

CHAPTER 15

Queuing Systems

Chapter Guide. The objective of queuing analysis is to offer a reasonably satisfactory service to waiting customers. Unlike the other tools of OR presented in the preceding chapters, queuing theory is not an optimization technique. Rather, it determines the measures of performance of waiting lines, such as the average waiting time in queue and the productivity of the service facility, which can then be used to design the service installation. This chapter emphasizes the implementation of queuing results in practice. However, to fully appreciate the practical side of queuing, you will need a reasonable background in the underlying theory. For this reason, the chapter starts with a presentation of the “total randomness” property of two important distributions: the Poisson and the exponential. This point is important because it helps identify the situations where queuing results apply in practice.

Queuing results involve computationally difficult formulas, and it is recommended that you use `exelPoissonQ.xls` or `TORA` to carry out these calculations. You will find `TORA` helpful in comparing multiple scenarios. Throughout the chapter, `TORA` is used to carry out the computations. The bulk of the discussion concentrates on the practical interpretations of the results. We recommend that you follow the same procedure when you work out the problems in this chapter. In this manner, you are not “bogged down” in the tedious computational details and can readily test different scenarios conveniently.

This chapter includes a summary of 2 real-life applications, 17 solved examples, 2 Excel templates, 137 end-of-section problems, and 5 cases. The cases are in Appendix E on the CD. The `AMPL/Excel/Solver/TORA` programs are in folder `ch15Files`.

Real-Life Application—Analysis of an Internal Transport System in a Manufacturing Plant

Three trucks are used in a manufacturing plant to transport materials. The trucks wait in a central parking lot until requested. A truck answering a request will travel to the customer location, carry a load to its destination, and then return to the central parking lot. The principal user of the service is production, followed by the workshop and maintenance. Other departments occasionally may request the use of the trucks.

Complaints about the long wait for a free truck have prompted users, especially production, to request adding a fourth truck to the fleet. This is an unusual application, because queuing theory is used to show that the source of the long delays is mainly logistical and that with a simple change in the operating procedure of the truck pool, a fourth truck is not needed. Case 14 in Chapter 24 on the CD provides the details of the study.

15.1 WHY STUDY QUEUES?

Waiting for service is part of our daily life. We wait to eat in restaurants, we “queue up” at the check-out counters in grocery stores, and we “line up” for service in post offices. And the waiting phenomenon is not an experience limited to human beings only: Jobs wait to be processed on a machine, planes circle in a stack before given permission to land at an airport, and cars stop at traffic lights. Waiting cannot be eliminated completely without incurring inordinate expenses, and the goal is to reduce its adverse impact to “tolerable” levels.

The study of queues deals with quantifying the phenomenon of waiting in lines using representative measures of performance, such as average queue length, average waiting time in queue, and average facility utilization. The following example demonstrates how these measures can be used to design a service facility.

Example 15.1-1

McBurger is a fast-food restaurant with three service counters. The manager has commissioned a study to investigate complaints about slow service. The study reveals the following relationship between the number of service counters and the waiting time for service:

No. of cashiers	1	2	3	4	5	6	7
Average waiting time (min)	16.2	10.3	6.9	4.8	2.9	1.9	1.3

An examination of these data shows a 7-minute average waiting time for the present 3-counter situation. Five counters are needed to reduce the waiting time to about 3 minutes.

Remarks. The results of queuing analysis can be used in the context of a cost optimization model, where we seek the minimization of the sum of two costs: the cost of offering the service and the cost of waiting. Figure 15.1 depicts a typical cost model (in dollars per unit time) where the cost of service increases with the increase in the level of service (e.g., the number of service counters). At the same time, the cost of waiting decreases with the increase in level of service. The main obstacle in implementing cost models is the difficulty of obtaining reliable estimates of the cost of waiting, particularly when human behavior is an integral part of the operation. This point is discussed in Section 15.9.

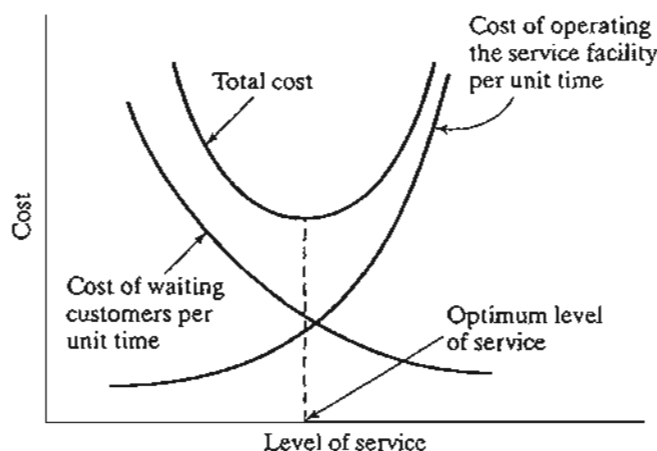


FIGURE 15.1
Cost-based queuing decision model

PROBLEM SET 15.1A

- *1. Suppose that further analysis of the McBurger restaurant reveals the following additional results:

No. of cashiers	1	2	3	4	5	6	7
Idleness (%)	0	8	12	18	29	36	42

- What is the productivity of the operation (expressed as the percentage of time the employees are busy) when the number of cashiers is five?
 - The manager wants to keep the average waiting time around 3 minutes and, simultaneously, maintain the efficiency of the facility at approximately 90%. Can the two goals be achieved? Explain.
2. Acme Metal Jobshop is in the process of purchasing a multipurpose drill press. Two models, *A* and *B*, are available with hourly operating costs of \$18 and \$25, respectively. Model *A* is slower than model *B*. Queuing analysis of similar machines shows that when *A* is used, the average number of jobs in the queue is 4, which is 30% higher than the queue size in *B*. A delayed job represents lost income, which is estimated by Acme at \$10 per waiting job per hour. Which model should Acme purchase?

15.2 ELEMENTS OF A QUEUING MODEL

The principal actors in a queuing situation are the **customer** and the **server**. Customers are generated from a **source**. On arrival at a service **facility**, they can start service immediately or wait in a **queue** if the facility is busy. When a facility completes a service, it automatically “pulls” a waiting customer, if any, from the queue. If the queue is empty, the facility becomes idle until a new customer arrives.

From the standpoint of analyzing queues, the arrival of customers is represented by the **interarrival time** between successive customers, and the service is described by the **service time** per customer. Generally, the interarrival and service times can be

probabilistic, as in the operation of a post office, or deterministic, as in the arrival of applicants for job interviews.

Queue size plays a role in the analysis of queues, and it may have a finite size, as in the buffer area between two successive machines, or it may be infinite, as in mail order facilities.

Queue discipline, which represents the order in which customers are selected from a queue, is an important factor in the analysis of queuing models. The most common discipline is **first come, first served** (FCFS). Other disciplines include **last come, first served** (LCFS) and **service in random order** (SIRO). Customers may also be selected from the queue based on some order of **priority**. For example, rush jobs at a shop are processed ahead of regular jobs.

The queuing behavior of customers plays a role in waiting-line analysis. "Human" customers may **jockey** from one queue to another in the hope of reducing waiting time. They may also **balk** from joining a queue altogether because of anticipated long delay, or they may **renege** from a queue because they have been waiting too long.

The design of the service facility may include parallel servers (e.g., post office or bank operation). The servers may also be arranged in series (e.g., jobs processed on successive machines), or they may be networked (e.g., routers in a computer network).

The source from which customers are generated may be finite or infinite. A **finite source** limits the customers arriving for service (e.g., machines requesting the service of a repairperson). An **infinite source** is forever abundant (e.g., calls arriving at a telephone exchange).

Variations in the elements of a queuing situation give rise to a variety of queuing models. This chapter provides examples of these models.

PROBLEM SET 15.2A

1. In each of the following situations, identify the customer and the server:
 - *(a) Planes arriving at an airport.
 - *(b) Taxi stand serving waiting passengers.
 - (c) Tools checked out from a crib in a machining shop.
 - (d) Letters processed in a post office.
 - (e) Registration for classes in a university.
 - (f) Legal court cases.
 - (g) Check-out operation in a supermarket.
 - *(h) Parking lot operation.
2. For each of the situations in Problem 1, identify the following: (a) nature of the calling source (finite or infinite), (b) nature of arriving customers (individually or in bulk), (c) type of the interarrival time (probabilistic or deterministic), (d) definition and type of service time, (f) queue capacity (finite or infinite), and (g) queue discipline.
3. Study the following system and identify the associated queuing situations. For each situation, define the customers, the server(s), the queue discipline, the service time, the maximum queue length, and the calling source.

Orders for jobs are received at a workshop for processing. On receipt, the supervisor decides whether it is a rush or a regular job. Some orders require the use of one of

several identical machines. The remaining orders are processed in a two-stage production line, of which two are available. In each group, one facility is assigned to handle rush jobs.

Jobs arriving at any facility are processed in order of arrival. Completed orders are shipped on arrival from a shipping zone having a limited capacity.

Sharpened tools for the different machines are supplied from a central tool crib. When a machine breaks down, a repairperson is summoned from the service pool to make the repair. Machines working on rush orders always receive priorities both in acquiring new tools from the crib and in receiving repair service.

4. True or False?
 - (a) An impatient waiting customer may elect to renege.
 - (b) If a long waiting time is anticipated, an arriving customer may elect to balk.
 - (c) Jockeying from one queue to another is exercised to reduce waiting time.
5. In each of the situations in Problem 1, discuss the possibility of the customers jockeying, balking, and renegeing.

15.3 ROLE OF EXPONENTIAL DISTRIBUTION

In most queuing situations, the arrival of customers occurs in a *totally random* fashion. Randomness here means that the occurrence of an event (e.g., arrival of a customer or completion of a service) is not influenced by the length of time that has elapsed since the occurrence of the last event.

Random interarrival and service times are described quantitatively in queuing models by the **exponential distribution**, which is defined as

$$f(t) = \lambda e^{-\lambda t}, t > 0$$

Section 12.4.3 shows that for the exponential distribution

$$\begin{aligned} E\{t\} &= \frac{1}{\lambda} \\ P\{t \leq T\} &= \int_0^T \lambda e^{-\lambda t} dt \\ &= 1 - e^{-\lambda T} \end{aligned}$$

The definition of $E\{t\}$ shows that λ is the rate per unit time at which events (arrivals or departures) are generated. The fact that the exponential distribution is **completely random** is illustrated by the following example: If the time now is 8:20 A.M. and the last arrival has occurred at 8:02 A.M., the probability that the next arrival will occur by 8:29 is a function of the interval from 8:20 to 8:29 only, and is totally independent of the length of time that has elapsed since the occurrence of the last event (8:02 to 8:20). This result is referred to as the **forgetfulness** or **lack of memory** of the exponential.

Let the exponential distribution, $f(t)$, represent the time, t , between successive events. If S is the interval since the occurrence of the last event, then the *forgetfulness property* implies that

$$P\{t > T + S | t > S\} = P\{t > T\}$$

To prove this result, we note that for the exponential with mean $\frac{1}{\lambda}$,

$$P\{t > Y\} = 1 - P\{t < Y\} = e^{-\lambda Y}$$

Thus,

$$\begin{aligned} P\{t > T + S | t > S\} &= \frac{P\{t > T + S, t > S\}}{P\{t > S\}} = \frac{P\{t > T + S\}}{P\{t > S\}} \\ &= \frac{e^{-\lambda(T+S)}}{e^{-\lambda S}} = e^{-\lambda T} \\ &= P\{t > T\} \end{aligned}$$

Example 15.3-1

A service machine always has a standby unit for immediate replacement upon failure. The time to failure of the machine (or its standby unit) is exponential and occurs every 5 hours, on the average. The machine operator claims that the machine "has the habit" of breaking down every night around 8:30 P.M. Analyze the operator's claim.

The average failure rate of the machine is $\lambda = \frac{1}{5} = .2$ failure per hour. Thus, the exponential distribution of the time to failure is

$$f(t) = .2e^{-.2t}, t > 0$$

Regarding the operator's claim, we know offhand that it cannot be correct because it conflicts with the fact that the time between breakdowns is exponential and, hence, totally random. The probability that a failure will occur by 8:30 P.M. cannot be used to support or refute the operator's claim, because the value of such probability depends on the time of the day (relative to 8:30 P.M.) at which it is computed. For example, if the time now is 8:20 P.M., the probability that the operator's claim will be right tonight is

$$P\{t < \frac{10}{60}\} = 1 - e^{-2(\frac{10}{60})} = .03278$$

which is low. If the time now is 1:00 P.M., the probability that a failure will occur by 8:30 P.M. increases to approximately .777 (verify!). These two extreme values show that the operator's claim cannot be supported.

PROBLEM SET 15.3A

1. (a) Explain your understanding of the relationship between the arrival rate λ and the average interarrival time. What are the units describing each variable?
- (b) In each of the following cases, determine the average arrival rate per hour, λ , and the average interarrival time in hours.
 - (i) One arrival occurs every 10 minutes.
 - (ii) Two arrivals occur every 6 minutes.
 - (iii) Number of arrivals in a 30-minute period is 10.
 - (iv) The average interval between successive arrivals is .5 hour.

- (c) In each of the following cases, determine the average service rate per hour, μ , and the average service time in hours.
- *(i) One service is completed every 12 minutes.
 - (ii) Two departures occur every 15 minutes.
 - (iii) Number of customers served in a 30-minute period is 5.
 - (iv) The average service time is .3 hour.
2. In Example 15.3-1, determine the following:
- (a) The average number of failures in 1 week, assuming the service is offered 24 hours a day, 7 days a week.
 - (b) The probability of at least one failure in a 2-hour period.
 - (c) The probability that the next failure will *not* occur within 3 hours.
 - (d) If no failure has occurred 3 hours after the last failure, what is the probability that interfailure time is at least 4 hours?
3. The time between arrivals at the State Revenue Office is exponential with mean value .05 hour. The office opens at 8:00 A.M.
- *(a) Write the exponential distribution that describes the interarrival time.
 - *(b) Find the probability that no customers will arrive at the office by 8:15 A.M.
 - (c) It is now 8:35 A.M. The last customer entered the office at 8:26. What is the probability that the next customer will arrive before 8:38 A.M.? That the next customer will not arrive by 8:40 A.M.?
 - (d) What is the average number of arriving customers between 8:10 and 8:45 A.M.?
4. Suppose that the time between breakdowns for a machine is exponential with mean 6 hours. If the machine has worked without failure during the last 3 hours, what is the probability that it will continue without failure during the next hour? That it will break down during the next .5 hour?
5. The time between arrivals at the game room in the student union is exponential with mean 10 minutes.
- (a) What is the arrival rate per hour?
 - (b) What is the probability that no students will arrive at the game room during the next 15 minutes?
 - (c) What is the probability that at least one student will visit the game room during the next 20 minutes?
6. The manager of a new fast-food restaurant wants to quantify the arrival process of customers by estimating the fraction of interarrival time intervals that will be (a) less than 2 minutes, (b) between 2 and 3 minutes, and (c) more than 3 minutes. Arrivals in similar restaurants occur at the rate of 35 customers per hour. The interarrival time is exponentially distributed.
- *7. Ann and Jim, two employees in a fast-food restaurant, play the following game while waiting for customers to arrive: Jim pays Ann 2 cents if the next customer does not arrive within 1 minute; otherwise, Ann pays Jim 2 cents. Determine Jim's average payoff in an 8-hour period. The interarrival time is exponential with mean 1.5 minute.
8. Suppose that in Problem 7 the rules of the game are such that Jim pays Ann 2 cents if the next customer arrives after 1.5 minutes, and Ann pays Jim an equal amount if the next arrival is within 1 minute. For arrivals within the range 1 to 1.5 minutes, the game is a draw. Determine Jim's expected payoff in an 8-hour period.

9. In Problem 7, suppose that Ann pays Jim 2 cents if the next arrival occurs within 1 minute and 3 cents if the interarrival time is between 1 and 1.5 minutes. Ann receives from Jim 5 cents if the interarrival time is between 1.5 and 2 minutes and 6 cents if it is larger than 2 minutes. Determine Ann's expected payoff in an 8-hour period.
- *10. A customer arriving at a McBurger fast-food restaurant within 4 minutes of the immediately preceding customer will receive a 10% discount. If the interarrival time is between 4 and 5 minutes, the discount is 6%. If the interarrival time is longer than 5 minutes, the customer gets 2% discount. The interarrival time is exponential with mean 6 minutes.
 - (a) Determine the probability that an arriving customer will receive the 10% discount.
 - (b) Determine the average discount per arriving customer.
11. The time between failures of a Kencore refrigerator is known to be exponential with mean value 9000 hours (about 1 year of operation) and the company issues a 1-year warranty on the refrigerator. What are the chances that a breakdown repair will be covered by the warranty?
12. The U of A runs two bus lines on campus: red and green. The red line serves north campus, and the green line serves south campus with a transfer station linking the two lines. Green buses arrive randomly (exponential interarrival time) at the transfer station every 10 minutes. Red buses also arrive randomly every 7 minutes.
 - (a) What is the probability distribution of the waiting time for a student arriving on the red line to get on the green line?
 - (b) What is the probability distribution of the waiting time for a student arriving on the green line to get on the red line?
13. Prove that the mean and standard deviation of the exponential distribution are equal.

