CHAPTER 4

More Complex Single-Channel Models

In many service businesses and manufacturing line applications, the basic M/M/1 model does not provide useful results that appear to agree with reality. There are several possible reasons for this poor performance:

- Chief and foremost, the waiting line formulas are based on the waiting line achieving a steady-state condition. This takes time and should lead you to the observation that the formulas are likely to be a poor predictor of what occurs at the beginning of a business day.
- The service time distribution is not best described by an exponential distribution.
- The capacity for accommodating the waiting line is limited.
- The calling population is not infinite.
- The interarrival time distribution is not best described by an exponential distribution.
- The M/M/1 model is based on a single-service phase, where the server performs all the service required. How do we handle situations where there is more than one phase or step? In some cases, we can combine the separate service steps to approximate a single phase. In others, we require either more complicated analysis approaches or simulations to obtain the performance data we want.

There are several adjustments we can make in the performance measures to cope with many of these situations. Some will work out well; others will give us only rough estimates of what is happening. In such cases, discrete simulation models may be the only way to gain more

accurate insight into what is occurring. Some examples of simple simulations that can be done in Excel to analyze these situations are discussed in Chapter 7.

Instead of the normal choice of an exponential distribution for the interarrival and service times, we have a variety of probability distributions that can be used to more accurately represent a particular service business situation. Some of these distributions will make the performance measures more complicated to determine. Some will require the use of simulations because of the lack of suitable, closed-form (analytical) solutions. The good news is that because Little's Law applies regardless of our choices for service and arrival distributions, all we have to be able to do is to calculate L, L_q, W, or W_q to determine the other values.

Alternate Service Time Distributions

The M/G/1 and M/D/1 models allow for other choices of service time distributions than the exponential distribution. Some common situations that require a different choice are when

- The minimum service time is significantly greater than the nearly zero times allowed by an exponential distribution;
- The service time may have a much smaller variance than allowed by the exponential distribution;
- The service time may even be consistent enough to be considered a constant value;
- The service time can have a discrete distribution, such as in the case of a business offering only a few standardized services;
- Or the service time can be a collection of exponentially distributed service times, such as in the case of different classes of customers being served by the same facility.

Fortunately, we can obtain the steady-state average performance measures for many of these situations by using the Pollaczek-Khintchine (P-K) formula.[1] This formula uses the coefficient of variation (C_x) for a probability distribution of x to calculate the average waiting time in line (W_q). Because we are dealing only with the average steady-state value for

the waiting time, the shape of the probability distribution used for the service time is unimportant as long as we know the distribution's mean value and standard deviation or variance (the standard deviation squared). Here we assume that for an M/G/1 model, x represents the mean service time, and the mean service time is the reciprocal of the average service rate (μ):

$$W_q = \frac{1}{\mu} \times \frac{\lambda}{(\mu - \lambda)} \times \frac{\left(1 + C_{(1/\mu)}^2\right)}{2}, \quad \text{where } C_{(1/\mu)} = \frac{\text{standard deviation}}{\text{mean}}. \tag{4.1}$$

Examine this formula carefully. For a constant service time where the standard deviation is zero, the time spent waiting in line is exactly half of what is predicted for the M/M/1 model using the same average service time.

You may very well say, "No service time is exactly constant where people are involved, what then?" In that case, we can use a normal distribution, where we collect samples of actual service times and calculate their average (1/μ) and corresponding standard deviation (σ). This would give a C(1/μ) value of μσ, and Equation 4.1 for a M/G/1 model using a normal service time distribution becomes

$$W_q = \frac{1}{\mu} \times \frac{\lambda}{(\mu - \lambda)} \times \frac{\left(1 + (\mu\sigma)^2\right)}{2} = \frac{\rho(1 + \mu^2\sigma^2)}{2(\mu - \lambda)}. \tag{4.2}$$

Let us examine this expression more closely. It says that as the variance in the service time increases, the wait becomes longer even though the average service time does not change.

Here is an important corollary: Reducing variances in the service time reduces the average waiting time. This corollary is important because most textbook examples focus on shortening the average service time to reduce the average waiting time. Reducing variability in manufacturing or production step service times, particularly when there are many steps required can often be more beneficial than trying to shorten service time. This is one of the benefits of applying statistical process control (SPC) methods whose focus is on reducing variability in processes. Some other approaches for addressing service time variability are discussed in Chapter 6.

