cycle to the next. Based on an analysis of past order cycles, it is estimated that the standard deviation of the lead time is 3.3 weeks. All other relevant figures are given in parts (*a*) and (*b*). Find the mean and the standard deviation of lead time demand in this case.

d. If Crazy Charlie's uses a Type 1 service objective of 98 percent to control the replenishment of this item, what is the value of the reorder level? (Assume that the lead time demand has a normal distribution.)

39. The home appliance department of a large department store is using a lot size–reorder point system to control the replenishment of a particular model of FM table radio. The store sells an average of 10 radios each week. Weekly demand follows a normal distribution with variance 26.

    The store pays $20 for each radio, which it sells for $75. Fixed costs of replenishment amount to $28. The accounting department recommends a 20 percent interest rate for the cost of capital. Storage costs amount to 3 percent and breakage to 2 percent of the value of each item.

    If a customer demands the radio when it is out of stock, the customer will generally go elsewhere. Loss-of-goodwill costs are estimated to be about $25 per radio. Replenishment lead time is three months.

    a. If lot sizes are based on the EOQ formula, what reorder level should be used for the radios?

    b. Find the optimal values of (Q, R).

    c. Compare the average annual costs of holding, ordering, and stock-out for the policies that you found in parts (*a*) and (*b*).

    d. Re-solve the problem using Equations (1) and (2′) rather than (1) and (2). What is the effect of including lost sales explicitly?

40. Re-solve the problem faced by the department store mentioned in Problem 39, replacing the stock-out cost with a 96 percent Type 1 service level.

41. Re-solve the problem faced by the department store mentioned in Problem 39, replacing the stock-out cost with a 96 percent Type 2 service level. What is the imputed shortage cost?

42. Consider the equation giving the expected average annual cost of the policy (Q, R) in a continuous-review inventory control system from Section 5.4:

$$G(Q, R) = h\left(\frac{Q}{2} + r - \lambda T\right) + \frac{K\lambda}{Q} + \frac{P\lambda n(R)}{Q}.$$

    Design a spreadsheet to compute $G(Q, R)$ for a range of values of $Q \geq \text{EOQ}$ and $R \geq \mu$. Use the following approximation formula for $L(z)$ to avoid table look-ups:

$$L(z) = \exp(-0.92 - 1.19z - 0.37z^2).$$

    (This formula is from Parr, 1972.) Store the problem parameters c, h, p, μ, σ, and λ in cell locations. Visually search through the tabled values of $G(Q, R)$ to discover the minimum value and estimate the optimal (Q, R) values in this manner. Compare your results to the true optimal found from manual calculation.

    a. Solve Example 5.4.

    b. Solve Problem 13 in this manner.

    c. Solve Problem 14 in this manner.

43. The daily demand for a spare engine part is a random variable with a distribution, based on past experience, given by

Number of Demands per Day	Probability
0	.21
1	.38
2	.19
3	.14
4	.08

The part is expected to be obsolete after 400 days. Assume that demands from one day to the next are independent. The parts cost $1,500 each when acquired in advance of the 400-day period and $5,000 each when purchased on an emergency basis during the 400-day period. Holding costs for unused parts are based on a daily interest rate of 0.08 percent. Unused parts can be scrapped for 10 percent of their purchase price. How many parts should be acquired in advance of the 400-day period? (Hint: Let $D_1, D_2, \ldots, D_{400}$ represent the daily demand for the part. Assume each D_i has mean μ and variance σ^2. The central limit theorem says that the total demand for the 400-day period, ΣD_i, is approximately normally distributed with mean 400μ and variance $400\sigma^2$.)

44. Cassorla's Clothes sells a large number of white dress shirts. The shirts, which bear the store label, are shipped from a manufacturer in New York City. Hy Cassorla, the proprietor, says, "I want to be sure that I never run out of dress shirts. I always try to keep at least a two months' supply in stock. When my inventory drops below that level, I order another two-month supply. I've been using that method for 20 years, and it works."

The shirts cost $6 each and sell for $15 each. The cost of processing an order and receiving new goods amounts to $80, and it takes three weeks to receive a shipment. Monthly demand is approximately normally distributed with mean 120 and standard deviation 32. Assume a 20 percent annual interest rate for computing the holding cost.

a. What value of Q and R is Hy Cassorla using to control the inventory of white dress shirts?

b. What fill rate (Type 2 service level) is being achieved with the current policy?

c. Based on a 99 percent fill rate criterion, determine the optimal values of Q and R that he should be using. (Assume four weeks in a month for your calculations.)

d. Determine the difference in the average annual holding and setup costs between the policies in parts (*b*) and (*c*).

e. Estimate how much time would be required to pay for a $25,000 inventory control system, assuming that the dress shirts represent 5 percent of Hy's annual business and that similar savings could be realized on the other items as well.

45. The Poisson distribution is discussed in Appendix 5–D at the end of this chapter. Assume that the distribution of bagels sold daily at Billy's Bakery in Problem 8 follows a Poisson distribution with mean 16 per day. Using Table A–3 in the back of the book or the Poisson distribution function built into Excel, determine the optimal number of bagels for Billy's to bake each day.

46. Consider the Crestview Printing Company described in Problem 9. Suppose that sales of cards (in units of 50,000) follow a Poisson distribution with mean 6.3. Using Table A–3 in the back of the book or the Poisson distribution function built into Excel find the optimal number of cards for Crestview to print for the next Christmas season.
47. The Laplace distribution is discussed in Appendix 5–D. As noted there, the Laplace distribution could be a good choice for describing demand for slow-moving items and for fast-moving items with high variation. The cdf of the Laplace distribution is given by

$$F(x) = 0.5[1 + \text{sgn}(x - \mu)](1 - \exp(-|x - \mu|/\theta),$$

and the inverse of the cdf is given by

$$F^{-1}(p) = \mu - \theta \text{ sgn } (p - 0.5) \ln (1 - 2|p - 0.5|),$$

where sgn(x) is the sign of x. The mean is μ and the variance is $2\theta^2$. Since the inverse distribution function can be written down explicitly, one does not have to resort to tables to solve newsvendor problems when demand follows the Laplace distribution.

Solve Problem 10, part (a), assuming the demand for the EX123 follows a Laplace distribution with parameters $\mu = 60$ and $\theta = 3\sqrt{2}$ (which will give exactly the same mean and variance).

48. Solve Problem 11 assuming the demand for fans over the selling season follows a Laplace distribution with the same mean and variance as you computed in Problem 11(a).
49. Solve Problem 12(b) assuming the demand for handbags over the selling season follows a Laplace distribution with mean 150 and standard deviation 20.
50. Solve Problem 13 assuming that the lead time demand follows a Laplace distribution with mean and variance equal to that which was computed in Problem 13. What difference do you see in the (Q, R) values as compared to those for the normal case?
51. Solve Problem 14 assuming that the lead time demand follows a Laplace distribution with mean and variance equal to the mean and variance of lead time demand you computed for Problem 14. What difference do you see in the (Q, R) values as compared to those for the normal case?

Appendix **5–A**

Notational Conventions and Probability Review

Demand will be denoted by D, which is assumed to be a random variable. The cumulative distribution function (cdf) of demand is $F(x)$ and is defined by

$$F(x) = P\{D \leq x\} \qquad \text{for } -\infty < x < +\infty.$$

When D is continuous, the probability density function (pdf) of demand, $f(x)$, is defined by

$$f(x) = \frac{dF(x)}{dx}.$$

When D is discrete, $f(x)$ is the probability function (pf) defined by

$$f(x) = P\{X = x\} = F(x) - F(x - 1).$$

Note that in the continuous case the density function is not a probability and the value of $f(x)$ is not necessarily less than 1, although it is always true that $f(x) \geq 0$ for all x.

The expected value of demand, $E(D)$, is defined as

$$E(D) = \int_{-\infty}^{+\infty} xf(x)\,dx$$

in the continuous case, and

$$E(D) = \sum_{x=-\infty}^{+\infty} xf(x)$$

in the discrete case.

We use the symbol μ to represent the expected value of demand [$E(D) = \mu$]. In what follows we assume that D is continuous; similar formulas hold in the discrete case. Let $g(x)$ be any real-valued function of the real variable x. Then

$$E(g(D)) = \int_{-\infty}^{+\infty} g(x)f(x)dx.$$

In particular, let $g(D) = \max(0, Q - D)$. Then

$$E(g(D)) = \int_{-\infty}^{+\infty} \max(0, Q - x)f(x)\,dx.$$

Because demand is nonnegative, it must be true that $f(x) = 0$ for $x < 0$. Furthermore, when $x > Q$, $\max(0, Q - x) = 0$, so we may write

$$E(g(D)) = \int_{0}^{Q} (Q - x)f(x)dx.$$

In the analysis of the newsvendor model, Leibniz's rule is used to determine the derivative of $G(Q)$. According to Leibniz's rule:

$$\frac{d}{dy}\int_{a_1(y)}^{a_2(y)} h(x, y)\,dx = \int_{a_1(y)}^{a_2(y)} [\partial h(x, y)/\partial y]\,dx + h(a_2(y), y)a'_2(y) - h(a_1(y), y)a'_1(y).$$

Appendix **5–B**

Additional Results and Extensions for the Newsvendor Model

1. INTERPRETATION OF THE OVERAGE AND UNDERAGE COSTS FOR THE SINGLE PERIOD PROBLEM

Define

S = Selling price of the item.

c = Variable cost of the item.

h = Holding cost per unit of inventory remaining in stock at the end of the period.

p = Loss-of-goodwill cost plus bookkeeping expense (charged against the number of back orders on the books at the end of the period).

We show how c_u and c_o should be interpreted in terms of these parameters. As earlier, let Q be the order quantity and D the demand during the period. Assume without loss of generality that starting inventory is zero. Then the cost incurred at the end of the period is

$$cQ + h\max(Q - D, 0) + p\max(D - Q, 0) - S\min(Q, D).$$

The expected cost is

$$G(Q) = cQ + h\int_0^Q (Q - x)f(x)\,dx + p\int_Q^\infty (x - Q)f(x)\,dx$$

$$- S\int_0^Q xf(x)\,dx - SQ\int_Q^\infty f(x)\,dx.$$

Using

$$\int_0^Q xf(x)dx = \int_0^\infty xf(x)\,dx - \int_Q^\infty xf(x)\,dx = \mu - \int_Q^\infty xf(x)\,dx,$$

the expected cost may be written

$$G(Q) = cQ + h\int_0^Q (Q - x)f(x)\,dx + (p + S)\int_Q^\infty (x - Q)f(x)\,dx - S\mu.$$

The optimal order quantity satisfies

$$G'(Q) = 0$$

or

$$c + hF(Q) - (p + S)(1 - F(Q)) = 0,$$

which results in

$$F(Q) = \frac{P + S - c}{p + S + h}.$$

Setting $c_u = p + S - c$ and $c_o = h + c$ gives the critical ratio in the form $c_u/(c_u + c_o)$.

2. THE NEWSVENDOR COST WHEN DEMAND IS NORMAL

When the one period demand follows a normal distribution, we can obtain an explicit expression for the optimal one period cost for the newsvendor model from Section 5.3. We know that the expected single period cost function is

$$G(Q) = c_o\int_0^Q (Q - t)\,f(t)\,dt + c_u\int_Q^\infty (t - Q)\,f(t)\,dt,$$

and the optimal value of Q satisfies

$$F(Q^*) = \frac{c_u}{c_u + c_o}.$$

It is convenient to express $G(Q)$ in a slightly different form. Note that

$$\int_0^Q (Q - t)\,f(t)\,dt = \int_0^\infty (Q - t)\,f(t)\,dt - \int_Q^\infty (Q - t)\,f(t)\,dt = Q - \mu + n(Q) \text{ where}$$

$$n(Q) = \int_Q^\infty (t - Q)\,f(t)\,dt.$$

It follows that *G(Q)* can be written in the form

$$G(Q) = c_o(Q - \mu) + (c_u + c_o)n(Q).$$

where μ is the expected demand.

We know from Section 5.4 that one can write an explicit expression for *n(Q)* when demand is normal. In particular, we showed that $n(Q) = \sigma L(z)$ where $L(z)$ is the standard loss integral. As noted in Section 5.4,

$$L(z) = \phi(z) - z[1 - \phi(z)].$$

where $\phi(z)$ is the standard normal density, $\phi(z)$ the standard normal distribution function and $z = (Q - \mu)/\sigma$ is the standardized value of the order quantity Q.

It follows that in the normal case we can write

$$G(Q) = c_o(Q - \mu) + (c_u + c_o)\sigma[\phi(z) - z(1 - \phi(z)].$$

The standard normal density function can be computed directly, and the cumulative distribution function is available through tables and is a built in function in Excel.

It might also be of interest to know the minimum value of the expected cost. This is the value of the function G when $Q = Q^*$. Note that at $Q = Q^*$ the cumulative distribution function $\phi(z^*)$ is equal to the critical ratio $c_u/(c_u + c_o)$, which means the complementary cumulative distribution function $1 - \phi(z^*)$ is equal to the ratio $c_o/(c_u + c_o)$. Also, since $z = (Q - \mu)/\sigma$, it follows that $Q - \mu = \sigma z$. Making these two substitutions in the expression above for $G(Q)$ we obtain

$$G(Q^*) = c_o\sigma z^* + (c_u + c_o)\sigma\left[\phi(z^*) - z^*\left(\frac{c_o}{c_o + c_u}\right)\right] = (c_u + c_o)\sigma\phi(z^*).^1$$

Note that if one knows z^* no table look up is required to compute this expression since we know that

$$\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-0.5z^2}.$$

The expression for *G(Q*)* shows that the optimal newsvendor cost increases linearly in both the underage and overage costs as well as in the standard deviation of demand.

3. EXTENSION TO INFINITE HORIZON ASSUMING FULL BACK-ORDERING OF DEMAND

Let $D_1, D_2, \ldots$ be an infinite sequence of demands. Assume that the demands are independent identically distributed random variables having common distribution function $F(x)$ and density $f(x)$. The policy is to order up to Q each period. As all excess demand is back-ordered, the order quantity in any period is exactly the previous period's demand. The number of units sold will also equal the demand. In order to see this, consider the number of units sold in successive periods:

Number of units sold in period 1 = $\min(Q, D_1)$,

Number of units sold in period 2 = $\max(D_1 - Q, 0) + \min(Q, D_2)$,

Number of units sold in period 3 = $\max(D_2 - Q, 0) + \min(Q, D_3)$,

and so on.

[1]Porteus (2002) has derived essentially the same expression on page 13. I am grateful to Gerard Cachon for helpful discussions on this point.

These relationships follow because back-ordering of excess demand means that the sales are made in the subsequent period. Now, notice that

$$\min(Q, D_i) + \max(D_i - Q, 0) = D_i \qquad \text{for } i = 1, 2, \ldots,$$

which follows from considering the cases $Q < D_i$ and $Q \geq D_i$.

Hence, the expected cost over n periods is

$$cQ + (c - S)E(D_1 + D_2 + \ldots + D_{n-1}) - (S)E[\min(Q, D_n)] + nL(Q)$$

$$= cQ + (c - S)(n - 1)\mu - (S)E[\min(Q, D_n)] + nL(Q)$$

where

$$L(Q) = h\int_0^Q (Q - x)f(x)dx + p\int_Q^\infty (x - Q)f(x)\,dx.$$

Dividing by n and letting $n \to \infty$ gives the average cost over infinitely many periods as

$$(c - S)\mu + L(Q).$$

The optimal value of Q occurs where $L'(Q) = 0$, which results in

$$F(Q) = p/(p + h).$$

4. EXTENSION TO INFINITE HORIZON ASSUMING LOST SALES

If excess demand is lost rather than back-ordered, then the previous argument is no longer valid. The number of units sold in period 1 is $\min(Q, D_1)$, which is also the number of units ordered in period 2; the number of units sold in period 2 is $\min(Q, D_2)$ (since there is no back-ordering of excess demand), which is also the number of units ordered in period 3; and so on. As shown in Section 1 of this appendix,

$$E[\min(Q, D)] = \mu - \int_Q^\infty (x - Q)f(x)\,dx.$$

Hence, it follows that the expected cost over n periods is given by

$$cQ + [(n - 1)c - nS]\left[\mu - \int_Q^\infty (x - Q)f(x)\,dx\right] + nL(Q)$$

If we divide by n and let n $\to$, we obtain the following expression for the average cost per period:

$$(c - S)\left[\mu - \int_Q^\infty (x - Q)f(x)\,dx\right] + L(Q).$$

Differentiating with respect to Q and setting the result equal to zero gives the following condition for the optimal Q:

$$F(Q) = \frac{p + S - c}{p + S + h - c}.$$

Setting $c_u = p + S - c$ and $c_\circ = h$ gives the critical ratio in the form $c_u/(c_u + c_\circ)$. Hence, we interpret c_u as the cost of the loss of goodwill plus the lost profit per sale, and c_o as the cost of holding only.

Appendix 5–C

Derivation of the Optimal (Q, R) Policy

From Section 5.4, the objective is to find values of the variables Q and R to minimize the function

$$G(Q, R) = h(Q/2 + R - \lambda\tau) + K\lambda/Q + p\lambda n(R)/Q. \tag{1}$$

Because this function is to be minimized with respect to the two variables (Q, R), a necessary condition for optimality is that $\partial G/\partial Q = \partial G/\partial R = 0$. The two resulting equations are

$$\frac{\partial G}{\partial Q} = \frac{h}{2} - \frac{K\lambda}{Q^2} - \frac{p\lambda n(R)}{Q^2} = 0, \tag{2}$$

$$\frac{\partial G}{\partial R} = h + p\lambda n'(R)/Q = 0. \tag{3}$$

Note that since $n(R) = \int_R^\infty (x - R)f(x)\,dx$, one can show that

$$n'(R) = -(1 - F(R)).$$

From Equation (2) we obtain

$$\frac{1}{Q^2}[K\lambda + p\lambda n(R)] = \frac{h}{2}$$

or

$$Q^2 = \frac{2K\lambda + 2p\lambda n(R)}{h},$$

which gives

$$Q = \sqrt{\frac{2\lambda[K + pn(R)]}{h}}. \tag{4}$$

From Equation (3) we obtain

$$h + p\lambda[-(1 - F(R))]/Q = 0,$$

which gives

$$1 - F(R) = Qh/p\lambda. \tag{5}$$

Author's note: We use the term *optimal* somewhat loosely here. Technically speaking, the model that gives rise to these two equations is only approximate for several reasons. For one, the use of the average expected inventory is an approximation, since this is not the same as the expected average inventory (because one should not charge holding costs when inventory goes negative). As we saw in the discussion of negative safety stock in this chapter, there are cases where the right-hand side of Equation (5) can exceed 1, and the model "blows up." This would not be the case for a truly exact model, which is beyond the scope of this book.

Appendix 5–D

Probability Distributions for Inventory Management

In this chapter we have made frequent reference to the normal distribution as a model for demand uncertainty. Although the normal distribution certainly dominates applications, it is not the only choice available. In fact, it could be a poor choice in some circumstances. In this appendix we discuss other distributions for modeling demand uncertainty.

1. THE POISSON DISTRIBUTION AS A MODEL OF DEMAND UNCERTAINTY

One situation in which the normal distribution may be inappropriate is for slow-moving items, that is, ones with small demand rates. Because the normal is an infinite distribution, when the mean is small it is possible that a substantial portion of the density curve extends into the negative half line. This could give poor results for safety stock calculations. A common choice for modeling slow movers is the Poisson distribution. The Poisson is a discrete distribution defined on the positive half line only. Let X have the Poisson distribution with parameter μ. Then

$$f(x) = \frac{e^{-\mu}\mu^x}{x!} \qquad \text{for } x = 0, 1, 2, \ldots .$$

(The derivation of the Poisson distribution and its relationship to the exponential distribution are discussed in detail in Section 12.3).

An important feature of the Poisson distribution is that both the mean and the variance are equal to μ (giving $\sigma = \sqrt{\mu}$). Hence, it should be true that the observed standard deviation of periodic demand (or lead time demand) is close to the square root of the mean periodic demand (or lead time demand) for the Poisson to be an appropriate choice. Table A–3 at the back of the book is a table of the complementary cumulative Poisson distribution. This table allows one to compute optimal policies for the newsvendor and the (Q, R) models assuming that demand follows the Poisson distribution.

We show how to find optimal policies for both the newsvendor and the (Q, R) models when demand is Poisson. For the newsvendor case, one simply applies the method outlined in Section 5.3 for discrete demand and obtains the probabilities from Table A–3. We illustrate with an example.

Example *5D.1* Consider Mac's newsstand, discussed in several examples in Chapter 5. A few regular customers have asked Mac to stock a particular stereo magazine. Mac has agreed even though sales have been slow. Based on past data, Mac has found that a Poisson distribution with mean 5 closely fits the weekly sales pattern for the magazine. Mac purchases the magazines for \$0.50 and sells them for \$1.50. He returns unsold copies to his supplier, who pays Mac \$0.10 for each. Find the number of these magazines he should purchase from his supplier each week.

Solution The overage cost is $c_o = \$0.50 - \$0.10 = \$0.40$ and the underage cost is $c_u = \$1.50 - \$0.50 = \$1.00$. It follows that the critical ratio is $c_u/(c_o + c_u) = 1/1.4 = .7143$. As Table A–3 gives the complementary cumulative distribution, we subtract table entries from 1 to obtain values of the cumulative distribution function. From the table we see that

$$F(5) = 1 - .5595 = .4405,$$
$$F(6) = 1 - .3840 = .6160,$$
$$F(7) = 1 - .2378 = .7622.$$

Because the critical ratio is between $F(6)$ and $F(7)$, we move to the larger value, thus giving an optimal order quantity of 7 magazines.

It is not difficult to calculate optimal (Q, R) policies when the demand follows a Poisson distribution, but the process requires an iterative solution similar to that for the normal distribution described in Section 5.5. Define $P(x)$ as the complementary cumulative probability of the Poisson. That is,

$$P(x) = \sum_{k=x}^{\infty} f(k).$$

Table A–3 gives values of $P(x)$. It can then be shown that for the Poisson distribution

$$n(R) = \mu P(R) - RP(R + 1),$$

which means that all the information required to compute optimal policies appears in Table A–3.[1] That is, special tables for computing $n(R)$ are not required in this case. This relationship allows one to compute the (Q, R) policy using the pair of equations (1) and (2) from Section 5.4.

Example *5D.2* A department store uses a (Q, R) policy to control its inventories. Sales of a pocket FM radio average 1.4 radios per week. The radio costs the store $50.00, and fixed costs of replenishment amount to $20.00. Holding costs are based on a 20 percent annual interest charge, and the store manager estimates a $12.50 cost of lost profit and loss of goodwill if the radio is demanded when out of stock. Reorder lead time is 2 weeks, and statistical studies have shown that demand over the lead time is accurately described by a Poisson distribution. Find the optimal lot sizes and reorder points for this item.

Solution The relevant parameters for this problem are

$$\lambda = (1.4)(52) = 72.8.$$
$$h = (50)(0.2) = 10.0.$$
$$K = \$20.$$
$$\mu = (1.4)(2) = 2.8.$$
$$p = \$12.50.$$

To start the solution process, we compute the EOQ. It is

$$\text{EOQ} = \sqrt{\frac{2K\lambda}{h}} = \sqrt{\frac{(2)(20)(72.8)}{10}} = 17.1.$$

The next step is to find R solving

$$P(R) = Qh/p\lambda.$$

Substituting the EOQ for Q and solving gives $P(R_0) = .1868$. From Table A–3, we see that $R = 4$ results in $P(R) = .3081$ and $R = 5$ results in $P(R) = .1523$. Assuming the conservative strategy of rounding to the *larger* R, we would choose $R = 5$ and $P(R) = .1523$. It now follows that

$$n(R_0) = (2.8)(.1523) - (5)(.0651) = 0.1009.$$

Hence Q_1 is

$$Q_1 = \sqrt{\frac{(2)(72.8)[20 + (12.5)(0.1009)]}{10}} = 17.6.$$

It follows that $P(R_1) = .1934$, giving $R_1 = R_0 = 5$. Hence the solution has converged, because successive R (and, hence, Q) values are the same. The optimal solution is $(Q, R) = (18, 5)$.

2. THE LAPLACE DISTRIBUTION

A continuous distribution, which has been suggested for modeling slow-moving items or ones with more variance in the tails than the normal, is the Laplace distribution. The Laplace distribution has been called the pseudo-exponential distribution, as it is mathematically an exponential distribution with a symmetric mirror image. (The exponential distribution is discussed at length in Chapter 12 in the context of reliability management.)

The mathematical form of the Laplace pdf is

$$f(x) = \frac{1}{2\theta} \exp(-|x - \mu|/\theta) \qquad \text{for } -\infty < x < +\infty.$$

Because the pdf is symmetric around μ, the mean is μ. The variance is $2\theta^2$. The Laplace distribution is also a reasonable model for slow-moving items, and is an

[1] See Hadley and Whitin (1963), p. 441.

alternative to the normal distribution for fast-moving items when there is more spread in the tails of the distribution than the normal distribution gives. As far as inventory applications are concerned, Presutti and Trepp (1970) noticed that it significantly simplified the calculation of the optimal policy for the (Q, R) model.

One can show that for any value of $R > \mu$, the complementary cumulative distribution ($P(R)$) and the loss integral ($n(R)$) are given by

$$P(R) = 0.5 \exp(-[(R - \mu)/\theta])$$
$$n(R) = 0.5\theta \exp(-[(R - \mu)/\theta])$$

so that the ratio $n(R)/P(R) = \theta$, independent of R. This fact results in a very simple solution for the (Q, R) model. Recall that the two equations defining the optimal policy were

$$Q = \sqrt{\frac{2\lambda[K + pm(R)]}{h}}$$

$$P(R) = Qh/p\lambda \qquad [\text{where } P(R) = 1 - F(R)].$$

The simplification is achieved by using the SOQ formula presented in Section 5.5. The SOQ representation is an alternative representation for Q that does not include the stock-out cost, p. Using $P(R) = 1 - F(R)$, the SOQ formula is

$$Q = \frac{n(R)}{P(R)} + \sqrt{\frac{2K\lambda}{h} + \left(\frac{n(R)}{P(R)}\right)^2}$$
$$= \theta + \sqrt{\frac{2K\lambda}{h} + \theta^2}$$

independently of R! Hence, the optimal Q and R can be found in a simple one-step calculation. When using a cost model, find the value of $P(R)$ from $P(R) = Qh/p\lambda$. Then, using the representation $P(R) = \exp[-(R - \mu)/\theta)]/2$, it follows that $R = -\theta \ln(2P(R)) + \mu$. Using a service level model, one simply uses the formulas for R given in Section 5.5. We illustrate with an example.

Example *5D.3*

Consider Example 5D.2, but suppose that we wish to use the Laplace distribution to compute the optimal policy rather than the Poisson distribution. As both the mean and the variance of the lead time demand are 2.8, we set $\mu = 2.8$ and $2\theta^2 = 2.8$, giving $\theta = 1.1832$. It follows that

$$Q = 1.1832 + \sqrt{\frac{(2)(20)(72.8)}{10} + (1.1832)^2} = 18.3.$$

As with Example 5D.2, we obtain $P(R) = .1934$. Using $R = -\theta \ln(2P(R)) + \mu$ and substituting $P(R) = .1934$ gives $R = 3.92$, which we round to 4. Notice that this solution differs slightly from the one we obtained assuming Poisson demand. However, recall that using the Poisson distribution we found that the optimal R was between 4 and 5, which we chose to round to 5 to be conservative.

3. OTHER LEAD TIME DEMAND DISTRIBUTIONS

Many other probability distributions have been recommended for modeling lead time demand. Some of these include the negative binomial distribution, the gamma distribution, the logarithmic distribution, and the Pearson distribution. (See Silver and Peterson,

1985, p. 289, for a list of articles that discuss these distributions in the context of inventory management.) The normal distribution probably accounts for the lion's share of applications, with the Poisson accounting for almost all the rest. We included the Laplace distribution because of the interesting property that optimal (Q, R) policies can be found without an iterative solution, and because it could be a good alternative to the Poisson for low-demand items.

Appendix **5–E**

Glossary of Notation for Chapter 5

α = Desired level of Type 1 service.

β = Desired level of Type 2 service.

c_o = Unit overage cost for newsvendor model.

c_u = Unit underage cost for newsvendor model.

D = Random variable corresponding to demand. It is the one-period demand for the newsvendor model and the lead time demand for the (Q, R) model.

EOQ = Economic order quantity.

$F(t)$ = Cumulative distribution function of demand. Values of the standard normal CDF appear in Table A–4 at the end of the book.

$f(t)$ = Probability density function of demand.

$G(Q)$ = Expected one-period cost associated with lot size Q (newsvendor model).

$G(Q, R)$ = Expected average annual cost for the (Q, R) model.

h = Holding cost per unit per unit time.

I = Annual interest rate used to compute holding cost.

K = Setup cost or fixed order cost.

λ = Expected demand rate (units per unit time).

$L(z)$ = Normalized loss function. Used to compute $n(R) = \sigma L(z)$. Tabled values of $L(z)$ appear in Table A–4 at the end of the book.

μ = Mean demand [lead time demand for (Q, R) model].

$n(R)$ = Expected number of stock-outs in the lead time for (Q, R) model.

p = Penalty cost per unit for not satisfying demand.

Q = Lot size or size of the order.

S = Safety stock; $S = R - \lambda\tau$ for (Q, R) model.

SOQ = Service order quantity.

T = Expected cycle time; mean time between placement of successive orders.

τ = Order lead time.

Type 1 service = Proportion of cycles in which all demand is satisfied.

Type 2 service = Proportion of demands satisfied.

Bibliography

Agrawal, N., and S. A. Smith, *Retail Supply Chain Management.* Springer, 2008.

Arrow, K. A.; T. E. Harris; and T. Marschak. "Optimal Inventory Policy." *Econometrica* 19 (1951), pp. 250–72.

Arrow, K. A.; S. Karlin; and H. E. Scarf, eds. *Studies in the Mathematical Theory of Inventory and Production.* Stanford, CA: Stanford University Press, 1958.

Bessler, S. A., and A. F. Veinott, Jr. "Optimal Policy for a Dynamic Multiechelon Inventory Model." *Naval Research Logistics Quarterly* 13 (1966), pp. 355–89.

Brown, R. G. *Statistical Forecasting for Inventory Control.* New York: McGraw-Hill, 1959.

Brown, R. G. *Decision Rules for Inventory Management.* Hinsdale, IL: Dryden Press, 1967.

Clark, A., and H. E. Scarf. "Optimal Policies for a Multiechelon Inventory Problem." *Management Science* 6 (1960), pp. 475–90.

Cohen, M. A., and D. Pekelman. "LIFO Inventory Systems." *Management Science* 24 (1978), pp. 1150–62.

Deuermeyer, B. L., and L. B. Schwarz. "A Model for the Analysis of System Service Level in Warehouse-Retailer Distribution Systems: Identical Retailer Case." In *Multilevel Production/Inventory Control Systems: Theory and Practice,* ed. L. B. Schwarz, pp. 163–94. Amsterdam: North Holland, 1981.

Dvoretsky, A.; J. Kiefer; and J. Wolfowitz. "The Inventory Problem: I. Case of Known Distributions of Demand." *Econometrica* 20 (1952), pp. 187–222.

Ehrhardt, R. "The Power Approximation for Computing (*s*, *S*) Inventory Policies." *Management Science* 25 (1979), pp. 777–86.

Eppen, G., and L. Schrage. "Centralized Ordering Policies in a Multi-Warehouse System with Lead Times and Random Demand." In *Multilevel Production/Inventory Control Systems: Theory and Practice,* ed. L. B. Schwarz, pp. 51–68. Amsterdam: North Holland, 1981.

Federgruen, A., and P. Zipkin. "Computational Issues in an Infinite Horizon Multi-Echelon Inventory Model." *Operations Research* 32 (1984), pp. 818–36.

Fishman, G. S. *Concepts and Methods in Discrete Event Digital Simulation.* New York: John Wiley & Sons, 1973.

Freeland, J. R., and E. L. Porteus. "Evaluating the Effectiveness of a New Method of Computing Approximately Optimal (*s*, *S*) Inventory Policies." *Operations Research* 28 (1980), pp. 353–64.

Graves, S. C. "A Multiechelon Inventory Model for a Repairable Item with One for One Replenishment." *Management Science* 31 (1985), pp. 1247–56.

Graves, S. C.; A. H. G. Rinnooy Kan; and P. Zipkin, eds. *Handbooks in Operations Research and Management Science.* Volume 4, *Logistics of Production and Inventory.* Amsterdam: Elsevier Science Publishers, 1992.

Hadley, G. J., and T. M. Whitin. *Analysis of Inventory Systems.* Englewood Cliffs, NJ: Prentice Hall, 1963.

Hartung, P. "A Simple Style Goods Inventory Model." *Management Science* 19 (1973), pp. 1452–58.

Hausman, W. H., and R. Peterson. "Multiproduct Production Scheduling for Style Goods with Limited Capacity, Forecast Revisions, and Terminal Delivery." *Management Science* 18 (1972), pp. 370–83.

Iglehart, D. L., "Optimality of (*s*, *S*) Inventory Policies in the Infinite Horizon Dynamic Inventory Problem." *Management Science* 9 (1963), pp. 259–67.

Kaplan, R. "A Dynamic Inventory Model with Stochastic Lead Times." *Management Science* 16 (1970), pp. 491–507.

Love, S. F. *Inventory Control.* New York: McGraw-Hill, 1979.

Mohan, S. "EDI's Move to Prime Time Stalled by Cost Perception." *Computerworld* 29 (February 20, 1995), p. 91.

Muckstadt, J. M., and L. J. Thomas. "Are Multiechelon Inventory Models Worth Implementing in Systems with Low Demand Rate Items?" *Management Science* 26 (1980), pp. 483–94.

Murray, G. R., and E. A. Silver. "A Bayesian Analysis of the Style Goods Inventory Problem." *Management Science* 12 (1966), pp. 785–97.

Murray, J. E. "The EDI Explosion." *Purchasing* 118 (February 16, 1995), pp. 28–30.

Nahmias, S. "Inventory Models." In *The Encyclopedia of Computer Science and Technology, Volume 9,* ed. J. Belzer, A. G. Holzman, and A. Kent, pp. 447–83. New York: Marcel Dekker, 1978.

Nahmias, S. "Simple Approximations for a Variety of Dynamic Lead Time Lost-Sales Inventory Models." *Operations Research* 27 (1979), pp. 904–24.

Nahmias, S. "Managing Reparable Item Inventory Systems: A Review." In *Multilevel Production/Inventory Control Systems: Theory and Practice,* ed. L. B. Schwarz, pp. 253–77. Amsterdam: North Holland, 1981.

Nahmias, S. "Perishable Inventory Theory: A Review." *Operations Research* 30 (1982), pp. 680–708.

Nahmias, S., and S. Smith. "Mathematical Models of Retailer Inventory Systems: A Review." In *Perspectives in Operations Management: Essays in Honor of Elwood S. Buffa,* ed. R. K. Sarin. Boston: Kluwer, 1992.

Parr, J. O. "Formula Approximations to Brown's Service Function." *Production and Inventory Management* 13 (1972), pp. 84–86.

Porteus, E. L. *Foundations of Stochastic Inventory Theory*. Stanford, CA: Stanford Business Books, 2002.

Porteus, E. L. "Numerical Comparisons of Inventory Policies for Periodic Review Systems." *Operations Research* 33 (1985), pp. 134–52.

Presutti, V., and R. Trepp. "More Ado about EOQ." *Naval Research Logistics Quarterly* 17 (1970), pp. 243–51.

Ramberg, J. S., and B. W. Schmeiser. "An Approximate Method for Generating Symmetric Random Variables." *Communications of the ACM* 15 (1972), pp. 987–89.

Scarf, H. E. "The Optimality of (s, S) Policies in the Dynamic Inventory Problem." In *Mathematical Methods in the Social Sciences*, ed. K. J. Arrow, S. Karlin, and P. Suppes. Stanford, CA: Stanford University Press, 1960.

Scarf, H. E. "Analytical Techniques in Inventory Theory." In *Multi-Stage Inventory Models and Techniques*, ed. H. E. Scarf, D. M. Gilford, and M. W. Shelly. Stanford, CA: Stanford University Press, 1963.

Schmidt, C. P., and S. Nahmias. "Optimal Policy for a Single Stage Assembly System with Stochastic Demand." *Operations Research* 33 (1985), pp. 1130–45.

Sherbrooke, C. C. "METRIC: Multiechelon Technique for Recoverable Item Control." *Operations Research* 16 (1968), pp. 122–41.

Silver, E. A., and R. Peterson. *Decision Systems for Inventory Management and Production Planning*. 2nd ed. New York: John Wiley & Sons, 1985.

Veinott, A. F. "The Status of Mathematical Inventory Theory." *Management Science* 12 (1966), pp. 745–77.

Veinott, A. F., and H. M. Wagner. "Computing Optimal (*s*, *S*) Inventory Policies." *Management Science* 11 (1965), pp. 525–52.

Whitin, T. M. *The Theory of Inventory Management*. Rev. ed. Princeton, NJ: Princeton University Press, 1957.

Chapter **Six**

Supply Chain Management

"Supply chains cannot tolerate even 24 hours of disruption. So if you lose your place in the supply chain because of wild behavior, you could lose a lot. It would be like pouring cement down one of your oil wells."

—Thomas Friedman

Chapter Overview

Purpose

To understand what a modern supply chain is, how supply chains are organized and managed, and to review the newest developments in this important area.

Key Points

1. *What is a supply chain?* A supply chain is the entire network comprising the activities of a firm that links suppliers, factories, warehouses, stores, and customers. It requires management of goods, money, and information among all the relevant players. While the specific term *supply chain management* (SCM) emerged only in the late 1980s, managing the flow of goods has been an issue since the industrial revolution, and was traditionally simply called logistics.
2. *Supply chain strategy.* For most products it is not possible to have a supply chain that is both low cost and highly responsive. A supply chain strategy must therefore align with the product's positioning in the marketplace. In particular, a product that competes on price must be delivered through a highly efficient supply chain, while for an innovative or high-fashion item it is most important to be able to respond quickly to changes in customer demand.
3. *The role of information in supply chains.* As has been noted, a supply chain involves the transfer of goods, money, and information. Modern supply chain management seeks to eliminate the inefficiencies that arise from poor information flows. One way to ameliorate this problem is vendor-managed inventories, where vendors, rather than retailers, are responsible for keeping inventory on the shelves. Advances in technology have also improved the availability of information in supply chains.
4. *The transportation problem.* The transportation problem is one of the early applications of linear programming. Assume m production facilities (sources) and n demand points (sinks). The unit cost of shipping from every source to every sink is known, and the objective is to determine a shipping plan that satisfies the supply

and demand constraints at minimum cost. The linear programming formulation of the transportation problem has been successfully solved with hundreds of thousands of variables and constraints. A generalization of the transportation problem is the transshipment problem, where intermediate nodes can be used for storage as well as be demand or supply points. Transshipment problems can also be solved with linear programming.

5. *Routing in supply chains.* Consider a delivery truck that must make deliveries to several customers. The objective is to find the optimal sequence of deliveries that minimizes the total distance required. This problem is known as the traveling salesman problem, and turns out to be very difficult to solve optimally. The calculations required to find the optimal solution grow exponentially with the problem size (known mathematically as an NP hard problem). In this section, we present a simple heuristic for obtaining approximate solutions known as the savings method.
6. *Risk pooling.* One key technique for mitigating uncertainty and improving planning is risk pooling. In essence, this principle states that the sum of a number of uncertain variables is inherently less variable than each individual variable, and as such is easier to plan for, schedule, and manage. There are a variety of ways to operationalize risk pooling, including product postponement, regional warehouses, aggregate capacity plans, and flexible capacity.
7. *Designing products for supply chain efficiency.* "Thinking outside the box" has become a cliché. It means looking at a problem in a new way, often not taking constraints at face value. An example of thinking outside the box is postponement in supply chains. The first application of this idea is due to Benetton, a well-known manufacturer of fashion knitwear. Benetton must predict consumers' color preferences in advance of the selling season. Because wool is dyed first and then later weaved into sweaters, the color mix must be decided upon well in advance. If their predictions about consumers' color preferences are wrong (which they invariably are), popular colors would sell out quickly and unpopular colors would sit on the shelves. Their solution was to reverse the order of the weaving and dying operations. Sweaters were woven from undyed wool (gray stock) and then dyed to specific colors as late as possible. This provided Benetton with more time to observe which colors were selling best. Hewlett Packard discovered a similar solution in their printer division. Printers must be configured for local markets due to language and other differences. By producing "gray stock" printers that had all common parts, and then configuring export printers on site in local markets, they were able to delay product differentiation and better balance their inventories.

   Another example of designing products for supply chain efficiency is Ikea. Ikea is a Swedish-based firm that sells inexpensive home furnishings. To reduce costs Ikea designs their furniture to be easily stored directly at the retail outlets. This means that customers can take their purchases with them, thus removing the long delays and customization required by more traditional furniture outlets.
8. *Multilevel distribution systems.* Typically in large systems, stock is stored at multiple locations. Distribution centers (DCs) receive stock from plants and factories and then ship to either smaller local DCs or directly to stores. Some of the advantages of employing DCs include economies of scale, tailoring the mix of product to a particular region or culture, and safety stock reduction via risk pooling.
9. *Incentives in the supply chain.* Consider a clothing designer whose goods are sold at a chain of high-end boutiques. The designer contracts the manufacturing to a firm that subcontracts to a plant in China. The Chinese plant manager is paid

a bonus based on the quantity of output. As a result, he provides incentives for his workers to produce as quickly as they can. However, this results in slipshod quality and a high rate of defective pieces. Ultimately, the designer has to answer to the boutique chain that carries her designs. This is an example of misaligned incentives. How could this problem be ameliorated? One possible answer is to have careful inspection at the plant level, and pay the plant manager based on nondefective items only. What this means is that each player in the supply chain needs to have its incentives aligned with what is best for the system as a whole.

10. *Global supply chain management.* Today, most firms are multinational. Products are designed for, and shipped to, a wide variety of markets around the world. As an example, consider the market for automobiles. Fifty years ago, virtually all the automobiles sold in the United States were produced here. Today, that number is probably closer to 50 percent. Global market forces are shaping the new economy. Vast markets, such as China, are now emerging, and the major industrial powers are vying for a share. Technology, cost considerations, and political and macroeconomic forces have driven globalization. Selling in diverse markets presents special problems for supply chain management.

We take a trip to the grocery store at 10 P.M. to get a jar of peanut butter for our child's lunch the next day. Not only is the store open, but there are a large variety of brands, styles, and sizes of peanut butter available. Americans (and residents of most modern countries) take such things for granted. However, there is a complex myriad of activities that must be carefully coordinated to ensure that the peanut butter will be there when we need it. And this goes for clothes, hardware, and all the other consumer goods we purchase. Do we appreciate the fact that many of these goods are produced and shipped all over the world before they make it to our homes? The logistics of coordinating all the activities that afford us this convenience is the essence of supply chain management.

The term *supply chain management* (SCM) seems to have emerged in the late 1980s and continues to gain interest at an increasing rate. The trade literature abounds with book titles and articles relating to some aspect of SCM. Software and consulting firms specializing in SCM solutions are now commonplace. These companies have grown at remarkable rates and include giants SAP and Oracle, which offer SCM solutions as part of comprehensive information retrieval systems. Although the SCM label is somewhat new, the problems considered are not. Nearly all the material in Chapters 2 to 5 involves SCM. So what is different about SCM? The simple answer is that SCM looks at the problem of managing the flow of goods as an integrated system. Many definitions of SCM have been proposed, and it is instructive to examine some of them. The simplest and most straightforward appears at the Stanford Supply Chain Forum (1999) Web site and is probably due to Hau Lee, the head of the Forum. It is

> Supply chain management deals with the management of materials, information and financial flows in a network consisting of suppliers, manufacturers, distributors, and customers.

While short, this definition is fairly complete. It indicates that it is not only the flow of goods that is important, but the flow of information and money as well.

A definition due to Simchi-Levi et al. (1999, p. 1) that focuses on only the flow of goods is

> Supply chain management is a set of approaches utilized to efficiently integrate suppliers, manufacturers, warehouses, and stores, so that merchandise is produced and distributed at the right quantities, to the right locations, and at the right time, in order to minimize systemwide costs while satisfying service level requirements.

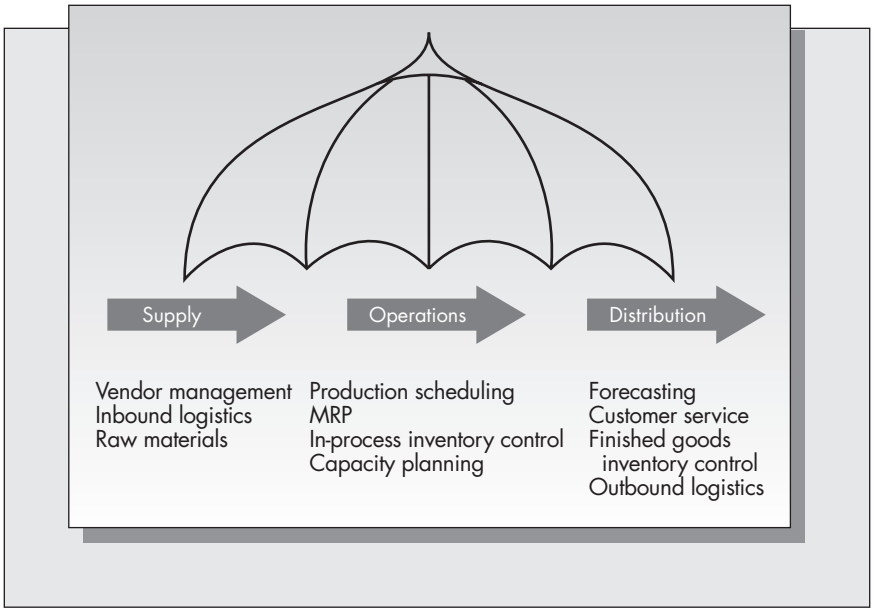


FIGURE 6–1
The supply chain umbrella

Since this definition focuses on the flow of goods, it implies that SCM is relevant only to manufacturing firms. Is SCM also relevant to service organizations? These days it is well recognized that many aspects of SCM do indeed apply to service organizations, although typically these organizations refer to the *value chain* instead of the supply chain.

Most writers agree that "logistics" deals with essentially the same issues as supply chain management. The term logistics originated in the military and concerned problems of moving personnel and materiel in critical times. It was later adopted by business and became a common moniker for professional societies and academic programs. One might argue that although classical logistics treated the same set of problems as SCM, it did not consider the supply chain as an integrated system (Copacino, 1997). A summary of the "umbrella" of activities composing SCM appears in Figure 6–1.

In a sense, one might consider this entire text to deal with some aspect of SCM if one takes the broad view of the field depicted in Figure 6–1. However, most of the chapters focus on specific topics, like inventory control or job shop scheduling. SCM, on the other hand, treats the entire product delivery system from suppliers of raw materials to distribution channels of finished goods. While important and useful in many contexts, simple formulas such as the EOQ are unlikely to shed much light on effective management of complex supply chains.

The Supply Chain as a Strategic Weapon

Where does the supply chain fit into the overall business strategy of the firm? In Chapter 1, it was noted that an important part of the firm's competitive edge is its strategic positioning in the marketplace. Examples of strategic positioning include being the low-cost provider (such as Hyundai automobiles) or the high-quality provider (such as Mercedes-Benz) or exploiting market segmentation to achieve shares of both markets (as does General Motors with different brands aimed at different market segments). The design of the supply chain also reflects a firm's strategic positioning.

In a supply chain, the primary trade-off is between cost and response time. Ground transportation (boat, truck, or rail) is less expensive but air freight is faster. Will deliveries be more reliable if the product is moved via the firm's internal system or would it be better to subcontract the logistics operation to a third party? Third-party logistics (abbreviated 3PL) is becoming more common for the same reason that third-party manufacturing has become so widespread. Companies such as Foxconn, a Taiwanese multinational electronics manufacturer, are able to achieve significant economies of scale by providing manufacturing services to large numbers of firms producing similar products. In this way, it can be less expensive to subcontract manufacturing than to do it in-house. Cost is not the only issue. For example, Apple Inc. outsources to Foxconn because they have long realized that their core competencies lie in aesthetics and innovation and not in manufacturing. Cisco Systems Inc., also outsources to Foxconn. They have transitioned from a manufacturing-focused company to a design-focused firm. They have also embraced a servicization strategy (see Chapter 1).

Key elements of effective supply chain management include the use of information, analytics, and incentives. The Snapshot Application featured in this section highlights Wal-Mart, whose extraordinary success was achieved to some extent because of its sophisticated supply chain strategies and effective use of information, analytics, and incentives.

As previously mentioned, one of the key challenges in supply chain management is aligning incentives. Much of the research is concerned with developing appropriate contracts to align incentives. Modern technologies make this easier. In fact, it has been almost 20 years since Gibson, the then chairman of Manugistics Inc., stated in a *Wall Street Journal* article that "strategic partnering used to mean stealing revenue or pushing cost onto someone else in the supply chain. You are a pig at the trough if you view it that way. With technology, there are so many efficiencies that can be shared." (*WSJ*, April 12, 1996)

In this chapter we provide the reader with an appreciation of the most important issues encountered in managing complex supply chains. In addition, we present a sampling of the mathematical models (analytics) used in SCM analysis.

6.1 SUPPLY CHAIN STRATEGY

As just discussed, one of the key trade-offs in supply chain design is between cost and speed. It is important to recognize that the best choices on this dimension depend on the product being produced. Fisher (1997) presents a simple framework for this choice. He breaks products into two categories: *functional* or *innovative*, and supply chains into two types: *efficient* versus *responsive*. He argues that functional products should be produced using efficient supply chains while innovative products require responsive supply chains.

A functional product is a commodity-type product that competes primarily on cost. A canonical functional product is Campbell's Soup. Only 5 percent of the products they produce are new each year and most products have been in the market for many years. It has highly predictable demand and long life cycles (particularly relative to lead times). This type of product is best produced using an efficient supply chain where the primary focus is on costs, which include inventory holding, transportation, handling, and manufacturing costs. Large manufacturing batches achieve scale economies, and minimum order quantities reduce handling/order processing costs. Supply chain performance is evaluated using traditional cost measures, such as inventory turns, and factory and transportation utilization.

Snapshot Application

WAL-MART WINS WITH SOLID SUPPLY CHAIN MANAGEMENT

Although there are many examples of companies winning (or losing) because of either good or bad supply chain strategies, perhaps none is more dramatic than the stories of Wal-Mart and Kmart. Both companies were founded in the same year, 1962. In only a few years, Kmart had become a household term, while Wal-Mart was largely unknown except for a few communities in the South. Wal-Mart stores were typically located in rural areas, so they rarely competed head-on with the larger, better-known chains like Sears and Kmart, most of whose stores were located in large cities and surrounding suburbs. In 1987 Kmart's sales were almost twice those of Wal-Mart (about $25 billion annually versus about $15 billion annually). By 1990 Wal-Mart had overtaken Kmart and in 1994 Wal-Mart's annual sales almost tripled Kmart's (about $80 billion versus about $27 billion)! Wal-Mart is now the largest discount retailer in the United States, surpassing Sears as well as Kmart, who have now merged. What could have possibly accounted for this dramatic turn-around?

While one can point to several factors leading to Wal-Mart's success, there is no question that the firm's emphasis on solid supply chain management was one of the most important. According to Duff and Ortega (1995) in comparing the management decisions of Kmart's former CEO, Joseph Antonini, and Wal-Mart's Sam Walton:

> When Mr. Antonini took the reins of Kmart in 1987, he had his hands full . . . Also, his predecessors neglected to implement the sophisticated computer systems that were helping Wal-Mart track and replenish its merchandise swiftly and efficiently.
>
> A self-promoter with a boisterous voice and wide smile, Mr. Antonini invested heavily in national television campaigns and glamorous representatives such as Jaclyn Smith . . . Mr. Walton avoided publicity. And instead of marketing, he became obsessed with operations. He invested tens of millions of dollars in a companywide computer system linking cash registers to headquarters, enabling him to quickly restock goods selling off the shelves. He also invested heavily in trucks and distribution centers. Besides enhancing his control, these moves sharply reduced costs.
>
> Mr. Antonini tried bolstering growth by overseeing the purchase of other types of retailers: the Sports Authority sporting goods chain, OfficeMax office supply stores, Borders bookstores, and Pace Membership Warehouse clubs . . .
>
> But the least visible difference between Wal-Mart and Kmart was beginning to matter a lot. Wal-Mart's incredibly sophisticated distribution inventory and scanner systems meant that customers almost never encountered depleted shelves or price-check delays at the cash register.
>
> The halls of Kmart, meanwhile, were filled with distribution horror stories. Joseph R. Thomas, who oversaw distribution, said that in retrospect, he should have smelled trouble when he found warehouses stuffed full of merchandise on December 15, the height of the Christmas season.

Wal-Mart is the world's largest public corporation, according to the Fortune Global 500 list in 2014, the biggest private employer in the world with over two million employees, and the largest retailer in the world. It has over 11,000 stores in 27 countries, under a total of 55 different names, and continues to expand. Its Chinese operations are growing particularly quickly with 100 stores planned for the next few years and a projected market of 2,000 stores.