15.4 PURE BIRTH AND DEATH MODELS (RELATIONSHIP BETWEEN THE EXPONENTIAL AND POISSON DISTRIBUTIONS)

This section presents two queuing situations: the **pure birth** model in which arrivals only are allowed, and the **pure death** model in which departures only can take place. An example of the pure birth model is the creation of birth certificates for newly born babies. The pure death model may be demonstrated by the random withdrawal of a stocked item in a store.

The exponential distribution is used to describe the interarrival time in the pure birth model and the interdeparture time in the pure death model. A by-product of the development of the two models is to show the close relationship between the exponential and the Poisson distributions, in the sense that one distribution automatically defines the other.

15.4.1 Pure Birth Model

Define

$$p_0(t) = \text{Probability of no arrivals during a period of time } t$$

Given that the interarrival time is exponential and that the arrival rate is λ customers per unit time, then

$$\begin{aligned} p_0(t) &= P\{\text{interarrival time} \geq t\} \\ &= 1 - P\{\text{interarrival time} \leq t\} \\ &= 1 - (1 - e^{-\lambda t}) \\ &= e^{-\lambda t} \end{aligned}$$

For a sufficiently small time interval $h > 0$, we have

$$p_0(h) = e^{-\lambda h} = 1 - \lambda h + \frac{(\lambda h)^2}{2!} - \dots = 1 - \lambda h + O(h^2)$$

The exponential distribution is based on the assumption that during $h > 0$, at most one event (arrival) can occur. Thus, as $h \rightarrow 0$,

$$p_1(h) = 1 - p_0(h) \approx \lambda h$$

This result shows that the probability of an arrival during h is directly proportional to h , with the arrival rate, λ , being the constant of proportionality.

To derive the distribution of the number of arrivals during a period t when the interarrival time is exponential with mean $\frac{1}{\lambda}$, define

$$p_n(t) = \text{Probability of } n \text{ arrivals during } t$$

For a sufficiently small $h > 0$,

$$\begin{aligned} p_n(t+h) &\approx p_n(t)(1 - \lambda h) + p_{n-1}(t)\lambda h, \quad n > 0 \\ p_0(t+h) &\approx p_0(t)(1 - \lambda h), \quad n = 0 \end{aligned}$$

In the first equation, n arrivals will be realized during $t+h$ if there are n arrivals during t and no arrivals during h , or $n-1$ arrivals during t and one arrival during h . All other combinations are not allowed because, according to the exponential distribution, at most one arrival can occur during a very small period h . The product law of probability is applicable to the right-hand side of the equation because arrivals are independent. For the second equation, zero arrivals during $t+h$ can occur only if no arrivals occur during t and h .

Rearranging the terms and taking the limits as $h \rightarrow 0$, we get

$$\begin{aligned} p'_n(t) &= \lim_{h \rightarrow 0} \frac{p_n(t+h) - p_n(t)}{h} = -\lambda p_n(t) + \lambda p_{n-1}(t), \quad n > 0 \\ p'_0(t) &= \lim_{h \rightarrow 0} \frac{p_0(t+h) - p_0(t)}{h} = -\lambda p_0(t), \quad n = 0 \end{aligned}$$

where $p'_n(t)$ is the first derivative of $p_n(t)$ with respect to t .

The solution of the preceding difference-differential equations yields

$$p_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, n = 0, 1, 2, \dots$$

This is a **Poisson distribution** with mean $E\{n|t\} = \lambda t$ arrivals during t .

The preceding result shows that if the time between arrivals is exponential with mean $\frac{1}{\lambda}$ then the number of arrivals during a specific period t is Poisson with mean λt . The converse is true also.

The following table summarizes the strong relationships between the exponential and the Poisson given an arrival rate of λ arrivals per unit time:

	Exponential	Poisson
Random variable	Time between successive arrivals, t	Number of arrivals, n , during a specified period T
Range	$t \geq 0$	$n = 0, 1, 2, \dots$
Density function	$f(t) = \lambda e^{-\lambda t}, t \geq 0$	$p_n(T) = \frac{(\lambda T)^n e^{-\lambda T}}{n!}, n = 0, 1, 2, \dots$
Mean value	$\frac{1}{\lambda}$ time units	λT arrivals during T
Cumulative probability	$P\{t \leq A\} = 1 - e^{-\lambda A}$	$p_{n \leq N}(T) = p_0(T) + p_1(T) + \dots + p_N(T)$
$P\{\text{no arrivals during period } A\}$	$P\{t > A\} = e^{-\lambda A}$	$p_0(A) = e^{-\lambda A}$

Example 15.4-1

Babies are born in a sparsely populated state at the rate of one birth every 12 minutes. The time between births follows an exponential distribution. Find the following:

- The average number of births per year.
- The probability that no births will occur in any one day.
- The probability of issuing 50 birth certificates in 3 hours given that 40 certificates were issued during the first 2 hours of the 3-hour period.

The birth rate per day is computed as

$$\lambda = \frac{24 \times 60}{12} = 120 \text{ births/day}$$

The number of births per year in the state is

$$\lambda t = 120 \times 365 = 43,800 \text{ births/year}$$

The probability of no births in any one day is computed from the Poisson distribution as

$$p_0(1) = \frac{(120 \times 1)^0 e^{-120 \times 1}}{0!} = e^{-120} = 0$$

Another way to compute the same probability is to note that no birth in any one day is equivalent to saying that the *time between successive births* exceeds one day. We can thus use the exponential distribution to compute the desired probability as

$$P\{t > 1\} = e^{-120} = 0$$

To compute the probability of issuing 50 certificates by the end of 3 hours given that 40 certificates were issued during the first 2 hours is equivalent to having 10 ($= 50 - 40$) births in one ($= 3 - 2$) hour because the distribution of the number of births is Poisson.

Given $\lambda = \frac{60}{12} = 5$ births per hour, we get

$$p_{10}(1) = \frac{(5 \times 1)^{10} e^{-5 \times 1}}{10!} = .01813$$

Excel Moment

The calculations associated with the Poisson distribution and, indeed, all queuing formulas are tedious and require special programming skill to secure reasonable computational accuracy. You can use Excel POISSON, POISSONDIST, and EXPONDIST functions to compute the individual and cumulative probabilities Poisson and exponential probabilities. These functions are also automated in *exceStatTables.xls*. For example, for a birth rate of 5 babies per hour, the probability of *exactly* 10 births in .5 hour is computed by entering 2.5 in F16, 10 in J16 to obtain the answer .000216 in M16. The cumulative probability of *at most* 10 births is given in O16 ($= .999938$). To determine the probability of the time between births being less than or equal to 18 minutes, use the exponential distribution by entering 2.5 in F9 and .3 in J9. The answer, .527633, is found in O9.

TORA/Excel Moment

You can also use TORA (file *toraEx15.4-1.txt*) or template *excelPoissonQ.xls* to determine all significant ($>10^{-5}$ in TORA and 10^{-7} in Excel) Poisson probabilities automatically. In both cases, the input data are the same. For the pure birth model of Example 15.4-1 the data are entered as follows:

Lambda	Mu	c	System limit	Source limit
5	0	(not applicable)	infinity	infinity

Note the entry under Lambda is $\lambda t = 5 \times 1 = 5$ births per day.

PROBLEM SET 15.4A

- *1. In Example 15.4-1, suppose that the clerk who enters the information from birth certificates into the computer normally waits until at least 5 certificates have accumulated. Find the probability that the clerk will be entering a new batch every hour.

2. An art collector travels to art auctions once a month on the average. Each trip is guaranteed to produce one purchase. The time between trips is exponentially distributed. Determine the following:
 - (a) The probability that no purchase is made in a 3-month period.
 - (b) The probability that no more than 8 purchases are made per year.
 - (c) The probability that the time between successive trips will exceed 1 month.
3. In a bank operation, the arrival rate is 2 customers per minute. Determine the following:
 - (a) The average number of arrivals during 5 minutes.
 - (b) The probability that no arrivals will occur during the next .5 minute.
 - (c) The probability that at least one arrival will occur during the next .5 minute.
 - (d) The probability that the time between two successive arrivals is at least 3 minutes.
4. The time between arrivals at L&J restaurant is exponential with mean 5 minutes. The restaurant opens for business at 11:00 A.M. Determine the following:
 - * (a) The probability of having 10 arrivals in the restaurant by 11:12 A.M. given that 8 customers arrived by 11:05 A.M.
 - (b) The probability that a new customer will arrive between 11:28 and 11:33 A.M. given that the last customer arrived at 11:25 A.M.
5. The Springdale Public Library receives new books according to a Poisson distribution with mean 25 books per day. Each shelf in the stacks holds 100 books. Determine the following:
 - (a) The average number of shelves that will be stacked with new books each (30-day) month.
 - (b) The probability that more than 10 bookcases will be needed each month, given that a bookcase has 5 shelves.
6. The U of A runs two bus lines on campus: red and green. The red line serves north campus and the green line serves south campus with a transfer station linking the two lines. Green buses arrive randomly (according to a Poisson distribution) at the transfer station every 10 minutes. Red buses also arrive randomly every 7 minutes.
 - * (a) What is the probability that two buses will stop at the station during a 5-minute interval?
 - (b) A student whose dormitory is located next to the station has a class in 10 minutes. Either bus will take the student to the classroom building. The ride takes 5 minutes, after which the student will walk for about 3 minutes to reach the classroom. What is the probability that the student will make it to class on time?
7. Prove that the mean and variance of the Poisson distribution during an interval t equal λt , where λ is the arrival rate.
8. Derive the Poisson distribution from the difference-differential equations of the pure birth model. *Hint:* The solution of the general differential equation

$$y' + a(t)y = b(t)$$

is

$$y = e^{-\int a(t) dt} \left\{ \int b(t) e^{\int a(t) dt} dt + \text{constant} \right\}$$

15.4.2 Pure Death Model

In the pure death model, the system starts with N customers at time 0 and no new arrivals are allowed. Departures occur at the rate μ customers per unit time. To develop

the difference-differential equations for the probability $p_n(t)$ of n customers remaining after t time units, we follow the arguments used with the pure birth model (Section 15.4.1). Thus,

$$p_N(t+h) = p_N(t)(1-\mu h)$$

$$p_n(t+h) = p_n(t)(1-\mu h) + p_{n+1}(t)\mu h, 0 < n < N$$

$$p_0(t+h) = p_0(t)(1) + p_1(t)\mu h$$

As $h \rightarrow 0$, we get

$$p'_N(t) = -\mu p_N(t)$$

$$p'_n(t) = -\mu p_n(t) + \mu p_{n+1}(t), 0 < n < N$$

$$p'_0(t) = \mu p_1(t)$$

The solution of these equations yields the following **truncated Poisson** distribution:

$$p_n(t) = \frac{(\mu t)^{N-n} e^{-\mu t}}{(N-n)!}, n = 1, 2, \dots, N$$

$$p_0(t) = 1 - \sum_{n=1}^N p_n(t)$$

Example 15.4-2

The florist section in a grocery store stocks 18 dozen roses at the beginning of each week. On the average, the florist sells 3 dozens a day (one dozen at a time), but the actual demand follows a Poisson distribution. Whenever the stock level reaches 5 dozens, a new order of 18 new dozens is placed for delivery at the beginning of the following week. Because of the nature of the item, all roses left at the end of the week are disposed of. Determine the following:

- The probability of placing an order in any one day of the week.
- The average number of dozen roses that will be discarded at the end of the week.

Because purchases occur at the rate of $\mu = 3$ dozens per day, the probability of placing an order by the end of day t is given as

$$\begin{aligned} p_{n \leq 5}(t) &= p_0(t) + p_1(t) + \dots + p_5(t) \\ &= p_0(t) + \sum_{n=1}^5 \frac{(3t)^{18-n} e^{-3t}}{(18-n)!}, t = 1, 2, \dots, 7 \end{aligned}$$

The calculations of $p_{n \leq 5}(t)$ are best done using excelPoissonQ.xls or TORA. TORA's multiple scenarios may be more convenient in this case. The associated input data for the pure death model corresponding to $t = 1, 2, \dots$, and 7 are

$$\text{Lambda} = 0, \text{Mu} = 3t, c = 1, \text{System Limit} = 18, \text{and Source Limit} = 18$$

Note that t must be substituted out numerically as shown in file toraEx15.4-2.txt.

The output is summarized as follows:

t (days)	1	2	3	4	5	6	7
μt	3	6	9	12	15	18	21
$p_{n \leq 5}(t)$.0000	.0088	.1242	.4240	.7324	.9083	.9755

The average number of dozen roses discarded at the end of the week ($t = 7$) is $E\{n|t = 7\}$. To calculate this value we need $p_n(7)$, $n = 0, 1, 2, \dots, 18$, which can be determined using provided software, which yields

$$E\{n|t = 7\} = \sum_{n=0}^{18} n p_n(7) = .664 \approx 1 \text{ dozen}$$

PROBLEM SET 15.4B

- In Example 15.4-2, use excelPoissonQ.xls or TORA to compute $p_n(7)$, $n = 1, 2, \dots, 18$, and then verify manually that these probabilities yield $E\{n|t = 7\} = .664$ dozen.
- Consider Example 15.4-2. In each of the following cases, first write the answer algebraically, and then use excelPoissonQ.xls or TORA to provide numerical answers.
 - The probability that the stock is depleted after 3 days.
 - The average number of dozen roses left at the end of the second day.
 - The probability that at least one dozen is purchased by the end of the fourth day, given that the last dozen was bought at the end of the third day.
 - The probability that the time remaining until the next purchase is at most half a day given that the last purchase occurred a day earlier.
 - The probability that no purchases will occur during the first day.
 - The probability that no order will be placed by the end of the week.
- The Springdale High School band is performing a benefit jazz concert in its new 400-seat auditorium. Local businesses buy the tickets in blocks of 10 and donate them to youth organizations. Tickets go on sale to business entities for 4 hours only the day before the concert. The process of placing orders for tickets is Poisson with a mean 10 calls per hour. Any (blocks of) tickets remaining after the box office is closed are sold at a discount as "rush tickets" 1 hour before the concert starts. Determine
 - The probability that it will be possible to buy rush tickets.
 - The average number of rush tickets available.
- Each morning, the refrigerator in a small machine shop is stocked with two cases (24 cans per case) of soft drinks for use by the shop's 10 employees. The employees can quench their thirst at any time during the 8-hour work day (8:00 A.M. to 4:00 P.M.), and each employee is known to consume approximately 4 cans a day, but the process is totally random (Poisson distribution). What is the probability that an employee will not find a drink at noon (the start of the lunch period)? Just before the shop closes?
- A freshman student receives a bank deposit of \$100 a month from home to cover incidentals. Withdrawal checks of \$20 each occur randomly during the month and are spaced according to an exponential distribution with a mean value of 1 week. Determine the probability that the student will run out of incidental money before the end of the fourth week.

6. Inventory is withdrawn from a stock of 80 items according to a Poisson distribution at the rate of 5 items per day. Determine the following:
 - (a) The probability that 10 items are withdrawn during the first 2 days.
 - (b) The probability that no items are left at the end of 4 days.
 - (c) The average number of items withdrawn over a 4-day period.
7. A machine shop has just stocked 10 spare parts for the repair of a machine. Stock replenishment that brings the stock level back to 10 pieces occurs every 7 days. The time between breakdowns is exponential with mean 1 day. Determine the probability that the machine will remain broken for 2 days because no spare parts are available.
8. Demand for an item occurs according to a Poisson distribution with mean 3 per day. The maximum stock level is 25 items, which occurs on each Monday immediately after a new order is received. The order size depends on the number of units left at the end of the week on Saturday (business is closed on Sundays). Determine the following:
 - *(a) The average weekly size of the order.
 - *(b) The probability of incurring shortage when the business opens on Friday morning.
 - (c) The probability that the weekly order size exceeds 10 units.
9. Prove that the distribution of the time between departures corresponding to the truncated Poisson in the pure death model is an exponential distribution with mean $\frac{1}{\mu}$ time units.
10. Derive the truncated Poisson distribution from the difference-differential equations of the pure death model using induction. [Note: See the hint in Problem 8, Set 15.4a.]

15.5 GENERALIZED POISSON QUEUING MODEL

This section develops a general queuing model that combines both arrivals and departures based on the Poisson assumptions—that is, the interarrival and the service times follow the exponential distribution. The model is the basis for the derivation of the specialized Poisson models in Section 15.6.

The development of the generalized model is based on the long-run or **steady-state** behavior of the queuing situation, which is achieved after the system has been in operation for a sufficiently long time. This type of analysis contrasts with the **transient** (or warm-up) behavior that prevails during the early operation of the system. One reason for not discussing the transient behavior in this chapter is its analytical complexity. Another reason is that the study of most queuing situations occurs under steady-state conditions.