What about the problem with the probability of unrealistically short service times using the exponential distribution? Unfortunately, the steady-state performance measures do not illustrate this problem; one must use simulation to see its effect. However, the subtle effects on average waiting time and line lengths can be taken into account by choosing general service time distributions that have a lower probability of shorter service times. In addition to the constant and normal distributions, we can use the Erlang probability density function, which is defined by the following formula:

$$f(t) = \frac{t^{k-1}e^{-t/\alpha}}{\alpha^k(k-1)!} \text{ for } 0 \le t \le \infty. \tag{4.3}$$

where α = the scale factor,[2] k = the shape factor, αk = the mean, and the variance = $\alpha^2 k$. In a gamma distribution, k is a continuous value; in an Erlang distribution, k is restricted to integer values greater than zero and, in essence, represents the number of identical but still independent exponential distributions that are added together to form an Erlang distribution. When k = 1, this distribution defaults to the exponential distribution, where the average time is α (substitute $1/\lambda$ for the interarrival time or $1/\mu$ for the service time). This leads to another version of Equation 4.3 that is easier to understand when using it for service time distribution in $M/E_k/1$ models because it is expressed in terms of the service rate instead of α:

$$f(t) = \frac{\mu^k t^{k-1}e^{-\mu t}}{(k-1)!} \text{ for } 0 \le t \le \infty. \tag{4.4}$$

Therefore, the mean becomes k/μ, and the variance becomes k/μ^2. Plots of Equation 4.4 for an average service rate of 5 and using values of k = 1, 2, and 4 are shown in Figure 4.1. This service rate corresponds to an average service time of 0.2; while you may observe that the peak of the probability distribution for k = 2 roughly occurs at that time, that fact does not mean that the best value for k is 2. However, the mean for that curve is $k/\mu = 2/5 = 0.4$, indicating that there are a larger number of possible times greater than the most probable time compared to the number of possible times less than the most probable time.

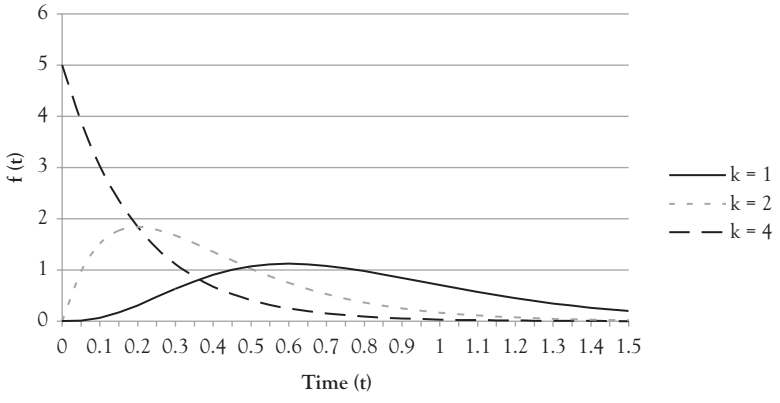

Figure 4.1 Probability density functions for an average service rate of 5 using an Erlang distribution with k = 1, 2, and 4

This illustrates a common area of confusion when trying to understand the mathematics behind the waiting line performance equations. My students often confused the maximum value of a probability density function with the average they obtained from a collection of observations. They also forgot that the Central Limit Theorem states that the distribution of sample averages is normally distributed regardless of the underlying distribution being sampled. I demonstrated to students that knowing the mean is not enough to determine which probability density function applies by giving them two sets of data that provided the same mean and standard deviation but showed significantly different frequency distributions when the raw sample results were plotted in histograms.

So, let us use the mean and variance for an Erlang distribution in the P-K formula given in Equation 4.1. In this case, $C_{(1/\mu)} = \frac{1}{\sqrt{k}}$, resulting in

$$W_q = \frac{1}{\mu} \times \frac{\lambda}{(\mu - \lambda)} \times \frac{(1 + (1/k))}{2} = \frac{\rho(k+1)}{2k(\mu - \lambda)}. \quad (4.5)$$

As a check, you can see that when k = 1, we obtain the same equation for an exponential distribution in the M/M/1 model. What if we try k = 2 to reduce the possibility of really short service times? Then using Equation 4.5, W_q = 75 percent of the W_q for the M/M/1 model. If the choice for k is much larger, we converge on the same result as for a constant service time.

