CHAPTER 1

Concepts, Probabilities, Models, and Costs

Standing in line for some service is a universal human experience. We all have had the experience of choosing one line that appears to be moving the fastest only to observe later that an adjacent line is now moving faster. The reality is that all customers do not have the same service requirements, and even if they do, that service does not always take the same amount of time to complete.

Some of us are prepared to ask for what we need when we reach the server; others are still making up their minds as to exactly what they want. The process of paying for the service also adds to service time variability. How many of us have watched the person being served in front of us take a considerable amount of time collecting belongings and finding money or a credit card to pay the server while another customer is more organized, has exact change, and moves out of the line quickly when the service is completed. Such observations are important because they illustrate that one way to improve customer service is to help customers be better prepared when they reach the server. An illustration of this in practice is the security line process at many airports.

An important consideration for service businesses is being able to estimate how many customers might arrive during a given time period T. It is also important to know the nature of customer arrivals. Are they at regular intervals, random, one at a time, in groups, or in some other way? Given this information and the internal knowledge of how long service usually takes, businesses can use waiting line analysis to determine how many servers are needed to provide an acceptable level of performance at a reasonable cost. Some other performance measures of possible interest are as follows:

- How long is the average line?
- What is the typical customer waiting time before being served?
- What is the probability that the line will exceed the available waiting space?
- What is the probability that a customer will not have to wait in line?
- Are customer arrivals relatively consistent during the hours that the service is available or are they variable at predictable times during the hours of operation?
- If one server is not enough to satisfy demand in a timely manner, how many more servers are needed?
- How much do we have to reduce average service time to avoid adding another server?
- Which is more cost-effective—buying a new automated machine to handle part of the demand or hiring more servers?
- What effect would setting up separate lines for different classes of customers have on overall service performance and operating costs?

When applying queuing theory to factory applications, some waiting line parameters are more easily controlled; others become more complex, particularly when more than one process step (phase) is involved. Production scheduling can reduce variability in the arrival rate. The service rate variability is usually constrained when manufacturing standardized items but can vary much the same as when dealing with customers if the factory is providing repairs or custom items with varying work flows. Moving items between batch and one-at-a-time processes are also analytical challenges. Performance measures of possible interest in such applications include the following:

- What is the average throughput?
- What is the typical processing time?
- What is the line capacity and which step limits it?
- What is the average amount of work in process (WIP)?

- How much storage capacity is needed for the inventory (queue) before each step?
- How are mixed job flows handled?

In some cases, basic queuing equations can provide rough estimations for some of these manufacturing questions. For more accurate estimates, however, simulation methods are required. For those who do not want to develop their own applications, a variety of vendors have produced simulation programs for common situations. However, it is important if you take either path toward using simulation that you are aware of the assumptions and queuing models available. Selecting the wrong model for your application will not provide useful results no matter how sophisticated the simulation package is.

Managerial considerations regarding the previous performance questions and others are discussed in Chapter 6. Many of them have different options depending on the waiting line model(s) used. Chapter 7 discusses some simple simulation applications and how a simple model can be used as a building block for more complex simulations. Appendix D provides information for using Excel for some basic simulations. Appendix B defines the symbols used throughout this monograph. Because three of these symbols are necessary to any discussion of waiting line situations, analysis, models, or concepts, they need to be defined here:

λ: the *average* arrival rate of customers or items seeking service
μ: the *average* service rate
ρ: the ratio λ/μ, often referred to as the utilization factor

The fundamental assumption of waiting line analysis is that the behavior of customer arrivals and service times can be described by appropriate probability distributions given the average interarrival time $1/\lambda$, the average service time $1/\mu$, and some knowledge of the pool of possible customers (the "calling population") to be served.[1] Many of the waiting line performance measure formulas discussed in this chapter are a result of this assumption combined with the insight and the contributions of many talented individuals and organizations.