The generalized model assumes that both the arrival and departure rates are **state dependent**, meaning that they depend on the number of customers in the service facility. For example, at a highway toll booth, attendants tend to speed up toll collection during rush hours. Another example occurs in a shop with a given number of machines where the rate of breakdown decreases as the number of broken machines increases (because only working machines are capable of generating new breakdowns).

Define

n = Number of customers in the system (in-queue plus in-service)

λ_n = Arrival rate given n customers in the system

μ_n = Departure rate given n customers in the system

p_n = Steady-state probability of n customers in the system

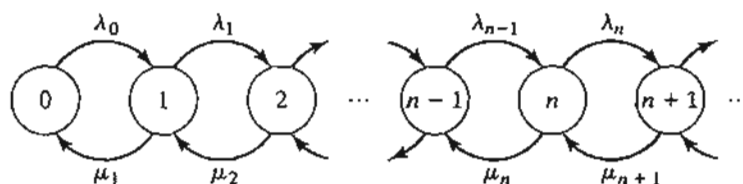


FIGURE 15.2
Poisson queues transition diagram

The generalized model derives p_n as a function of λ_n and μ_n . These probabilities are then used to determine the system's measures of performance, such as the average queue length, the average waiting time, and the average utilization of the facility.

The probabilities p_n are determined by using the **transition-rate diagram** in Figure 15.2. The queuing system is in state n when the number of customers in the system is n . As explained in Section 15.3, the probability of more than one event occurring during a small interval h tends to zero as $h \rightarrow 0$. This means that for $n > 0$, state n can change only to two possible states: $n - 1$ when a departure occurs at the rate μ_n , and $n + 1$ when an arrival occurs at the rate λ_n . State 0 can only change to state 1 when an arrival occurs at the rate λ_0 . Notice that μ_0 is undefined because no departures can occur if the system is empty.

Under steady-state conditions, for $n > 0$, the *expected* rates of flow into and out of state n must be equal. Based on the fact that state n can be changed to states $n - 1$ and $n + 1$ only, we get

$$\left(\begin{array}{l} \text{Expected rate of} \\ \text{flow into state } n \end{array} \right) = \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1}$$

Similarly,

$$\left(\begin{array}{l} \text{Expected rate of} \\ \text{flow out of state } n \end{array} \right) = (\lambda_n + \mu_n)p_n$$

Equating the two rates, we get the following **balance equation**:

$$\lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1} = (\lambda_n + \mu_n)p_n, n = 1, 2, \dots$$

From Figure 15.2, the balance equation associated with $n = 0$, is

$$\lambda_0 p_0 = \mu_1 p_1$$

The balance equations are solved recursively in terms of p_0 as follows: For $n = 0$, we have

$$p_1 = \left(\frac{\lambda_0}{\mu_1} \right) p_0$$

Next, for $n = 1$, we have

$$\lambda_0 p_0 + \mu_2 p_2 = (\lambda_1 + \mu_1) p_1$$

Substituting $p_1 = \left(\frac{\lambda_0}{\mu_0}\right)p_0$ and simplifying, we get (verify!)

$$p_2 = \left(\frac{\lambda_1 \lambda_0}{\mu_2 \mu_1}\right)p_0$$

In general, we can show by induction that

$$p_n = \left(\frac{\lambda_{n-1} \lambda_{n-2} \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1}\right)p_0, \quad n = 1, 2, \dots$$

The value of p_0 is determined from the equation $\sum_{n=0}^{\infty} p_n = 1$

Example 15.5-1

B&K Groceries operates with three check-out counters. The manager uses the following schedule to determine the number of counters in operation, depending on the number of customers in store:

No. of customers in store	No. of counters in operation
1 to 3	1
4 to 6	2
More than 6	3

Customers arrive in the counters area according to a Poisson distribution with a mean rate of 10 customers per hour. The average check-out time per customer is exponential with mean 12 minutes. Determine the steady-state probability p_n of n customers in the check-out area.

From the information of the problem, we have

$$\begin{aligned} \lambda_n &= \lambda = 10 \text{ customers per hour,} & n &= 0, 1, \dots \\ \mu_n &= \begin{cases} \frac{60}{12} = 5 \text{ customers per hour,} & n = 0, 1, 2, 3 \\ 2 \times 5 = 10 \text{ customers per hour,} & n = 4, 5, 6 \\ 3 \times 5 = 15 \text{ customers per hour,} & n = 7, 8, \dots \end{cases} \end{aligned}$$

Thus,

$$\begin{aligned} p_1 &= \left(\frac{10}{5}\right)p_0 = 2p_0 \\ p_2 &= \left(\frac{10}{5}\right)^2 p_0 = 4p_0 \\ p_3 &= \left(\frac{10}{5}\right)^3 p_0 = 8p_0 \\ p_4 &= \left(\frac{10}{5}\right)^3 \left(\frac{10}{10}\right)p_0 = 8p_0 \\ p_5 &= \left(\frac{10}{5}\right)^3 \left(\frac{10}{10}\right)^2 p_0 = 8p_0 \\ p_6 &= \left(\frac{10}{5}\right)^3 \left(\frac{10}{10}\right)^3 p_0 = 8p_0 \\ p_{n \geq 7} &= \left(\frac{10}{5}\right)^3 \left(\frac{10}{10}\right)^3 \left(\frac{10}{15}\right)^{n-6} p_0 = 8\left(\frac{2}{3}\right)^{n-6} p_0 \end{aligned}$$

The value of p_0 is determined from the equation

$$p_0 + p_0\{2 + 4 + 8 + 8 + 8 + 8 + 8\left(\frac{2}{3}\right) + 8\left(\frac{2}{3}\right)^2 + 8\left(\frac{2}{3}\right)^3 + \dots\} = 1$$

or, equivalently

$$p_0\{31 + 8\left(1 + \left(\frac{2}{3}\right) + \left(\frac{2}{3}\right)^2 + \dots\right)\} = 1$$

Using the geometric sum series

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1-x}, |x| < 1$$

we get

$$p_0\left\{31 + 8\left(\frac{1}{1-\frac{2}{3}}\right)\right\} = 1$$

Thus, $p_0 = \frac{1}{55}$.

Given p_0 , we can now determine p_n for $n > 0$. For example, the probability that only one counter will be open is computed as the probability that there are at most three customers in the system:

$$p_1 + p_2 + p_3 = (2 + 4 + 8)\left(\frac{1}{55}\right) \approx .255$$

We can use p_n to determine measures of performance for the B&K situation. For example,

$$\begin{aligned} \left(\begin{array}{l} \text{Expected number} \\ \text{of idle counters} \end{array} \right) &= 3p_0 + 2(p_1 + p_2 + p_3) + 1(p_4 + p_5 + p_6) \\ &\quad + 0(p_7 + p_8 + \dots) \\ &= 1 \text{ counter} \end{aligned}$$

PROBLEM SET 15.5A

- In Example 15.5-1, determine the following:
 - The probability distribution of the number of open counters.
 - The average number of busy counters.
- In the B&K model of Example 15.5-1, suppose that the interarrival time at the check-out area is exponential with mean 5 minutes and that the checkout time per customer is also exponential with mean 10 minutes. Suppose further that B&K will add a fourth counter and that counters will open based on increments of two customers. Determine the following:
 - The steady-state probabilities, p_n for all n .
 - The probability that a fourth counter will be needed.
 - The average number of idle counters.
- *In the B&K model of Example 15.5-1, suppose that all three counters are always open and that the operation is set up such that the customer will go to the first empty counter. Determine the following:
 - The probability that all three counters will be in use.
 - The probability that an arriving customer will not wait.

4. First Bank of Springdale operates a one-lane drive-in ATM machine. Cars arrive according to a Poisson distribution at the rate of 12 cars per hour. The time per car needed to complete the ATM transaction is exponential with mean 6 minutes. The lane can accommodate a total of 10 cars. Once the lane is full, other arriving cars seek service in another branch. Determine the following:
 - (a) The probability that an arriving car will not be able to use the ATM machine because the lane is full.
 - (b) The probability that a car will not be able to use the ATM machine immediately on arrival.
 - (c) The average number of cars in the lane.
5. Have you ever heard someone repeat the contradictory statement, "The place is so crowded no one goes there any more"? This statement can be interpreted as saying that the opportunity for balking increases with the increase in the number of customers seeking service. A possible platform for modeling this situation is to say that the arrival rate at the system decreases as the number of customers in the system increases. More specifically, we consider the simplified case of M&M Pool Club, where customers usually arrive in pairs to "shoot pool." The normal arrival rate is 6 pairs (of people) per hour. However, once the number of pairs in the pool hall exceeds 8, the arrival rate drops to 5 pairs per hour. The arrival process is assumed to follow the Poisson distribution. Each pair shoots pool for an exponential time with mean 30 minutes. The pool hall has a total of 5 tables and can accommodate no more than 12 pairs at any one time. Determine the following:
 - (a) The probability that customers will balk.
 - (b) The probability that all tables are in use.
 - (c) The average number of tables in use.
 - (d) The average number of pairs waiting for a pool table to be available.
- *6. A barbershop serves one customer at a time and provides three seats for waiting customers. If the place is full, customers go elsewhere. Arrivals occur according to a Poisson distribution with mean four per hour. The time to get a haircut is exponential with mean 15 minutes. Determine the following:
 - (a) The steady-state probabilities.
 - (b) The expected number of customers in the shop.
 - (c) The probability that customers will go elsewhere because the shop is full.
7. Consider a one-server queuing situation in which the arrival and service rates are given by

$$\lambda_n = 10 - n, n = 0, 1, 2, 3$$

$$\mu_n = \frac{n}{2} + 5, n = 1, 2, 3, 4$$

This situation is equivalent to reducing the arrival rate and increasing the service rate as the number in the system, n , increases.

- (a) Set up the transition diagram and determine the balance equation for the system.
 - (b) Determine the steady-state probabilities.
8. Consider the single queue model where only one customer is allowed in the system. Customers who arrive and find the facility busy never return. Assume that the arrivals distribution is Poisson with mean λ per unit time and that the service time is exponential with mean $\frac{1}{\mu}$ time units.
 - (a) Set up the transition diagram and determine the balance equations.
 - (b) Determine the steady-state probabilities.
 - (c) Determine the average number in the system.

9. The induction proof for deriving the general solution of the generalized model is applied as follows. Consider

$$p_k = \prod_{i=0}^{k-1} \left(\frac{\lambda_i}{\mu_{i+1}} \right) p_0, k = 0, 1, 2, \dots$$

We substitute for p_{n-1} and p_{n-2} in the general difference equation involving p_n , p_{n-1} , and p_{n-2} to derive the desired expression for p_n . Verify this procedure.

15.6 SPECIALIZED POISSON QUEUES

Figure 15.3 depicts the specialized Poisson queuing situation with c parallel servers. A waiting customer is selected from the queue to start service with the first available server. The arrival rate at the system is λ customers per unit time. All parallel servers are identical, meaning that the service rate for any server is μ customers per unit time. The number of customers in the system is defined to include those *in service* and those *in queue*.

A convenient notation for summarizing the characteristics of the queuing situation in Figure 15.3 is given by the following format:

$$(a/b/c):(d/e/f)$$

where

a = Arrivals distribution

b = Departures (service time) distribution

c = Number of parallel servers ($= 1, 2, \dots, \infty$)

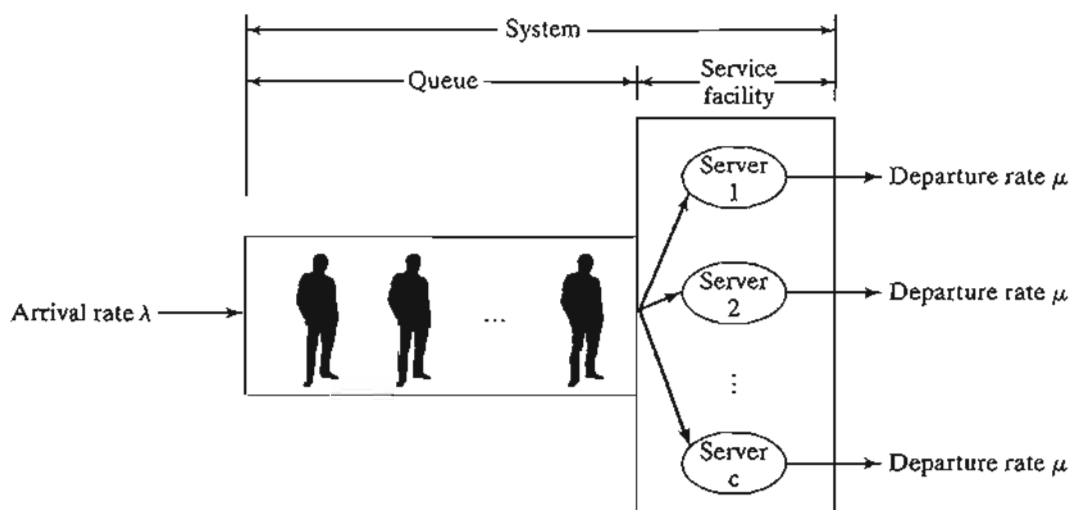
d = Queue discipline

e = Maximum number (finite or infinite) allowed in the system
(in-queue plus in-service)

f = Size of the calling source (finite or infinite)

FIGURE 15.3

Schematic representation of a queuing system with c parallel servers



The standard notation for representing the arrivals and departures distributions (symbols a and b) is

M = Markovian (or Poisson) arrivals or departures distribution
(or equivalently exponential interarrival or service time distribution)

D = Constant (deterministic) time

E_k = Erlang or gamma distribution of time (or, equivalently, the sum of independent exponential distributions)

GI = General (generic) distribution of interarrival time

G = General (generic) distribution of service time

The queue discipline notation (symbol d) includes

$FCFS$ = First come, first served

$LCFS$ = Last come, first served

$SIRO$ = Service in random order

GD = General discipline (i.e., any type of discipline)

To illustrate the use of the notation, the model $(M/D/10):(GD/20/\infty)$ uses Poisson arrivals (or exponential interarrival time), constant service time, and 10 parallel servers. The queue discipline is GD , and there is a limit of 20 customers on the entire system. The size of the source from which customers arrive is infinite.

As a historical note, the first three elements of the notation $(a/b/c)$, were devised by D.G. Kendall in 1953 and are known in the literature as the **Kendall notation**. In 1966, A.M. Lee added the symbols d and e to the notation. This author added the last element, symbol f , in 1968.

Before presenting the details of the specialized Poisson queues, we show how the steady-state measures of performance of the generalized queuing situation can be derived from the steady-state probabilities p_n given in Section 15.5.

15.6.1 Steady-State Measures of Performance

The most commonly used measures of performance in a queuing situation are

L_s = Expected number of customers in system

L_q = Expected number of customers in queue

W_s = Expected waiting time in system

W_q = Expected waiting time in queue

\bar{c} = Expected number of busy servers

Recall that the *system* includes both the *queue* and the *service facility*.

We show now how these measures are derived (directly or indirectly) from the steady-state probability of n in the system, p_n as

$$L_s = \sum_{n=1}^{\infty} n p_n$$

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

The relationship between L_s and W_s (also L_q and W_q) is known as **Little's formula**, and is given as

$$L_s = \lambda_{\text{eff}} W_s$$

$$L_q = \lambda_{\text{eff}} W_q$$

These relationships are valid under rather general conditions. The parameter λ_{eff} is the *effective* arrival rate at the system. It equals the (nominal) arrival rate λ when all arriving customers can join the system. Otherwise, if some customers cannot join because the system is full (e.g., a parking lot), then $\lambda_{\text{eff}} < \lambda$. We will show later how λ_{eff} is determined.

A direct relationship also exists between W_s and W_q . By definition,

$$\left(\begin{array}{c} \text{Expected waiting} \\ \text{time in system} \end{array} \right) = \left(\begin{array}{c} \text{Expected waiting} \\ \text{time in queue} \end{array} \right) + \left(\begin{array}{c} \text{Expected service} \\ \text{time} \end{array} \right)$$

This translates to

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

Next, we can relate L_s to L_q by multiplying both sides of the last formula by λ_{eff} , which together with Little's formula gives

$$L_s = L_q + \frac{\lambda_{\text{eff}}}{\mu}$$

By definition, the difference between the average number in the system, L_s , and the average number in the queue, L_q , must equal the average number of *busy* servers, \bar{c} . We thus have,

$$\bar{c} = L_s - L_q = \frac{\lambda_{\text{eff}}}{\mu}$$

It follows that

$$\left(\begin{array}{c} \text{Facility} \\ \text{utilization} \end{array} \right) = \frac{\bar{c}}{c}$$

Example 15.6-1

Visitors' parking at Ozark College is limited to five spaces only. Cars making use of this space arrive according to a Poisson distribution at the rate of six cars per hour. Parking time is exponentially distributed with a mean of 30 minutes. Visitors who cannot find an empty space on arrival

may temporarily wait inside the lot until a parked car leaves. That temporary space can hold only three cars. Other cars that cannot park or find a temporary waiting space must go elsewhere. Determine the following:

- The probability, p_n , of n cars in the system.
- The effective arrival rate for cars that actually use the lot.
- The average number of cars in the lot.
- The average time a car waits for a parking space inside the lot.
- The average number of *occupied* parking spaces.
- The average utilization of the parking lot.