How can this be? We have reduced the probability of short service times, so should the wait not be longer? This illustrates how our intuition is sometimes misleading when it comes to analyzing waiting lines. First, keep in mind that the exponential distribution was not representing realistic service times for Ken's Caffeine Fix, so comparing the new W_q to an incorrect W_q is inappropriate. Second, the higher k factor could have also reduced the probability of really long service times. What really matters is which service time distribution is the best representation of reality for a business like Ken's Caffeine Fix.

Example 4.1 Coffee Shop Using Different Service Distributions

Let us compare the differences in performance using different service distributions for Ken's Caffeine Fix analyzed in Example 2.1. The values for ρ and P_0 remain the same, but W_q, L_q, W, and L will vary. The average service rate of 30 customers per hour will be the same for all distributions, and we will assume a standard deviation of service time of 15 seconds for the normal distribution. Table 4.1 shows the results.

Observe that the best performance is with a normal service time distribution. Why is the constant service time not providing the best performance? Once again we must consider that when there is some variance in service time, we can have better results because there is a reasonable probability that the shorter service times will occur at the same time as when there is a higher than average arrival rate, enabling the line to move faster in such cases. Similarly, there is also a probability of longer service times occurring when the arrival rate is lower than normal, which will reduce their usual effect in increasing line lengths.

Table 4.1 Ken's caffeine fix coffee shop performance measures for different service distributions with an average service rate of 30 customers per hour

Performance	Exponential	Constant	Normal	Erlang (k = 2)
W_q (minutes)	8	4	3.05	4.5
W (minutes)	10	5	3.18	5.63
L_q (customers)	3.2	1.6	1.22	1.8
L (customers)	4	2	1.52	2.25

The combination of these two possibilities results in slightly improved performance as long as the variance in service time is not too large. For example, if the standard deviation increases from 15 seconds to 30 seconds, W_q increases to 3.19 minutes, and L_q increases to 1.28 customers—as expected.

What about the beta distribution commonly used to represent completion times in project management planning? This asymmetrical distribution could be used because it allows the selection of a minimum time and the probability for much longer times. However, its mean and variance terms are much more complex compared to the choice of an Erlang distribution with the proper value of k. In project management analysis, the beta distribution is often approximated by a normal distribution to simplify the mathematics. This approximation and its use for probabilistic estimates of project completion time can be found in many project management texts.

Finally, sometimes there is not enough information to determine what might be the best distribution to use. In such cases, we can take what historical operational data we do have regarding service times to form a discrete distribution that can be used in a simulation program. An example of this will be discussed in Chapter 7.

Alternate Arrival Time Distributions

For most service business situations, the exponential distribution is a very good representation of customer arrival behavior. But, for some businesses, a better representation is needed. The Kendall notation for this situation using a single-channel, single-phase system is G/M/1:

- The arrival distribution may not be best represented by a Poisson distribution.
- Arrival rates can vary during the day and often have regularly occurring peak and slack periods depending on the nature of the service(s) provided.
- The arrival population could be a combination of different types of customers.

When appointments or production schedules are used to control the arrival rate, the notation is D/M/1. You may ask why we would want to use waiting line analysis for such a predictable situation. If we want to select the best time between successive appointments, it is necessary to evaluate the effects of varying service times. If the appointment duration is too short, we end the day with a long line of disgruntled customers still waiting to be served. If the appointment time is too long, we reduce the number of customers we can accommodate per day and will have more idle time than we can profitably use to do other tasks, such as preparation, cleanup, record keeping, and so forth. In other words, we must ensure that the average arrival rate is sufficiently lower than the average service rate to prevent unnecessary backups but not so low that we unnecessarily limit our capacity and create excess idle time. Selecting the optimal value is a cost decision that we will discuss in Chapter 6.

Dealing with varying arrival rates is difficult because the waiting line equations assume steady-state conditions. For small businesses, there is likely to be insufficient volume to achieve a steady-state situation each day. For businesses such as government driver's license bureaus or the post office, where the customer really does not have a choice of an alternate supplier, you can average the total number of arrivals during the day to determine the level of service required. For businesses such as Ken's Caffeine Fix, staffing has to better match the ebb and flow of customers to provide acceptable service for customers who are not as tolerant of long lines and waiting times.

