qquad P_5 = \rho^2 P_3, \qquad \ldots. \tag{4.14}$$

On the theory that it is best to do the easier parts of a problem first, we now derive a simple formula that calculates L_q in terms of P_3:

$$L_q = P_5 + 2P_6 + 3P_7 + \cdots = (1 + 2\rho + 3\rho^2 + \cdots)P_5 = \frac{P_5}{(1-\rho)^2} = \frac{\rho^2}{(1-\rho)^2}P_3.$$

This formula is convenient because it generalizes easily to the case where s is unspecified. The current version uses P_3 because P_3 begins the chain appearing in formula (4.14). In general, the result is

$$L_q = \frac{\rho^2}{(1-\rho)^2}P_{s-1}. \tag{4.15}$$

Hence, we need only find a formula for P_{s-1}, and then we'll have L_q.

To calculate P_{s-1} with $s = 4$, we can start by adding the probabilities P_3 and up:

$$P_3 + P_4 + P_5 + \cdots = (1 + \rho + \rho^2 + \cdots)P_3 = \frac{P_3}{1-\rho}.$$

In the case of general s, the corresponding formula will have to start at state $s - 1$:

$$P_{s-1} + P_s + P_{s+1} + \cdots = \frac{P_{s-1}}{1-\rho}.$$

The hard part of calculating P_{s-1} is the portion of the probability sum that precedes P_{s-1}. For $s = 4$, we have

$$P_1 = \gamma P_0, \qquad P_2 = \frac{\gamma}{2}P_1 = \frac{\gamma^2}{2}P_0, \qquad P_3 = \frac{\gamma^3}{6}P_0.$$

Thus,

$$P_0 + P_1 + P_2 = \left(1 + \gamma + \frac{\gamma^2}{2}\right)P_0.$$

Combining the two sums yields

$$1 = (P_0 + P_1 + P_2) + (P_3 + P_4 + \cdots) = \left(1 + \gamma + \frac{\gamma^2}{2}\right)P_0 + \frac{P_3}{1-\rho}.$$

The usual procedure here would be to use $P_3 = (\gamma^3/6)P_0$ to replace P_3 and then have an equation for P_0. But since it is P_3 that we want, we can save effort by instead using the relationship between P_3 and P_0 to replace P_0 by $(6/\gamma^3)P_3$, giving us

$$1 = (P_0 + P_1 + P_2) + (P_3 = P_4 + \cdots) = \frac{6}{\gamma^3}\left(1 + \gamma + \frac{\gamma^2}{2}\right)P_3 + \frac{P_3}{1-\rho}.$$

We can now solve this equation for P_3, with the messy result

$$P_3 = \frac{1}{\frac{1}{1-\rho} + \frac{6}{\gamma^3}\left(1 + \gamma + \frac{\gamma^2}{2}\right)}.$$

Generalizing to arbitrary s, we get

$$P_{s-1} = \frac{1}{\frac{1}{1-\rho} + R_s}, \tag{4.16}$$

where, to separate messy parts into different formulas, we have defined R_s by

$$R_s = \frac{(s-1)!}{\gamma^{s-1}} \sum_{n=0}^{s-2} \frac{\gamma^n}{n!}. \tag{4.17}$$

While it has taken a lot of work to get there, the result is a set of relatively simple formulas that we can use to calculate L: Given s, λ, and μ, we can calculate γ and ρ and then get R_s from formula (4.17), P_{s-1} from formula (4.16), and L_q from formula (4.15). L then follows from

$$L = L_q + \gamma, \tag{4.18}$$

which follows from the general formulas $L = \lambda S$, $L_q = \lambda W$, and $S - W = 1/\mu$.

The reader should confirm that these formulas yield formula (4.11) for $s = 2$.

Example

Suppose $\rho = 0.8$ for the M/M/3 queue. Since $\rho = \gamma/s$, this means $\gamma = 2.4$. Formulas (4.17), (4.16), (4.15), and (4.18) yield

$$R_3 = \frac{2}{\gamma^2}(1 + \gamma) \approx 1.181,$$

$$P_2 = \frac{1}{5 + 1.181} \approx 0.1618,$$

$$L_q = \frac{\rho^2 P_2}{(1-\rho)^2} = 2.589, \qquad L = L_q + \gamma = 4.989.$$

For cases where $s > 3$ or values are desired for a variety of ρ values, one would not want to calculate the results manually. it is better to write a computer program that defines a function that implement these formulas to calculate L from input values γ and s. Repeatedly calling this program with a range of γ values (making sure $\gamma < s$) allows a large number of results to be presented in the form of a graph. The results of such a program are displayed in Figure 4.4. If we wanted to emphasize the relationships between the formulas for different numbers of servers, it would be better to use ρ as the horizontal coordinate. Using γ as the horizontal coordinate produces a plot that better illustrates the effect of adding servers, making the plot more useful in a decision-making context.

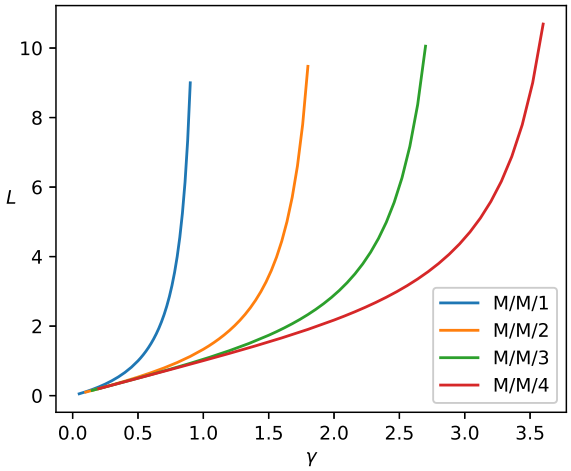

Figure 4.4: Dependence of expected system size on arrival-service ratio $\gamma = \lambda/\mu$ for M/M/s queues.

5 The Erlang Distribution (see Hillier and Lieberman 17.7)

Learning Objectives

1. Be able to determine standard deviation for the Erlang distribution.
2. Be able to use measured data to choose parameter values for the Erlang distribution.
3. Understand the theoretical connection between the Erlang distribution and the exponential distribution.
4. Be able to prepare and explain the rate diagram for the M/E$_2$/1 system and write down the balance equations.

The Erlang distribution (known outside of queueing theory as the gamma distribution) is defined by its probability density function

$$f(t) = \frac{(k\mu)^k}{(k-1)!}\, t^{k-1} e^{-k\mu t}, \tag{5.1}$$

leading to the survival function

$$P\{T > t\} = \int_t^\infty \frac{(k\mu)^k}{(k-1)!}\, t^{k-1} e^{-k\mu t}\, dt. \tag{5.2}$$

Note that if $k = 1$, the probability density function is just

$$f(t) = \mu e^{-\mu t},$$

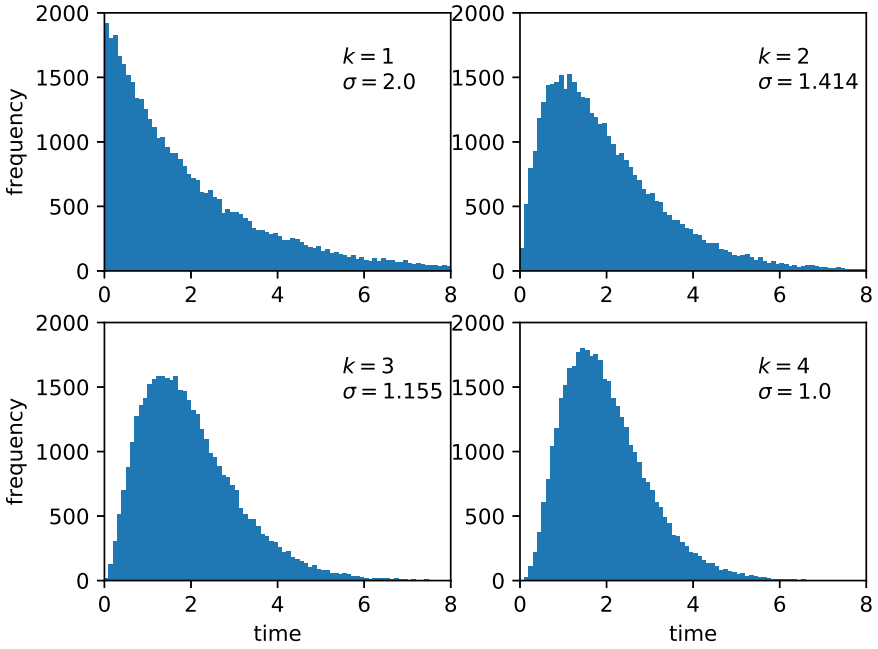

Figure 5.1: Histograms of 40000 values drawn from the Erlang distribution E_k, all with $\mu = 0.5$. Note that the mean in each case is $\mu_T = 1/\mu = 2$.

which is the exponential distribution; hence, the Erlang distribution generalizes the latter by adding the parameter k, which must be a positive integer.

