CHAPTER 2

The Basics—Single-Channel, Single-Phase Model

A single line of people waiting for some service is arguably the most common business model on our planet (see Figure 2.1). The variations of this simple theme are nearly infinite when you consider the forms that the service can take and the nature of the customers desiring that service.

We will use a typical coffee shop where customers from different walks of life enter the shop; wait in line; and, when they reach the service window, order anything from a simple cup of coffee to a more complex mixture of coffee and other ingredients to larger orders of multiple combinations of these. To analyze the basic behavior of this model, we must make some assumptions where the full Kendall notation for the model is M/M/1/∞/∞/FCFS:

- The customer arrival rate is described by a Poisson distribution using an average rate λ, which means that the interarrival times can be characterized by an exponential distribution with an average interarrival time of $1/\lambda$. Interarrival times are independent of the number of customers in the system.
- The variability in service time is defined by an exponential distribution with an average service time of $1/\mu$, where μ represents the average service rate. Service times are independent of the number of customers in the system. While one could argue that the server would be under more pressure to work faster when the customer line is longer, in practice this is not a good assumption for a service business to make.

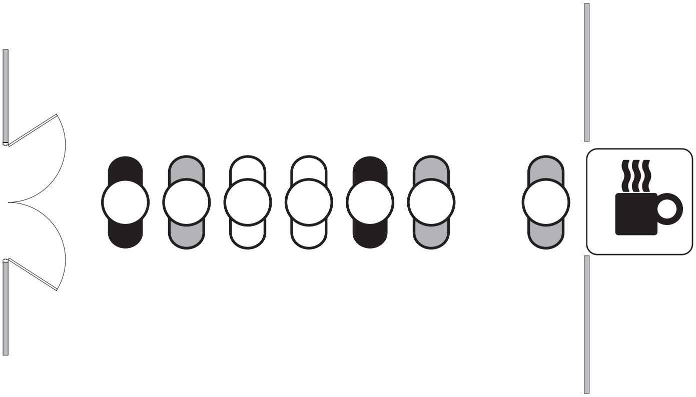

Figure 2.1 M/M/1 configuration

Such actions increase the likelihood of mistakes that require additional time to correct and offset any potential reduction in average service time.

- The number of channels (servers) is one, and the number of phases in the service is one. We assume here that a single server does everything for the customer in the coffee shop: taking the order, preparing the coffee, and collecting the payment. Obviously, in many coffee shops with a higher volume of customers, more than one person likely performs these activities. We will discuss these complexities in Chapter 4.
- The arrival or calling population is infinite in size. This avoids complications introduced by the possibility of having served all available customers. Limited or fixed customer situations are discussed in Chapters 4 and 5.
- The length of the waiting line can be infinite. Although not really possible in real-world situations, this avoids analysis complications introduced by the rare possibility that some customers are blocked from entering the line. We will discuss the effects of limited line lengths in later chapters.
- The priority rule is first come, first served (FCFS).
- Balking or reneging by customers is not considered in the analysis.

- The average arrival rate is less than the average service rate ($\lambda < \mu$). That is, the utilization factor $\rho = \lambda/\mu$ is less than 1.

This last assumption should be intuitively obvious because the average line length will increase significantly when the arrival rate approaches the service rate, as shown in Figure 2.2.

One thing that is often confusing when reviewing the equations presented by different authors for a given waiting line model is that at first glance they do not always appear to be the same. However, with closer inspection, we can see that one version is equivalent to another version because some author(s) substituted ρ for some combinations of λ and μ or have applied Little's Law (defined in a moment). The different versions are provided here for your reference with what is considered to be the most useful version listed first. The various performance measures for the basic M/M/1 model are listed in the same order in which they will be presented for other waiting line models later in the book.

- Utilization factor: $\rho = \lambda/\mu$; for a basic M/M/1 model, ρ must be less than 1.
- Probability of 0 customers in the system (this is also the probability that a customer will experience no waiting for service): $P_0 = 1 - \rho$.