In both cases, but more so for Ken's and places like banks and grocery stores, it is important to have staff that can perform other business tasks when customer arrival volume is low. Knowing how much time is required for those tasks and the estimate of P_0 from the queuing analysis can provide guidance as to the total staffing needed.

Many businesses have collected point-of-sale data regarding the nature of the services they provide and the types of customers. Such data can be used to improve waiting line performance in several ways, particularly when more than one server is available, as will be discussed in more detail in Chapter 5. For single server operations, this information can be used to ensure that adequate supplies are available for the different services

requested by customers and help identify those parts of the service that might be delegated to some of the customers. For example, a coffee shop, noting that about 30 percent of its customers want a cup of house coffee only, could set up a self-serve station for those customers so that customers wanting more complex coffee mixtures could be served more quickly.

In factory applications, manufacturing custom or semicustom products creates different classes of arrival and service rates. Depending on the overall demand separate production lines or appropriate production scheduling could be used to provide such products efficiently. Another application where different classes of customers are likely is the machine repair model discussed later in this chapter when we cover the limited calling population model.

From an analysis viewpoint, it is relatively easy to combine separate classes of customers if the arrival behavior of each class can be described by a separate exponential distribution. In this case, the average arrival rate for the total distribution is just the sum of the individual average arrival rates. That is, $\lambda_{total} = \lambda_1 + \lambda_2 + \lambda_3 + \ldots$. Similarly, a total exponential distribution can be separated into individual exponential distributions according to their probability of occurrence.

Finally, there is one common arrival scenario that is difficult to classify, let alone analyze. Consider the restaurant situation where there is the usual exponential arrival of individual customers combined with regularly scheduled arrivals who have made reservations. So how does the restaurant handle an unscheduled tour bus unloading a large group of passengers for lunch or dinner? Does the restaurant mix them in with individual customers to determine the appropriate service level or do they treat the group separately? What if the tour bus stops are scheduled?

A closed-form solution does not exist for many of these arrival situations, and those that have been developed for very special situations are beyond the mathematical scope covered by this monograph. What we can do here is to attempt to break such a situation into separate parts that can be analyzed individually. For example, in the restaurant scenario, management could choose to require that groups above a certain size make a reservation to be served—in essence, creating two separate service operations that could be analyzed independently.

Limited Capacity

There can be physical limits on how many customers may be in line. This may take the form of how many customers can be accommodated on hold when a single operator at a call center is busy or when there is not enough room to accommodate a long line in a small shop. The Kendall notation for this situation is M/M/1/K/∞/FCFS, where K is the maximum number of people in a queue, including the customers being served at the moment. A specific example of this model was analyzed by the state diagram discussed in Chapter 1 and shown in Figure 1.3. In that example, the value for K was two customers.

The performance measures for an M/M/1/K model are similar to those for the M/M/1 model but with slight modifications to account for a system capacity of K customers:

- Utilization factor: $\rho = \lambda/\mu$; for a M/M/1/K model, ρ must be ≤ 1
- Probability of zero customers in the system: $P_0 = (1 - \rho)/(1 - \rho^{K+1})$ for $\lambda < \mu$ and $P_0 = 1/(1 + K)$ for $\lambda = \mu$
- Probability of exactly n customers in the system: $P_n = P_0 \rho^n$ for $n \leq K$
- Probability that the server is busy: $P_{n>0} = 1 - P_0 = \rho$
- Probability that a customer will be turned away: $P_K = P_0 \rho^K$
- Average number of customers in the system: $L = K/2$ for $\lambda = \mu$ and

$$L = \frac{\rho}{(1-\rho)} - \frac{(K+1)\rho^{(K+1)}}{\left(1-\rho^{(K+1)}\right)} \text{ for } \lambda < \mu \tag{4.6}$$

- Effective arrival rate: $\lambda' = \lambda(1 - P_K)$
- Average total time customers spend in the system: $W = L/\lambda' = L/(\lambda(1 - P_K))$
- Average number of customers waiting in the queue (not yet being served): $L_q = L - (\lambda'/\mu) = L - (\lambda(1 - P_K)/\mu) = \lambda' W_q$
- Average time customers wait in the queue before being served: $W_q = W - (1/\mu) = L_q/\lambda' = L_q/(\lambda(1 - P_K))$

Recall that for Little's Law, we need to use an effective arrival rate (λ'), which is the arrival rate multiplied by the probability of acceptance (one minus the probability that a customer will be turned away). There are three possible assumptions here: (1) We lose the customers who are turned away because they have alternate service choices with our competitors, (2) those customers return later to try again as would be the case for a call center where they have no other choice for the support they want, or (3) the blocked customers give up trying.