Wal-Mart is known for both its sophisticated business analytics and advanced information systems. They have been a leader in vendor-managed inventory, sharing data with suppliers for better decisions, implementing effective cross-docking operations, and lately they have been proponents of sustainable supply chain management. Although much of the information about their systems is proprietary, it is clear that they invest heavily in business intelligence.

There are some important lessons here. When problems arose, Kmart's management responded by putting money into marketing and acquisitions, while Wal-Mart invested in better stores and a better logistics system for making sure that shelves were stocked with what customers wanted. One of the key themes of this book is that operations matter, and indeed effective operations and supply chain management can be a strategic weapon for a firm. While innovative designs can delay the need for effective supply chain management, sooner or later inefficiencies catch up with a firm.

The story of Kmart and Wal-Mart is reminiscent of the story of the American and Japanese auto industries. In the 1950s and 1960s when American auto makers had a virtual monopoly on much of the world's auto market, where did the Big Three put their resources? Into producing better engineered cars at lower cost? No. Into slick marketing campaigns and annual cosmetic makeovers. The Japanese, on the other hand, invested in the latest automotive technology and sophisticated logistics systems such as just-in-time. The rest, as they say, is history.

An innovative product is one that typically has high margin but highly variable demand and short life cycle. An example is Sport Obermeyer's fashion skiwear. Each year, 95 percent of the products are completely new designs and demand forecasts err by as much as 200 percent. There is a short retail season, and cost concerns focus around inventory obsolescence and lost sales (see the Snapshot Application in Chapter 2). This type of product is best produced using a market responsive supply chain, where a premium is paid for flexibility, including faster transportation, lower transportation and factory utilization, and smaller batches. Supply chain performance is best measured not with traditional cost measures but by recognizing that the opportunity costs of lost sales and poor service are key metrics to consider.

Cost and speed are both key measures of supply chain effectiveness. However, Lee (2004) argues that they are not the whole story. He proposes the "Triple A" supply chain, which is *agile*, *adaptable*, and *aligned*. Here agility is defined as the ability to respond to short-term changes in demand or supply quickly. Adaptability is defined as the ability to adjust the supply chain design to accommodate market changes. Finally, alignment relates to the existence of incentives for supply chain partners to improve performance of the entire chain. Further details may be found within his article but clearly these are desirable traits for a supply chain. Like any desirable feature they are not costless, and the cost to implement must be balanced against strategic necessity.

6.2 THE ROLE OF INFORMATION IN THE SUPPLY CHAIN

How often have we heard it said that we are living in the "information age"? The availability of information is increasing at an exponential rate. New sources of information in the form of academic and trade journals, magazines, newsletters, and so on are introduced every day. The explosion of information availability on the Web has been truly phenomenal. Web searches are now the first place many go for information on almost anything.

Knowledge is power. In supply chains, information is power. It provides the decision maker the power to get ahead of the competition, the power to run a business smoothly and efficiently, and the power to succeed in an ever more complex environment. Information plays a key role in the management of the supply chain. As we saw in the earlier chapters of this book, many aspects of operations planning start with the forecast of sales and build a plan for manufacturing or inventory replenishment from that forecast. Forecasts, of course, are based on information.

An excellent example of the role of information in supply chains is the Harvard Business School cases Barilla SpA (A and B) (1994) written by Jan Hammond. In many operations management courses, these cases provide an introduction to the role of information in supply chains. Barilla is an Italian company that specializes in the production of pasta. In the late 1980s, Barilla's head of logistics tried to introduce a new approach to dealing with distributors, which he called just-in-time distribution (JITD). Briefly, the idea was to obtain sales data directly from the distributors (Barilla's customers) and use these data to allow Barilla to determine when and how large the deliveries should be. At that time, Barilla was operating in the traditional fashion. Distributors would independently place weekly orders based on standard reorder point methods. This led to wide swings in the demand on Barilla's factories due to the "bullwhip effect," which is discussed in detail next. The JITD idea met with staunch resistance, both within and outside Barilla.

Snapshot Application

ANHEUSER-BUSCH RE-ENGINEERS THEIR SUPPLY CHAIN

Anheuser-Busch is the leading American brewer, holding close to a half share of U.S. beer sales, and producing iconic brands such as Budweiser and Bud Light. In the late 1990s they undertook a major reorganization of their supply chain (John & Willis, 1998).They found that less than 10 percent of their volume accounted for around 80 percent of the brand and package combinations. Management decided to switch to focused production. Some breweries were dedicated for the efficient production of large volume items, such as Bud and Bud Light. Others were responsible for producing lower volume niche products, such as wheat beers or products targeted at holidays. (Refer to the discussion of the focused factory in Chapter 1.)

By aligning supply chain metrics to the type of product being produced, their decisions were in line with Fisher's recommendations for matching product type with supply chain type. In particular, they made sure that their high-volume breweries were measured by efficiency metrics, such as utilization and volume of output; whereas, their niche product breweries were measured by how effectively they met demand. Previously, benchmarking had been across all breweries, which did not create the correct incentives for either the high volume or niche products.

In addition to focused facilities and supply chains, Anheuser-Busch also undertook a number of initiatives in line with topics covered later in this chapter. In particular, a mixed integer programming model (i.e., supply chain analytics) was used to identify the optimal number and location of distribution points in the supply chain. Contracts with a thousand different trucking companies (across the whole network) were largely replaced by a single contract with a single dedicated carrier, decreasing the complexity of the network. Inventory pooling was used to reduce both risk and cost. Finally, replenishment agreements were renegotiated and vendor-managed inventory agreements were put in place, which improved the alignment of the incentives in the supply chain.

The Barilla sales and marketing organizations, in particular, were most threatened by the proposed changes. Distributors were concerned that their prerogatives would be compromised.

Without going into the details [which can be found in Barilla SpA (B)], management eventually did prevail, and the JITD system was implemented with several of Barilla's largest distributors. The results were striking. Delivery reliability improved, and the variance of orders placed on the factory were substantially reduced. The program proved a win-win for Barilla and its customers.

Barilla's success with its JITD program is one example of what today is known as vendor-managed inventory (VMI). VMI programs have been the mainstay of many successful retailers in the United States. For example, Procter & Gamble has assumed responsibility for keeping track of inventories for several of its major clients, such as Wal-Mart. Years ago, grocery store managements were responsible for keeping track of their inventories. Today, it is commonplace to see, in our local grocery stores, folks checking shelves who are not store employees but employees of the manufacturers. With VMI programs, it becomes the manufacturer's responsibility to keep the shelves stocked. While stock-outs certainly hurt the retailer, they hurt the manufacturer even more. When a product stocks out, customers typically will substitute another, so the store makes the sale anyway. It is the manufacturer that really suffers the penalty of lost sales, so the manufacturer has a strong incentive to keep shelves stocked.

The Bullwhip Effect

Barilla's experience prior to the implementation of their VMI program is one example of the bullwhip effect. It has become a topic of considerable interest among both practitioners and academics alike. The history of bullwhip appears to be the following.

Executives at Procter & Gamble (P&G) were studying replenishment patterns for one of their best-selling products: Pampers disposable diapers. They were surprised to see that the orders placed by distributors had much more variation than sales at retail stores. Furthermore, orders of materials to suppliers had even greater variability. Demand for diapers is pretty steady, so one would assume that variance would be low in the entire supply chain. However, this was clearly not the case. P&G coined the term "bullwhip" effect for this phenomenon. It also has been referred to as the "whiplash" or "whipsaw" effect.

This phenomenon was observed by other firms as well. HP experienced the bullwhip effect in patterns of sales of its printers. Orders placed by a reseller exhibited wider swings than retail sales, and orders placed by the printer division to the company's integrated circuit division had even wider swings. Figure 6–2 shows how variance increases as one moves up the supply chain.

Where does the bullwhip effect come from? One cause stems from a phenomenon discussed in Chapter 8 in MRP systems. Consider a basic two-level MRP system in which the pattern of final demand is fixed and constant. Since items are produced in batches of size Q, the demand pattern one level down is much spikier. The tendency for lower levels in the supply chain to batch orders in this way is one of the root causes of the bullwhip effect.

High levels of inventories, caused in part by the bullwhip effect, are common in the grocery industry as well. To address this problem, the industry has adopted the efficient consumer response (ECR) initiative (see, for example, Crawford, 1994). The total food delivery supply chain, from the point at which products leave the manufacturer to the point when they are stocked on the shelves at the retailer, has an average of over 100 days of supply. The stated goal of the ECR initiative is to save $30 billion annually by streamlining food delivery logistics.

Another example of the bullwhip effect is the popular "beer game" due to Sterman (1989). Participants play the roles of retailers, wholesalers, and manufacturers of beer. Communication among participants is prohibited: each player must make ordering

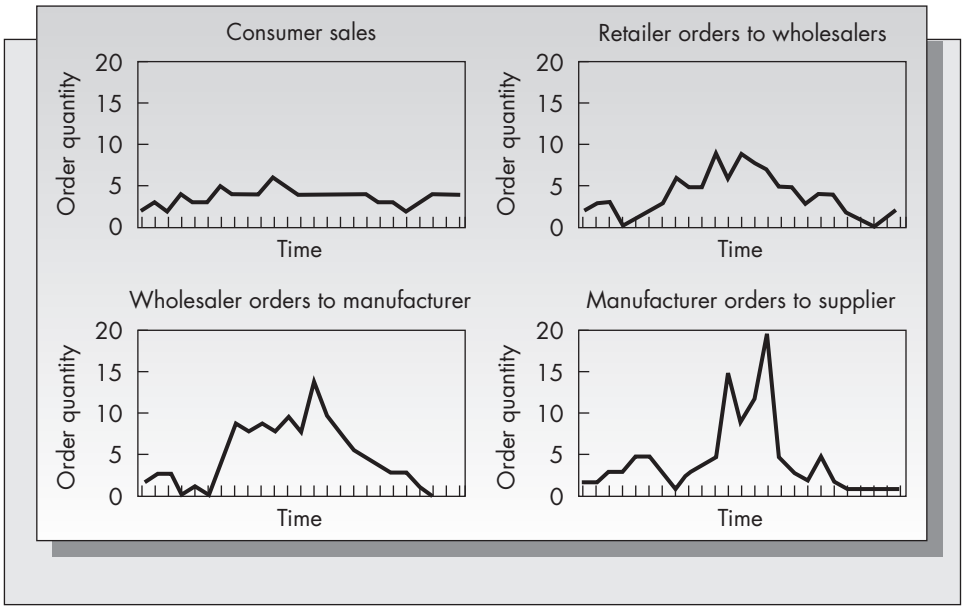


FIGURE 6–2
Increasing variability of orders up the supply chain

Source: H. L. Lee, P. Padmanabhan, and S. Whang, 1997.

decisions based on what is demanded from the downstream player only. What one observes are wild swings in orders placed downstream even when the original demands are fairly stable. This is a consequence of the bullwhip effect.

The telescoping variation in the demand patterns in a supply chain results in a planning nightmare for many industries. What can be done to alleviate these effects? First, we need to understand the causes of this phenomenon. According to Lee, Padmanabhan and Whang (1997), there are four primary causes of the bullwhip effect:

- Demand forecast updating
- Order batching
- Price fluctuations
- Shortage gaming

We consider each of these effects in turn. *Demand forecasts* at each stage of the supply chain are the result of demands observed one level downstream (as in the beer game). Only at the final stage of the chain (the retailer) are consumer demands observed directly. When each individual in a serial supply chain determines his or her demand forecasts individually, the bullwhip effect results. The retailer builds in safety stocks to protect against uncertainty in consumer demand. These safety stocks cause the retailers' orders to have greater variance than the consumers'. The distributor observes these swings in the orders of the retailer and builds in even larger safety stocks, and so on.

Order batching is the phenomenon we saw in MRP systems that results in smooth demand patterns being translated to spiky demand patterns at lower levels of the product structure. The natural tendency to save fixed costs by ordering less frequently (which is the idea behind the EOQ formula) gives rise to order batching. Given the cost structure at each level, this is certainly a reasonable response.

When prices *fluctuate*, there is a speculative motive for holding inventories. (This was first discussed in the inventory management context by Arrow, 1958, and relates to the motivations for holding cash postulated by the economist John Maynard Keynes.) In the food industry, the majority of transactions from the manufacturer to distributors are made under a "forward buy" arrangement. This refers to the practice of buying in advance of need because of an attractive price offered by manufacturers. Such practices also contribute to the bullwhip effect. Large orders are placed when promotions are offered.

Shortage gaming occurs when product is in short supply and manufacturers place customers on allocation. When customers find out that they may not get all the units they request, they simply inflate orders to make up for the anticipated shortfall. For example, if a computer manufacturer expects to receive only half a request for a CPU in short supply, they can simply double the size of their order. If anticipated demands don't materialize, orders can be canceled. The result is that the manufacturer gets an inflated picture of the real demand for the product. This could have dire consequences for a manufacturer placing large amounts of capital into capacity expansion based on "phantom" demands.

Clearly, the bullwhip effect is not the result of poor planning or irrational behavior on the part of players in the supply chain. Each individual acts to optimize his or her position. What, then, can be done to alleviate the situation? There are several potential remedies, which we will discuss. However, these remedies must take into account that people behave selfishly. The motivation for bullwhip-inducing behavior must be eliminated.

Lee, Padmanabhan, and Whang (1997) recommend four initiatives. They are

1. Information sharing
2. Channel alignment
3. Price stabilization
4. Discouragement of shortage gaming

1. *Information sharing* means that all parties involved share information on point-of-sale (POS) data and base forecasts on these data only. Information sharing can be accomplished by several techniques, one of which is EDI (electronic data interchange), discussed in detail later. The trend toward information sharing is beginning to take hold. Computer manufacturers, for example, are now requiring sell-through data from resellers (these are data on the withdrawal of stocks from the central warehouse). At some point, we expect to see manufacturers linked directly to sources of POS data.

2. *Channel alignment* is the coordination of efforts in the form of pricing, transportation, inventory planning, and ownership between upstream and downstream sites in the supply chain. One of the tendencies that defeats channel alignment is order batching. As we noted earlier, fixed costs motivate order batching. Reducing fixed costs will result in smaller order sizes. One of the important components of fixed costs is the paperwork required to process an order. With new technologies, such as EDI, these costs could be reduced substantially. Another factor motivating large batches is transportation economies of scale. It is cheaper on a per-unit basis to order a full truckload than a partial one. If manufacturers allow customers to order an assortment of items in a single truckload, the motivation to order large batches of single items is reduced. Another trend encouraging small batch ordering is the outsourcing of logistics to third parties. Logistics companies can consolidate loads from multiple suppliers. Outsourcing of logistics (like outsourcing of manufacturing) is expanding rapidly.

3. Pricing promotions motivate customers to buy in large batches and store items for future use; this is very common in the grocery industry. By stabilizing pricing, sales patterns will have less variation. In retailing, the effect of stable pricing is evident by comparing sales patterns at retailers such as Macy's that run frequent promotions and warehouse stores such as Costco that offer everyday low pricing. The warehouse stores experience steadier sales than do the department stores. In fact, for many department stores, promotional sales account for most of their business. Major grocery manufacturers such as P&G, Kraft, and Pillsbury are moving toward a value-pricing strategy and away from promotional pricing for the same reason.

4. One way to minimize excessive orders as a result of shortage gaming is to allocate based on past sales records rather than on orders. This will reduce the tendency of customers to exaggerate orders. Several companies (including General Motors) are moving in this direction.

The bullwhip effect has been observed in a variety of settings and is also theoretically predicted to occur by so-called systems dynamics models (Sterman, 1989). However, a well-designed production system should work to smooth variability, particularly that associated with seasonality. Cachon et al. (2007) perform an empirical study on demand variability across a wide range of U.S. firms. They find that "industries with seasonality tend to smooth production relative to

demand, whereas industries without seasonality tend to amplify." They also show that retailers, in particular, tend to perform a smoothing rather than amplifying effect on demand variability.

The study by Cachon et al. is primarily at the industry level, whereas Bray and Mendelson (2012) perform a firm level study that finds "65 percent of firms exhibit a positive overall bullwhip." They confirm the findings of Cachon et al. that in highly seasonal settings the production smoothing effect outweighs the inherent uncertainty amplification caused by demand shocks. Interestingly, they also show that the bullwhip effect is significantly reduced for data after 1995. While much of this can be explained by improved information systems, perhaps the increasing awareness of the bullwhip effect is also a contributing factor. Bray and Mendelson (2012) empirically verify the impressive bullwhip reduction achieved by Caterpillar Inc., which is known to have focused its efforts on initiatives similar to the four outlined above, for bullwhip reduction.

In summary, the bullwhip effect is the consequence of individual agents in the supply chain acting in their own best interests. To mitigate bullwhip effects, several changes must be made. Incentives must be put in place to reduce demand forecast errors, reduce excessive order sizes in allocation situations, and encourage information sharing and system alignment. As these initiatives become policy, everyone, especially the consumer, will benefit from reduced costs and improved supply chain efficiency.

Electronic Commerce

Electronic commerce, or simply e-commerce, is a catch-all term for a wide range of methods for effecting business transactions without the traditional paper-based systems. It includes EDI, e-mail, electronic funds transfers, electronic publishing, image processing, electronic bulletin boards, shared databases, and all manner of Internet-based business systems (Handfield and Nichols, 1999). The point-of-sale barcode system, now ubiquitous in supermarkets and retailers, is another type of e-commerce.

Barcodes were first introduced as a way to speed up supermarket checkouts. The first retail scanning of a barcode appears to be a pack of Wrigley chewing gum in June 1974 (Varchaver, 2004). However, a powerful advantage of barcodes is their role in information gathering. Assuming the check-out operator scans the products correctly, the retailer now has accurate information on sales. The reason for the caveat in the previous sentence is that a major source of data error is operators who scan, for example, one can of mushroom soup and hit times three on the till, rather than separately scanning the mushroom, tomato, and cream of asparagus soups that the customer has actually purchased. Retailers have put significant educational effort into ensuring correct scanning techniques by staff. Further, with the introduction of customer loyalty cards, retailers now have information on individual purchasing behavior and can target promotions to the individual customer.

The internet has fundamentally changed supply chain management. First, it allows easy transmission of information to users within the firms and to customers and trading partners. This includes point-of-sale demand information, purchase orders, and inventory status information. Originally, such information sharing was through electronic data interchange (EDI) systems that allow the transmission of standard business documents in a predetermined format between computers. Dedicated EDI systems have now been mostly replaced by internet-based systems. Cloud computing has also facilitated information sharing and enterprise systems. SAP and Oracle, for example, offer cloud

computing options for their systems. One of the key advantages of such systems from a supply chain management perspective is speed. By transmitting information quickly, lead times are reduced. As noted in Chapter 5, it is the uncertainty of demand over the replenishment lead time that results in the need to carry safety stock.

In addition to facilitating information sharing, the internet has also provided both business to consumer (B2C) and business to business (B2B) opportunities. The rise of B2C commerce was rocky with the now infamous dot-com bust in the early 21st century. Many internet-only companies failed at that time; however, Amazon.com, which is now the world's largest online retailer, has thrived. Amazon started as a discount bookseller. By selling from a single location, it could reap the benefits of stock centralization (to be discussed in Section 6.9) and avoid the costly capital investment in brick and mortar stores.

The model for internet-based retailing is essentially the same as that for catalog shopping. Catalog retailers have been around for many years, with Lands' End and L.L.Bean being perhaps the most successful. However, catalogs need to be developed, printed, and mailed regularly, which is quite expensive, especially considering that most catalogs are thrown away. For this reason, the internet has a significant advantage over the traditional mail-order catalog business, and both Lands' End and L.L.Bean have put significant investment into their websites. Further, the move away from paper catalogs has allowed them to reach consumers around the world. Lands' End boasts that they "proudly ship to over 170 countries."

Along with the growth of B2C commerce there has also been a less visible growth of B2B commerce. Hundreds, if not thousands, of firms offer internet-based supply chain solutions. Many tailor their products to specific industry segments, such as PrintingForLess.com (printing) and ShowMeTheParts.com (automotive aftermarket). There are also intermediaries, such as Li & Fung Lmt., who specialize in third party sourcing. While Li & Fung started as a traditional trading company, today they provide access to a network of over 15,000 suppliers in more than 40 economies. This would not be possible without e-commerce.

It is interesting to speculate what effect 3D printing will have on e-commerce. 3D printing, or additive manufacturing, allows the construction of three dimensional objects by a "printer" just as regular printers allow for the production of two dimensional text. Such printers are currently used widely in the production of manufacturing prototypes but less widely in consumer goods. The range of materials that may be printed is limited but increasing. Examples of products produced by 3D printing include custom jewelry and home decor, such as lampshades. In time, instead of buying items from the internet (or store) we may simply buy a design and print it up at home or in our local print shop. Such a development could drastically simplify supply chain management!

RFID Technology

Radio frequency identification (RFID) tags are an emerging technology that will change the way information is stored and transmitted in a supply chain. Unlike barcodes, which are the same across all items of the same stock-keeping unit (SKU), RFID tags can contain item specific information, such as farm of origin (for produce) or date of production.

RFID tags were invented in 1973, but are only now becoming commercially viable. They are microchips powered by batteries (active tags) or radio signals (passive tags).

The passive device is smaller and cheaper and is likely to emerge as the device of choice for inventory management applications. Passive tags receive power from the readers (at which time they are "woken up") and transmit a unique code. Passive tags can only be read at close distances but they provide a simple means of electronic identification.

Passive RFID tags cost between $.07 and $.15 each, which makes them too expensive for small grocery items, such as gum, but practical for larger items, such as jeans. JCPenney announced in 2012 that they were moving to 100 percent RFID tagging of all items in their stores. While their roll out was postponed, at the time of this writing they have moved to item-level tagging in bras, footwear, fashion jewelry, and men's and women's denim. Wal-Mart and American Apparel have also moved to tagging clothing. One of the main advantages of such tagging is in drastically speeding up and improving accuracy in stock-taking, thereby reducing stock-outs and increasing sales. For example, in a pilot program at American Apparel Inc. in 2007, sales increased by 14.3 percent (Bustillo, 2010). When all items are tagged retailers will also see savings in the checkout process because RFID readers can replace the need for item by item barcode scanning.

Beyond retail item tagging, other applications of RFID technology include (1) EZ Pass for paying bridge or highway tolls, (2) tagging of luggage on flights, and (3) tagging of cargo containers at most of the world's ports. As the cost of RFID tags declines, we will see a much wider range of applications in the context of supply chain management. Reconciling shipments against bills-of-lading (i.e., cargo records) or packing and pick lists can be performed quickly and accurately electronically, eliminating the need to perform these functions manually.

Of course, RFID technology has far broader implications than its application to supply chains. For example, Solusat, the Mexican distributor of the VeriChip—a rice-sized microchip that is injected beneath the skin—is marketing its device as an emergency identification. The interest in Mexico for this product is a consequence of the more than 10,000 Mexican children abducted annually. VeriChip manufacturer, Applied Digital Solutions, said it plans to roll out the VeriKid service in other countries, including the United States (Scheeres, 2003). Such applications of RFID technology potentially have enormous benefits, but also can be used in ways to threaten our privacy.

RFID technology is quickly making inroads into many other industries. Potential applications discussed in a recent book on the subject (Schuster, Allen, and Brock, 2007) include the following:

- *Warehousing*. By tagging inventory stored in a warehouse, improved tracking of items and order fulfillment can lead to significant improvements in several measures of customer service.
- *Maintenance*. Maintenance programs require keeping track of the location, suitability, and condition of spare parts. RFID tags can provide significant improvements in monitoring the installed base of parts and components.
- *Pharmaceuticals*. Tracking and tracing pharmaceuticals can help ameliorate the problems of counterfeit drugs and theft in the industry. These problems have been estimated to run to $30 billion annually worldwide.
- *Medical devices*. RFID technology can provide continuous access to the identity, location, and state of medical devices. This has the potential to significantly improve patient care.

- *Animal tracking*. The appearance of mad cow disease (BSE) in the United Kingdom, Canada, and the United States raised an alarm worldwide about food safety. Being able to track individual livestock can be invaluable when trying to trace the source of problems.
- *Shelf-life tracking*. We are all familiar with expiration dates on foods such as dairy products, packaged meats, fish, and canned goods. RFID technology provides the opportunity for "dynamic" shelf-life tracking, that is, updating the shelf-life indicator to take into account environmental conditions such as temperature.
- *Retailing*. It is common for expensive retail items, such as leather jackets, to be tagged with a transmitter that sounds an alarm if removed from the store. Inexpensive RFID chips provide the opportunity to tag almost all retail items, virtually eliminating theft altogether.
- *Defense*. Logistic support has always been a key to securing victory in battle, in both modern and ancient times. It is currently estimated that the U.S. Department of Defense manages an inventory valued at $67 billion. Keeping track of this inventory is obviously a top priority, and RFID technology can go a long way toward accomplishing this goal.

Problems for Sections 6.1 and 6.2

1. Name two products you are familiar with, that are *not* discussed in this chapter, one of which is best classed as functional and the other as innovative.
2. What product characteristics would be necessary for a product to be able to have a supply chain that is simultaneously both efficient and responsive?
3. What is the bullwhip effect? What is the origin of the term?
4. Do you think that eliminating intermediaries would always get rid of the bullwhip effect? Under what circumstances might it not work?
5. Discuss the four initiatives recommended for ameliorating the bullwhip effect in supply chains.
6. Amazon.com is a purely e-commerce company, yet Barnes & Noble maintains both retail bookstores and a significant internet business. What are the advantages and disadvantages of Barnes and Noble's strategy (as contrasted to Amazon's)?
7. What are the key benefits that the internet has provided in aiding efficiency or effectiveness in supply chain management?
8. Give three examples of revenue or cost benefits made possible by RFID tags when contrasted to barcodes.

6.3 THE TRANSPORTATION PROBLEM

With good information comes the opportunity to optimize within the supply chain. The transportation problem is a mathematical model for optimally scheduling the flow of goods from production facilities to distribution centers. Assume that a fixed amount of product must be transported from a group of sources (plants) to a group of sinks (warehouses). The unit cost of transporting from each source to each sink is assumed

to be known. The goal is to find the optimal flow paths and the amounts to be shipped on those paths to minimize the total cost of all shipments.

The transportation problem can be viewed as a prototype supply chain problem. Although most real-world problems involving the shipment of goods are more complex, the model provides an illustration of the issues and methods one would encounter in practice.

Example *6.1*

The Pear Tablet Corporation produces several types of tablet computers. In 2013, Pear produced tablets with capacities from gigabytes (GB), all 10-inch display. The most popular product is the 64 GB tablet. Pear produces the tablets in three plants located in Sunnyvale, California; Dublin, Ireland; and Bangkok, Thailand. Periodically, shipments are made from these three production facilities to four distribution warehouses located in the United States in Amarillo, Texas; Teaneck, New Jersey; Chicago, Illinois; and Sioux Falls, South Dakota. Over the next month, it has been determined that these warehouses should receive the following proportions of the company's total production of the 64 GB tablets.

Warehouse	Percentage of Total Production
Amarillo	31
Teaneck	30
Chicago	18
Sioux Falls	21

The production quantities at the factories in the next month are expected to be (in thousands of units)

Plant	Anticipated Production (in 1,000s of units)
Sunnyvale	45
Dublin	120
Bangkok	95

Since the total production at the three plants is 260 units, the amounts shipped to the four warehouses will be (rounded to the nearest unit)

Warehouse	Total Shipment Quantity (1,000s)
Amarillo	80
Teaneck	78
Chicago	47
Sioux Falls	55

While the shipping cost may be lower between certain plants and distribution centers, Pear has established shipping routes between every plant and every warehouse. This is in case of unforeseen problems such as a forced shutdown at a plant, unanticipated swings in regional

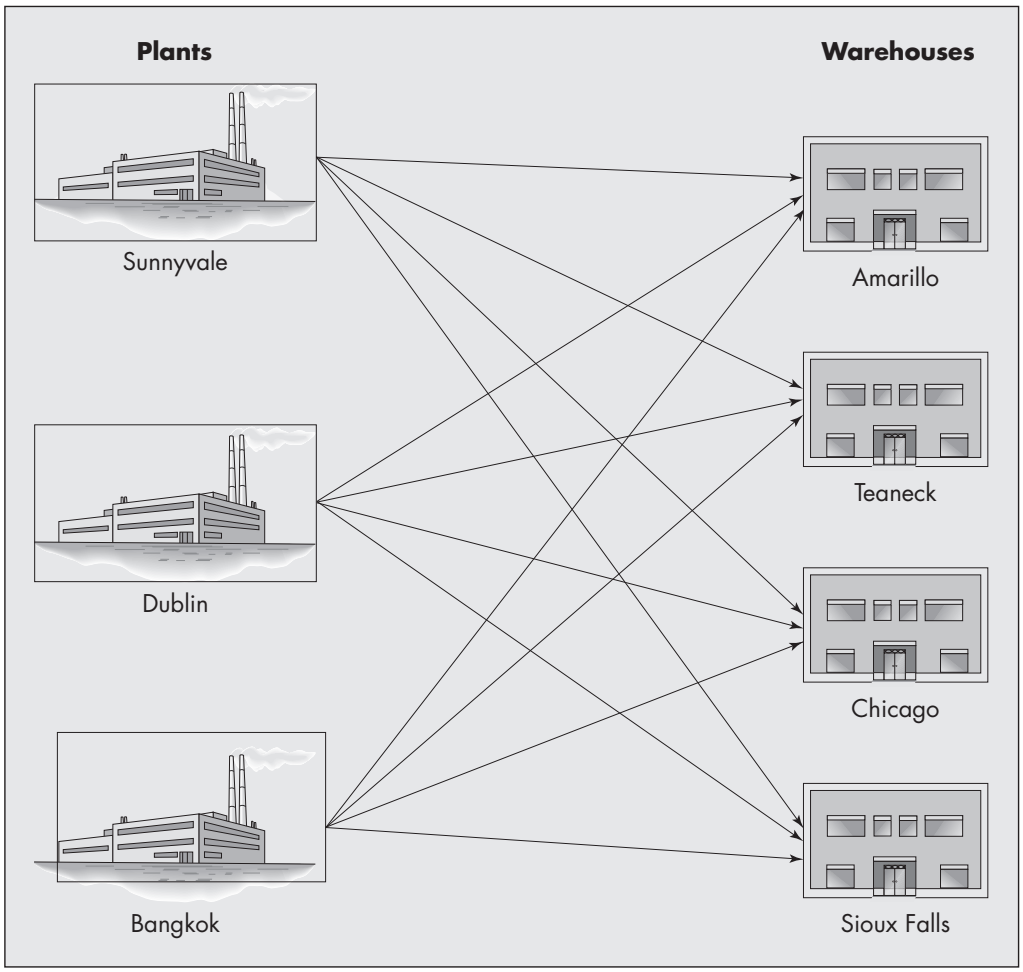
demands, or poor weather along some routes. The unit costs for shipping 1,000 units from each plant to each warehouse is given in the following table.

Shipping Costs per 1,000 Units in $

		TO			
		Amarillo	**Teaneck**	**Chicago**	**Sioux Falls**
FROM	**Sunnyvale**	250	420	380	280
	Dublin	1,280	990	1,440	1,520
	Bangkok	1,550	1,420	1,660	1,730

The goal is to determine a pattern of shipping that minimizes the total transportation cost from plants to warehouses. The network representation of Pear's distribution problem appears in Figure 6–3.

FIGURE 6–3
Pear Tablet transportation problem



Several heuristics for solving transportation problems have been proposed, such as a greedy heuristic that allocates capacity first to the cheapest options. However, it is unlikely that anyone with a real problem would use a heuristic, since optimal solutions can be found efficiently by linear programming. In fact, because of the special structure of the transportation problem, today's specialized codes can solve problems with millions of variables.

Let m be the number of sources and n the number of sinks. (In Example 6.1, $m = 3$ and $n = 4$.) Recall the definition of the decision variables from Section 6.1:

$$x_{ij} = \text{flow from source } i \text{ to sink } j \text{ for } 1 \le i \le m \text{ and } 1 \le j \le n,$$

and define c_{ij} as the cost of shipping one unit from i to j. It follows that the total cost of making all shipments is

$$\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}$$

For the case of Pear Company, described in Example 6.1, the objective function is

$$250x_{11} + 420x_{12} + 380x_{13} + 280x_{14} + \cdots + 1{,}730x_{34}.$$

Since many routes are obviously not economical, it is likely that many of the decision variables will equal zero at the optimal solution.

The constraints are designed to ensure that the total amount shipped out of each source equals the amount available at that source, and the amount shipped into any sink equals the amount required at that sink. Since there are m sources and n sinks, there are a total of $m + n$ constraints (excluding nonnegativity constraints). Let a_i be the total amount to be shipped out of source i and b_j the total amount to be shipped into sink j. Then linear programming constraints may be written:

$$\sum_{j=1}^{n} x_{ij} = a_i \qquad \text{for } 1 \le i \le m$$

$$\sum_{i=1}^{m} x_{ij} = b_j \qquad \text{for } 1 \le j \le n$$

$$x_{ij} \ge 0 \qquad \text{for } 1 \le i \le m \quad \text{and} \quad 1 \le j \le n.$$

For the Pear Tablet Company problem, we obtain the following seven constraints:

$$\begin{aligned}
x_{11} + x_{12} + x_{13} + x_{14} &= 45 && \text{(shipments out of Sunnyvale)} \\
x_{21} + x_{22} + x_{23} + x_{24} &= 120 && \text{(shipments out of Dublin)} \\
x_{31} + x_{32} + x_{33} + x_{34} &= 95 && \text{(shipments out of Bangkok)} \\
x_{11} + x_{21} + x_{31} &= 80 && \text{(shipments into Amarillo)} \\
x_{12} + x_{22} + x_{32} &= 78 && \text{(shipments into Teaneck)} \\
x_{13} + x_{23} + x_{33} &= 47 && \text{(shipments into Chicago)} \\
x_{14} + x_{24} + x_{34} &= 55 && \text{(shipments into Sioux Falls)}
\end{aligned}$$

and the nonnegativity constraints required in linear programming:

$$x_{ij} \ge 0 \qquad \text{for } 1 \le i \le 3 \quad \text{and} \quad 1 \le j \le 4.$$

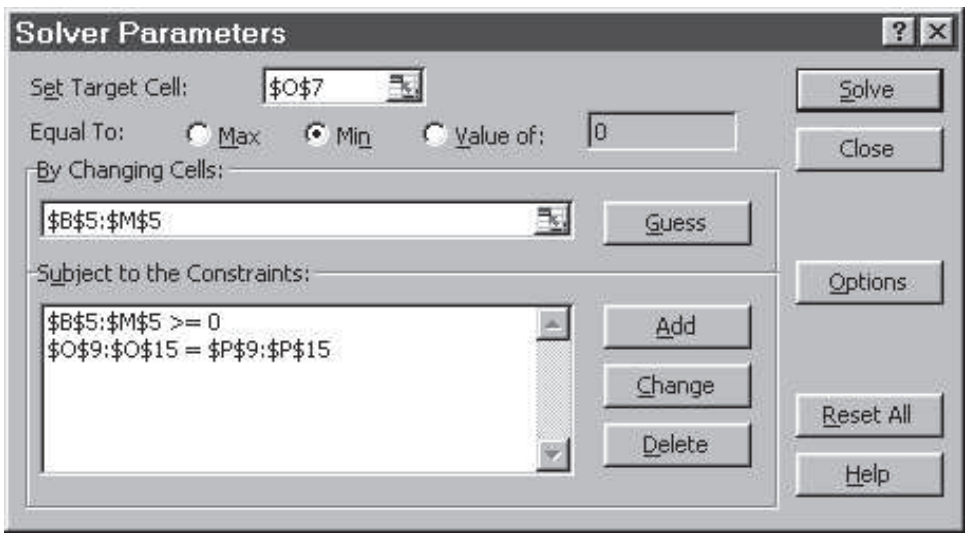
The problem was entered in Excel Solver. The spreadsheet used and the solution appear in Figure 6–4. The solution obtained is

$$x_{14} = 45, \qquad x_{21} = 42, \qquad x_{22} = 78, \qquad x_{31} = 38,$$
$$x_{33} = 47, \quad \text{and} \quad x_{34} = 10,$$

with all other values equaling zero. The total cost of this solution is $297,800.

FIGURE 6–4
Solution of Pear's transportation problem using Excel Solver

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1			**Solution of Example 6.1 (Pear's Transportation Problem)**													
2																
3	Variables	×11	×12	×13	×14	×21	×22	×23	×24	×31	×32	×33	×34	Operator	Value	RHS
4																
5	Values	0	0	0	45	42	78	0	0	38	0	47	10			
6																
7	Objective Coeff	250	420	380	280	1280	990	1440	1520	1550	1420	1660	1730	Min	297800	
8	st															
9	Constraint 1	1	1	1	1									=	45	45
10	Constraint 2					1	1	1	1					=	120	120
11	Constraint 3									1	1	1	1	=	95	95
12	Constraint 4	1				1				1				=	80	80
13	Constraint 5		1				1				1			=	78	78
14	Constraint 6			1				1				1		=	47	47
15	Constraint 7				1				1				1	=	55	55
16																
17																
18		Notes: Formula for Cell O9: =SUMPRODUCT (B9:M9,$B$5:$M$5). Copied to O10 to O15.														
19			Changing cells for Solver are $B$5:$M$5.													
20																



6.4 GENERALIZATIONS OF THE TRANSPORTATION PROBLEM

The Pear Company example is the simplest type of transportation problem. Every link from source to sink is feasible, and the total amount available from the sources exactly equals the total demand at the sinks. Several of these requirements can be relaxed without making the problem significantly more complicated.

Infeasible Routes

Suppose in the example that the firm has decided to eliminate the routes from Dublin to Chicago and from Bangkok to Sioux Falls. This would be accounted for by placing very high costs on these arcs in the network. Traditionally, uppercase M has been used to signify a very high cost. In practice, of course, one would have to assign a number to these locations. As long as that number is much larger than the other costs, an optimal solution will never assign flow to these routes. For the example, suppose we assign costs of \$1,000,000 to each of these routes and re-solve the problem. The reader can check that one now obtains the following solution:

$$x_{14} = 45, \quad x_{21} = 32, \quad x_{22} = 78, \quad x_{31} = 48, \quad \text{and} \quad x_{33} = 47,$$

with all other values equaling zero. The cost of the new solution is \$298,400, only slightly larger than the cost obtained when all the routes were feasible.

Unbalanced Problems

An unbalanced transportation problem is one in which the total amount shipped from the sources is not equal to the total amount required at the sinks. This can arise if the demand exceeds the supply or vice versa. There are two ways of handling unbalanced problems. One is to add either a dummy row or a dummy column to absorb the excess supply or demand. A second method for solving unbalanced problems is to alter the appropriate set of constraints to either $\leq$ or $\geq$ form. Both methods will be illustrated.

Suppose in Example 6.1 that the demand for the tablets was higher than anticipated. Suppose that the respective requirements at the four warehouses are now Amarillo, 90; Teaneck, 78; Chicago, 55; and Sioux Falls, 55. This means that the total demand is 278 and the total supply is 260. To turn this into a balanced problem, we add an additional fictitious factory to account for the 18-unit shortfall. This can be labeled as a dummy row in the transportation tableau and all entries for that row assigned an arbitrarily large unit cost. Note that when supply exceeds demand and one adds a dummy column, the costs in the dummy column do *not* have to be very large numbers, but they do all have to be the same. (In fact, one can assign zero to all costs in the dummy column.) In the example, we assigned a cost of 10^6 to each cell in the dummy row. The resulting Excel spreadsheet and Solver solution (shown in the row labeled "Values") appear in Figure 6–5. The optimal solution calls for assigning the shortfall to two warehouses: 8 units to Chicago and 10 units to Sioux Falls.

Unbalanced transportation problems also can be formulated as linear programs by using inequality constraints. In the previous example, where there is excess demand, one would use equality for the first three constraints, to be sure that all the supply is shipped, and less than or equal to constraints for the last four, with the slack accounting for the shortfall. The reader should check that one obtains the same solution by doing so as was obtained by adding a dummy row. This method has the advantage of giving an accurate value for the objective function. (In the case where the supply exceeds the demand, the principle is the same, but the details differ. The first three supply constraints are converted to *greater than or equal to* form, while the last four demand constraints are still equality constraints. The slack in the first three constraints corresponds to the excess supply.)

FIGURE 6–5
Solution of Example 6.1 with excess demand and dummy row

				Solution of Example 6.1 with Excess Demand and a Dummy Row															
Variables	×11	×12	×13	×14	×21	×22	×23	×24	×31	×32	×33	×34	×41	×42	×43	×44	Oper	Value	RHS
Values	0	0	0	45	42	78	0	0	48	0	47	0	0	0	8	10			
Obj Func	250	420	380	280	1280	990	1440	1520	1550	1420	1660	1730	1.E+06	1.E+06	1.E+06	1.E+06	Min	2E+07	
st																			
Constraint 1	1	1	1	1													=	45	60
Constraint 2					1	1	1	1									=	120	130
Constraint 3									1	1	1	1					=	95	95
Constraint 4													1	1	1	1	=	18	18
Constraint 5	1				1				1				1				=	90	90
Constraint 6		1				1				1				1			=	78	78
Constraint 7			1				1				1				1		=	55	55
Constraint 8				1				1				1				1	=	55	55

6.5 MORE GENERAL NETWORK FORMULATIONS

The transportation problem is a special type of network where all nodes are either supply nodes (also called sources) or demand nodes (also called sinks). Linear programming also can be used to solve more complex network distribution problems as well. One example is the *transshipment problem.* In this case, one or more of the nodes in the network are transshipment points rather than supply or demand points. Note that a transshipment node also can be either a supply or a demand node as well (but no node is both a supply and a demand node).

For general network flow problems, we use the following balance of flow rules:

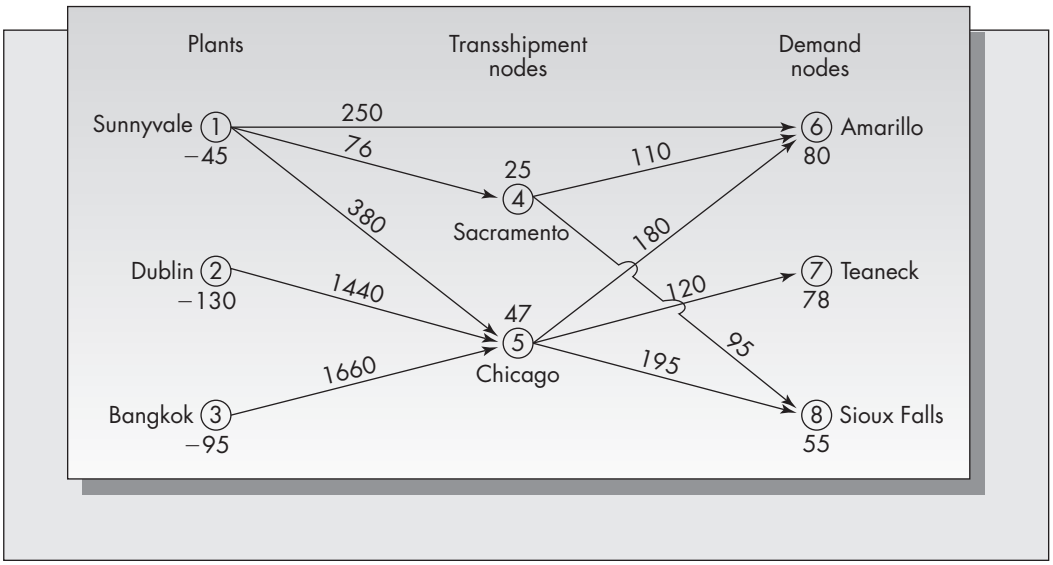
If	Apply the Following Rule at Each Node:
1. Total supply > total demand	Inflow − outflow ≥ supply or demand
2. Total supply < total demand	Inflow − outflow ≤ supply or demand
3. Total supply = total demand	Inflow − outflow = supply or demand

The decision variables are defined in the same way as with the simple transportation problem. That is, x_{ij} represents the total flow from node i to node j. For general network flow problems, we represent the supply as a negative number attached to that node and the demand as a positive number attached to that node. This convention along with the flow rules will result in the correct balance-of-flow equations.

Example *6.2*

Consider the example of Pear Tablets. The company has decided to place a warehouse in Sacramento to be used as a transshipment node and has expanded the Chicago facility to also allow for transshipments. Suppose that in addition to being transshipment nodes, both Chicago and Sacramento are also demand nodes. The new network is pictured in Figure 6–6. Note that several of the old routes have been eliminated in the new configuration.

FIGURE 6–6
Pear Tablets problem with transshipment nodes



We define a decision variable for each arc in the network. In this case, there are a total of 10 decision variables. The objective function is

$$\begin{aligned}\text{Minimize} \quad & 250x_{16} + 76x_{14} + 380x_{15} + 1{,}440x_{25} + 1{,}660x_{35} \\ & + 110x_{46} + 95x_{48} + 180x_{56} + 120x_{57} + 195x_{58}\end{aligned}$$

The total supply available is still 260 units, but the demand is 285 units (due to the additional 25 units demanded at Sacramento). Hence, this corresponds to case 2 of the flow rules in which total demand exceeds total supply. Applying rule 2 to each node gives the following eight constraints for this problem:

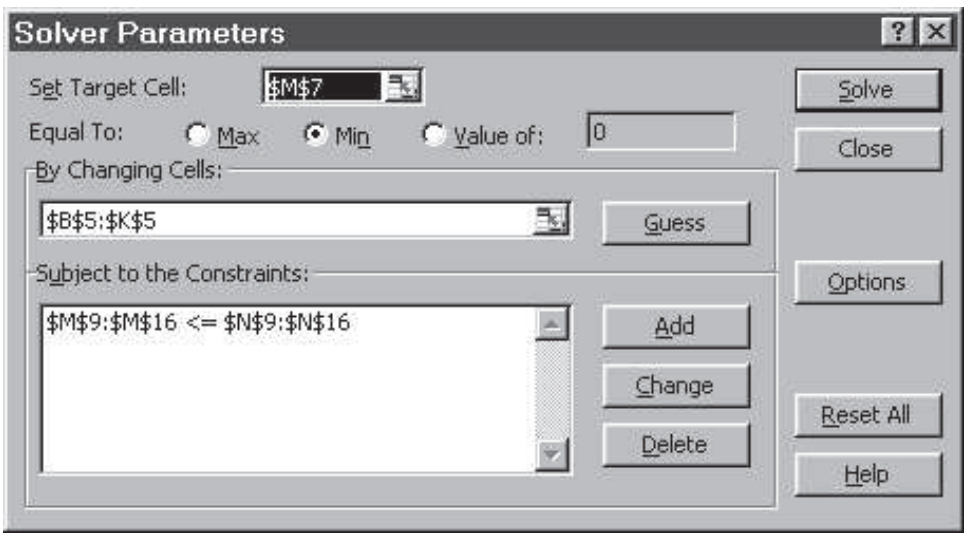
$$\begin{aligned}\text{Node 1:} \quad & -x_{14} - x_{15} - x_{16} \leq -45 \\ \text{Node 2:} \quad & -x_{25} \leq -120 \\ \text{Node 3:} \quad & -x_{35} \leq -95 \\ \text{Node 4:} \quad & x_{14} - x_{46} - x_{48} \leq 25 \\ \text{Node 5:} \quad & x_{16} + x_{46} + x_{56} - x_{56} - x_{57} - x_{58} \leq 47 \\ \text{Node 6:} \quad & x_{16} + x_{46} + x_{56} \leq 80 \\ \text{Node 7:} \quad & x_{57} \leq 78 \\ \text{Node 8:} \quad & x_{48} + x_{58} \leq 55.\end{aligned}$$

For some linear programming codes, all right-hand sides would have to be nonnegative. In those cases, one would multiply the first three constraints by −1. However, this is not required in Excel, so we can enter the constraints just as they appear.

The Excel spreadsheet and solution for Pear's transshipment problem appears in Figure 6–7. Note that the solution calls for shipping all units from the sources. Since this problem had more supply than demand, it is interesting to see where the shortfall occurs. The amount shipped into Sacramento is 45 units (all from Sunnyvale), and the total shipped out of Sacramento is 20 units. Thus, all the demand (25 units) is satisfied in Sacramento. For the other transshipment point, Chicago, the shipments into Chicago are 120 units from Dublin and 95 units from Bangkok, and

FIGURE 6–7
Excel spreadsheet for Pear transshipment problem in Example 6.2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1			**Excel Spreadsheet for Pear Transshipment Problem in Example 6.2**											
2														
3	Variables	×14	×15	×16	×25	×35	×46	×48	×56	×57	×58	Operator	Value	RHS
4														
5	Values	45	0	0	120	95	0	20	80	78	10			
6														
7	Obj Funct	76	380	250	1440	1660	110	95	180	120	195	Min	361530	
8	st													
9	Node 1	−1	−1	−1								<=	−45	−45
10	Node 2				−1							<=	−120	−120
11	Node 3					−1						<=	−95	−95
12	Node 4	1					−1	−1				<=	25	25
13	Node 5		1		1	1			−1	−1	−1	<=	47	47
14	Node 6			1			1		1			<=	80	80
15	Node 7									1		<=	78	78
16	Node 8							1			1	<=	30	55



the shipments out of Chicago are 48 units to Amarillo and 120 units to Teaneck. The difference is 120 + 95 − (48 + 120) = 47 units. Hence, there is also no shortage in Chicago. Total shipments into the demand nodes at Amarillo, Teaneck, and Sioux Falls are respectively 80, 78, and 30 units. Hence, all the shortage (25 units) is absorbed at the Sioux Falls location at the optimal solution.

Real networks can be extremely complex by virtue of their sheer magnitude. As Simchi-Levi et al. (1999) note, a typical soft drink distribution system could involve anywhere between 10,000 and 120,000 accounts. Some retailers have thousands of stores and hundreds of thousands of products. For this reason, efficient data aggregation may be required to solve problems of this magnitude. Customer aggregation is generally

Snapshot Application

IBM STREAMLINES ITS SEMICONDUCTOR SUPPLY CHAIN USING SOPHISTICATED MATHEMATICAL MODELS

An important part of IBM's success is its focus on servicization (see Chapter 1). IBM has also been a leader in the use of optimization and other business analytics tools, both selling them as software and services and applying them to their own supply chains. In the mid-1980s they applied such tools to their supply chain for spare parts (Cohen et al., 1990) realizing a 10 percent improvement in parts availability and savings of approximately $20 million annually. More recently, they have applied a combination of optimization and heuristics to improve the planning of their semiconductor supply chain (Degbotse et al., 2013).

IBM has been in the semiconductor business since 1957 and has manufacturing and contract manufacturing facilities in Asia and North America. These facilities make products that range from silicon wafers to complex semiconductor devices. Until the 1990s, IBM facilities were separated by regions. A North American facility would supply component parts to local assembly plants in North America, for example. The regional supply chains were managed and planned independently, in part because enterprise supply chain optimization was not feasible.

Aided by more powerful computers and algorithms, Degbotse et al. (2013) developed a central planning engine to coordinate planning across the extended supply chain. Its purpose is "to determine a production and shipment plan for the enterprise by using limited material inventories and capacity availability to satisfy a prioritized demand statement." Because such a problem is still too large-scale to solve optimally, they used heuristic decomposition and mixed integer programming. They state that the result is "is a unified production, shipping, and distribution plan with no evidence of the original decomposition."

They found the following benefits: (a) on-time deliveries to commit date increased by 15 percent, (b) asset utilization increased by 2–4 percent of costs, and (c) inventory decreased by 25–30 percent. Notice that by coordinating and planning the supply chain as a whole they are effectively pooling the different regions into one centralized system. Further, notice how the benefits include both increased service and decreased costs, rather than a trade-off between the two. This is the ideal outcome for any process improvement.

What is the lesson learned from this case? As supply chains get larger and more complex, more and more sophisticated methods will be required to manage those systems. Generic software products may not be able to provide sufficient power and customization to be effective in such environments. IBM's experience is only one example of how the management of complex supply chain structures can be improved with the aid of the modeling methods discussed in this text.

accomplished by combining accounts that are nearby. Combining customers with like or similar zip codes is a common means of geographic aggregation. Product aggregation is discussed in detail in Chapter 3 in the context of manufacturing. Product aggregation rules for a supply chain are likely to be different from those discussed in Chapter 3. For example, one might aggregate products according to where they are picked up or where they are delivered. We refer the interested reader to Simchi-Levi et al. (1999) for a more comprehensive discussion of the practical issues surrounding implementation of supply chain networks.

Mathematical modeling has been used successfully in many supply chain applications. Sophisticated models lie at the heart of several commercial software products such as those offered by Texas-based i2 Technologies. The Snapshot Application for this section discusses a successful application of advanced inventory control models to IBM's supply chain for semiconductors.

Problems for Sections 6.3–6.5

9. Consider Example 6.2 of the Pear Company assuming the following supplies and demands:

Plant	Production	Warehouse	Requirement
Sunnyvale	60	Amarillo	100
Dublin	145	Teaneck	84
Bangkok	125	Chicago	77
		Sioux Falls	69

Use Excel's Solver (or other linear programming code) to determine the optimal solution.