Figure 5.1 shows histograms of E_1 to E_4, all with mean time $\mu_T = 2$. The plots show that choosing $k > 1$ changes the shape of the distribution. The exponential distribution, corresponding to $k = 1$, is highly skewed, meaning that the peak at $t = 0$ is far from the mean value of $\mu_T = 1/\mu$. The skewness decreases as k increases.

5.1 Significance of the Erlang distribution

As seen in the first panel of Figure 5.1, smaller times are always more common than larger times for the exponential distribution, which is generally plausible for arrival times but can be problematic for service times. For example, if the mean time required to make a purchase at the grocery store is 5 minutes, then surely it is more likely for a purchase to be completed in the second minute than the first. A less skewed distribution is usually more realistic.

While skewness has no natural metric, we can use the standard deviation as a surrogate since a high degree of skewness corresponds to a larger standard deviation. There is a simple formula for the standard deviation of the Erlang distribution:

$$\sigma = \frac{1}{\mu\sqrt{k}}. \tag{5.3}$$

Rearranging this formula allows us to calculate a best fit value of k for a measured mean and

standard deviation as

$$k = \left(\frac{\mu_T}{\sigma}\right)^2. \tag{5.4}$$

With empirical data, this calculation will not produce an integer, but we can round the result off to the nearest integer to obtain a suitable Erlang model for the unknown distribution.[1]

Even when the standard deviation is quite a bit less than the mean, it is common in practice to use the exponential distribution without any justification, although this could yield poor results.[2] The decision between a distribution that makes for easy calculation and one that better approximates reality should be based on the difference seen in the final results. If the simpler distribution gives results that are not much different than the better-fitting distribution, then there is no harm in using it, but the choice should not be made without investigating this question.

Among the alternatives to the exponential distribution, the Erlang distribution has a significant theoretical advantage. Unless the queue system can be represented by a rate diagram, the only way to obtain results for $s > 1$ is with a simulation. This is not very satisfactory, as simulations take a long time to converge to the mean.[3] But since $\mathrm{M/E}_k/s$ queue systems can be represented using rate diagrams, they ultimately lead to formulas that allow the probabilities to be calculated using large sums rather than simulations. (See Section 5.2 and Appendix B.)

5.2 Probabilities for $\mathrm{M/E}_2/1$

The distribution E_k with overall mean time $1/\mu$ corresponds to a service process that consists of a sequence of k subtasks, each exponentially distributed with mean time $1/(k\mu)$. This means that rate diagrams and steady-state equations are possible, but only with each service phase considered separately. The state of the system needs to define both the number of customers in the system (n) and the phases for each of the servers. For the $\mathrm{M/E}_3/3$ system, there will be just one state with $n = 0$. With $n = 1$, there will be three states because the server could be in any of the three phases. There will be nine different states when $n = 2$ because each of the two servers could be in any of the three phases. Similarly, there will be 27 different states for each $n > 2$. Clearly this is going to be difficult for $k > 2$ or more than a couple of servers. Here we consider only the simplest case of the E_2 distribution with one server. The $\mathrm{M/E}_2/2$ system is analyzed in Appendix B.

For the queue system with $\mathrm{M/E}_2/1$ service, we can use a single parameter p to represent the current service phase. Thus, we need states $n = 0$ and (n, p) with each $n > 0$ and $p = 1$ or 2. These states can be conveniently arranged in the rate diagram of Figure 5.2. Note the patterns

[1] In the real world, one must always guard against trying to draw too strong a conclusion from uncertain data. I call this the "measure it with your hand, mark it with chalk, cut it with a laser" fallacy . The Erlang distribution will fit any likely service distribution well enough for any practical purpose. Similarly, one could probably get a better fit to data with a family of distributions that has more than two parameters, but the data is almost certainly too uncertain to justify this level of assumed precision.

[2] This is the mathematical equivalent of looking for your lost keys where the light is bright rather than where you think you might have left them.

[3] Each of the jagged curves in Figure 5.4 at the end of this section was obtained by running a simulation for a duration corresponding to the expected time required for 100,000 service completions. For a service process with a mean time of 5 minutes, the server can do a little less than 100 customers in a day, so 100,000 service completions represents almost three years! Of course the real situation restarts at the beginning of each day, so we should not expect any particular day to be very close to the average.

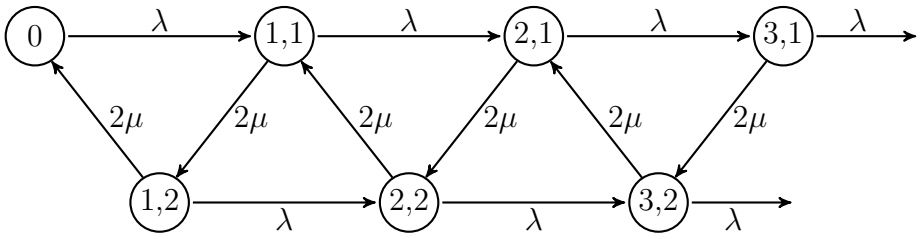

Figure 5.2: A schematic diagram of the M/E$_2$/1 queue

of arrivals and service. Arrival from state 0 results in state (1,1) because the new customer will begin phase 1 service. Any arrivals in other states will be customers who must wait in the queue; thus, arrivals when $n > 0$ increase n while leaving p unchanged. Arrivals are marked by left-right arrows on each horizontal row of nodes in the graph. To see what happens with service completions, consider the example with current state (3,1), meaning that there is one customer in phase 1 of service and two customers waiting in the queue. Now suppose we have a string of service completions with no intervening arrivals. The first service completion moves the customer being served from phase 1 to phase 2; hence, the system moves from (3,1) to (3,2). The next completion finishes that customer, which decreases the system size by one. The next customer in line begins phase 1 service, so the system moves from (3,2) to (2,1). Similarly, consecutive service completions move the system from (2,1) to (2,2) to (1,1) to (1,2) to 0. Of course the chain can be broken by an arrival, but that merely serves to move the system to a state we have already studied. Service arrows in the rate diagram point from right to left, always moving from one row to the other. These arrows are labeled with the rate 2μ because the mean completion time for half of a full service sequence is $\mu_T/2 = 1/(2\mu)$. Alternatively, we can think of a half service process as running twice as fast as a full service process.

Once we have the rate diagram, we can write down the steady-state equations, working from left to right through the diagram:

$$2\mu P_{12} = \lambda P_0 \tag{5.5}$$
$$2\mu P_{11} = (2\mu + \lambda) P_{12} \tag{5.6}$$
$$2\mu P_{22} + \lambda P_0 = (2\mu + \lambda) P_{11} \tag{5.7}$$
$$2\mu P_{21} + \lambda P_{12} = (2\mu + \lambda) P_{22} \tag{5.8}$$
$$2\mu P_{32} + \lambda P_{11} = (2\mu + \lambda) P_{21} \tag{5.9}$$
$$2\mu P_{31} + \lambda P_{22} = (2\mu + \lambda) P_{32} \tag{5.10}$$
$$2\mu P_{42} + \lambda P_{21} = (2\mu + \lambda) P_{31} \tag{5.11}$$

for states 0, (1,2), (1,1), (2,2), (2,1), (3,2), and (3,1), respectively, and so on.