We will accept most of these formulas without derivation except when a partial or a complete derivation is necessary to gain a better understanding of what a particular formula does or does not address. For those interested in such derivations and some good application examples, the books by Laguna and Marklund[2] on business process modeling and Hillier and Lieberman[3] on operations research are useful references. Another reference is Nelson's book[4] on modeling stochastic processes.

The most commonly used probability distributions are the Poisson distribution for discrete values and the exponential distribution for continuous values with the assumption that the calling population is infinite. Other distributions are used when some control over the arrival rate or the service time is possible and when the population pool or the waiting line capacity is limited.

In selecting distributions that best represent a given queuing situation, it is important to select an appropriate time interval for data collection and the analysis on which the average values for arrival and service rates can be based. The average waiting line performance measures are independent of the time interval chosen, which will be shown later when using the formulas. However, you will lose sight of how much the arrival rate can vary during the day if you choose an interval that is too large.

Poisson Distribution

The Poisson distribution is a discrete distribution because there are no fractional arrivals. It is described by Equation 1.1, where P(n) is the probability that n arrivals will arrive during time interval T given an average arrival rate of λ:

$$P(n) = \frac{(\lambda T)^n e^{-\lambda T}}{n!} \quad \text{for } n = 0, 1, 2, \ldots . \tag{1.1}$$

The Poisson distribution for an average arrival rate of 4 is shown in Figure 1.1. It is important to note that there is a finite probability that no arrivals will occur during the time interval on which the average rate is based. Knowing this value—often designated as P_0 rather than P(0)—is very useful in business decisions, as we will discuss in Chapter 6.

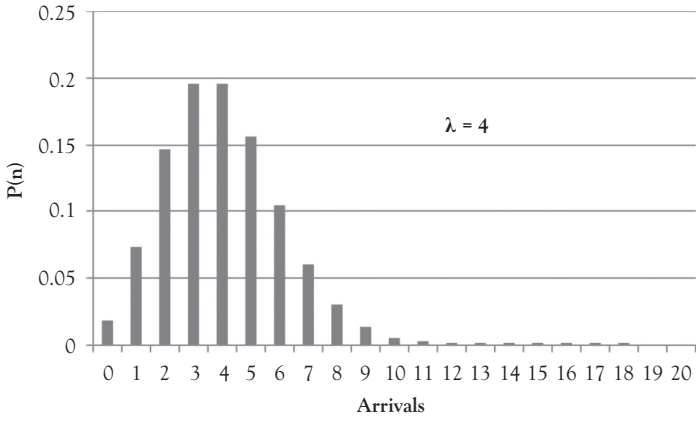

Figure 1.1 Poisson distribution for an average arrival rate of 4

Another useful property is that adding individual Poisson distributions results in another Poisson distribution. Conversely, breaking down a Poisson distribution into two or more separate distributions results in a set of two or more Poisson distributions. This is advantageous when adding together the known arrival rates of individual classes of customers or determining the effect of diverting a part of an existing arrival distribution to a new branch office. We will discuss this in more detail in Chapter 6.

Exponential Distribution

An exponential distribution is a versatile probability distribution that is used to describe both service times and the times between arrivals in waiting line scenarios. It is most often expressed by Equation 1.2, where P(time > t) is the probability that the service time or the interarrival time will be greater than time t, setting $\alpha = \mu$ for service time probabilities and $\alpha = \lambda$ for interarrival time probabilities:

$$P(\text{time} > t) = e^{-\alpha t} \text{ for } t \geq 0. \tag{1.2}$$

The exponential distributions for two average service rates of 3 and 6.5 are shown in Figure 1.2. Your intuition that the probability of a task requiring a given completion time should decrease as the service rate