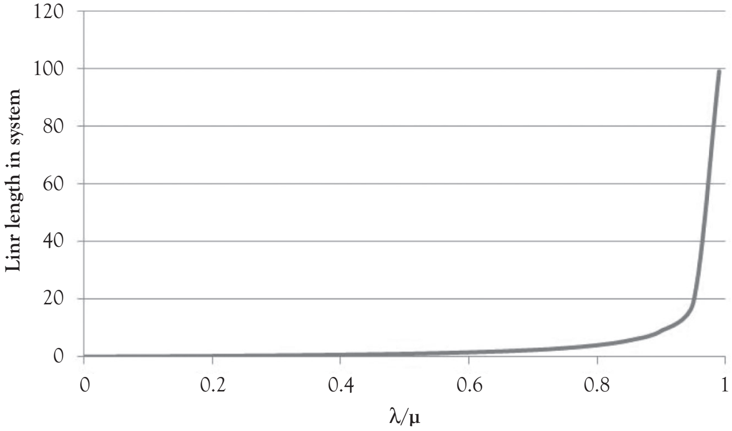

Figure 2.2 Increase in average line length as λ → μ

- Probability of exactly n customers in the system: $P = (1 - \rho)\rho^n = P_0\, \rho^n$.
- Probability that the number of customers in the system is greater than k:
$$P_{n>k} = \rho^{k+1}.$$
- Probability that the server is busy: $P_{n>0} = 1 - P_0 = \rho$.
- Average number of customers in the system:
$$L = \lambda/(\mu - \lambda) = \rho/(1 - \rho) = \lambda W.$$
- Average total time customers spend in the system: $W = 1/(\mu - \lambda) = L/\lambda$.
- Average number of customers waiting in the queue (not yet being served):
$$L_q = \rho L = \rho\lambda/(\mu - \lambda) = \lambda^2/[\mu(\mu - \lambda)] = \rho^2/(1 - \rho) = L - \rho = \lambda W_q.$$
- Average time customers wait in the queue before being served:
$$W_q = \rho W = W - (1/\mu) = \rho/(\mu - \lambda) = \lambda/[\mu(\mu - \lambda)] = L_q/\lambda.$$

The term *system* applies to not only the number of customers waiting in line but also any customer(s) being served. Also useful to remember is that the sum of all possible probabilities for a parameter must equal 1. This allows you to determine a probability that may be difficult to calculate directly. In such a case, you add together all the other possible probabilities excluding the one you want to determine and subtract that sum from 1. An example of this is shown in the list of performance measures above for the probability that a server is busy.

If you study the formulas for line lengths and waiting time carefully, you can see that the ratio of the average number of people in the system to the average waiting time in that system is equal to the average arrival rate when the system is in a steady-state condition. Expressed in the form $L = \lambda W$, this ratio is referred to as Little's Law.[1] This ratio also applies when considering just the number of people waiting in line and the time they spend waiting in that line before being served. As we will see in subsequent chapters, Little's Law applies across a wide range of waiting line models regardless of the probability distributions chosen to represent the arrival and the service rates. Hence, Little's Law is especially useful for manufacturing applications of queuing analysis where such probability distributions are often unknown.

Another significant observation can be drawn by looking at the formulas for L, L_q, and P_0 that use only the utilization factor $\rho = \lambda/\mu$. This means that these values vary only with ρ; that is, as long as the arrival and service rates increase in direct proportion to each other (ρ remains constant), the values for L, L_q, and P_0 will remain unchanged. However, W and W_q will decrease with increasing values of λ and μ. Figure 2.2 shows the value for L versus ρ, and you can see that the line length is essentially flat for $\rho < 0.4$. In Chapter 3, you will see that this characteristic also holds true for multiple-channel lines.