Our goal is to obtain formulas for the total probability P_n for each system size. We don't actually care about how that probability is divided between phases. After a fair amount of algebra (see Appendix A), we obtain a multi-formula computational scheme to compute the probabilities P_n in terms of intermediate quantities a_n and b_n:

$$a_1 = \delta, \qquad \delta \equiv \frac{\lambda}{2\mu}, \tag{5.12}$$

	$M/E_2/1$			$M/M/1$		$M/E_2/1$			$M/M/1$
n	a_n	b_n	P_n	P_n	n	a_n	b_n	P_n	P_n
0			0.4000	0.4000					
1	0.3000	0.6900	0.2760	0.2400	7	0.0085	0.0146	0.0058	0.0112
2	0.2070	0.3861	0.1544	0.1440	8	0.0044	0.0075	0.0030	0.0067
3	0.1158	0.2043	0.0817	0.0864	9	0.0023	0.0039	0.0016	0.0040
4	0.0613	0.1062	0.0425	0.0518	10	0.0012	0.0020	0.0008	0.0024
5	0.0319	0.0549	0.0220	0.0311	11	0.0006	0.0010	0.0004	0.0015
6	0.0165	0.0283	0.0113	0.0187	12	0.0003	0.0005	0.0002	0.0009

Table 5.1: Intermediate values and probabilities for $M/E_2/1$ as compared to $M/M/1$ for $\rho = 0.6$.

$$a_2 = (1+\delta)^2 a_1 - \delta, \tag{5.13}$$

$$a_3 = (1+\delta)^2 a_2 - 2\delta(1+\delta)a_1, \tag{5.14}$$

$$a_n = (1+\delta)^2 a_{n-1} - 2\delta(1+\delta)a_{n-2} + \delta^2 a_{n-3}, \qquad n > 3, \tag{5.15}$$

$$b_1 = \delta(2+\delta), \tag{5.16}$$

$$b_n = (2+\delta)a_n - \delta a_{n-1}, \qquad n > 1. \tag{5.17}$$

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} b_n}, \tag{5.18}$$

$$P_n = b_n P_0, \qquad n \geq 1. \tag{5.19}$$

The only difficulty with the implementation of this scheme is that the infinite sum must be approximated numerically. The good news is that this problem is significant only when ρ is too close to 1. For $\rho \leq 0.8$, thirty-one terms are enough to produce results that have less than a 0.1% error.

Other $M/E_k/s$ systems can be analyzed in the same fashion; however, there are more subdivisions for each state n. The $M/E_2/2$ system is analyzed in Appendix B; more servers or a larger k make for a much messier derivation.

Example

A queue system has exponentially distributed arrival times with a mean rate of 18 per hour and one server with an average service completion time of 2 minutes and standard deviation of 1.2 minutes. Since the standard deviation is significantly less than the mean service time, we model this system with an Erlang distribution. Formula (5.4) yields a nominal value of 2.8. The E_3 distribution is probably the best choice for accuracy, but we choose E_2 as a reasonable compromise between accuracy and tractability. The actual system performance should be slightly better than what we find with our model because our model has a larger standard deviation of 1.414.

From the data in the narrative, the parameter values are

$$\lambda = 18, \quad \mu = 30, \quad \rho = 0.6, \quad \delta = 0.3.$$

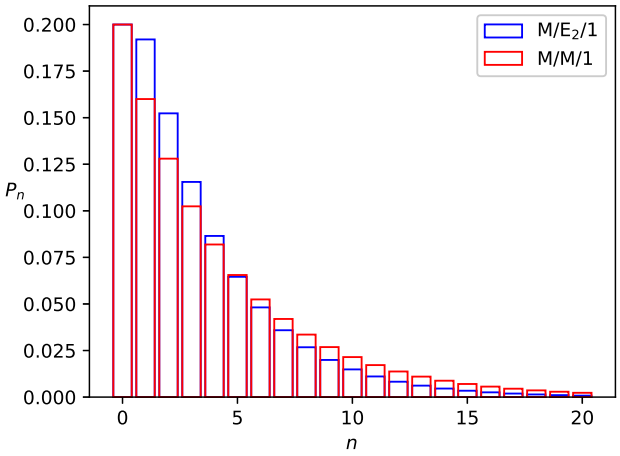

Figure 5.3: Steady-state probabilities for the M/E$_2$/1 and M/M/1 systems with $\rho = 0.8$.

Table 5.1 shows the results of the computation, along with the corresponding probabilities for the M/M/1 system.

Figure 5.3 presents a visual comparison of M/E$_2$/1 and M/M/1 for $\rho = 0.8$. Note that the probabilities for large states are greater for M/M/1 than for M/E$_2$/1. This is because higher variability of service times results in worse performance. Curiously, P_0 is the same for both service distributions. This would seem to be a coincidence, but it appears to hold for all values of ρ, so there is probably some subtle reason. While the other probabilities seem very similar, they are enough different that the expected system sizes differ by 15%, with 1.5 for M/M/1 and 1.275 for M/E$_2$/1.

5.3 Results

Figure 5.4 compares the performance of systems with 1, 2, and 3 servers and service distributions M, E_2, and D. Four different computation methods were used to plot these curves:

1. The M/M/s analytical results were used for M/M/1, M/M/2, and M/M/3.

2. The formulas developed in this section were used for M/E$_2$/1, and similar formulas were used for M/E$_2$/2.

3. The M/G/1 formula was used for M/D/1.

4. Simulations with durations equivalent to 100000 service completions were used for M/E$_2$/3, M/D/2, and M/D/3. These three curves are jagged in appearance because a duration of 100000 service completions is not quite enough to guarantee that the observed mean will match the theoretical mean. These curves would eventually be smooth if the simulations were run for a much longer period of time.

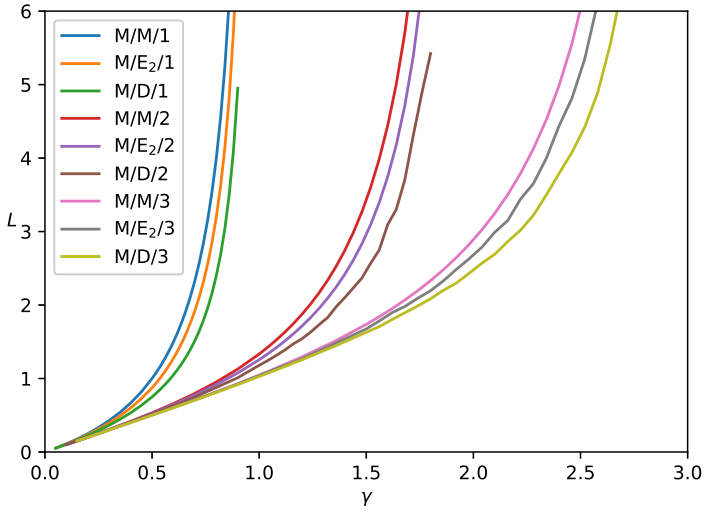

Figure 5.4: Dependence of expected system size on arrival-service ratio $\gamma = \lambda/\mu$ for M/M/s, M/E$_2$/s and M/D/s queue systems.

6 Queue System Costs (see Hillier and Lieberman 17.10, 26.3)

Learning Objectives

1. Be able to explain the difference between direct and indirect costs of queue systems.
2. Be able to identify specific direct and indirect costs for queue system examples.
3. Be able to choose an appropriate model to represent direct and indirect costs for queue system examples.
4. Be able to compute indirect costs using weighted averages with either analytical summation formulas or numerical computation as appropriate.

Now that we have acquired some familiarity with the mathematics of queue systems, it is time to consider questions about how to use the results to make decisions. We'll do this in two steps: this section is about ways to quantify the cost of a queue system, and the following section is about types of queue system optimization problems that use this section's cost models to derive the objective function to be optimized. There are two types of costs involved in a queue system. Direct costs are those that are associated with the operation of the system, such as the wages paid to servers or routine maintenance for machines used for service. Indirect costs are those associated with performance, such as lost business caused by slow service.[1] We consider each of these in turn.

[1]The cost is actually based on the "poorness" of the performance, but there is no simple way to say this.

6.1 Direct (Operational) Cost

For most queue system design problems, the direct cost (DC) is simply proportional to the number of servers:

$$\text{DC} = C_s s, \tag{6.1}$$

where C_s is the cost per unit time for each server. In some cases, however, the cost per unit time for each server is not a fixed quantity. For example, we might buy better tools so that service is more efficient, in which case the cost per unit server depends in some way on the mean service completion rate:

$$\text{DC} = C_s f(\mu) s. \tag{6.2}$$

This form assumes that there is some standard service rate μ_0 for which the cost of one server per unit time is C_s. Then f is chosen so that $f(\mu_0) = 1$. It should also be an increasing function, since faster service should cost more.