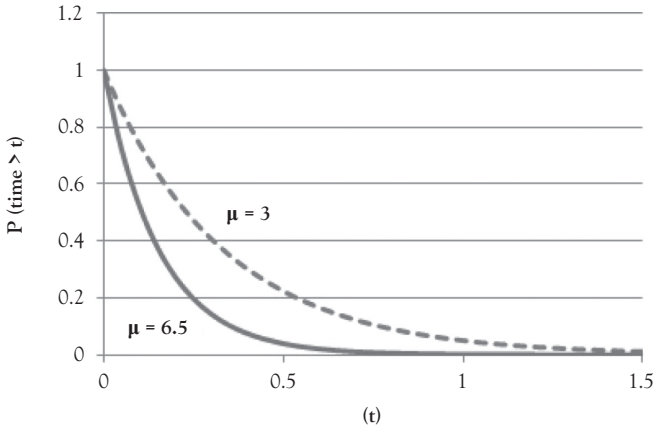

Figure 1.2 Exponential distributions for service rates of 3 and 6.5

increases is now validated. You should also note that a probability of 36.79 percent[5] corresponds to the average service time of 1/μ.

An exponential probability distribution has the property of being "memoryless." That is, its predictions of what happens next are independent of what has happened before. For example, the probability at any given moment that the next customer will arrive in two minutes is the same whether the previous customer arrived a few seconds ago or an hour ago. Such a distribution is also referred to as being "Markovian" and is indicated by the symbol M in the waiting line model notation described later in this chapter. More common examples of this property are the probability of the next coin flip being heads or the next roll of two dice adding up to five. The odds of either occurring are unaffected by any knowledge of the results of previous coin flips or dice throws.

Other Probability Distributions

In many service scenarios, the performance data collected will indicate that either the arrival or the service characteristics are not well represented by Poisson and exponential distributions. For example, consider a coffee shop where there is a mix of customers, some who just want a standard cup of coffee and others who want more customized lattes and mochas. In addition, while the exponential distribution allows the probability of very short service times, in practice, the minimum time required to serve

a customer is greater than those values. If we want to evaluate more than the average performance values, this mix of relatively constant and widely varying service times plus a minimum service time usually does not fit well with using a single exponential distribution.

Also, consider standardized service situations where the service times are more predictable (deterministic). In such cases, normal distributions or even constant values can be used. This situation is discussed in more detail in Chapter 4.

A variant of the Erlang[6] distribution can be used to determine the number of customers turned away by insufficient capacity, and phase-type distributions can be used to characterize waiting lines with more than one phase (step) in sequence. When there is more than one phase in a channel, the mathematics for analytical expressions describing waiting line performance becomes much more challenging. In such cases, we can make some approximations or use computer simulation to provide more useful insight for business applications. One example is a production line with several assembly operations. This situation is discussed in more detail in Chapters 4 and 7.

State Diagrams and Balance Equations

The intent of this monograph is not to make you an expert on deriving waiting line expressions; however, it is useful to spend some time discussing how state diagrams are used to develop some of the simpler formulas. This then provides a better understanding of the relationships between arrival rates, service rates, and the probabilities of different line conditions, such as nobody in line (P_0), two people in line (P_2), and so forth.

Figure 1.3 shows a generic state diagram where each hexagon represents a specific number of customers in a single-channel waiting line. In this case, we will keep it simple by limiting the maximum number of customers in the system to two. This is not a far-fetched simplification because it could represent an independent stockbroker's telephone with a capacity for only one call on hold.

To obtain an intuitive feel for what this diagram represents, we first consider the first two states on the left (state 2 is ignored for now). One can see that the *maximum* flow from the state of no customers in the