10. Resolve Pear's transportation problem assuming that the following routes are eliminated: Sunnyvale to Teaneck, Dublin to Chicago, and Bangkok to Amarillo. What is the percentage increase in total shipping costs at the optimal solution due to the elimination of these routes?
11. Resolve Pear's transshipment problem assuming that an additional transshipment point is located at Oklahoma City. Assume that the unit costs of shipping from the three plants to Oklahoma City are Sunnyvale, \$170; Dublin, \$1,200; and Bangkok, \$1,600; and the respective costs of shipping from Oklahoma City to the demand nodes are Amarillo, \$35; Teaneck, \$245; and Sioux Falls, \$145. Assume that Oklahoma City is only a transshipment point and has no demand of its own. Find the new shipping pattern with the addition of the new transshipment point and the savings, if any, of introducing this additional node.
12. Major Motors produces its Trans National model in three plants located in Flint, Michigan; Fresno, California; and Monterrey, Mexico. Dealers receive cars from regional distribution centers located in Phoenix, Arizona; Davenport, Iowa; and Columbia, South Carolina. Anticipated production at the plants over the next month (in 100s of cars) is 43 at Flint, 26 at Fresno, and 31 at Monterrey. Based on firm orders and other requests from dealers, Major Motors has decided that it needs to have the following numbers of cars at the regional distribution centers at month's end: Phoenix, 26; Davenport, 28; and Columbia, 30. Suppose that the cost of shipping 100 cars from each plant to each distribution center is given in the following matrix (in \$1,000s):

		TO		
		Phoenix	**Davenport**	**Columbia**
FROM	**Flint**	12	8	17
	Fresno	7	14	21
	Monterrey	18	22	31

a. Convert the problem to a balanced problem by adding an appropriate row or column and find the optimal solution using Solver.

b. Now find the optimal solution using Solver and inequality constraints.

c. Do the solutions in (a) and (b) match? Why or why not?

d. Suppose that the route between Monterrey and Columbia is no longer available due to a landslide on a key road in Mexico. Modify the model in (b) and resolve to find the optimal solution. Has the objective value increased or decreased? Explain why.

13. Consider the problem of Major Motors described in Problem 4. In order to be able to deal more effectively with unforeseen events (such as the road closing), Major Motors has established two transshipment points between the factories and the regional distribution centers at Santa Fe, New Mexico, and Jefferson City, Missouri. The cost of shipping 100 cars to the transshipment points is (in $1,000s):

		TO	
		Santa Fe	**Jefferson City**
FROM	**Flint**	8	6
	Fresno	6	9
	Monterrey	9	14

while the cost of shipping from the transshipment points to the distribution centers is

		TO		
		Phoenix	**Davenport**	**Columbia**
FROM	**Santa Fe**	3	8	10
	Jefferson City	5	5	9

Assuming that none of the direct routes between the factories and the distribution centers is available, find the optimal flow of cars through the transshipment points that minimizes the total shipping costs.

14. Toyco produces a line of Bonnie dolls and accessories at its plants in New York and Baltimore that must be shipped to distribution centers in Chicago and Los Angeles. The company uses Air Freight, Inc., to make its shipments. Suppose that it can ship directly or through Pittsburgh and Denver. The daily production rates at the plants are respectively 5,000 and 7,000 units daily, and the demands at the distribution centers are respectively 3,500 and 8,500 units daily. The costs of shipping 1,000 units are given in the following table. Find the optimal shipping routes and the associated cost.

		TO			
		Pittsburgh	**Denver**	**Chicago**	**Los Angeles**
FROM	**New York**	$182	$375	$285	$460
	Baltimore	77	290	245	575
	Pittsburgh	—	275	125	380
	Denver	—	—	90	110

15. Reconsider Problem 6 if there is a drop in the demand for dolls to 3,000 at Chicago and 7,000 at Los Angeles. Find the optimal shipping pattern in this case. How much of the total decrease in demand of 2,000 units is absorbed at each factory at the optimal solution?
16. Reconsider Problem 6 assuming that the maximum amount that can be shipped from either New York or Baltimore through Pittsburgh is 2,000 units due to the size of the plane available for this route.

6.6 DETERMINING DELIVERY ROUTES IN SUPPLY CHAINS

An important aspect of supply chain logistics is efficiently moving product from one place to another. The transportation and transshipment problems discussed earlier in this chapter deal with this problem at a macro or firmwide level. At a micro level, deliveries to customers also must be planned efficiently. Because of the scale of the problem, efficient delivery schedules can have a very significant impact on the bottom line. As a result, they become an important part of designing the entire supply chain.

Determining optimal delivery schedules turns out to be a very difficult problem in general, rivaling the complexity of job shop scheduling problems discussed in detail in Chapter 8. Vehicle scheduling is closely related to a classical operations research problem known as the traveling salesman problem. The problem is described in the following way. A salesman starts at his home base, labeled city 1. He must then make stops at $n - 1$ other cities, visiting each city exactly once. The problem is to determine the optimal sequence in which to visit the cities to minimize the total distance traveled. Although this problem is easy to state, it turns out to be very hard to solve. If the number of cities is small, it is possible to enumerate all the possible tours. There are $n!$ orderings of n objects.[1] For modest values of n, one can enumerate all the tours and compute their distances directly. For example, for $n = 5$ there are 120 sequences. This number grows very fast, however. For $n = 10$, the number of sequences grows to over 3 million, and for $n = 25$ it grows to more than 1.55×10^{25}. To get some idea of how large this number is, suppose that we could evaluate 1 trillion sequences per second on a supercomputer. Then, for a 25-city problem, it would take nearly 500,000 years to evaluate all the sequences!

Total enumeration is hopeless for solving all but the smallest traveling salesman problems. Problems such as this are known in mathematics as *NP hard*. The NP stands for no polynomial, meaning that the time required to solve such problems is an exponential function of the number of cities rather than a polynomial function. We will not dwell on the traveling salesman problem here, but note that methods of solution have been proposed that are vast improvements over total enumeration. However, finding optimal solutions to even moderate-sized problems is still difficult.

Finding optimal routes in vehicle scheduling is a similar, but more complex, problem. Assume that there is a central depot with one or more delivery vehicles and n customer locations, each having a known requirement. The question is how to assign vehicles to customer locations to meet customer demand and satisfy whatever constraints there might be at minimum cost. More real vehicle scheduling problems are too large and complex to solve optimally.

[1] $n!$ is equal to n times $(n - 1)$ times $(n - 2)$ . . . times 1.

Because optimality may be impossible to achieve, methods for determining "good" solutions are important. We will discuss a simple technique for finding good routes, known as the savings method and developed by Clarke and Wright (1964).

Suppose that there is a single depot from which all vehicles depart and return. Customers' locations and needs are known. Identify the depot as location 0 and the customers as locations $1, 2, \ldots, n$. We assume that there are known costs of traveling from the depot to each customer location, given by

$$c_{0j} = \text{Cost of making one trip from the depot to customer } j.$$

To implement the method, we will also need to know the costs of trips between customers. This means that we will assume that the following constants are known as well:

$$c_{ij} = \text{Cost of making a trip from customer location } i \text{ to customer location } j.$$

For our purposes we consider only the case in which $c_{ij} = c_{ji}$ for all $1 \leq i, j \leq n$. This does not necessarily hold in all situations, however. For example, if there are one-way streets, the distance from i to j may be different from the distance from j to i. The method proceeds as follows: Suppose initially that there is a separate vehicle assigned to each customer location. Then the initial solution consists of n separate routes from the depot to each customer location and back. It follows that the total cost of all round trips for the initial solution is

$$2\sum_{j=1}^{n} c_{0j}.$$

Now, suppose that we link customers i and j. That is, we go from the depot to i to j and back to the depot again. In doing so, we would save one trip between the depot and location i and one trip between the depot and location j. However, there would be an added cost of c_{ij} for the trip from i to j (or vice versa). Hence, the savings realized by linking i and j is

$$s_{ij} = c_{0i} + c_{0j} - c_{ij}.$$

The method is to compute s_{ij} for all possible pairs of customer locations i and j, and then rank the s_{ij} in decreasing order. One then considers each of the links in descending order of savings and includes link (i, j) in a route if it does not violate feasibility constraints. If including the current link violates feasibility, one goes to the next link on the list and considers including that on a single route. One continues in this manner until the list is exhausted. Whenever link (i, j) is included on a route, the cost savings is s_{ij}.

The total number of calculations of s_{ij} required is

$$\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2}.$$

(When c_{ij} and c_{ji} are not equal, twice as many savings terms must be computed.)

The savings method is feasible to solve by hand for only small values of n. For example, for $n = 10$ there are 45 terms, and for $n = 100$ there are nearly 5,000 terms. However, as long as the constraints are not too complex, the method can be implemented easily on a computer.

We illustrate the method with the following example.

Example *6.3*

Whole Grains is a small bakery that supplies five major customers with bread each morning. If we locate the bakery at the origin of a grid [i.e., at the point (0, 0)], then the five customer locations and their daily requirements are

Customer	Location	Daily Requirements (loaves)
1	(15, 30)	85
2	(5, 30)	162
3	(10, 20)	26
4	(5, 5)	140
5	(20, 10)	110

The relative locations of the Whole Grains bakery and its five customers are shown in Figure 6–8.The bakery has several delivery trucks, each having a capacity of 300 loaves. We shall assume that the cost of traveling between any two locations is simply the straight-line or Euclidean distance between the points. The formula for the straight-line distance separating the points (x_1, y_1) and (x_2, y_2) is

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

The goal is to find a delivery pattern that both meets customer demand and minimizes delivery costs, subject to not exceeding the capacity constraint on the size of the delivery trucks.

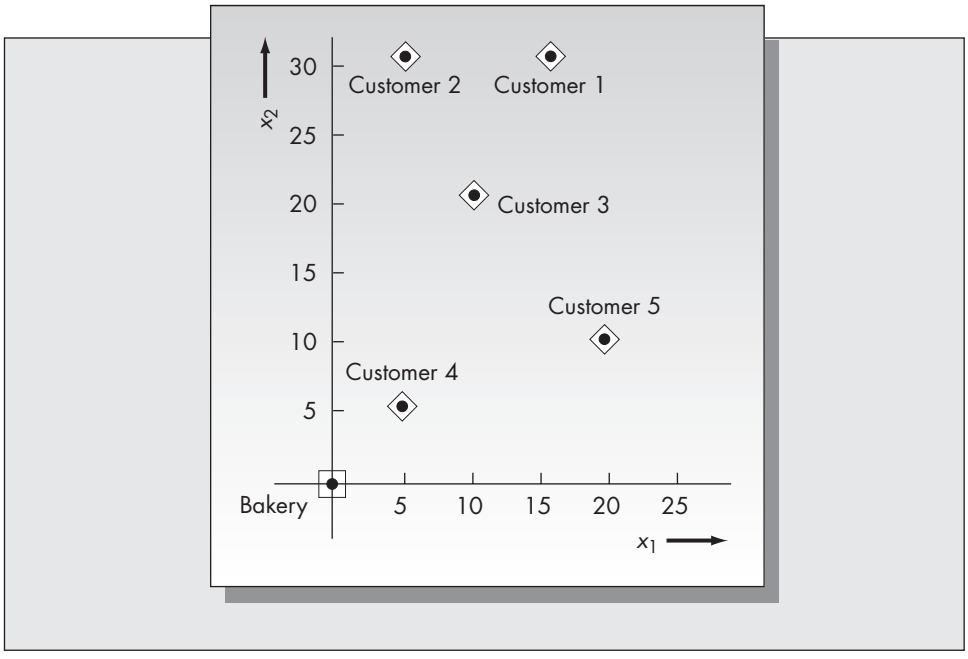


FIGURE 6–8
Customer locations in Example 6.3

Solution

The first step is to compute the cost for each pair (i, j) where i and j vary from 0 to 5. We are assuming that this cost is the straight-line distance between the points representing customer locations. The straight-line distances are given in the following matrix.

Cost Matrix (c_{ij})

		TO					
		0	**1**	**2**	**3**	**4**	**5**
FROM	**0**		33.5	30.4	22.4	7.1	22.4
	1			10.0	11.2	26.9	20.6
	2				11.2	25.0	25.0
	3					15.8	14.1
	4						15.8

Next, we compute the savings for all pairs (i, j), $1 \leq i < j \leq 5$. There are a total of 10 savings terms to compute for this example:

$$s_{12} = c_{01} + c_{02} - c_{12} = 33.5 + 30.4 - 10 = 53.9,$$
$$s_{13} = c_{01} + c_{03} - c_{13} = 33.5 + 22.4 - 11.2 = 44.7.$$

The remaining terms are computed in the same way, with the results

$$s_{14} = 13.7, \quad s_{25} = 27.8,$$
$$s_{15} = 35.3, \quad s_{34} = 13.7,$$
$$s_{23} = 41.6, \quad s_{35} = 30.7,$$
$$s_{24} = 12.5, \quad s_{45} = 13.7.$$

The next step is to rank the customer pairs in decreasing order of their savings values. This results in the ranking

(1, 2), (1, 3), (2, 3), (1, 5), (3, 5), (2, 5), (1, 4), (3, 4), (4, 5), and (2, 4).

Note that (1, 4), (3, 4), and (4, 5) have the same savings. Ties are broken arbitrarily, so these three pairs could have been ranked differently. We now begin combining customers and creating vehicle routes by considering the pairs in ranked order, checking each time that we do not violate the problem constraints. Because (1, 2) is first on the list, we first try linking customers 1 and 2 on the same route. Doing so results in a load of 85 + 162 = 247 loaves. Next, we consider combining 1 and 3, which means including 3 on the same route. This results in a load of 247 + 26 = 273, which is still feasible. Hence, we have now constructed a route consisting of customers 1, 2, and 3. The next pair on the list is (2, 3). However, 2 and 3 are already on the same route. Next on the list is (1, 5). Linking customer 5 to the current route is infeasible, however. As the demand at location 5 is 110 loaves, adding location 5 to the current route would exceed the truck's capacity. The next feasible pair on the list is (4, 5) which we make into a new route. The solution recommended by the savings method consists of two routes, as shown in Figure 6–9.

We should point out that the savings method is only a heuristic. It does not necessarily produce an optimal routing. The problem is that forcing the choice of a highly ranked link may preclude other links that might have slightly lower savings but might be better choices in a global sense by allowing other links to be chosen downstream. Several authors have suggested modifications of the savings method to attempt to overcome this difficulty. The interested reader should refer to Eilon et al. (1971) for a discussion on these methods. However, the authors point out that these modifications do not always result in a more cost-effective solution.

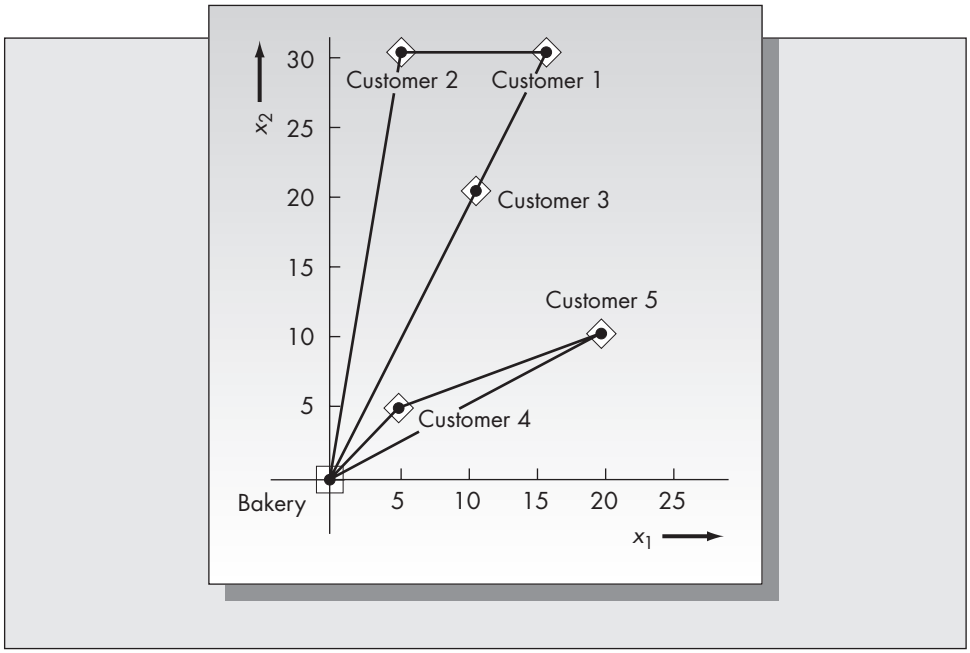


FIGURE 6–9
Vehicle routing found from the savings method for Example 6.3

Practical Issues in Vehicle Scheduling

We may classify vehicle scheduling problems as one of two types: arc-based or node-based. Arc-based problems are ones in which the goal is to cover a certain collection of arcs in a network. Typical examples are snow removal and garbage collection. The type of distribution problems we have discussed in this section are node-based problems. The objective is to visit a specified set of locations. Problems also may be a combination of both of these.

Most real vehicle scheduling problems are much more complex than that described in Example 6.3. Schrage (1981) lists 10 features that make real problems difficult to solve. These include the following seven features:

1. *Frequency requirements.* Visits to customers may have to occur at a certain frequency, and that frequency may vary from customer to customer. In our example, bread deliveries are made daily to each customer, so frequency is not an issue. Consider, however, the problem of delivering oil or gas for residential use. The frequency of delivery depends on the usage rate, so delivery frequency will vary from customer to customer.
2. *Time windows.* This refers to the requirement that visits to customer locations be made at specific times. Dial-a-ride systems and postal and bank pickups and deliveries are typical examples.
3. *Time-dependent travel time.* When deliveries are made in urban centers, rush-hour congestion can be an important factor. This is an example of the case in which travel time (and hence the cost associated with a link in the network) depends on the time of day.
4. *Multidimensional capacity constraints.* There may be constraints on weight as well as on volume. This can be a thorny issue, especially when the same vehicles are used to transport a variety of different products.

Snapshot Application

J.B. HUNT SAVES BIG WITH ROUTING AND SCHEDULING ALGORITHM

J.B. Hunt Transport Services, Inc., is one of the largest transportation logistics companies in North America. A significant portion of their business lies in *drayage*, which is the transport of goods from an origin to a destination within the same urban area (i.e., it does not include long haul operations). The term originated from transport by dray horses, but of course trucks are used exclusively for such operations these days. One of the most common types of drayage is the transport of containerized cargo between and among rail ramps and shipping docks.

Pazour and Neubert (2013) describe a routing and scheduling project done for J.B. Hunt Transport to determine cross-town drayage moves between rail ramps. While this scheduling was originally done manually, the size of the fleet and scale of movements had grown too large for this to remain practical. The combination of fleet and movements made for 4.13×10^{32} routes, although not all will be feasible because they ignore availability constraints for ramps and drivers and geographical considerations. Because of the scale of the problem, the authors produced a heuristic to decide the routing of trucks and the scheduling of drivers on the routes.

J.B. Hunt retains a fleet of drivers and also hires third-party drivers to cover overloads. Since third-party drivers are paid by the load, the primary objective of the heuristic is to maximize the number of loads covered by company drivers. However, there may be many solutions with identical numbers of company driver loads; therefore, the heuristic also tries to minimize total empty travel miles, which can cost a significant amount of money in terms of fuel usage and vehicle wear and tear.

The constraints in the optimization are that every load must be covered, either by a company driver or a third-party contractor and every company driver must be assigned a route. The heuristic generates feasible routes that consist of combinations of legs that either move a container across town or move the empty truck to where it is needed. A truck may do twelve or more loads in a day. Third-party contractors are not assigned to routes. Instead, they are assumed to simply do a single container load from an origin to a destination, which is a conservative assumption. The heuristic also considers operational constraints, including the number of loads per driver schedule, driver start times, driver start and end locations, hourly traffic patterns, load time windows, and required driver service hours.

As reported by the authors, the implementation of the cross-town application has positively impacted J.B. Hunt's intermodal drayage operation by automating and enhancing planning work flow for dispatchers, allowing the fleet size to grow without making planning impossible, reducing the number of costly outsourced loads, and significantly improving operational efficiency. J.B. Hunt has documented the annualized cost savings of the cross-town heuristic implementation at $581,000. This is yet another example of the application of sophisticated operations research methods.

5. *Vehicle types*. Large firms may have several vehicle types from which to choose. Vehicle types may differ according to capacity, the cost of operation, and whether the vehicle is constrained to closed trips only (where it must return to the depot after making deliveries). When several types of vehicles are available, the number of feasible alternatives increases dramatically.
6. *Split deliveries*. If one customer has a particular requirement, it could make sense to have more than one vehicle assigned to that customer.
7. Uncertainty. Routing algorithms invariably assume that all the information is known in advance. In truth, however, the time required to cross certain portions of a network could be highly variable, depending on factors such as traffic conditions, weather, and vehicle breakdowns.

Problems for Section 6.6

17. Re-solve Example 6.3 assuming that the capacity of the vehicles is only 250 loaves of bread.
18. Re-solve Example 6.3 assuming that the distance between any two locations is the rectangular distance rather than the Euclidean distance. (See Section 10.7 for a definition of rectangular distance.)

19. Add the following customer locations and requirements to Example 6.3 and re-solve:

Customer	Location	Daily Requirement
6	(12, 12)	78
7	(23, 3)	126

20. Suppose that one wishes to schedule vehicles from a central depot to five customer locations. The cost of making trips between each pair of locations is given in the following matrix. (Assume that the depot is location 0.)

Cost Matrix (c_{ij})

		TO					
		0	**1**	**2**	**3**	**4**	**5**
	0		20	75	33	10	30
F	**1**			35	5	20	15
R	**2**				18	58	42
O	**3**					40	20
M	**4**						25

Assume that these costs correspond to distances between locations and that each vehicle is constrained to travel no more than 50 miles on each route. Find the routing suggested by the savings method.

21. All-Weather Oil and Gas Company is planning delivery routes to six natural gas customers. The customer locations and gas requirements (in gallons) are given in the following table.

Customer	Location	Requirements (gallons)
1	(5, 14)	550
2	(10, 25)	400
3	(3, 30)	650
4	(35, 12)	250
5	(10, 7)	300

Assume that the depot is located at the origin of the grid and that the delivery trucks have a capacity of 1,200 gallons. Also assume that the cost of travel between any two locations is the straight-line (Euclidean) distance between them. Find the route schedule obtained from the savings method.

6.7 RISK POOLING

A key strategy for increasing efficiency and mitigating uncertainty in supply chains is known as *risk pooling*. Put simply, risk pooling means that when one adds multiple sources of variability, the whole is inherently less variable. A common measure of relative variability is known as the coefficient of variation (CV). The coefficient of variation of a random variable is the ratio of the standard deviation over the mean. In symbols, if X is a random variable with mean μ and standard deviation σ, then $CV(X) = \dfrac{\sigma}{\mu}$.

Now, consider n independent sources of variation, represented by the random variables (X_1, X_2, …, X_n). Assume that these are independent and identically distributed and each have mean μ and standard deviation σ. These might represent demands at n stores of a single retailer. Now consider *CV(W)* where $W = \Sigma_{i=1}^{n} X_i$ (the sum of the demands at the n stores). Since the random variables are assumed to be independent, it follows that the variance of W is $n\sigma^2$ and hence the standard deviation of W is $\sigma\sqrt{n}$. Since the expected value (that is, the mean) of W is $n\mu$, it follows that the coefficient of variation of W is $CV(W) = \dfrac{\sigma}{\mu\sqrt{n}}$. Clearly this is decreasing in n. Even if the random variables (X_1, X_2, …, X_n) do not have the same distribution, a similar phenomenon will hold. (However, if the demands are positively correlated the relative variation can actually increase.) Pooling is analogous to portfolio diversification in finance. In that case, one seeks financial products (e.g., stocks and bonds) that are negatively correlated and that are of sufficient variety (i.e., large enough n) such that the risk of the total portfolio is decreased. In operations management, the goal is to aggregate sources of uncertainty so that the whole is easier to manage. There are three key versions of risk pooling that will be discussed below.

1. Inventory/location pooling.
2. Product pooling and postponement.
3. Capacity pooling.

Inventory/Location Pooling

If two geographic sources of demand can be served by the same supply then pooling will mean that the aggregate demand will have lower coefficient of variation. This in turn implied that the inventory needed to achieve a target service level will be less. Of course, there are declining marginal returns to pooling so if the coefficient of variation is already low it will have little benefit.

Inventory pooling can either be achieved by a centralized warehouse, which has the disadvantage of moving inventory away from customers, or by *virtual pooling*, where items are transshipped from one location to another for resupply, which may increase transportation costs.

Example *6.4*

Consider a toy retailer with warehouses in St. Louis and Kansas City, Missouri. The warehouses stock an identical popular toy delivered to stores in the two cities. Assume a warehouse serves only stores in its city. Weekly demand for St. Louis is normally distributed with mean 2,000, and standard deviation 400 (that is N(μ = 2,000, σ = 400)). Weekly demand for Kansas City is distributed N(μ = 2,000, σ = 300). We assume that the two cities are far enough apart so that demand in the two cities is independent.

The following parameters are seen by both warehouses:

Replenishment lead time in weeks ~ N(μ =2, σ = 0.1);

Fixed shipping cost of replenishment: $500;

Cost per toy: $10; and

Holding cost per toy 20 percent of toy's value per year.

How many toys should each location order at a time and when should they reorder if they want a 99 percent cycle service level? What if they pool the two locations?

Solution

From the EOQ equation of Section 4.5 of Chapter 4, shown in the spreadsheet below, both warehouses should order 7,221 toys at a time. Notice how weekly demand has been converted to yearly demand so it is in the same time units as the holding cost.

	A	B	C	D	E	F	G
1	λ	104280	Demand rate of the item, in units/unit time				
2	K	$500	Fixed cost incurred with each replenishment, in dollars				
3	h	2	Holding cost per unit per unit time				
4	EOQ	7220.8	Economic order quantity				
5	EOQ rounded up	7221					
6							

Cell Formulas

Cell	Formula
B1	=2000*52.14
B3	=0.2*10
B4	=SQRT(2*B2*B1/B3)

From the type-1 service reorder point model of Section 5.5, shown in the spreadsheet below, St. Louis should keep a safety stock of 1,396 toys and Kansas City should keep a safety stock of 1,092 toys (only calculations for St Louis are shown but Kansas City is similar with cell B3 equal to 300 instead of 400).

	A	B	C	D	E	F	G
1	Service level = (1-α)	0.01	Probability of a stockout in a reorder cycle				
2	λ	2000	Demand rate of the item, in units/unit time (E[D])				
3	σ(λ)	400	Standard deviation of the demand per unit time				
4	μ(t)	2	Expected lead time, in unit time				
5	σ(t)	0.1	Standard deviation of the leadtime				
6							
7							
8	Solve for safety stock and reorder point						
9	σ(LTD)	600.00	Standard deviation of demand during lead time, in units				
10	z	2.33	Safety factor				
11	Safety Stock	1395.81	zσ(LTD)				
12	Reorder Point	5395.81	Safety stock + λμ(t)				
13							

Cell Formulas

Cell	Formula
B9	=SQRT(B4*B3*B3+B2*B2*B5*B5)
B10	=NORMSINV(1−B1)
B11	=B10*B9
B12	=B11+B2*B4

If the two warehouses are pooled (either physically or virtually through an information system) then weekly demand becomes 4,000 (=2*2000) and the standard deviation of weekly demand becomes $500\left(=\sqrt{300^2+400^2}\right)$. Substituting these numbers into the above spreadsheets show that the company should order 10,199 toys at a time (a 29 percent reduction) and keep a safety stock of 1,890 toys (a 24 percent reduction).

Thus it can be seen that pooling inventory can result in a significant reduction of inventory in the system. As another illustration consider the following simple scenario. Assume n independent retail locations stock similar items. For example, these could be Macy's department stores located in different cities. Let us further assume that the stock level of a particular item is determined from the newsvendor model. Referring to the discussion in Section 5.5, there are known values of the unit overage cost, c_o, and the unit underage cost, c_u. In Chapter 5 it was proven that the optimal stocking level is the $c_u/(c_u + c_o)$ fractile of the demand distribution. (If we assume that stocking levels are determined based on service levels instead of costs, then the critical ratio *is* the service level and c_o and c_u need not be known.)

To simplify the analysis, let us suppose that the demand for this item follows the same normal distribution at each store with mean μ and standard deviation σ, and the demands are independent from store to store. Let z^* be the value of the standard normal variate that corresponds to a left-tail probability equal to the critical ratio. Then, as is shown in Section 5.3, the optimal policy is to order up to $Q = m + \sigma z^*$ at each location. The safety stock held at each location is σz^*, so the total safety stock in the system is $n\sigma z^*$. Refer to this case as the decentralized system.

Alternatively, suppose that all inventory for this item is held at a single distribution center and shipped overnight on an as-needed basis to the stores. Let's consider the amount of safety stock needed in this case to provide the same level of service as with the decentralized system. Since store demands are independent normal random variables with mean μ and variance σ^2, the aggregated demand from all n stores is also normal, but with mean $n\mu$ and variance $n\sigma^2$. The standard deviation of the aggregated demand therefore has standard deviation $\sigma\sqrt{n}$. This means that to achieve the same level of service, the warehouse needs to stock up to the level Q_w given by $Q_w = n\mu + z^*\sigma\sqrt{n}$. The total safety stock is now $z^*\sigma\sqrt{n}$. This corresponds to a centralized system.

Forming the ratio of the safety stock in the decentralized system over the safety stock in the centralized system gives $z^*\sigma n/z^*\sigma\sqrt{n} = \sqrt{n}$. Hence, the decentralized system will have to hold $\sqrt{n}$ times more safety stock than the centralized system to achieve the same level of service. Even for small values of n, this difference is significant, and for large retailers such as Wal-Mart that have thousands of stores, this difference is enormous.

Of course, this example is a simplification of reality. For most products, it is impractical for every sale to come from a distribution center, and it is impractical for a single distribution center to serve the entire country. Furthermore, the assumption that demands at different stores for the same item are independent is also unlikely to be true. Demands for some items, such as desirable fashion items, are likely to be positively correlated, while others, such as seasonal items like bathing suits, might be negatively correlated. However, even with these caveats, the advantages of centralization are substantial and account for the fact that multilevel distribution systems are widespread in many industries, especially retailing. These results are based on the work of Eppen (1979) (who allowed for correlated demands). This model was extended by Eppen and Schrage (1981) and Erkip, Hausman, and Nahmias (1990) to more general settings.

Several authors have examined the issue of how item characteristics affect the optimal breakdown of DC versus in-store inventory. Muckstadt and Thomas (1980) showed that high-cost, low-demand items derived the greatest benefit from centralized stocking, while Nahmias and Smith (1994) showed how other factors, such as the probability of a lost sale and the frequency of shipments from the DC to the stores, also affect the optimal breakdown between store and DC inventory. A comprehensive review of inventory control models for retailing can be found in Nahmias and Smith (1993), and an excellent collection of articles on general multi-echelon inventory models in Schwarz (1981).

Product Pooling and Postponement

A universal design may be used to pool demand between two product categories. We see the reverse of this effect when products are tailored to a specific market (e.g., boys and girls diapers). The likely negative effect on variability, and hence inventory and production planning, is often overlooked when making such product decisions. One way to mitigate such effects is using delayed differentiation or postponement where the configuration of the final product is delayed as long as possible.

The first application of this principle known to this writer was implemented by Benetton (Signorelli and Heskett, 1984). The Benetton Group is an Italian-based maker of fashion clothes that had, by 1982, become the world leader in the field of knitwear. About 60 percent of the garments sold by the firm are made of wool. Traditionally, wool is dyed before it is knitted. In 1972 Benetton undertook a unique strategy: to dye the garments *after* they were knitted. One might well question this strategy since labor and production costs for garments dyed after manufacture are about 10 percent higher than for garments knitted from dyed thread.

The advantage of reversing the order of the dying and the knitting operations is that it provides additional time before committing to the final mix of colors. This time gives the firm a chance to gain additional data on consumer preferences for colors. Benetton's knitwear included nearly 500 color and style combinations. Undyed garments are referred to as *gray stock*. Keeping inventories of gray stock rather than dyed garments had several advantages. First, if a specific color became more popular than anticipated, Benetton could meet the demand for that color. Second, the company would run less risk of having large unsold stockpiles of garments in unpopular colors. These advantages more than offset the higher costs of dying the knitted garments rather than the raw wool.

How does postponement correlate with inventory management theory? As we saw in Chapter 5, safety stock is retained to protect against demand uncertainty over the replenishment lead time. The lead time for garments of a specific color is reduced by postponing the dying operation. It follows that the uncertainty is reduced as well, thus achieving comparable service levels with less safety stock. Further, because intermediate inventories for different end items are pooled (e.g., all sweaters have the same base garment at Benetton) the inherent demand uncertainty has been reduced.

Postponement has become a key strategy in many diverse industries. Another example discussed in the literature is that of Hewlett-Packard (HP) (Lee, Billington, and Carter, 1994). HP is one of the world's leading producers of inkjet and laser printers, among other products. Printers are sold worldwide. While the basic mechanisms of the printers sold overseas are the same as the American versions, subassemblies, such as power supplies, must be customized for local markets. HP's original strategy was to configure printers for local requirements (i.e., localize them) at the factory. That is, printers with the correct power supplies, plugs, and manuals would be produced at the factory, sorted, and shipped overseas as final products. The result was that HP needed to carry large safety stocks of all printer configurations.

In order to reduce inventories and improve the service provided by the distribution centers (DCs) to retail customers, HP adopted a strategy similar to Benetton's. Printers sent from the factory would be generic or gray stock. Localization would be done at the DC level rather than at the factory. As with Benetton, this had the ultimate effect of reducing necessary safety stocks while improving service. Printers are shipped to overseas DCs by boat, requiring one-month transit time. With local customization of the product, the replenishment lead time for locally configured printers was dramatically reduced. Also, the demand on the factory is now the aggregate of the demands at the DCs, which has relatively smaller variation due to pooling. Lee, Billington, and Carter

showed that DC localization for HP Deskjet-Plus printers would lead to an 18 percent reduction in inventories with no reduction in service levels.

According to Harold E. Edmondson, an HP vice president (as quoted in Lee, Billington, and Carter, 1994):

> The results of this model analysis confirmed the effectiveness of the strategy for localizing the printers at remote distribution centers. Such a design strategy has significant benefits in terms of increased flexibility to meet customer demands, as well as savings in both inventory and transportation costs . . . I should add that the design for localization concept is now part of our manufacturing and distribution strategy.

The notion of postponing customization of a product appears to be gaining acceptance. For example, semiconductors are produced in generic form when possible and customized through programming or other means after firm orders are received (Barrone, 1996). Firms in many industries are becoming aware of the risk-pooling and lead time reduction benefits from postponement and local customization of products.

Capacity Pooling

Almost all production systems require capacity that is larger than expected demand to allow for fluctuations in either demand or supply. This excess capacity is called *safety capacity* and forms a buffer against variability, similar to how inventory buffers variation**.** If capacity can be shifted among products, the different product demands are effectively one pool. In this case, less safety capacity is needed. However, flexible capacity is usually either more expensive or less efficient than dedicated capacity, so such pooling needs to be done carefully. There has been significant research on the "right" level of flexibility for various types of systems.

For manufacturing systems, capacity pooling typically involves investing in more flexible machines. Setup times for changing between product types need to be carefully managed in any nondedicated system. Toyota has led the way in designing equipment to produce multiple products with minimal setup times between different types of products. (Toyota's single minute exchange of dies, SMED, were described in Chapter 1.) For assembly or service system, capacity pooling typically involves cross-training workers. In this case, the workers may be more expensive because they need a higher skill set. Also, there are cognitive limits on how much cross-training one person can absorb.

The good news on capacity pooling is that its benefits can be achieved without a fully flexible system. That is, as shown originally by Jordan and Graves (1995), a little flexibility goes a long way. Consider Figure 6–10, which reproduces Figure 2 in Jordan and Graves (1995). On the right is a fully flexible system where each of 10 plants can produce any of 10 products. Not shown is the fully dedicated system where each product is only produced by one plant. On the left is a strategy known as *chaining* where each product is connected to two plants in such a fashion that the whole system becomes linked. Jordan and Graves (1995) show that such a chain is almost as effective as full flexibility.

Notice that there are many more options for capacity pooling than just the two described above, particularly when capacity corresponds to people rather than machines. For example, instead of training all staff to either work on products (or customer types) 1 and 10 or 1 and 2, as shown in Figure 6–10, some staff may be trained on types 1 and 3, 1 and 4, etc. In addition, some staff may be dedicated to a product or customer type, while others are cross-trained. Such a strategy is called *tailored pairing* by Bassamboo et al. (2010a, 2010b) and is shown to be close to optimal for the systems they consider.

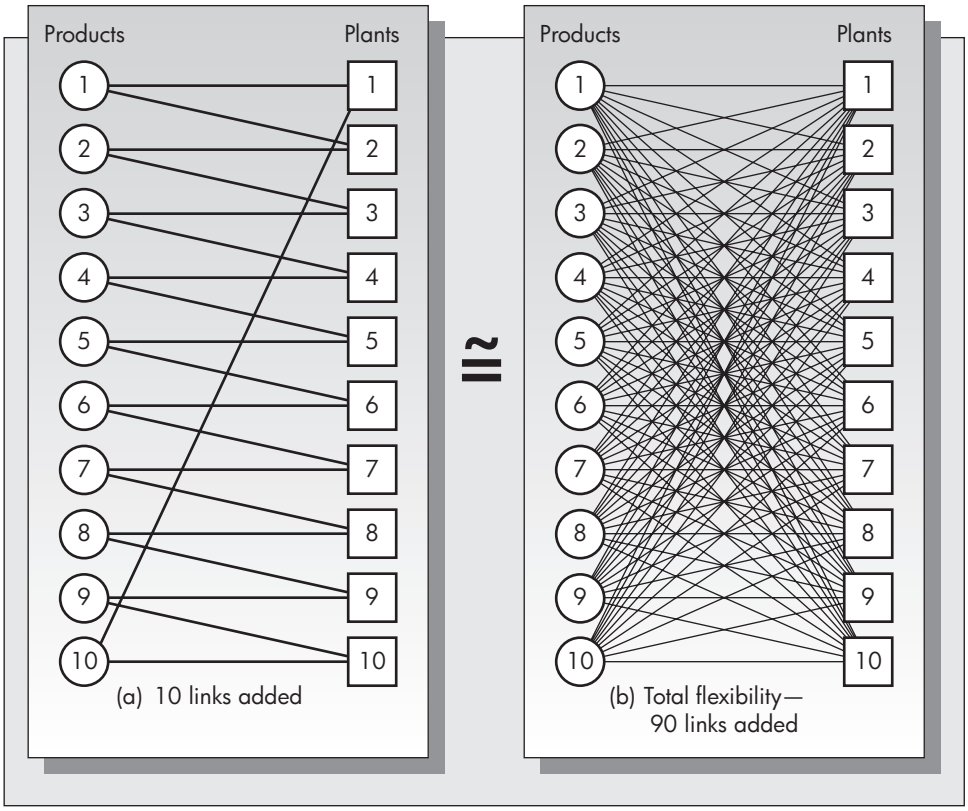


FIGURE 6–10
Flexibility Configurations.

Source: Jordan and Graves (1995)

The key finding in most capacity pooling research is that more pooling is better but there are decreasing returns to scope (i.e., number of types of product pooled). Furthermore, it is important to try to create a circuit with the pooling that encompasses as broad a range of capacity and products as possible. In practice, pooling is constrained by the day-to-day realities of running a production system. But, as we saw above, a little pooling can go a long way.

6.8 DESIGNING PRODUCTS FOR SUPPLY CHAIN EFFICIENCY

Product design was traditionally a function that was totally divorced from more mundane operations management issues. Designers would be concerned primarily with aesthetics and marketability. Little attention would be given to nuts-and-bolts concerns such as manufacturing and logistics at the design stage. As quality and reliability advanced to the forefront, it became clear that product reliability and product design are closely linked. The design for manufacturability (DFM) movement developed out of a need to know why products fail and how those failures could be minimized. Once it was understood that reliability could be factored in at the design stage, the link between design and manufacturing was forged. Another term for DFM is concurrent engineering.

In recent years, firms have realized that the logistics of managing the supply chain can have as much impact on the bottom line as product design and manufacturing. More than ever, we are seeing innovative designs that take supply chain

considerations into account. One way of describing this concept is design for logistics (DFL). Another is three-dimensional concurrent engineering (3-DCE), a term adopted by Fine (1998). The three dimensions here are product, process, and supply chains. It carries the concept of concurrent engineering one step further. Concurrent engineering means that product-related issues (functionality, marketability) and process-related issues (how the product is produced, reliability and quality of the final product) are joint considerations in the design phase. Three-dimensional concurrent engineering means that supply chain logistics is also considered in the product design phase.

Two significant ways that logistics considerations enter into the product design phase are

1. Product design for efficient transportation and shipment.
2. Postponement of final product configuration (as discussed earlier).

Products that can be packed, moved, and stored easily streamline both manufacturing and logistics. Buyers prefer products that are easy to store and easy to move. Some products tend to be large and bulky and present a special challenge in this regard. An example is furniture. Swedish-based Ikea certainly did an excellent job of designing products that are modular and easily stored. Furniture is sold in simple-to-assemble kits that allow Ikea retailers to store furniture in the same warehouselike locations at which they are displayed. Simchi-Levi et al. (1999) discuss several other examples of products whose success is partially based on their ease of shipment and storage.

Additional Issues in Supply Chain Design

While products can be better designed for efficient supply chain operation, there are several important issues to consider in the design of the supply chain itself. Fine (1998) refers to supply chain design as the "ultimate core competency." Three important relevant issues are

- The configuration of the supplier base.
- Outsourcing arrangements.
- Channels of distribution.

Configuration of the Supplier Base

The number of suppliers, their locations, and their sizes are important considerations for efficient supply chain design. In recent years, the trend has been to reduce the number of suppliers and develop long-term arrangements with the existing supplier base. An example is the Xerox Corporation. Jacobson and Hillkirk (1986) discuss several reasons for the impressive turnaround of Xerox during the 1980s. One of the company's strategies was to streamline the supply chain on the procurement side by reducing the number of its suppliers from 5,000 to only about 400. Those 00 suppliers included a mix of both overseas and local suppliers. The overseas suppliers were chosen primarily on a cost basis, while the local suppliers could provide more timely deliveries when necessary.

Cooperative efforts between manufacturers and suppliers (as well as manufacturers and retailers on the other end of the chain) have been gaining popularity. Traditionally this relationship was an adversarial one. Today, with the advent of arrangements like vendor-managed inventories, suppliers and manufacturers work closely and often share what was once proprietary data to improve performance on both ends.

Outsourcing Arrangements

Outsourcing of manufacturing has become a popular trend in recent years. Successful contract manufacturers such as Foxxcon experienced rapid growth in the past decade. Many firms are outsourcing supply chain functions as well. Third-party logistics (3PL) is becoming big business. For example, Wal-Mart now outsources a major portion of its logistics operation.

A 3PL agreement might only outsource transport operations or it might also outsource warehousing, purchasing, and inventory management. A new term, 4PL, has been coined to reflect 3PL providers that offer end-to-end logistics support. Originally the term was only applied to firms that manage external 3PL providers (as well as providing full logistics support). However, more common usage allows the 3PL activities to take place internally to the 4PL. In some sense, Amazon.com is a 4PL provider as it sells products for other companies providing both the web interface as well as the pick-and-pack and dispatch operations.

Dreifus (1992) describes a case of a smaller firm that decided to outsource purchasing. The Singer Furniture Company contracted with the Florida-based IMX Corporation to handle its procurements. In the agreement, IMX handles all negotiations from consolidated shipments, takes care of the paperwork, conducts quality-control inspection, and searches for new suppliers in exchange for a fixed commission. In a trial run, IMX suggested a switch from a Taiwanese supplier of bedposts to one based in the Caribbean, which resulted in a 50 percent reduction in price to Singer. From Singer's point of view, the primary advantages of entering into this arrangement were

Buying economies of scale. IMX buys for several clients (including other furniture makers), which allows it to negotiate lower unit costs on larger orders. There are also economies of scale in shipping costs from overseas. IMX claims that it saves at least 20 percent for clients in this way.

Reduced management overhead. Singer was able to eliminate its international purchasing unit and several related management functions.

Lower warehousing costs. IMX bears much of the risk on several lines of product by absorbing local storage costs.

Above-board accounting. The firms agreed on open procedures and disclosures to be sure that the quality programs are functioning and charges are in line with costs.

Channels of Distribution

Failure to establish solid channels of distribution can spell the end for many firms. The design of the distribution channel includes the number and the configuration of distribution centers; arrangements with third-party wholesalers; licensing and other agreements with retailers, including vendor-managed inventory programs; and establishment of the means to move product quickly through the channel. Direct selling to consumers has been an enormously successful strategy for some retailers, especially for products that become obsolete quickly. By eliminating intermediaries, manufacturers reduce the risk of product dying in the supply channel. This has been a successful strategy for computer maker Dell and for Amazon.com. Dell Computer's phenomenal success can be attributed partially to its outstanding supply chain design, highlighted in the Snapshot Application in this section.

Snapshot Application

DELL COMPUTER DESIGNS THE ULTIMATE SUPPLY CHAIN

Dell Computer was one of the biggest success stories of the 1990s. How successful? The stock price increased a whopping 27,000 percent during the decade. There is no question that a sleek, efficient supply chain design was a contributing factor to Dell's phenomenal success.

Michael Dell, the firm's founder, began by assembling and reselling IBM PCs from his dorm room at the University of Texas in the early 1980s. Later, in the mid-1980s, he formed PC's Limited, which offered one of the first of the mail order "clones." The computers were clones of the IBM XT, which was selling for several thousand dollars at the time. PC's Limited sold a basic box consisting of a motherboard with an Intel 8088 chip, a power supply, a floppy drive, and a controller for $795. The buyer had to add a graphics card, a monitor, and a hard drive to make the system functional, but the final product cost less than half as much as an equivalent IBM XT and ran faster. (This was, in fact, the first computer purchased by this writer.) From this modest beginning grew the multibillion-dollar Dell Computer Corporation which dominates the PC and laptop market today.

To understand Dell's success, one must understand the nature of the personal computer marketplace. The central processor is really the computer, and its power determines the power of the PC. In the PC marketplace, Intel has been the leader in designing every new generation of processor chip. A few companies, such as AMD and Texas Instruments, have had limited success in this market, but the relentless new-product introduction by Intel has resulted in competitors' processors becoming obsolete before they can garner significant market share. Each new generation of microprocessors renders old technology obsolete. For that reason, computers and computer components have a relatively short life span. As new chips are developed at an ever-increasing pace, the obsolescence problem becomes even more critical.

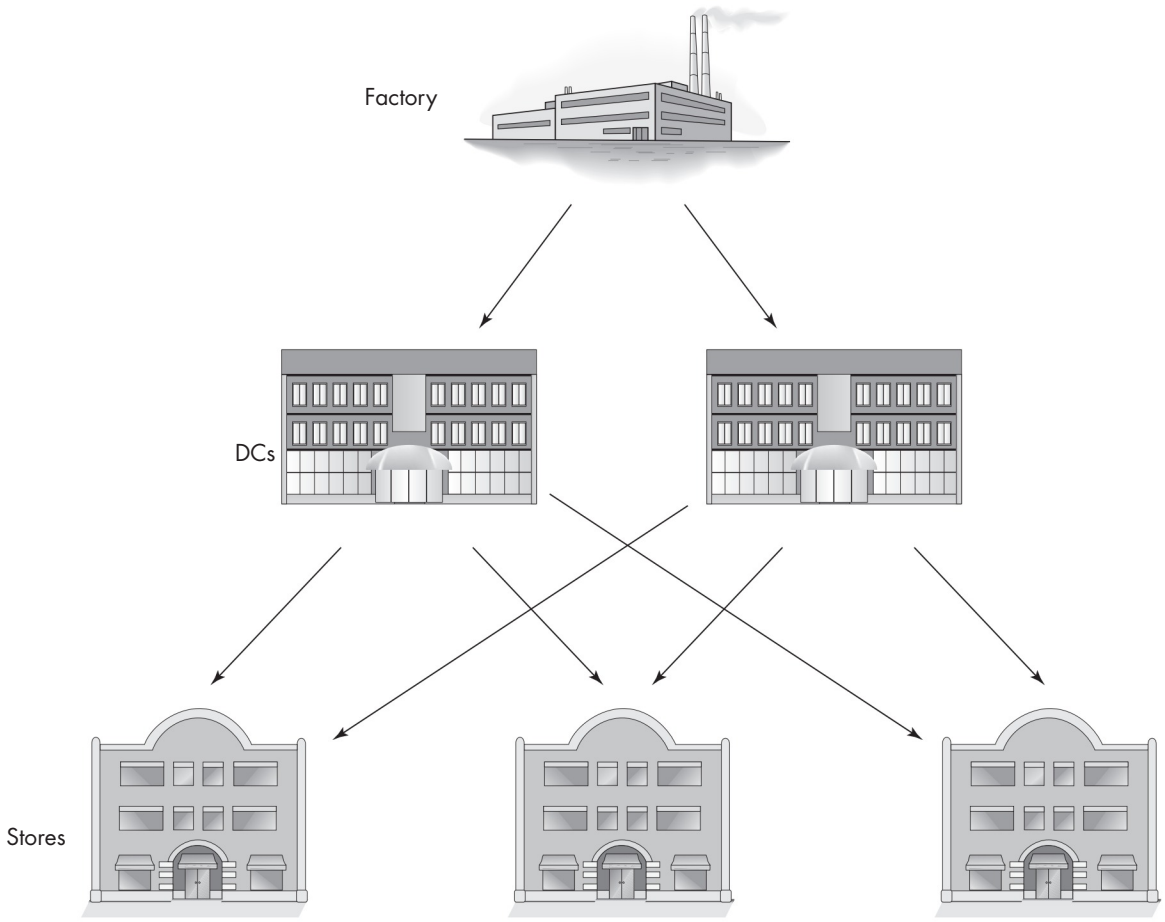
How does one avoid getting stuck with obsolete inventory? Dell's solution was simple and elegant. Don't keep *any* inventory! All PCs purchased from Dell are made to order. Dell does not build to stock. They do store components, however. Although Dell can't guarantee all components are used, their market strategy is designed to move as much of the component inventory as possible. Dell focuses only on the latest technology, and both systems and components are priced to sell quickly. Furthermore, because of their high volume, they can demand quantity discounts from suppliers. As the so-called clockspeed of the industry (a term coined by Fine, 1998) increases, Dell's advantage over manufacturers with more traditional supply chain designs also increases.

Dell's supply chain strategy is not designed to produce the least expensive computer. Less expensive brands, such as Asus and Acer, manufacture in low cost labor markets and in large production quantities. Dell's more agile supply chain allows them to quickly design and market products with the latest technology. However, it is true that personal computers have evolved into a much more commoditized market. For this reason, Dell has moved away from their configure-to-order strategy for many of its products. Also, this has prompted Dell and other U.S.-based manufacturers to expand into other areas, such as peripherals, servers, software, and services. Dell frequently contracts with large organizations (e.g., universities) to provide all of their computing needs. This is another example of servicization, which was discussed in Chapter 1.

6.9 MULTILEVEL DISTRIBUTION SYSTEMS

In many large commercial and military inventory systems, it is common for stock to be stored at multiple locations. For example, inventory may flow from manufacturer to regional warehouses, from regional warehouses to local warehouses, and from local warehouses to stores or other point-of-use locations. Determining the allocation of inventory among these multiple locations is an important strategic consideration, and can have a significant impact on the bottom line. A typical multilevel distribution system is pictured in Figure 6–11.

FIGURE 6–11
Typical multilevel distribution system



In the parlance of traditional inventory theory, such systems are referred to as multi-echelon inventory systems. The interest in this area was originally sparked by a seminal paper by Clark and Scarf (1960) that was part of a major initiative at the Rand Corporation to study logistics and inventory management problems arising in the military. This paper laid the theoretical framework for what later became a very large body of research in multi-echelon inventory theory. (There is a brief overview of multi-echelon inventory models in Chapter 5.)

Including intermediate storage locations in a large-scale distribution system has advantages and disadvantages. The advantages include

1. Risk pooling.
2. Distribution centers that can be designed to meet local needs.
3. Economies of scale in storage and movement of goods.
4. Faster response time.

Suppose a distribution center serves 50 retail outlets. As seen earlier, by holding the majority of the stock at the distribution center rather than at the stores, pooling implies that the same systemwide service level could be achieved with less total inventory.

In retailing, the mix of products sold depends on the location of the store. For example, one sells more short-sleeve shirts in Arizona than in Maine. Hence, *distribution centers* located in the Southwest would stock a different assortment of products than distribution centers located in the Northeast. Distribution centers could be designed to take these differences into account.

Distribution centers allow for *economies of scale* in the following way. If all product were shipped directly from the factory to the local outlet, shipment sizes would be relatively small. Large bulk shipments from factories located overseas could be made to distribution centers at a lower unit cost. In that way, the smaller shipments would be made over shorter distances from distribution centers to stores.

Finally, because distribution centers can be located closer to the customer than can factories, demands can be met more quickly. Retailers try to locate distribution centers in the United States within one day's drive away from any store. (This is why Ohio is a popular place for retail distribution centers, since this is the geographic center of the country.) In this way, store inventories can be replenished overnight if necessary.

Multilevel distribution systems could have several disadvantages as well. These include that they

1. May require more inventory than simpler distribution systems.
2. May increase total order lead times from the plant to the customer.
3. Could ultimately result in greater costs for storage and movement of goods.
4. Could contribute to the bullwhip effect.

Depending on the number of distinct locations used, it is likely that a multilevel system requires more overall total inventory than a single-level system, since safety stock is likely built in at each level. This means that more money is tied up in the pipeline.