It is also possible that the arrival parameter λ is not fixed. This would occur for an internal queue system in which the potential customers are machines in a factory. The factory operator could institute a preventive maintenance program that decreases λ. The cost might then be of the form

$$\text{DC} = C_s s + C_m f(\lambda) N, \tag{6.3}$$

where N is the number of potential customers and $C_m f(\lambda)$ is the cost per unit time per potential customer.[2] This function assumes a fixed cost per unit time per server for the service system itself and a preventive maintenance cost of $C_m f(\lambda)$ per unit time per potential customer. The modeling is most convenient if there is a maximum arrival rate λ_0 associated with no preventive maintenance program and a minimum arrival rate λ_m for which the cost is C_m. Then f must be chosen so that $f(\lambda_0) = 0$ and $f(\lambda_m) = 1$.

6.2 Indirect (Performance) Cost

Generally more thought is needed to choose a model for the cost of performance. Doing so requires two decisions, one qualitative and one quantitative. The qualitative decision is the choice of a performance measure to use as the basis for the indirect cost model. Any performance measure can be used, including the probability distributions of the system state, the queue length, the total time the customer is in the system, the time the customer is in the queue, the expected values of these distributions (L, L_q, S, and W, respectively), the fraction of potential customers who balk, the fraction of customers whose time in the queue exceeds some threshold value, and numerous other similar choices. The quantitative decision is the choice of a mathematical model to prescribe the indirect cost as a function of the chosen performance measure. We consider some examples.

[2]This function f has nothing to do with the f in (6.2). As observed in Section 0.1, there simply aren't enough easy-to-read symbols to allow them to be dedicated to one single usage, except for the ubiquitous π and e. We could distinguish the f functions with subscripts, perhaps f_m for (6.2) and f_l for (6.3). But this would only deal with the present instance, as the reader might encounter these symbols in another subject. The only real solution to this problem is to appreciate that symbols need to be understood in the context in which they appear. Readers who have not learned this yet would be well advised to learn it now.

6.2.1 Indirect cost proportional to system size

The simplest choice is to make the indirect cost (IC) a linear function of the system size n. As an example, consider the case of customers that are machines in a factory. Here it makes sense to associate the cost of being in the system with the lost productivity of the machines. If each working machine produces a value of C_n per time unit, then the total cost of lost productivity for n machines is

$$\text{IC} = C_n n. \tag{6.4}$$

This is conceptually more complicated than it sounds. While the actual direct cost is predictable, the actual indirect cost is not. The cost in this formula assumes that we know the state n of the system. The actual indirect cost of the system depends on the mix of system states, which is a random variable. As with everything governed by probability distributions, we need to combine the costs of the different states together into an expected indirect cost, using the probabilities of the states as the weights for the averaging. Fortunately, this works out nicely in this case because we already have a performance measure that is the expected value of n. Thus,

$$E(\text{IC}) = E(C_n n) = C_n E(n) = C_n L. \tag{6.5}$$

6.2.2 Indirect cost proportional to waiting time

For a more typical queue system, where the customers are external to the organization, the connection between the state of the system and the indirect cost is less clear, and a different model might be better. Perhaps the indirect cost for each customer in this case should be some function of the time the customer spends in the system (the sojourn time). Let's use t as the random variable that indicates the sojourn time for a randomly chosen customer. For the simplest case, let's assume it to be linear, so

$$(\text{IC})_c(n) = C_t t, \tag{6.6}$$

where C_t is the cost per unit time for a customer who spends total time t in the system. Thus, $C_t t$ has the dimension of cost per unit time multiplied by time per customer. So $(\text{IC})_c$ is the cost *per customer*. What we actually need is the cost per unit time, so we need to multiply by the number of customers per unit time, which is λ. Thus,

$$\text{IC}(n) = C_t \lambda t. \tag{6.7}$$

Now we can compute the expected value as before.

To calculate the expected indirect cost, we will have to address both the problem of averaging the cost over all the customers and the problem of changing the units. We will work out the details in Section 7.2.3, but here we'll anticipate the answer with a clever guess. The idea is that the cost function is linear, so the expected cost per customer ought to be $C_t S$, since S is the expected value of t. The cost per unit time can be seen on dimensional grounds to be (cost/customer)*(customer/time), and the expected number of customers per unit time is λ, so the final result is

$$E(\text{IC}) = E(C_t \lambda t) = C_t \lambda E(t) = C_t \lambda S. \tag{6.8}$$

6.2.3 Indirect cost function $g(n)$

Suppose the indirect cost of a system is associated with business that is lost when customers see a long line ahead of them. The probability that a given customer balks might be a linear function of the state n, but it is more likely that the function is concave up; that is, the probability a customer leaves increases faster as the system size increases. Put another way, perhaps the indirect cost when there are four customers in the system is more than double the indirect cost when there are only two. In such cases, the indirect cost is a nonlinear function of n. The modeling issue in this form of indirect cost is to choose a function $g(n)$ to represent the cost per unit time for system state n. There are some standard properties we expect such a function to have. There should be no indirect cost when the system is empty, so $g(0) = 0$. The cost cannot go down as the system size increases, so $g'(n) \geq 0$. Most likely the cost increases at a growing rate, so $g''(n) \geq 0$.

For an optimization model, what we need is the expected value of the cost, which is a weighted average of the values $g(0)$, $g(1)$, and so on, using the probability of each state as the weight. Thus,

$$E(\text{IC}) = E(g(n)) = \sum_{n=0}^{\infty} g(n)P_n. \tag{6.9}$$

The sum in formula (6.9) can sometimes be computed analytically. This would always be true for an M/M/s/K system because the probabilities can always be computed analytically (Section 3) and the sum is finite. For an M/M/s queue, only a few cost functions permit an analytical solution formula. One example is the cost function

$$g(n) = C_n \begin{cases} 0, & n \leq s \\ n - s, & n > s \end{cases}, \tag{6.10}$$

which simply says that only customers in the queue count toward the cost. In this case, we have a similar calculation to formula (6.5), with the result

$$E(\text{IC}) = C_n L_q. \tag{6.11}$$

Another example is the cost function

$$g(n) = C_n n^2 \tag{6.12}$$

with an M/M/1 or M/M/2 system. Both of these queue systems have a simple relationship among the probabilities:

$$P_n = \rho^{n-1} P_1.^3 \tag{6.13}$$

The expected value of $g(n)$ can then be written as

$$E(g(n)) = C_n(P_1 + 4P_2 + 9P_3 + \cdots) = C_n P_1(1 + 4\rho + 9\rho^2 + 16\rho^3 + \cdots).$$

We can complete the calculation using the sum formula (see Appendix C)

$$1 + 4\rho + 9\rho^2 + 16\rho^3 + \cdots = \frac{1+\rho}{(1-\rho)^3} \tag{6.14}$$

[3] See formulas (4.4) and (4.8).

to get

$$E(C_n n^2) = C_n \frac{1+\rho}{(1-\rho)^3} P_1. \tag{6.15}$$

For M/M/s queues with $s > 2$, the probabilities can be computed analytically, but then the sum in formula (6.9) must be computed numerically.

6.3 Indirect cost function $h(t)$

One of my colleagues and I recently needed to choose a restaurant for a group dinner. I suggested a restaurant that has been frequently used by the math department for such events. My colleague said that he had decided to boycott that restaurant because in his last group dinner there it took a whole hour for the group to get their food. Most likely different customers have different amounts of tolerance for slow service, and the net effect of this is that restaurants pay an indirect cost for slow service through lost future business. In cases like this, the indirect cost depends on the probability distribution of sojourn times (total time in the system) or waiting times in the queue rather than the probability distribution of system sizes. This makes the mathematics rather more complicated. We consider sojourn times here and waiting times in the next section.

Formula (6.9) calculates expected values as weighted averages of discrete quantities. But sojourn times are continuously distributed, so the expected value formula requires an integral rather than an infinite sum. In principle, think of the range of sojourn times to be divided into infinitely many infinitesimal times. The probability for the time t is $f_s(t)\,dt$, where f_s is the probability density function for the distribution of sojourn times. Each possible value of t corresponds to the cost $h(t)$, and these costs are weighted by the probabilities. Adding these up as a definite integral rather than a discrete infinite sum yields

$$E(\mathrm{IC}) = \lambda E(h(t)) = \lambda \int_0^\infty h(t) f_s(t)\,dt. \tag{6.16}$$

The extra factor of λ is needed because $E(h(t))$ is the expected waiting cost per customer while $E(\mathrm{IC})$ is the expected waiting cost per unit time.