Because we lose some customers in the first assumption or can smooth out peak arrival rates for the second assumption by forcing customers to call back again later, we can allow the average arrival rate to be as high as the average service rate, which explains the conditional equations for P_0 and L.

P_K gives us a value for lost business that can be used to compare with the cost of adding additional line capacity. P_K is sometimes called the blocking probability.

The astute reader may notice that the equation given for P_0 when K = 2 is not the same as the state diagram equation derived for P_0 in Chapter 1. Let us compare them:

$$\text{Does } 1/(1 + \rho + \rho 2) = (1 - \rho)/(1 - \rho^3)?$$

Multiplying both sides by the denominators to remove the fractions, we get

$$1 - \rho^3 = (1 - \rho)(1 + \rho + \rho^2) = 1 + \rho + \rho^2 - \rho - \rho^2 - \rho^3 = 1 - \rho^3.$$

Okay, they match! (Before I get too cocky, maybe we should also check the respective equations for L just to be sure.)

Once we have the probability for each state in Figure 1.3, we can determine the average number of persons in the system by multiplying the number of persons represented by each state by that state's probability. That is, $L = (P_0 \times 0) + (P_1 \times 1) + (P_2 \times 2)$ for the state diagram in Figure 1.3, which gives $L = (\rho + 2\rho^2)/(1 + \rho + \rho^2)$. Does that equate to the expression in Equation 4.6 when K = 2? That is, does $(\rho + 2\rho^2)/(1 + \rho + \rho^2) = [\rho/(1 - \rho)] - [3\rho^3/(1 - \rho^3)]$?

Multiplying both sides by the denominators to remove the fractions and then combining the terms on both sides, we can see that the two expressions do equate to the same value.

It is suggested that you do this check for yourself as practice in verifying formulas. When there are several terms with subscripts and superscripts, it is easy for typographical errors to occur. Verifying equations and making sure that the units of measure cancel out to the desired set of units are methods for detecting many typographical errors in the literature and when drafting your own presentations.

Example 4.2 Coffee Shop with Limited Capacity

Returning to Ken's Caffeine Fix, the owner is concerned that the shop is turning away customers because the maximum customer capacity is seven customers. Using the same arrival rate of 24/hour, a service rate of 30/hour, and a capacity K = 7, we determine P_0 to be $(1 - 0.8)/ (1 - 0.8^8) = 0.240319$. Compare this value with the 20 percent value obtained for P_0 in Chapter 2; you can see that the loss of customers turned away has increased the probability of idle time.

To determine the percentage of customers lost, we need to determine the value for $P_K = P_0 \rho^K = 0.240319 \times 0.8^7 = 0.050398 = 5\%$. This loss creates an effective arrival rate $\lambda' = 24 \times (1 - P_K) = 22.8$/hour, and the average line length L = 4 from Chapter 2 is reduced using Equation 4.6 to $4 - [(8 \times 0.8^8)/(1 - 0.8^8)] = 2.39$ customers. The total wait in the system has been reduced from 10 minutes to 2.39/22.8 = 0.1048 hours = 6.29 minutes.

The lesson to be learned here is that customers lost because of capacity limitations are not often noticed because the line is longest at those times when the server is busiest. This also results in the average line length and overall waiting time being shorter, creating an illusion that the service is being performed more efficiently than it is. A 5 percent customer loss may not seem to be that significant, but when you compare its lost revenue with the profit margin for many small service businesses …

Limited Calling Population

The available customer population may be limited, as in the case for a service operation maintaining a fleet of airplanes for a major airline or a group servicing the copiers at a company. The Kendall notation for this situation when there is only one server is M/M/1/∞/N/FCFS, where N represents the number of customers or items to be served. In many textbooks and articles, this situation is often referred to as the machine repairman problem.