When distribution centers experience stock-outs, the response time of the system from manufacturer to point of sale could actually be worse in a multilevel system. This is a consequence of the fact that in the multilevel case, the total lead time from plant to store is the sum of the lead times at each level. (However, in a well-managed system, stock-outs at the distribution center should be very rare occurrences.)

One must not ignore the cost of the distribution center itself. Building a modern, large-scale facility could cost upward of $500 million. In addition, there are recurring costs such as rent and labor. As a result, building and maintaining a multilocation distribution system is expensive.

The *bullwhip effect*, discussed earlier, is the propensity of the variance of orders to increase as one moves up the supply chain. Adding additional levels to a distribution system could cause a bullwhip effect.

Problems for Sections 6.7–6.9

22. Describe how inventory pooling works in a multiechelon inventory system. Under what circumstances does one derive the greatest benefit from pooling? The least benefit?

23. Cross training workers is one way to achieve capacity pooling in a manufacturing environment. What are the advantages and disadvantages to such cross-training?
24. Describe the concept of postponement in supply chains. If you were planning a trip to a distant place, what decisions would you want to postpone as long as possible?
25. Many automobiles can be ordered in one of two engine sizes (examples are the Lincoln LS, the Lexus Coupe, and the Jaguar S Type) but are virtually identical in every other way. How might these automakers use the concept of postponement in their production planning?
26. Discuss why having too many suppliers can be troublesome. Can having too few suppliers also be a problem?
27. Many large companies that have their own manufacturing facilities and logistics organizations outsource a portion of their production (such as IBM) or their supply chain operations (such as Saturn). Why do you suppose this is?
28. A 3PL provider may have only transport services or may also provide warehouse and inventory management services. What are the advantages and disadvantages to a firm in outsourcing its warehouse and inventory management operations along with transportation?
29. What types of companies are best suited to using 4PL providers?
30. Why might a retailer want to consider developing a three-level multilevel system? (The three levels might be labeled National Distribution Center, Regional Distribution Center, and Store.)
31. What are the characteristics of items from which one derives the greatest benefit from centralized storage? the least benefit?

6.10 INCENTIVES IN THE SUPPLY CHAIN

One of the key challenges in SCM is that a supply chain is rarely owned by any one firm. It generally involves different players, often with competing objectives. Consider the following example.

Example *6.5* Suppose that the retailer SnowInc buys ski jackets from the Chinese manufacturer JacketCo. Due to the short selling season and long delivery lead time for this fashion good, this is a one-time purchase decision. SnowInc estimates the season's demand for one type of jacket to be normally distributed with mean 500 and standard deviation 100. Retail price for the jacket is $100. If not sold at the end of the season they unload them for $10 per jacket. The wholesale price JacketCo charges is $30 per jacket and JacketCo's per unit manufacturing cost is $12. How many jackets should SnowInc order and what are the two firms' profits? If both firms were owned by the same company, how would these answers change?

Solution We first consider the decision from SnowInc's perspective. From the newsvendor model (see Section 5.3), they have an overage cost of $30 − $10 = $20 and an underage cost of $100 − $30 = $70. The critical ratio is therefore $70/(20 + 70) = 7/9$ and it is best for SnowInc to order 576 jackets, as shown in the below spreadsheet.

	A	B
1	Price	$100
2	Cost	$30
3	Salvage Value	$10
4	Cost of inventory left (overage Cost)	$20
5	Cost of unsatisfied demand (underage Cost)	$70
6	Mean Demand	500
7	Standard Deviation of Demand	100
8		
9	Critical Fractile	0.7778
10	z	0.7647
11	Amount to Stock	576
12		
13	L(z) - loss function	0.1279
14	Expected lost sales	13
15	Expected sales	487
16	Expected left over inventory	89
17		
18	Expected cost	$ 2,680.21
19	Expected profit	$32,319.79

Cell Formulas

Cell	Formula
B4	=B2−B3
B5	=B1−B2
B9	=B5/(B4+B5)
B10	=NORMSINV(B9)
B11	=B6+B7*B10
B13	=NORM.S.DIST(B10,0)−B10*(1−NORM.S.DIST(B10,1))
B14	=B7*B13
B15	=B6−B14
B16	=B11−B15
B18	=(B4+B5)*B7*NORM.S.DIST(B10,0)
B19	=B5*B15−(B4*B16)

SnowInc has an expected profit of approximately $32,320 and JacketCo has a profit from this order of 576*($30 − $12) = $10,368 for a total channel profit of approximately $42,688. However, if SnowInc and JacketCo are owned by the same company, then by replacing the jacket cost of $30 with its true cost of $12 in the above spreadsheet it is seen that 701 jackets should be ordered for a total channel profit of approximately $43,524 ($30,906 to SnowInc and $12,618 to JacketCo), which is a 2 percent increase in total profit.

If 701 jackets are ordered then the supply chain creates more profit and availability improves. In fact, there is only a 2 percent chance that the retailer will stock out (one minus the critical ratio). The reason that so many jackets are ordered (relative to mean demand of 500 jackets) is because the margins are so high relative to the overstock costs.

Given that the supply chain makes more profit and customers are happier if the retailer orders 701 jackets in Example 6.5, why doesn't the retailer do this on their own initiative? The answer is clear from the number above. Ordering 576 jackets produces an expected profit of $32,320 for the retailer whereas ordering 701 jackets produces an expected profit of $30,906. There is no incentive for the retailer to make this change.

In general, the incentives in Example 6.5 can be written as follows. Suppose the manufacturer produces an item for c, sells it to the retailer for w, and the retailer sells it for p. Then the underage cost for the retailer is $p - w$, for the manufacturer is $w - c$, and for the supply-chain as a whole is $p - c$. Because $(p - w) > (p - c)$, the retailer will under stock from the standpoint of maximizing profit to the whole supply chain. In economics this is known as *double marginalization*. Because each firm naturally considers its own margin, the decisions are not what a centralized decision maker would do.

The issue of double marginalization arises because the firms are following a standard wholesale price contract agreement. With different contracts, it is often possible to align incentives of each party so that the optimal supply chain decision is reached; in this case, the supply chain is said to be *coordinated*. For example, many book publishers offer buy-back agreements to bookstores which allow the stores to return unsold copies for a full refund. In this way, the overage cost to retailers is reduced and they are encouraged to stock more. Pasternack (1985) showed how the right buy-back contract can coordinate the supply chain, resulting in system optimal profit; the question then becomes how to split the profits. Assuming that a win-win arrangement can be found, where both parties benefit, both parties will enter into such an agreement. However, buy- back contracts can be problematic in environments where the supplier really does not want to take the goods back or the goods may be damaged in transit.

Revenue sharing contracts are another type of contract that, with the right parameter choices, may coordinate the supply chain. In this case, the retailer pays much less for the goods up front but gives the supplier a portion of all revenue earned. Because there is less risk of overstock, the retailer is again incentivized to order more than they would with a simple wholesale price contract. Such contracts have proved particularly effective in the movie rental industry, where low upfront DVD (and previously video) costs incentivize the store to stock more, which both makes for happier customers and more rentals. The movie studios then take a cut from every movie rented.

Vendor-managed inventory is another way to mitigate double marginalization and coordinate the supply chain. Because the supplier is deciding inventory at the retailer's facility it need not consider the retailer's margins. Of course, restrictions such as shelf-space limitations need to be in place for such contracts to ensure the supplier doesn't overstock on the retailer's behalf.

There are many more kinds of contacts than those described here that, in the right situations, work to coordinate the supply chain. However, there has also been significant behavioral research to evaluate how such contacts perform in practice, when the parties may have considerations beyond expected profit, such as perceptions of fairness. Somewhat surprisingly, behavioral considerations often work against supply chain coordination, and in lab studies wholesale price contracts can do better than profit-maximizing theory would predict (e.g., Loch and Wu, 2008). This combined with their simplicity probably explains the continued widespread use of wholesale price contracts in practice.

Problems for Section 6.10

32. How is double marginalization affected by the profit margin of the retailer ($p - w$) relative to the profit margin of the whole supply chain ($p - c$)? Is its effect greatest when most of the profit margin sits with the retailer or with the supplier?
33. Why might it not be practical for the supplier in Example 6.5 to simply tell the retailer to order 701 jackets and give the supplier a kick-back for the profit differential? Can you think of different supply chains where such an arrangement might be effective?
34. What types of industries, beyond book publishing, are likely to find buy-back agreements effective?
35. What types of industries, beyond movie rentals, are likely to find revenue sharing agreements effective? When are they least likely to be practical?
36. Why might the supplier be likely to overstock (relative to what is supply chain optimal) in a VMI arrangement if shelf-space restrictions or similar are not in place by the retailer?

6.11 GLOBAL SUPPLY CHAIN MANAGEMENT

Economic barriers between countries are coming down. China, a long time holdout, now participates in a free trade area with the Association of Southeast Asian Nations (ASEAN) and in 2008 signed a bilateral free trade agreement with New Zealand (its first such agreement). Today, few industries only produce in and serve their home market.

One example of a dramatic shift in the marketplace occurred in the automobile industry. Consider the experience from Womack et al. (1990). As a child in the 1950s, there was no question that the family car would be an American nameplate because almost *everybody* in the United States purchased American cars in those years. Foreign automakers began to make inroads into the American market during the late fifties and early sixties and, by 1970, the American firms had given up about 20 percent of the domestic market. During the 1970s, the market share of U.S.-based firms eroded more quickly. While the American firms clung to old designs and old ways of doing business, foreign manufacturers took advantage of changing tastes. The big winner was the Japanese who saw their share of the world auto market soar from almost zero in the mid-fifties to about 30 percent by 1990 (Womack et al., 1990). In 2007 Toyota surpassed American General Motors to take the world's top spot for quarterly sales (Chozick & Shirouzui, 2007).

In addition to the fact that U.S. consumers are buying more foreign cars, it is also true that U.S. firms are producing more cars in other countries, and foreign competitors are producing more cars in the United States. Mexico, and to a lesser extent Canada, have also become key manufacturing locations for U.S.-headquartered companies' automobile production. However, as of the time of this writing, China is the world's largest automobile producer with annual production that exceeds that of the United States and Japan combined. Nevertheless, China's labor costs have been rapidly rising in recent years (by 500 percent between 2000 and 2012) and are expected to continue to rise at a rate of 18 percent per year (Mayer, 2013). Another country may yet emerge as the world's largest automobile producer.

Automobile supply chains are simply one example of supply chains that have undergone major changes in the last few decades. In general, supply chains have become both more global and more complex. For example, between 1995 and 2007 the number of transnational companies more than doubled from 38,000 to 79,000, and foreign subsidiaries

nearly tripled from 265,000 to 790,000, according to a supply chain survey by IBM (IBM, 2009). Further, managing global supply chains is made more challenging by the heterogeneity of government and local regulations and cultures across different countries.

A further complexity in managing global supply chains is managing volatile exchange rates. If revenue is reported in U.S. dollars but earned in a foreign currency then any changes in exchange rates can directly affect reported earnings. Similarly, if suppliers are paid in their local currency then exchange rate changes will affect the home country's reported costs. Companies often use *exchange rate hedging* to mitigate the risk from exchange rate movements. Currency options are purchased so that if the exchange rate moves in a favorable direction for the firm then there is no payout but if it moves in an unfavorable direction then the options pay out. The cost of the options depend in part on the *strike price* for the option, which is the point at which they begin to pay out. Southwest Airlines used fuel price options to successfully hedge against fuel price rises in the mid-2000s and is said to have, saved approximately $3.5 billion through fuel hedging between 1999 and 2008 (New York Times, 2008). Of course, when fuel prices dropped during the global financial crisis in 2009 these hedges were less successful, but such is the fundamental nature of options.

Fuel costs are a key issue in supply chain design. In particular, fuel prices have tripled between 2000 and 2012 (Mayer, 2013). One effect of rising fuel costs has been the decision by shipping companies to use *slow steaming*, where vessels designed to travel at 25 knots are deliberately slowed to between 18 and 20 knots to reduce the usage of bunker fuel. In fact, some companies are even using super slow steaming (12 to 14 knots) to reduce costs even further. Clearly, this has significant negative implications with respect to supply chain responsiveness and is particularly problematic for supply chains of perishable items.

One response to increasing labor costs in China and increasing shipping costs due to fuel is the rise of *on-shoring* or *re-shoring*, which is simply the returning of manufacturing operations that were offshored to back in-country (see also Chapter 1). This is most common in the United States where labor unions have become more flexible and automation has made routine tasks more cost effective (Booth, 2013). Another factor in this decision is to decrease risk, particularly in the face of increasing global tensions. In food supply chains, concerns over food miles, food safety, and sustainability have also led to changing supply chain configurations. Clearly, supply chain designs must be regularly evaluated and rethought as global trends change the key trade-offs that lead to one location being chosen over another.

Problems for Section 6.11

37. What do you see as the new emerging markets in the world in the next 20 years? For what reason has Africa been slow to develop as a new market and as a desirable location for manufacturing?
38. As noted earlier in this chapter, the share of U.S. sales accounted for by multinational firms is increasing. What events might reverse this trend?
39. What difficulties for supply chain management are created by the growth of globalization?
40. What industries do you think are particularly likely to have onshore manufacturing operations in the next few years?
41. What is the downside to using currency option hedging?

6.12 SUMMARY

With the rise of the information age has come the rise of interest in supply chain management. When firms can see data on supply chain costs they can work to reduce them, and there is anecdotal evidence to suggest that firms first worked to push costs out of manufacturing and then out of their supply chains (Henkoff, 1994). Because of this, supply chain software has become big business. Most of the biggest names in the information systems arena, including Oracle Corporation and German-based SAP, now offer supply chain modules as part of their total system solutions.

Both information systems and new technologies have allowed information to be gathered and shared in ways that were not previously possible. Supply chain partners can share data on market trends or even manage each other's inventory through vendor-managed inventory agreements. Point-of-sale entry systems based on barcoding are now ubiquitous and RFID technology solutions are becoming increasingly popular as well.

With information, or "big data" as it is often called these days, has come the rise of business analytics and mathematical modeling can play an important role in efficient supply chain management. The transportation and transshipment problems, discussed in this chapter, are examples of the kind of mathematical optimization models that can assist firms with determining efficient schedules for moving product from the factory to the market. There are also mathematically based techniques for efficient scheduling of delivery vehicles.

Delivering the product to the consumer was traditionally a secondary consideration for manufacturing firms. Today, however, supply chain considerations are taken into account even at the product design level. Products that are easily stacked can be shipped and stored more easily and less expensively. The strategy of postponement has turned out to be a fundamental design principle that has proven to be highly cost effective. By designing products whose final configuration can be postponed, firms can delay product differentiation. This allows them to gain valuable time in determining the realization of demand and also to pool intermediate inventory for a variety of end products.

An unusual phenomenon that was observed in the late 1980s is the bullwhip effect. The variance of orders seems to increase dramatically as one moves up the supply chain. While most agree that the bullwhip effect is the result of different agents acting to optimize their own positions, solutions for this problem are not trivial. Information sharing, reduced order batching, and decreased order lead times will help but they do not relieve the issue of each party operating under its own incentives.

Misaligned incentives in the supply chain are a major source of supply chain inefficiency. About the only way to mitigate these issues are through a change in approach to partnerships and contracts. When different parties in the supply chain consider themselves to be partners then they are more likely to work together to find win-win solutions, namely solutions that both increase the performance of the supply chain and also increase each party's individual profit. Moving beyond wholesale price contracts to agreements such as buy-back contracts, revenue sharing contracts, and vendor managed inventory agreements can all help to coordinate the supply chain if operated under appropriate parameter choices.

The problem of misaligned incentives becomes exacerbated in global supply chains because the parties are often located in different geographic areas, may come from different cultural backgrounds, and may have different local incentives provided by their governments. Variable exchange rates, rising fuel costs, and increased global tensions also form challenges for global supply chain management. Sustainability and risk management have also become of increasing importance to many global corporations.

Bibliography

Arntzen, B. C., G. G. Brown, T. P. Harrison, and L. L. Trafton. "Global Supply Chain Management at Digital Equipment Corporation." *Interfaces* 25 (1995), pp. 69–93.

Arrow, K. J. "Historical Background." Chapter 1 in *Studies in the Mathematical Theory of Inventory and Production*, ed. K. J. Arrow, S. Karlin, and H. Scarf. Stanford, CA: Stanford University Press, 1958.

Barrone, F. Private communication, 1996.

Bassamboo , A., R. S. Randhawa, and J. A. Van Mieghem. "A Little Flexibility Is All You Need: On the Asymptotic Value of Flexible Capacity in Parallel Queuing Systems." *Operations Research* 60 (2012), pp. 1423–1435.

Bassamboo , A., R. S. Randhawa, and J. A. Van Mieghem. "Optimal Flexibility Configurations in Newsvendor Networks: Going Beyond Chaining and Pairing." *Management Science* 56 (2010), pp. 1285–1303.

Bell, W. J. "Improving the Distribution of Industrial Gases with an Online Computerized Routing and Scheduling Optimizer." *Interfaces* 13, no. 6 (1983), pp. 4–23.

Booth, T. "Here, there and everywhere: After decades of sending work across the world, companies are rethinking their offshoring strategies." *The Economist,* Jan 19th 2013. Accessed from http://www.economist.com/news/special-report/21569572-after-decades-sending-work-across-world-companies-are-rethinking-their-offshoring

Bray, R. L., and H. Mendelson. "Information Transmission and the Bullwhip Effect: An Empirical Investigation." *Management Science* 58, no. 5 (2012), pp. 860–875.

Bustillo, M. "Wal-Mart Radio Tags to Track Clothing." *Wall Street Journal,* July 23, 2010.

Cachon, G. P., T. Randall, and G. M. Schmidt. "In Search of the Bullwhip Effect." *Manufacturing and Service Operations Management* 9, no. 4 (2007), pp. 457–479.

Cannon, E. EDI Guide. A Step by Step Approach. New York: Van Nostrand Reinhold, 1993.

Chozick, A. and Shirouzui, N. "GM Slips Into Toyota's Rearview Mirror. Japanese Firm Passes U.S. Rival for First Time in Quarterly Global Sales." *Wall Street Journal*, April 25, 2007. Accessed from http://online.wsj.com/article/SB117739853275580259.html

Clark, A. J., and H. E. Scarf. "Optimal Policies for a Multiechelon Inventory Problem." *Management Science* 6 (1960), pp. 475–90.

Clarke, G., and G. W. Wright. "Scheduling of Vehicles from a Central Depot to a Number of Delivery Points." *Operations Research* 12 (1964), pp. 568–81.

Cohen, M.; P. V. Kamesam; P. Kleindorfer; H. Lee; and A. Tekerian. "Optimizer: IBM's Multi-Echelon Inventory System for Managing Service Logistics." *Interfaces* 20, no. 1 (1990), pp. 65–82.

Cooke, J. A. "Software Takes Babel out of Vendor Managed Inventory." *Logistics Management & Distribution Report* 38, no. 2 (1999), p. 87.

Copacino, W. C. *Supply Chain Management: The Basics and Beyond*. Boca Raton, FL: St. Lucie Press, 1997.

Crawford, F. A. "ECR: A Mandate for Food Manufacturers?" *Food Processing*, February 1994.

Dauzere-Peres, S., et al. "Omya Hustadmarmor Optimizes Its Supply Chain for Delivering Calcium Carbonate Slurry to European Paper Manufacturers." *Interfaces* 37 (2007), pp. 39–51.

Degbotse , A., B. T. Denton, K. Fordyce, R. J. Milne, R. Orzell, C. T. Wang. "IBM Blends Heuristics and Optimization to Plan Its Semiconductor Supply Chain." *Interfaces* 43, no. 2 (2013), pp. 130–141.

Dinning, M., and E. W., Schuster. "Fighting Friction," *APICS—The Performance Advantage*, February 2003.

Dornier, P.-P.; R. Ernst; M. Fender; and P. Kouvelis. *Global Operations and Logistics, Text and Cases*. New York: John Wiley & Sons, 1998.

Dreifus, S. B., ed. *Business International's Global Desk Reference*. New York: McGraw-Hill, 1992.

Duff, C., and R. Ortega. "How Wal-Mart Outdid a Once-Touted Kmart in Discount Store Race." *The Wall Street Journal*, March 24, 1995.

Eilon, S.; C. D. T. Watson-Gandy; and N. Christofides. *Distribution Management: Mathematical Modeling and Practical Analysis*. London: Griffin, 1971.

Eppen, G. D. "Effects of Centralization on Expected Costs in a Multi-Location Newsboy Problem." *Management Science* 25 (1979), pp. 498–501.

Eppen, G. D., and L. Schrage. "Centralized Ordering Policies in a Multi-Warehouse System with Leadtimes and Random Demand." In *Multi-Level Production/Inventory Systems: Theory and Practice*, ed. L. B. Schwarz. New York: North Holland, 1981.

Erkip, N.; W. H. Hausman; and S. Nahmias. "Optimal Centralized Ordering Policies in Multi-Echelon Inventory Systems with Correlated Demands." *Management Science* 36 (1990), pp. 381–92.

Fine, C. H. Clockspeed. Reading, MA: Perseus Books, 1998.

Fisher, M. L. "What is the right supply chain for your product?" *Harvard Business Review* (1997), pp. 105–106.

Granneman, S. "RFID Chips Are Here," *The Register,* accessed June 27, 2003.

Hammond, Jan. Barilla SpA (A and B). Copyright © 1994 by the President and Fellows of Harvard Business School.

Handfield, R. B., and E. L. Nichols Jr. *Introduction to Supply Chain Management*. Upper Saddle River, NJ: Prentice Hall, 1999.

Henkoff, R. "Delivering the Goods." *Fortune*, November 25, 1994, pp. 64–78.

IBM. "The Smarter Supply Chain of the Future." *Global Chief Supply Chain Officer Study*. Accessed from http://www-07.ibm.com/events/my/industrialinsights/pdf/01_Randy_Sng_Smarter_SC_for_Mfg-10Dec-MY.pdf

Jacobson, G., and J. Hillkirk. *Xerox, American Samurai*. New York: Macmillan, 1986.

John, C. G. and M. Willis. "Supply chain re-engineering at Anheuser-Busch." *Supply Chain Management Review* 2 (1998), pp. 28–36.

Lee, H. L., C. Billington, and B. Carter. "Hewlett-Packard Gains Control of Inventory and Service through Design for Localization." *Interfaces* 23, no. 4 (1994), pp. 1–11.

Lee, H. L., P. Padmanabhan, and S. Whang. "The Bullwhip Effect in Supply Chains." *Sloan Management Review*, Spring 1997, pp. 93–102.

Loch, C. and Y. Wu. (2008). "Social Preferences and Supply Chain Performance: An Experimental Study." *Management Science* 54, no. 11 (2008), pp. 1835–1849.

Martin, A. J. *DRP: Distribution Resource Planning*. 2nd ed. Essex Junction, VT: Oliver Wight Limited Publications, 1990.

Mayer, L. "Why Onshoring High-tech Manufacturing Jobs Makes Economic Sense." *Huffington Post*, January 24, 2013. Accessed from http://www.huffingtonpost.com/linda-mayer/manufacturing-makes-economic-sense_b_2533593.html

Muckstadt, J. A., and L. J. Thomas. "Are Multi-Echelon Inventory Methods Worth Implementing in Systems with Low Demand Rate Items?" *Management Science* 26 (1980), pp. 483–94.

Nahmias, S., and S. A. Smith. "Mathematical Models of Retailer Inventory Systems: A Review." In *Perspectives in Operations Management*, ed. R. K. Sarin. Boston: Kluwer, 1993.

Nahmias, S., and S. A. Smith. "Optimizing Inventory Levels in a Two-Echelon Retailer System with Partial Lost Sales." *Management Science* 40 (1994), pp. 582–96.

New York Times. "Airlines try to hedge oil costs to stay in business." June 30, 2008. Accessed from http://www.nytimes.com/2008/06/30/business/worldbusiness/30iht-30hedge.14104427.html?_r=0

Palevich, R. F. "Supply Chain Management." *Hospital Materiel Management Quarterly* 20, no. 3 (1999), pp. 54–63.

Pasternack, B. "Optimal pricing and returns policies for perishable commodities." *Marketing Science* 4 (1985), pp. 166–176.

Pazour, J. A., and L. C. Neubert. "Routing and Scheduling of Cross-Town Drayage Operations at J.B. Hunt Transport." *Interfaces* 43, no. 2 (2013), pp. 117–129.

Ragsdale, C. T. *Spreadsheet Modeling and Decision Analysis*. 2nd ed. Cincinnati: South-Western, 1998.

Ross, David Frederick. *Competing through Supply Chain Management: Creating Market-Winning Strategies through Supply Chain Partnerships*. New York: Chapman & Hall, 1998.

Scheeres, J. "Tracking Junior with a Microchip," *Wired News*, October 10, 2003.

Schrage, L. "Formulation and Structure of More Complex/Realistic Routing and Scheduling Problems" *Networks* 11 (1981), pp. 229–32.

Schuster, E. W.; S. J. Allen; and D. L. Brock. *Global RFID: The Value of the EPCglobal Network for Supply Chain Management*. Berlin: Springer, 2007.

Schwarz, L. B., ed. *Multi-Level Production/Inventory Systems: Theory and Practice*. New York: North Holland, 1981.

Signorelli, S., and H. Heskett. "Benneton." Harvard Business School Case 9-685-014, 1984.

Simchi-Levi, D.; P. Kaminski; and E. Simchi-Levi. *Designing and Managing the Supply Chain: Concepts, Strategies, and Case Studies*. New York: McGraw-Hill/Irwin, 1999.

Sliwa, C. "Users Cling to EDI for Critical Transactions." *Computerworld* 33, no. 11 (1999), p. 48.

Sterman, R. "Modeling Managerial Behavior: Misperception of Feedback in a Dynamic Decision Making Experiment." *Management Science* 35, no. 3 (1989), pp. 321–39.

Supply Chain Forum. http://www.stanford.edu/group/scforum/, accessed August 1999.

Varchaver, N. "Scanning the Globe: The Humble Bar Code Began as an Object of Suspicion and Grew into a Cultural Icon. Today It's Deep in the Heart of the Fortune 500." *Fortune Magazine,* May 31, 2004. Accessed from http://money.cnn.com/magazines/fortune/fortune_archive/2004/05/31/370719/index.htm

Wilder, C., and M. K. McGee. "GE: The Net Pays Off." *Informationweek*, February 1997, pp. 14–16.

Womack, J. P., D. T. Jones, and D. Roos. *The Machine That Changed the World*. New York: Harper Perennial, 1990.

Yansouni, C. Private communication, 1999.

Chapter **Seven**

Service Operations Management

> "The goal as a company is to have customer service that is not just the best, but legendary."
>
> —Sam Walton, Founder of Wal-Mart

Chapter Overview

Purpose

To understand the challenges unique to managing service operations and to learn key tools for matching supply with demand in services.

Key Points

1. *What is a service?* To be considered a service there must be an intangible portion to the offering; that is, it involves something that is not a good, or more informally, something that cannot be dropped on your foot. A service is also usually time-perishable; it cannot be stored. Further, it frequently involves the customer as co-producer; the service cannot take place without the customer's involvement.
2. *Service operations strategy.* While strategy within service operations is fundamentally similar to general operations strategy, as discussed in Chapter 1, there are two particular features that make it more challenging for a service firm. First, in part, because capacity in services tends to correspond to people rather than machines, there is a "fear of focus" that develops for many firms as they try to be all things to all customers. Second, defining and measuring quality is more difficult for services than for goods because it tends to be more subjective.
3. *Bottleneck analysis.* A system's capacity is the capacity of its bottleneck; that is, the step or resource in the system that can process the fewest customers per hour. In order to find the bottleneck, the capacity of all resources must be calculated. The utilization of a resource is the rate at which customers arrive divided by the capacity (i.e., the rate at which customers can be processed). Thus, the bottleneck resource will have the highest percentage utilization.
4. *Poisson arrivals.* Arrivals to many service systems are both unscheduled and highly variable. The Poisson process is often a good model for these types of systems, particularly over the short term; it can be used to make system predictions. The model is based on an assumption of independent behavior by a large number of potential customers.

5. *Pooling.* As in Chapter 6, pooling is a key technique for mitigating uncertainty and improving planning. In service systems, pooling strategies involve either combining variable streams of arrivals into one larger, inherently less variable, stream or cross-training staff so that they may serve multiple classes of customer.
6. *Queueing systems.* A queue (waiting line) represents customers waiting for service. The structural aspects of queueing models include: arrivals, service, queue discipline, capacity of the queue, number of servers, and the network structure. Measures of performance of queueing systems can be determined analytically for simpler cases, and by simulation for more complex cases.
7. *The M/M/1 queue.* One of the simplest queueing models is known is the M/M/1 queue. It assumes Poisson arrivals and exponentially distributed service times. Although rarely completely accurate, it yields simple formulas for various system measures.
8. *Little's law.* This "law" states that the average number of customers in the system is equal to the average customer's time in system multiplied by the customer arrival rate. Because it is an equation, any one of the values can be calculated if the other two are known. For example, average time in system can be calculated by dividing the average number of customers in the system by the arrival rate.
9. *Incentives in services.* Like any operational system, metrics drive performance and incentivize behavior in service systems. A common contract in services is a Service Level Agreement (SLA) where the firm agrees to meet a specified goal for service a certain percentage of the time. A widely used type of SLA is a delay percentile contract, which puts an upper bound on waiting time for a certain percentage of customers. Unfortunately, as discussed in the chapter, this type of contract can lead to perverse incentives for the service provider; there are more effective contracts available.
10. *The human element.* Human behavior is more important to consider in the design of service systems than for most production systems, because services frequently involve the customer as co-producer, rather than simply being a consumer. For example, a self-service system must be intuitive and pleasant for new customers, who have never encountered the system before, while being sufficiently fast to use for existing customers.
11. *Revenue management.* An important technique for managing variable customer arrivals and perishable capacity, as are typical for service systems, is to use differentiated prices, or revenue management. For example, airlines will price seats higher if the plane has little spare capacity and lower if there are a large number of unsold seats. Further, they will try to segment the customers by willingness to pay, so that business travelers, who need more flexibility and perhaps a more comfortable trip, will typically pay more for a ticket than highly price-sensitive leisure travelers, who are going to the same destination.

One of the benefits of social media is instant access to the opinions of others. A popular site for finding out what others think about consumer services is Yelp. If you find yourself in a new city and want to find a restaurant, Yelp provides you with feedback from other patrons. Moreover, as we know, restaurants are rated on both the quality of the food and the quality of the service. What used to be a simple word of mouth process has now been elevated to a new level. However, Yelp reviews can be misleading as well. One of the authors recently went to Yelp to read the reviews of local

pool services. One service was located with five glowing reviews on Yelp. Unfortunately, these reviews were not very reliable. During the month the firm was treating the pool, the algae became so bad that the pool turned an emerald green and they broke some equipment to boot. It is likely that they or family members and friends wrote their reviews. (And yes, the author did his civic duty and posted a very negative review.)

The previous example illustrates the difficulty of measuring quality in service systems. In part, this is because the fundamental principle that differentiates a service from a good is an element of *intangibility*, which means something that cannot be perceived by the sense of touch. This intangibility also means that services typically cannot be inventoried in advance of consumption and hence are *time-perishable*. For example, a movie shown in a theater is of no use to a customer if they are not there at the same time as it is being shown.

Another distinctive feature of services is that they typically involve the customer as *co-producer*. Whether it is the customer selecting and consuming the food in a restaurant or riding the roller coaster in a theme park, the service does not happen without the customer helping to "produce" the service.

The line between goods and services is not solid, especially in light of the trend towards servicization, as discussed in Chapter 1. Restaurants and retailers are usually classed as service businesses, despite the fact that the former needs food for the transaction and the latter sells goods. A typical representation of this idea is that there is a continuum of industries from pure experience services (such as movies) on one end, to commodity goods producers (such as wheat farmers) on the other. Retailers fall around the middle of the continuum and are to the right of most services but to the left of most production industries.

In marketing literature, a popular trend is known as *service dominant logic* (SDL). It states that "all firms are service firms" because all firms provide "the applications of competences (knowledge and skills) for the benefit of a party" (Lusch and Vargo, 2014). In some sense, this idea is similar to the process view of a firm, often taken in operations, which is that firms use process competencies to convert input to outputs. The term *services* (as opposed to simply "service") is used to reflect service operations as defined in this chapter.

As mentioned above, services typically cannot be inventoried. Therefore, when there is the inevitable mismatch between supply and demand, customers either wait or leave without service. The presence of variability exacerbates this mismatch. There are five key types of variability that must be planned for in most service systems as follows (Frei, 2006):

1. *Arrival variability*. There is variability in the timing of customer arrivals to the service by day of the week, time of day, number of customers in a group, and even from minute to minute.
2. *Request variability*. There is variability in customer expectations and the type of service they wish to consume.
3. *Capability variability*. There is variability both in the ability of staff to provide the service and in customer abilities when they serve as co-producer.
4. *Effort variability*. There is variability in the effort put into the service both by staff and by customers.
5. *Subjective preference variability*. Even if customers receive identical services, there is still variability in how they perceive the service due to individual and subjective preferences. For example, the mood of the customer affects his perceptions, as do his individual tastes.

Improving the performance of a service system typically involves either reducing or better accommodating at least one of these types of variability. This chapter considers both possibilities. It provides tools for mitigating the mismatch between supply and demand that adds cost to any service system.

The two key goals for the chapter are:

- to foster readers' abilities to analyze services with regards to their potential to deliver the services promised; and
- to provide readers with tools that they can apply to the design and improvement of service systems.

7.1 SERVICE OPERATIONS STRATEGY

As discussed in Chapters 1 and 6, operations strategy involves trade-offs; managing service operations is no different. The most common trade-off in services is between quality and cost. However, quality is more difficult to measure in services and customers may not be willing to pay for it. For example, Bose can charge a premium for their headphones because their superior fidelity can be measured. Whereas, McDonald's cannot charge a premium for their friendly service, even though one of the authors had a friend who was fired from McDonald's for failing to smile at an undercover McDonald's quality assessment employee who was posing as a customer!

In this section, we first discuss where services fit in the economic landscape. Then, we consider what is meant by service quality. We discuss the key decisions that must be made in positioning a service in the marketplace and how the five key types of variability are best planned for. We conclude by discussing service competition and how it differs from competition among goods-producing firms.

The Service Economy

The services sector of the economy is also called the *tertiary sector* in a three-sector framework developed by economists Fisher (1935), Clark (1940), and Fourastié (1949). Under this model of the economy, the *primary sector* is extractive, including mining, agriculture, fishing, forestry, etc. The *secondary sector* is goods producing. The tertiary sector is all services. This tertiary sector is sometimes broken up further to include *domestic services* (including restaurants and hotels, barber and beauty shops, laundry and dry cleaning, and maintenance and repair), *trade and commerce services* (including transportation, retailing, real estate, communication, and finance and insurance), *refining and extending human capacities* (including health, education, research, recreation, and the arts), and the *experience economy*, that in its purest form provides entertainment value only (e.g., theme parks).

In 1850 the primary sector in the United States made up over 60 percent of all employment, with the secondary and tertiary sectors accounting for under 20 percent each. Throughout the 20th century, the percentage of manufacturing jobs grew as the percentage of jobs in agriculture shrunk. Then, later in the century, the percentage of service jobs grew as the percentage of manufacturing jobs decreased. In 2012, the primary sector was 1.5 percent of all employment in the United States and the secondary sector was 17.2 percent (down from a high of above 30 percent in the middle of the 20th century). As was discussed in Chapter 1, there is an argument that civilizations move naturally from the primary sector to the tertiary sector, and indeed this can be seen in the following table of percentage employment in service jobs for various developed countries.

TABLE 7–1 **Percentage Employment in Service Jobs**

	1965	1975	1985	1995	2005	2012
U.S.	59.5	66.4	70.0	74.1	78.6	81.2
U.K.	51.3	58.3	64.1	71.4	77.0	81.2
Canada	57.8	65.8	70.6	74.8	76.0	78.6
Australia	54.6	61.5	68.4	73.1	75.8	77.7
France	43.9	51.9	61.4	70.0	74.8	76.5
Japan	44.8	52.0	57.0	61.4	68.6	71.5
Italy	36.5	44.0	55.3	62.2	65.5	69.8

(Source: U.S. Bureau of Labor Statistics)

While the numbers in Table 7–1 show a clear increasing trend, it seems highly unlikely that for any large country these numbers will eventually reach 100 percent. (See Problem 3 for whether 100 percent services might be practical for a small country.) This is because a pure service economy would use no labor to produce food or goods. Further, as was argued in Chapter 1, manufacturing matters for a country's economy. It is interesting to speculate whether the numbers will eventually reach an asymptote, and if so, at what value? What is the long-term stable mix between production and services going to prove to be? It is clear that more leisure time and/or wealth naturally lead to increased demand for services. However, whether we will continue to see more of the former is beyond the scope of this text.

Over 20 years ago, Fortune magazine predicted that "in the new U.S. economy, service—bold, fast, unexpected, innovative, and customized—is the ultimate strategic imperative . . . . Everyone has become better at developing products. The one place you can differentiate yourself is in the service you provide" (Henkoff, 1994). While these predictions were a little premature, they do emphasize the importance of service as a competitive strategy. While novel products will always have their place in the marketplace, service quality is indeed an important piece of a firm's competitive strategy.

Service Quality

As mentioned earlier, in most service operations the primary trade-off is between cost and quality; however, how quality is defined differs from application to application. In a fast food restaurant, it is speed of service and consistency of food and experience; while in a romantic restaurant, it includes atmosphere and attentiveness of staff. Therefore, quality in services must be defined relative to customers' expectations.

Key elements that define service quality include the following (cf., Metters et al., 2002): the *consistency* of the service; the *delay* incurred before service; the ability of staff to perform the promised service *dependably*, *accurately*, *promptly*, *courteously*, and with a *friendly demeanor*; the *appearance* of physical facilities, equipment, personnel, and communication materials, which help set the *atmosphere* for the service; the appropriate level of *communication* with the customer; the ease with which the service is *accessed*, which includes hours of operation, location, and availability of the appropriate server; the level of *personalization* of the service; the *pleasure* or "fun" level of the service (where appropriate); the *credibility* and *technical level* of the service; and the *safety* and *security* of the service.

The elements of service quality that are most important will obviously depend on the industry, but all will be important to some extent. For example, many people will choose a bank depending on its convenience, banking product offerings, and security that it offers, but may switch banks if the staff treat them poorly or fail to be sufficiently friendly. Who among us has not been put-off by an unhelpful or unfriendly staff member in some service environment? As humans, we tend to avoid unnecessary sources of conflict or stress and will simply switch to another firm rather than risk another unpleasant encounter. Most problematic for the firm is that we are unlikely to tell them what occurred, preventing them from taking corrective action.

Measuring Quality

Because there are so many elements that define quality, it is typically more difficult to measure quality for services than for goods. Even defining a "defect" can be problematic. For example, if the service was performed as prescribed but the customer was not happy because he misunderstood part of the experience or simply entered into the experience in a negative mood, should that be considered a faulty service? The answer to this question will depend on what the service provider is going to do with this information. Clearly, the staff member involved should not be penalized. However, the information should be collected so that the process can be redesigned to be more "foolproof."

Quality in services is often measured by customer surveys, yet this can be a problem if unhappy customers disengage and therefore do not fill in the survey at the same rate that satisfied customers do (or if the reverse occurs and only dissatisfied customers give feedback). In addition, if there is separation between the time the service is consumed and the time the survey is asked, the customer may not accurately remember his perception of the service at the time. One tool for mitigating this issue is to place push-button perception collectors at the end of the service. For example, London's Heathrow Airport has small kiosks with four buttons ranging from a very sad face to a very happy face. Arriving international travelers are encouraged to push one of the buttons to indicate how their customs and immigration experience was. It is not clear how this information is used, but clearly, Heathrow has eliminated the immediacy issue.

Controlling Quality

Another complication in service quality is that controlling defects when customers are involved in co-production can be highly problematic. Even if the issue is the customer's fault, they will quite naturally blame the process, and hence the service provider. Some service providers try to "train" customers in order to speed up service, which will improve the quality of the experience. For example, Starbucks coffee shops have the servers repeat the customer's order in the preferred sequence (e.g., size before coffee type) and jargon (skinny rather than skim milk) in the deliberate desire to make regular customers follow this "language"; this both increases process efficiency and grows brand loyalty.

Paying for Quality

High quality services are usually more costly than lower quality services and therefore must have a revenue source. While most people expect to pay more at a romantic restaurant than at McDonald's, most customers also expect servers from any business to have a pleasant demeanor and sufficient skills to provide the service properly. Many service providers struggle with the issue of whether to hire staff

based on ability, personality, or both (which will cost more). If they have both able and personable staff, they must work out how to retain them through either higher salary or increased job benefits, and also how to fund this. If customers are not willing to pay for both ability and personality, then the firm needs to decide which dimension they are willing to compromise on. For example, many banks by necessity hire for ability, but Commerce Bank set out to distinguish itself by simplifying its account offerings and therefore being able to hire for personality (Frei & Hajim, 2002). Banks are an example of a service industry where it is difficult to extract revenue for high quality service.

A commonly cited statistic, used to emphasize the importance of quality service, is that it costs five times more money to acquire a new customer than to retain a current customer (e.g., Hart et al., 1990). Further, Reichheld (1996) argues that it is more profitable to serve long-time customers because they purchase more frequently and can be served more efficiently. Service businesses must therefore work hard to understand customers' preferences and retain current customers.

Key Decisions for Service Businesses

There are four key decisions that any service business must make (Frei, 2008).

1. *The offering*. What precisely is the service provider going to offer customers? Will there be a variety of options for customers to choose among or one standard service offering? Can customers customize their experience or not? What types of customers will be served? Is this designed to be a long-term customer relationship or a more transactional service?
2. *The funding mechanism*. There are usually more ways to charge customers than simply fee-for-service. For example, banks charge account fees, have flexibility in the interest they both pay out and charge, and choose the transaction fees they apply. Airlines charge passengers for the ticket but also often for bags or seat booking preferences.
3. *The employee management system*. What management structure will be used in the organization? How will employees be trained in the various processes? How will the processes be structured to accommodate employee variation and to foolproof the service? What type of environment or atmosphere will be created in the work place?
4. *The customer management system*. What type of environment will the customer experience? How much of the service will the customer be expected to participate in? How will the customer be communicated with and how will customers communicate their needs? Is this to be a long-term relationship with the customer, and if so, how will this be promoted?

These decisions define how the service provider competes in the marketplace. There are a couple of key points for managers to keep in mind. First, a firm's culture is not happenstance—it is the result of the company's (hopefully deliberate) decisions. As described in the Snapshot Application below, Southwest Airlines made a deliberate decision to be a "fun" place to work and has aligned its policies to promote this. Second, firms can design customer policies in ways that help to improve the work experience of their employees. For example, most airlines have moved to self-service check-in for domestic travel. This reduces the workload for ticket agents. It can also lead to more satisfied customers because such terminals allow customers to select their preferred seating.

Snapshot Application

SOUTHWEST AIRLINES COMPETES WITH SERVICE

Southwest Airlines is the darling of operations management texts for a number of reasons. First, it has been successful partly due to effective operations management. Second, it is one of the very few U.S. airlines that have never filed for bankruptcy. Finally, it has a very well-thought-out corporate and business strategy that match well with the firm's operational strategy.

Southwest's effective operations are based on their overarching goal of short turnaround times at airports. Because planes are very capital intensive, and only planes in the air earn money, being able to have short turnaround times has led to extra flights and hence extra revenue. Many of their decisions are guided by this goal. They only operate Boeing 737s, which (a) reduces the complexity of their maintenance and spare-parts systems; (b) reduces the training needed for their pilots; and (c) makes it easier for one plane to be substituted for another should there be a problem. They also typically fly in and out of less congested secondary airports (e.g., Chicago's Midway airport rather than Chicago O'Hare). They do not charge for checked baggage. This is due in part to the company's desire to avoid departure delays caused by congestion in the passenger compartment.

The mission of Southwest Airlines is "dedication to the highest quality of customer service delivered with a sense of warmth, friendliness, individual pride, and Company Spirit." Further, Southwest states the following: "We are committed to provide our employees a stable work environment with equal opportunity for learning and personal growth. Creativity and innovation are encouraged for improving the effectiveness of Southwest Airlines. Above all, employees will be provided the same concern, respect, and caring attitude within the organization that they are expected to share externally with every Southwest Customer." Their golden rule of "do unto others as you would have them do unto you" is stated explicitly and taught to all employees.

Southwest has realized that high service quality can only come through empowered, happy, and loyal employees. Southwest works hard to achieve such a workforce. They state that they "hire for attitude and train for skill." Notice they do not pretend that they can hire for both attitude and skill and still be a low-cost airline. Instead, they hire for attitude and make sure they have processes in place to ensure employees can succeed. Their CEO, Gary Kelly, has stated that "our people are our single greatest strength and most enduring long-term competitive advantage." They have innovative hiring processes, such as asking applicants to share their most embarrassing moment and then observing the empathy of the other participants, to ensure that they are indeed hiring for attitude.

Southwest's definition of "high-quality service" is, of course, informed by it being a low cost airline. While they do still offer free in-flight nonalcoholic beverages and snacks, they do not offer any bells and whistles. Their employees dress informally and do not try to coddle customers. Instead, they are friendly and often quirky. Safety announcements sometimes involve singing and games are often played on-board. One of the cleverest of these seen by one of the authors was a competition to see which section of the aircraft could have the most passengers pack their snack boxes into their peanut bags. Not only was it an amusing game but it significantly decreased the trash to be collected; thus, it will have had a small impact on decreasing turnaround time through decreased trash to be removed from the plane.

It is easier to rally employees around the goal of short turnaround times than the goal of lowering costs. This is another example of the alignment of their culture, their metrics, and their processes. Southwest is also very focused on what they do and do not offer. They are very deliberate about not trying to be a full service airline servicing all cities. Other innovative processes followed by Southwest include their boarding process, where seats are not pre-assigned, and the fact that they typically fly point-to-point, rather than using a hub and spoke system as used by most airlines. All of this has paid off as in 2013 Southwest was named as the seventh most admired company in the world by Fortune Magazine.

Source: Frei and Hajim (2001) and www.southwest.com.

Managing Variability

The introduction to this chapter gave five key types of variability that must be planned for: arrival, request, capability, effort, and subjective preference variability. Frei (2006) has provided the following table discussing how such variability can be accommodated or reduced in a variety of situations.

TABLE 7–2 Strategies for Managing Customer-Introduced Variability

	Classic Accommodation	Low-Cost Accommodation	Classic Reduction	Uncompromised Reduction
Arrival	• Make sure plenty of employees are on hand	• Hire lower-cost labor • Automate tasks • Outsource customer contact • Create self-service options	• Require reservations • Provide off-peak pricing • Limit service availability	• Create complementary demand to smooth arrivals without requiring customers to change their behavior
Request	• Make sure many employees with specialized skills are on hand • Train employees to handle many kinds of requests	• Hire lower-cost specialized labor • Automate tasks • Create self-service options	• Require customers to make reservations for specific types of service • Persuade customers to compromise their requests • Limit service breadth	• Limit service breadth • Target customers on the basis of their requests
Capability	• Make sure employees are on hand who can adapt to customers' varied skill levels • Do work for customers	• Hire lower-cost labor • Create self-service options that require no special skills	• Require customers to increase their level of capability before they use the service	• Target customers on the basis of their capability
Effort	• Make sure employees are on hand who can compensate for customers' lack of effort • Do work for customers	• Hire lower-cost labor • Create self-service options with extensive automation	• Use rewards and penalties to get customers to increase their effort	• Target customers on the basis of motivation • Use a normative approach to get customers to increase their effort
Subjective Preference	• Make sure employees are on hand who can diagnose differences in expectations and adapt accordingly	• Create self-service options that permit customization	• Persuade customers to adjust their expectations to match the value proposition	• Target customers on the basis of their subjective preferences

(Frei, 2006. With permission from Harvard Business Publishing.)

Notice how many of the uncompromised reductions of variability in Table 7–2 relate to the targeting of specific types of customers. For service firms to have an effective strategy, they cannot be all things to all people, even if their employees are, in theory, capable of such flexibility. For any firm, where service firms are no exception, a well-designed strategy is key for long-term success.

Service Competition

Competition in service environments is complicated by several factors considered by Porter in his five forces model of strategy (cf., Porter, 1979; Fitzsimmons & Fitzsimmons, 2006).

- *Relatively low entry barriers.* It is usually much easier to set up a service business than a production system, which requires equipment and specialized materials. This means that it may be easy for new entrants to enter the market. Thus, the existing service business will need to find a way to grow brand loyalty. Starbucks is a service company that has had unusual success along these lines.

- *Minimal opportunities for economies of scale.* Most services are highly labor-dependent. Thus, if more capacity is needed then more staff must be hired. Therefore, there are few opportunities for economies of scale in such proportional scaling of capacity. Like the issue of relatively low entry barriers, this factor means that incumbent firms are constantly at threat from new entrants in a way that firms with economies of scale, who can underprice the competition, are not.
- *Product substitution.* Services are often highly substitutable, in that customers can easily find an alternative for meeting their needs. For example, there is often very little difference between banking providers. Customers can easily switch banks with few negative consequences (other than the hassle involved). Further, innovations can sometimes replace the need for the service. For example, the internet has meant that in many cases customers can perform the service themselves (e.g., search for travel information) rather than rely on a service provider (e.g., a travel agent).
- *Exit barriers.* Some who start a service-based business do so for the love of the job, rather than because they have a particularly innovative service offering. For example, boutique owners, bed and breakfast operators, and art gallery owners may all have had a dream of owning their own operation. This makes them less likely to exit the service marketplace even if the firm is not very successful financially.

A firm positioning itself in the services marketplace should consider and evaluate all these factors. Note that, just as for production firms, it is important for firms to deliberately choose a strategic position in the marketplace. In service firms, customers participate more and staff are usually more adaptable than manufacturing equipment. Therefore, some service firms develop a *fear of focus*, trying to avoid a specific competitive position within the marketplace; this rarely works well. By trying to appeal to everyone, they can end up appealing to no one.

Problems for Section 7.1

1. Give three examples of services you are familiar with that include all of the following: there is an intangible portion to the offering, it is time-perishable, and it involves the customer as co-producer. Explain your answers.
2. The statement was made in this section that "more leisure time and/or wealth naturally lead to increased demand for services." Explain why this is the case.
3. What would a pure services economy look like? Do you think this is practical for a small country with little land area to consider? Why or why not?
4. Choose five of the dimensions of service quality and for each of them name a service firm that you believe competes effectively on this dimension. Explain your answers.
5. Choose a service firm you are familiar with and describe how it has made the four key decisions of: offering, funding mechanism, employee management system, and customer management system.
6. Choose a service firm you are familiar with and describe how each of the five key types of variability from Table 7–2 apply in its setting (e.g., make concrete its source of arrival variability). For each of the five types, describe what sort of accommodation and/or reduction is typically applied.
7. Give an example of a service firm, outside the coffee shop industry, that has managed to compete in an industry with low entry barriers using brand loyalty.

7.2 FLOW SYSTEMS

The first step in analyzing a process within a service system is to consider the aggregate flow of customers or their orders. This is often done using a process flow diagram to identify the flow and then calculating the capacity of the system to meet the demand for the service. System utilization, which at a broad level is demand divided by capacity, provides a measure of system efficiency and it is useful in both services and good-producing systems. This section outlines these concepts further.

Process Flow Diagrams

There is a variety of types of process flow diagrams from the very high level to the very detailed. Lean production systems apply a process known as *value stream mapping*; a similar process can be used for service systems. Section 11.2 describes patterns of flow within facilities. Readers interested in learning more about such tools are referred to Rother and Shook (2003).

Figure 7–1 depicts a simple flow diagram for a doctor's office. Patients check-in with the receptionist and then wait to be called by a nurse. At the nurse's station, they are weighed and various vital statistics are taken before they are shown to an examination room. The doctor then joins them in the examination room. After treatment, some proportion of patients must see the receptionist to schedule a follow-up visit. Finally, the patient leaves the office.

A more sophisticated flow diagram will also show the *resources* used for a customer to complete service. Resources may correspond to staff who perform the service or equipment that is used to process the customer's order. For example, in a doctor's office resources include the receptionist, the nurses, the doctor (who may treat multiple patients in different examination rooms), and the equipment used. In a restaurant, resources include tables, wait staff, the maître d', chefs, ovens, and other cooking equipment.