In practice, there are several different ways to evaluate the expected waiting cost from formula (6.16):

1. If h is linear, then formula (6.16) becomes

$$E(\mathrm{IC}) = \lambda \int_0^\infty C_s t f_s(t)\,dt = C_s \lambda S, \tag{6.17}$$

   where we have used the fact that $\int_0^\infty t f_s(t)\,dt$ is the weighted average of the sojourn times, which is S. (We previously did this by a simpler method as formula (6.7). If possible, one should always check more general results by confirming that the give the right answer for known special cases.)

2. If the queue system is M/M/s with infinite queue size and calling population, then we can use the formulas

$$f_s(t) = \mu(1-\gamma)e^{-\mu(1-\gamma)t}, \qquad s = 1, \tag{6.18}$$

$$f_s(t) = \mu e^{-\mu t}\left[1 + \frac{\gamma^s P_0}{s!(1-\rho)}\,\frac{1-(s-\gamma)e^{-\mu(s-1-\gamma)t}}{s-1-\gamma}\right], \qquad s>1, \quad \gamma \neq s-1, \tag{6.19}$$

and

$$f_s(t) = \mu e^{-\mu t}\left[1 + \frac{\gamma^s P_0}{s!(1-\rho)}\,(\mu t - 1)\right], \qquad s>1, \quad \gamma = s-1, \tag{6.20}$$

which we will not derive here. Given these analytical forms for the probability density function, we might be able to calculate the integral by hand and can use numerical integration otherwise.

3. If the waiting times are given by a simulation, then we have a finite list of waiting times instead of a theoretical probability distribution. In this case, the method used for $g(n)$ works:

$$E(\text{IC}) = \frac{\lambda}{J}\sum_{j=1}^{J} h(t_j), \tag{6.21}$$

where J is the number of customers in the database and t_j is the waiting time for customer j. Note that the sum is divided by J to get an average waiting cost per customer and then multiplied by λ to convert the result into an expected waiting cost per unit time.

6.4 Indirect cost function $h(t_q)$

In some instances, it makes more sense to associate the indirect cost with the waiting time in the queue rather than the total time in the system. An example would be a hospital emergency room, in which time spent waiting to be seen is far more important than time spent in treatment. The mathematics here is similar to that for $h(t)$, except that the formulas for the probability distribution of waiting time in a queue are simpler. The general result is

$$E(\text{IC}) = \lambda E(h(t_q)) = \lambda \int_0^\infty h(t_q) f_w(t_q)\, dt_q, \tag{6.22}$$

where f_w is the probability density function for waiting times. Similar to formula (6.17), the linear case ends up as

$$E(\text{IC}) = \lambda \int_0^\infty C_w t_q f_w(t)\, dt_q = C_w \lambda W. \tag{6.23}$$

In place of the formulas (6.18)–(6.20) for the probability density function, we have just one formula,

$$f_w(t_q) = P_q s\mu(1-\rho)e^{-s\mu(1-\rho)t_q}, \tag{6.24}$$

where P_q is the overall probability that a new arrival has to enter the queue rather than being served immediately; that is,

$$P_q = 1 - \sum_{n=0}^{s-1} P_n. \tag{6.25}$$

Example

Consider a system with one server and a service cost that depends on the speed of service. We assume that the cost is 2 per unit time at standard service rate $\mu = 1$ and increases linearly up to a maximum service rate of $\mu = 2$ with cost 6 per unit time. The indirect cost for each customer is 1 unit for each unit of time spent in the system. The mean arrival rate is 0.8. We assume that the service times are exponentially distributed.

The description of the direct cost matches formula (6.2) with $s = 1$. With costs of 2 for $\mu = 1$ and 6 for $\mu = 2$, we have a slope of 4 cost units per speed unit. Thus, the total direct cost is

$$\text{DC} = 2 + 4(\mu - 1) = 4\mu - 2.$$

The indirect cost is a linear function of the time in system, which matches formula (6.7). $C_t = 1$ and $\lambda = 0.8$ were given. For this M/M/1 system, we have the formula

$$L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda} = \frac{0.8}{\mu - 0.8};$$

thus, the expected indirect cost is

$$E(\text{IC}) = (1)(0.8)\frac{0.8}{\mu - 0.8} = \frac{0.64}{\mu - 0.8}.$$

Combining these gives the expected total cost

$$E(\text{TC}) = 4\mu - 2 + \frac{0.64}{\mu - 0.8}, \qquad 1 \le \mu \le 2. \tag{6.26}$$

This result appears in Figure 6.1.

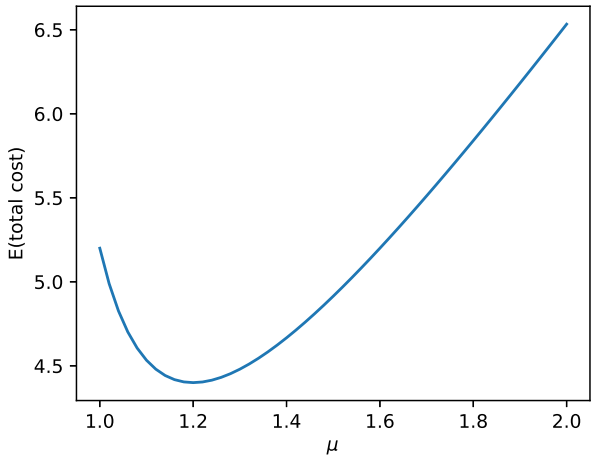

Figure 6.1: Expected total cost for the variable service speed example from formula (6.25).

7 Queue System Optimization (see Hillier and Lieberman 26.2, 4)

Learning Objectives

1. Understand the difference between design parameters and fixed parameters.
2. Develop some facility in constructing optimization models for queue systems.
3. Understand how to obtain an objective function from an expected total cost by removing parameters.
4. Be able to solve some optimization problems with a discrete design parameter by plotting curves that mark indifference between alternatives.
5. Be able to use calculus to solve optimization problems with a continuous design parameter and a simple formula for the objective function.

Design issues arise because of trade-offs. An alternative that is better than the others in every way is clearly the best choice, with no need for mathematical decision making, but that seldom occurs in practice. In the case of queue systems, it is reasonable to expect that better performance costs more. There is usually a principle of diminishing returns: each increment of additional spending produces progressively less improvement in indirect cost. Thus it is common that small improvements are worth the cost while large improvements are not.

In any design problem, there are certain features of the setting that can be chosen, perhaps with constraints on their possible values. There are generally also some features that are inherent and cannot be altered. The critical modeling tasks in a design problem are to distinguish these elements and to identify a mathematical quantity that can be used to compare alternatives. Specifically, the modeler must

1. Identify qualitative features and parameters in the real world setting that can be chosen and determine whether or not there is a limited range of possible values;
2. Identify features of the real world setting that are unalterable, keeping in mind that unalterable features may still have a range of values that need to be considered;
3. Identify a quantity whose value is to be maximized or minimized.

Once these modeling tasks are completed, we are left with a mathematical optimization problem. In this phase we need to

4. Use calculation or simulation to determine the value of the objective function for any given set of parameter values;
5. Use some mathematical, graphical, or numerical method to identify the design parameter values that yield the best value of the objective function.

A common issue in modeling is the trade-off between accuracy and tractability. Sometimes we have a more accurate model that is hard to analyze and a simplified model that is easy to analyze. Which of these we choose depends on how certain we are of the scenario facts. In the case of queueing theory, the mathematical results are expected values of probabilistic quantities rather than actual values of deterministic quantities. The probabilistic nature of queueing theory also affects the reliability of the values we choose for the fixed parameters. For example, given that simulations need to run for a duration on the order of 100,000 service completion times in order to converge to the expected value, we cannot realistically expect to measure the mean service completion time with a high degree of reliability. These considerations make the connection between the input and output quantities less certain than would be the case for a deterministic scenario, suggesting that tractability might sometimes be more important relative to accuracy in queue systems than other optimization settings. Accordingly, we will want to try to make our models reasonably accurate, but we'll accept a modest amount of error in exchange for analysis that we can reasonably do without simulations.

7.1 Modeling Issues

In queueing theory, it is natural to use money as a currency in which to measure value, so optimization problems are usually about minimizing cost or maximizing profit. We developed several different models in Section 6, each consisting of a formula for direct costs, which are associated with the cost of operation, or for indirect costs, which are associated with the inefficiency of the system. In many cases, the increase in operational cost is proportional to the level of service, but the corresponding increase in system performance usually shows diminishing returns. Thus, the optimal level of service is commonly some intermediate level rather than a level of extremely high or extremely low service.