We note first that a parking space acts as a server, so that the system has a total of $c = 5$ parallel servers. Also, the maximum capacity of the system is $5 + 3 = 8$ cars.

The probability p_n can be determined as a special case of the generalized model in Section 15.5 using

$$\lambda_n = 6 \text{ cars/hour}, n = 0, 1, 2, \dots, 8$$

$$\mu_n = \begin{cases} n\left(\frac{60}{30}\right) = 2n \text{ cars/hour}, & n = 1, 2, 3, 4, 5 \\ 5\left(\frac{60}{30}\right) = 10 \text{ cars/hour}, & n = 6, 7, 8 \end{cases}$$

From Section 15.5, we get

$$p_n = \begin{cases} \frac{3^n}{n!} p_0, & n = 1, 2, 3, 4, 5 \\ \frac{3^n}{5! 5^{n-5}} p_0, & n = 6, 7, 8 \end{cases}$$

The value of p_0 is computed by substituting p_n , $n = 1, 2, \dots, 8$, in the following equation

$$p_0 + p_1 + \dots + p_8 = 1$$

or

$$p_0 + p_0 \left(\frac{3}{1!} + \frac{3^2}{2!} + \frac{3^3}{3!} + \frac{3^4}{4!} + \frac{3^5}{5!} + \frac{3^6}{5! 5} + \frac{3^7}{5! 5^2} + \frac{3^8}{5! 5^3} \right) = 1$$

This yields $p_0 = .04812$ (verify!). From p_0 , we can now compute p_1 through p_8 as

n	1	2	3	4	5	6	7	8
p_n	.14436	.21654	.21654	.16240	.09744	.05847	.03508	.02105

The effective arrival rate λ_{eff} can be computed by observing the schematic diagram in Figure 15.4, where customers arrive from the source at the rate λ cars per hour. An arriving car may enter the parking lot or go elsewhere with rates λ_{eff} or λ_{lost} , which means that $\lambda = \lambda_{\text{eff}} + \lambda_{\text{lost}}$. A car will not be able to enter the parking lot if 8 cars are already in. This means that the proportion of cars that will not be able to enter the lot is p_8 . Thus,

$$\lambda_{\text{lost}} = \lambda p_8 = 6 \times .02105 = .1263 \text{ cars per hour}$$

$$\lambda_{\text{eff}} = \lambda - \lambda_{\text{lost}} = 6 - .1263 = 5.8737 \text{ cars per hour}$$

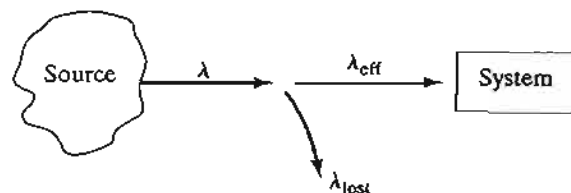


FIGURE 15.4

Relationship between λ , λ_{eff} , and λ_{lost}

The average number of cars in the lot (those waiting for or occupying a space) equals L_s , the average number in the system. We can compute L_s from p_n as

$$L_s = 0p_0 + 1p_1 + \dots + 8p_8 = 3.1286 \text{ cars}$$

A car waiting in the temporary space is actually a car in queue. Thus, its waiting time until a space is found is W_q . To determine W_q we use

$$W_q = W_s - \frac{1}{\mu}$$

Thus,

$$W_s = \frac{L_s}{\lambda_{\text{eff}}} = \frac{3.1286}{5.8737} = .53265 \text{ hour}$$

$$W_q = .53265 - \frac{1}{2} = .03265 \text{ hour}$$

The average number of occupied parking spaces is the same as the average number of busy servers,

$$\bar{c} = L_s - L_q = \frac{\lambda_{\text{eff}}}{\mu} = \frac{5.8737}{2} = 2.9368 \text{ spaces}$$

From \bar{c} , we get

$$\text{Parking lot utilization} = \frac{\bar{c}}{c} = \frac{2.9368}{5} = .58736$$

PROBLEM SET 15.6A

1. In Example 15.6-1, do the following:

*(a) Compute L_q directly using the formula $\sum_{n=c+1}^{\infty} (n - c)p_n$.

(b) Compute W_s from L_q .

*(c) Compute the average number of cars that will not be able to enter the parking lot during an 8-hour period.

*(d) Show that $c - (L_s - L_q)$, the average number of empty spaces, equals

$$\sum_{n=0}^{c-1} (c - n)p_n$$

2. Solve Example 15.6-1 using the following data: number of parking spaces = 6, number of temporary spaces = 4, $\lambda = 10$ cars per hour, and average parking time = 45 minutes.

15.6.2 Single-Server Models

This section presents two models for the single server case ($c = 1$). The first model sets no limit on the maximum number in the system, and the second model assumes a finite system limit. Both models assume an infinite-capacity source. Arrivals occur at the rate λ customers per unit time and the service rate is μ customers per unit time.

The results of the two models (and indeed of all the remaining models in Section 15.6) are derived as special cases of the results of the generalized model of Section 15.5.

The Kendall notation will be used to summarize the characteristics of each situation. Because the derivations of p_n in Section 15.5 and of all the measures of performance in Section 15.6.1 are totally independent of a specific queue discipline, the symbol GD (general discipline) will be used with the notation.

$(M/M/1):(GD/\infty/\infty)$. Using the notation of the generalized model, we have

$$\left. \begin{aligned} \lambda_n &= \lambda \\ \mu_n &= \mu \end{aligned} \right\}, n = 0, 1, 2, \dots$$

Also, $\lambda_{\text{eff}} = \lambda$ and $\lambda_{\text{lost}} = 0$, because all arriving customers can join the system.

Letting $\rho = \frac{\lambda}{\mu}$, the expression for p_n in the generalized model then reduces to

$$p_n = \rho^n p_0, n = 0, 1, 2, \dots$$

To determine the value of p_0 , we use the identity

$$p_0(1 + \rho + \rho^2 + \dots) = 1$$

Assuming $\rho < 1$, the geometric series will have the finite sum $(\frac{1}{1-\rho})$, thus

$$p_0 = 1 - \rho, \text{ provided } \rho < 1.$$

The general formula for p_n is thus given by the following geometric distribution

$$p_n = (1 - \rho)\rho^n, n = 1, 2, \dots (\rho < 1)$$

The mathematical derivation of p_n imposes the condition $\rho < 1$, or $\lambda < \mu$. If $\lambda \geq \mu$, the geometric series will not converge, and the steady-state probabilities p_n will not exist. This result makes intuitive sense, because unless the service rate is larger than the arrival rate, queue length will continually increase and no steady state can be reached.

The measure of performance L_q can be derived in the following manner:

$$\begin{aligned} L_s &= \sum_{n=0}^{\infty} n p_n = \sum_{n=0}^{\infty} n (1 - \rho) \rho^n \\ &= (1 - \rho) \rho \frac{d}{d\rho} \sum_{n=0}^{\infty} \rho^n \\ &= (1 - \rho) \rho \frac{d}{d\rho} \left(\frac{1}{1 - \rho} \right) = \frac{\rho}{1 - \rho} \end{aligned}$$

Because $\lambda_{\text{eff}} = \lambda$ for the present situation, the remaining measures of performance are computed using the relationships in Section 15.6.1. Thus,

$$W_s = \frac{L_s}{\lambda} = \frac{1}{\mu(1 - \rho)} = \frac{1}{\mu - \lambda}$$

$$W_q = W_s - \frac{1}{\mu} = \frac{\rho}{\mu(1 - \rho)}$$

$$L_q = \lambda W_q = \frac{\rho^2}{1 - \rho}$$

$$\bar{c} = L_s - L_q = \rho$$

Example 15.6-2

Automata car wash facility operates with only one bay. Cars arrive according to a Poisson distribution with a mean of 4 cars per hour, and may wait in the facility's parking lot if the bay is busy. The time for washing and cleaning a car is exponential, with a mean of 10 minutes. Cars that cannot park in the lot can wait in the street bordering the wash facility. This means that, for all practical purposes, there is no limit on the size of the system. The manager of the facility wants to determine the size of the parking lot.

For this situation, we have $\lambda = 4$ cars per hour, and $\mu = \frac{60}{10} = 6$ cars per hour. Because $\rho = \frac{\lambda}{\mu} < 1$, the system can operate under steady-state conditions.

The TORA or excelPoissonQ.xls input for this model is

Lambda	Mu	c	System limit	Source limit
4	6	1	infinity	infinity

The output of the model is shown in Figure 15.5. The average number of cars waiting in the queue, L_q , is 1.33 cars.

Generally, using L_q as the sole basis for the determination of the number of parking spaces is not advisable, because the design should, in some sense, account for the maximum possible length of the queue. For example, it may be more plausible to design the parking lot such that an arriving car will find a parking space at least 90% of the time. To do this, let S represent the number of parking spaces. Having S parking spaces is equivalent to having $S + 1$ spaces in the system (queue plus wash bay). An arriving car will find a space 90% of the time if there are *at most* S cars in the system. This condition is equivalent to the following probability statement:

$$p_0 + p_1 + \dots + p_S \geq .9$$

From Figure 15.5, cumulative p_n for $n = 5$ is .91221. This means that the condition is satisfied for $S \geq 5$ parking spaces.

The number of spaces S can be determined also by using the mathematical definition of p_n —that is,

$$(1 - \rho)(1 + \rho + \rho^2 + \dots + \rho^S) \geq .9$$

Scenario1: (M/M/1):(GD/infinity/infinity)

Lambda = 4.00000

Mu = 6.00000

Lambda eff = 4.00000

Rho/c = 0.66667

Ls = 2.00000

Lq = 1.33333

Ws = 0.50000

Wq = 0.33333

n	Probability pn	Cumulative Pn	n	Probability pn	Cumulative Pn
0	0.33333	0.33333	13	0.00171	0.99657
1	0.22222	0.55556	14	0.00114	0.99772
2	0.14815	0.70370	15	0.00076	0.99848
3	0.09877	0.80247	16	0.00051	0.99899
4	0.06584	0.86831	17	0.00034	0.99932
5	0.04390	0.91221	18	0.00023	0.99955
6	0.02926	0.94147	19	0.00015	0.99970
7	0.01951	0.96098	20	0.00010	0.99980
8	0.01301	0.97399	21	0.00007	0.99987
9	0.00867	0.98266	22	0.00004	0.99991
10	0.00578	0.98844	23	0.00003	0.99994
11	0.00385	0.99229	24	0.00002	0.99996
12	0.00257	0.99486	25	0.00001	0.99997

FIGURE 15.5

TORA output of Example 15.6-2 (file toraEx15.6-2.txt)

The sum of the truncated geometric series equals $\frac{1-\rho^{S+1}}{1-\rho}$. Thus the condition reduces to

$$(1 - \rho^{S+1}) \geq .9$$

Simplification of the inequality yields

$$\rho^{S+1} \leq .1$$

Taking the logarithms on both sides (and noting that $\log(x) < 0$ for $0 < x < 1$, which reverses the direction of the inequality), we get

$$S \geq \frac{\ln(.1)}{\ln(\rho)} - 1 = 4.679 \approx 5$$

PROBLEM SET 15.6B

1. In Example 15.6-2, do the following.

- Determine the percent utilization of the wash bay.
- Determine the probability that an arriving car must wait in the parking lot prior to entering the wash bay.
- If there are seven parking spaces, determine the probability that an arriving car will find an empty parking space.
- How many parking spaces should be provided so that an arriving car may find a parking space 99% of the time?

- *2. John Macko is a student at Ozark U. He does odd jobs to supplement his income. Job requests come every 5 days on the average, but the time between requests is exponential. The time for completing a job is also exponential with mean 4 days.
- What is the probability that John will be out of jobs?
 - If John gets about \$50 a job, what is his average monthly income?
 - If at the end of the semester, John decides to subcontract on the outstanding jobs at \$40 each. How much, on the average, should he expect to pay?
3. Over the years, Detective Columbo, of the Fayetteville Police Department, has had phenomenal success in solving every single crime case. It is only a matter of time before any case is solved. Columbo admits that the time per case is "totally random," but, on the average, each investigation will take about a week and half. Crimes in peaceful Fayetteville are not very common. They occur randomly at the rate of one crime per (four-week) month. Detective Columbo is asking for an assistant to share the heavy work load. Analyze Columbo's claim, particularly from the standpoint of the following points:
- The average number of cases awaiting investigation.
 - The percentage of time the detective remains busy.
 - The average time needed to solve a case.
4. Cars arrive at the Lincoln Tunnel toll gate according to a Poisson distribution, with a mean of 90 cars per hour. The time for passing the gate is exponential with mean 38 seconds. Drivers complain of the long waiting time, and authorities are willing to reduce the average passing time to 30 seconds by installing automatic toll collecting devices, provided two conditions are satisfied: (1) the average number of waiting cars in the present system exceeds 5, and (2) the percentage of the gate idle time with the new device installed does not exceed 10%. Can the new device be justified?
- *5. A fast-food restaurant has one drive-in window. Cars arrive according to a Poisson distribution at the rate of 2 cars every 5 minutes. The space in front of the window can accommodate at most 10 cars, including the one being served. Other cars can wait outside this space if necessary. The service time per customer is exponential, with a mean of 1.5 minutes. Determine the following:
- The probability that the facility is idle.
 - The expected number of customers waiting to be served.
 - The expected waiting time until a customer reaches the window to place an order.
 - The probability that the waiting line will exceed the 10-space capacity.
6. Customers arrive at a one-window drive-in bank according to a Poisson distribution, with a mean of 10 per hour. The service time per customer is exponential, with a mean of 5 minutes. There are three spaces in front of the window, including the car being served. Other arriving cars line up outside this 3-car space.
- What is the probability that an arriving car can enter one of the 3-car spaces?
 - What is the probability that an arriving car will wait outside the designated 3-car space?
 - How long is an arriving customer expected to wait before starting service?
 - *How many car spaces should be provided in front of the window (including the car being served) so that an arriving car can find a space there at least 90% of the time?
7. In the $(M/M/1):(GD/\infty/\infty)$, give a plausible argument as to why L_s does not equal $L_q + 1$, in general. Under what condition will the equality hold?
8. For the $(M/M/1):(GD/\infty/\infty)$, derive the expression for L_q using the basic definition $\sum_{n=2}^{\infty} (n-1)p_n$.

9. For the $(M/M/1):(GD/\infty/\infty)$, show that

- (a) The expected number in the queue given that the queue is not empty $= \frac{1}{(1-\rho)}$.
 (b) The expected waiting time in the queue for those who must wait $= \left(\frac{1}{\mu - \lambda}\right)$.

Waiting Time Distribution for $(M/M/1):(FCFS/\infty/\infty)$.¹ The derivation of p_n in the generalized model of Section 15.5 is *totally* independent of the queue discipline. This means that the average measures of performance (w_s , w_q , L_s , and L_q) apply to all queue disciplines.

Although the *average* waiting time is independent of the queue discipline, its probability density function is not. We illustrate this point by deriving the waiting-time distribution for the $(M/M/1)$ model based on the FCFS discipline.

Let τ be the amount of time a person *just arriving* must be in the system (i.e., until the service is completed). Based on the FCFS discipline, if there are n customers in the system ahead of an arriving customer, then

$$\tau = t'_1 + t_2 + \dots + t_{n+1}$$

where t'_1 is the time needed for the customer currently in service to complete service and t_2, t_3, \dots, t_n are the service times for the $n - 1$ customers in the queue. The time t_{n+1} represents the service time for the arriving customer.

Define $w(\tau|n + 1)$ as the conditional density function of τ given n customers in the system ahead of the arriving customer. Because the distribution of the service time is exponential, the forgetfulness property (Section 15.3) tells us that t'_1 is also exponential with the same distribution. Thus, τ is the sum of $n + 1$ identically distributed and independent exponential random variables. From probability theory, $w(\tau|n + 1)$ follows a gamma distribution with parameters μ and $n + 1$. We thus have

$$\begin{aligned} w(\tau) &= \sum_{n=0}^{\infty} w(\tau|n + 1)p_n \\ &= \sum_{n=0}^{\infty} \frac{\mu(\mu\tau)^n e^{-\mu\tau}}{n!} (1 - \rho)\rho^n \\ &= (1 - \rho)\mu e^{-\mu\tau} \sum_{n=0}^{\infty} \frac{(\lambda\tau)^n}{n!} \\ &= (1 - \rho)\mu e^{-\mu\tau} e^{\lambda\tau} \\ &= (\mu - \lambda)e^{-(\mu - \lambda)\tau}, \tau > 0 \end{aligned}$$

Thus, $w(\tau)$ is an exponential distribution with mean $W_s = \frac{1}{(\mu - \lambda)}$.