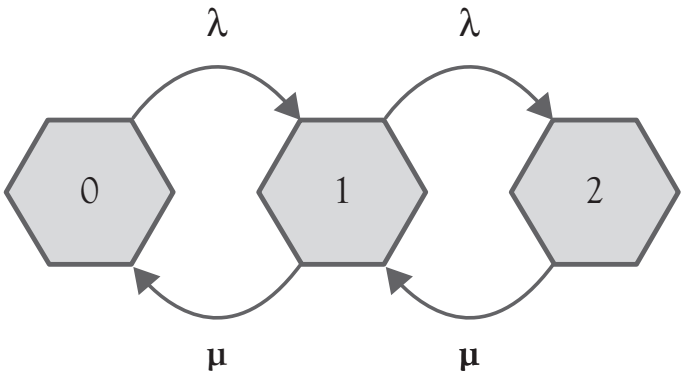

Figure 1.3 State diagram for a single-channel waiting line

system to the state of one customer in the system is the arrival rate λ. The *maximum* flow back to a state of no customers is the service rate μ. If both the arrival and the service rates are constant with no variability, then the proportion of time each state exists is determined by the difference between the arrival and the service rates. This then requires the service rate to be greater than the arrival rate to avoid the need for further states to account for inadequate capacity.

So, what is missing in this state diagram? While it may be obvious to the frequent state diagram user, it often is not obvious to many business students. The key factor here and the magic behind using state diagrams to develop queuing analysis formulas is that each state has a probability of existence that allows the rates into a state to equal the rates from that state. Such balance equations allow one to determine the probability for each state in terms of the given average arrival and service rates. This becomes particularly important when the distributions for the arrival and the service rates are taken into account. While their probabilities of occurring are low, there will be instances when the arrival rate is greater than the service rate. Then an overflow to higher states is necessary for situations when a momentary burst of customers overwhelms the service rate and results in more customers in the system.

For example, the balance equations (inputs equal outputs) for the three states in Figure 1.3 are as follows:

$$\begin{aligned} \text{State 0: } & (P_1 \times \mu) = (P_0 \times \lambda), \\ \text{State 1: } & (P_0 \times \lambda) + (P_2 \times \mu) = (P_1 \times \lambda) + (P_1 \times \mu), \\ \text{State 2: } & (P_1 \times \lambda) = (P_2 \times \mu). \end{aligned} \tag{1.3}$$

Given the above set of balance equations for all the possible states, the individual probabilities for steady-state behavior can be derived. For example, we first use the equation for State 0 to solve for P_0 in terms of P_1. Then, using the requirement that the sum of all probabilities must be equal to 1, we solve for P_2 in terms of P_1 and P_0. That is, $P_2 = 1 - P_1 - P_0 = 1 - P_1 - [(\mu/\lambda) \times P_1]$. Remembering that μ/λ is $1/\rho$ and substituting these results for P_0 and P_2 in the equation for State 1, we can then determine the value for P_1 in terms of ρ. Knowing P_1, we can then obtain the values for P_0 and P_2 using the equations for States 0 and 2. The following results are produced:

$$\begin{aligned} P_0 &= 1/(1 + \rho + \rho^2), \\ P_1 &= \rho/(1 + \rho + \rho^2), \\ P_2 &= \rho^2/(1 + \rho + \rho^2). \end{aligned} \tag{1.4}$$

If our calculations are correct, the sum of P_0, P_1, and P_2 should equal 1, which they do. What is also useful here is having answers requiring only the utilization factor ρ. This may not always be the case, but it simplifies the analysis when it is.

For a larger number of states, the mathematics involved can be quite daunting. However, for a queuing situation, where the number of states is limited by physical constraints such as a finite calling population, limited line length, or the number of phone lines, state diagrams can be quite useful in obtaining useful expressions without having to employ extensive mathematical methods. We will return to this state diagram when we discuss the more complex aspects of single-channel waiting lines in Chapter 4.

Waiting Line Models and Notation

Although there is some commonality across various waiting line models, such as Little's Law which is discussed in Chapter 7, many of the formulas predicting various performance measures are dependent on the type of model used. Most references use Kendall's notation[7] to identify which waiting line situation they are discussing. The original notation reportedly had just three characteristics, A/B/C, indicating, in order, the nature of the arrival distribution, the service distribution, and the number

of channels or servers. This notation has been expanded over time to five or six characteristics, A/B/C/d/e/f, to include values for the limit on line length, the size of the calling population, and the priority rule used to process customers.