You may observe what appears to be an inconsistency in the expressions for L_q, where it is shown as being equal to both ρL and L – ρ. Is this possible? Most textbooks avoid this question by presenting only one version for L_q; however, both versions were included in my class lectures as a check on whether students were reading the material. To answer the question: If both equations are true, then ρL must equal L – ρ. Moving all the L terms to one side of the equation gives the result that $L(1 - \rho) = \rho$. Dividing both sides by $(1 - \rho)$ gives $L = \rho/(1 - \rho)$, which is the equation for L.

Example 2.1 Coffee Shop with One Server

To gain a more comfortable understanding of how the preceding performance measures can be used, consider a small coffee shop we will call Ken's Caffeine Fix. The shop is located in a downtown area and provides various forms of coffee to nearby office workers who stop in for a cup of their favorite brew at various times of the day. We will make two assumptions: (1) The average arrival rate does not vary during the day, and (2) the owner performs all parts of the service provided: takes the order, prepares the coffee, and collects the payment. The issues created by arrival rates varying during the day and having a helper do parts of the service will be discussed in later chapters. To assign some values to this operation, let us assume that it takes an average of two minutes to serve each customer, and the average number of customers per hour is 24.

This gives us an average arrival rate λ of 24 customers per hour and an average service rate μ of 1 customer per 2 minutes = 30 customers/hour. This illustrates a critical consideration when analyzing waiting line situations for small businesses—the arrival rate must be less than the service rate. Or put another way, the service rate must be greater than the expected arrival rate for the business.

Many businesses collect operating data in this way: the average number of customers per some time period and the typical service time. Thus some conversion of the data is necessary because the time references for the arrival and service rates must be the same. A real-life example of collecting data for coffee shops on a college campus will be given in Chapter 7.

The utilization factor ρ is 24/30 = 0.8 or 80 percent. This is the probability that the owner will be busy serving a customer when the next customer enters the shop. Hence the probability of no customers in the shop (P_0) is 1 – ρ = 20 percent. This is a useful value to know because the owner needs some time during each hour for support activities such as replenishing the cream-and-sugar station, brewing more regular coffee, and general cleanup.

The average number of customers in the shop (L) is expected to be 24/(30 – 24) = 4 customers, and the average number waiting in line (Lq) is 0.8 × 4 = 3.2 customers. You should recognize that these values are independent of the time period chosen. If the owner had collected data on customer arrivals and service times over a sequence of 15-minute periods rather than hours, the results for L and L_q would be the same. In other words, because the utilization factor is unchanged, L and L_q are unchanged. That is, the average line length is independent of the time reference for the average arrival and service rates.

The average total time spent by customers getting their coffee is 60/(30 – 24) = 10 minutes, and the time spent waiting in line is 0.8 × 10 = 8 minutes. This obviously is too long for an office worker just wanting a quick cup of regular coffee. Ways to shorten the wait for this class of customer are discussed in Chapter 6. Of interest here is that Little's Law indicates the waiting time decreases as the arrival rate increases, *provided* that the utilization factor remains constant (that is, the service rate increases proportionately with the arrival rate).

Figure 2.3 shows typical individual waiting times and service (brewing) times along with the number of customers already in the shop when each customer arrives. These values were obtained using a simple Excel simulation program. This set of sample data does not show any values for L that are greater than 5, which could lead a person to assume that the line lengths for the business are not too long. However, subsequent simulation runs using the same interarrival and service time distributions and the