Example 15.6-3

In the car wash facility model of Example 15.6-2, it is reasonable to assume that this service is performed based on FCFS discipline. Assess the reliability of using W_s as an estimate of the waiting time in the system.

¹This material may be skipped without loss of continuity.

One way of answering this question is to estimate the proportion of customers whose waiting time exceeds W_s . Noting that $W_s = \frac{1}{(\mu - \lambda)}$, we get

$$\begin{aligned} P\{\tau > W_s\} &= 1 - \int_0^{W_s} w(\tau) d\tau \\ &= e^{-(\mu - \lambda)W_s} = e^{-1} = .368 \end{aligned}$$

Thus, under FCFS discipline, about 37% of the customers will wait longer than W_s . This appears excessive, particularly since the current W_s for the car wash facility is already high ($= .5$ hour). We note that the computed probability ($= e^{-1} \approx .368$) is independent of the rates λ and μ for any $(M/M/1):(FCFS/\infty/\infty)$, which means that its value cannot be reduced. Thus, if we design the system based on the average W_s , then we should expect 36.8% of the customers to wait more than the average waiting time.

The situation can be improved in two ways: (1) we can increase the service rate μ to bring the value of W_s down to an acceptable level, or (2) we can select the service rate such that the probability that the waiting time exceeds a prespecified value (say, 10 minutes) remains under a reasonably small percentage (say, 10%). The first method is equivalent to finding μ such that $W_s < \bar{T}$, and the second method finds μ by solving the inequality $P\{\tau > \bar{T}\} < \alpha$, where \bar{T} and α are specified by the analyst.

PROBLEM SET 15.6C

- *1. In Problem 3, Set 15.6b, determine the probability that detective Columbo will take more than 1 week to solve a crime case.
2. In Example 15.6-3, compute the following:
 - (a) The standard deviation of the waiting time τ in the system.
 - (b) The probability that the waiting time in the system will vary by half a standard deviation around the mean value.
3. In Example 15.6-3, determine the service rate μ that satisfies the condition $W_s < 10$ minutes.
4. In Example 15.6-3, determine the service rate μ that will satisfy the condition $P\{\tau > 10 \text{ minutes}\} < .1$.
- *5. Consider Problem 5, Set 15.6b. To attract more business, the owner of the restaurant will give free soft drinks to any customer who waits more than 5 minutes. Given that a drink costs 50 cents, how much will it cost daily to offer free drinks? Assume that the restaurant is open for 12 hours a day.
6. Show that for the $(M/M/1):(FCFS/\infty/\infty)$, the distribution of waiting time in the queue is

$$w_q(t) = \begin{cases} 1 - \rho, & t = 0 \\ \mu\rho(1 - \rho)e^{-(\mu - \lambda)t}, & t > 0 \end{cases}$$

Then find W_q from $w_q(t)$.

$(M/M/1):(GD/N/\infty)$. This model differs from $(M/M/1):(GD/\infty/\infty)$ in that there is a limit N on the number in the system (maximum queue length $= N - 1$). Examples include manufacturing situations in which a machine may have a limited buffer area, and a one-lane drive-in window in a fast-food restaurant.

When the number of customers in the system reaches N , no more arrivals are allowed. Thus, we have

$$\lambda_n = \begin{cases} \lambda, & n = 0, 1, \dots, N-1 \\ 0, & n = N, N+1 \end{cases}$$

$$\mu_n = \mu, \quad n = 0, 1, \dots$$

Using $\rho = \frac{\lambda}{\mu}$, the generalized model in Section 15.5 yields

$$p_n = \begin{cases} \rho^n p_0 & n \leq N \\ 0, & n > N \end{cases}$$

The value of p_0 is determined from the equation $\sum_{n=0}^{\infty} p_n = 1$, which yields

$$p_0(1 + \rho + \rho^2 + \dots + \rho^N) = 1$$

or

$$p_0 = \begin{cases} \frac{(1 - \rho)}{1 - \rho^{N+1}}, & \rho \neq 1 \\ \frac{1}{N+1}, & \rho = 1 \end{cases}$$

Thus,

$$p_n = \begin{cases} \frac{(1 - \rho)\rho^n}{1 - \rho^{N+1}}, & \rho \neq 1 \\ \frac{1}{N+1}, & \rho = 1 \end{cases}, n = 0, 1, \dots, N$$

The value of $\rho = \frac{\lambda}{\mu}$ need not be less than 1 in this model, because arrivals at the system are controlled by the system limit N . This means that λ_{eff} , rather than λ , is the rate that matters in this case. Because customers will be lost when there are N in the system, then, as shown in Figure 15.4,

$$\lambda_{\text{lost}} = \lambda p_N$$

$$\lambda_{\text{eff}} = \lambda - \lambda_{\text{lost}} = \lambda(1 - p_N)$$

In this case, $\lambda_{\text{eff}} < \mu$.

The expected number of customers in the system is computed as

$$\begin{aligned} L_s &= \sum_{n=1}^N n p_n \\ &= \frac{1 - \rho}{1 - \rho^{N+1}} \sum_{n=1}^N n \rho^n \\ &= \left(\frac{1 - \rho}{1 - \rho^{N+1}} \right) \rho \frac{d}{d\rho} \sum_{n=0}^N \rho^n \\ &= \frac{(1 - \rho)\rho}{1 - \rho^{N+1}} \frac{d}{d\rho} \left(\frac{1 - \rho^{N+1}}{1 - \rho} \right) \\ &= \frac{\rho[1 - (N+1)\rho^N + N\rho^{N+1}]}{(1 - \rho)(1 - \rho^{N+1})}, \rho \neq 1 \end{aligned}$$

Scenario 1: (M/M/1): (GD/5/infinity)					
Lambda =		4.00000	Mu =		6.00000
Lambda eff =		3.80752	Rho/c =		0.66667
Ls =		1.42256	Lq =		0.78797
Ws =		0.37362	Wq =		0.20695
n	Probability pn	Cumulative Pn	n	Probability pn	Cumulative Pn
0	0.36541	0.36541	3	0.10827	0.87970
1	0.24361	0.60902	4	0.07218	0.95188
2	0.16241	0.77143	5	0.04812	1.00000

FIGURE 15.6

TORA output of Example 15.6-4 (file toraEx15.6-4.txt)

When $\rho = 1$, $L_s = \frac{N}{2}$ (verify!). We can derive W_s , W_q , and L_q from L_s using λ_{eff} as shown in Section 15.6.1.

The use of a hand calculator to compute the queuing formulas is at best cumbersome (the formulas will get more complex in later models!). The use of TORA or template excelPoissonQ.xls to handle these computations is recommended.

Example 15.6-4

Consider the car wash facility of Example 15.6-2. Suppose that the facility has a total of four parking spaces. If the parking lot is full, newly arriving cars balk to other facilities. The owner wishes to determine the impact of the limited parking space on losing customers to the competition.

In terms of the notation of the model, the limit on the system is $N = 4 + 1 = 5$. The following input data provides the output in Figure 15.6.

Lambda	Mu	c	System limit	Source limit
4	6	1	5	infinity

Because the limit on the system is $N = 5$, the proportion of lost customers is $p_5 = .04812$, which, based on a 24-hour day, is equivalent to losing $(\lambda p_5) \times 24 = 4 \times .04812 \times 24 = 4.62$ cars a day. A decision regarding increasing the size of the parking lot should be based on the value of lost business.

Looking at the problem from a different angle, the expected total time in the system, W_s , is .3736 hour, or approximately 22 minutes, down from 30 minutes in Example 15.6-3 when all arriving cars are allowed to join the facility. This reduction of about 25% is secured at the expense of losing about 4.8% of all potential customers because of the limited parking space.

PROBLEM SET 15.6D

- *1. In Example 15.6-4, determine the following:
 - (a) Probability that an arriving car will go into the wash bay immediately on arrival.
 - (b) Expected waiting time until a service starts.
 - (c) Expected number of empty parking spaces.

- (d) Probability that all parking spaces are occupied.
 (e) Percent reduction in average service time that will limit the average time in the system to about 10 minutes. (*Hint:* Use trial and error with excelPoissonQ.xls or TORA.)

2. Consider the car wash facility of Example 15.6-4. Determine the number of parking spaces such that the percentage of cars that cannot find a space does not exceed 1%.
3. The time barber Joe takes to give a haircut is exponential with a mean of 12 minutes. Because of his popularity, customers usually arrive (according to a Poisson distribution) at a rate much higher than Joe can handle: six customers per hour. Joe really will feel comfortable if the arrival rate is effectively reduced to about four customers per hour. To accomplish this goal, he came up with the idea of providing limited seating in the waiting area so that newly arriving customers will go elsewhere when they discover that all the seats are taken. How many seats should Joe provide to accomplish his goal?
- *4. The final assembly of electric generators at Electro is produced at the Poisson rate of 10 generators per hour. The generators are then conveyed on a belt to the inspection department for final testing. The belt can hold a maximum of 7 generators. An electronic sensor will automatically stop the conveyor once it is full, preventing the final assembly department from assembling more units until a space becomes available. The time to inspect the generators is exponential, with a mean of 15 minutes.
 - (a) What is the probability that the final assembly department will stop production?
 - (b) What is the average number of generators on the conveyor belt?
 - (c) The production engineer claims that interruptions in the assembly department can be reduced by increasing the capacity of the belt. In fact, the engineer claims that the capacity can be increased to the point where the assembly department can operate 95% of the time without interruption. Is this claim justifiable?
5. A cafeteria can seat a maximum of 50 persons. Customers arrive in a Poisson stream at the rate of 10 per hour and are served (one at a time) at the rate of 12 per hour.
 - (a) What is the probability that an arriving customer will not eat in the cafeteria because it is full?
 - (b) Suppose that three customers (with random arrival times) would like to be seated together. What is the probability that their wish can be fulfilled? (Assume that arrangements can be made to seat them together as long as three seats are available.)
6. Patients arrive at a 1-doctor clinic according to a Poisson distribution at the rate of 20 patients per hour. The waiting room does not accommodate more than 14 patients. Examination time per patient is exponential, with a mean of 8 minutes.
 - (a) What is the probability that an arriving patient will not wait?
 - (b) What is the probability that an arriving patient will find a seat in the room?
 - (c) What is the expected total time a patient spends in the clinic?
7. The probabilities p_n of n customers in the system for an $(M/M/1):(GD/5/\infty)$ are given in the following table:

n	0	1	2	3	4	5
p_n	.399	.249	.156	.097	.061	.038

The arrival rate λ is five customers per hour. The service rate μ is eight customers per hour. Compute the following:

- *(a) Probability that an arriving customer will be able to enter the system.
 - *(b) Rate at which the arriving customers will not be able to enter the system.
 - (c) Expected number in the system.
 - (d) Average waiting time in the queue.
8. Show that when $\rho = 1$ for the $(M/M/1):(GD/N/\infty)$ the expected number in the system, L_s , equals $\frac{N}{2}$. (Hint: $1 + 2 + \dots + i = \frac{i(i+1)}{2}$.)
9. Show that λ_{eff} for the $(M/M/1):(GD/N/\infty)$ can be computed from the formula

$$\lambda_{\text{eff}} = \mu(L_s - L_q)$$

15.6.3 Multiple-Server Models

This section considers three queuing models with multiple parallel servers. The first two models are the multiserver versions of the models in Section 15.6.2. The third model treats the self-service case, which is equivalent to having an infinite number of parallel servers.

Real-Life Application—Telephone Sales Manpower Planning at Qantas Airways

To reduce operating costs, Qantas Airways seeks to staff its main telephone sales reservation office efficiently while providing convenient service to its customers. Traditionally, staffing needs are estimated by forecasting future telephone calls based on historical increase in business. The increase in staff numbers is then calculated based on the projected average increase in telephone calls divided by the average number of calls an operator can handle. Because the calculations are based on averages, the additional number of hired staff does not take into account the fluctuations in demand during the day. In particular, long waiting time for service during peak business hours has resulted in customer complaints and lost business. The problem deals with the determination of a plan that strikes a balance between the number of hired operators and the customer needs. The solution uses $(M/M/c)$ queuing analysis imbedded into an integer programming model. Savings from the model in the Sydney office alone were around \$173,000 in fiscal year 1975–1976. The details of the study are given in Case 15, Chapter 24 on the CD.

$(M/M/c):(GD/\infty/\infty)$. In this model, there are c parallel servers. The arrival rate is λ and the service rate per server is μ . Because there is no limit on the number in the system, $\lambda_{\text{eff}} = \lambda$.

The effect of using c parallel servers is a proportionate increase in the facility service rate. In terms of the generalized model (Section 15.5), λ_n and μ_n are thus defined as

$$\lambda_n = \lambda, \quad n \geq 0$$

$$\mu_n = \begin{cases} n\mu, & n < c \\ c\mu, & n \geq c \end{cases}$$

Thus,

$$p_n = \begin{cases} \frac{\lambda^n}{\mu(2\mu)(3\mu)\dots(n\mu)} p_0 = \frac{\lambda^n}{n! \mu^n} p_0 = \frac{\rho^n}{n!} p_0, & n < c \\ \frac{\lambda^n}{(\prod_{i=1}^c i\mu)(c\mu)^{n-c}} p_0 = \frac{\lambda^n}{c! c^{n-c} \mu^n} p_0 = \frac{\rho^n}{c! c^{n-c}} p_0, & n \geq c \end{cases}$$

Letting $\rho = \frac{\lambda}{\mu}$, and assuming $\frac{\rho}{c} < 1$, the value of p_0 is determined from $\sum_{n=0}^{\infty} p_n = 1$, which gives,

$$\begin{aligned} p_0 &= \left\{ \sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!} \sum_{n=c}^{\infty} \left(\frac{\rho}{c}\right)^{n-c} \right\}^{-1} \\ &= \left\{ \sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!} \left(\frac{1}{1 - \frac{\rho}{c}}\right) \right\}^{-1}, \frac{\rho}{c} < 1 \end{aligned}$$

The expression for L_q can be determined as follows:

$$\begin{aligned} L_q &= \sum_{n=c}^{\infty} (n - c) p_n \\ &= \sum_{k=0}^{\infty} k p_{k+c} \\ &= \sum_{k=0}^{\infty} k \frac{\rho^{k+c}}{c^k c!} p_0 \\ &= \frac{\rho^{c+1}}{c! c} p_0 \sum_{k=0}^{\infty} k \left(\frac{\rho}{c}\right)^{k-1} \\ &= \frac{\rho^{c+1}}{c! c} p_0 \frac{d}{d(\frac{\rho}{c})} \sum_{k=0}^{\infty} \left(\frac{\rho}{c}\right)^k \\ &= \frac{\rho^{c+1}}{(c-1)! (c-\rho)^2} p_0 \end{aligned}$$

Because $\lambda_{\text{eff}} = \lambda$, $L_s = L_q + \rho$. The values of W_s and W_q can be determined by dividing L_s and L_q by λ .

Example 15.6-5

A community is served by two cab companies. Each company owns two cabs and both share the market equally, as evidenced by the fact that calls arrive at each company's dispatching office at the rate of eight per hour. The average time per ride is 12 minutes. Calls arrive according to a Poisson distribution, and the ride time is exponential. The two companies recently were bought by an investor who is interested in consolidating them into a single dispatching office to provide better service to customers. Analyze the new owner's proposal.

From the standpoint of queuing, the cabs are the servers, and the cab ride is the service. Each company can be represented by the model $(M/M/2):(GD/\infty/\infty)$ with $\lambda = 8$ calls per

Comparative analysis

c	Lambda	Mu	L'da eff	p0	Ls	Ws	Lq	Wq
2	8.000	5.000	8.00	0.110	4.444	0.556	2.844	0.356
4	16.000	5.000	16.00	0.027	5.586	0.349	2.386	0.149

FIGURE 15.7

TORA output for Example 15.6-5 (file toraEx15.6-5.txt)

hour and $\mu = \frac{60}{10} = 5$ rides per cab per hour. Consolidation will result in the model $(M/M/4):(GD/\infty/\infty)$ with $\lambda = 2 \times 8 = 16$ calls per hour and $\mu = 5$ rides per cab per hour.

A suitable measure for comparing the two models is the average time a customer waits for a ride, W_q . TORA comparative analysis input data are given as follows:

Scenario	Lambda	Mu	c	System limit	Source limit
1	8	5	2	infinity	infinity
2	16	5	4	infinity	infinity

Figure 15.7 provides the output for the two scenarios. The results show that the waiting time for a ride is .356 hour (≈ 21 minutes) for the two-cab situation and .149 (≈ 9 minutes) for the consolidated situation, a remarkable reduction of more than 50% and a clear evidence that the consolidation of the two companies is warranted.

Remark. The conclusion from the preceding analysis is that **pooling services** is *always* a more efficient mode of operation. This result is true even if the separate installations happen to be “very” busy (see Problems 2 and 10, Set 15.6e).