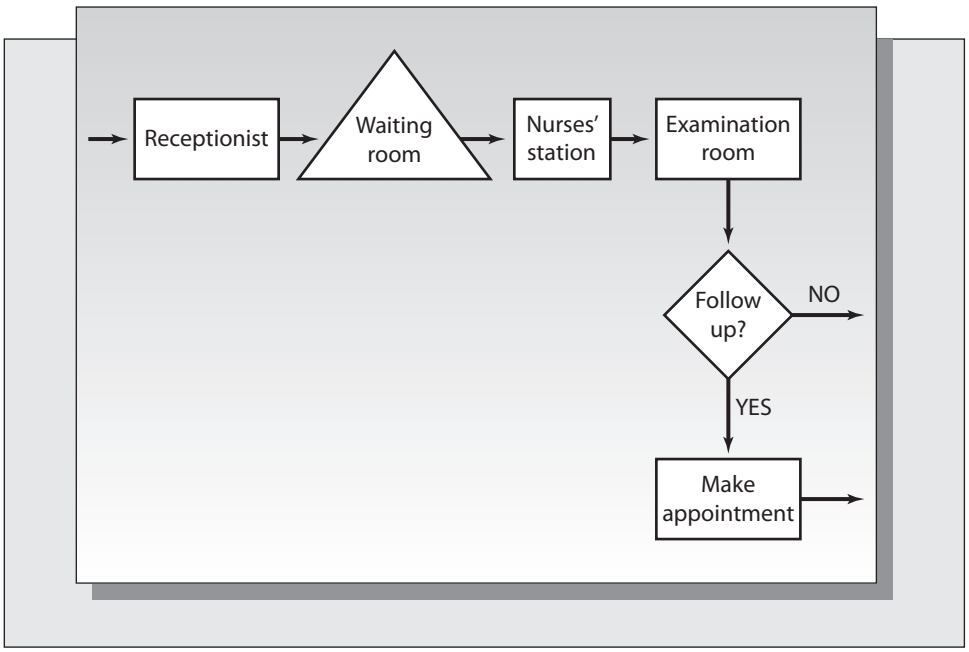


FIGURE 7–1
Patient flow at a doctor's office

Capacity

Key to determining the performance of any system is the *capacity* of the relevant resources. This is defined as the number of customers per hour (or per any relevant time unit) that may be processed by the resource. If the resource is only available some fraction of the time, this should be taken into account when calculating capacity. Similarly, if the resource, or set of resources, can process customers in parallel (i.e., in batches) then the capacity calculation should also consider this. Capacity is therefore defined as:

$$\text{Number of parallel resources} \times \frac{\text{Units per batch}}{\text{Time per batch}} \times \text{Fraction of time avavailable.}$$

In many cases, both the batch size and the fraction of time available are equal to one. For example, if a single server takes 5 minutes to process a customer then its capacity is 0.2 customers per minute, or 12 customers per hour. For two servers, this is then doubled.

Example *7.1*

Suppose a theme park roller coaster has a train with six carriages. Each carriage takes eight customers. Trains arrive to the loading point one every 30 seconds. However, each hour, on the hour, the roller coaster is shut for 3 minutes for a systems' check. What is the capacity of the roller coaster?

Solution

We will calculate capacity as customers per hour. Therefore, the time per batch is calculated as $30/(60 \times 60)$ hours, rather than just 30 seconds. Total capacity is as follows:

$$6 \times \frac{8}{30/(60 \times 60)} \times \frac{57}{60} = \frac{6 \times 8 \times 60 \times 57}{30} = 5472 \text{ customers per hour.}$$

The system *bottleneck* is defined as the resource group with the smallest capacity (assuming that all customers go through all resources). If some group of required resources can only process 10 customers per hour then there is no way to get more than 10 customers per hour through the system as a whole. Therefore, the *system capacity* is defined as the capacity of the bottleneck resource group.

Calculations are more complex if customers follow different paths through the system or have different service requirements. Also, if there are multiple paths that must be completed in parallel for the customer's service to be complete (e.g., if a patient must receive the results from both x-rays and blood tests before moving on to treatment), all paths must be considered separately. Finally, the customer mix will determine the precise resource capacity when different types of customers have differing processing requirements. Similar principles to those discussed here may be used for such systems, although the calculations are more complicated. Readers interested in more detail are referred to Anupindi et al. (2011).

Flow Rates and Utilization

In most flow analysis, we assume that the *arrival rate* of customers to the system is less than the system capacity. Let λ denote the number of customers per unit time that arrive to the system. While it is possible to do transient (that is, short-term) calculations with an arrival rate greater than the system capacity, one usually assumes *steady state*. In steady state, the system is assumed to have been running long enough, under a similar set of conditions, so that the effect of the starting state disappears. Further, the conditions under which the system is running are assumed to be relatively constant with no large effect from seasonality.

A common cause of confusion is the calculation of the arrival rate to a set of resources within the process (e.g., the arrival rate to the examination room in Figure 7–1). Unless

customers are leaving without service or being created somehow (e.g., a maternity ward), the arrival rate to any resource group in the process must be the same as the arrival rate to the start of the process, which must also equal (on average) the rate at which customers depart the system. Customers do not depart at the processing rate of the final server because, so long as arrivals do not exceed its capacity (the usual assumption), the server will be idle for portions of time when there are no customers to process.

If there are s resources that can process μ customers per unit time then the resources have capacity $s\mu$ and the *utilization* of this set of resources is given by

$$\rho = \frac{\lambda}{s\mu}.$$

Notice that utilization ρ is always nonnegative and is less than one so long as the arrival rate is less than the capacity. It is a dimensionless measure that represents the fraction of time the set of resources is busy. It can also be calculated as

$$\frac{\text{Amount of work that needs to be done}}{\text{Time available to work}}.$$

Example *7.2* Consider the roller coaster in Example 7.1 and suppose that customers arrive to the ride at a rate of 90 customers per minute and all customers wait for service. What is the roller coaster's average utilization?

Solution Using the capacity calculated in Example 7.1, the service rate $s\mu = 5472$ customers per hour. We must convert the arrival rate, λ, to the same time units, namely $90 \times 60 = 5400$ customers per hour. Utilization is therefore

$$\rho = \frac{\lambda}{s\mu} = \frac{5400}{5472} = 0.9868.$$

Thus, each seat in the roller coaster will be full 98.68 percent of the time and empty 1.32 percent of the time.

Problems for Section 7.2

8. Consider the flow in Figure 7–1 but assume that no patients require follow-up appointments. All patients first check in with the single receptionist, which takes an average of 6 minutes. They are then seen by one of two nurses who take weight, blood pressure, etc., which takes an average of 10 minutes. Finally, they are seen by a GP in an examination room, which takes an average of 16 minutes. There are three GPs. The table below summarizes this data. Assume the clinic has no waiting room or examination room constraints.

   Assume that there are eight patients arriving an hour during working hours.

Resource	Minutes per patient	Number of resources
Receptionist	6	1
Nurse	10	2
Doctor	16	3

   a. What is the capacity of each resource?

   b. Which resource is the bottleneck and hence what is the capacity of the system?

c. Suppose they wish to increase system capacity, what actions do you recommend?

d. What are the utilizations of the receptionist, nurses, and GPs, respectively (during working hours)?

e. Now suppose that 20 percent of patients require follow-up appointments, which take 2 minutes to book on average. This means that, on average, the receptionist spends $6 + 0.2 \times 2 = 6.4$ minutes per patient. What is the new capacity and utilization of the receptionist?

9. Does increasing the capacity of a resource ever increase the arrival rate to the resource group that follows it in the process flow? Explain your answer.
10. What does a utilization of greater than one mean in practice?

7.3 MODELING UNSCHEDULED ARRIVALS

One of the key features of a service system is that unscheduled arrivals must be buffered using waiting time, rather than inventory. Who has not experienced hours of waiting for service in their life? We line up to wait our turn at banks, supermarkets, hair stylists, and restaurants. Estimates of the average time a person spends in line in their lifetime range from six months to five years (Chicago Tribune, 1988; The Telegraph, 2013). Regardless, it is a lot of time waiting in line, or *queueing*, as it is more formally known. As we will see in Section 7.5, queueing is exacerbated by variability. This section examines arrival process variability.

Figure 7–2 shows the arrivals of calls to a call center. The variation seen in the first three boxes from month to month, week to week, and hour to hour is relatively predictable seasonal variation. Usually, staffing can account for seasonal variation of a predictable nature. However, the fourth box shows an extraordinary amount of variation in the number of calls from minute to minute. Such variability is difficult to predict, leading to difficulties in determining suitable staffing levels (even if it could be predicted). It causes delays in call centers and queueing in other service systems. A good model for arrivals of this type is the Poisson process, which is covered next.

Poisson Process

Let $N(t)$ be the number of arrivals to some service facility between time zero and t. Then $\{N(t): 0 \leq t < \infty\}$ is called a *stochastic process* because it is random (i.e., stochastic) and it evolves over time (so can be called a process). Arrivals to service systems can follow many different types of stochastic processes but a common model to use for unscheduled arrivals is the *Poisson process*. The Poisson process is an arrival process that satisfies the following three key assumptions:

1. The number of arrivals in disjoint intervals are independent.
2. The number of arrivals in an interval depends only on the interval's length.
3. For a very short interval (of duration h):

   the probability of one arrival is approximately λh; and
   the probability of more than one arrival is negligible.

A more formal statement of these three assumptions is made in Supplement 2.1. Together, they imply that for any $s, t \geq 0$, the distribution of the number of

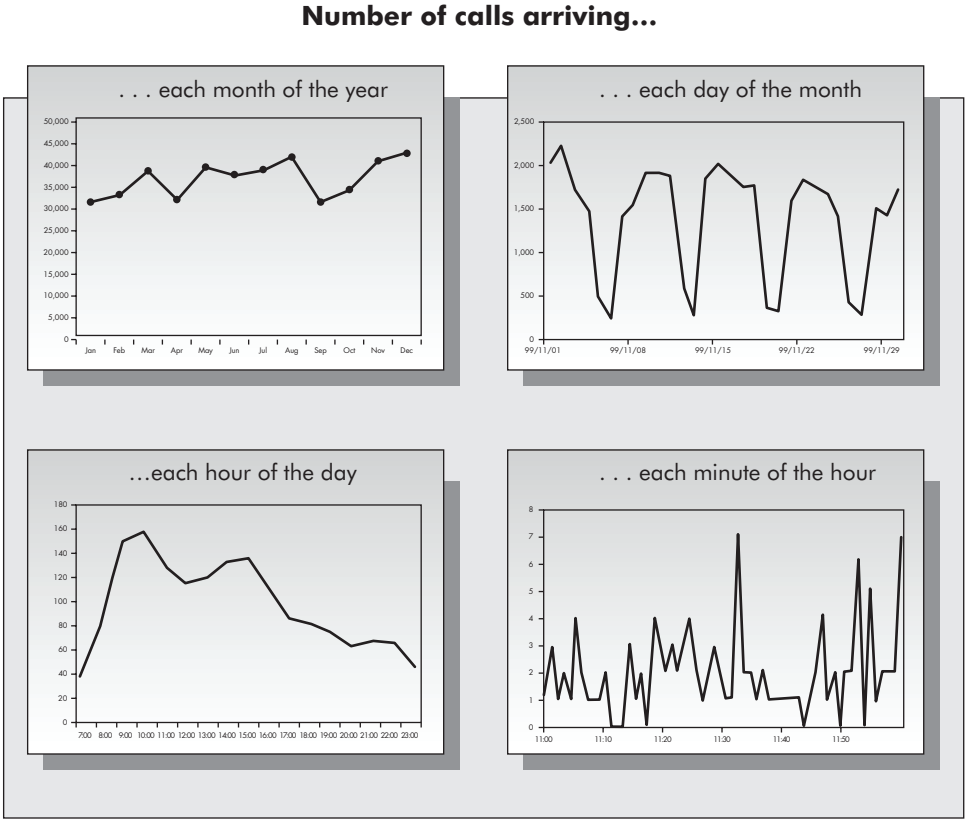


FIGURE 7–2
Unscheduled arrivals to a call center.

(From Gans et al., 2003.)

arrivals in an interval [s, s + t] (i.e., $N(t + s) - N(s)$) is *Poisson* with mean λt, which means that

$$P\{N(t) = n\} = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \text{ for } n = 0,1,2,...$$

The proof that the three assumptions above result in the Poisson distribution for the number of arrivals in an interval is given in Supplement 2.1. However, some comments are in order on their practicality. The first assumption implies that if the service provider observes an unusually high or low number of arrivals between 9am and 10am, this does not affect the distribution of the number of customers likely to arrive between 10am and 11am. This assumption is reasonable in a system where customer arrivals are not driven by some underlying force but instead are the result of random individual customer behavior.

The second assumption implies that the number of arrivals between 9am and 10am should have the same distribution as the number between 10am and 11am because both intervals are an hour long. This assumption can be problematic for systems with seasonality (see the discussion on steady state in Section 7.2), but is avoided by studying the system over a short enough time period (e.g., just during lunchtime) so that the arrival rate to the system is reasonably constant in the period of study. If arrival rates are changing slowly enough, the assumption of steady state may still form a reasonable approximation for an accurate analysis of the given time interval (e.g., Green & Kolesar, 1991).

The final assumption implies that a very short time interval should have at most one arrival. Thus, if the system experiences batch arrivals then each batch must be counted as one arrival, otherwise the probability of more than one arrival in a short interval will not be negligible.

Other than the assumptions above, there are two further models of arrivals that result in a Poisson process. First, Lariviere and Van Mieghem (2004) show that Poisson arrivals can also occur as the result of strategic customers trying to avoid congestion if the population is large and the time horizon is long. Second, as discussed next, if the time between arrivals has an exponential distribution then the number of arrivals will follow a Poisson process.

Exponential Interarrival Times

Let $N(t)$ be a Poisson process with rate λ, and let $T_1, T_2, \ldots$ be successive interarrival times; that is, the first customer arrives at time T_1, the second at time $T_1 + T_2$, the third at $T_1 + T_2 + T_3$, and so on. If arrivals follow a Poisson process then the interarrival times $T_1, T_2, \ldots$ have an exponential distribution. That is, if X is a random variable representing the time between successive arrivals, then

$$P\{X > t\} = \exp(-\lambda t).$$

Note that $E[X] = 1/\lambda$ and $Var[X] = 1/\lambda^2$ so that the coefficient of variation (see Section 6.7) $CV[X] = 1$.

Not only does a Poisson process result in exponential interarrival times, but the relationship also goes the other direction. That is, if the time between any two consecutive arrivals is exponential with mean $1/\lambda$ (and independent of all other interarrival times), then the number of arrivals in any interval of length t must be Poisson with mean λt. This equivalence is proven in Chapter 13 (see also Figure 13–5 for further understanding).

In Chapter 13 (on reliability modeling) we discuss the *memoryless property* of the exponential distribution and its relationship to the Poisson process. Both the exponential and the Poisson distribution play a key role in queueing theory, just as they do in reliability theory. When we talk about purely random arrivals in queueing, we mean that the arrival process is a Poisson process. A purely random service process means that service times have the exponential distribution. We use the term "purely random" because of the memoryless property of the exponential distribution.

The memoryless property states that no matter how much time has passed, the distribution of the time until the next arrival is the same as the distribution of the time of first arrival, T_1. (This idea is formalized in Supplement 2.1.) Such a property may make little intuitive sense until one considers it in the light of coin flips. Even if one has thrown 20 heads in a row, the probability of a head on the next throw (from a fair coin, of course) is still 50/50. Alternatively, think of the property with regard to forgetfulness around the home. Even though you have not forgotten to turn a light off recently, this has little effect on the likelihood that you will be careless and forget to turn one off today (or at least, such is the authors' experience, where lights being left on by family members seem to follow a highly random process that appears Poisson to an observer).

As will be discussed further in Chapter 13, because the exponential distribution has the memoryless property it is the continuous equivalent of the geometric (coin flip) distribution. This is equivalent to describing arrivals as "random" in that knowing something about one period of time tells nothing about the following time interval (refer also to Assumption 1 above). As a final example, suppose that cabs arrive at a

cabstand according to a Poisson process. You arrive at the stand at some random time and wait for the next cab. Your waiting time is exponential, with exactly the same distribution as the time between two successive arrivals of cabs. However, time has already passed, so if you add the time that has passed to your expected wait it will be longer than the original exponential distribution. This is what is termed the *inspection paradox* and is described further in Supplement 2.1.

Example *7.3*

Suppose customers arrive to a 7–11 convenience store according to a Poisson process with rate 10 per hour. What are the probabilities of (a) no customers in an hour; (b) exactly 5 customers in an hour; (c) exactly 10 customers in two hours; and (d) at least two customers in half an hour?

Solution

Here $\lambda = 10$ per hour and we must compute:

$$\text{(a) } P\{N(1) = 0\} = e^{-10 \times 1} = 0.00005;$$

$$\text{(b) } P\{N(1) = 5\} = \frac{(10)^5 e^{-10 \times 1}}{5!} = 0.038;$$

$$\text{(c) } P\{N(2) = 10\} = \frac{(10 \times 2)^{10} e^{-10 \times 2}}{10!} = 0.0058;$$

$$\text{(d) } P\{N(0.5) \geq 2\} = 1 - P\{N(0.5) = 0\} - P\{N(0.5) = 1\} = 1 - e^{-5} - \frac{(5)^1 e^{-5}}{1!} = 0.96.$$

Notice that the probability of 10 arrivals in two hours is not twice the probability of 5 arrivals in one hour; in fact, it is less than the probability of 5 arrivals in one hour. Explaining why this is the case is left as an exercise for the reader (see Exercise 12).

General Arrival Processes

If arrivals are not purely random (Poisson) then an important metric is the arrival process variation. Let c_a^2 be the squared coefficient of variation associated with the arrival process (see also Section 6.7). That is, $c_a^2 = \lambda^2 \sigma_A^2$, where σ_A^2 is the variance of the times between arrivals and λ is the arrival rate. For a Poisson process $\sigma_A^2 = \lambda^2$ and $c_a^2 = 1$. Notice that if the average time between arrivals stays the same, but the variance increases, then c_a^2 increases. Also, c_a^2 is dimensionless, in that it is just a number that says something about how variable the system is.

Thus, $c_a^2 = 1$ when arrivals are independent, such as customers walking in to a store. However, when there is a schedule, we expect to see $c_a^2 < 1$. For example, patients arriving for doctor's appointments should be much less variable than unscheduled emergency department patients. It is possible to have $c_a^2 > 1$ when the system is highly variable, which can occur when there are batch arrivals. Examples of batch arrivals include the arrival of people from tour buses to some monument or students arriving at the local coffee shop when arrivals right before or after classes are much more common than when classes are in session.

If there is no data on the exact interarrival times, a reasonably good estimate can be calculated as follows. Let N be the number of arrivals per time period (e.g., per hour). Calculate the mean *E[N]* and variance *Var[N]*. Then $c_a^2 \approx Var[N] / E[N]$. This estimate is exact for the Poisson process because if N is Poisson then $Var[N] = E[N] = \lambda t$, where t is the time period considered. The lack of a square on the denominator of this estimate is correct and follows from the central limit theorem for renewal processes (e.g., Wolff, 1989). Although this result is an approximation, it is often easier to calculate than the exact form. This is because it relies on the mean and variance of the number of arrivals per time period, rather than the mean and variance of the time between two arrivals, which requires an accurate stopwatch.

Pooling in Services

We saw uses of pooling in Chapter 6, but it can also be effective in service systems. For example, suppose we have only one bed and unscheduled patient arrivals. Then, if we want to be 95 percent sure the bed is available for a new patient, its utilization can be at most 5 percent. However, suppose we have thousands of beds and we want to be 95 percent sure a bed is available when a new patient arrives. We should be able to keep each bed at above 99 percent utilization and still be reasonably sure at least one will be available for the new arrival. Thus, if we can pool streams of patients into one set of bed resources then we will be able to keep the bed utilizations higher than if each type of patient gets a specialized bed type.

This question is not just academic. Hospitals often wish to separate patients by type and emergency departments often separate the arrival streams of low and high priority patients. Such service lines are usually more efficient and provide higher quality of care than the more general pooled layouts. However, their lack of pooling can mean that they will have either low utilizations or poor service levels. Therefore, service environments should treat specialization with extreme caution. Even a relatively small amount of cross training or cross sharing of resources can help mitigate utilization problems caused by dedicated service lines. These ideas are related to decisions of product versus process layouts as discussed in Section 11.3.

Probability of Delay

A useful service metric is the probability that a customer is delayed. If there is only one server and arrivals are Poisson, the probability a customer is delayed is simply ρ, where ρ is the utilization rate. However, if arrivals are not Poisson then this is only an approximation. As an extreme example, suppose customers arrive in *exactly* one-minute intervals and each takes *exactly* 45 seconds to serve. Then $\rho = 0.75$ but no customer is ever delayed. The server is busy 75 percent of the time, but an arriving customer never sees the server busy because he arrived exactly 15 seconds after the previous customer left the system.

When arrivals are Poisson, the probability that a customer is delayed is the same as the probability all servers are busy at any random instant. This is because, if arrivals are truly random (Poisson), an arriving customer will see the system in its usual or typical state. If arrivals are not truly random, the probability that all servers are busy at the instant the customer arrives, so that the customer is delayed, is not exactly the same as the probability all servers are busy at a randomly chosen instant in time (although the difference is likely to be small). The so-called *PASTA*—Poisson Arrivals See Time Averages—property guarantees that customers arriving according to a Poisson process indeed see the system in its usual or "time-averaged" state (e.g., see Wolff, 1989).

A useful and simple approximation for the probability of delay, p^d, has been developed by Sakasegawa (1977). It is

$$p^d \approx \rho^{-1+\sqrt{2(s+1)}},$$

where ρ is the utilization and s is the number of servers. This approximation is used for both Poisson and non-Poisson arrivals, although is usually more accurate for Poisson arrivals. If the number of servers $s = 1$ then the right-hand-side equals ρ, which is the exact probability of delay for Poisson arrivals. Further, the right-hand-side approaches zero as s grows to infinity; implying that, with a large enough number of servers, customers will not be delayed. The formula is based on a least squares fit to the exact formula for the so-called M/M/s queue (see Section 7.4). Although the derivation of this formula is outside the scope of this text, the intuition behind it should be clear to the reader.

FIGURE 7–3
Scaling of utilization in scale

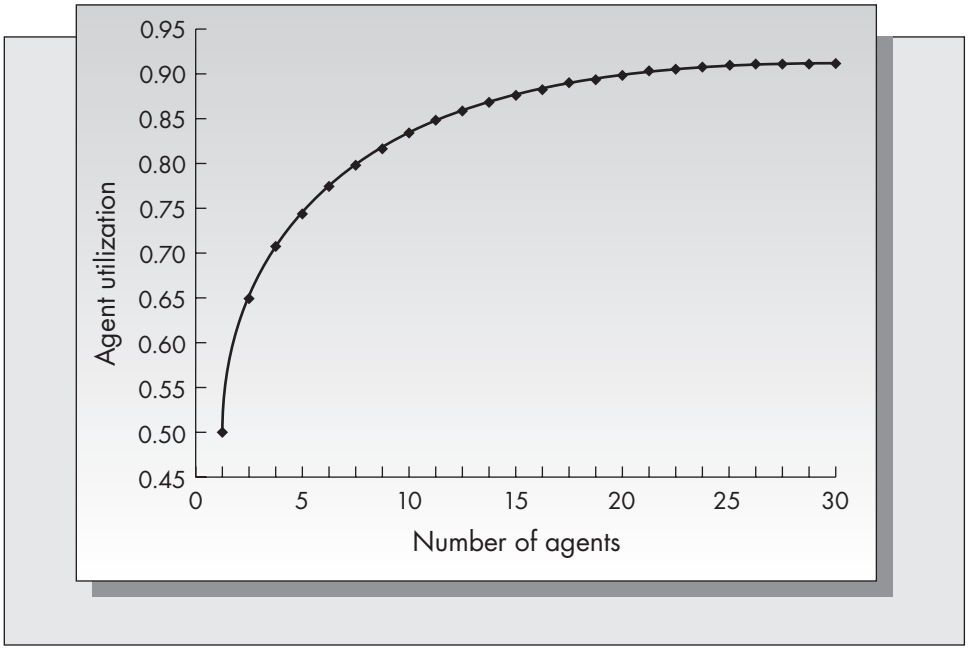


Figure 7–3 uses the Sakasegawa formula to derive the number of agents needed to achieve a 50 percent no hold rate (i.e., no delay) in a call center. If there is only one agent, the utilization is only 50 percent. However, as the scale grows so can agent utilization.

While the Sakasegawa approximation is simple and intuitive, the actual probability of delay is dependent on both the staffing decisions made by the firm and the arrival and service processes. Whitt (2004) has defined an *efficiency-driven* regime as one where p^d tends to one as both arrival rate and s grow to infinity, whereas a *quality-driven* regime has p^d tending to zero (as in the Sakasegawa formula). The so-called *quality-and-efficiency* (QED) regime is one where staffing is matched to arrival rate growth so that p^d tends to a value strictly between 0 and 1. An approximation useful for this scenario is given in Supplement 2.4.

Example *7.4*

Consider the roller coaster example of Examples 7.1 and 7.2. What is the probability of delay?

Solution

In Example 7.1 we saw that the number of server $s = 6 \times 8 = 48$; in Example 7.2 we calculated $\rho = 0.9868$. Under the Sakasegawa formula, $p^d = \rho^{-1+\sqrt{2(s+1)}} = 0.9869^{-1+\sqrt{2\times(48+1)}} = 0.889$. This probability is much less than $\rho = 0.9868$ due to the pooling effects of the servers. Therefore, an arriving customer has only an 89 percent chance of being delayed even though the seats in the ride will have an occupancy of over 98 percent.

Problems for Section 7.3

11. Suppose that arrivals to a coffee shop follow a Poisson process with average rate of one customer every 5 minutes. What are the probabilities of (a) no customers in 5 minutes; (b) exactly one customer in a minute; (c) exactly two customers in 2 minutes; and (d) at least two customers in 10 minutes?
12. Suppose that arrivals to an emergency room follow a Poisson process with mean two patients every 5 minutes. Calculate the probabilities of (a) no customers in

2 minutes; (b) exactly 2 customers in a minute; (c) exactly 5 customers in 3 minutes; and (d) no more than 2 customers in 5 minutes.

13. In Example 7.3 we saw that the probability of 10 arrivals in two hours is not double the probability of 5 arrivals in one hour; in fact, it is less than the probability of 5 arrivals in one hour. Explain why this is the case.
14. The following data has been collected on the interarrival times of calls to a call center: 3.33, 4.21, 5.12, 1.24, 0.11, 10.23, 7.65, 1.23, 4.44, 2.89, 4.92, 3.87, 2.67, 3.51, and 5.90 minutes. Estimate the squared coefficient of variation of the arrival process. Is the arrival process likely to be Poisson? Why or why not?
15. The following data has been collected on the number of customers seen to arrive to a museum in a succession of 5-minute intervals: 5, 1, 3, 7, 5, 5, 6, 7, 5, 7, 4, 8, 1, 5, 2, 3, and 5. Estimate the squared coefficient of variation of the arrival process. If this data was known to come from a Poisson process, what would be your estimate of λ, the rate of customer arrivals?
16. Recreate Figure 7–3 for a call center that wants to achieve a 40 percent no hold rate. Suppose that agents only become cost effective (where the revenue they generate exceeds their cost) if they have 90 percent utilization. How large, in terms of number of agents, does the call center need to be to have agents that are generating a positive profit? If mean call time is 5 minutes, what would the arrival rate need to be for 90 percent agent utilization with this many agents?
17. What does question 16 imply about whether an online and catalog retailer, such as L.L.Bean, should have regional or centralized call centers? What other considerations might come into this decision?
18. Suppose a bank has three tellers that are each busy 80 percent of the time. Estimate the probability of delay for a randomly arriving customer.

7.4 QUEUEING SYSTEMS

Queueing theory is the science of waiting line processes. Virtually all the results in queueing theory assume that both the arrival and service processes are random. The interaction between these two processes makes queueing an interesting and challenging area. Figure 7–4 shows a typical queueing system. Customers arrive at one or more service facilities. If other customers are already waiting for service, depending on the service discipline, newly arriving customers would wait their turn for the next available server and then exit the system when service is completed.

Queueing problems are common in operations management. In the context of manufacturing, most complex systems can be thought of as networks of queues. However, queueing problems occur most frequently in service systems. Call or contact centers are a good example of complex queueing systems. Telephone calls are routed through switching systems, where they queue up until they are either switched to the

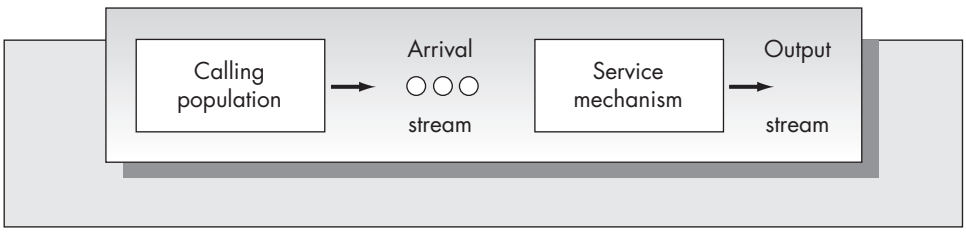


FIGURE 7–4
Typical single-server queueing system

next switching station or routed to their final destination. In fact, it was A. K. Erlang, a Danish telephone engineer, who was responsible for many of the early theoretical developments in the area of queueing.

Structural Aspects of Queueing Models

Queueing systems share a number of common structural elements as follows.

1. *Arrival process*. This is the process describing arrivals of customers to the system. Arrival processes were described in Section 7.3.
2. *Service process*. The service process is characterized by the distribution of the time required to serve a customer. The easiest case to analyze is when the distribution of service times is exponential; other more general service distributions can also provide queueing results.
3. *Service discipline*. This is the rule by which customers in the queue are served. Most queueing problems occurring in service systems are first-come, first-served (FCFS). This is the rule we usually think of as "fair." However, other service disciplines are also common. When we buy milk, we may check the dates of the bottles and buy the one with the latest expiration date. Thinking of the milk as the queue, this means that the service discipline is last-come, first-served (LCFS). Hospital emergency rooms will give priority to patients with a life-threatening condition, such as trauma from an automobile accident, over patients with less severe problems. This is referred to as a priority service discipline.
4. *Capacity of the queue*. In some cases, the size of the queue might be limited. For example, restaurants and movie theaters can accommodate only a limited number of customers. From a mathematical point of view, the simplest assumption is that the queue size is unlimited. Even where there is a finite capacity, it is reasonable to ignore the capacity constraint if the queue is unlikely to fill.
5. *Number of servers*. Queues may be either single-server or multiserver. A bank is the most common example of a multiserver queue. Customers form a single line and are served by the next available server. By contrast, the checkout area of a typical supermarket is *not* a multiserver queue. Because a shopper must commit to a specific checkout line, this is a parallel system of (possibly dependent) single-server queues. Another example of a multiserver queue is the landing area of the airport; planes may take off or land on one of several runways.
6. *Network structure*. A network of queues results when the output of one queue forms the input of another queue. Most manufacturing processes are generally some form of a queueing network. Highway systems, telephone switching systems, and medical facilities are other examples. Network queueing structures are often too complex to analyze mathematically.

Notation

A shorthand notation for single-station queueing systems, due originally to Kendall (1953), is of the form

Label 1/ Label 2/ Number,

where Label 1 is an abbreviation for the arrival process, Label 2 is an abbreviation for the service process, and Number indicates the number of servers.[1] The letter "M" is

[1] More complex notations exist that include capacity restrictions and specification of the queueing discipline. See, for example, Gross and Harris (1985), p. 9.

used to denote pure random arrivals or pure random service. This means that interarrival times are exponential (i.e., the arrival process is Poisson) or service times are exponential. The "M" stands for "Markovian," a reference to the memoryless property of the exponential distribution. The simplest queueing problem is the one labeled M/M/1. Another symbol that is commonly used is "G," for general distribution. Hence, G/G/s would correspond to a queueing problem in which the interarrival distribution is general, the service distribution is general, and there are s servers. There are other labels for other distributions but we do not consider them here. Some useful notation, some of which has already been covered, is as follows:

λ = Arrival rate to system.
μ = Service rate per server.
s = Number of servers.
ρ = Utilization rate $= \lambda/(s\mu)$.
w = Expected time a customer spends in the system in steady state.
w_q = Expected time a customer spends in the queue in steady state.
l = Expected number of customers in the system in steady state.
l_q = Expected number of customers in the queue in steady state.
W = Variable representing time in system for an arbitrary customer in steady state; $E[W] = w$.
L = Variable representing number of customers in system in steady state; $E[\mathrm{L}] = l$.
p_n = Steady-state probability of n customers in the system; $p_n = P\{L = n\}$.
p^d = Steady-state probability an arbitrary customer in steady state is delayed; $p^d = P\{W > 0\}$.

Little's law

In this section we show some useful relationships between the steady state expected values l, l_q, w, and w_q. Because w_q is the expected time in the queue only, whereas w is the expected time in the queue plus the expected time in service, it follows that w_q and w differ by the expected time in service. That is,

$$w = w_q + 1/\mu.$$

(If the mean service rate is μ, it follows that the mean service time is $1/\mu$.)

Little's law is named for John D. C. Little of the Massachusetts Institute of Technology, who proved that it holds under very general circumstances. It is a simple but very useful relationship between the l's and the w's. The basic result is

$$l = \lambda w.$$

We will not present a formal proof of this result, but provide only the following intuitive explanation. Consider a customer who joins the queue in steady state. At the instant the customer is about to complete service, he looks over his shoulder at the customers who have arrived behind him. There will be, on average, l customers in the system. The expected amount of time that has elapsed since he joined the queue is, by definition, w. Because customers arrive at a constant rate λ, it follows that during a time w, on average, there will have been λw arrivals, giving $l = \lambda w$. For example, if customers arrive at the rate of 2 per minute and each spends an average of 5 minutes in the system, then there will be 10 customers in the system on average. Another version of Little's law is

$$l_q = \lambda w_q.$$

The argument here is essentially the same, except that the customer looks over his shoulder as he enters service, rather than when completing service.

The M/M/1 Queue

The M/M/1 queue assumes Poisson arrivals, exponential service times, and a single server serving customers in a FCFS fashion. As discussed above, Poisson arrivals are a reasonably good assumption for unscheduled systems. Further, if there is a mix of many different types of jobs, the exponential distribution can be realistic for service times. Otherwise, it tends to be too variable of a distribution. However, it will often provide a reasonable upper bound, because its extra variability leads to the overestimation of most system statistics.

It is shown in Supplement 2.2 that for the M/M/1 queue the steady state probability of n customers in the system is given by

$$p_n = \rho^n(1 - \rho) \text{ for } n = 0,1,2,...$$

This distribution, known as the geometric distribution, is pictured in Figure 7–5. Several aspects to this result are both interesting and surprising. First, the geometric distribution is the discrete analog of the exponential distribution. Second, the probability of state n is a decreasing function of n, as pictured in Figure 7–5, so long as $\rho < 1$. As ρ gets close to one, the variance increases and the distribution "spreads out" (large values become more likely). As ρ gets close to zero, the probabilities associated with larger values drop off to zero more quickly. This means that the most likely state is *always* state 0 (as long as $\rho < 1$)! This is an extremely surprising result! As ρ approaches one, the queues get longer and longer. One would have thought that the probability of n in the system for some large value of n would be higher than the probability of zero in the system when ρ is near one. This turns out not to be the case. What is true is that, for ρ close to one, the probability that the system is in state zero is close to the probability that the system is in state 1 or state 2. For example, whereas for ρ close to zero the probability that the system

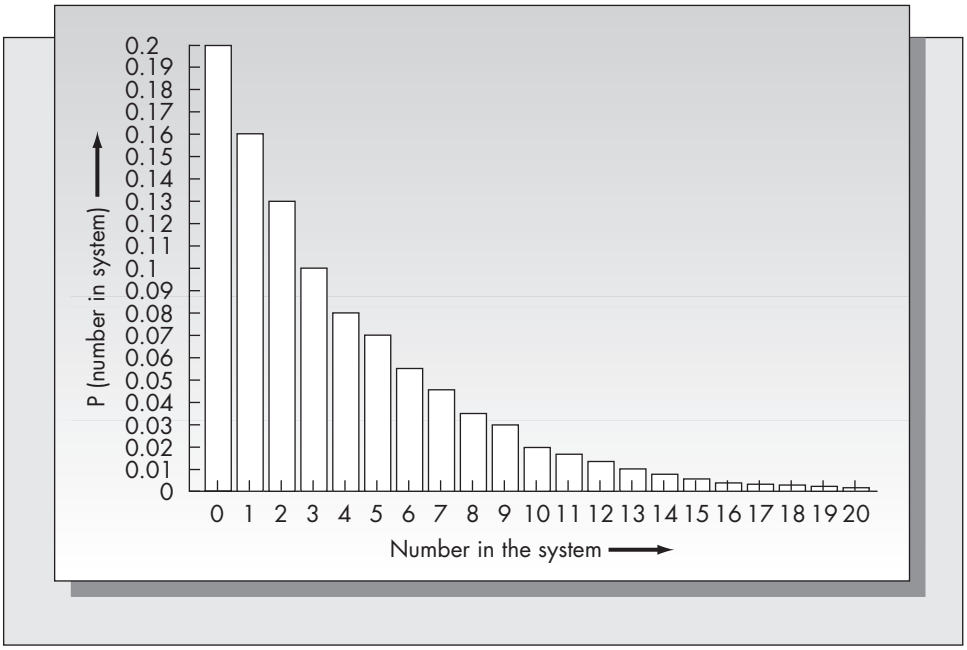


FIGURE 7–5
Geometric distribution of number in system for M/M/1 queue ($\rho = .8$)

is in state zero is much larger than the probability that it is state 1 or 2. This phenomenon holds *only* for exponential services and Poisson arrivals, however.

The distribution of p_n may be used to calculate l, l_q, w, and w_q. The expected value of a random variable is the sum of its outcomes weighted by the probabilities of those outcomes. It follows that the average, or expected, number of customers in the system in steady state, l, is:

$$l = \sum_{i=0}^{\infty} ip_i = \sum_{i=0}^{\infty} i(1 - \rho)\rho^i = (1 - \rho)\rho \sum_{i=0}^{\infty} i\rho^{i-1}.$$

To complete the calculation, we use the fact that

$$\sum_{i=0}^{\infty} i\rho^{i-1} = \frac{d}{d\rho}\left(\sum_{i=0}^{\infty} \rho^i\right) = \frac{d}{d\rho}\left(\frac{1}{1 - \rho}\right) = \frac{1}{(1 - \rho)^2}.$$

It follows that

$$l = \frac{(1 - \rho)\rho}{(1 - \rho)^2} = \frac{\rho}{(1 - \rho)}.$$

For the case of l_q, we note that the number in the queue is exactly one less than the number in the system as long as there is at least one in the system. It follows that

$$\begin{aligned} l_q &= \sum_{i=1}^{\infty} (i - 1)p_i = \sum_{i=1}^{\infty} ip_i - \sum_{i=1}^{\infty} p_i \\ &= l - (1 - p_0) = l - \rho = \rho^2/(1 - \rho). \end{aligned}$$

Given knowledge of l and l_q, we can obtain w and w_q directly from Little's law. From Little's law, $w = l/\lambda$, giving

$$w = \frac{\rho}{\lambda(1 - \rho)} = \frac{1/\mu}{(1 - \rho)} = \frac{1}{(\mu - \lambda)}.$$

Similarly, $w_q = l_q/\lambda$, which gives

$$w_q = \frac{\rho^2}{\lambda(1 - \rho)}.$$

Let W be the random time a customer spends in the system, so that $E[W] = w$. Then, for the M/M/1 queue, the distribution of W is known and remarkably turns out also to be an exponential distribution; it has mean $w = 1/(\mu - \lambda)$ and hence rate $(\mu - \lambda)$. That is,

$$P\{W \leq t\} = 1 - e^{-(\mu - \lambda)t} \text{ for all } t \geq 0.$$

The derivation of this result is given in Supplement 2.2.

Example *7.5*

Customers arrive one at a time, completely at random, at an ATM at the rate of six per hour. Customers take an average of 4 minutes to complete their transactions. However, ATM tasks are highly variable ranging from simple withdrawals to complex deposits; thus service times may be considered truly random. Customers queue up on a first-come, first-served basis and no customers leave without service. Assume there is only one ATM.

a. Find the following expected measures of performance for this system: the expected number of customers in the system, the expected number of customers waiting for service, the expected time in the system, and the expected time in the queue.

b. What is the probability that there are more than five people in the system at a random point in time?

c. What is the probability that the waiting time in the system exceeds 10 minutes?

d. Given these results, do you think that management should consider adding another ATM?

Solution The statement that customers arrive one at a time completely at random implies that the input process is a Poisson process. The arrival rate is $\lambda = 6$ per hour. The statement that service times are also truly random implies service times may be modeled by an exponential distribution. The mean service time is 4 minutes $= 1/15$ hour, so that the service rate is $\mu = 15$ per hour. Thus, this is an M/M/1 queue with utilization rate $\rho = \lambda/\mu = 6/15 = 2/5 = 0.4$.

a. $l = \rho/(1 - \rho) = (2/5)/(3/5) = 2/3$ (= 0.6667 customers)
$l_q = \rho L = (2/5)(2/3) = 4/15$ (= 0.2667 customers)
$w = l/\lambda = (2/3)/6 = 2/18 = 1/9$ hour (= 6.6667 minutes)
$w_q = l_q/\lambda = (4/15)/6 = 4/90 = 2/45$ hour (= 2.6667 minutes)

b. Here we are interested in $P\{L > 5\}$. In general,

$$P\{L > k\} = \sum_{i=k+1}^{\infty} p_i = \sum_{i=k+1}^{\infty} (1 - \rho)\rho^i = (1 - \rho) \sum_{i=k+1}^{\infty} \rho;$$

$$= (1 - \rho)\rho^{k+1} \sum_{i=0}^{\infty} \rho^i = (1 - \rho)\rho^{k+1}(1/(1 - \rho)) = \rho^{k+1}.$$

Hence, $P\{L > 5\} = \rho^6 = (0.4)^6 = 0.0041$.

c. Here we are interested in $P\{W > 1/6\}$.

$$P\{W > t\} = e^{-(\mu - \lambda)t} = e^{-(15-6)1} = e^{-1.5} = 0.223.$$

d. The answer is not obvious. Looking at the expected measures of performance, it would appear that the service provided is reasonable. The expected number of customers in the system is fewer than one and the average waiting time in the queue is less than 3 minutes. However, from part (c) we see that the proportion of customers who have to spend more than 10 minutes in the system is more than 20 percent. This means that there are probably plenty of irate customers, even though, on average, the system looks good. This illustrates a pitfall of only considering expected values when evaluating queueing service systems.

Problems for Section 7.4

19. A supermarket manager notices that there are 20 customers at the checkouts and also knows that arrivals to the checkout at that time of day are at a rate of about two per minute. About how long are customers spending in the checkout process (queueing and being served) on average?
20. Suppose that the billing cycle for a firm is 60 days and they invoice on average \$5000 per day. What is the average total dollar amount of outstanding invoices that they carry?
21. Which of the three variables in Little's law do you think is generally the most difficult to estimate (and why)?
22. A teller works at a rural bank. Customers arrive to complete their banking transactions on average one every 10 minutes; their arrivals follow a Poisson arrival process. Because of the range of possible transactions, the time taken to serve each customer may be assumed to follow an exponential distribution with a mean time of 7 minutes. Customers wait in a single queue to get their banking done and no customer leaves without service.

    a. Calculate the average utilization of the teller.
    b. Calculate how long customers spend on average to complete their transactions at the bank (time in queue plus service time). What percentage of that time is spent queueing?

c. How many customers are in the bank on average?

d. Calculate the probability a customer will spend less than 30 minutes at the bank (time in queue plus service time).

e. Calculate the probability that there are more than two customers in the bank.

f. What do a–e imply about customer service at the bank?

23. Customers arrive to a local bakery with an average time between arrivals of 5 minutes. However, there is quite a lot of variability in the customers' arrivals, as one would expect in an unscheduled system. The single bakery server requires an amount of time having the exponential distribution with mean 4.5 minutes to serve customers (in the order in which they arrive). No customers leave without service.

a. Calculate the average utilization of the bakery server.

b. Calculate how long customers spend on average to complete their transactions at the bakery (time in queue plus service time). What percentage of that time is spent queueing?

c. How many customers are in the bakery on average?

d. Calculate the probability a customer will spend more than an hour at the bakery (time in queue plus service time).

e. What is the probability that there are fewer than two customers in the bakery?

f. Why are the estimated waits in this system so long? Are the assumptions behind them reasonable? Why or why not?

7.5 GENERAL QUEUEING MODELS

This section considers results for G/G/s queues and other general queueing models. We also discuss simulation as a tool if the model is too complex for queueing analysis. The supplement on Queueing Techniques contains further queueing theory results and covers some of the more technical results that are too detailed for this chapter.

In the M/M/1 model of the previous section, the distribution of the service time is exponential. In many cases, this assumption is unwarranted. One would expect that service times would rarely be exponential because the exponential distribution has the memoryless property: the amount of time remaining in service would have to be independent of the time already spent. One would think that a modal distribution, such as the normal or Erlang, would be a more accurate model of service times in most circumstances. For that reason, models with general service times are of great interest.

Define c_s^2 to be the *squared coefficient of variation* associated with the service process. As for the arrival process (see Section 7.3), we can compute it with two alternate formulas. First, $c_s^2 = \mu^2\sigma_s^2$ where σ_s^2 is the variance of the service times. Second, $c_s^2 \approx \text{Var[S]}/\text{E[S]}$, where S is the number of services per unit time (e.g., per hour), not including idle time. This second value is less convenient to calculate than for arrivals because of the need to exclude idle time. If we observe that a server has served a certain number of customers in an hour, we must ensure that the entire hour was used for processing for the calculation to be correct. Then, $c_s^2 < 1$ when there is some uniformity in customer service times, $c_s^2 \approx 1$ when the service tasks are very customer specific (with $c_s^2 = 1$ for exponential service distributions), and $c_s^2 > 1$ when the customers have unusually variable service requirements.

Expected Time in System for a Single Server System

For the G/G/1 system, an approximation for the expected time in system is the following.

$$w \approx \frac{1}{\mu}\left(\frac{c_a^2 + c_s^2}{2}\right)\frac{\rho}{1-\rho} + \frac{1}{\mu}.$$

The expected time in queue, w_q, is the first term of this equation. (Recall that $w = w_q + 1/\mu$.) This result is exact, and known as the *Pollaczek-Khintchine* (P-K) formula, if arrivals follow a Poisson process (i.e., when $c_a^2 = 1$). If service times are also exponentially distributed then the formula is the same as the one for the M/M/1 queue. If arrivals are not Poisson then it is an approximation; it becomes proportionately more accurate as utilization, ρ, tends to 1.

This formula implies that, even if a server is not fully utilized and even if a server is after (or before) the system bottleneck (as defined in Section 7.2), significant waiting (queueing) can occur. This implies that, in service systems, where the customer's experience is influenced by waiting time, a bottleneck analysis is likely not sufficient for good system design. The formula shows that waits increase linearly with variability and exponentially with utilization. Figure 7–6 shows the growth of expected time in system for a range of variabilities as utilization grows.

The impact of not understanding the relationships shown in Figure 7–6 has huge implications with respect to staffing decisions. Many managers consider utilizations of less than 100 percent inefficient, yet such utilizations are only feasible in a system with no variability (which almost never occurs in the real world). Many hospital emergency departments target utilizations of 80 percent, because they understand that high utilizations result in long waits. However, as can be seen from the formula, it is not really possible to have a one-size-fits-all target for utilization (because it should also depend on variability).

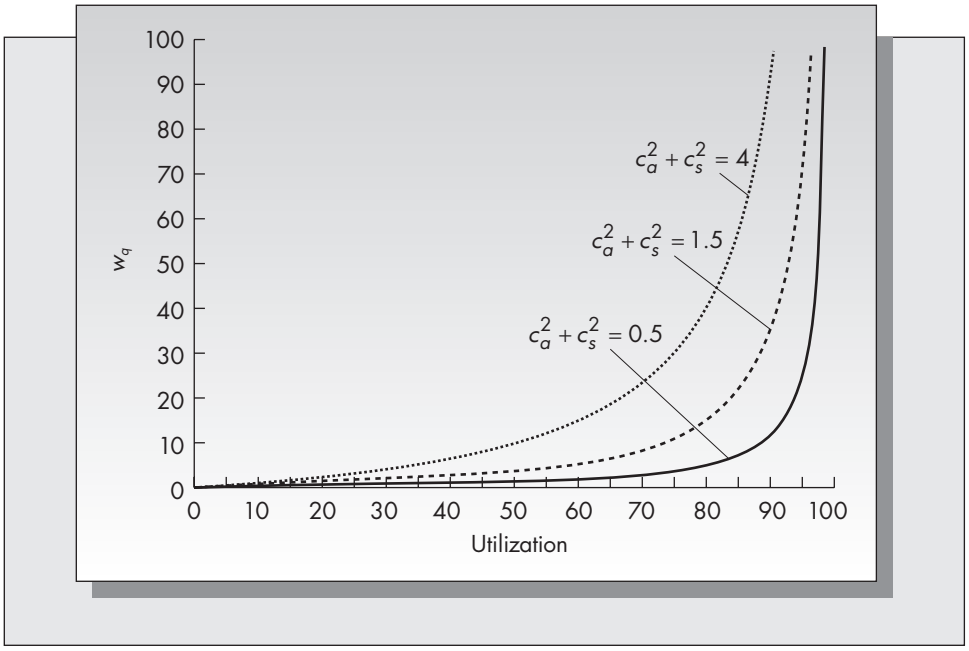


FIGURE 7–6
Growth in time in queue for a G/G/1 queue.

Example 7.6

A large discount warehouse is assessing the number of check stands it needs. During a period of the day when the arrival rate of customers is about one every 12 minutes, there is only one check stand open. It takes an average of 8 minutes to check out one customer. The checkout time follows a normal distribution with standard deviation 1.3 minutes. The arrival process may be assumed to be a Poisson process. Find l, l_q, w, and w_q for this system. How far off would your calculations be if you assumed that the service distribution was exponential?

Solution

The arrival rate is $\lambda = 5$ per hour, and the service rate is $\mu = 60/8 = 7.5$ per hour, giving $\rho = 5/7.5 = 2/3$. Notice how we put λ and μ into consistent units to compute ρ. We also must use consistent units when computing the squared coefficient of variation associated with the service process, using $c_s^2 = \mu^2\sigma_s^2$. The standard deviation of the service time is 1.3 minutes, or 1.3/60 = 0.02167 hour. It follows that the variance of the service time is $0.02167^2 = 4.6944 \times 10^{-4}$ hours^2. Therefore $c_s^2 = 4.6944 \times 10^{-4} \times (7.5)^2 = 0.0264$. (If we had computed both μ and σ_s in minutes then we would have got the same answer for c_s^2, which is unitless.) Since arrivals are Poisson, $c_a^2 = 1$ and the results will be exact, not approximate, it therefore follows that

$$w_q = \frac{1}{\mu}\left(\frac{c_a^2 + c_s^2}{2}\right)\frac{\rho}{1-\rho} = \frac{1}{7.5}\left(\frac{1 + 0.0264}{2}\right)\frac{2/3}{1 - 2/3} = 0.13686 \text{ hour} = 8.21 \text{ minutes}.$$

Therefore,

$$w = w_q + \frac{1}{\mu} = 8.21 + 8 \text{ minutes} = 16.21 \text{ minutes}.$$

Then, using Little's law (with the units expressed in hours),

$$l_q = \lambda w_q = 5 \times 0.1368 = 0.6843 \text{ customers}.$$
$$l = \lambda w = 5 \times 0.2702 = 1.351 \text{ customers}.$$

Hence, each customer should expect to wait in line about 8 minutes and expect to be in the system about twice that time. On average, there is less than one customer in the queue.

Now, suppose that we had assumed that the service distribution was exponential. We can either use the same method substituting $c_s^2 = 1$, or more simply, use the performance measures for an M/M/1 queue as follows.

$$l = \rho/(1-\rho) = (2/3)/(1/3) = 2 \text{ customers}.$$
$$l_q = l - \rho = 2 - 2/3 = 4/3 \text{ customers}.$$
$$w = l/\lambda = 2/5 = 0.4 \text{ hours (24 minutes)}.$$
$$w_q = l_q/\lambda = 1.3333/5 = 0.2667 \text{ hours (16 minutes)}.$$

We see that assuming that the service distribution is exponential results in substantial errors in the system performance measures. In particular, w_q is too high by 100 percent! The M/M/1 model is very sensitive to the assumption that the service time is exponential and typically overestimates the performance measures for most real systems (i.e., it is rare for a system to be more variable than an M/M/1 system, although it is possible). If the results of a queueing analysis are to be accurate, it is vitally important that any assumptions regarding the form of the service or the arrival process be verified by direct observation of the system.

One interesting special case of the M/G/1 queue is when the service distribution is deterministic (labeled M/D/1). In this case $\sigma_s^2 = c_s^2 = 0$.

As discussed in the next subsection, when the number of servers exceeds one, exact algebraic expressions are not known in general. The interested reader should refer to a more complete coverage of queueing, such as that given by Gross and Harris (1985) or Kleinrock (1975). Explicit results for many versions of the M/M queueing model

are available, however. Examples include systems with priority services, jockeying (switching between queues), and impatient customers, just to mention a few.

Multiple Parallel Servers

Consider a queue with s servers in parallel as shown in Figure 7–7. When customers arrive they queue up in a single line. The next customer in the line is served by the next available server. This is the G/G/s queue.

For the G/G/s system, an approximation for expected time in system is as follows.

$$w \approx \frac{1}{s\mu}\left(\frac{c_a^2 + c_s^2}{2}\right)\frac{p^d}{1 - \rho} + \frac{1}{\mu},$$

where p^d is the probability of delay and may be approximated using the formula in Section 7.3 or the one in Supplement 2.4. As in the G/G/1 system, this approximation for w becomes proportionately more accurate as utilization, ρ, tends to 1. However, it is no longer exact even for Poisson arrivals. Exact results do exist for M/M/s queues and may be found in Supplement 2.3.