Because the costs of downtime can be quite high, there are often two or more servers to get equipment back in service quickly. This complicates the closed-form solutions considerably. This situation will be discussed in much more detail in Chapter 5. Here, we will discuss the single-server model because it provides some insight into the fundamental issues involved and is appropriate for many small service operations responsible for maintaining and repairing a small set of equipment. This also applies to those operations where one professional serves a select group of customers.

The arrival rate for a repair operation supporting N machines is usually dictated by three things: the recommended preventive maintenance period, the expected failure rate, and the equipment usage rate (supplies replenishment). Equivalents of these rate components also apply to handling a select group of clients, such as at a financial advisor brokerage: regularly scheduled status appointments, typical percentage of emergency consultations, and volume-related requests.

We can attempt to analyze these classes as one combined calling population, or we can assign separate servers for each class. Consider that the preventive maintenance arrivals are essentially appointment based, and their respective service times are likely to be relatively constant or at least normally distributed with small standard deviations. The failure rate and subsequent repair times are more likely to be described by exponential distributions.

The base arrival rate for an M/M/1/∞/N model is defined by the needs of one customer, item, or machine for the situation under consideration. In the case of a machine, it can be the expected time between repairs

or scheduled maintenance. The effective arrival rate is the base arrival rate per unit λ_u multiplied by the number of units (N − L) in the finite population N that are not already in line for service in the system. The calculations for P_0 with a finite population of N customers (machines) are considerably nastier because now we have to account for the reduced probability of future arrivals as current arrivals enter the system and are being taken care of. This requires using N summation terms, as shown in the following set of performance measures:

- Effective arrival rate: $\lambda' = \lambda_u(N - L)$
- Unit utilization factor $\rho_u = \lambda_u/\mu$
- Probability of zero customers in the system:

$$P_0 = \frac{1}{\sum_{n=0}^{N}\left[\frac{N!}{(N-n)!} \times \rho_u^n\right]} \tag{4.7}$$

- Probability of exactly n customers in the system:

$$P_n = P_0 \times \left[\frac{N!}{(N-n)!} \times \rho_u^n\right] \text{for } n \leq N;\ P_n = 0 \text{ for } n > N$$

- Average number of customers waiting in the queue (not yet being served):

$$L_q = N - \left[\left(\frac{\lambda_u + \mu}{\lambda_u}\right)(1 - P_0)\right]$$

- Average number of customers in the system: $L = L_q + (1 - P_0)$
- Average time customers wait in the queue before being served:

$$W_q = L_q/(\lambda_u(N - L)$$

- Average total time customers spend in the system:

$$W = W_q + (1/\mu) = L/(\lambda_u(N - L)$$

Looking at the equation for P_0, it increases in value as N becomes smaller, which is what we would expect with a lighter workload. However,

as N becomes larger, there will be many terms in the denominator. This leads some business students to ask the question, "How large would N have to be where we could assume that an infinite calling population model could give us a ballpark estimate with only a small percentage of error? The calculations would certainly be easier to do."

If we attempt to do this, students are reminded that they must recognize that we are not dealing with the same definition of arrival rate. In a repair service situation, the average arrival rate is based on some expected rate of failure per unit; as the number of possible items to be repaired increases, the effective arrival rate increases. The M/M/1/∞/N model takes these increases in N into account in its performance measures, but an M/M/1/∞/∞ model assumes an average arrival rate that is independent of population size. That is, some of that infinite population can choose to never arrive, or at least take a very long time before they choose to do so.

Example 4.3 Repair Service With One Server

Consider an in-house service activity staffed with a single person. Campus Reboot is responsible for maintaining a set of network printers located in several buildings clustered together as part of a liberal arts college. Campus Reboot is only responsible for repairs because routine reloading of paper, installing new toner cartridges, and clearing simple paper jams are done when needed by the respective department secretaries. At the moment, there are only 10 of these printers in use, but their popularity with the faculty for printing exams and quizzes is increasing, and the dean of the college wants to know what effect increasing the number of printers to 25 or even 50 units will have on the repair response. For now, we will ignore the associated costs but will return to this example in Chapter 6 to discuss the cost tradeoffs. The failure rate for the current printer model used is 250 hours between repairs, and the typical repair takes an average of four hours, which includes any time spent to order parts or travel between buildings. This information converts into an arrival rate λ_u per printer of 1/250 = 0.004 printer/hour and an average service rate of 0.25 printer/hour. This results in a value for $\rho_u = 0.004/0.25 = 0.016$.