PROBLEM SET 15.6E

- Consider Example 15.6-5.
 - Show that the remarkable reduction in waiting time by more than 50% for the consolidated case is coupled with an increase in the percentage of time the servers remain busy.
 - Determine the number of cabs that the consolidated company should have to limit the average waiting time for a ride to 5 minutes or less.
- *2. In the cab company example, suppose that the average time per ride is actually about 14.5 minutes, so that the utilization $(= \frac{\lambda}{\mu c})$ for the 2- and 4-cab operations increases to more than 96%. Is it still worthwhile to consolidate the two companies into one? Use the average waiting time for a ride as the comparison measure.
- Determine the minimum number of parallel servers needed in each of the following (Poisson arrival/departure) situations to guarantee that the operation of the queuing situation will be stable (i.e., the queue length will not grow indefinitely):
 - Customers arrive every 5 minutes and are served at the rate of 10 customers per hour.
 - The average interarrival time is 2 minutes, and the average service time is 6 minutes.
 - The arrival rate is 30 customers per hour, and the service rate per server is 40 customers per hour.
- Customers arrive at Thrift Bank according to a Poisson distribution, with a mean of 45 customers per hour. Transactions per customer last about 5 minutes and are exponentially

distributed. The bank wants to use a single-line multiple-teller operation, similar to the ones used in airports and post offices. The manager is conscious of the fact that customers may switch to other banks if they perceive that their wait in line is "excessive." For this reason, the manager wants to limit the average waiting time in the queue to no more than 30 seconds. How many tellers should the bank provide?

- *5. McBurger fast food restaurant has 3 cashiers. Customers arrive according to a Poisson distribution every 3 minutes and form one line to be served by the first available cashier. The time to fill an order is exponentially distributed with a mean of 5 minutes. The waiting room inside the restaurant is limited. However, the food is good, and customers are willing to line up outside the restaurant, if necessary. Determine the size of the waiting room inside the restaurant (excluding those at the cashiers) such that the probability that an arriving customer does not wait outside the restaurant is at least .999.
6. A small post office has two open windows. Customers arrive according to a Poisson distribution at the rate of 1 every 3 minutes. However, only 80% of them seek service at the windows. The service time per customer is exponential, with a mean of 5 minutes. All arriving customers form one line and access available windows on an FCFS basis.
 - (a) What is the probability that an arriving customer will wait in line?
 - (b) What is the probability that both windows are idle?
 - (c) What is the average length of the waiting line?
 - (d) Would it be possible to offer reasonable service with only one window? Explain.
7. U of A computer center is equipped with four identical mainframe computers. The number of users at any time is 25. Each user is capable of submitting a job from a terminal every 15 minutes, on the average, but the actual time between submissions is exponential. Arriving jobs will automatically go to the first available computer. The execution time per submission is exponential with mean 2 minutes. Compute the following:
 - * (a) The probability that a job is not executed immediately on submission.
 - (b) The average time until the output of a job is returned to the user.
 - (c) The average number of jobs awaiting execution.
 - (d) The percentage of time the entire computer center is idle.
 - * (e) The average number of idle computers.
8. Drake Airport services rural, suburban, and transit passengers. The arrival distribution for each of the three groups is Poisson with mean rates of 15, 10, and 20 passengers per hour, respectively. The time to check in a passenger is exponential with mean 6 minutes. Determine the number of counters that should be provided at Drake under each of the following conditions:
 - (a) The total average time to check a customer in is less than 15 minutes.
 - (b) The percentage of idleness of the counters does not exceed 10%.
 - (c) The probability that all counters are idle does not exceed .01.
9. In the United States, the use of single-line, multiple-server queues is common in post offices and in passenger check-in counters at airports. However, both grocery stores and banks (especially in smaller communities) tend to favor single-line, single-server setups, despite the fact that single-line, multiple-server queues offer a more efficient operation. Comment on this observation.
10. For the $(M/M/c):(GD/\infty/\infty)$ model, Morse (1958, p. 103) shows that as $\frac{\rho}{c} \rightarrow 1$,

$$L_q = \frac{\rho}{c - \rho}$$

Noting that $\frac{\rho}{c} \rightarrow 1$ means that the servers are extremely busy, use this information to show that the ratio of the average waiting time in queue in the $(M/M/c):(GD/\infty/\infty)$ model to that in the $(M/M/1):(GD/\infty/\infty)$ model approaches $\frac{1}{c}$ as $\frac{\rho}{c} \rightarrow 1$. Thus, for $c = 2$, the average waiting time can be reduced by 50%. The conclusion from this exercise is that it is always advisable to pool services regardless of how "overloaded" the servers may be.

11. In the derivation of p_n for the $(M/M/c):(GD/\infty/\infty)$ model, indicate which part of the derivation requires the condition $\frac{\rho}{c} < 1$. Explain verbally the meaning of the condition. What will happen if the condition is not satisfied?
12. Prove that $L_s = L_q + \bar{c}$ starting with the definition $L_q = \sum_{n=c+1}^{\infty} (n-c)p_n$, where \bar{c} is the average number of busy servers. Hence, show that $\bar{c} = \frac{\lambda_{eff}}{\mu}$.
13. Show that p_n for the $(M/M/1):(GD/\infty/\infty)$ model can be obtained from that of the $(M/M/c):(GD/\infty/\infty)$ by setting $c = 1$.
14. Show that for the $(M/M/c):(GD/\infty/\infty)$ that

$$L_q = \frac{c\rho}{(c-\rho)^2} p_c$$

15. For the $(M/M/c):(GD/\infty/\infty)$ model, show that
 - (a) The probability that a customer is waiting is $\frac{\rho}{(c-\rho)} p_c$.
 - (b) The average number in the queue given that it is not empty is $\frac{c}{(c-\rho)}$.
 - (c) The expected waiting time in the queue for customers who must wait is $\frac{1}{\mu(c-\rho)}$.
16. Prove that the probability density function of waiting time in the queue for the $(M/M/c):(GD/\infty/\infty)$ model is given as

$$w_q(T) = \begin{cases} 1 - \frac{\rho^c}{(c-1)!(c-\rho)} p_0, & T = 0 \\ \frac{\mu \rho^c e^{-\mu(c-\rho)T}}{(c-1)!} p_0, & T > 0 \end{cases}$$

(Hint: Convert the c -channel case into an equivalent single channel for which

$$P\{t > T\} = P\left\{\min_{1 \leq i \leq c} t_i > T\right\} = (e^{-\mu T})^c e^{-\mu c T}$$

where t is the service time in the equivalent single channel.)

17. Prove that for $w_q(T)$ in Problem 16

$$P\{T > y\} = P\{T > 0\} e^{-(c\mu-\lambda)y}$$

where $P\{T > 0\}$ is the probability that an arriving customer must wait.

18. Prove that the waiting time in the system for the $(M/M/c):(FCFS/\infty/\infty)$ model has the following probability density function:

$$w(\tau) = \mu e^{-\mu\tau} + \frac{\rho^c \mu e^{-\mu\tau}}{(c-1)!(c-\rho-1)} \left\{ \frac{1}{c-\rho} - e^{-\mu(c-\rho-1)\tau} \right\} p_0, \tau \geq 0$$

(Hint: τ is the convolution of the waiting time in queue, T [see Problem 16], and the service time distribution.)

$(M/M/c):(GD/N/\infty)$, $c \leq N$. This model differs from that of the $(M/M/c):(GD/\infty/\infty)$ model in that the system limit is finite and equal to N . This means that the maximum queue size is $N - c$. The arrival and service rates are λ and μ . The effective arrival rate λ_{eff} is less than λ because of the system limit, N .

In terms of the generalized model (Section 15.5), λ_n and μ_n for the current model are defined as

$$\lambda_n = \begin{cases} \lambda, & 0 \leq n \leq N \\ 0, & n > N \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & 0 \leq n \leq c \\ c\mu, & c \leq n \leq N \end{cases}$$

Substituting λ_n and μ_n in the general expression in Section 15.5 and noting that $\rho = \frac{\lambda}{\mu}$, we get

$$p_n = \begin{cases} \frac{\rho^n}{n!} p_0, & 0 \leq n < c \\ \frac{\rho^n}{c! c^{n-c}} p_0, & c \leq n \leq N \end{cases}$$

where

$$p_0 = \begin{cases} \left(\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c (1 - (\frac{\rho}{c})^{N-c+1})}{c! (1 - \frac{\rho}{c})} \right)^{-1}, & \frac{\rho}{c} \neq 1 \\ \left(\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c! (N - c + 1)} \right)^{-1}, & \frac{\rho}{c} = 1 \end{cases}$$

Next, we compute L_q for the case where $\frac{\rho}{c} \neq 1$ as

$$\begin{aligned} L_q &= \sum_{n=c}^N (n - c) p_n \\ &= \sum_{j=0}^{N-c} j p_{j+c} \\ &= \frac{\rho^c \rho}{c! c} p_0 \sum_{j=0}^{N-c} j \left(\frac{\rho}{c} \right)^{j-1} \\ &= \frac{\rho^{c+1}}{cc!} p_0 \frac{d}{d(\frac{\rho}{c})} \sum_{j=0}^{N-c} \left(\frac{\rho}{c} \right)^j \\ &= \frac{\rho^{c+1}}{(c-1)!(c-\rho)^2} \left\{ 1 - \left(\frac{\rho}{c} \right)^{N-c+1} - (N-c+1) \left(1 - \frac{\rho}{c} \right) \left(\frac{\rho}{c} \right)^{N-c} \right\} p_0 \end{aligned}$$

It can be shown that for $\frac{\rho}{c} = 1$, L_q reduces to

$$L_q = \frac{\rho^c (N - c)(N - c + 1)}{2c!} p_0, \quad \frac{\rho}{c} = 1$$

To determine W_q and hence W_s and L_s , we compute the value of λ_{eff} as

$$\lambda_{\text{lost}} = \lambda p_N$$

$$\lambda_{\text{eff}} = \lambda - \lambda_{\text{lost}} = (1 - p_N) \lambda$$

Scenario1: (M/M/4):(GD/10/infinity)					
Lambda =		16.00000	Mu =		5.00000
Lambda eff =		15.42815	Rho/c =		0.80000
Ls =		4.23984	Lq =		1.15421
Ws =		0.27481	Wq =		0.07481
n	Probability pn	Cumulative Pn	n	Probability pn	Cumulative Pn
0	0.03121	0.03121	6	0.08726	0.79393
1	0.09986	0.13106	7	0.06981	0.86374
2	0.15977	0.29084	8	0.05584	0.91958
3	0.17043	0.46126	9	0.04468	0.96426
4	0.13634	0.59760	10	0.03574	1.00000

FIGURE 15.8

TORA output of Example 15.6-6 (file toraEx15.6-6.txt)

Example 15.6-6

In the consolidated cab company problem of Example 15.6-5, suppose that new funds cannot be secured to purchase additional cabs. The owner was advised by a consultant that one way to reduce the waiting time is for the dispatching office to inform new customers of potential excessive delay once the waiting list reaches 6 customers. This move is certain to get new customers to seek service elsewhere, but will reduce the waiting time for those on the waiting list. Assess the friend's advice.

Limiting the waiting list to 6 customers is equivalent to setting $N = 6 + 4 = 10$ customers. We are thus investigating the model $(M/M/4):(GD/10/\infty)$, where $\lambda = 16$ customers per hour and $\mu = 5$ rides per hour. The following input data provide the results in Figure 15.8.

Lambda	Mu	c	System limit	Source limit
16	5	4	10	infinity

The average waiting time, W_q , before setting a limit on the capacity of the system is .149 hour (≈ 9 minutes) (see Figure 15.7), which is about twice the new average of .075 hour (≈ 4.5 minutes). This remarkable reduction is achieved at the expense of losing about 3.6% of potential customers ($p_{10} = .03574$). However, this result does not reflect the effect of possible loss of customer goodwill on the operation of the company.

PROBLEM SET 15.6F

1. In Example 15.6-6, determine the following:
 - (a) The expected number of idle cabs.
 - (b) The probability that a calling customer will be the last on the list.
 - (c) The limit on the waiting list if it is desired to keep the waiting time in the queue to below 3 minutes.

2. Eat & Gas convenience store operates a two-pump gas station. The lane leading to the pumps can house at most 3 cars, excluding those being serviced. Arriving cars go elsewhere if the lane is full. The distribution of arriving cars is Poisson with mean 20 per hour. The time to fill up and pay for the purchase is exponential with mean 6 minutes. Determine the following:
 - (a) Percentage of cars that will seek business elsewhere.
 - (b) Percentage of time one pump is in use.
 - *(c) Percent utilization of the two pumps.
 - *(d) Probability that an arriving car will not start service immediately but will find an empty space in the lane.
 - (e) Capacity of the lane that will ensure that, on the average, no more than 10% of the arriving cars are turned away.
 - (f) Capacity of the lane that will ensure that the probability that both pumps are idle is .05 or less.
3. A small engine repair shop is run by three mechanics. Early in March of each year, people bring in their tillers and lawn mowers for service and maintenance. The shop is willing to accept all the tillers and mowers that customers bring in. However, when new customers see the floor of the shop covered with waiting jobs, they go elsewhere for more prompt service. The floor shop can house a maximum of 15 mowers or tillers, excluding those being serviced. The customers arrive at the shop every 10 minutes on the average, and it takes a mechanic an average of 30 minutes to complete each job. Both the interarrival and the service times are exponential. Determine the following:
 - (a) Average number of idle mechanics:
 - (b) Amount of business lost to competition per 10-hour day because of the limited capacity of the shop.
 - (c) Probability that the next arriving customer will be serviced by the shop.
 - (d) Probability that at least one of the mechanics will be idle.
 - (e) Average number of tillers or mowers awaiting service.
 - (f) A measure of the overall productivity of the shop.
4. At U of A, newly enrolled freshmen students are notorious for wanting to drive their cars to class (even though most of them are required to live on campus and can conveniently make use of the university free transit system). During the first couple of weeks of the fall semester, traffic havoc prevails on campus as freshmen try desperately to find parking spaces. With unusual dedication, the students wait patiently in the lanes of the parking lot for someone to leave so they can park their cars. Let us consider a specific scenario: The parking lot has 30 parking spaces but can also accommodate 10 more cars in the lanes. These additional 10 cars cannot park in the lanes permanently and must await the availability of one of the 30 parking spaces. Freshman students arrive at the parking lot according to a Poisson distribution, with a mean of 20 cars per hour. The parking time per car averages about 60 minutes but actually follows an exponential distribution.
 - *(a) What is the percentage of freshmen who are turned away because they cannot enter the lot?
 - *(b) What is the probability that an arriving car will wait in the lanes?
 - (c) What is the probability that an arriving car will occupy the only remaining parking space on the lot?

- *(d) Determine the average number of occupied parking spaces.
 - (e) Determine the average number of spaces that are occupied in the lanes.
 - *(f) Determine the number of freshmen who will not make it to class during an 8-hour period because the parking lot is totally full.
5. Verify the expression for p_0 for the $(M/M/c):(GD/N/\infty)$ given that $\frac{\rho}{c} \neq 1$.
 6. Prove the following equality for the $(M/M/c):(GD/N/\infty)$

$$\lambda_{\text{eff}} = \mu \bar{c}$$

where \bar{c} is the number of busy servers.

7. Verify the expression for p_0 and L_q for the $(M/M/c):(GD/N/\infty)$ when $\frac{\rho}{c} = 1$.
8. For the $(M/M/c):(GD/N/\infty)$ model in which $N = c$, define λ_n and μ_n in terms of the generalized model (Section 15.5), then show that the expression for p_n is given as

$$p_n = \frac{\rho^n}{n!} p_0, n = 1, 2, \dots, c$$

where

$$p_0 = \left(1 + \sum_{n=1}^c \frac{\rho^n}{n!} \right)^{-1}$$

$(M/M/\infty):(GD/\infty/\infty)$ —Self-Service Model. In this model, the number of servers is unlimited because the customer is also the server. A typical example is taking the written part of a driver's license test. Self-service gas stations and 24-hour ATM banks do not fall under this model's description because the servers in these cases are actually the gas pumps and the ATM machines. The model assumes steady arrival and service rates, λ and μ , respectively.

In terms of the generalized model of Section 15.5, we have

$$\lambda_n = \lambda, n = 0, 1, 2, \dots$$

$$\mu_n = n\mu, n = 0, 1, 2, \dots$$

Thus,

$$p_n = \frac{\lambda^n}{n! \mu^n} p_0 = \frac{\rho^n}{n!} p_0, n = 0, 1, 2, \dots$$

Because $\sum_{n=0}^{\infty} p_n = 1$, it follows that

$$p_0 = \frac{1}{1 + \rho + \frac{\rho^2}{2!} + \dots} = \frac{1}{e^\rho} = e^{-\rho}$$

As a result,

$$p_n = \frac{e^{-\rho} \rho^n}{n!}, n = 0, 1, 2, \dots$$

which is Poisson with mean $L_s = \rho$. As should be expected, L_q and W_q are zero because it is a self-service model.