Some comments are in order on this formula. First, w scales linearly in $1/\mu$. If we double the service time and scale arrivals to match, the waiting time will double. If we transform waits from hours to minutes by multiplying $1/\mu$ by sixty, w will also become sixty times larger and be transformed from hours to minutes. Second, the delay in queue scales down by the number of servers, s, while the final $1/\mu$ does not. This is because no matter how many servers there are, the customer will always require an average of $1/\mu$ time to be served; however, the pooling of the servers cuts down the queueing delay. Third, as for the G/G/1 queue, although variability does not affect average utilization it does have a large impact on delays. The key measure of variability is $c_a^2 + c_s^2$, so both arrival and service time variability contribute equally to delays. Finally, waiting time is dependent on both utilization rates and

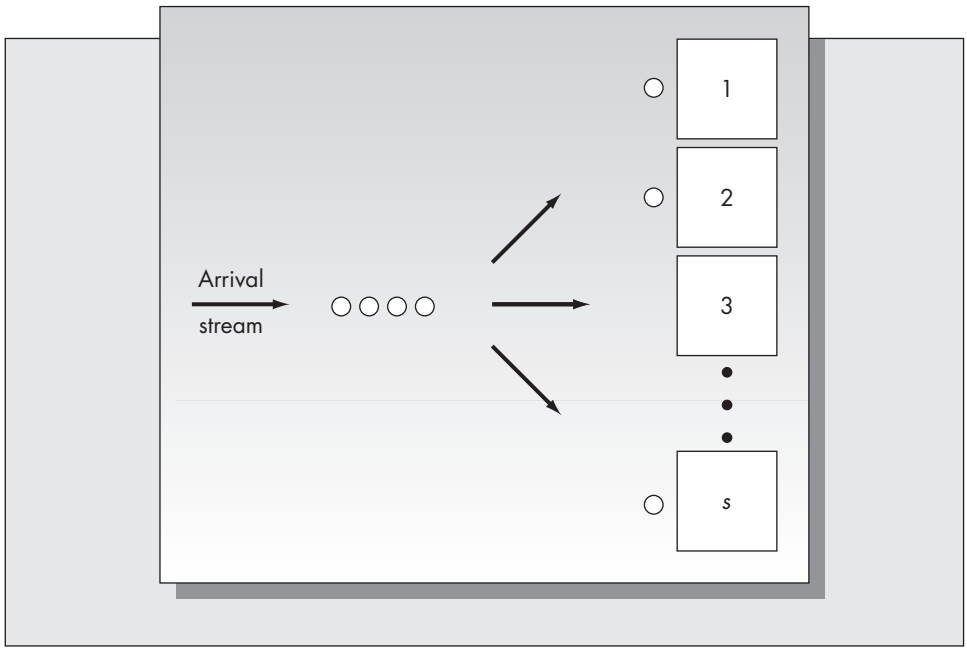


FIGURE 7–7
s servers in parallel

variability and when utilization gets close to 1, waits (and hence lines by Little's law) get long.

This formula provides insight into the circumstances that would give rise to exceptionally long queues. First, anywhere capacity is expensive (e.g., theme parks) or the service is measured primarily in terms of cost (e.g., government departments), we expect to see high utilizations and hence long waits. Second, any system with high variability either in arrivals (e.g., tourist attractions receiving tour buses) or service times (e.g., emergency departments) will also experience long waits. A final caveat is that this formula assumes that no one leaves because the line is too long. Systems with customers departing without service are described in the next subsection.

Example *7.7*

Suppose the Department of Motor Vehicles (DMV) employs three servers that serve customers in one (virtual) line. Assume customers never leave once they have taken a number, which holds their position in the queue. Customers arrive one at a time completely at random at a rate of one every 3 minutes. Service times are also quite variable with a mean of 8.5 minutes and a standard deviation of 7.5 minutes. Approximate w, w_q, l, and l_q for this system.

Solution

The above description implies that arrivals may be modeled by a Poisson process with rate $\lambda = 1/3$ per minute or 20 per hour. Service rate $\mu = 60/8.5 = 7.05882$ per hour. The standard deviation of 7.5 minutes implies that $c_s^2 = (7.5)^2/(8.5)^2 = 0.7785$. There are 3 servers. Thus, $s = 3$ and $\rho = 20/3(3 \times 7.05882) = 0.94444$. We approximate the probability of delay as

$$p^d \approx \rho^{-1+\sqrt{2(s+1)}} = (0.94444)^{-1+\sqrt{2\times(3+1)}} = 0.9008.$$

Notice how this is significantly smaller than ρ due to the pooling effects of three servers. We can then calculate

$$w_q \approx \frac{1}{s\mu}\left(\frac{c_a^2 + c_s^2}{2}\right)\frac{p^d}{1-\rho} = \frac{8.5}{3}\left(\frac{1 + 0.7785}{2}\right)\frac{0.9008}{1 - 0.9444} = 40.85 \text{ minutes} = 0.681 \text{ hour.}$$

Therefore,

$$w = w_q + \frac{1}{\mu} \approx 40.85 + 8.5 \text{ minutes} = 49.35 \text{ minutes.}$$

Then, using Little's law (with the units expressed in minutes),

$$l_q = \lambda w_q \approx 1/3 \times 40.85 = 13.62 \text{ customers.}$$
$$l = \lambda w \approx 1/3 \times 49.35 = 16.45 \text{ customers.}$$

Hence, each customer should expect to spend almost 50 minutes in the system, and on average there are more than 13 customers in the queue. Even though there are three servers and each will be idle more than 5 percent of the time, there is significant waiting in this system. Given the cost emphasis placed on government departments, utilizations of 94 percent are not unreasonable. Therefore, spending close to an hour in the DMV (and if the average time is 50 minutes then many customers actually spend more than an hour) should not be a surprise. As discussed in the next subsection, waits are also exacerbated by the fact that in this system we assume no one leaves without service. Indeed, for many government services people do wait indefinitely because the service is not discretionary.

Systems with Abandonment

The queueing models above all show that the expected wait grows towards infinity as utilization approaches 100 percent. Further, the formulas assume $\rho < 1$. What happens in reality when inflow is greater than outflow ($\rho > 1$)? In most service systems,

customers will not actually wait indefinitely but will leave without the service. This is known as *abandonment* or *customer impatience*. If the customer leaves as soon as he observes the queue length without joining, this is termed *balking*; whereas, if he leaves after having joined the queue and has waited some (usually random) length of time, this is termed *reneging*. Either way, this represents lost revenue to the firm and lost customer goodwill.

In most systems abandonment is not observed so a manager may make the assumption that if $\rho > 1$ then abandonment is not occurring. This is a very poor assumption for two reasons. First, abandonment can occur in any system even when queues are short. Second, even when utilization is less than one, long queues can still form, leading to significant abandonment. Service firms can find themselves in a downward cycle of service if managers place too much importance on high server utilization and do not recognize the need for buffer capacity to make sure the lines do not get too long. In such a scenario, the manager may cut staffing because seemingly there is slack in the system. This results in longer waits and more abandonment, resulting in more slack, and so the downward cycle continues.

There are some limited analytic models of abandonment (e.g., see Wang et al., 2010). Most models typically assume that customers have a "patience" distribution and when their patience tolerance has been exceeded they leave, if they have not yet been placed in service. M/M/1 queueing models with exponential abandonment are particularly tractable. In practice, customer behavior is actually more complex than a simple distribution and may depend on factors like the length of the queue (e.g., Bolandifar et al., 2013). Section 7.6 discusses the psychology of queueing.

There has also been work on systems with balking. Here, the customer will join the system if the value of service minus the combined cost of service and waiting is non-negative. The earliest work in this area is due to Naor (1969), who showed that, at equilibrium, customers join the queue only if the length of the queue is below a threshold. Hassin and Haviv (2003) extensively review game theoretic results of this sort in their book *To Queue or Not to Queue*.

Priorities

While first-come, first-served is considered fair, it is not always the order of service. Some customers may simply be higher priority than others (e.g., severe trauma cases in the emergency department versus minor ailments). Furthermore, it may not be efficient to treat all customers equally if they have different service requirements. This is illustrated in the following example.

Example *7.8*

Suppose we have a system with two types of Poisson arrival streams. Type 1 customers arrive at a rate of 9 arrivals per hour and each requires exactly 6 minutes of work. Type 2 customers arrive at a rate of 1.5 arrivals per hour and each requires exactly 36 minutes of work. What is the expected waiting time in the system if each type has a line to itself? Then, what is the expected time in system if there is only a single FCFS line with two servers? Finally, suppose the system gives priority to Type 1 arrivals but still shares the two servers. It can be shown (e.g., Kleinrock, 1976) that the wait in system for type 1, w_1, equals 18.0 minutes and the time in system for type 2, w_2, equals 117.8 minutes. How do these waits compare to the ones previously computed without priorities?

Solution

For type 1 customers, the arrival rate is $\lambda = 9$ per hour, and the service rate is $\mu = 60/6 = 10$ per hour, giving $\rho = 9/10 = 0.9$. Poisson arrivals and exact service times imply that $c_a^2 = 1$ and $c_s^2 = 0$ respectively. Using the formula for an M/G/1 queue, the wait in system for

type 1, w_1, equals $\frac{1}{\mu}\left(\frac{c_a^2 + c_s^2}{2}\right)\frac{\rho}{1-\rho} + \frac{1}{\mu} = \frac{60}{10}\left(\frac{1+0}{2}\right)\frac{0.9}{1-0.9} + 6 = 33$ minutes. For type 2 customers, the arrival rate is $\lambda = 1.5$ per hour and the service rate is $\mu = 60/36 = 1.667$ per hour, giving $\rho = 1.5/1.667 = 0.9$. Again, $c_a^2 = 1$ and $c_s^2 = 0$. Then, w_2 equals $\frac{1}{\mu}\left(\frac{c_a^2 + c_s^2}{2}\right)\frac{\rho}{1-\rho} + \frac{1}{\mu} = 36\left(\frac{1+0}{2}\right)\frac{0.9}{1-0.9} + 36 = 198$ minutes. Finally, the average customer wait, w, is computed as $9/10.5 \times 33 + 1.5/10.5 \times 198 = 56.6$ minutes, where $9/10.5$ is the fraction of type 1 customers and $1.5/10.5$ is the fraction of type 2 customers.

If the customers are pooled into one queue, we must first calculate c_s^2, because the presence of two types of customers implies that service is no longer deterministic. In particular, the expected service time is $E[S] = 9/10.5 \times 6 + 1.5/10.5 \times 36 = 10.286$ and the second moment of service time $E[S^2] = 9/10.5 \times 6^2 + 1.5/10.5 \times 36^2 = 216$. Therefore, $c_s^2 = Var[S]/(E[S])^2 = (216 - 10.286)^2/(10.286)^2 = 1.042$. It is known that the combination of two streams of Poisson arrivals is also Poisson (e.g., Wolff, 1989), so that $c_a^2 = 1$. Then, putting this into the M/G/2 system with $p^d = \rho^{-1+\sqrt{2(s+1)}} = 0.9^{-1+\sqrt{6}} = 0.858$, we have that $w_1 = w_2 = w = \frac{1}{s\mu}\left(\frac{c_a^2 + c_s^2}{2}\right)\frac{p^d}{1-\rho} + \frac{1}{\mu} = \frac{10.286}{2}\left(\frac{1+1.042}{2}\right)\frac{0.858}{1-0.9} + 10.286 = 55.4$ minutes. Thus, total expected wait has decreased very slightly at significant expense to Type 1 customers. Pooling servers improves overall system performance but at a significant cost to some customers. However, pooling the servers *significantly* increased the variability of service time from zero to above 1. Therefore, there is only a very modest decrease in overall wait obtained by the pooling. It is even possible to construct examples where pooling increases overall wait, if the gap between the two types of service is large enough and so enough variability has been added to the system by such pooling (see Exercise 7.26).

Finally, suppose the system gives priority to Type 1 arrivals but still shares the two servers. As noted (the formulas are beyond the scope of this text) $w_1 = 18.0$ minutes and $w_2 = 117.8$ minutes. Therefore, the time in system averaged across both types, w, equals $9/10.5 \times 18 + 1.5/10.5 \times 117.8 = 32.2$ minutes. Compared to two separate lines (no pooling) both Type 1 and Type 2 are better off. Compared to the FCFS system Type 1 is significantly better off but Type 2 is not. Further, notice how much the overall average wait w has decreased by using this priority system. We explain how this is possible next.

The order of service does not affect server utilization (unless there is abandonment or some similar effect); however, it can have a large effect on waiting times. If the goal is to minimize the average wait in the system, short jobs should be served first. We all intuitively know this when we let the person with one sheet of paper at the photocopier ahead of us when we have a large job. We understand that the effect on our own wait is minimal but the effect on his is large. In general, giving priority to customers with the shortest expected service time will minimize w when it is computed across all customer classes. Chapter 9 contains more discussion on the optimal scheduling of jobs or customers.

In call centers it is easy to give priority based on customer specific information, because customers do not observe the queue. However, in other service environments, deviating from FCFS can be problematic because customers see it as unfair. Further, customers have a good reason to be upset because FCFS minimizes delay variance (across all customers) and so indeed is the most "fair" system (see Section 9.9). One proposed rule, given in Ayhan and Olsen (2000), is to serve the next customer with the largest value of $a\mu$, where a is how long the customer has waited thus far and $1/\mu$ is the customer's expected service time. This is shown to minimize the second moment of delay as utilization approaches 1. It provides a balance between efficiency (a small w) and fairness.

In short, queueing theory tells us it is better to pool customers and to give priority to customers with short service times. However, there are a variety of other considerations that may make this unattractive or impractical. Regardless, customer prioritization can have a significant impact on service system performance. Optimization of queueing systems is discussed further in Supplement 2.6.

Other Queueing Extensions

Many more variants of queueing models are possible than those that have been discussed thus far. In some cases there is a finite limit K on the number of customers that can be present in the system at any time. Supplement 2.3 presents results for the M/M/1 queue with finite capacity K. The interested reader is referred to Gross and Harris (1985) for a comprehensive treatment of traditional queueing models.

Queueing networks are not generally very tractable but there are a number of notable exceptions discussed in Supplement 2.5. Beyond these, other studied queueing networks include closed queueing networks and polling models. In a closed network, customers never leave but instead are recycled back to the beginning (this can represent the interaction between computer users and the core). Polling models have a single server that cycles through multiple classes of customers. Setup times between classes can be accommodated as can time for the server to walk between the different classes. In general, the availability of results depends on the restrictiveness of the assumptions.

Simulation

Most real queueing problems are not amenable to the type of mathematical analysis discussed in this section. Some of the reasons include:

1. The system is a queueing network. Queueing networks are common in manufacturing systems where the output of one process is the input to one or more other processes. They are also common in complex service systems such as emergency departments. Some systems are very complex with feedback loops and other unusual features. Such systems are generally too complicated to be exactly analyzed mathematically.
2. The interest is in transient rather than steady state behavior. The results discussed here are for steady state only. As stated earlier, steady state means that the system has evolved for a sufficiently long period so that initial conditions do not affect the probability distribution of system states. In many real problems, short-term behavior is an important part of proper systems management.
3. Interarrival times and/or service times are not exponential. Only approximations are available even for G/G/s queueing systems. Additional features, such as a finite waiting room, priority service, abandonment from the queue, and so on, can make such systems too complex to analyze mathematically.
4. Human behavior needs to be considered. The behavior of customers and/or servers may not obey a mathematical model. For example, in many service systems it has been observed that servers will speed up if the line is long. Also, customer behavior in switching/jockeying between parallel queues can be quite complex. Such behavior is usually too complex for queueing analysis.

One way to deal with a complex queueing system is simulation. A simulation (in our context) is a computer program that recreates the essential steps of a service process. In some sense, the computer "experiences" the process. In this way, one can see how the system responds to different parameter settings available to the system designer.

Simulations evolve over time; systems can be simulated for years in the space of a few seconds on a computer.

Problems amenable to simulation almost always involve some element of randomness. Computer-based simulators that incorporate randomness are called *Monte Carlo simulators*. At the heart of a Monte Carlo simulation are random numbers. See Appendix A–7 for more details on how such random numbers are generated.

Simulation can be used to analyze many types of complex problems with uncertainty but probably has been used most often for queueing problems. Simulators can be written in a general purpose programming language (such as Java or C#), or constructed using special-purpose simulation packages, such as ProModel, ARENA, or GPSS. Many packages are specifically designed for certain types of applications, such as queueing or network problems; MedModel is a package designed specifically for simulating medical facilities.

More recently, two approaches for developing simulations have become popular, opening up the use of simulation to a much wider audience. Spreadsheet programs have gained a great deal of popularity in recent years. Because many spreadsheets have a random number generator built in, they can be used to construct simulators. Excel, in particular, can generate observations from many distributions. The add-ins @Risk, Crystal Ball, or Risk Solver Platform facilitate building simulations in Excel; they are all very similar. These packages include a much wider array of distributions and convenient report generation (e.g., see Winston and Albright, 2012; Powell and Baker, 2013).

For complex queueing networks, graphical-based simulation packages are far easier to use than spreadsheets. These programs allow the user to construct a model of the system using graphical icons. Icons represent service facilities and waiting areas and arrows represent the direction of flow. Random arrivals and random service can be incorporated easily into the model. Both ProModel and ARENA, mentioned above, are graphics-based, and allow experienced users to build simulations of complex systems very quickly. Such programs employ live animation to show the simulation in action, and summary statistics are collected after the simulation has run its course. Further details on the type of simulation engine used in these packages may be found in Appendix 7–A.

Improving a Service Process

The queueing principles in this section yield insights into how to improve the performance of a service system. These are easiest seen using the following elements of the process.

1. *The arrival process*. The focus should be on reducing variability, possibly by scheduling appointments or using peak load pricing. Further, the arrival rate can be decreased by eliminating the need for the service, possibly by moving work offline.
2. *The service process*. If process rates can be increased then this will improve performance. Often lean programs can be used to improve efficiency. However, frequently forgotten is the improvement to be gotten from reducing service variability, which can be achieved by improved process design or further standardization of options.
3. *The scheduling rules and process flow*. As discussed above, priorities make a large difference to the service experience. Giving priority to customers with short service times improves overall expected waits. One way to achieve this and still appear fair is to have dedicated servers for small jobs, which is the approach taken by supermarkets.

If these servers can also be used for larger tasks when there are no small tasks available then this will eliminate any inefficiency caused by decreasing pooling.

4. *Pooling for added efficiencies*. Pooling allows servers not to sit idle while other servers are overworked. Even a small number of cross-trained servers can work to effectively pool work without the need for all servers to be trained in all tasks.
5. *Transform customers into servers*. If the customers can serve themselves then this will decrease the work content for paid servers. Examples include salad bars and self-check-in kiosks at airports.
6. *Make the wait "feel" less long*. Sometimes a firm does not actually need to decrease the waiting time if it can make it feel less long to the customer. Techniques for this are discussed in the next section.

Problems for Section 7.5

24. Customers send emails to a help desk that has three employees who answer the emails (and this is their only responsibility). Customer requests arrive according to a Poisson process with a rate of 30 per hour. It takes on average 4 minutes to write a response email; the standard deviation of this service times is 2 minutes.
    a. What is the utilization of the employees?
    b. What is the average time an email spends waiting before an employee starts working on it? What is the average time to complete the emails requests (from the time the email is sent to when it is answered)?
    c. How many emails on average are in the system queue waiting to be worked on?
    d. If the manager wants to decrease customer waiting, what options are available to him/her? Which do you recommend?
25. At the SuperSpeedy drive-through the time between consecutive customer arrivals has a mean of 50 seconds and a standard deviation of 30 seconds. There are two servers whose service time averages 80 seconds with a standard deviation of 20 seconds. Assume that no customers leave the drive-through after entry.
    a. What is the utilization of the employees?
    b. What is the average time a customer spends at the drive-through? What fraction of that is waiting in the queue?
    c. How many cars on average are in the drive-through lane (including those in service)?
    d. What suggestions would you have for the drive-through to improve customer satisfaction?
26. Modify Example 7.8 to find an example where pooling increases average waiting time across the two customer types.
27. What are the advantages and disadvantages of queueing analysis versus simulation?

7.6 THE HUMAN ELEMENT IN SERVICE SYSTEMS

As discussed earlier, one key characteristic of service processes is that they tend to involve the customer as co-producer. This means that human psychology and decision biases should be considered when designing service systems. This section outlines

Snapshot Application

USING QUEUEING TO MAKE STAFFING DECISIONS SAVES LIVES

Emergency departments (EDs) are complex queueing systems where too much waiting time can have severe negative health consequences. Delays in treatment have been shown to both increase mortality rates as well as cause patients to abandon the queue, known in the medical literature as leaving without being seen (LWBS). LWBS by itself is also associated with negative consequences because some patients leave because they feel too ill to wait, and actually need treatment. It is therefore not a stretch to say that improved staffing of EDs, if it results in lower waits and lower LWBS rates, saves lives.

Of course, one way to improve waits and LWBS rates is simply to add more staff. Unfortunately, because capacity is expensive and budgets are limited, this is usually not possible. Green et al. (2006) have provided a queueing model for ED staffing that decreased LWBS rates in an urban hospital even for periods *without* increased staffing levels. They modeled the nonlinear queueing effects that simple nurse to patient ratios, as frequently used by hospitals, do not account for.

Arrivals to EDs are, of course, highly nonstationary. To get around this they used a stationary independent period by period (SIPP) model where each independent period of time (e.g., each hour) is assumed to be stationary. They also made a so-called "lag" adjustment, which accounts for the fact that in nonstationary queueing models the arrival rate peaks before the waiting time.

A simple M/M/s queueing model was used to estimate the staff needed during each staffing interval. This estimate was based on the requirement that no more than 20 percent of patients wait more than one hour (i.e., a delay percentile metric). The model showed that staffing on weekdays should be increased from 55 to 58 provided hours but on weekends 53 provider hours were sufficient. However, the model also indicated a significant shift in when the hours were provided, from the middle of the night to much earlier in the day. Unfortunately, the ED did not have the budget to add all the additional staffing. However, they did shift hours around and add a few hours. Further, there were practical considerations to be considered. For example, it was not deemed possible to have different daily schedules and hence there were only two final schedules, namely, one for weekdays and one for weekends.

Even though not all of the recommended hours were given, the queueing model did result in better placement of the limited resources. Four hours fewer were provided on both Saturdays and Sundays with those extra hours going to the weekday schedule. Further, there is a four-day subset for which there were no more and no fewer hours, simply a rearrangement of schedules that was guided by the queueing model. In this time interval LWBS events declined from 9.2 percent to 7.2 percent even though the number of patients to arrive increased by 5.5 percent (548 patients) from the initial study period to the one with the new schedule. The schedule was a success!

Source: Green et al. (2006) and Wiler et al. (2013)

the psychology of queueing, guidelines for introducing technology into services, and principles for giving service guarantees and refunds.

The Psychology of Queueing

Maister (1985) has detailed a number of psychological principles for the design of queueing systems, as follows.

1. *Unoccupied time feels longer than occupied time*. If a customer is simply waiting not doing anything then his wait feels longer. Mirrors at elevators are often introduced to make the elevator wait seem shorter. There is an airport that increased the walk to baggage claim to cut down on complaints about delayed bags; unfortunately for customers, this strategy worked!
2. *Pre-process waits feel longer than in-process waits*. If the customer feels like he is progressing then he is more tolerant to the delay. For example, the wait in the doctor's waiting room is usually more frustrating than the wait in the examination room after having seen a nurse.

3. *Anxiety makes waits seem longer*. Worry makes customers more sensitive to their surrounding and less tolerant of waiting. For example, emergency room waits are in general long, but are certainly not helped by the patient being anxious and likely also in pain.
4. *Uncertain waits are longer than known, finite waits*. Uncertainty is a form of anxiety and will make the waiting less tolerable. Waits for a reservation are not a bother when the customer has arrived early but become increasingly unpleasant once the reservation time passes. Many call centers have understood this principle and now provide estimates of time on hold when the customer calls.
5. *Unexplained waits are longer than explained waits*. This principle again follows from the anxiety associated with uncertainty. Many airlines are getting better at understanding this principle and will provide customers with an explanation of why their flight is delayed, rather than simply telling them the estimated time of departure.
6. *Unfair waits are longer than equitable waits*. This principle relates to the mood of the customer. A negative mood, whether it is anxiety or anger, will make the wait less tolerable. Seeing someone cut in front of oneself in line makes most people angry!
7. *The more valuable the service, the longer I will wait*. This is natural. People will wait a long time for needed medical treatment but much less time for a coffee. Further, a customer's tolerance for waiting in queue is proportional to the complexity or quality of service anticipated by that customer.
8. *Solo waiting feels longer than group waiting*. This is a corollary to principle number 1 on unoccupied time. Waiting in a group is simply more pleasant than solo waiting.

The above together implies that it is not so much the duration of the delay that matters; it is what the customers experience, particularly relative to expectations, that matters most. Sensible firms consider these principles when they design their service. Often this can be done quite simply. For example, putting in ropes delineating the line so that customers cannot jump the queue will keep the queue fair.

Introducing Technology into Services

With the increasing role of technology in services (e.g., self-service checkouts, online ordering, helpful applications, etc.) companies must consider how to introduce technology carefully. Frei (2008) recommends the following three key principles.

1. *Be helpful before being intrusive*. If customers can see the obvious value of the technology then they will be much more tolerant of data gathering than if they view the technology as solely operating for the firm's benefit. For example, Facebook has been slowly increasing its intrusiveness; it is unlikely it would have reached today's popularity if it had launched with its current level of intrusiveness. As a second example, customer loyalty cards have provided customers with benefits even though they are used by firms to collect and analyze data on customer behavior. If the technology makes the service easier or more pleasant then customers are often willing to share data, which can be very helpful to firms looking to optimize their offerings.
2. *Roll out functions at a pace consumers can absorb*. Customers can get overwhelmed by too much technology. Firms need to consider what the customers can successfully navigate as they introduce technology. The firm may need to deliberately hold back functionality if it feels customers will not be able to absorb it. They should also consider the demographics of the customer base when introducing technology. Frei (2008) describes the case of Audi versus BMW. BMW overwhelmed

customers of its 7-series automobiles with too many new features (e.g., joystick control and keyless entry) and received significant customer backlash; Audi rolled out similar technology successfully.

3. *Framing matters*. If consumers feel the technology is simply being used for cost savings then they may revolt or turn away from the service provider. However, if the technology is presented to the customer as something that will improve their service experience then they will be much more likely to view it positively. For example, self-checkout machines are clearly designed to save on labor costs but, so long as customers see improved waiting time with these machines, they will be tolerant of the technology.

Guidelines for Service Guarantees and Refunds

Because the definition of good service is usually relative, firms often advertise service guarantees. Fitzsimmons and Fitzsimmons (2006) provide the following guidelines for such service guarantees.

1. *Focuses on customers*. The customer does not care about firm centric metrics and so service guarantees should be relative to customer-centric metrics, such as delay.
2. *Sets clear standards*. The guarantee should not be vague about what constitutes service within the guidelines.
3. *Guarantees feedback*. If customers are rewarded for reporting service failures that do not meet the guarantee then the firm has just implemented an effective method of quality control, as well as keeping customers happy.
4. *Promotes an understanding of the service delivery system.* A firm cannot introduce a service guarantee without knowing what is achievable. Therefore, introduction of the service guarantee must be preceded by a thorough understanding of the service process.
5. *Builds customer loyalty*. Happy customers are usually loyal customers. Thus, done right, service guarantees can build customer loyalty.

Unfortunately, service failures are all too frequent. In such cases, firms need to remedy the failure. Some key guidelines on such remedies, due to Hart (1988), are the following:

1. *Unconditional*. Customer satisfaction should not come with exceptions. Therefore, the refund should not come with strings attached.
2. *Easy to understand and communicate*. The customer should not be confused by the offer or misunderstand what is available.
3. *Meaningful*. If the offer is too trivial to have any real meaning to the customer then it may be better to not offer it at all.
4. *Easy to invoke*. Such offers should not consume large amounts of firm or customer resources to initiate.
5. *Easy to collect*. It should not be a headache for the customer to actually receive the discount or other remedy.

Of course, offering refunds that are conditional or difficult to collect will save money, but usually such savings are shortsighted. Recall the statistic that it costs five times more money to acquire a new customer than to retain a current customer; it becomes apparent why handling service refunds properly is so important.

Snapshot Application

DISNEY USES BOTH THE SCIENCE AND THE PSYCHOLOGY OF QUEUEING

A company that is a world leader in the science of queueing is Disney Corporation. They provide a service that is only of value for the experience it provides. It is therefore critical that customers leave Disney's parks feeling like they have had a great day.

According to Disney, they employ more than 75 industrial engineers who help with queue management at their parks around the world. They measure the capacity of the rides, optimize the flow within the rides, set the capacity of the queues, etc. However, Disney is also very aware of the psychology of queueing. Knowing that uncertain waits feel longer, they will post the expected waiting time at the start of each line. Further, knowing that unoccupied time also feels longer, they provide significant entertainment for customers in line for each ride. They have even started to introduce interactive entertainment in some of their queues.

One of the interesting technological innovations Disney has introduced is their Fastpass system. Available to anyone with a ticket, this system is in effect a reservation system for rides. Customers scan their ticket at a ride and are given a time to return within a half hour window. Because the time windows advance 5 minutes at a time, customers in groups will always have overlapping times (assuming they scan their tickets shortly after each other). The return times are displayed on a board and may be anywhere from an hour to many hours after the current time. At some point during the day, the system will run out of reservations and no more Fastpass tickets will be available for that ride. Customers can get a new Fastpass reservation (at the same or a different ride) at the sooner of two hours or when the reservation time has passed; thus, they are not able to run through the park getting reservations at all rides and will typically have only one outstanding Fastpass ticket at a time. Disney has arranged separate entries at rides with the Fastpass option; customers using those entries have little wait. They strictly enforce no entering before the allotted time but are usually less strict on customers who come after their allotted time.

There is a lot of science behind which rides Disney offers a Fastpass on and how much capacity gets allocated to the Fastpass reservations, most of which is, unfortunately, proprietary. One likely effect of the system is that many customers spend less time queueing and more time wandering the park and probably buying consumables (i.e., food and souvenirs), which is clearly good for Disney's bottom line. Some customers will of course simply stand in line for other rides while waiting for their Fastpass time, and those customers will probably end up riding more rides during a day than without the system. It is likely that the system has therefore made the lines for less attractive rides longer than they were before. Fully understanding the queueing implications of Fastpass is an interesting open research question.

Source: Disney.com and Pawlowski, A. "Queuing psychology: Can waiting in line be fun?" *CNN* (2008).

7.7 CALL AND CONTACT CENTERS

An important class of service system is the call or contact center. These can be *outgoing*, where agents sell products and services to potential clients, or *incoming* where customers call for service; we will focus on the latter. A call center is entirely phone call focused, whereas a contact center will answer emails and likely participate in web chats.

Call Center Basics

In a typical call center, calls come in to one of a number of parallel trunk lines. If no trunk lines are free, the caller will receive a busy signal; otherwise, they will reach the interactive voice response (IVR) system. Firms pay telecommunication companies for the number of truck lines they reserve. Thus, they may want to limit the number of lines to both save cost and because they prefer a customer to receive a busy signal rather than have a very long wait.

The IVR system interacts with the Automatic Call Distributor (ACD), which routes the call based on customer selections. This system may or may not be linked to the

central data server. Have you ever had the experience of typing in your customer ID number to the IVR system only to find that you have to tell your ID to the agent answering your call? If so, this is because the two computer systems were not connected; the first input of your customer ID was simply used to route your call based on your priority as a customer. Priority routing will be discussed later. Appropriate design of IVR systems is very important but outside the scope of this text.

As noted by Gans et al. (2003), in best practice call centers there are many hundreds of agents catering to many thousands of phone callers per hour. Agent utilization levels could average between 90 percent to 95 percent with no customer encountering a busy signal and about half of the customers receiving an answer immediately. The waiting time of those delayed is measured in seconds and the fraction that abandon while waiting varies from the negligible to a mere 1 to 2 percent. Clearly, many of the call centers we interact with are not best practice!

Just as nurse staffing decisions can benefit from queueing theory, so can call center staffing. In fact, call centers tend to have more access to data than hospitals; therefore, these days, call center staffing is much more a science than an art. Interested readers are referred to Green et al. (2007) for an excellent practical overview of using queueing models for staffing decisions.

Metrics

Because there is so much data available in call centers, and there is often a pressing requirement to drive out cost, metrics matter even more in call centers than in other operations systems. In Gans et al. (2003) it is detailed how agents were hanging up on customers, which of course made their average talk time metric look better. Clearly, there were not the proper systems in place to catch such undesirable behavior (e.g., random call monitoring), yet this is not a one-off occurrence. The second author has taught students with firsthand experience of similar behavior and there is even a Dilbert comic based on this phenomenon (with Dogbert being the agent hanging up on customers). Further, while such undesirable behavior can happen within a firm, it is even more likely when the metrics are driven by misaligned incentives across firms.

Just as in Section 6.10, contracts will drive incentives. Call centers are typically contracted by some form of *service-level agreement* (SLA). In such an SLA, firms agree to meet some desired level of service a certain percentage of the time. For example, firms may contract to have fewer than a given percentage of customers abandon their calls or that average delay across the day is no more than a given number 99 percent of the time. A very common contract for call centers is the *delay percentile contract* where the service level requirement is on customer delay.

Under a delay percentile contract, the provider agrees to meet a certain fraction of the calls, say 80 percent, within a certain window of time, say 20 seconds. However, because the contract leaves unspecified what should happen to the other 20 percent, the provider has an incentive to effectively drop calls that have exceeded the 20 second threshold, assigning them the lowest possible priority in the system. Milner and Olsen (2010) study this phenomenon and show that a better contract is one where there is a convex increasing cost to delays; of course, such contracts are less common in practice in part because they are more difficult to implement.

Delay percentile contracts have also been documented as leading to undesirable behavior in the British health system (BBC, 2011). According to this article, hospitals are required to treat patients within 18 weeks. This resulted in patients who had waited longer than 18 weeks being downgraded in priority (rather than upgraded as

one would want) relative to those who had not yet exceeded the (somewhat arbitrary) 18-week threshold. If the hospitals were instead measured on the total cumulative days that patients had waited beyond 18 weeks (a convex increasing cost), there would have been no incentive for this undesirable behavior. It should come as no surprise to the reader that metrics matter in service system design.

Call Routing

As mentioned earlier, calls to call centers are frequently routed by the priority of the customer to the firm. High priority customers receive both little to no wait and better trained agents than low priority customers. It is documented in Brady (2000) how different treatment is for the best customers versus the least valuable customers. If a strict segmentation strategy is followed, where high-priority customer agents are reserved only for high-priority customers and never used for lower priority customers, then this decreased pooling may result in very long delays for the low priority customers. It may also result in very low utilizations for the high-priority agents (to ensure good service). Therefore, most call centers will allow more flexibility in assignments of customers to agents.

While preferential treatment for priority clients is not a new phenomenon it does appear that asking customers to pay for priority is increasing in popularity. For example, United Airlines has introduced a portfolio of Visa cards that gives customers the benefits of preferred status without them needing to fly any miles. In some sense, paying for priority is a type of revenue management as discussed in the next section.

7.8 REVENUE MANAGEMENT

According to Boyd (2002), *revenue management* is "the science of maximizing profits through market demand forecasting and the mathematical optimization of pricing and inventory." There are a number of key terms to pull out of this definition. First, the decisions involved in revenue management surround both inventory and pricing. Here, "inventory," in the sense of revenue management, may mean seats of a certain class on an airplane or tickets for a certain area of a sports arena. Second, revenue management is a science, in that it uses data more than gut instinct in order to make these decisions. Finally, revenue management's key goal is to maximize the firm's profit, not the customers' welfare or what is "socially optimal." This section contains a number of examples where, indeed, revenue management may not be good for the consumer.

Since having been developed in the early 1980s as a result of U.S. airline deregulation, revenue management tools have spilled over to many other industries. Many companies find it particularly attractive because revenue management addresses the revenue side of the balance sheet rather than the cost side. Hence, it is less painful than downsizing or process re-engineering. Other application areas for revenue management include hotels, car rentals, rail tickets, tour operators, cargo freight, energy, entertainment, and restaurants. In fact, any service industry that faces finite perishable capacity and uncertain demand may benefit from revenue management tools.

Airline Revenue Management Overview

In the airline industry, the fundamental revenue management question is, at any point of time, how many seats of a certain class should be made available at what price? Seat classes may mean different types of seats in the airplane (e.g., business versus economy) but they also refer to the type of restrictions attached to the seat. For example,

an economy seat that can be purchased at any time with no restrictions and that is fully refundable will be more expensive than an economy seat that requires three week advanced purchase where there are penalties for changing the ticket after purchase.

Consider the following example due to Boyd (2002). A mid-size carrier might have 1000 daily departures with an average of 200 seats per flight leg, which results in $200 \times 1000 = 200{,}000$ seats per network day. If there are 365 network days maintained in inventory then there are $365 \times 200{,}000 = 73$ million seats in inventory at any given time. Therefore, the mechanics of managing final inventory represents a challenge simply due to volume. Airline revenue management therefore often treats price for a class as given and then optimizes inventory availability for that price. Price for each class is then set at a higher level; although, this is changing as computers get more powerful. Even setting inventory availability requires sophisticated forecasting techniques that can determine customer demand and willingness-to-pay for each class.

Effective revenue management can have a significant impact but requires sophisticated tools to be done correctly. One of the key firms operating in this space is *PROS Revenue Management.* It was founded in 1985, and claims to be the "largest and most experienced company focused only on pricing and revenue management analytics, execution, and optimization." According to their website, as of 2013 they have over 700 employees with over 30 PhDs, and 2012 revenues of $118 million.

Revenue Management Basics

Revenue management involves dividing customers by their willingness to pay (i.e., *market segmentation*). Consider the following simple example. Suppose that there are 1000 potential customers for some service. The customers differ in their valuation of the service and are willing to pay between $0 and $100 for it, with all values in between equally likely (i.e., customer valuation follows a uniform distribution). If the firm charges $0 then all 1000 customers will buy the service but the firm will make no money. If the firm charges $100 then no one will buy the service and again they make no money. If they charge $x ($0 < x < $100) then they make $x $(100 - x)1000$. Thus, it is easy to show that, if they have no capacity constraints and costs are fixed, then the optimal amount to charge is $50 so that half the customers buy service and they make revenue of $25,000.

Now, suppose the firm is able to introduce a service-plus option that they charge $75 for, the standard service still costs $50, and they offer service-minus for $25. Further, suppose that service-plus is designed to appeal to all customers with high valuations, so that all 250 customers with valuations over $75 buy it. Customers with valuations between $50 and $75 buy standard service. Finally, service minus is bought by customers with valuations between $25 and $50. (This is not very realistic but is intended as an example only.) Now the firm makes revenue of $\$25 \times 250 + \$50 \times 250 + \$75 \times 250 = \$27{,}500$ (a 10 percent revenue increase). The difficulty is setting the conditions for service-minus so that they are not attractive to any (or too many) customers willing to pay at least $50 and conditions for regular service so that they are not attractive to any (or too many) customers willing to pay at least $75. Otherwise, all customers will simply buy service-minus and the firm will only make $\$25 \times 750 = \$18{,}750$ revenue.

Revenue management involves finding offerings that appeal to price-sensitive customers and more expensive offerings that will appeal to high-value customers. Hotels use premium offerings such as executive lounges to extract revenue from nonprice-sensitive customers, and use discounting sites such as priceline.com to sell unbooked hotel rooms to price-sensitive customers.

Thus, revenue management is both a science and an art. The art lies in appropriately defining product or service offerings that will appeal to different customer segments. The science is in forecasting the best prices and quantity of offerings for the different segments at different points in time. The optimization algorithms behind revenue management are beyond the scope of this text but the interested reader is referred to Talluri and van Ryzin (2004) for a good overview.

Lead Time Pricing

Another potential application for revenue management tools is dynamic lead time pricing. Amazon has already worked out how to segment customers by urgency of delivery; their "super saver" shipping is free but if you want the item in two days, it will be expensive. However, Amazon has close to unlimited capacity for delivery. In smaller service firms, or make-to-order manufacturing, capacity is more limited. Thus, it would be dangerous for a small firm to guarantee a short lead time for a single fixed price as Amazon does.

As a concrete example, many higher-end U.S. furniture stores offer customers a choice of fabrics for sofa and chair purchases. The customer places their order and then frequently waits three months or more for delivery. This author has yet to see a store where customers are offered the option to pay more to get their custom furniture sooner, yet shouldn't this be an option, at least for furniture made domestically? Further, shouldn't the price paid depend on the firm's current order backlog (which should be easily assessable through the ERP system)?

As another example, consider the automobile industry. Both Ford and GM have publically stated a goal of allowing people to custom-order cars online and receive delivery in three to five days. GM has also discussed having a "premium" delivery service of one to four days. Further, most high-end cars in Europe are made-to-order, yet lead time dependent pricing does not seem to be much practiced yet. Revenue management does not appear to have yet made it to the automobile industry.

Dynamic lead time pricing allows the quotation to depend on current congestion. With static quotation/pricing the firm will need to work to a worst case bound, which may be very long. For example, consider an M/M/1 queue at 90 percent utilization. Using the formulas from Section 7.4, the expected lead time is 10 times the expected service time and the 95th percentile of lead time is around 30 times the expected service time. Yet, 10 percent of customers wait no time, and 39 percent of customers wait five times the expected service time or less. Clearly, there is value to be had by offering short lead times to customers willing to pay for them, especially when congestion is light. Conversely, when congestion is heavy, offering a discount to customers willing to wait may alleviate the need for overtime costs. One suggestion for implementing such lead time dependent pricing, motivated by the work by Akan et al. (2013), is as follows. First, decide on a time unit to quote lead time in (e.g., days or weeks). Then predict what fraction of customers will pay for "premium" delivery, which is the fewest number of time units that it is practical to produce and deliver the product to the customer if the system was empty. Next, reserve approximately that fraction of capacity per time unit (e.g., one couch per day). However, the firm should never waste the reserved capacity, but rather use it for backlogged nonpremium customers. They can either deliver early or store until it is time to ship, depending on whether they think early delivery will cannibalize demand for premium delivery. The firm should not offer premium delivery if the capacity has already been committed. Further, they should consider a "super saver" discount for long lead times when they are busy, which will act as a tool for smoothing

demand. The second author believes that there is a true revenue building opportunity available for firms willing to consider such dynamic pricing techniques for lead times.

Nontraditional Applications for Revenue Management

There is a current trend towards more nontraditional applications for revenue management. Lead time pricing, as described above, is one such application but there are many others. For example, ticket pricing for baseball games has become quite sophisticated (e.g., Biderman, 2010) as has ticketing for other events such as movies, concerts, and other sporting events. Promoters often use revenue management principles to decide on their pricing strategies.

Of course, airlines have long used revenue management to price their seats, but they are now getting creative about using it in other ways as well. Offers are made at the time of check-in for upgrades. Air New Zealand runs a type of auction for upgrades where passengers bid what they are willing to pay for a one-class upgrade and the highest bids win the upgrades. They also run a reverse auction system where fixed seats on a pair of flights are offered for two fixed dates (with the gap between the flights being sufficient for a vacation). The auction runs from a high value down to the reserve value and stops when a single customer has bid on, and therefore bought, the flight pair (this is known as a Dutch auction).

It is interesting to speculate on where revenue management is leading in the future. The rise of "big data" and business analytics is likely to only increase its opportunities, not all of which will be good for the consumer. The wise manager will keep an eye out for opportunities, but also be wary of alienating existing customers.

Problems for Sections 7.6–7.8

28. Give an example (other than Disney) of a firm that you think is good at using queueing psychology. Explain your answer.
29. Consider your most recent unpleasant waiting experience. What went wrong and how could it have been improved?
30. Give an example of a firm that has recently introduced or increased the use of technology in its service offering. How did they perform relative to the three guidelines given in Section 7.6?
31. Give an example of a service guarantee you are familiar with. How does it rate relative to the guidelines in Section 7.6?
32. Give an example of a service failure you have experienced. Was there any remedy given? What could the firm have done better?
33. Write (and send) a letter or email to a firm detailing a recent service failure you have personally experienced. Analyze the reply (if any) with respect to the guidelines in Section 7.6.
34. Disney's Fastpass system is free to all ticket holders but other theme parks, such as Universal Studios, charge a fee for priority queue access. Discuss the advantages and disadvantages of such charging for priority.
35. List the possible metrics that may be used by call centers to measure performance. Which of these are easiest to measure? Which are most important from the customer's perspective?
36. Give an example of an application of revenue management that you have seen or heard of outside of airlines and hotels. Describe how it is offered.
37. Give an example of an industry you have not seen revenue management applied to but that you believe it could be applied to. Why do you think it has not been used?

7.9 HISTORICAL NOTES AND ADDITIONAL READINGS

Service operations management is a significantly younger distinct research field than either manufacturing or inventory management. For example, it wasn't until the late 1980s that it was recognized as a discipline within the *Decision Science Institute* (Fitzsimmons & Fitzsimmons, 2004, p. xv). Moreover, it wasn't until 2007, within the *Institute for Operations Research and Management Science*, both the special interest group on *Service Management* (within the society of *Manufacturing and Services Operations Management*) and a new section on *Service Science* were founded. It is now a thriving area of research encompassing important subfields such as healthcare management, call center management, and revenue management.

There are a number of comprehensive textbooks on service management including Fitzsimmons and Fitzsimmons (2004), Haksever et al. (2000), and Metters et al. (2002). Such texts cover the material in this chapter in further detail. They also typically cover other operations topics found in this text, such as forecasting or quality management, but from a services slant. It should be noted that most of the building blocks for successful operations management are not fundamentally different whether one is applying them to service or production systems.

Queueing theory as a discipline is much older than service operations management. A. K. Erlang, a Danish telephone engineer, developed many of his theories in the early 1900s. His work, Erlang (1909), both modeled the number of calls with the Poisson distribution and solved for the mean delay in an M/D/1 queue. However, the Poisson distribution itself is around a century older than Erlang's work. Queueing theory became a thriving research field in the 1960s and 1970s, although often from a purely mathematical standpoint. Kleinrock (1975; 1976) is an important pair of early texts on the subject, covering theory and applications, respectively. Gross and Harris (1985) is a comprehensive text on different queueing models. Many textbooks devoted entirely to queueing exist.

Revenue management was originally called yield management. It was used primarily in the airline industry to fill seats. One of the earliest published applications is American Airlines, where even in the 1990s the program was generating close to $1 billion in revenue annually (Cook, 1998). Revenue management has also been called "revenue optimization," "demand management," and "demand chain management." As a research discipline it did not really take off until the 2000s. Talluri and van Ryzin (2004) is one of the earliest and most comprehensive texts on the field.

7.10 Summary

Services are an increasing proportion of the economies of developed nations, but have traditionally lagged production systems in operational efficiency. Two of the key reasons for this are their need to include the customer as a co-producer and the perishable nature of most service capacity. This chapter examined tools for analyzing and managing service systems. Some key tools that were described for mitigating the mismatch between supply and demand in service systems include the following:

1. Turn customers into servers (i.e., no supply/demand mismatch).
2. Better predictive models (i.e., better forecasting).
3. Pooling (unless it increases service variability significantly).
4. Queueing models to predict capacity to meet acceptable wait standards.

5. Queueing models and simulation to guide improvement.
   –Schedule different customer types differently.
6. Decrease or accommodate variability.
7. Make system performance visible to employees and align incentives accordingly.
8. Tools for "lean" operations (see Chapter 8).
9. Don't forget the psychological aspects.
10. Revenue management tools.

One of the most often forgotten tools above is the need to decrease or accommodate variability in order to improve performance. Variability has a large impact on delays, both on their mean and their distribution. In most service systems, long delays lead to customer frustration and/or abandonment and need to be avoided. An often ignored benefit of lean improvement programs is their drive towards standardization and hence the reduction of variation. However, many other tools, such as the careful use of pricing, can also be helpful in reducing variability.

Additional Problems for Chapter 7

38. Consider the following commercial bread making process. First, the dough is mixed in batches in the single mixer and it takes 15 minutes for a batch of dough that will produce 100 loaves. The batch of dough is then proofed, which takes an hour but has no effective capacity constraint. Then the dough is baked in the single oven, which takes half an hour, and is again done in batches of 100 loaves. Finally, the loaves are sliced and packed one at a time into bags, which takes 30 seconds per loaf on one of two slicing and bagging machines. What is the capacity of the mixer, oven, and the bagging machines? What is the bottleneck step? What is the capacity of the system?
39. Suppose that arrivals to a hairdresser follow a Poisson process with mean 12 customers per hour. Calculate the probabilities of (a) no customers in 20 minutes; (b) exactly one customer in 5 minutes; (c) exactly 12 customers in a hour; and (d) fewer than three customers in 10 minutes.
40. The following data has been collected on the interarrival times of patients to an emergency department: 3.772, 1.761, 0.743, 15.988, 0.412, 7.541, 6.900, 3.447, 7.024, 1.061, 5.449, 0.309, 0.766, 4.807, 8.143, 0.093, 9.524, 0.012, 4.634, and 0.195 minutes. Estimate the squared coefficient of variation of the arrival process. Is the arrival process likely to be Poisson? Why or why not? Estimate the arrival rate?
41. The following data has been collected on the *number* of customers seen to arrive to a doctor's office in a succession of 15 minute intervals: 4, 5, 4, 3, 3, 3, 1, 2, 1, 3, 2, 4, 1, 2, and 4. Estimate the squared coefficient of variation of the arrival process. Is the arrival process likely to be Poisson? Why or why not? Estimate the arrival rate?
42. Suppose that on average we observe 50 people at the beach and a rate of arrivals of one person every 3 minutes. How long, on average, do people spend at the beach?
43. Patients arrive to a small hospital emergency room according to a Poisson process with an average rate of 1.5 per hour. Because service times for these patients vary considerably, the service times are accurately described by an exponential distribution. Suppose that the average service time is 26 minutes. If there is only a single doctor working at any point in time, find the following measures of service for this emergency room:
    a. The expected total time in the system.
    b. The expected time each customer has to wait.

c. The expected number of patients waiting for service.

d. If the emergency room has a triage nurse who gives priority to more serious conditions, explain qualitatively how that would change your results to parts a, b, and c. What additional information would one need to know to analyze a system like this?

44. Students arrive one at a time, completely at random, to an advice clinic at a rate of 10 per hour. Students take on average 5 minutes of advice but there is wide variation in the time they need; this variation may be well modeled by the exponential distribution.

    a. Assume there is only one advisor serving in the clinic. Find the following expected measures of performance for this system: the expected time in the clinic, the expected time in the queue for advice, the expected number of students in the clinic, and the expected number of students waiting for advice.

    b. Again assuming one advisor, what is the probability that there are more than ten students in the clinic at a random point in time? What is the probability that the time spent in the clinic exceeds 30 minutes?

    c. Now suppose that a second advisor is hired. Repeat your answer for a. What are the advisors' utilizations? Would you recommend this second advisor is hired?

    d. What are some ways that the advice clinic could improve students' experiences in the advice clinic without hiring more staff?

45. An ice cream truck is parked at a local beach and customers queue up to buy ice creams at a rate of one per minute. The arrival pattern of people buying ice cream is essentially random. It takes 40 seconds on average to serve a customer ice cream, with a standard deviation of 20 seconds. Find the following expected measures of performance for this system: the expected time in the queue for ice cream, the expected total time to get an ice cream, and the expected number of customers waiting for ice cream.

46. Cars travelling to the George Washington Bridge from New Jersey to New York City must pay a toll. About 38 percent of the commuters use E-ZPass, which registers the toll to their account electronically. E-ZPass customers go quickly through the toll area averaging a wait time of 30 seconds because of the need to slow down. However, paying customers must queue up at the cash booths. They require an average service time of twenty seconds each with a standard deviation of 10 seconds. If cars are arriving to the toll area at an average rate of 8 per minute and there are 3 cash toll booths, what is the ratio of total time in the system for E-ZPass commuters versus cash commuters?

47. A local café has a single cash register, with a single assistant to work it, and three servers working to fill the customer orders. Customers arrive with exponential interarrival time an average of one every 2 minutes. The time to place their order and pay at the register is normally distributed with mean 90 seconds and standard deviation 20 seconds. Each customer's order is then passed to one of the servers who take on average 5 minutes with standard deviation 1.5 minutes, also normally distributed, to fill the order.

    a. Calculate the capacity of the register and the servers. What is the bottleneck in this system?

    b. Calculate the average utilizations of the register and the servers

    c. What is the probability a customer is delayed at the register?

    d. What is the expected time from a customer's arrival to the order being passed on to the servers (including any queueing time)?

    e. Estimate the probability that there is a delay between a customer placing his order and a server beginning to work on the order.

f. Using the formula in S2.5 in Supplement 2, estimate the squared coefficient of variation of arrivals of orders to the servers.

g. Estimate the expected time from the servers receiving an order to it being ready for the customer (including any queueing time).

h. If we add (d) and (g) we get the total time from a customer walking in to receiving their order. What assumptions have been made to compute this time? Which ONE of these assumptions is the least realistic for this system? Explain your answer.

i. List as many ways you can think of that would decrease the average time a customer spends waiting from placing their order to receiving their food.

Appendix 7–A

Simulation Implementation

This appendix discusses how simulations are implemented by computers. In particular, it discusses random number generation and entity-driven logic for process simulations.

Random Number Generation

At the heart of any Monte Carlo simulation are random numbers, which in this context are drawn from a uniform (0,1) distribution. That is, they are numbers between zero and one with the property that every number drawn has an equal likelihood of being selected. Random number generators are algorithms that produce what appear to be independent realizations of uniform variates. The algorithms used do not cycle for a very large number of steps, thus producing number sequences that appear random. However, because the recursive algorithms used are deterministic, the resulting string of numbers is referred to as "pseudorandom" numbers (e.g., see Fishman, 1973).