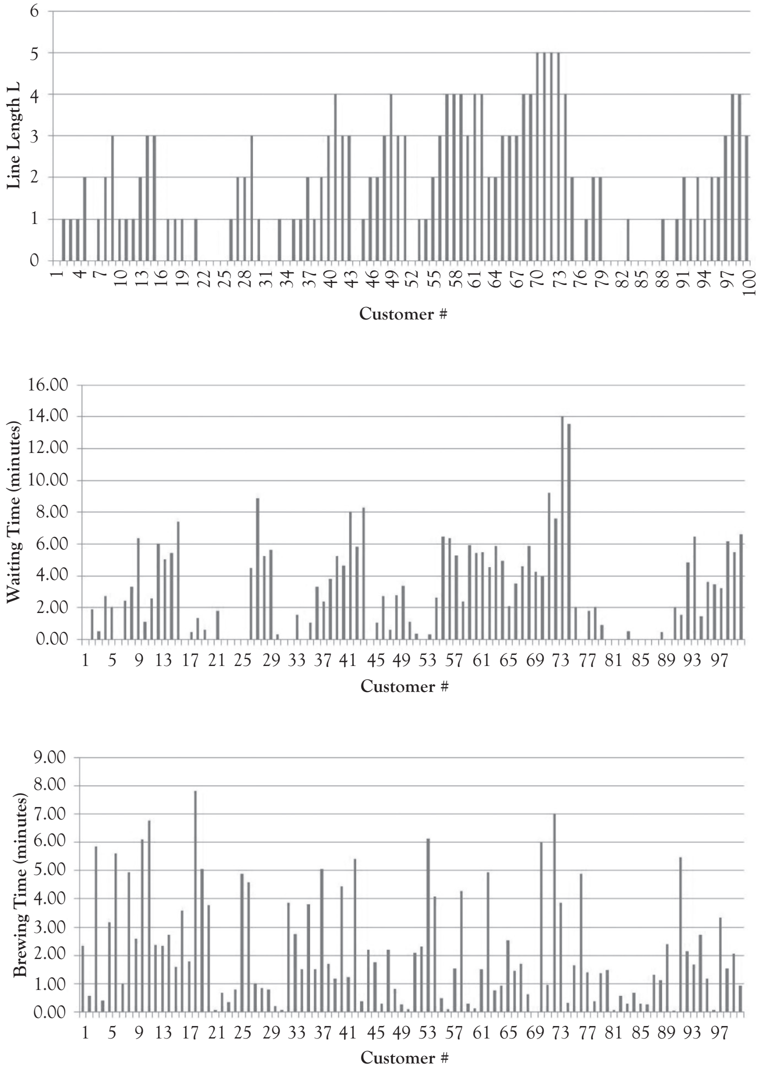

Figure 2.3 Some typical performance values for the first 100 customers entering a typical coffee shop

same number of customers have demonstrated possible line lengths as great as 14, brewing times as long as 16 minutes, and waiting times up to 20 minutes. To obtain more information regarding the variability of the performance measures, multiple simulation runs are required. These will be explored in detail in Chapter 7. You should note in Figure 2.3 that the average for each performance measure depicted rarely occurs, if at all, for any customer. *This leads to a very important observation that the average performance values derived from the queuing equations are primarily useful for the longer-term business perspective and are a very poor indicator of what a typical customer encounters.*

As mentioned in Chapter 1, an exponential distribution is not the most accurate distribution to represent the service time for this type of business because the minimum service time required for even a simple cup of house coffee is likely to be greater than 30 seconds when the time for payment is included. Yet the simulated data in Figure 2.3 show several occasions when the brewing time is less than 30 seconds (0.5 minute). The longer brewing times shown can be expected when one considers a customer ordering several coffees for a group of people. Perhaps an Erlang distribution with an appropriate k factor would be a better representation for this type of service, as discussed in Chapter 4.

Finally, market surveys can provide estimates of how long a line can be before customers looking into the shop are likely to decide to not even enter (balking). For example, let us say that a line longer than five people, including the person being served, is a definite turnoff for a potential coffee shop customer.[2] So, what is the probability of this occurring? Referring to the data in Example 2.1 and using the formula $P_{n>k} = \rho^{k+1}$, where k = 5, we obtain $0.8^6 = 0.2621$ or 26.21 percent. Thus, Ken's Caffeine Fix coffee shop is likely to lose at least one fourth of its possible customers. Conversely, the probability that the average number of customers in the shop is 5 or less is 1 – 0.2621 = 0.7379 or 73.79 percent.