There is an important consideration to note here. Some of you may be tempted to multiply the arrival rate per item by the total number of items to obtain the average arrival rate for λ. *Do not* do this because the derivation of the equations is based on the arrival rate per unit, which I designate by λ_u to avoid confusion with the more common understanding of what λ represents.

For example, consider the state diagram for a population of four items, as shown in Figure 4.2, where each state indicated by the hexagons represents the number of items L in the system. The arrival rate from state 0 to state 1 is $4\lambda_u$ because the entire population is available to move to that state. Similarly, the number of items available to move from state 3 to state 4 is only one because all the other items in the population are already in the system. The rate for moving the other direction, reducing the number of items in the system, is the service rate, which can be applied to only one item at a time. If there are two servers instead of one server, then the reduction rate is 2μ from states 2, 3, and 4 and μ from state 1.

If you use Excel to determine P_0, I recommend having separate columns as shown in Figure 4.3, for the terms in the summation portion of the denominator of Equation 4.7: n, $(\rho_u)^n$, (N – n)!, and the product within the brackets, rather than attempting to compose a single expression to calculate P_0. Using N as an input value and summing the column for the products allows a simpler algorithm for computing the final result. This helps minimize computation errors, which is critically important for P_0 because it is used for subsequent calculations. It also allows for easy expansion to larger N values by using Excel's AutoFill capability to copy the respective formulas in each column further down the sheet. Appendix E shows the formulas used to determine the results in each of the cells depicted in Figure 4.3 for those interested.

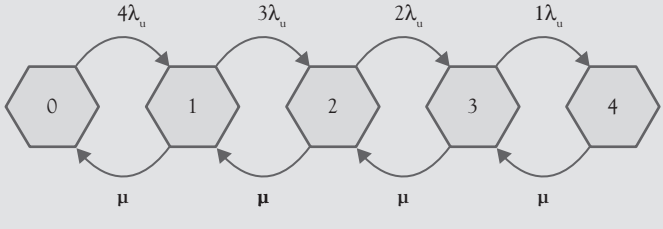

Figure 4.2 State diagram for a single-server system with a limited population of four items with individual arrival rates of λ_u

Table 4.2 Performance measures for M/M/∞/N model with N = 10, 25, and 50 units using the input data for campus reboot from Example 4.3

N	10	25	50
P_0	0.842919	0.609848	0.242301
L_q	0.025346	0.225374	1.886103
L	0.182427	0.615525	2.643802
λ'	0.0392702	0.0975379	0.1894248
W_q (hours)	0.645432	2.310626	9.957002
W (hours)	4.645432	6.310626	13.957

Table 4.2 summarizes the results for repair populations of 10, 25, and 50 printers in response to the dean's request regarding the effect of adding more printers to Campus Reboot's repair response. With 10 printers, the average wait before repair work begins is about 39 minutes. This average waiting time increases to about 2.3 hours for 25 printers and nearly 10 hours for 50 printers. The number of printers

n	(N – n)!	N!/(N – n)!	$(\rho_u)^n$	Product	P_n
0	3628800	1	1	1	0.842919
1	362880	10	0.016	0.16	0.134867
2	40320	90	0.000256	0.02304	0.019421
3	5040	720	4.1E-06	0.002949	0.002486
4	720	5040	6.55E-08	0.00033	0.000278
5	120	30240	1.05E-09	3.17E-05	2.67E-05
6	24	151200	1.68E-11	2.54E-06	2.14E-06
7	6	604800	2.68E-13	1.62E-07	1.37E-07
8	2	1814400	4.29E-15	7.79E-09	6.57E-09
9	1	3628800	6.87E-17	2.49E-10	2.1E-10
10	1	3628800	1.1E-18	3.99E-12	3.36E-12
			Sum	1.186354	1

Figure 4.3 Excel setup to determine P_0 (shaded cell at upper right) and P_n for Example 4.3 with a population of 10 network printers. P_0 is the inverse of the sum of the products in the fifth column, and the other P_ns are the respective product for the value of n in the first column multiplied by the value for P_0. As a check on the accuracy of the computations, the sum of the probabilities P_n in the sixth column is 1.0 (See the Excel formula view at the beginning of Appendix E for the calculations used in this table.)

awaiting repair increases from the current value of 0.025 printer to nearly two printers for a 50-printer population.