Example 15.6-7

An investor invests \$1000 a month on average in one type of stock market security. Because the investor must wait for a good “buy” opportunity, the actual time of purchase is totally random. The investor usually keeps the securities for about 3 years on the average but will sell them at random times when a “sell” opportunity presents itself. Although the investor is generally recognized as a shrewd stock market player, past experience indicates that about 25% of the securities decline at about 20% a year. The remaining 75% appreciate at the rate of about 12% a year. Estimate the investor’s (long-run) average equity in the stock market.

This situation can be treated as an $(M/M/\infty):(GD/\infty/\infty)$ because, for all practical purposes, the investor does not have to wait in line to buy or to sell securities. The average time between order placements is 1 month, which yields $\lambda \approx 12$ securities per year. The rate of selling securities is $\mu = \frac{1}{3}$ security per year. You can secure the model output using the following input:

Lambda	Mu	c	System limit	Source limit
12	.3333333	infinity	infinity	infinity

Given the values of λ and μ , we obtain

$$L_s = \rho = \frac{\lambda}{\mu} = 36 \text{ securities}$$

The estimate of the (long-run) average *annual* net worth of the investor is

$$(.25L_s \times \$1000)(1 - .20) + (.75L_s \times \$1000)(1 + .12) = \$63,990$$

PROBLEM SET 15.6G

- In Example 15.6-7, compute the following:
 - The probability that the investor will sell out completely.
 - The probability that the investor will own at least 10 securities.
 - The probability that the investor will own between 30 and 40 securities, inclusive.
 - The investor’s net annual equity if only 10% of the securities depreciate by 30% a year, and the remaining 90% appreciate by 15% a year.
- New drivers are required to pass written tests before they are given a road driving test. These tests are usually administered by the city police department. Records at the City of Springdale show that the average number of written tests is 100 per 8-hour day. The average time needed to complete the test is about 30 minutes. However, the actual arrival of test takers and the time each spends on the test are totally random. Determine the following:
 - The average number of seats the police department should provide in the test hall.
 - The probability that the number of test takers will exceed the average number of seats provided in the test hall.
 - The probability that no tests will be administered in any one day.
- Show (by using excelPoissonQ.xls or TORA) that for small $\rho = .1$, the values of L_s , L_q , W_s , W_q , and p_n for the $(M/M/c):(GD/\infty/\infty)$ model can be estimated reliably using the less cumbersome formulas of the $(M/M/\infty):(GD/\infty/\infty)$ model for c as small as 4 servers.
- Repeat Problem 3 for large $\rho = 9$ and show that the same conclusion holds except that the value of c must be higher (at least 14). From the results of Problems 3 and 4, what

general conclusion can be drawn regarding the use of the $(M/M/\infty):(GD/\infty/\infty)$ to estimate the results of the $(M/M/c):(GD/\infty/\infty)$ model?

15.6.4 Machine Servicing Model— $(M/M/R):(GD/K/K), R < K$

The setting for this model is a shop with K machines. When a machine breaks down, one of R available repairpersons is called upon to do the repair. The rate of breakdown *per machine* is λ breakdowns per unit time, and a repairperson will service broken machines at the rate of μ machines per unit time. All breakdowns and services are assumed to follow the Poisson distribution.

This model differs from all the preceding ones because it has a finite calling source. We can see this point by realizing that when all the machines in the shop are broken, no more calls for service can be generated. In essence, only machines in working order can break down and hence can generate calls for service.

Given the rate of breakdown per machine, λ , the rate of breakdown for the *entire shop* is proportional to the number of machines that are in working order. In terms of the queuing model, having n machines *in the system* signifies that n machines are broken. Thus, the rate of breakdown for the entire shop is

$$\lambda_n = (K - n)\lambda, 0 \leq n \leq K$$

In terms of the generalized model of Section 15.5, we have

$$\lambda_n = \begin{cases} (K - n)\lambda, & 0 \leq n \leq K \\ 0, & n \geq K \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & 0 \leq n \leq R \\ R\mu, & R \leq n \leq K \end{cases}$$

From the generalized model, we can then obtain (verify!)

$$p_n = \begin{cases} C_n^K \rho^n p_0, & 0 \leq n \leq R \\ C_n^K \frac{n! \rho^n}{R! R^{n-R}} p_0, & R \leq n \leq K \end{cases}$$

$$p_0 = \left(\sum_{n=0}^R C_n^K \rho^n + \sum_{n=R+1}^K C_n^K \frac{n! \rho^n}{R! R^{n-R}} \right)^{-1}$$

There is no closed form expression for L_s , and hence it must be computed using the following basic definition:

$$L_s = \sum_{n=0}^K n p_n$$

The value of λ_{eff} is computed as

$$\lambda_{\text{eff}} = E\{\lambda(K - n)\} = \lambda(K - L_s)$$

Using the formulas in Section 15.6.1, we can compute the remaining measures of performance W_s , W_q , and L_q .

Example 15.6-8

Toolco operates a machine shop with a total of 22 machines. Each machine is known to break down once every 2 hours, on the average. It takes an average of 12 minutes to complete a repair. Both the time between breakdowns and the repair time follow the exponential distribution. Toolco is interested in determining the number of repairpersons needed to keep the shop running "smoothly."

The situation can be analyzed by investigating the productivity of the machines as a function of the number of repairpersons. Such productivity measure can be defined as

$$\begin{aligned} \left(\frac{\text{Machines}}{\text{productivity}} \right) &= \frac{\text{Available machines} - \text{Broken machines}}{\text{Available machines}} \times 100 \\ &= \frac{22 - L_s}{22} \times 100 \end{aligned}$$

The results for this situation can be obtained using the following input data: $\lambda = .5$, $\mu = 5$, $R = 1, 2, 3$, or 4 , system limit $= 22$, and source limit $= 22$. Figure 15.9 provides the output. The associated productivity is computed as follows:

Repairperson, R	1	2	3	4
Machines productivity (100%)	45.44	80.15	88.79	90.45
Marginal increase (100%)	—	34.71	8.64	1.66

The results show that with one repairperson the productivity is low ($= 45.44\%$). By increasing the number of repairpersons to two, the productivity jumps by 34.71% to 80.15%. When we employ three repairpersons, the productivity increases only by about 8.64% to 88.79%, whereas four repairpersons will increase the productivity by a meager 1.66% to 90.45%.

Judging from these results, the use of two repairpersons is justifiable. The case for three repairpersons is not as strong because it raises the productivity by only 8.64%. Perhaps a monetary comparison between the cost of hiring a third repairperson and the income attributed to the 8.64% increase in productivity can be used to settle this point (see Section 15.10 for discussion of cost models). As for hiring a fourth repairperson, the meager increase of 1.66% in productivity does not justify such an action.

FIGURE 15.9
TORA comparative analysis output for Example 15.6-8 (file toraEx15.6-8.txt)

Comparative Analysis

c	Lambda	Mu	L'da eff	p0	Ls	Lq	Ws	Wq
1	0.500	5.00	4.9980	0.0004	12.0040	11.0044	2.4018	2.2018
2	0.500	5.00	8.8161	0.0564	4.3677	2.6045	0.4954	0.2954
3	0.500	5.00	9.7670	0.1078	2.4660	0.5128	0.2525	0.0525
4	0.500	5.00	9.9500	0.1199	2.1001	0.1102	0.2111	0.0111

PROBLEM SET 15.6H

1. In Example 15.6-8, do the following:
 - (a) Verify the values of λ_{eff} given in Figure 15.9.
 - *(b) Compute the expected number of idle repairpersons given $R = 4$.
 - (c) Compute the probability that all repairpersons are idle given $R = 3$.
 - *(d) Compute the probability that the majority (more than half) of repairpersons are idle given $R = 3$.
2. In Example 15.6-8, define and compute the productivity of the repairpersons for $R = 1, 2, 3$, and 4. Use this information in conjunction with the measure of machine productivity to decide on the number of repairpersons Toolco should hire.
3. In the computations in Figure 15.9, it may appear confusing that the average rate of machine breakdown in the shop, λ_{eff} , increases with the increase in R . Explain why the increase in λ_{eff} should be expected.
- *4. An operator attends five automatic machines. After each machine completes a batch run, the operator must reset it before a new batch is started. The time to complete a batch run is exponential with mean 45 minutes. The setup time is also exponential with mean 8 minutes.
 - (a) Determine the average number of machines that are awaiting setup or are being set up.
 - (b) Compute the probability that all machines are working.
 - (c) Determine the average time a machine is down.
5. Kleen All is a service company that performs a variety of odd jobs, such as yard work, tree pruning, and house painting. The company's four employees leave the office with the first assignment of the day. After completing an assignment, the employee calls the office requesting instruction for the next job to be performed. The time to complete an assignment is exponential, with a mean of 45 minutes. The travel time between jobs is also exponential, with a mean of 20 minutes.
 - (a) Determine the average number of employees who are traveling between jobs.
 - (b) Compute the probability that no employee is on the road.
- *6. After a long wait, the Newborns were rewarded with quintuplets, two boys and three girls, thanks to the wonders of new medical advances. During the first 5 months, the babies' life consisted of two states: awake (and mostly crying) and asleep. According to the Newborns, the babies "awake-asleep" activities never coincide. Instead, the whole affair is totally random. In fact, Mrs. Newborn, a statistician by profession, believes that the length of time each baby cries is exponential, with a mean of 30 minutes. The amount of sleep each baby gets also happens to be exponential, with a mean of 2 hours. Determine the following:
 - (a) The average number of babies who are awake at any one time.
 - (b) The probability that all babies are asleep.
 - (c) The probability that the Newborns will not be happy because more babies are awake (and crying) than are asleep.
7. Verify the expression for p_n for the $(M/M/R):(GD/K/K)$ model.
8. Show that the rate of breakdown in the shop can be computed from the formula

$$\lambda_{\text{eff}} = \mu \bar{R}$$

where \bar{R} is the average number of busy repairpersons.

9. Verify the following results for the special case of one repairperson ($R = 1$):

$$p_n = \frac{K! \rho^n}{(K-n)!} p_0$$

$$p_0 = \left(1 + \sum_{n=1}^R \frac{K! \rho^n}{(K-n)!} \right)^{-1}$$

$$L_s = K - \frac{(1-p_0)}{\rho}$$

15.7 (M/G/1):(GD/∞/∞)—POLLACZEK-KHINTCHINE (P-K) FORMULA

Queuing models in which the arrivals and departures do not follow the Poisson distribution are complex. In general, it is advisable to use simulation as an alternative tool for analyzing these situations (see Chapter 16).

This section presents one of the few non-Poisson queues for which analytic results are available. It deals with the case in which the service time, t , is represented by any probability distribution with mean $E\{t\}$ and variance $\text{var}\{t\}$. The results of the model include the basic measures of performance L_s , L_q , W_s , and W_q . The model does not provide a closed-form expression for p_n because of analytic intractability.

Let λ be the arrival rate at the single-server facility. Given $E\{t\}$ and $\text{var}\{t\}$ of the service time distribution and that $\lambda E\{t\} < 1$, it can be shown using sophisticated probability/Markov chain analysis that

$$L_s = \lambda E\{t\} + \frac{\lambda^2 (E^2\{t\} + \text{var}\{t\})}{2(1 - \lambda E\{t\})}, \lambda E\{t\} < 1$$

The probability that the facility is empty (idle) is computed as

$$p_0 = 1 - \lambda E\{t\} = 1 - \rho$$

Because $\lambda_{\text{eff}} = \lambda$, the remaining measures of performance (L_q , W_s , and W_q) can be derived from L_s , as explained in Section 15.6.1.

Template excelPKFormula.xls automates the calculations of this model.

Example 15.7-1

In the Automata car wash facility of Example 15.6-2, suppose that a new system is installed so that the service time for all cars is constant and equal to 10 minutes. How does the new system affect the operation of the facility?

From Example 15.6-2, $\lambda_{\text{eff}} = \lambda = 4$ cars per hour. The service time is constant so that $E\{t\} = \frac{10}{60} = \frac{1}{6}$ hour and $\text{var}\{t\} = 0$. Thus,

$$L_s = 4\left(\frac{1}{6}\right) + \frac{4^2\left(\left(\frac{1}{6}\right)^2 + 0\right)}{2\left(1 - \frac{4}{6}\right)} = 1.33 \text{ cars}$$

$$L_q = 1.333 - \left(\frac{4}{6}\right) = .667 \text{ cars}$$

$$W_s = \frac{1.333}{4} = .333 \text{ hour}$$

$$W_q = \frac{.667}{4} = .167 \text{ hour}$$

It is interesting that even though the arrival and departure rates are the same as in the Poisson case of Example 15.6-2 ($\lambda = 4$ cars per hour and $\mu = \frac{1}{E(t)} = 6$ cars per hour), the expected waiting time is lower in the current model because the service time is constant, as the following table shows.

	(M/M/1):(GD/∞/∞)	(M/D/1):(GD/∞/∞)
W_s (hr)	.500	.333
W_q (hr)	.333	.167

The results make sense because a constant service time indicates *more certainty* in the operation of the facility.

PROBLEM SET 15.7A

- In Example 15.7-1, compute the percentage of time the facility is idle.
- Solve Example 15.7-1 assuming that the service-time distribution is given as follows:
 - Uniform between 8 and 20 minutes.
 - Normal with $\mu = 12$ minutes and $\sigma = 3$ minutes.
 - Discrete with values equal to 4, 8, and 15 minutes and probabilities .2, .6, and .2, respectively.
- Layson Roofing Inc. installs shingle roofs on new and old residences in Arkansas. Prospective customers request the service randomly at the rate of nine jobs per 30-day month and are placed on a waiting list to be processed on a FCFS basis. Homes sizes vary, but it is fairly reasonable to assume that the roof areas are uniformly distributed between 150 and 300 squares. The work crew can usually complete 75 squares a day. Determine the following:
 - Layson's average backlog of roofing jobs.
 - The average time a customer waits until a roofing job is completed.
 - If the work crew is increased to the point where they can complete 150 squares a day, how will this affect the average time until a job is completed?
- *Optica, Ltd., makes prescription glasses according to orders received from customers. Each worker is specialized in certain types of glasses. The company has been experiencing unusual delays in the processing of bifocal and trifocal prescriptions. The worker in charge receives 30 orders per 8-hour day. The time to complete a prescription is normally distributed, with a mean of 12 minutes and a standard deviation of 3 minutes. After spending between 2 and 4 minutes, uniformly distributed, to inspect the glasses, the worker can start on a new prescription. Determine the following:
 - The percentage of time the worker is idle.
 - The average backlog of bifocal and trifocal prescriptions in Optica.
 - The average time until a prescription is filled.

5. A product arrives according to a Poisson distribution at the rate of one every 45 minutes. The product requires two tandem operations attended by one worker. The first operation uses a semiautomatic machine that completes its cycle in exactly 28 minutes. The second operation makes adjustments and minor changes, and its time depends on the condition of the product when it leaves operation 1. Specifically, the time of operation 2 is uniform between 3 and 6 minutes. Because each operation requires the complete attention of the worker, a new item cannot be loaded on the semiautomatic machine until the current item has cleared operation 2.
- Determine the number of items that are awaiting processing on the semiautomatic machine.
 - What is the percentage of time the worker will be idle?
 - How much time is needed, on the average, for an arriving item to clear operation 2?
6. $(M/D/1):(GD/\infty/\infty)$. Show that for the case where the service time is constant, the P-K formula reduces to

$$L_s = \rho + \frac{\rho^2}{2(1 - \rho)}$$

where $\mu = \frac{1}{E\{t\}}$ and $\rho = \frac{\lambda}{\mu} = \lambda E\{t\}$.

7. $(M/E_m/1):(GD/\infty/\infty)$. Given that the service time is Erlang with parameters m and μ (i.e., $E\{t\} = \frac{m}{\mu}$ and $\text{var}\{t\} = \frac{m}{\mu^2}$), show that the P-K formula reduces to

$$L_s = m\rho + \frac{m(1 + m)\rho^2}{2(1 - m\rho)}$$

8. Show that the P-K formula reduces to L_s of the $(M/M/1):(GD/\infty/\infty)$ when the service time is exponential with a mean of $\frac{1}{\mu}$ time units.
9. In a service facility with c parallel servers, suppose that customers arrive according to a Poisson distribution, with a mean rate of λ . Arriving customers are assigned to servers (busy or free) on a strict rotational basis.
- Determine the probability distribution of the interarrival time.
 - Suppose in part (a) that arriving customers are assigned randomly to the c servers with probabilities α_i , $\alpha_i \geq 0$, $i = 1, 2, \dots, c$, and $\alpha_1 + \alpha_2 + \dots + \alpha_c = 1$. Determine the probability distribution of the interarrival time.

15.8 OTHER QUEUING MODELS

The preceding sections have concentrated on the Poisson queuing models. Queuing literature is rich with other types of models. In particular, queues with priority for service, network queues, and non-Poisson queues form an important body of the queuing theory literature. These models can be found in most specialized books on queuing theory.