You will find in various references that there is no consistent order for indicating the last three characteristics or even for using all three. Hence, it is important to note that here we will use the d/e/f sequence defined in the previous paragraph. The symbols used to designate the various probability distributions are defined in Appendix B. For example, the single-channel, single-phase model discussed in Chapter 2 is given the notation M/M/1/∞/∞/FCFS, where M indicates the choice of Markovian distributions for the arrival and the service rates. In most references, this notation is shortened to M/M/1.

Priority Rules

Unless specified otherwise, a first-come, first served (FCFS) priority rule is assumed for most waiting line situations. This rule can also be expressed as a first-in, first-out (FIFO) priority for single-line situations.[8] Customers consider this to be the fairest approach, particularly when other customers waiting in line are visible.

However, there are situations where an FCFS rule is not the best approach, such as the processing of patients in a hospital emergency room. Obviously, there will be some patients with more urgent needs for care than others, regardless of their place in the arrival sequence. Another example is travelers waiting in line to check in at the airport. When the lines at airline ticket counters become long at peak periods and the waiting time becomes greater than the time left before flight departure for some travelers, some process is required to expedite the ticket and baggage check-in processing for those passengers.

Not so obvious is the desire by many service operations to give some preference to their more important customers. In situations where customers can see other customers waiting, this desire can be satisfied by having dedicated servers for the more important customers. Examples are the frequent flyer lines at the airport, a business-only teller window at the bank, or a window dedicated for package pickup at the post office.

Note that it is a good business practice to encourage the servers for the dedicated lines to take care of customers in other lines when there are no preferential customers waiting to be served.

More solutions for providing preferential treatment to selected customers are available when customers cannot view other customers waiting. One example is a call center for a financial institution. Such solutions are discussed in more detail in Chapter 5.

In a manufacturing line situation, you may want a system for expediting critical orders, often referred to as "hot lots," while also taking care to avoid any given item from being delayed too long because of preemptions by expedited orders. This is especially important if the customer-ordering process frequently accepts too many rush orders. Part of the solution is a carefully considered managerial policy regarding the use of expediting, which is discussed in Chapter 6. In addition, you can take advantage of computer methods for managing the sequence of items being processed, which is described in Chapter 7.

While not often considered to be a typical waiting line situation, the boarding of passengers can be characterized using both single and multiple line configurations with an arrival rate determined by a mix of people arriving at random and determined by priority. What makes the analysis challenging is that the service rate is also a mix of typical servers (agents checking tickets) and the passengers doing some of the work (stowing personal items and handling carry-on baggage). In addition to the obvious priorities for allowing first-class passengers, frequent flyers, and persons needing assistance to board earlier, various priority schemes have been developed to help expedite the boarding of the main cabin.[9] Reducing boarding time helps increase the efficiency of the transportation equipment since time spent on the ground or standing at the station or terminal is time not spent moving passengers. At airports longer boarding times can also cause planes to miss scheduled departure windows and affect the ability of an airport to handle the amount of traffic required. In such cases the increased operational costs can be significant.

Finally, although rarely discussed in most waiting line textbooks, effective methods are needed to deal with rude, unruly, or disruptive customers. Considering how your business addresses such incidents

before they actually occur is especially important when such behavior is visible to other customers.

Cost Curves

The relationship between operation costs and the costs of waiting is illustrated in many references by similar versions of the simplified graph shown in Figure 1.4. As more servers are hired, customers have to wait less, but the costs of doing business increase. In addition to the obvious increase in salary and benefits costs additional equipment and facility increases are likely to be needed, particularly if the improved service performance attracts additional customers. This trade-off between customer service and operating costs is shown in Figure 1.4 to have a minimum value that intuitively would appear to be the desired business solution.