Some of the elements of a queue system model are usually outside the control of the designer. These include the types of probability distribution for arrival and service. Others may be within control in some circumstances. A system with a size limit could be relocated to provide room for a longer queue, or a telephone system with a limited number of lines could have more made available. The number of servers is generally flexible. The mean service rate could perhaps be increased with better training or equipment. When the system is a repair service for a population of machines, it might be possible to decrease the arrival rate by performing regular maintenance or updating the machines. In a system where balking affects performance, it might be possible to decrease balking by making the wait in the queue less undesirable, as when a restaurant offers a waiting area with a bar or an airline offers a special airport facility for its regular customers. Clearly, there is a large variety of optimization models that can result from queue system scenarios. Since we cannot do a systematic survey, we will focus on two examples, one with a discrete decision variable and one with a continuous decision variable.

Example 1 – Choosing the number of servers In the simplest case, the only design choice is the number of servers. Other features, such as the distributions and mean rates for arrival and service, are unalterable. Consider the case where the indirect cost is a linear function of the system size. We can write the expected total cost as

$$E(\mathrm{TC}) = C_s s + C_n L(s, \gamma), \tag{7.1}$$

where the cost per time for a server C_s, the cost per time for each customer in the system C_n, and the arrival to service ratio γ are fixed parameters. Note that these parameters are "fixed" in the sense that the designer cannot change them, not in the sense that they have one specific value. This objective function represents a class of problems rather than a single problem with a specific set of parameters. We might therefore want to know the solution for many sets of parameter values or we might want to know how the solution changes as a parameter value changes. For this reason, it is worth the effort needed to look over a problem to see if all the parameters are really necessary. In this case, we could define a cost ratio parameter by $C = C_s/C_n$ and then rewrite the expected total cost as

$$E(\text{TC}) = C_n(Cs + L(s, \gamma)). \tag{7.2}$$

In this formulation, the parameter C_n is a multiplicative factor. We need it to calculate the expected total cost, but it plays no role in determining the optimal server number . Thus, we can replace the expected total cost with the simpler objective function

$$Z(s; \gamma, C) = Cs + L(s, \gamma). \tag{7.3}$$

In this notation, any quantities in front of the semicolon are true independent variables, while quantities after the semicolon are fixed parameters whose effects we might want to study. This notation serves to define a class of problems. In this case, we can think of the optimal strategy as a function $s(\gamma, C)$ that tells us how many servers to choose for a given set of parameter values. We'll return to this idea later when we analyze the model.

Note that each value of s has its own formula for $L(s, \gamma)$. We'll discuss methods for analyzing this optimization problem below.

This particular example uses the simplest indirect cost model. Other scenarios might call for different choices.

Example 2 – Choosing the service rate In some cases, it may be possible to improve the service rate through better maintenance of machines used for service, better server training, or some other modification. We briefly considered an example in Section 6 where there was only one server with a range of possible μ values. Generalizing that example a little, suppose the parameter μ is confined to the range $\mu_0 \le \mu \le \mu_1$, with the cost a linear function that runs from $C_s = C_0$ for μ_0 to $C_s = C_1$ for μ_1. The rate of change of C_s is

$$C_s' = \frac{C_1 - C_0}{\mu_1 - \mu_0}. \tag{7.4}$$

We can then use the point-slope form for a straight line to write the direct cost as

$$\text{DC} = C_0 + C_s'(\mu - \mu_0) = [C_0 - C_s'\mu_0] + C_s'\mu \equiv A + C_s'\mu. \tag{7.5}$$

Taking the simplest form for the indirect cost, as in Example 1, we have

$$E(\text{IC}) = C_n L(\lambda, \mu) = \frac{C_n\rho}{1-\rho} = \frac{C_n\lambda}{\mu - \lambda}, \tag{7.6}$$

where we have separated out the factors λ and μ since one of these is fixed and the other is variable, and we have substituted in the formula for L when $s = 1$. We therefore have expected total cost of

$$E(\text{TC}) = A + C_s'\mu + \frac{C_n\lambda}{\mu - \lambda}. \tag{7.7}$$

In addition to the design parameter μ, this formula has four fixed parameters, but only two of them are necessary. We can define

$$C' = \frac{C'_s}{C_n} \tag{7.8}$$

so as to obtain the formula

$$E(\text{TC}) = A + C_n \left[C'\mu + \frac{\lambda}{\mu - \lambda} \right]. \tag{7.9}$$

Only the portion of this formula inside the brackets depends on μ , so this is all we need for the objective function. We therefore have

$$z(\mu; \lambda, C') = C'\mu + \frac{\lambda}{\mu - \lambda}, \qquad \mu_0 \leq \mu \leq \mu_1. \tag{7.10}$$

7.2 Analysis Issues

Textbook problems generally deal with specific examples, where all of the fixed parameter values are given. These are relatively straightforward to solve, but the answers are correct only for that specific set of parameters. Real world problems are different. In most practical operations research cases, the problems are posed by managers and solved by employees or consultants. The fixed parameters may not be fully known or may be reassessed periodically. If management changes its estimate of the parameter values or wants to apply the solution to a different instance with different values, our limited solution ceases to be useful and we have to rework the entire problem.

Modeling requires a different mindset than what is required for textbook problems. Instead of *solving problems*, we want to *analyze models*. In general, this means obtaining the most thorough solution that we can find for the problem class or developing a tool, such as a computer program, that can be used to easily solve any new instance of the problem class.

In the remainder of this section, we illustrate the analysis of models using the two examples presented earlier. In the first example, the problem is to find the value of a discrete design parameter that minimizes a function that includes two fixed parameters. Our analysis will reduce the problem to a matter of using the fixed parameters to plot a point on a graph. In the second example, the problem is to find the value of a continuous design parameter, again to minimize a function that includes two fixed parameters. Here we will obtain a solution formula that gives the optimal design parameter value as a function of the fixed parameter values, reducing the whole class of problems to a single calculation.

Example 1 – Choosing the number of servers For a complete analysis of problem (7.3), we need to produce a result that identifies the correct number of servers for all possible values of γ and C. To see how this can be done, let's do a thought experiment. Suppose we solve the problem with one set of parameters and find that 1 server is better than 2. We plot that point in the γC plane with a red dot. We keep solving the problem with different sets of γ and C, plotting red dots if 1 server is better and blue dots if 2 servers are better. Near the γ axis, C is small, meaning that servers are really cheap; hence, 2 servers will be better than 1. Similarly, when C is large enough 1 server will be better than 2. For each value of γ there will be a "purple" point that marks the boundary between red and blue. These purple points

combine to make a purple curve. If we can plot this purple curve, we'll know that 1 server is optimal for points above it and 2 servers are better than 1 for points below it. We won't need to keep solving the problem with different sets of parameters.

The purple curve in our thought experiment marks the points for which the objective function values are the same for both cases; that is, the values of γ and C satisfy the equation

$$2C + L(2, \gamma) = C + L(1, \gamma). \tag{7.11}$$

For any particular γ, there will be one value of C that satisfies this equation because higher values of C always give greater preference to fewer servers. We can calculate this value of C as a function of γ, by solving equation (7.11) for C, with result

$$C = C_{12}(\gamma) \equiv L(1, \gamma) - L(2, \gamma). \tag{7.12}$$

It is a simple matter to plot this curve using a computer by choosing a set of γ values and calculating the corresponding C value.

With the same reasoning, we can identify the point of indifference between 2 servers and 3, which is

$$C_{23}(\gamma) = L(2, \gamma) - L(3, \gamma). \tag{7.13}$$

Two servers are optimal if $C_{23}(\gamma) < C < C_{12}(\gamma)$ because the first inequality marks 2 servers as better than 3 while the second marks 2 servers as better than 1. In a similar fashion, we can compute curves that mark points of indifference between any adjacent server numbers. By plotting these curves on common axes, we obtain a graphical solution for the model (see Figure 7.1). Given any particular set of parameter values, we obtain the optimal number of servers by computing γ and C, plotting the point on the graph, and seeing which of the regions the point lies in .

Example 2 – Choosing the service rate Example 2 differs from Example 1 in two important ways: (1) the design parameter is continuous rather than discrete, which makes calculus a possibility, and (2) the function to be optimized has just one formula used for all cases .