15.9 QUEUING DECISION MODELS

The *service level* in a queuing facility is a function of the service rate, μ , and the number of parallel servers, c . This section presents two decision models for determining "suitable" service levels for queuing systems: (1) a cost model, and (2) an aspiration-level

model. Both models recognize that higher service levels reduce the waiting time in the system. Both models aim at striking a balance between the conflicting factors of service level and waiting.

15.9.1 Cost Models

Cost models attempt to balance two conflicting costs:

1. Cost of offering the service.
2. Cost of delay in offering the service (customer waiting time).

The two types of costs are in conflict because an increase in one automatically causes reduction in the other, as demonstrated earlier in Figure 15.1.

Letting x ($= \mu$ or c) represent the *service level*, the cost model can be expressed as

$$ETC(x) = EOC(x) + EWC(x)$$

where

ETC = Expected total cost *per unit time*

EOC = Expected cost of operating the facility *per unit time*

EWC = Expected cost of waiting *per unit time*

The simplest forms for EOC and EWC are the following linear functions:

$$EOC(x) = C_1x$$

$$EWC(x) = C_2L_s$$

where

C_1 = *Marginal* cost per unit of x per unit time

C_2 = Cost of waiting per unit time per (waiting) customer

The following two examples illustrate the use of the cost model. The first example assumes x to equal the service rate, μ , and the second assumes x to equal the number of parallel servers, c .

Example 15.9-1

KeenCo Publishing is in the process of purchasing a high-speed commercial copier. Four models whose specifications are summarized below have been proposed by vendors.

Copier model	Operating cost (\$/hr)	Speed (sheets/min)
1	15	30
2	20	36
3	24	50
4	27	66

Jobs arrive at KeenCo according to a Poisson distribution with a mean of four jobs per 24-hour day. Job size is random but averages about 10,000 sheets per job. Contracts with the customers specify a penalty cost for late delivery of \$80 per jobs per day. Which copier should KeenCo purchase?

Let the subscript i represent copier model i ($i = 1, 2, 3, 4$). The total expected cost *per day* associated with copier i is

$$\begin{aligned} ETC_i &= EOC_i + EWC_i \\ &= C_{1i} \times 24 + C_{2i}L_{si} \\ &= 24C_{1i} + 80L_{si}, i = 1, 2, 3, 4 \end{aligned}$$

The values of C_{1i} are given by the data of the problem. We determine L_{si} by recognizing that, for all practical purposes, each copier can be treated as an $(M/M/1):(GD/\infty/\infty)$ model. The arrival rate is $\lambda = 4$ jobs/day. The service rate μ_i associated with model i is computed as

Model i	Service rate μ_i (jobs/day)
1	4.32
2	5.18
3	7.20
4	9.50

Computation of the service rate is demonstrated for model 1.

$$\text{Average time per job} = \frac{10,000}{30} \times \frac{1}{60} = 5.56 \text{ hours}$$

Thus,

$$\mu_1 = \frac{24}{5.56} = 4.32 \text{ jobs/day}$$

The values of L_{si} , computed by TORA or excelPoissonQ.xls, are given in the following table:

Model i	λ_i (Jobs/day)	μ_i (Jobs/day)	L_{si} (Jobs)
1	4	4.32	12.50
2	4	5.18	3.39
3	4	7.20	1.25
4	4	9.50	0.73

The costs for the four models are computed as follows:

Model i	EOC_i (\$)	EWC_i (\$)	ETC_i (\$)
1	360.00	1000.00	1360.00
2	480.00	271.20	751.20
3	576.00	100.00	676.00
4	648.00	58.40	706.40

Model 3 produces the lowest cost.

PROBLEM SET 15.9A

1. In Example 15.9-1, do the following:
 - (a) Verify the values of μ_2 , μ_3 , and μ_4 given in the example.
 - (b) Suppose that the penalty of \$80 per job per day is levied only on jobs that are *not* "in progress" at the end of the day. Which copier yields the lowest total cost per day?
- *2. Metalco is in the process of hiring a repairperson for a 10-machine shop. Two candidates are under consideration. The first candidate can carry out repairs at the rate of 5 machines per hour and earns \$15 an hour. The second candidate, being more skilled, receives \$20 an hour and can repair 8 machines per hour. Metalco estimates that each broken machine will incur a cost of \$50 an hour because of lost production. Assuming that machines break down according to a Poisson distribution with a mean of 3 per hour and that repair time is exponential, which repairperson should be hired?
3. B&K Groceries is opening a new store boasting "state-of-the-art" check-out scanners. Mr. Bih, one of the owners of B&K, has limited the choices to two scanners: scanner *A* can process 10 items a minute, and the better-quality scanner *B* can scan 15 items a minute. The daily (10 hours) cost of operating and maintaining the scanners are \$25 and \$35 for models *A* and *B*, respectively. Customers who finish shopping arrive at the cashier according to a Poisson distribution at the rate of 10 customers per hour. Each customer's cart carries between 25 and 35 items, uniformly distributed. Mr. Bih estimates the average cost per waiting customer per minute to be about 20 cents. Which scanner should B&K acquire? (*Hint:* The service time per customer is not exponential. It is uniformly distributed.)
4. H&I Industry produces a special machine with different production rates (pieces per hour) to meet customer specifications. A shop owner is considering buying one of these machines and wants to decide on the most economical speed (in pieces per hour) to be ordered. From past experience, the owner estimates that orders from customers arrive at the shop according to a Poisson distribution at the rate of three orders per hour. Each order averages about 500 pieces. Contracts between the owner and the customers specify a penalty of \$100 per late order per hour.
 - (a) Assuming that the actual production time per order is exponential, develop a general cost model as a function of the production rate, μ .
 - *(b) From the cost model in (a), determine an expression for the optimal production rate.
 - *(c) Using the data given in the problem, determine the optimal production rate the owner should request from H&I.
5. Jobs arrive at a machine shop according to a Poisson distribution at the rate of 80 jobs per week. An automatic machine represents the bottleneck in the shop. It is estimated that a unit increase in the production rate of the machine will cost \$250 per week. Delayed jobs normally result in lost business, which is estimated to be \$500 per job per week. Determine the optimum production rate for the automatic machine.
6. Pizza Unlimited sells two franchised restaurant models. Model *A* has a capacity of 20 groups of customers, and model *B* can seat 30 groups. The monthly cost of operating model *A* is \$12,000 and that of model *B* is \$16,000. An investor wants to set up a buffet-style pizza restaurant and estimates that groups of customers, each occupying one table, arrive according to a Poisson distribution at a rate of 25 groups per hour. If all the tables are occupied, customers will go elsewhere. Model *A* will serve 26 groups per hour, and model *B* will serve 29 groups per hour. Because of the variation in group sizes and in the types of orders, the service time is exponential. The investor estimates that the average

cost of lost business per customer group per hour is \$15. A delay in serving waiting customers is estimated to cost an average of \$10 per customer group per hour.

- (a) Develop an appropriate cost model.
 - (b) Assuming that the restaurant will be open for business 10 hours a day, which model would you recommend for the investor?
7. Suppose in Problem 6 that the investor can choose any desired restaurant capacity based on a specific marginal cost for each additional capacity unit requested. Derive the associated general cost model, and define all its components and terms.
 8. Second Time Around sells popular used items on consignment. Its operation can be viewed as an inventory problem in which the stock is replenished and depleted randomly according to Poisson distributions with rates λ and μ items per day. Every time unit the item is out of stock, Second Time loses $\$C_1$ because of lost opportunities, and every time unit an item is held in stock, a holding cost $\$C_2$ is incurred.
 - (a) Develop an expression for the expected total cost per unit time.
 - (b) Determine the optimal value of $\rho = \frac{\lambda}{\mu}$. What condition must be imposed on the relative values of C_1 and C_2 in order for the solution to be consistent with the assumptions of the $(M/M/1):(GD/\infty/\infty)$ model?

Example 15.9-2

In a multiclerk tool crib facility, requests for tool exchange occur according to a Poisson distribution at the rate of 17.5 requests per hour. Each clerk can handle an average of 10 requests per hour. The cost of hiring a new clerk in the facility is \$12 an hour. The cost of lost production per waiting machine per hour is approximately \$50. Determine the optimal number of clerks for the facility.

The situation corresponds to an $(M/M/c):(GD/\infty/\infty)$ model in which it is desired to determine the optimum value of c . Thus, in the general cost model presented at the start of this section, we put $x = c$, resulting in the following cost model:

$$\begin{aligned} ETC(c) &= C_1c + C_2L_s(c) \\ &= 12c + 50L_s(c) \end{aligned}$$

Note that $L_s(c)$ is a function of the number of (parallel) clerks in the crib.

We use the $(M/M/c):(GD/\infty/\infty)$ model with $\lambda = 17.5$ requests per hour and $\mu = 10$ requests per hour. In this regard, the model will reach steady state only if $c > \frac{\lambda}{\mu}$ —that is, for the present example, $c \geq 2$. The following table provides the necessary calculation for determining optimal c . The values of $L_s(c)$ (determined by excelPoissonQ.xls or TORA) given below show that the optimum number of clerks is 4.

c	$L_s(c)$ (requests)	$ETC(c)$ (\$)
2	7.467	397.35
3	2.217	146.85
4	1.842	140.10
5	1.769	148.45
6	1.754	159.70

PROBLEM SET 15.9B

1. Solve Example 15.9-2, assuming that $C_1 = \$20$ and $C_2 = \$45$.
- *2. Tasco Oil owns a pipeline booster unit that operates continuously. The time between breakdowns for each booster is exponential with a mean of 20 hours. The repair time is exponential with mean 3 hours. In a particular station, two repairpersons attend 10 boosters. The hourly wage for each repairperson is \$18. Pipeline losses are estimated to be \$30 per broken booster per hour. Tasco is studying the possibility of hiring an additional repairperson.
 - (a) Will there be any cost savings in hiring a third repairperson?
 - (b) What is the schedule loss in dollars per breakdown when the number of repairpersons on duty is two? Three?
3. A company leases a wide-area telecommunications service (WATS) telephone line for \$2000 a month. The office is open 200 working hours per month. At all other times, the WATS line service is used for other purposes and is not available for company business. Access to the WATS line during business hours is extended to 100 salespersons, each of whom may need the line at any time but averages twice per 8-hour day with exponential time between calls. A salesperson will always wait for the WATS line if it is busy at an estimated inconvenience of 1 cent per minute of waiting. It is assumed that no additional needs for calls will arise while the salesperson waits for a given call. The normal cost of calls (not using the WATS line) averages about 50 cents per minute, and the duration of each call is exponential, with a mean of 6 minutes. The company is considering leasing (at the same price) a second WATS line to improve service.
 - (a) Is the single WATS line saving the company money over a no-WATS system? How much is the company gaining or losing per month over the no-WATS system?
 - (b) Should the company lease a second WATS line? How much would it gain or lose over the single WATS case by leasing an additional line?
- *4. A machine shop includes 20 machines and 3 repairpersons. A working machine breaks down randomly according to a Poisson distribution. The repair time per machine is exponential with a mean of 6 minutes. A queuing analysis of the situation shows an average of 57.8 calls for repair per 8-hour day for the entire shop. Suppose that the production rate per machine is 25 units per hour and that each produced unit generates \$2 in revenue. Further, assume that a repairperson is paid at the rate of \$20 an hour. Compare the cost of hiring the repairpersons against the cost of lost revenue when machines are broken.
5. The necessary conditions for $ETC(c)$ (defined earlier) to assume a minimum value at $c = c^*$ are

$$ETC(c^* - 1) \geq ETC(c^*) \text{ and } ETC(c^* + 1) \geq ETC(c^*)$$

Show that these conditions reduce to

$$L_s(c^*) - L_s(c^* + 1) \leq \frac{C_1}{C_2} \leq L_s(c^* - 1) - L_s(c^*)$$

Apply the result to Example 15.9-2 and show that it yields $c^* = 4$.

15.9.2 Aspiration Level Model

The viability of the cost model depends on how well we can estimate the cost parameters. Generally, these parameters are difficult to estimate, particularly the one associated

with the waiting time of customers. The aspiration level model seeks to alleviate this difficulty by working directly with the measures of performance of the queuing situation. The idea is to determine an acceptable range for the service level (μ or c) by specifying reasonable limits on *conflicting* measures of performance. Such limits are the *aspiration levels* the decision maker wishes to reach.

We illustrate the procedure by applying it to the multiple-server model, where it is desired to determine an "acceptable" number of servers, c^* . We do so by considering the following two (conflicting) measures of performance:

1. The average time in the system, W_s .
2. The idleness percentage of the servers, X .

The idleness percentage can be computed as follows:

$$X = \frac{c - \bar{c}}{c} \times 100 = \frac{c - (L_s - L_q)}{c} \times 100 = \left(1 - \frac{\lambda_{\text{eff}}}{c\mu}\right) \times 100$$

(See Problem 12, Set 15.6e for the proof.)

The problem reduces to determining the number of servers c^* such that

$$W_s \leq \alpha \text{ and } X \leq \beta$$

where α and β are the levels of aspiration specified by the decision maker. For example, we may stipulate that $\alpha = 3$ minutes and $\beta = 10\%$.

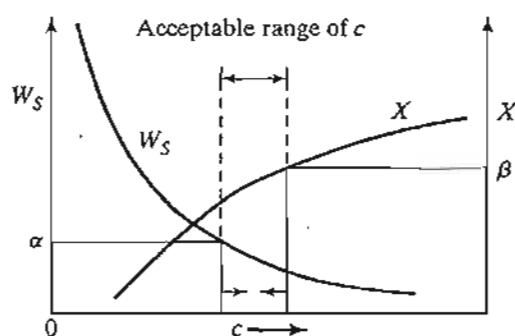
The solution of the problem may be determined by plotting W_s and X as a function of c , as shown in Figure 15.10. By locating α and β on the graph, we can immediately determine an acceptable range for c^* . If the two conditions cannot be satisfied simultaneously, then one or both must be relaxed before a feasible range can be determined.

Example 15.9-3

In Example 15.9-2, suppose that it is desired to determine the number of clerks such that the expected waiting time until a tool is received stays below 5 minutes. Simultaneously, it is also required to keep the percentage of idleness below 20%.

Offhand, and before any calculations are made, an aspiration limit of 5 minutes on the waiting time until a tool is received (i.e., $W_s \leq 5$ minutes) is definitely unreachable because, accord-

FIGURE 15.10
Application of aspiration levels in queuing decision-making



ing to the data of the problem, the average service time alone is 6 minutes. The following table summarizes W_s and X as a function of c :

c	2	3	4	5	6	7	8
W_s (min)	25.4	7.6	6.3	6.1	6.0	6.0	6.0
X (%)	12.5	41.7	56.3	65.0	70.8	75.0	78.0

Based on these results, we should either reduce the service time or recognize that the source of the problem is that tools are being requested at an unreasonably high rate ($\lambda = 17.5$ requests per hour). This, most likely, is the area that should be addressed. For example, we may want to investigate the reason for such high demand for tool replacement. Could it be that the design of the tool itself is faulty? Or could it be that the operators of the machines are purposely trying to disrupt production to express grievances?

PROBLEM SET 15.9C

- *1. A shop uses 10 identical machines. Each machine breaks down once every 7 hours on the average. It takes half an hour on the average to repair a broken machine. Both the breakdown and repair processes follow the Poisson distribution. Determine the following:
 - (a) The number of repairpersons needed such that the average number of broken machines is less than 1.
 - (b) The number of repairpersons needed so that the expected delay time until repair is started is less than 10 minutes.
2. In the cost model in Section 15.9.1, it is generally difficult to estimate the cost parameter C_2 (cost of waiting). As a result, it may be helpful to compute the cost C_2 implied by the aspiration levels. Using the aspiration level model to determine c^* , we can then estimate the implied C_2 by using the following inequality:

$$L_s(c^*) - L_s(c^* + 1) \leq \frac{C_1}{C_2} \leq L_s(c^* - 1) - L_s(c^*)$$

(See Problem 5, Set 15.9b, for the derivation.) Apply the procedure to the problem in Example 15.9-2, assuming $c^* = 3$ and $C_1 = \$12.00$.

REFERENCES

- Bose, S., *An Introduction to Queuing Systems*, Kluwer Academic Publishers, Boston, 2001.
- Hall, R., *Queuing Methods for Service and Manufacturing*, Prentice Hall, Upper Saddle River, NJ, 1991.
- Lee, A., *Applied Queuing Theory*, St. Martin's Press, New York, 1966.
- Lipsky, L., *Queuing Theory, A Linear Algebraic Approach*, Macmillan, New York, 1992.
- Morse, P., *Queues, Inventories, and Maintenance*, Wiley, New York, 1958.
- Parzen, E., *Stochastic Processes*, Holden-Day, San Francisco, 1962.
- Saaty, T., *Elements of Queuing Theory with Applications*, Dover, New York, 1983.
- Tanner, M., *Practical Queuing Analysis*, McGraw-Hill, New York, 1995.
- Tijms, H. C., *Stochastic Models—An Algorithmic Approach*, Wiley, New York, 1994.