From uniform variates, one obtains observations of random variables having virtually any distribution using results from probability theory. For example, the central limit theorem says that the sums of independent random variables are approximately normally distributed. (This is a *very* loose statement of the central limit theorem.) Therefore, the sum of a reasonable number of independent draws from a uniform (0,1) distribution will be approximately normal. (Convergence occurs very quickly so the number does not have to be very large.) For example, if we let U_1, U_2,… be successive draws from a uniform (0,1) distribution, then

$$Z = \sum_{i=1}^{12} U_i - 6$$

is approximately standard normal. (This is meant for illustrative purposes only. In practice, there are more efficient ways to generate normal variates.)

Entity Driven Logic

The interface for most graphical based high-level simulation packages (such as ProModel or ARENA) is known as *entity-based logic*. An entity is any object or customer that moves through the processes in the simulation. Entities are the "brains" in the

simulation. If the modeler needs anything to happen that is not a predefined function in the package, an entity needs to do it. Therefore, an entity may serve as a breakdown demon or a lunchtime angel, if such events need to depend on more than just time elapsed.

In general, entities move from one station or location in the simulation to the next. Locations may represent servers, queues, decision points, transportation, etc. Most simulation packages will try to move an entity as far as possible through the process before the entity encounters a delay. At that point, the entity is placed on a queue and the next entity ready for movement is picked up. This continues until there is nothing more that can occur at the current time and the simulation clock is advanced to the earliest next event (which may be an entity arrival, an entity completing service, or any other time-flagged event). This is the reason such simulation models are often called *discrete event simulation*.

The implication of this type of logic is that some innocuous looking processes can be remarkably difficult to simulate, if they are not well modeled by entity-driven logic. For example, consider customers queueing at supermarket checkouts who are willing to jockey between queues, if the line next to them gets shorter. At any given point in time, there will be customer entities queueing at each of the server resources. If there was no jockeying allowed then the server resources would simply pick the next customer in line whenever a departure occurs from their station, which is very easy to implement in any major simulation package. However, with jockeying, whenever a departure occurs all entities need to re-evaluate whether they are going to switch queues. Entity-driven logic means that this involves releasing all waiting entities from some sort of gate process, routing them through a decision node, and then routing them back to the gate process. The gate process needs to be set up by the user, rather than using built-in resource queues. There is no simple way to put customers in line for a checkout server resource and then release them to jockey in any commercial simulation package that this author has used. Thus, simulations can get complicated quite quickly, if the process consists of more than simple flows of entities through locations.

Bibliography

Anupindi, R., S. Chopra, S. D. Deshmukh, J. A. Van Mieghem, and E. Zemel. *Managing Business Process Flows.* 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 2011.

Ayhan, H., and T. L. Olsen. "Scheduling of a Multi-Class Single Server Queue Under Non-Traditional Performance Measures." *Operations Research* 48 (2000), pp. 482–489.

BBC News. "Royal Cornwall Hospital Patients 'Jumping Waiting List'" August 27, 2011. Accessed from http://www.bbc.co.uk/news/uk-england-cornwall-14686568

Biderman, D. "When Did Buying Tickets Get So Complicated?" *Wall Street Journal*. January 4, 2010. Accessed from http://online.wsj.com/article/SB10001424052748704065404574636622642639610.html

Bolandifar, E., N. Dehoratius, T. L. Olsen, and J. Wiler. "Modeling Abandonment from the Emergency Department." Working paper, (2013).

Boyd, E. A. *Revenue Management and Dynamic Pricing: Part I.* 2002. https://www.ima.umn.edu/talks/workshops/ 9-9-13.2002/boyd/boyd.pdf

Brady, D. "Why Service Stinks." *Businessweek*, October 23, 2000.

Clark, C. *The Conditions of Economic Progress.* London: MacMillan & Co. Ltd., 1940 (revised and reprinted in 1951).

Cook, T. S. "Sabre Soars." *OR/MS Today* (1998), pp. 26–31.

Erlang, A. K. "The Theory of Probabilities and Telephone Conversations." *Nyt Tidskrift for Matematik B* 20 (1909), p. 33.

Feller, W. *An Introduction to Probability Theory and Its Applications.* Vol. 2. New York: John Wiley & Sons, 1966.

Fisher, A. G. B. *The Clash of Progress and Security.* London: MacMillan & Co. Ltd.,1935.

Fishman, G. *Concepts and Methods in Discrete Event Digital Simulation.* New York: John Wiley & Sons, 1973.

Fitzsimmons, J. A., and M. J. Fitzsimmons. *Service Management: Operations, Strategy, and Information Technology.* New York: McGraw-Hill/Irwin, 2004.

Fortino, M. *Chicago Tribune*, June 21, 1988.

Fourastié, J. *Le Grand Espoir du XXe Siècle.* Paris: Presses Universitaires de France, 1949.

Frei, F. X. "Breaking the Trade-Off Between Efficiency and Service." *HBR Magazine*, November 2006.

Frei, F. X. "The Four Things a Service Business Must Get Right." *HBR Articles*, April 2008.

Frei, F. X. "Commerce Bank." *HBR Cases*, December 2002.

Frei, F. X. "Rapid Rewards at Southwest Airlines." *HBR Cases*, August 2001.

Frei, F. X., and H. Rodriguez-Farrar. "Innovation at Progressive (A): Pay-As-You-Go Insurance." *HBS Case Teaching Note* 5-608-044. April 2008.

Gans, N.; G. Koole; and A. Mandelbaum. "Telephone Call Centers: Tutorial, Review, and Research Prospects." *Manufacturing & Service Operations Management (M&SOM)* 5 (2003), pp. 79–141.

Green, L., and P .J. Kolesar. "The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals." *Management Science* 37 (1991) pp. 84-97

Green, L.; P. J. Kolesar; and W. Whitt. "Coping with Time-Varying Demand when Setting Staffing Requirements for a Service Systems." *Production and Operations Management* 16 (2007), pp. 13–39.

Green, L.; J. Soares; J. Giulio; and R. Green. "Using Queueing Theory to Increase the Effectiveness of Physician Staffing in the Emergency Department." *Academic Emergency Medicine* 13 (2006), pp. 61–68.

Gross, D. and C. M. Harris. *Fundamentals of Queueing Theory*. 2nd ed. New York: John Wiley & Sons, 1985.

Haksever, C.; B. Render; R. S. Russell; and R. G. Murdick. *Service Management and Operations*. 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 2000.

Hart, C. W. L. "The Power of Unconditional Service Guarantees." *HBR Articles*, July 1988.

Hart, C. W. L.; J. L. Heskett; and W. E. Sasser Jr. "The Profitable Art of Service Recovery." *HBR Articles*, July 1990.

Hassin, R., and M. Haviv. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems.* Boston: Kluwer Academic Publishers, 2003.

Henkoff, R. "Service is Everybody's Business." *FORTUNE Magazine*, June 27, 1994.

Kendall, D. G. "Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain." *The Annals of Mathematical Statistics* 24 (1953), pp. 338–354.

Kleinrock, L. *Queueing Systems. Vol. I: Theory.* New York: Wiley Interscience, 1975.

Kleinrock, L. *Queueing Systems. Vol. II: Computer Applications.* New York: Wiley Interscience, 1976.

Lariviere, M. and J. A. Van Mieghem. "Strategically Seeking Service: How Competition Can Generate Poisson Arrivals." *Manufacturing & Service Operations Management* 6 (2004), pp. 23–40.

Lusch, R. F., and S. L. Vargo. *Service-Dominant Logic: Premises, Perspectives, Possibilities*. Cambridge, U.K.: Cambridge University Press, 2014.

Maister, D. H. "The Psychology of Waiting Lines." *Technical Report* (1985). Accessed from http://davidmaister.com/articles/the-psychology-of-waiting-lines/

Metters, R. D.; K. H. King-Metters; and M. Pullman. *Successful Service Operations Management*. Cincinnati, OH: South-Western Publishing, 2002.

Milner, J. M., and T. L. Olsen. "Service-Level Agreements in Call Centers: Perils and Prescriptions." *Management Science* 54 (2010), pp. 238–252.

Naor, P. "The Regulation of Queue Size by Levying Tolls." *Econometrica* 37 (1969) pp. 15–24.

Porter, M. E. "How Competitive Forces Shape Strategy." *Harvard Business Review* (March/April 1979).

Powell, S. G., and K. R. Baker. *Management Science: The Art of Modeling with Spreadsheets*. 4th ed. New York: Wiley, 2013.

Reichheld, F. F. *The Loyalty Effect.* Watertown, MA: Harvard Business School Press, 1996.

Rother, M. and J. Shook. *Learning to See: Value-Stream Mapping to Create Value and Eliminate Muda*. Brookline, MA: Lean Enterprise Institute, 2003.

Sakasegawa, H. "An Approximation Formula $L_q \simeq \alpha \cdot \rho^{\beta}/(1 - \rho)$," *Annals of the Institute of Statistical Mathematics* 29 (1977), pp. 67–75.

Talluri, K, and G. van Ryzin. *The Theory and Practice of Revenue Management*. New York: Springer-Verlag, 2004.

The Telegraph. "Britons Spend Six Months Queuing." September 15, 2013. Accessed from http://www.telegraph.co.uk/news/newstopics/howaboutthat/5052956/Britons-spent-six-months-queuing.html

U.S. Bureau of Labor Statistics. "International Comparisons of Annual Labor Force Statistics, 1970–2012." Accessed from http://www.bls.gov/fls/flscomparelf.htm

Wang, K.; N. Li; and Z. Zhiang. "Queueing Systems with Impatient Customers: A Review." *Proceedings of the IEEE international conference on service operations and logistics and informatics (SOLI)*, 2010.

Winston, W. L., and S. C. Albright. *Practical Management Science*. 4th ed. Nashville, TN: South-Western Cengage Learning, 2012.

Whitt, W. "Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments." *Management Science* 50 (2004), pp. 1449–1461.

Wolff, R. W. *Stochastic Modeling and the Theory of Queues.* Englewood Cliffs, NJ: Prentice Hall, 1989.

Supplement **Two**

Queueing Techniques

Chapter 7 described those aspects of queueing theory that are most critical for managing service systems. However, queueing theory is a very large field and is the subject of numerous stand-alone textbooks. This supplement contains further important queueing theory results and covers some of the more technical details that were deemed too distracting for Chapter 7. In particular, it covers further details on the Poisson process and exponential distribution, derives the results given for the M/M/1 queue, covers further M/M queue results, gives some infinite server queueing results, briefly covers queueing networks, and touches on the optimization of queueing systems.

S2.1 DETAILS OF THE POISSON PROCESS AND EXPONENTIAL DISTRIBUTION

This section details some of the more technical details around Poisson arrival processes. In particular, as discussed in Section 7.3, the key assumptions for a Poisson process $\{N(t): t \geq 0\}$ are as follows:

1. The number of arrivals in disjoint intervals are independent;
2. The number of arrivals in an interval depends only on the interval's length; and
3. For a very short interval (of duration h):
   a. the probability of one arrival is approximately λh; and
   b. the probability of more than one arrival is negligible.

These assumptions can be used to derive the Poisson distribution. Mathematically we write them as follows:

a) $P\{N(t+s) - N(t) = n;\ N(t) - N(0) = m\} = P\{N(t+s) - N(t) = n\}\ P\{N(t) - N(0) = m\}$ for any $s,\ t \geq 0$ and integers $m,\ n \geq 0$.

b) $P\{N(t+h) - N(t) = 1\} = \lambda h + o(h)$.

c) $P\{N(t+h) - N(t) > 1\} = o(h)$.

Where $o(h)$ is a function such that $\lim_{h \to 0} o(h)/h = 0$. We can then write

$$
\begin{aligned}
P\{N(t+h) = n\} &= \sum_{m=0}^{n} P\{N(t) = m\} P\{N(t+h) - N(t) = n - m\} \\
&= P\{N(t) = n\}(1 - \lambda h + o(h)) + P\{N(t) = n-1\}(\lambda h + o(h)) + o(h) \\
&= P\{N(t) = n\}(1 - \lambda h) + P\{N(t) = n-1\}(\lambda h + o(h)).
\end{aligned}
$$

Defining $P_n(t) = P\{N(t) = n\}$, and letting $n = 0$ we have that

$$\frac{p_0(t+h) - p_0(t)}{h} = -\lambda p_0(t) + o(h)/h.$$

Taking the limit as $h \downarrow 0$ gives (through using differential equations and the fact that $p_0(0) = 1$)

$$p_0(t) = e^{-\lambda t}.$$

Further,

$$\frac{p_n(t+h) - p_n(t)}{h} = -\lambda(p_n(t) - p_{n-1}(t)) + o(h)/h.$$

Again taking the limit as $h \downarrow 0$, this equation can be used recursively to find that (as desired)

$$P\{N(t) = n\} = p_n(t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \text{ for } n = 0, 1, 2...$$

As noted in Section 7.3 and shown in Section 13.3, the Poisson process can also be derived from an assumption of exponential interarrival times. The exponential distribution has a property related to the memoryless property that is particularly useful in queueing analysis. It has to do with what are known as forward and backward recurrence times. Let $N(t)$ be a Poisson process with rate λ, and T_1, T_2, ... be successive interarrival times. Consider some deterministic time t that falls between the two success interarrival times, say T_{i-1}, and T_i. The forward recurrence time is the random variable $T_i - t$, or the time that elapses from t until the next arrival. The exponential distribution is the only distribution that has the property that the distribution of the forward recurrence time also has the exponential distribution with rate λ *independent of t.* In queueing, this means that if a server is busy when a customer arrives, the amount of time that elapses until the completion of service is still exponential with rate μ. This leads to an apparent paradox.

It turns out that the backward recurrence time of a Poisson process with rate λ, $t - T_{i-1}$, also has the exponential distribution with rate λ. The astute reader will sense something wrong here. Adding $t - T_{i-1}$ and $T_i - t$ gives $T_i - T_{i-1}$, which is just one interarrival time. However, if $t - T_{i-1}$, is exponential distribution with rate λ and $T_i - t$ is also exponential with rate λ, it should follow that $E(t - T_{i-1} + T_i - t) = E(T_i - T_{i-1}) = 2/\lambda$, contradicting the assumption that interarrival times are exponential with rate λ! This apparent contradiction is known as the **waiting time paradox** or the **inspection paradox**. It has to do with the fact that we picked a point in time at random and found the interval that included that point rather than picking an interval at random. Intervals covering a random point are twice as long, on average. We will not dwell on this point here, but note that it perplexed mathematicians (as it is probably perplexing the reader) for many years. We hope that the interested reader will follow up on their own. A good starting point is the excellent discussion in Feller (1966, p.11). Kleinrock (1975, p. 169) also discusses the paradox in the context of queueing.

S2.2 ANALYSIS OF THE M/M/1 QUEUE

This section derives the results given in Section 7.2 for the M/M/1 queue using what is known as a **birth and death analysis**. The process $N(t)$, the number of arrivals up until time t, is a pure birth process. It increases by one at each arrival. The process $L(t)$,

the number of customers in the system at time t, is known as a birth and death process because it both increases and decreases. It increases by one at each arrival and decreases by one at each completion of a service. A realization of $L(t)$ is shown in Figure S2–1.

Notice that the state of the system either increases by one or decreases by one. The intensity or rate at which the system state increases is λ and the intensity at which the system state decreases is μ.[1] This means that we can represent the rate at which the system changes state by the diagram in Figure S2–2.

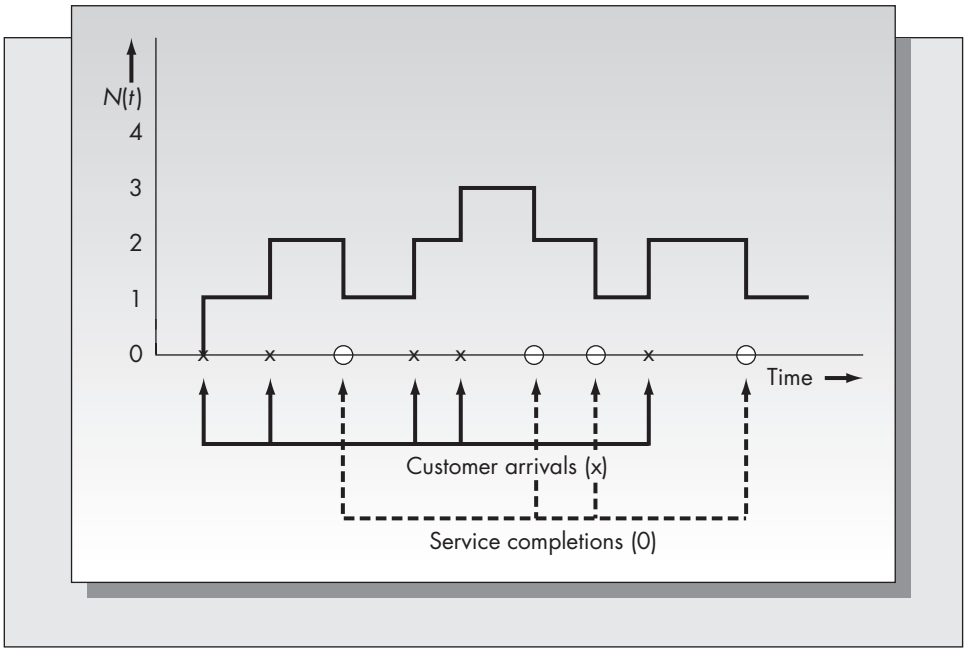


FIGURE S2–1
The process $L(t)$

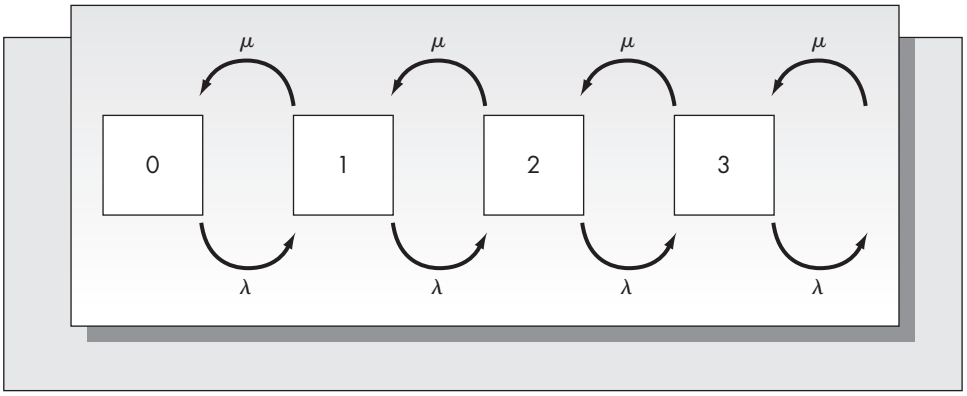


FIGURE S2–2
State changes for the M/M/1 queue

[1] At this point we consider only the case in which the arrival and the service rates are fixed and independent of the state of the system. The extension to the more general case will be considered in Supplement 2.3.

Let us suppose that the system has evolved to a steady-state condition. That means that the state of the system is independent of the starting state. Because we are in steady state, we consider only the stationary probabilities p_n. The following derivation is based on the Balance Principle:

> **Balance Principle:** In the steady state, the rate of entry into a state must equal the rate of entry out of a state if a steady state probability distribution exits.

Consider the application of the Balance Principle to state 0. We enter state 0 only from state 1. Given that we are in state 1, we move from state 1 to state 0 at a rate μ (see Figure S2–1). The probability of being in state 1 is p_1. It follows that the rate at which we move into state 0 is μp_1. Consider the rate at which we move out of state 0. When we are in state 0, we can only move to state 1, which we do (when a customer arrives) at rate λ. As the probability of being in state 0 is p_0, it follows that the overall rate at which we move out of state 0 is λp_0. From this we obtain our first balance equation:

$$\mu p_1 = \lambda p_0.$$

Consider state 1. From Figure S2–2, we see that we can enter state 1 in two ways: from state 0 or from state 2. Given that we are in state 0, we enter state 1 at rate λ, and given that we are in state 2, we enter state 1 at rate μ. It follows that the rate at which we enter state 1 is $\lambda p_0 + \mu p_2$. We can leave state 1 by going either to state 0 if an arrival occurs or state 2 if a service completion occurs. Hence, the rate at which we leave state 1 is $\lambda p_1 + \mu p_1 = (\lambda + \mu)p_1$. It follows that the second balance equation is

$$\mu p_2 + \lambda p_0 = (\lambda + \mu)p_1.$$

The form of the remaining balance equations is essentially the same as that of the second balance equation. In general,

$$\mu p_{i+1} + \lambda p_{i-1} = (\lambda + \mu)p_i \qquad \text{for } 1 \leq i \leq \infty.$$

These equations, along with one other condition, allow us to obtain an explicit solution for the steady-state probabilities. The method of solution is to first express each p_i in terms of is p_0. From the first balance equation we have

$$p_1 = (\lambda/\mu)p_0.$$

The second balance equation gives

$$\begin{aligned}\mu p_2 &= (\lambda + \mu)p_1 - \lambda p_0 = (\lambda + \mu)(\lambda/\mu)\, p_0 - \lambda p_0 \\ &= (\lambda^2/\mu)\, p_0 + \lambda\, p_0 - \lambda\, p_0 = (\lambda^2/\mu)p_0.\end{aligned}$$

Dividing both sides by μ gives

$$p_2 = (\lambda/\mu)^2 p_0.$$

Similarly, we will find in general that

$$p_i = (\lambda/\mu)^i p_0.$$

The solution is obtained by using the condition that

$$\sum_{i=0}^{\infty} p_i = 1,$$

since $p_0, p_1, p_2, \ldots$ forms a probability distribution on the states of the system.

Substituting for each p_i, we have

$$\sum_{i=0}^{\infty} (\lambda/\mu)^i p_0 = 1.$$

Let $\rho = \lambda/\mu$ be the utilization rate. For a solution to exist, it must be true that $\rho < 1$. In that case

$$\sum_{i=0}^{\infty} \rho^i = 1/(1 - \rho),$$

known as the geometric series, from which we obtain

$$p_0 = (1 - \rho)$$

and

$$p_i = \rho^i(1 - \rho) \qquad \text{for } i = 1, 2, 3\ldots$$

as given in Section 7–4. (This formula is also valid when $i = 0$.)

Waiting Time Distribution

We now derive the distribution of the waiting time W for a random customer joining the queue in steady state. To derive this distribution, we condition on the number of customers in the system at steady state, n, and uncondition by multiplying by the probability p_n. Suppose that a customer joining the queue at a random point in time finds n customers already in the system. Then that customer must wait for n service completions before entering service himself. As W is the total time in the system, this means that in this case W will be the sum of $n + 1$ service completions. Let $S_1, S_2, \ldots$ be the times of the successive services. By assumption, these random variables are mutually independent and exponentially distributed with common mean μ. The time for $n + 1$ service completions is $S_1 + S_2, + \ldots + S_{n+1}$, which we know has the Erlang distribution with parameters μ and $n + 1$ (see Section 13.3).

That is,

$$P\{W > t \,|\, n \text{ in the system}\} = \sum_{k=0}^{n} \frac{e^{-\mu t}(\mu t)^k}{k!}.$$

We know from the previous subsection that the unconditional probability of n in the system in the steady state, p_n, has the geometric distribution. Substituting $\rho = \lambda/\mu$, we may write p_n in the form

$$p_n = \left(\frac{\mu - \lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n.$$

Unconditioning on p_n gives

$$\begin{aligned} P\{W > t\} &= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{e^{-\mu t}(\mu t)^k}{k!} \left(\frac{\mu - \lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \\ &= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{e^{-\mu t}(\mu t)^k}{k!} \left(\frac{\mu - \lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \\ &= \frac{\mu - \lambda}{\mu} e^{-\mu t} \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \sum_{n=k}^{\infty} \left(\frac{\lambda}{\mu}\right)^n. \end{aligned}$$

Using the fact that

$$\sum_{n=k}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = \left(\frac{\lambda}{\mu}\right)^k \frac{1}{1-\lambda/\mu},$$

and substituting this into the earlier equation gives, after simplifying,

$$P\{W > t\} = e^{-\mu t} \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} = e^{-\mu t} e^{+\lambda t} = e^{-(\mu-\lambda)t}.$$

The summation term equals $e^{+\lambda t}$ because it is the Taylor series expansion for e (and because Poisson probabilities sum to one). What we have shown is the surprising result that W has the exponential distribution with parameter $\mu - \lambda$. This implies that W has the memoryless property. That is, suppose that a customer has already been waiting for s units of time. The probability that he will have to wait at least an additional t units of time is the same as the probability that a newly joining customer waits at least t units of time. This result is not intuitive and is rather depressing for the poor customer, who has already spent a substantial amount of time waiting for service!

We will not present the derivation (it is similar to the one previously given), but state that the distribution of W_q is essentially exponential with the complementary cumulative distribution function

$$P\{W_q > t\} = \rho e^{-(\mu-\lambda)t} \qquad \text{for all } t \geq 0.$$

Note that the probability that the waiting time in the queue is zero (i.e., there is no delay) is positive. It is equal to the probability that the system is empty, p_0. That is

$$P\{W_q = 0\} = p_0 = 1 - \rho = 1 - p^d.$$

S2.3 FURTHER RESULTS FOR M/M QUEUES

This section gives further known results for M/M queues. In particular, we derive results for when transitions are state dependent, when there are multiple servers, and when there is a finite system capacity.

Consider first the case where *both* the arrival and the service rates depend on the state. Several versions of the M/M/1 model are special cases of this one. The transition diagram is the same as that pictured in Figure S2–2, except that both λ and μ are state dependent (see Figure S2–3). The balance equation principle applied to this system yields

$$\begin{aligned} \mu_1 p_1 &= \lambda_0 p_0, \\ \lambda_0 p_0 + \mu_2 p_2 &= (\lambda_1 + \mu_1) p_1, \\ \lambda_1 p_1 + \mu_3 p_3 &= (\lambda_2 + \mu_2) p_2, \end{aligned}$$

and so on.

Expressing each of the state probabilities in terms of p_0, as we did earlier, results in the following:

$$\begin{aligned} p_1 &= \frac{\lambda_0}{\mu_1} = p_0, \\ p_2 &= \frac{\lambda_0 \lambda_1}{\mu_2 \mu_1} p_0, \\ p_3 &= \frac{\lambda_0 \lambda_1 \lambda_2}{\mu_3 \mu_2 \mu_1} p_0, \end{aligned}$$

and so on.

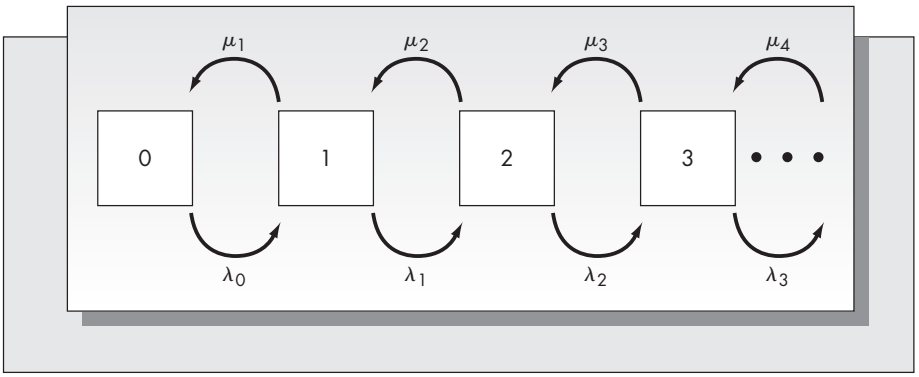


FIGURE S2–3
State changes for the M/M/1 queue with state-dependent service and arrival rates

Define

$$a_n = \frac{\lambda_{n-1}\lambda_{n-2} \ldots \lambda_0}{\mu_n\mu_{n-1} \ldots \mu_1}$$

so that

$$p_n = a_n p_0 \text{ for } n = 1, 2, 3 \ldots$$

Again using the fact that $p_0, p_1, \ldots$ is a probability distribution, we have that

$$\sum_{n=0}^{\infty} p_n = 1.$$

This translates to the defining condition for p_0 as

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} a_n}.$$

The various measure of service can be obtained by applying their definitions. In particular, l, the expected number in the system, is given by

$$l = \sum_{n=0}^{\infty} n p_n.$$

If there are assumed to be s servers then l_q, the expected number in the queue, is given by

$$l_q = \sum_{n=s}^{\infty} (n-s) p_n.$$

Little's law still applies and can be used to find the expected waiting times given the expected number in the system and the expected number in the queue. However, to apply Little's law when the arrival rate is state dependent, we must determine the *overall* average expected arrival rate, or the effective arrival rate, which we call λ_{eff}. Because the arrival rate is λ_n, when the system is in state n, it follows that the effective arrival rate is

$$\lambda_{eff} = \sum_{n=0}^{\infty} \lambda_n p_n.$$

We can use these general results for a variety of configurations of the queue with exponential interarrival times and exponential service times. In particular, they can be used to find results for the M/M/s queue and queues with finite capacity, as shown next.

The M/M/s Queue

Consider the M/M/s queue; that is, the case in which there are s servers in parallel. This case is pictured in Figure 7–7 from Section 7.5. In order to apply the results from the state-dependent model we need to establish the following result.

Suppose that m servers are busy at a random point in time and also suppose that the distribution of each of the servers is exponential with rate μ. The question is, what is the distribution of the time until the next service completion? Let T_1, T_2, …,T_m be the service times of the customers who are currently in service. By assumption, these are independent exponential random variables. Furthermore, if t is a random point in time, the remaining time in service from t to the end of the service completion for each of the customers is also exponential with the same distribution. (This is a consequence of the properties of the exponential distribution discussed earlier.)

It follows that the time until the next service completion, say T, is distributed as the *minimum* of $T_1, T_2, \ldots, T_m$. The result we need to analyze this case is

> **Result:** Let T_1, T_2, …,T_m be independent exponential random variables with common exponential distribution with rate μ, and define $T = \min(T_1, T_2, \ldots, T_m)$. Then T is also exponentially distributed with rate $m\mu$.

(This result is proven in Chapter 13, in the context of series systems of components subject to exponential failure.)

Why is this result important? It means that the distribution between customer departures is still exponential and that the methods just derived for state-dependent queues still apply. If there are s servers, then it follows that the transition rate diagram is as pictured in Figure S2–4.

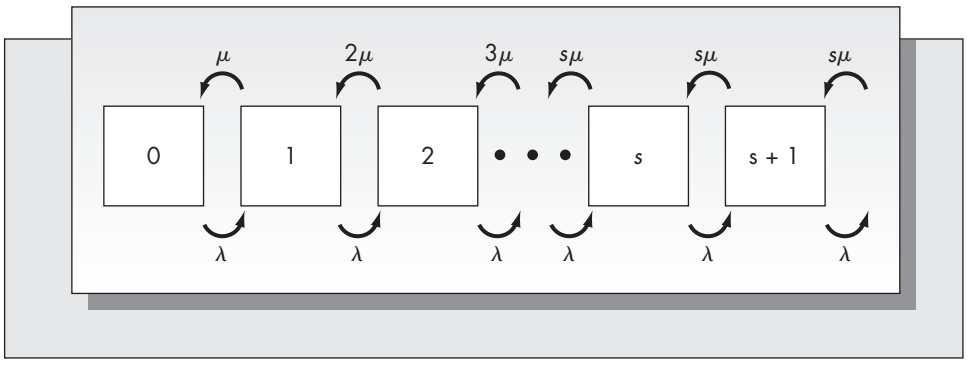


FIGURE S2–4
Transition rate diagram when there are s servers in parallel

Comparing Figures S2–3 and S2–4, we see that

$$\begin{aligned}\mu_1 &= \mu\\ \mu_2 &= 2\mu\\ &\vdots\\ \mu_s &= s\mu\\ \mu_{s+1} &= s\mu\\ \mu_{s+2} &= s\mu\end{aligned}$$

and $\lambda_i = \lambda$ for all $I = 0, 1, 2, \ldots$

Substituting, it follows that

$$\begin{aligned}p_1 &= \frac{\lambda}{\mu}p_0,\\ p_2 &= \frac{1}{2}\left(\frac{\lambda}{\mu}\right)^2 p_0,\\ p_3 &= \frac{1}{(3)(2)}\left(\frac{\lambda}{\mu}\right)^3 p_0,\\ &\vdots\\ p_n &= \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0 \quad \text{for } 0 \le n \le s.\end{aligned}$$

When $n > s$, we obtain

$$\frac{1}{s!s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n p_0, \quad \text{for } n > s.$$

Substituting these state probabilities gives the following for p_0:

$$p_0 = \left\{\sum_{n=0}^{s-1}\frac{1}{n!}\frac{\lambda^n}{\mu} + \sum_{n=q}^{\infty}\frac{1}{s!s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n\right\}^{-1}.$$

This can be simplified by noting that the second term is a geometric series. Letting $\rho = \lambda/s\mu$, one can show after a bit of algebraic manipulation that

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

When computing the standard performance measures, it turns out that l_q has the simplest form. Again, we will not present the details, but only the results. The derivations are similar to those presented in earlier sections.

$$\begin{aligned}l_q &= \frac{s^s\rho^{s+1}}{s!(1-\rho)^2}p_0.\\ l &= l_q + s\rho.\\ w_q &= l_q/\lambda.\\ w &= w_q + 1/\mu.\end{aligned}$$

As with the single-server queue, the condition that $\rho < 1$ is required to guarantee that the queue does not grow without bound.

Example *S2.1*

Tony's Barbershop is run, owned, and operated by Anthony Jones, who has been cutting hair for more than 20 years. Anthony does not take appointments, so the arrival pattern of customers is essentially random. Traditionally, the arrival rate had been about one customer every 50 minutes. Two months ago, the local paper ran an article about Anthony that improved business substantially. Currently, the arrival rate is closer to one customer every 35 minutes. Haircuts require an average of 25 minutes, but the times vary considerably depending on customer needs. A trim might require as little as 5 minutes, but a shampoo and styling could take as long as an hour or more. For this reason, the exponential distribution seems to provide a reasonably good fit of the service time distribution.

Anthony's customers have always been patient, but ever since business picked up, some have complained that the wait is too long. Anthony is considering taking his cousin Marvin into the business to improve customer service. Assume that Marvin cuts hair at the same rate as Anthony.

a. How much has the quality of service declined since more customers have started using the shop?

b. How much improvement in the performance of the system are customers likely to see with an additional barber in the shop?

Solution

a. First, we will determine the various performance measures for the system prior to the appearance of the newspaper article. The average time between arrivals was one every 50 minutes, which gives an arrival rate of

$$\lambda = 60/50 = 1.2 \text{ arrivals per hour.}$$

Each haircut requires an average of 25 minutes, which translates to a service rate of

$$\mu = 60/25 = 2.4 \text{ haircuts per hour.}$$

It follows that $\rho = \lambda/\mu = 1.2/2.4 = 0.5$. (That is, Tony was busy half the time.) The values of the performance measures are

$$L = \rho/(1 - \rho) = 0.5/0.5 = 1.$$
$$L_q = \rho L = 0.5.$$
$$W = L/\lambda = 1/1.2 = 0.8333 \text{ hour.}$$
$$W_q = L_q/\lambda = 0.5/1.2 = 0.4167 \text{ hour.}$$

This means that originally customers waited (0.4167)(60) = 25 minutes for a haircut on average.

After the article appeared, the arrival rate increased to one customer every 35 minutes. This means that λ became 60/35 = 1.7143 and $\rho = 0.7143$. The performance measures are now

$$L = 0.7143/(1 - 0.7143) = 2.5.$$
$$L_q = \rho L = (0.7143)(2.5) = 1.7857.$$
$$W = L/\lambda = 2.5/1.7143 = 1.458 \text{ hours.}$$
$$W_q = L_q/\lambda = 1.78/1.7143 = 1.0383 \text{ hours.}$$

The customers clearly have a valid gripe. A customer has to wait an average of more than an hour before getting a haircut. In fact, because the distribution of W_q is exponential, many would have to wait quite a bit longer than this.

b. Adding an additional barber improves the system performance dramatically. With two barbers, we have

$$\rho = \lambda/(s\mu) = 1.7143/(2)(2.4) = 0.3571 \qquad (s\rho = 0.7143).$$

$$p_0 = \left\{1 + 0.7143 + \frac{(0.7143)^2}{2!}\ \frac{1}{1 - 0.3571}\right\}^{-1}$$

$$= (2.111)^{-1} = 0.4737.$$

It follows that

$$l_q = \frac{(2)^2(0.3571)^{s+1}}{2!(1 - 0.3571)^2}\,(0.4737) = 0.0522.$$

$$l = l_q + s\rho = 0.0522 + 0.7143 = 0.7665.$$

$$w_q = l_q/\lambda = 0.0522/1.7143 = 0.3004 \text{ hour (1.82 minutes).}$$

$$w = w_q + 1/\mu = 0.3004 + 0.4167 = 0.4471 \text{ hour (about 27 minutes).}$$

With only a single barber customers could expect to wait more than an hour for a haircut. With the addition of another barber, time is reduced to less than 2 minutes on average.

The M/M/1 Queue with a Finite Capacity

Another special version of the general M/M/1 queue with state-dependent service and arrival rates is the case in which there is a finite waiting area. If arrivals occur when the waiting area is full, they are turned away. Problems of this type are common in service systems such as restaurants, movie theaters, and concert halls. They can also occur in manufacturing systems in which buffers between work centers have a finite capacity. This is the case, for example, with JIT systems. (See the discussion of JIT in Chapter 8.)

Suppose that the maximum number of customers permitted in the system is K. The transition rate diagram for this case is exactly the same as that pictured in Figure S2–2 except that the transitions do not occur beyond state K. Because the transition rate diagram is the same up to state K, the balance equations will yield the same relationship between p_n and p_0 for $n = 1, 2, \ldots, K$. That is,

$$p_n = \rho^n p_0 \qquad \text{for } n = 1, 2, 3, \ldots, K.$$

Then p_0 is found from

$$\sum_{n=0}^{K} p_n = 1,$$

which gives

$$p_0 = \left(\sum_{n=0}^{K} \rho^n\right)^{-1}.$$

An explicit expression for the finite geometric series is obtained in the following way:

$$\sum_{n=0}^{K} \rho^n = \sum_{n=0}^{\infty} \rho^n - \sum_{n=k+1}^{\infty} \rho^n = \frac{1}{1-\rho} - \frac{\rho^K}{1-\rho}$$
$$= \frac{1-\rho^K}{1-\rho}.$$

It follows that

$$p_0 = \frac{1-\rho}{1-\rho^{K+1}},$$

from which we obtain

$$p_n = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}} \qquad \text{for } n = 0, 1, 2, \ldots, K.$$

For the case of the finite waiting room, it is not necessary that $\rho < 1$. In fact, P_n has this value for *all* values of $\rho \neq 1$. When $\rho = 1$, it turns out that all states are equally likely, so that

$$p_n = 1/(K+1) \qquad \text{for } 0 \leq n \leq K \qquad \text{(when } \rho = 1 \text{ only)}.$$

Little's law still applies, but we must use a modified value for the arrival rate because not all arriving customers are permitted to enter the system. When there are K or more in the system, the arrival rate is zero, so the overall arrival rate is less than λ. The effective arrival rate, λ_{eff}, is computed as follows:

$$\begin{aligned}\lambda_{eff} &= \lambda P\{\text{Number in the system} < K\} + 0P\{\text{Number in the system} = K\} \\ &= \lambda(1 - P\{\text{number in the system} = K\}) \\ &= \lambda(1 - p_K).\end{aligned}$$

The measures of performance are obtained from l, the expected number in the system in steady state, which is found from

$$l = \sum_{n=0}^{K} n p_n$$
$$= \sum_{n=0}^{K} \frac{1-\rho}{1-\rho^{K+1}} n\rho^n.$$

The calculation proceeds by noting that

$$\sum_{n=0}^{K} n\rho^{n-1} = \frac{d}{d\rho} \sum_{n=0}^{K} \rho^n.$$

Using the earlier expression for the finite geometric sum, we eventually obtain.

$$l = \frac{\rho}{1-\rho} - \frac{(\mathrm{K}+1)\rho^{K+1}}{1-\rho^{K+1}}.$$

The remaining measures of performance are found from

$$l_q = l - (1 - p_0).$$
$$w = l/\lambda_{eff}.$$
$$w_q = l_q/\lambda_{eff}.$$

Similar formulas can be derived for the case of a finite capacity queue and multiple parallel servers. (See, for example, Hillier and Lieberman, 1990).

Example *S2.2*

A popular attraction at the New Jersey Shore is a street artist who will paint a caricature in about five minutes. However, because the times required for each drawing vary considerably, they are accurately described by an exponential distribution. People are willing to wait their turn, but when there are more than 10 waiting for a picture, customers are turned away and asked to return at a later time. At peak times one can expect as many as 20 customers per hour. Assume that customers arrive completely at random at the peak arrival rate.

a. What proportion of the time is the queue at maximum capacity?

b. How many customers are being turned away on average? Determine the measures of performance for this queueing system.

c. If the waiting area were doubled in size, how would that affect your answers to parts (a) and (b)?

Solution

The arrival rate is $\lambda = 20$ per hour and the service rate is $\mu = 12$ per hour, so that $\rho = 20/12 = 1.667$. The maximum number in the system is $K = 11$ (10 in the queue plus the customer being served).

a. The probability that the system is full is p_K, which is given by

$$p_K = \frac{(1 - \rho)\rho^K}{1 - \rho^{K+1}} = \frac{(1 - 1.667)(1.667)^{11}}{1 - 1.667^{12}} = \frac{-184.25}{-459.5} = 0.40.$$

b. As the arrival rate is 20 per hour and 40 percent of the time the system is full, there are, during peak periods, (20)(0.40) = 8 customers per hour being turned away. This gives λ_{eff} = 12 per hour.

$$\begin{aligned} l &= \frac{\rho}{1 - \rho} - \frac{(K + 1)\rho^{K+1}}{1 - \rho^{K+1}} = \frac{1.667}{1 - 1.667} - \frac{(12)(1.667)^{12}}{1 - 1.667^{12}} \\ &= -2.5 - (-12.03) \\ &= 9.53. \end{aligned}$$

We need to determine p_0 to compute l_q.

$$p_0 = \frac{(1 - \rho)}{1 - \rho^{K+1}} = \frac{1 - 1.667}{1 - 1.667^{12}} = 0.00145,$$

which gives $l_q = I - (1 - p_0) = 9.53 - (1-0.00145) = 8.53$.

Notice that the small value of p_0 means that the system is rarely empty. In particular, the artist is idle only 0.145 percent of the time!

We showed earlier that $\lambda_{eff} = 12$, so that

$$W = l/\lambda_{eff} = 9.53/12 = 0.7942 \text{ hour (about 48 minutes).}$$

$$W_q = l_q/\lambda_{eff} = 8.53/12 = 0.7108 \text{ hour (about 43 minutes).}$$

c. If the waiting area were doubled in size, then $K = 21$. In that case ρ_K is given by

$$p_K = \frac{(1 - 1.667)(1.667)^{21}}{1 - 1.667^{22}} = \frac{-30{,}533.28}{-76{,}309.3} = 0.40.$$

Interestingly, doubling the capacity of the queue makes no difference relative to the probability that the system is full. The reason is that because the arrival rate exceeds the service rate, the system reaches capacity quickly in either case. In both cases the effective arrival rate λ_{eff} is approximately equal to the service rate (although it will always be true that $\lambda_{eff} < \mu$). Even for much larger values of K, p_k is 0.4 in this example.

S2.4 INFINITE SERVER RESULTS

This section considers results for queues with an infinite number of servers. It gives exact results for the M/G/∞ queue and also limiting results as the number of servers, s, approaches ∞. These limiting results provide an approximation for the probability of delay that is typically more accurate (although a little more complicated) than the Sakasegawa (1977) formula given in Section 7.3.

The M/G/∞ queue

Another version of the queueing problem with general service distribution for which there are explicit results is the case in which there are an infinite number of servers. Customers arrive at the system completely at random according to a Poisson process with rate λ. The service time distribution is arbitrary with service rate μ. At the instant of arrival, the customer enters service. An infinite number of servers means that there is always a server available, no matter how many customers are in the system. Although this might seem unrealistic, many real problems can be modeled in this way. Because there is no queue of customers waiting for service, there is no waiting time for service. Hence, both measures of performance l_q and w_q, are zero. However, the number of customers in the system, l, is not zero. Note that the number of customers in the system is equal to the number of busy servers. The result of interest is the following:

Result: For the M/G/∞ queue with arrival rate λ and service rate μ, the distribution of the number of customers in the system (or the number of busy servers) in steady state is Poisson with rate λ/μ. That is

$$P\{L = k\} = \frac{e^{-(\lambda/\mu)}(\lambda/\mu)^k}{k!} \qquad \text{for } k = 0, 1, 2, \ldots$$

This is a powerful result. It follows that both the mean and the variance of the number of customers in the system in steady state is λ/μ.

Example *S2.3* A common inventory control policy for high-value items is the one-for-one policy, also known as the $(S - 1, S)$ policy. This means that the target stock is S and at each occurrence of a demand, a reorder for one unit is placed. Suppose that demands are generated by a stationary Poisson process with rate λ and that the lead time required for replenishment is a random variable with arbitrary distribution with mean $1/\mu$. This problem is exactly an M/G/∞ queue. The number of customers in the system is equivalent to the number of outstanding orders, which by the earlier result has the Poisson distribution with mean λ/μ. If the lead time is fixed at τ, the expected number of units on order is just $\lambda\tau$. It follows that the expected number of units in stock is $S - \lambda\tau$. This result can be used to determine an expression for the expected holding, stock-out, and replenishment costs, which can then be optimized with a choice of S. (See Hadley and Whitin, 1963, p. 212, for example.)

Example *S2.4* Personnel planning is an important function for many firms. Consider a company's department with a desired headcount of 100 positions. Suppose that employees leave their positions at a rate of 3.4 per month, and that it requires an average of 4 months for the firm to fill open positions. Analysis of past data shows that the number of employees leaving the firm per month has the Poisson distribution, and that the time required to fill positions follows Weibull distribution. What is the probability that there are more than 15 positions unfilled at any point in time? How many jobs within the department are filled on average? How many positions should the firm have in order for the head count of working employees to be 100 on average?

Solution To determine the distribution of filled positions, we model the problem as an M/G/∞ queue. Each time that an employee leaves, his or her position enters the queue of unfilled positions. Assuming that the search for a replacement starts immediately, the correct model is an infinite number of servers. According to the theory, the expected number of unfilled positions is independent of the time required to replace each employee. The expected number of unfilled positions is λ/μ. For this application, λ corresponds to the rate at which employees leave their positions, which is 3.4 per month, and μ the rate that jobs are filled, which is 1/4 per month. Thus, the mean number of unfilled jobs is $\lambda/\mu = 3.4/(1/4) = (3.4)(4) = 13.6$. Hence, there are $100 - 13.6 = 86.4$ positions filled on average.

The probability that there are more than 15 unfilled positions is the probability that a Poisson random variable with mean 13.6 exceeds 15. Interpolating from Table A–3, this probability is approximately 0.29.

It also follows that if the department were allotted 114 positions rather than 100, there would be, on average, 100 positions filled at any point in time (although the actual number is a random variable).[2]

Infinite Server Limits

Useful queueing approximations can be developed by taking the limit as the number of servers tends towards infinity. However, as the number of servers is increased we must also scale either the arrival rate or service rate; otherwise delays would go to zero. Turning this around, we must decide the rate at which the number of servers are increased, relative to the arrival rate scaling up. As mentioned in Section 7.3, one useful scenario is where staffing is matched to arrival rate growth so that the probability of delay p^d tends to a value strictly between 0 and 1. This occurs if the scaling in the number of servers, s, occurs such that

$$\sqrt{s}(1 - \rho_s) \to \beta \text{ for } 0 < \beta < 1,$$

where ρ_s is the utilization when there are s servers. This is the so-called quality-and-efficiency (QED) driven regime, or Halfin-Whitt scaling. This scaling results in the following approximation for the probability of delay p^d, due to Whitt (2004). Define

$$\beta = (1 - \rho)\sqrt{s},$$

and a measure of peakedness

$$z = 1 + (c_a^2 - 1)(1 - (c_s^2/2)) \text{ for } 0 \le c_s^2 \le 1.$$

Note that the above equation for z is only valid for $c_s^2 \le 1$ and a more accurate (but more complicated) estimate of this value, which is not limited by the range of c_s^2, may be found as equation (1.6) in Whitt (2004). The estimate for probability of delay is then given by

$$p^d = \frac{1}{1 + \beta\Phi(\beta/\sqrt{z})/(\sqrt{z}\phi(\beta/\sqrt{z}))}.$$

[2] I am grateful to John Peterson of Smith-Kline-Beecham for bringing this application to my attention.

where $\Phi(.)$ is the cumulative standard normal distribution and $\phi(.)$ is the standard normal density (see Section 5.1). Numerical tests show this approximation to be reasonably accurate across a wide range of G/G/s systems. It can also be extended to the case of a finite capacity waiting room (see Whitt, 2004).

Example *S2.5* In Example 7–4, suppose that customer arrivals are Poisson and service is deterministic. What is the probability of delay using Whitt's estimate? How does it compare to the Sakasegawa approximation? Now repeat this for Example 7–6.

Solution For Example 7–4, given the above description, $c_a^2 = 1$ and $c_s^2 = 0$. Therefore, $z = 1$. Further, $\beta = (1 - \rho)\sqrt{s} = (1 - 0.9868)\sqrt{48} = 0.09145$. Therefore,

$$p^d = \frac{1}{1 + 0.09145\Phi(0.09145)/\phi(0.09145)} = 0.8901.$$

Under the Sakasegawa formula, $p^d = 0.889$, so the two values are very close. For Example 7–6, $c_a^2 = 1$ and $c_s^2 = 0.7785$, which imply $z = 1$. Further, $\beta = (1 - \rho)\sqrt{s} = (1 - 0.94444)\sqrt{3} = 0.09623$. Therefore,

$$p^d = \frac{1}{1 + 0.09623\Phi(0.09623)/\phi(0.09623)} = 0.8846.$$

Under the Sakasegawa formula, $p^d = 0.9008$, which is 1.8 percent larger. It is encouraging that these probabilities are not entirely different.

S2.5 QUEUEING NETWORKS

As previously discussed, queueing networks are difficult to analyze, particularly exactly, and therefore simulation is often used to obtain an accurate estimation of system statistics. However, there are some notable exceptions to this statement that are discussed below.

An interesting result for M/M/s queues is that the departure stream of customers from the system forms a Poisson process (of rate λ if no customers are created or lost by the server). This means that if another station is downstream from this queue then it receives a stream of Poisson customers arriving, which will result in it being able to be analyzed exactly.

A **Jackson network** is a network of J M/M/s queueing nodes. Each node i receives a Poisson stream of external arrivals at rate a_i, $1 \le i \le J$. Jobs that are completed at node i are routed to node j with probability p_{ij} and out of the system with probability $1 - \sum_{j=1}^{J} p_{ij}$. The flow balance equations for the aggregate arrival rates, λ_i, are as follows

$$\lambda_i = \alpha_i + \sum_{j=1}^{J} p_{ji}\lambda_j \text{ for } 1 \le i \le J.$$

This implies that each node i forms its own M/M/s queueing system with arrival rate λ_i. Jackson (1957) showed that the distribution of customers in the system is simply the product of the probabilities for each queueing node. That is, the queues act as if they are independent of each other, even though they are clearly not.

In general, the departure stream from general queueing models is not Poisson, which of course means that there are typically not exact results for performance of downstream stations. However, an approximation technique was developed by Whitt (1983), called the **queueing network analyzer** (QNA), for quite general networks of G/G/s queues. It assumes an open network (where customers eventually leave the system), no capacity constraints, FCFS service, but that customers can be created or destroyed at stations, and the routing can be quite general.

One of the useful approximations in the QNA is an expression for the variability of departures from a G/G/1 queue. This variability is given by

$$\rho^2 c_s^2 + (1 - \rho^2) c_a^2.$$

Notice how if $\rho = 1$ then this is equal to c_s^2, the variability of the service process; whereas, if $\rho = 0$ then it equals c_a^2, the variability of the arrival process. This makes intuitive sense because if $\rho = 1$ then the server is consistently busy and the customers flowing out look like the service process, whereas if $\rho = 0$ (and there are arrivals) then the service time must be negligible and arrivals are just passed straight through. This expression can then be used as the variability of arrivals to a downstream station if the queues are in series.

S2.6 OPTIMIZATION OF QUEUEING SYSTEMS

Classical queueing analysis is descriptive rather than prescriptive. In practice, this means that given the various input and service distributions, one determines the measures of performance. These measures of performance do not directly translate to optimal decisions concerning the design of the system. This section shows how one would go about developing models for determining the optimal configuration of a queueing system.[3]

Let us consider some typical design problems arising in queueing service systems and how one would go about using the results of queueing theory to determine optimal system configurations.

Typical Service System Design Problems

1. The State Highway Board must determine the number of tollbooths to have available on a new interstate toll road. The more tollbooths open at any point in time, the less wait commuters will experience. However, additional tollbooths require an additional one-time cost to build and additional ongoing costs of salary for the toll taker.
2. A plant is being built by a major manufacturer of solid-state (memory) drives. The company management is considering several options for the manufacturing equipment. A new machine for the drives has double the throughput of the conventional equipment, but at more than triple the cost. Is the investment justified?
3. A translation service is considering how large of a client base to develop. The company wishes to have a large enough number of clients to make it busy, but not so many that it cannot provide reasonable turnaround times.