Before solving an optimization problem with parameters, it is helpful to look at some illustrations of what the function can look like. Figure 6.1 shows the expected total cost for one particular set of parameter values: $\lambda = 0.8$ and $C' = 4$. In this case, the minimum cost occurs at a point between $\mu_0 = 1$ and $\mu_1 = 2$. This point can be found by setting the derivative equal to 0. In other cases, however, the graph might be monotone increasing, as would be the case with the same λ and C' but having $\mu_0 = 1.3$. In this case, the optimal μ is at the left end point rather than the point where the derivative is 0.

We start by finding the value $\mu = M$ where the derivative of the objective function is 0. From formula (7.11), we have

$$\frac{dz}{d\mu} = C' - \frac{\lambda}{(\mu - \lambda)^2}.$$

Setting this equal to 0 at $\mu = M$ eventually yields the equation

$$(M - \lambda)^2 = \frac{\lambda}{C'},$$

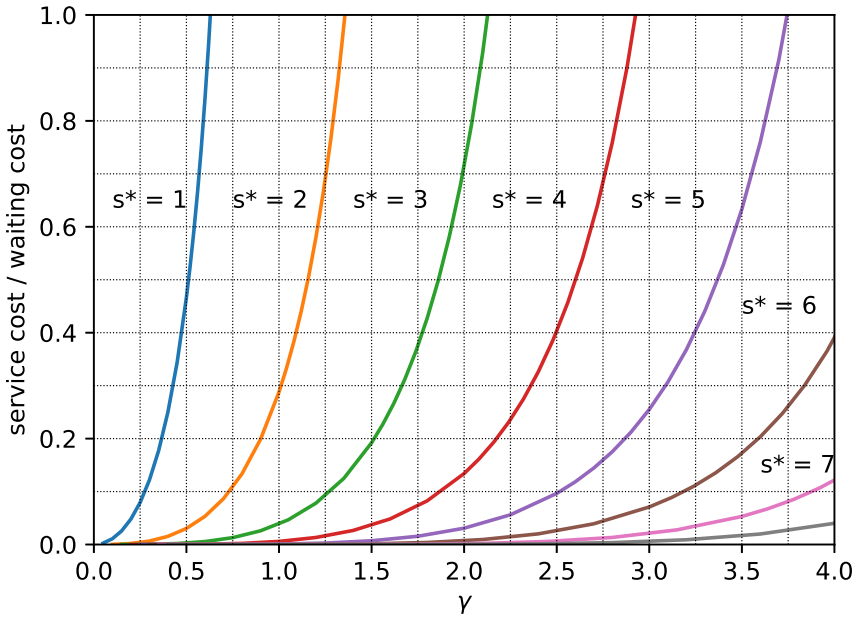

Figure 7.1: Cost-Stress tradeoffs for M/M/s systems with $g(n) = C_n n$.

from which we obtain the result

$$M = \lambda + \sqrt{\frac{\lambda}{C'}}. \tag{7.14}$$

If we consider the effect of different endpoints on the solution for the graph of Figure 6.1, we can see that the optimal μ is given in all cases as

$$\mu^* = \begin{cases} \mu_0, & M < \mu_0 \\ M, & \mu_0 \leq M \leq \mu_1 \\ \mu_1, & M > \mu_1 \end{cases}. \tag{7.15}$$

APPENDICES

A Detailed Analysis of the M/E$_2$/1 System

This section presents the details in the derivation of the computational scheme (5.13)–(5.20) from the balance equations (5.6)–(5.12).

We start by rearranging the balance laws to express the unknown in each successive equation in terms of the previously known values, taking P_0 to be known:

$$\begin{aligned}
P_{12} &= \delta P_0 \\
P_{11} &= (1+\delta)P_{12} \\
P_{22} &= (1+\delta)P_{11} - \delta P_0 \\
P_{21} &= (1+\delta)P_{22} - \delta P_{12} \\
P_{32} &= (1+\delta)P_{21} - \delta P_{11} \\
P_{31} &= (1+\delta)P_{32} - \delta P_{22} \\
P_{42} &= (1+\delta)P_{31} - \delta P_{21}
\end{aligned}$$

and so on, where for convenience we have defined

$$\delta = \frac{\lambda}{2\mu}. \tag{A.1}$$

Next, we combine equations to obtain formulas for P_{n2} that do not include any P_{n1}. For example,

$$P_{42} = (1+\delta)[(1+\delta)P_{32} - \delta P_{22}] - \delta[(1+\delta)P_{22} - \delta P_{12}].$$

This procedure yields the results

$$P_{12} = \delta P_0, \tag{A.2}$$

$$P_{22} = (1+\delta)^2 P_{12} - \delta P_0, \tag{A.3}$$

$$P_{32} = (1+\delta)^2 P_{22} - 2\delta(1+\delta)P_{12}, \tag{A.4}$$

and

$$P_{n2} = (1+\delta)^2 P_{(n-1)2} - 2\delta(1+\delta)P_{(n-2)2} + \delta^2 P_{(n-3)2}, \qquad n > 3. \tag{A.5}$$

The overall probabilities P_n are just the sums of P_{n1} and P_{n2}, conveniently written as

$$P_1 = (2+\delta)P_{12}, \tag{A.6}$$

and

$$P_n = (2+\delta)P_{n2} - \delta P_{(n-1)2}, \qquad n > 1. \tag{A.7}$$

The formulas (A.2)–(A.7) express each of the other probabilities in terms of the unknown probability P_0. If we define $a_n = P_{n2}/P_0$ and $b_n = P_n/P_0$, we have

$$a_1 = \delta, \tag{A.8}$$

$$a_2 = (1+\delta)^2 a_1 - \delta, \tag{A.9}$$

$$a_3 = (1+\delta)^2 a_2 - 2\delta(1+\delta)a_1, \tag{A.10}$$

$$a_n = (1+\delta)^2 a_{n-1} - 2\delta(1+\delta)a_{n-2} + \delta^2 a_{n-3}, \qquad n > 3, \tag{A.11}$$

and

$$b_1 = \delta(2+\delta), \tag{A.12}$$

$$b_n = (2+\delta)a_n - \delta a_{n-1}, \qquad n > 1. \tag{A.13}$$

The sum of probabilities P_n must equal 1, so (dividing by P_0),

$$\frac{1}{P_0} = 1 + b_1 + b_2 + \cdots .$$

P_0 is then given by the formula

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} b_n}, \tag{A.14}$$

after which all of the other probabilities follow from $b_n = P_n/P_0$.

B The M/E$_2$/2 Queue System

Formulas for queue systems with an Erlang service distribution become more complicated rapidly as the number of servers increases. Here we work through the development of formulas for the M/E$_2$/2 queue system. The key difficulty here is describing the different possible states associated with a given system size. When $n = 1$, the second server is idle, so the only states are $p = 1$ and $p = 2$, where p is the phase of the task being done by the working server. But when both servers are busy, we need to keep track of both of their service phases. There are three possibilities: both servers can be in phase 1, they could be split between the two phases, or both could be in phase 2. For $n > 1$, we can distinguish these three cases by defining p to be the sum of the phases of the two servers: $p = 2$ when both servers are in phase 1, $p = 3$ when they are split between phases, and $p = 4$ when both are in phase 2. The various states can be presented in a schematic diagram that is similar to the diagram for M/E$_2$/1, but with another row (see Figure B.1).