Reviewing the results in Figure 4.3, the likelihood of more than one printer awaiting repair (two or more printers total in the repair shop) by Campus Reboot is less than $(1 - P_0 - P_1) = 0.022214 = 2.22$ percent for the current population of 10 printers.

If Campus Reboot works only a normal 8-hour day, it is likely that it will often take more than two days before some departments get back their broken printer if the number of printers increases to 50. This would be unacceptable given that a normal exam week is only five days long. We will return to this example in Chapter 5 to determine the effect of adding another repair person to Campus Reboot's staff.

It should be noted here that many college textbooks still use lookup tables to solve this type of example for both single-channel and multiple-channel situations. As a result, the equations presented by those authors do not look anything like the equations used here. Instead, they use terms and expressions that work with these tables, often referred to in some textbooks as finite queuing tables (to be more correct some authors use the terms finite queuing for limited capacity models and finite sourcing for limited calling population models, which applies here). An example is the service factor X = average service (repair) time T divided by the sum of the average repair time T and the average time between repairs U. For Example 4.3, Campus Reboot's service factor would be $4/(250 + 4) = 0.0157$. The finite sourcing table for a given repair population level provides two variables, D and F, for each combination of X and number of servers. Using these variables, one can then determine the usual performance measures.

Such methods are a product of an earlier time when personal computers and spreadsheet programs were not yet available. At that time, calculations like those shown in Figure 4.3 using pencil-and-paper methods were tedious, time-consuming, and more prone to errors. This led to the development of lookup tables for various values of N, service factors, and the number of servers (repair persons) to expedite such analysis. Often-quoted examples of finite sourcing tables in college

textbooks are sample tables for N = 5 or 10 used with permission from the 1958 reference published by Peck and Hazelwood.[3]

In this monograph readers are encouraged to use Excel-based methods because they allow a wider range of choices and conditions to best suit a given business situation. Instead of providing lookup tables, the equations for their creation are provided. Readers wishing more details regarding the derivation of these equations are encouraged to consult the excellent books by Hillier and Lieberman (2010) or Laguna and Marklund (2005).

Multiple Phases

In factory applications and some service businesses, there is more than one step (phase) in the process. How do we deal with those cases? Because each phase is likely to have a different service rate distribution, a closed-form generic solution would be unwieldy, even if the solution were possible. Some references are provided for those who are interested in pursuing closed-form solutions when they are available. A *cautionary note is appropriate here*. Many of the analytical solutions involve some in-depth mathematics and statistics understanding to determine whether they are appropriate for your business situation.

What if you do not want to become a math whiz to obtain some insight into how a multiple-phase process works? Not to fear, there are some characteristics we can work with to gain some understanding about such processes without the need for complex mathematics.

- The average arrival rate into each phase will be the same, assuming no losses or additions at each phase and that the average service rate for the phase is greater than the average arrival rate, because what went into the previous phase must come out of that phase.
- If losses or additions do occur during the process, they can be handled if we know their percentage related to the initial input arrival rate.
- In the case of factory applications, we usually have control over the arrival rate distribution by using production scheduling.

- Each phase can have a different service rate without affecting the average arrival rates into each phase. WARNING! This is still subject to the condition that the average service rate is greater than the average arrival rate. If the average service rate for a step is less than the average arrival rate, that step will become a bottleneck, causing work to pile up before that step and that step's service rate will determine the capacity for the entire production sequence of steps.

Given these conditions, we can analyze or simulate a production process as a sequence of single-channel, single-phase waiting lines with the output of the preceding phase becoming the input to the following phase.

What do you need to look out for? A number of things can interfere with an otherwise successful simulation model. First, it is very, very difficult to include all of the disruptions that can occur in a multiple-phase process. Second, relatively predictable disruptions such as downtime for each phase for preventive maintenance, repair, operator training, late material deliveries, and tool setup time are often overlooked as to their effects on service and arrival rates. Third, while it is possible to include many disruptions in a simulation model many businesses do not collect sufficient operating data regarding their past frequency of occurrence or probability of happening. The need for such information and how to collect it is discussed in more detail in Chapter 6.

Because we can also accommodate a mixture of multiple-channel, single-phase process steps in manufacturing to deal with bottleneck capacity issues described above, there is a more detailed discussion of multiple-phase applications in Chapter 7.