However, this simplified model does not depict actual conditions for many service businesses, particularly since the operating costs rarely increase in a linear fashion. So, you may ask, how does one obtain the actual waiting costs for a particular situation? The costs related to adding more servers or items being out of service while they wait for maintenance are more easily determined than the costs of unhappy customers. Some marketing firms have done surveys about how much a typical customer is willing to wait for service or what length of line will discourage a customer

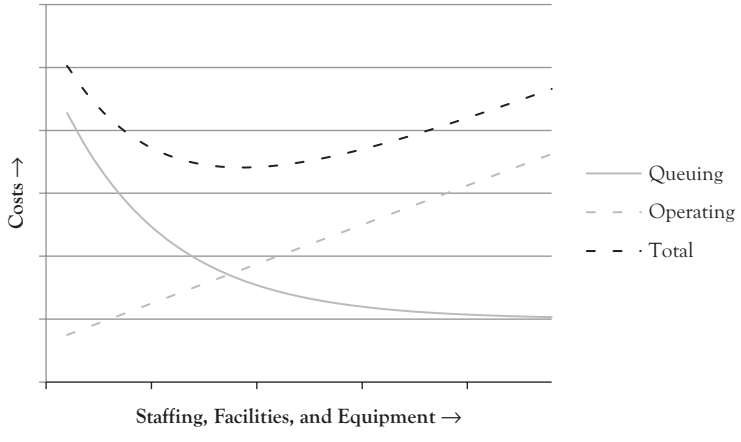

Figure 1.4 Waiting and operating cost relationships

from even entering it. But you need to be wary of this information unless such surveys have been done for your particular type of service in your locality. Even then, you should recognize that customer attitudes will vary from one day to the next, depending on whether or not customers are in a good mood, are in a hurry, or are with friends in line; the weather is sunny or miserable; and so forth. Did the marketing group survey people actually in line under different conditions or did they just interview a group of people about their service preferences?

Is your service one of choice—like buying a cup of coffee—or one of necessity—like obtaining a new driver's license? Do you have several competitors or are you the only choice for the service? In the latter situations, customers may not like the performance provided but must put up with it because there are no other alternatives available. This is often the case with government agencies, where the trade-off is better service or lower taxes.

In most business situations, operating costs are rarely linear, as implied in Figure 1.4. Increases in equipment and the number of servers create jumps in operating costs. Facility additions and other fixed costs required to support additional servers and equipment must be accounted for, and the effect on other behind-the-scenes (back office) support costs should be considered.

Waiting line costs are also not always linear. Waiting a few minutes longer may be inconvenient, but waiting long enough for a meal to get cold or to miss a deadline like a scheduled transportation departure can cause the cost for a customer to increase significantly.

In some situations the waiting line cost is also a significant part of the operating cost for a business and would be more accurately described as a queuing cost to make it clearer that the costs are just those caused by waiting. In the earlier section regarding priority rules we discussed passenger boarding processes and the potential increase in operational costs if they were not efficient. In essence, a transportation business is also a customer waiting in line for a service (boarding) to be completed. The range of cost components is broad and interactive between the passengers and the carrier. Boarding delays add to overall turnover time for equipment and reduce system capacity. Such delays can cause a passenger to miss a connecting flight or a critical business appointment or not be able

to attend an important family event. Complicating this situation is that normal approaches for reducing waiting time such as adding more servers or making existing servers more efficient are not an option. For example, adding another aisle on an aircraft could help speed up boarding, but would reduce passenger capacity. Since increasing server efficiency would involve making the passengers be more efficient in getting their items stowed and seated an airline could decide to prohibit carry-on luggage requiring an overhead bin and have it checked instead. However, this would increase baggage handling costs and could increase the time to load and unload the plane. The increased practice of charging for checked baggage to offset baggage handling costs has increased the use of carry-on luggage by passengers and as a consequence has made the boarding process less efficient and longer. This is not to say that there is no optimal solution for such cost trade-offs, but such a solution requires careful consideration of all system costs and their interactions.

These costs and some suggested methods for managing them are discussed in more detail in Chapter 6.