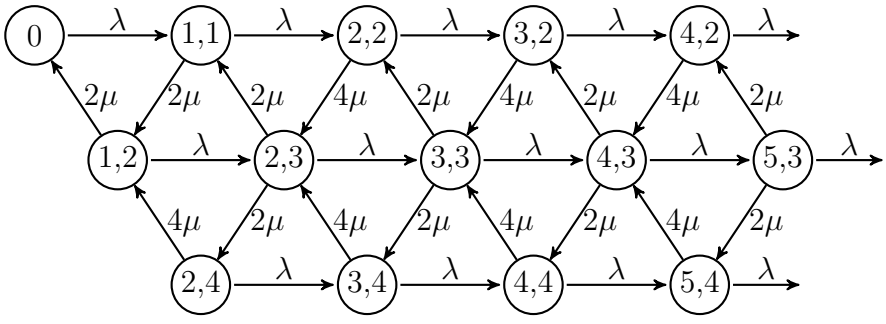

Figure B.1: A schematic diagram of the M/E$_2$/2 queue system

As in M/E$_2$/1, arrivals occur at rate λ and increase the size n. If the new size is 1 or 2, then service starts immediately in phase 1, so p also increases. There are a number of different cases for service completions. From (1,1), completions occur at rate 2μ and move the system

to state (1,2). From (1,2) the rate is the same and the resulting state is 0. These are the same as in $M/E_2/1$. From (2,2), (3,2), and so on, service completions move a customer from phase 1 to phase 2 without changing the size of the system, so (2,2) to (2,3), (3,2) to (3,3), and so on. The rates are 4μ rather than 2μ because each of the two servers has rate 2μ. Along the bottom row, service completions are always phase 2, so the size of the system decreases by 1. The phase decreases by 2 from (2,4) because the second server becomes idle, but decreases by 1 from (3,4) and upwards because the second server changes from phase 2 on the previous customer to phase 1 for the customer moving in from the queue. As with the top row, these transitions have rate 4μ because there are two different servers whose rates are combined. The middle row (not counting state (1,2)) is a bit more complicated because the two servers are in different phases. If the completion is for the server in phase 1, then that server moves on to phase 2 and the system has the same n while p increases by 1. If the completion is for the server in phase 2, then n decreases by 1 while p decreases by 2 if the transitioning server becomes idle and by 1 if a new customers comes in from the queue. All of the transitions from the middle row are at rate 2μ because there is only one server corresponding to each transition.

The procedure for obtaining equations to represent the $M/E_2/2$ system is similar to that for $M/E_2/1$, except that there is a lot more algebra. We start by defining

$$a_{np} = \frac{P_{np}}{P_0}, \quad b_n = \frac{P_n}{P_0}, \qquad n > 0. \tag{B.1}$$

Once we have formulas for a_{np}, we can complete the specifications with

$$b_1 = a_{11} + a_{12}, \qquad b_n = a_{n1} + a_{n2} + a_{n3}, \quad n > 1 \tag{B.2}$$

and

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} b_n}, \tag{B.3}$$

The balance equation for state 0 easily yields

$$a_{12} = \rho = \frac{\lambda}{2\mu}. \tag{B.4}$$

After that it gets difficult.

Working from left to right, we have the balance equation for state (1,2), which yields

$$a_{11} + 2a_{24} = (1 + \rho)a_{12} = \rho(1 + \rho). \tag{B.5}$$

This equation has two unknowns, so we need to combine it with the balance equations for the states in the next column: (1,1) and (2,4). These are

$$(1 + \rho)a_{11} = a_{23} + \rho, \qquad (2 + \rho)a_{24} = a_{23}; \tag{B.6}$$

subtracting the second from the first yields

$$(1 + \rho)a_{11} - (2 + \rho)a_{24} = \rho. \tag{B.7}$$

Equations B.5 and B.7 are a pair of equations for the unknowns a_{11} and a_{24}. Some careful algebra yields the results

$$a_{11} = \frac{4\rho + 3\rho^2 + \rho^3}{4 + 3\rho}, \qquad a_{24} = \frac{2\rho^2 + \rho^3}{4 + 3\rho}, \tag{B.8}$$

and then

$$a_{23} = (2+\rho)a_{24}. \tag{B.9}$$

The procedure for obtaining a_{22}, a_{34}, and a_{33} is similar, albeit messier. We need the balance equations for states (2,3), (2,2), and (3,4):

$$2a_{22} + 2a_{34} = (2+\rho)a_{23} - \rho a_{12}, \tag{B.10}$$

$$(2+\rho)a_{22} - a_{33} = \rho a_{11}, \tag{B.11}$$

$$(2+\rho)a_{34} - a_{33} = \rho a_{24}. \tag{B.12}$$

Subtracting equation (B.12) from equation (B.11) eliminates a_{33}:

$$(2+\rho)a_{22} - (2+\rho)a_{34} = \rho(a_{11} - a_{24}). \tag{B.13}$$

From here, we can rewrite equations (B.10) and (B.13) as

$$\frac{1}{2}a_{22} + \frac{1}{2}a_{34} = \frac{2+\rho}{4}a_{23} - \frac{\rho}{4}a_{12},$$
$$\frac{1}{2}a_{22} - \frac{1}{2}a_{34} = \frac{\rho}{2(2+\rho)}(a_{11} - a_{24}).$$

Finally, we can add and subtract these two equations to get a_{22} and a_{34} respectively:

$$a_{22} = \frac{2+\rho}{4}a_{23} - \frac{\rho}{4}a_{12} + \frac{\rho}{2(2+\rho)}(a_{11} - a_{24}), \tag{B.14}$$

$$a_{34} = \frac{2+\rho}{4}a_{23} - \frac{\rho}{4}a_{12} - \frac{\rho}{2(2+\rho)}(a_{11} - a_{24}), \tag{B.15}$$

and finish the set with

$$a_{33} = (2+\rho)a_{34} - \rho a_{24}. \tag{B.16}$$

The remaining groups of three values are similar, with all of the terms in the formulas replaced with the appropriate analogs (noting that states (1,1) and (1,2) are replaced by (n-1,2) and (n-1,3) in these formulas):

$$a_{n2} = \frac{2+\rho}{4}a_{n3} - \frac{\rho}{4}a_{n-1,3} + \frac{\rho}{2(2+\rho)}(a_{n-1,2} - a_{n4}), \qquad n \geq 2; \tag{B.17}$$

$$a_{n+1,4} = \frac{2+\rho}{4}a_{n3} - \frac{\rho}{4}a_{n-1,3} - \frac{\rho}{2(2+\rho)}(a_{n-1,2} - a_{n4}), \qquad n \geq 2; \tag{B.18}$$

and

$$a_{n+1,3} = (2+\rho)a_{n+1,4} - \rho a_{n4}, \qquad n \geq 2. \tag{B.19}$$

The full set of a values is given in turn by formulas (B.4), (B.8), (B.9), and (B.14) through (B.19), with the last three repeated for the increasing sequence $n = 2, 3, \ldots$ until the values are small enough not to matter.

Figure B.2 shows the steady-state probabilities for the M/$E_2$2 and M/M/2 systems. The differences are similar to those seen with one server in Figure 5.3.

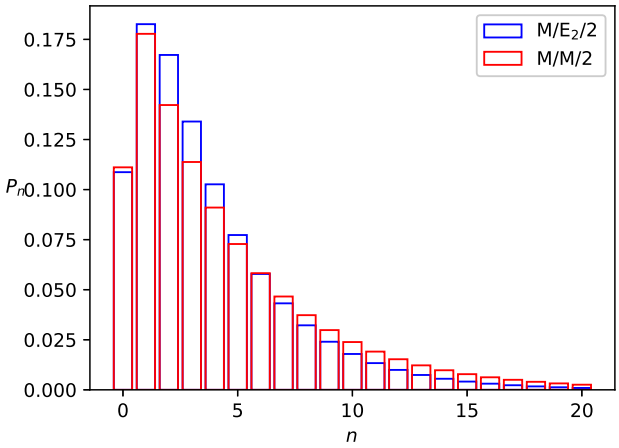

Figure B.2: Steady-state probabilities for the M/E$_2$2 and M/M/2 systems with $\rho = 0.8$.

C The Sum Formula $\sum_{n=1}^{\infty} n^2 x^{n-1}$ (6.13)

Let

$$S = \sum_{n=1}^{\infty} n^2 x^{n-1} = 1 + 4x + 9x^2 + 16x^3 + \cdots .$$

Then

$$xS = x + 4x^2 + 9x^3 + 16x^4 + \cdots .$$

Subtracting these yields

$$(1-x)S = 1 + 3x + 5x^2 + 7x^3 + \cdots ,$$

This can be further partitioned as

$$(1-x)S = (1 + 2x + 3x^2 + 4x^3 + \cdots) + (x + 2x^2 + 3x^3 + \cdots),$$

or

$$(1-x)S = (1+x)(1 + 2x + 3x^2 + 4x^3 + \cdots).$$

The infinite sum in this formula is known:

$$1 + 2x + 3x^2 + 4x^3 + \cdots = \frac{d}{dx}(1 + x + x^2 + x^3 + \cdots) = \frac{d}{dx}\left(\frac{1}{1-x}\right) = \frac{1}{(1-x)^2}.$$

Thus,

$$(1-x)S = \frac{1+x}{(1-x)^2},$$

or

$$S = \frac{1+x}{(1-x)^3}.$$