[3] The results of this section are based on Chapter 17 of Hillier and Lieberman (1990).

Modeling Framework

1. Consider the example of the state highway board. The more time commuters spend on the highway, the less time they spend working and contributing to society. If we view the goal as societal optimization, then there is clearly a direct economic benefit to reducing commute time. Suppose that an economic analysis of the highway problem resulted in an estimate of the cost incurred when a commuter spends w units of time in the system as the function $h(w)$. A typical case is pictured in Figure S2–5.

Let W be the time in system of a customer chosen at random. Then W is a random variable. For the M/M/1 queue, we showed that W has the exponential distribution with parameter $\mu - \lambda$. Given the distribution of W, it follows that the expected waiting cost of a customer chosen at random is

$$E(h(W)) = \int_0^\infty h(w)(\mu - \lambda)e^{-(\mu-\lambda)w}\, dw.$$

Because the arrival rate of customers is λ units per unit time, the overall waiting cost per unit time is $\lambda E(h(W))$. In the case of the tollbooths, one would determine the distribution of W for each number of servers being considered, say W_s. If the cost per unit time for maintaining each server is c, then the objective would be to find the optimal value of s to minimize

$$sc + \lambda E(W_s).$$

2. Consider the example of the firm producing solid-state drives. The firm can purchase a larger μ (service rate), but only for an increased cost. To determine the best decision in this case, the firm would have to be able to quantify the costs associated with various levels of service. Suppose that the annual cost of the manufacturing operation when the throughput rate of the process is μ is given by the function $f(\mu)$. Because the cost decreases as μ increases, this would be a monotonically decreasing

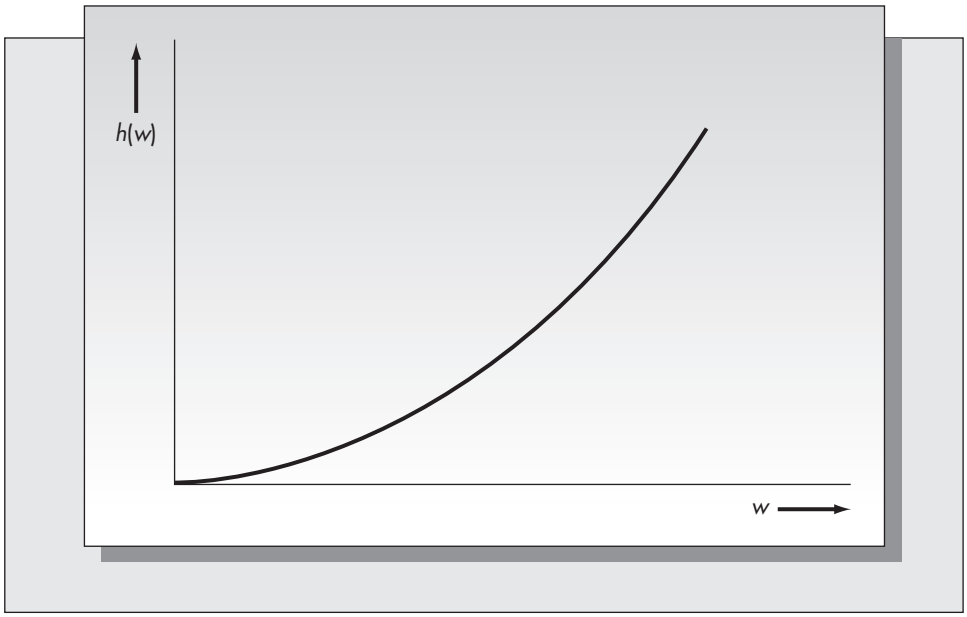


FIGURE S2–5
A typical waiting cost function

function of μ. Furthermore, suppose that the one-time cost of purchasing equipment with service rate μ is $C(\mu)$. We would expect that $C(\mu)$ would be a monotonically increasing function of μ. Let I be the annual interest rate of alternative investments. Then the total annual cost is

$$IC(\mu) + f(\mu).$$

This function will be convex in μ (see Figure S2–6), so an optimal minimizing μ will exist and can be found easily. When there are only several possible values of μ, the objective function can be evaluated at these values and the choice yielding the lowest cost can be made.

3. Consider the example of the translation service. In this case, the decision variable is the arrival rate λ. The larger the client base, the more jobs the firm will receive and the larger the value of λ. There are several possible formulations of this problem. One would be to determine the value of the expected number in the system that allows the firm to meet its obligations. In that case, we would assume that s and μ are given and that the objective is to determine λ so that l is equal to a target value. Another approach would be to find λ so that the probability that the number of customers in the system does not exceed a target level is at least some specified probability (such as .95).

To illustrate this case, consider the following specific example. Mary Worth runs her own translation service. She requires an average of 1.2 hours to complete a job, but the size of jobs varies considerably and can be described by an exponential distribution. Furthermore, the pattern of arrivals of jobs appears to be completely random. Mary works eight-hour days and does not want to have more than two days of work piled up at any point in time. Suppose that she wants the likelihood of this occurring to be no more than 5 percent. This means that the number of customers in the queue

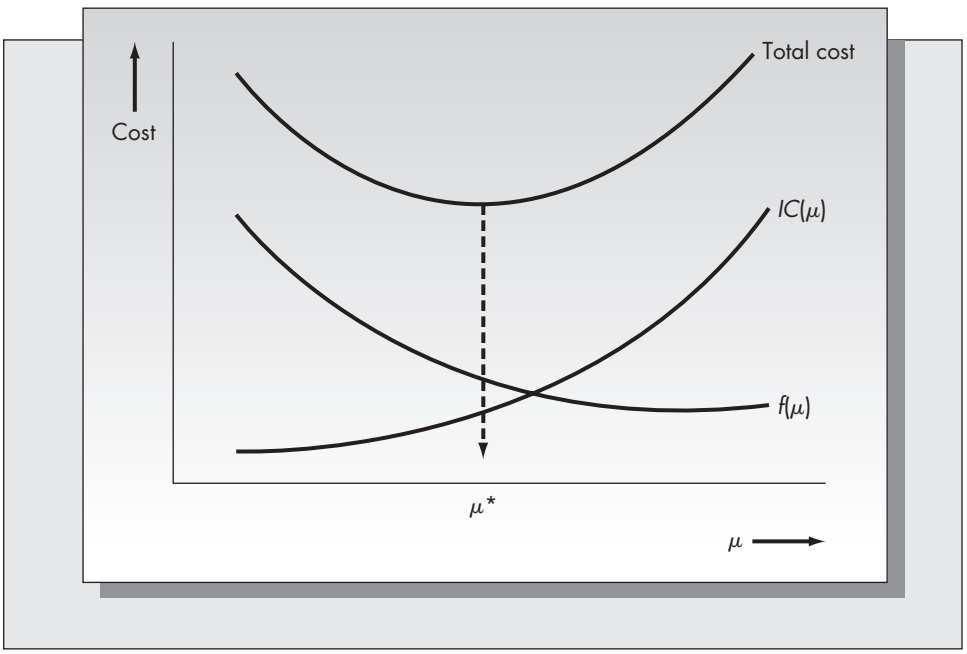


FIGURE S2–6
Optimizing the service rate μ

should not exceed 16/1.2 = 13.33 5 percent of the time. Hence, she wishes to have an arrival rate of jobs so that

$$P\{L > 13\} \leq 0.05.$$

We showed in the solution to Example 7–5 that

$$P\{L > k\} = \rho^{k+1}.$$

Hence, we wish to find λ to solve

$$(\lambda/\mu)^{13} = 0.05.$$

Using $\mu = 1/1.2 = 0.8333$ per hour, and taking natural logarithms of both sides, we obtain

$$\begin{aligned} 13\ \ln(\lambda/0.8333) &= \ln(0.05) \\ \ln(\lambda/0.8333) &= -0.23044 \\ \lambda/0.8333 &= \exp(-0.23044) = 0.7942, \end{aligned}$$

giving $\lambda = 0.662$.

The solution, then, is that she should plan to have enough clients to generate about 0.662 jobs each working hour, or about 5.3 jobs per day.

Assigning service territories to repairmen, or deciding which region a blood bank is to cover, are other examples of problems in which the objective is to find the optimal value of the arrival rate, λ.

Bibliography

Feller, W. *An Introduction to Probability Theory and Its Applications*. Vol. 2. New York: John Wiley & Sons, 1966.

Hadley, G., and T. M. Whitin. *Analysis of Inventory Systems*. Englewood Cliffs, NJ: Prentice Hall, 1963.

Hillier, F.S., and G.J. Lieberman (1990). *Introduction to Operations Research.* 5th ed. New York: McGraw-Hill, 1990.

Jackson, J. R. "Networks of Waiting Lines." *Operations Research* 5 (1957), pp. 518–521.

Kleinrock, L. *Queueing Systems. Vol.* 1*, Theory.* New York: Wiley Interscience, 1975.

Whitt, W. "A Diffusion Approximation for the G/GI/n/m Queue." *Operations Research* 52 (2004), pp. 922–941.

Whitt, W. "The Queueing Network Analyzer." *Bell System Technical Journal* 62 (1983), pp. 2779–2815.

Chapter **Eight**

Push and Pull Production Control Systems: MRP and JIT

"The most dangerous kind of waste is the waste we do not recognize."
—Shigeo Shingo

Chapter Overview

Purpose

To understand the push and pull philosophies in production planning and compare MRP and JIT methods for scheduling the flow of goods in a factory.

Key Points

1. *Push versus pull.* There are two fundamental philosophies for moving material through the factory. A push system is one in which production planning is done for all levels in advance. Once production is completed, units are pushed to the next level. A pull system is one in which items are moved from one level to the next only when requested. *Materials requirements planning* (MRP) is the basic push system. Based on forecasts for end items over a specified planning horizon, the MRP planning system determines production quantities for each level of the system. It relies on the so-called explosion calculus, which requires knowledge of the gozinto factor (i.e., how many of part A are required for part B), and production lead times. The earliest of the pull systems is *kanban* developed by Toyota, which has exploded into the *just-in-time* (JIT) and lean production movements. Here the fundamental goal is to reduce work-in-process to a bare minimum. To do so, items are only moved when requested by the next higher level in the production process. Each of the methods has particular advantages and disadvantages.
2. *MRP basics.* The MRP explosion calculus is a set of rules for converting a *master production schedule* (MPS) to a build schedule for all the components comprising the end product. The MPS is a production plan for the end item or final product by period. It is derived from the forecasts of demand adjusted for returns, on-hand inventory, and the like. At each stage in the process, one computes the production amounts required at each level of the production process by doing two basic operations: (1) offsetting the time when production begins by the lead time required at the current level and (2) multiplying the higher-level requirement by the gozinto factor. The simplest production schedule at each level is lot-for-lot

(L4L), which means one produces the number of units required each period. However, if one knows the holding and setup cost for production, it is possible to construct a more cost efficient lot-sizing plan. Three heuristics we consider are (1) EOQ lot sizing, (2) the Silver–Meal heuristic, and (3) the least unit cost heuristic. Optimal lot sizing requires dynamic programming and is discussed in Appendix 8–A. We also consider lot sizing when capacity constraints are explicitly accounted for. This problem is difficult to solve optimally, but can be approximated efficiently.

MRP as a planning system has advantages and disadvantages over other planning systems. Some of the disadvantages include (1) forecast uncertainty is ignored; (2) capacity constraints are largely ignored; (3) the choice of the planning horizon can have a significant effect on the recommended lot sizes; (4) lead times are assumed fixed, but they should depend on the lot sizes; (5) MRP ignores the losses due to defectives or machine downtime; (6) data integrity can be a serious problem; and (7) in systems where components are used in multiple products, it is necessary to peg each order to a specific higher-level item.

3. *JIT basics.* The JIT philosophy grew out of the kanban system developed by Toyota. Kanban is the Japanese word for card or ticket. Kanban controls the flow of goods in the plant by using a variety of different kinds of cards. Each card is attached to a palette of goods. Production cannot commence until production ordering kanbans are available. This guarantees that production at one level will not begin unless there is demand at the next level. This prevents work-in-process inventories from building up between work centers when a problem arises anywhere in the system. Part of what made kanban so successful at Toyota was the development of single minute exchange of dies (SMED), which reduced changeover times for certain operations from several hours to several minutes. Kanban is not the only way to implement a JIT system. Information flows can be controlled more efficiently with a central information processor than with cards.
4. *Comparison of JIT and MRP.* JIT has several advantages and several disadvantages when compared with MRP as a production planning system. Some of the advantages of JIT include (1) reduce work-in-process inventories, thus decreasing inventory costs and waste, (2) easy to quickly identify quality problems before large inventories of defective parts build up, and (3) when coordinated with a JIT purchasing program, ensures the smooth flow of materials throughout the entire production process. Advantages of MRP include (1) the ability to react to changes in demand, since demand forecasts are an integral part of the system (as opposed to JIT which does no look-ahead planning); (2) allowance for lot sizing at the various levels of the system, thus affording the opportunity to reduce setups and setup costs; and (3) planning of production levels at all levels of the firm for several periods into the future, thus affording the firm the opportunity to look ahead to better schedule shifts and adjust workforce levels in the face of changing demand.

The supply chain is the set of all activities that convert raw materials to the final product. One of the key activities in the supply chain is the actual production process. How well things are managed in the factory plays a fundamental role in the reliability and quality of the final product. There are two fundamentally different philosophies for managing the flow of goods in the factory. As we will see in this chapter, the methods developed in Chapters 4 and 5 for managing inventories are not always appropriate in the factory context.

The two approaches we consider are *materials requirements planning (MRP)* and *just-in-time (JIT)*. These are often referred to respectively as "push" and "pull" control systems. To appreciate exactly what distinguishes push and pull systems will require an understanding of exactly how these methods work, which will be covered in detail in this chapter. The simplest definition that this writer has seen (due to Karmarkar, 1989) is that "a pull system initiates production as a reaction to present demand, while a push system initiates production in anticipation of future demand." Thus, MRP incorporates forecasts of future demand while JIT does not.

To better understand the difference between MRP and JIT, consider the following simple example. Garden spades are produced by a plant in Muncie, Indiana. Each spade consists of two parts: the metal digger and the wooden handle. The parts are connected by two screws. The plant produces spades at an average rate of 100 per week. The metal digger is produced in batches of 400 on the first two days of each month, and the handles are ordered from an outside supplier. The assembly of spades takes place during the first week of each month.

Consider now the demand pattern for the screws. Exactly 800 screws are needed during the first week of each month. Assuming four weeks per month, the weekly demand pattern for the screws is 800, 0, 0, 0, 800, 0, 0, 0, 800, 0, 0, 0, and so on. Using a weekly demand rate of 200 and appropriate holding and setup costs, suppose that the EOQ formula gives an order quantity of 1,400. A little reflection shows that ordering the screws in lots of 1,400 doesn't make much sense. If we schedule a delivery of 1,400 screws at the beginning of a month, 800 are used immediately and 600 are stored for later use. At the beginning of the next month another order for 1,400 has to be made, since the 600 screws stored are insufficient to meet the next month's requirement. It makes more sense to either order 800 screws at the beginning of each month or some multiple of 800 every several months.

The EOQ solution was clearly inappropriate here. Why? Recall that in deriving the EOQ formula, we assumed that demand was known and constant. The demand pattern in this example is known, but it is certainly not constant. In fact, it is very spiky. If we were to apply the methods of Chapter 5, we would assume that the demand was random. It is easy to show that over a one-year period the weekly demand has mean 200 and standard deviation 350. These values could be used to generate (Q, R) values assuming some form of a distribution for weekly demand. But this solution would not make any sense either. The demand pattern for the screws is not random; it is predictable, since it is a consequence of the production plan for the spades, which is known. The demand is *variable,* but it is not *random.*

We still have not solved the problem of how many screws to buy and when they should be delivered. One approach might be to just order once at the beginning of the year to meet the demand for an entire year. This would entail a one-time delivery of 10,400 screws at the start of each year (assuming 200 per week). What would be the advantage of this approach? Screws are very inexpensive items. By purchasing enough for an entire year's production, we would incur the fixed delivery costs only once.

There is a completely different way to approach this problem. One could simply decide to schedule deliveries of screws at the beginning of every month. This approach might be more expensive than the once-a-year delivery strategy, since fixed costs would be 12 times higher. However, it could have other advantages that more than compensate for the higher fixed costs. Monthly deliveries eliminate the need to store screws in the plant. If usage rates vary, delivery sizes could be adjusted to match need. Also, if a problem arose with the screws caused by either a defect in production or a design change in the spades, the company would not be stuck with a large inventory of useless items.

These two policies illustrate the basic difference between MRP and JIT (although, as we will see, there is much more to these production control philosophies than this). In an MRP system, we determine lot sizes based on forecasts of future demands and possibly on cost considerations. In a JIT system, we try to reduce lot sizes to their minimum to eliminate waste and unnecessary buildups of inventory.

MRP may be considered to be a top-down planning system in that all production quantity decisions are derived from demand forecasts. Lot-sizing decisions are found for every level of the production system. Items are produced based on this plan and *pushed* to the next level. In JIT, requests for goods originate at a higher level of the system and are *pulled* through the various levels of production. This is the basic idea behind push and pull production control systems.

MRP Basics

In general, a production plan is a complete specification of the amounts of each end item or final product and subassembly produced, the exact timing of the production lot sizes, and the final schedule of completion. The production plan may be broken down into several component parts: (1) the **master production schedule (MPS),** (2) the **materials requirements planning (MRP) system,** and (3) the detailed **job shop schedule.** Each of these parts can represent a large and complex subsystem of the entire plan.

At the heart of the production plan are the forecasts of demand for the end items produced over the planning horizon. An end item is the output of the productive system; that is, products shipped out the door. Components are items in intermediate stages of production, and raw materials are resources that enter the system. A schematic of the productive system appears in Figure 8–1. It is important to bear in mind that raw materials, components, and end items are defined in a relative and not an absolute sense. Hence, we may wish to isolate a portion of a company's operation as a productive system. End items associated with one portion of the company may be raw materials for another portion. A single productive system may be the entire manufacturing operation of the firm or only a small part of it.

The master production schedule (MPS) is a specification of the exact amounts and timing of production of each of the end items in a productive system. The MPS refers to *unaggregated* items. As such, the inputs for determining the MPS are forecasts for

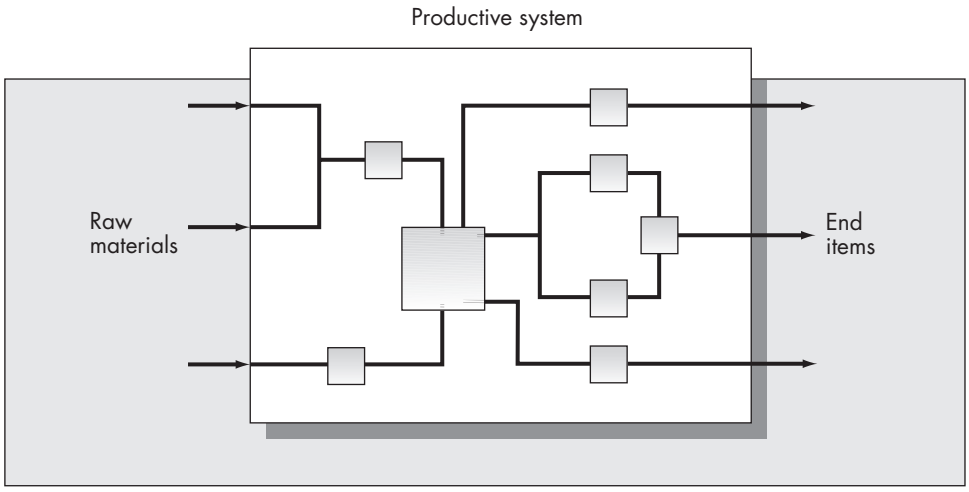


FIGURE 8–1
Schematic of the productive system

future demand by item rather than by aggregate items, as discussed in Chapter 3. The MPS is then broken down into a detailed schedule of production for each of the components that comprise an end item. The materials requirements planning (MRP) system is the means by which this is accomplished. Finally, the results of the MRP are translated into specific shop floor schedules (using methods such as those discussed in Chapter 9) and requirements for raw materials.

The data sources for determining the MPS include the following:

1. Firm customer orders.
2. Forecasts of future demand by item.
3. Safety stock requirements.
4. Seasonal plans.
5. Internal orders from other parts of the organization.

An important part of the success of MRP is the integrity and timeliness of the data. The information system that supports the MRP receives inputs from the production, marketing, and finance departments of the firm. A smooth flow of information among these three functional areas is a key ingredient to a successful production planning system.

We may consider the control of the production system to be composed of three major phases. Phase 1 is the gathering and coordinating of the information required to develop the master production schedule. Phase 2 is the determination of planned order releases using MRP, and phase 3 is the development of detailed shop floor schedules and resource requirements from the MRP planned order releases. Figure 8–2 is a schematic of these three control phases of the productive system.

This chapter is concerned with the way that the MPS is used as input to the MRP system. We will show in detail exactly how the MRP calculus works; that is, how product structures are converted to parent–child relationships between levels of the production system, how production lead times are used to obtain time-phased requirements, and how lot-sizing methods result in specific schedules. In the lot-sizing section we will

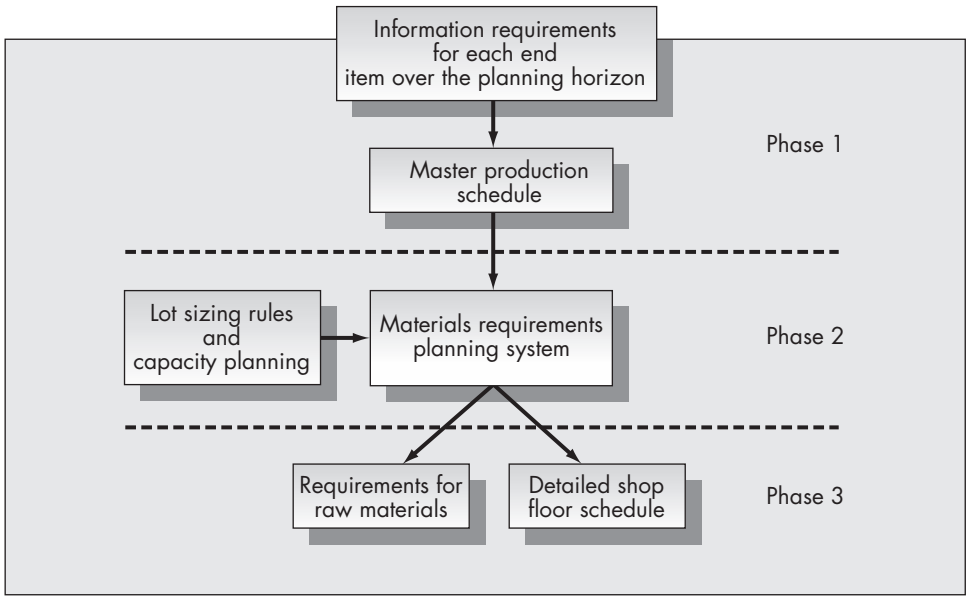


FIGURE 8–2
The three major control phases of the productive system

consider both optimal and heuristic lot-sizing techniques for uncapacitated systems and a straightforward heuristic technique for capacitated lot sizing.

JIT Basics

The just-in-time approach has its roots in the kanban system of material flow pioneered by Toyota. We will discuss the mechanics of kanban later in Section 8.6. The notion of JIT has grown significantly from its roots as a material flow technology. It has important strategic implications for firms, not just in manufacturing, but also in managing the supplier base and in distribution management.

The fundamental ideas behind JIT are

1. *Work-in-process (WIP) inventory is reduced to a bare minimum.* The amount of WIP inventory allowed is a measure of how tightly the JIT system is tuned. The less WIP designed in the system, the better balanced the various steps in the process need to be.
2. *JIT is a pull system.* Production at each stage is initiated only when requested. The flow of information in a JIT system proceeds sequentially from level to level.
3. *JIT extends beyond the plant boundaries.* Special relationships with suppliers must be in place to ensure that deliveries are made on an as-needed basis. Suppliers and manufacturers must be located in close proximity if the JIT design is to include the suppliers.
4. *The benefits of JIT extend beyond savings of inventory-related costs.* Plants can be run efficiently without the clutter of inventory of raw material and partially finished goods clogging the system. Quality problems can be identified before they build up to unmanageable proportions. Rework and inspection of finished goods are minimized.
5. *The JIT approach requires a serious commitment from top management and workers alike.* Workers need to maintain an awareness of their systems and products, and need to be empowered to stop the flow of production if they see something wrong. Management must allow these workers to have that flexibility.

Lately, the term "lean production" has been used to describe JIT. The term appears to have been coined by Womack et al. (1990) in their landmark study of the automobile industry, *The Machine That Changed the World.* In comparing the worst of American mass production and the best Japanese lean production, the authors show just how effective a properly implemented JIT philosophy can be. They described their experience at General Motors' Framingham, Massachusetts, plant in 1986:

> Next we looked at the line itself. Next to each work station were piles—in some cases weeks' worth—of inventory. Littered about were discarded boxes and other temporary wrapping material. On the line itself the work was unevenly distributed with some workers running madly to keep up and others finding time to smoke and even read a newspaper. . . . At the end of the line we found what is perhaps the best evidence of old-fashioned mass production: an enormous work area full of finished cars riddled with defects. All these cars needed further repair before shipment, a task that can prove enormously time-consuming and often fails to fix fully the problems now buried under layers of parts and upholstery.

Now contrast this with their experience at Toyota's Takaoka plant in Toyoda City:

> The differences between Takaoka and Framingham are striking to anyone who understands the logic of lean production. For a start hardly anyone was in the aisles. The armies of indirect workers so visible at GM were missing, and practically every worker in sight was actually

> adding value to the car . . . The final assembly line revealed further differences. Less than an hour's worth of inventory was next to each worker at Takaoka. The parts went on more smoothly and the work tasks were better balanced, so that every worker worked at about the same pace. . . . At the end of the line, the difference between lean and mass production was even more striking. At Takaoka we observed almost no rework area at all. Almost every car was driven directly from the line to the boat or the trucks taking cars to the buyer.

These differences are exciting and dramatic. We should note that GM has since closed Framingham and that the plants run by the "big three" (namely, GM, Ford, and Chrysler LLC) are far more efficient and better managed than Framingham. Still, we have a ways to go to duplicate the phenomenal success that the Japanese have had with lean production systems.

This chapter begins with a discussion of the basic explosion calculus of MRP and how lot-sizing strategies other than lot-for-lot are incorporated into a basic single-level MRP solution.

8.1 THE EXPLOSION CALCULUS

Explosion calculus is a term that refers to the set of rules by which gross requirements at one level of the product structure are translated into a production schedule at that level and requirements at lower levels. At the heart of any MRP system is the product structure. The product structure refers to the relationship between the components at adjacent levels of the system. The product structure diagram details the parent–child relationship of components and end items at each level, the number of periods required for production of each component, and the number of components required at the child level to produce one unit at the parent level.

A typical product structure appears in Figure 8–3. In order to produce one unit of the end item, two units of A and one unit of B are required. Assembly of A requires one week, and assembly of B requires two weeks. A and B are "children" of the end item. In order to produce A, one unit of C and two units of D are required. In order to produce B, two units of C and three units of E are required. The respective production lead times also appear on the product structure diagram. Product structure diagrams can be quite complex, with as many as 15 or more levels in some industries.

The explosion calculus (also known as the bill-of-materials explosion) follows a set of rules that translate the planned order releases for end items and components into production schedules for lower-level components. The method involves properly phasing requirements in time and accounting for the number of components required at the child level to produce a single parent item. The method is best illustrated by example.

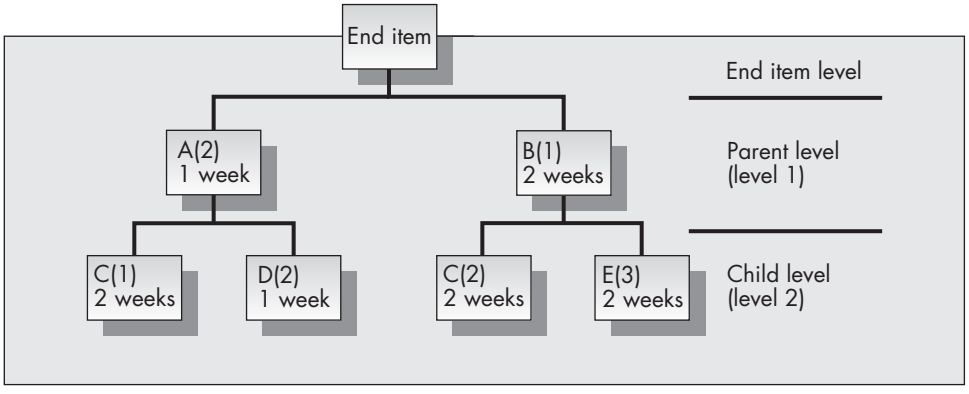


FIGURE 8–3
Typical product structure diagram

Example *8.1*

The Harmon Music Company produces a variety of wind instruments at its plant in Joliet, Illinois. Because the company is relatively small, it would like to minimize the amount of money tied up in inventory. For that reason production levels are set to match predicted demand as closely as possible. In order to achieve this goal, the company has adopted an MRP system to determine production quantities.

One of the instruments produced is the model 85C trumpet. The trumpet retails for $800 and has been a reasonably, if not spectacularly, profitable item for the company. Based on orders from music stores around the country, the production manager receives predictions of future demand for about four months into the future.

Figure 8–4 shows the trumpet and its various subassemblies. Figure 8–5 gives the product structure diagram for the construction of the trumpet. The bell section and the lead pipe and valve sections are welded together in final assembly. Before the welding, three slide assemblies and three valves are produced and fitted to the valve casing assembly. The forming and shaping of the bell section requires two weeks, and the forming and shaping of the lead pipe and valve sections require four weeks. The valves require three weeks to produce, and the slide assemblies two weeks.

The trumpet assembly problem is a three-level MRP system. Level 0 corresponds to the final product or end item, which is the completed trumpet. Level 1, the child level relative to the trumpet, corresponds to the bell and valve casing assemblies. Level 2 corresponds to the slide

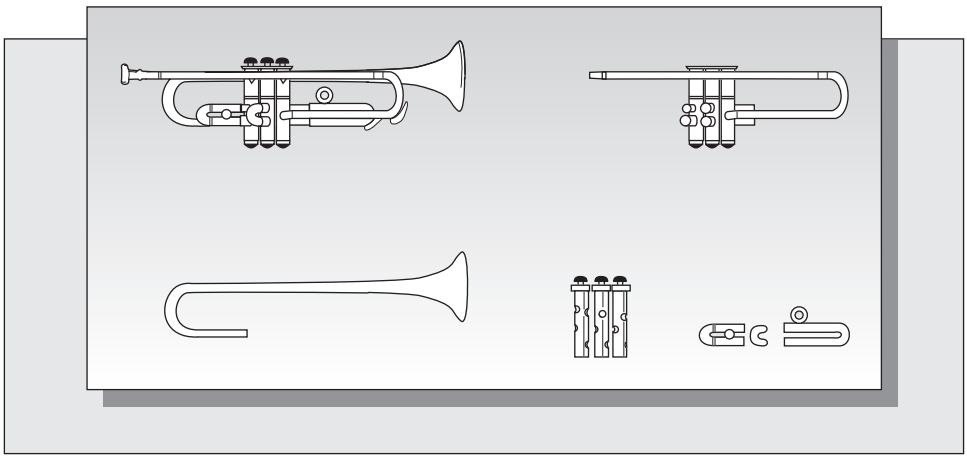


FIGURE 8–4
Trumpet and subassemblies

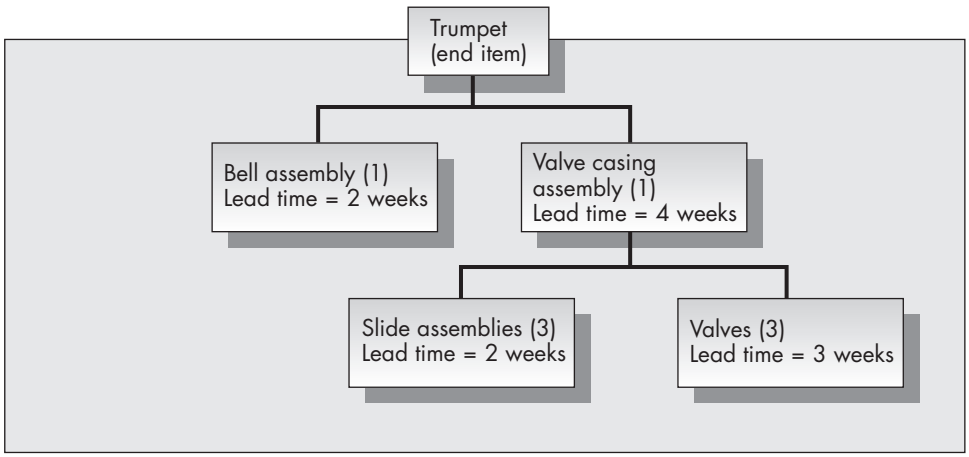


FIGURE 8–5
Product structure diagram for Harmon trumpet

and valve assemblies. The information in the product structure diagram is often represented as an indented bill of materials (BOM), which is a more convenient representation for preparation of computer input. The indented BOM for the trumpet is[1]

1 Trumpet
  1 Bell assembly
  1 Valve assembly
    3 Slide assemblies
    3 Valves

It takes seven weeks to produce a trumpet. Hence, the company must begin production now on trumpets to be shipped in seven weeks. For that reason we will consider only forecasts for demands that are at least seven weeks into the future. If we label the current week as week 1, then Harmon requires forecasts for the sales of trumpets for weeks 8 to 17. Suppose that the predicted demands for those weeks are

Week	8	9	10	11	12	13	14	15	16	17
Demand	77	42	38	21	26	112	45	14	76	38

These forecasts represent the numbers of trumpets that the firm would like to have ready to ship in the specified weeks. Harmon periodically receives returns from its various suppliers. These are instruments that are defective for some reason or are damaged in shipping. Once the necessary repairs are completed, the trumpets are returned to the pool of those ready for shipping. Based on the current and anticipated returns, the company expects the following schedule of receipts to the inventory:

Week	8	9	10	11
Scheduled receipts	12		6	9

In addition to the scheduled receipts, the company expects to have 23 trumpets in inventory at the end of week 7. The MPS for the trumpets is now obtained by netting out the inventory on hand at the end of week 7 and the scheduled receipts, in order to obtain the net predicted demand:

Week	8	9	10	11	12	13	14	15	16	17
Net predicted demand	42	42	32	12	26	112	45	14	76	38

Having determined the MPS for the end product, we must translate it into a production schedule for the components at the next level of the product structure. These are the bell assembly and the valve casing assembly. Consider first the bell assembly. The first step is to translate the MPS for trumpets into a set of gross requirements by week for the bell assembly. Because there is exactly one bell assembly used for each trumpet, this is the same as the MPS. The next step is to subtract any on-hand inventory or scheduled receipts to obtain the net requirements (here there are none). The net requirements are then translated back in time by the production lead time, which is two weeks for the bell assembly, to obtain the time-phased requirements. Finally, the lot-sizing algorithm is applied to the time-phased requirements to obtain the planned order release by period. Assuming a lot-for-lot production rule, we obtain the following MRP calculations for the bell assembly:

[1] The astute reader will know that the valves and the slides are not identical. Hence, each valve and each slide should be treated as a separate item. However, if we agree that slides and valves correspond to matching groups of three, our approach is valid. This allows us to demonstrate the multiplier effect when multiple components are needed for a single end item.

Week	6	7	8	9	10	11	12	13	14	15	16	17
Gross requirements			42	42	32	12	26	112	45	14	76	38
Net requirements			42	42	32	12	26	112	45	14	76	38
Time-phased net requirements	42	42	32	12	26	112	45	14	76	38		
Planned order release (lot for lot)	42	42	32	12	26	112	45	14	76	38		

Lot for lot means that the production quantity each week is just the time-phased net requirement. A lot-for-lot production rule means that no inventory is carried from one period to another. As we will see later, lot for lot is rarely an optimal production rule. Optimal and heuristic production scheduling rules will be examined in Section 8.2.

The calculation is essentially the same for the valve casing assembly, except that the production lead time is four weeks rather than two weeks. The calculations for the valve casing assembly are

Week	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Gross requirements					42	42	32	12	26	112	45	14	76	38
Net requirements					42	42	32	12	26	112	45	14	76	38
Time-phased net requirements	42	42	32	12	26	112	45	14	76	38				
Planned order release (lot for lot)	42	42	32	12	26	112	45	14	76	38				

Now consider the MRP calculations for the valves. Let us assume that the company expects an on-hand inventory of 186 valves at the end of week 3 and a receipt from an outside supplier of 96 valves at the start of week 5. There are three valves required for each trumpet. (Note that the valves are not identical, and hence are not interchangeable. We could display three separate sets of MRP calculations, but this is unnecessary because each trumpet has exactly one valve of each type.) One obtains gross requirements for the valves by multiplying the production schedule for the valve casing assembly by 3. Net requirements are obtained by subtracting on-hand inventory and scheduled receipts. The MRP calculations for the valves are

Week	2	3	4	5	6	7	8	9	10	11	12	13
Gross requirements			126	126	96	36	78	336	135	42	228	114
Scheduled receipts				96								
On-hand inventory		186	60	30								
Net requirements			0	0	66	36	78	336	135	42	228	114
Time-phased net requirements		66	36	78	336	135	42	228	114			
Planned order release (lot for lot)		66	36	78	336	135	42	228	114			

Net requirements are obtained by subtracting on-hand inventory and scheduled receipts from gross requirements. Because the on-hand inventory of 186 in period 3 exceeds the gross requirement in period 4, the net requirements for period 4 are 0. The remaining 60 units ($186 - 126$) are carried into period 5. In period 5 the scheduled receipt of 96 is added to the starting inventory of 60 to give 156 units. The gross requirements for period 5 are 126, so the net requirements for period 5 are 0, and the additional 30 units are carried over to period 6. Hence, the resulting net requirements for period 6 are $96 - 30 = 66$.

The net requirements are phased back three periods in order to obtain the time-phased net requirements and the production schedule. Note that the valves are produced internally. The scheduled receipt of 96 corresponds to defectives that were sent out for rework. A similar calculation is required for the slide assemblies.

Example 8.1 represents the essential elements of the explosion calculus. Note that we have assumed for the sake of the example that the production scheduling rule is lot for lot. That is, in each period the production quantity is equal to the net requirements for that period. However, such a policy may be suboptimal and even infeasible. For example, the schedule requires the delivery of 336 valves in week 9. However, suppose that the plant can produce only 200 valves in one week. If that is the case, a lot-for-lot scheduling rule is infeasible.

Problems for Section 8.1

1. The inventory control models discussed in Chapters 4 and 5 are often labeled *independent* demand models and MRP is often labeled a *dependent* demand system. What do the terms *independent* and *dependent* mean in this context?
2. What information is contained in a product structure diagram?
3. For the example of Harmon Music presented in this section, determine the planned order release for the slide assemblies. Assume lot-for-lot scheduling.
4. The Noname Computer Company builds a computer designated model ICU2. It imports the motherboard of the computer from Taiwan, but the company inserts the sockets for the chips and boards in its plant in Lubbock, Texas. Each computer requires a total of ninety 64K dynamic random access memory (DRAM) chips. Noname sells the computers with three add-in boards and two disk drives. The company purchases both the DRAM chips and the disk drives from an outside supplier. The product structure diagram for the ICU2 computer is given in Figure 8–6.

   Suppose that the forecasted demands for the computer for weeks 6 to 11 are 220, 165, 180, 120, 75, 300. The starting inventory of assembled computers in week 6 will be 75, and the production manager anticipates returns of 30 in week 8 and 10 in week 10.

   a. Determine the MPS for the computers.

   b. Determine the planned order release for the motherboards assuming a lot-for-lot scheduling rule.

   c. Determine the schedule of outside orders for the disk drives.
5. For Problem 4, suppose that Noname has 23,000 DRAM chips in inventory. It anticipates receiving a lot of 3,000 chips in week 3 from another firm that has gone out of

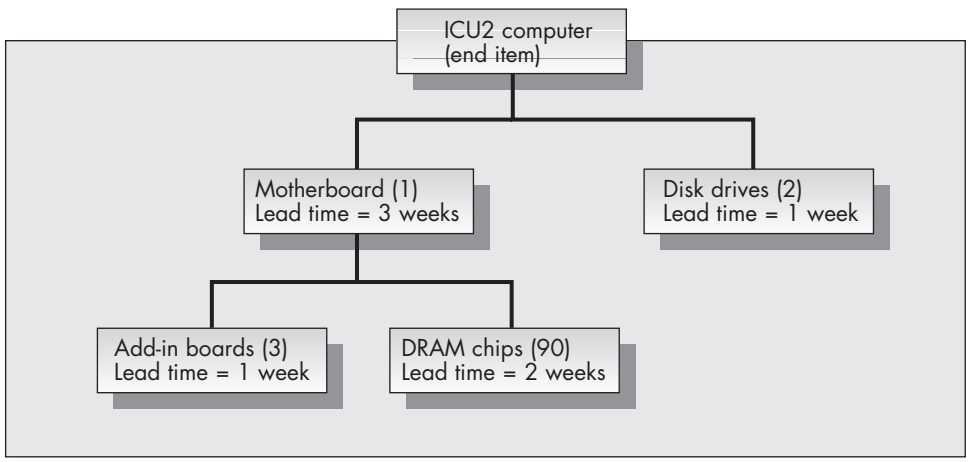


FIGURE 8–6
Product structure diagram for ICU2 computer (for Problem 4)

business. At the current time, Noname purchases the chips from two vendors, A and B. A sells the chips for less, but will not fill an order exceeding 10,000 chips per week.

a. If Noname has established a policy of inventorying as few chips as possible, what order should it be placing with vendors A and B over the next six weeks?

b. Noname has found that not all the DRAM chips purchased function properly. From past experience it estimates an 8 percent failure rate for the chips purchased from vendor A and a 4 percent failure rate for the chips purchased from vendor B. What modification in the order schedule would you recommend to compensate for this problem?

6. Consider the product structure diagram given in Figure 8–3. Assume that the MPS for the end item for weeks 10 through 17 is

Week	10	11	12	13	14	15	16	17
Net requirements	100	100	40	40	100	200	200	200

Assume that lot-for-lot scheduling is used throughout. Also assume that there is no entering inventory in period 10 and no scheduled receipts.

a. Determine the planned order release for component A.

b. Determine the planned order release for component B.

c. Determine the planned order release for component C. (Hint: Note that C is required for both A and B.)

7. What alternatives are there to lot-for-lot scheduling at each level? Discuss the potential advantages and disadvantages of other lot-sizing techniques.
8. One of the inputs to the MRP system is the forecast of demand for the end item over the planning horizon. From the point of view of production, what advantages are there to a forecasting system that smooths the demand (that is, provides forecasts that are relatively constant) versus one that achieves greater accuracy but gives "spiky" forecasts that change significantly from one period to the next?
9. An end item has the product structure diagram given in Figure 8–7.

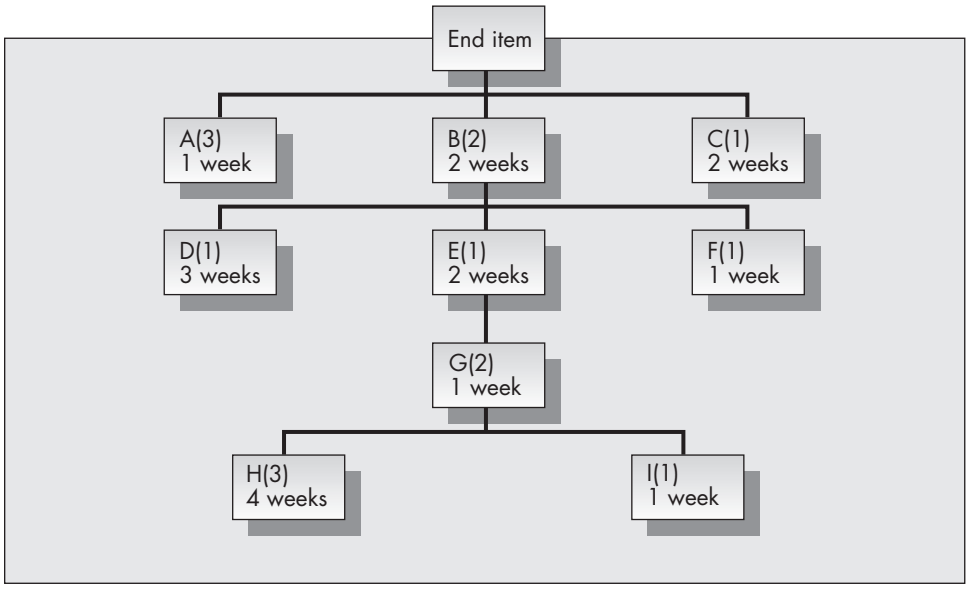


FIGURE 8–7
Product structure diagram (for Problem 9)

a. Write the product structure diagram as an indented bill-of-materials list.

b. Suppose that the MPS for the end item is

Week	30	31	32	33	34	35
MPS	165	180	300	220	200	240

If production is scheduled on a lot-for-lot basis, find the planned order release for component F.

c. Using the data in part (*b*), find the planned order release for component I.

d. Using the data in part (*b*), find the planned order release for component H.

8.2 ALTERNATIVE LOT-SIZING SCHEMES

In Example 8.1 we assumed that the production scheduling rule was lot for lot. That is, the number of units scheduled for production each period was the same as the net requirements for that period. In fact, this policy is assumed for convenience and ease of use only. It is, in general, not optimal. The problem of finding the best (or near best) production plan can be characterized as follows: we have a known set of time-varying demands and costs of setup and holding. What production quantities will minimize the total holding and setup costs over the planning horizon? Note that neither the methods of Chapter 4 (which assumes known but constant demands) nor those of Chapter 5 (which assumes random demands) are appropriate.

In this section we will discuss several popular heuristic (i.e., approximate) lot-sizing methods that easily can be incorporated into the MRP calculus.

EOQ Lot Sizing

To apply the EOQ formula, we need three inputs: the average demand rate, λ; the holding cost rate, h; and the setup cost, K. Consider the valve casing assembly in Example 8.1. Suppose that the setup operation for the machinery used in this assembly operation takes two workers about three hours. The workers average \$22 per hour. That translates to a setup cost of $(22)(2)(3) = \$132$.

The company uses a holding cost based on a 22 percent annual interest rate. Each valve casing assembly costs the company \$141.82 in materials and value added for labor. Hence, the holding cost amounts to $(141.82)(0.22)/52 = \$0.60$ per valve casing assembly per week.

The planned order release resulting from a lot-for-lot policy requires production in each week. Consider the total holding and setup cost incurred from weeks 6 through 15 when using this policy. If we adopt the convention that the holding cost is charged against the inventory each week, then the total holding cost over the 10-week horizon is zero. As there is one setup incurred each week, the total setup cost incurred over the planning horizon is $(132)(10) = \$1{,}320$.

This cost can be reduced significantly by producing larger amounts less often. As a "first cut" we can use the EOQ formula to determine an alternative production policy. The total of the time-phased net requirements over weeks 8 through 17 is 439, for an average of 43.9 per week. Using $\lambda = 43.9$, $h = 0.60$, and $K = 132$, the EOQ formula gives

$$Q = \sqrt{\frac{2K\lambda}{h}} = \sqrt{\frac{(2)(132)(43.9)}{0.6}} = 139.$$

If we schedule the production in lot sizes of 139 while guaranteeing that all net requirements are filled, the resulting MRP calculations for the valve casing assembly are

Week	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Net requirements					42	42	32	12	26	112	45	14	76	38
Time-phased net requirements	42	42	32	12	26	112	45	14	76	38				
Planned order release (EOQ)	139	0	0	0	139	0	139	0	0	139				
Planned deliveries					139	0	0	0	139	0	139	0	0	139
Ending inventory					97	55	23	11	124	12	106	92	16	117

One finds the ending inventory each period from the formula

$$\begin{matrix}\text{Ending}\\ \text{inventory}\end{matrix} = \begin{matrix}\text{Beginning}\\ \text{inventory}\end{matrix} + \begin{matrix}\text{Planned}\\ \text{deliveries}\end{matrix} - \begin{matrix}\text{Net}\\ \text{requirements.}\end{matrix}$$

Consider the cost of using EOQ lot sizing rather than lot for lot. During periods 8 through 17 there are a total of four setups, resulting in a total setup cost of (132)(4) = \$528. The most direct way to compute the holding cost is to simply accumulate the ending inventories for the 10 periods and multiply by h. The cumulative ending inventory is $97 + 55 + 23 + \cdots + 117 = 653$. Hence, the total holding cost incurred over the 10 periods is (0.6)(653) = \$391.80. The total holding and setup cost when lot sizes are computed from the EOQ formula is \$528 + \$391.80 = \$919.80. This is a considerable improvement over the cost of \$1,320 obtained when using lot-for-lot production scheduling. (However, this savings does not consider the cost impact that lot sizing at this level may have upon lower levels in the product tree. It is possible, though unlikely, that in a global sense lot for lot could be more cost effective than EOQ. This point will be explored in more depth in Section 8.5.) Note that the use of the EOQ to set production quantities results in an entirely different pattern of gross requirements for the valve and slide assemblies one level down. In particular, the gross requirements for the valves are now

Week	4	5	6	7	8	9	10	11	12	13
Gross requirements	417	0	0	0	417	0	417	0	0	417

In the remainder of this section, we discuss three other popular lot-sizing schemes when demand is known and time is varying. It should be pointed out that the problem of determining lot sizes subject to time-varying demand occurs in contexts other than MRP. We have included it here to illustrate how these methods can be linked to the MRP explosion calculus.

The Silver–Meal Heuristic

The Silver–Meal heuristic (named for Harlan Meal and Edward Silver) is a forward method that requires determining the average cost per period as a function of the number of periods the current order is to span, and stopping the computation when this function first increases.

Define $C(T)$ as the average holding and setup cost per period if the current order spans the next T periods. As above, let $(r_1, \ldots, r_n)$ be the requirements over the n-period horizon. Consider period 1. If we produce just enough in period 1 to meet the demand in period 1, then we just incur the order cost K. Hence,

$$C(1) = K.$$

If we order enough in period 1 to satisfy the demand in both periods 1 and 2, then we must hold r_2 for one period. Hence,

$$C(2) = (K + hr_2)/2.$$

Similarly,

$$C(3) = (K + hr_2 + 2hr_3)/3$$

and, in general,

$$C(j) = (K + hr_2 + 2hr_3 + \cdots + (j-1)hr_j)/j.$$

Once $C(j) > C(j-1)$, we stop and set $y_1 = r_1 + r_2 + \cdots + r_{j-1}$, and begin the process again starting at period j.

Example *8.2* A machine shop uses the Silver–Meal heuristic to schedule production lot sizes for computer casings. Over the next five weeks the demands for the casings are $\mathbf{r} = (18, 30, 42, 5, 20)$. The holding cost is \$2 per case per week, and the production setup cost is \$80. Find the recommended lot sizing.

Solution

Starting in period 1:

$C(1) = 80$,

$C(2) = [80 + (2)(30)]/2 = 70$,

$C(3) = [80 + (2)(30) + (2)(2)(42)]/3 = 102.67$. Stop because $C(3) > C(2)$.

Set $y_1 = r_1 + r_2 = 18 + 30 = 48$.

Starting in period 3:

$C(1) = 80$,

$C(2) = [80 + (2)(5)]/2 = 45$,

$C(3) = [80 + (2)(5) + (2)(2)(20)]/3 = 56.67$. Stop.

Set $y_3 = r_3 + r_4 = 42 + 5 = 47$.

Because period 5 is the final period in the horizon, we do not need to start the process again. We set $y_5 = r_5 = 20$. Hence, the Silver–Meal heuristic results in the policy $\mathbf{y} = (48, 0, 47, 0, 20)$. (Hint: You can streamline calculations by noting that $C(j+1) = [j/(j+1)][C(j) + hr_{j+1}]$.)

To show that the Silver–Meal heuristic will not always result in an optimal solution, consider the following counterexample.

Example *8.3* Let $\mathbf{r} = (10, 40, 30)$, $K = 50$, and $h = 1$. The Silver–Meal heuristic gives the solution $\mathbf{y} = (50, 0, 30)$, but the optimal solution is $(10, 70, 0)$.

In closing this section, we note that Silver and Peterson (1985, p. 238) recommend conditions under which the Silver–Meal heuristic should be used instead of the EOQ. The condition is based on the variance of periodic demand: the higher the variance, the better the improvement the heuristic gives. However, our feeling is that given today's computing technology and the ease with which the heuristic solution can be found, the additional computational costs of using Silver–Meal (or one of the following two methods described) instead of EOQ are minimal and not an important consideration.

Least Unit Cost

The least unit cost (LUC) heuristic is similar to the Silver–Meal method except that instead of dividing the cost over j periods by the number of periods, j, we divide it by the total number of units demanded through period j, $r_1 + r_2 + \cdots + r_j$. We choose the

order horizon that minimizes the cost per unit of demand rather than the cost per period.

Define $C(T)$ as the average holding and setup cost per unit for a T-period order horizon. Then,

$$\begin{aligned} C(1) &= K/r_1, \\ C(2) &= (K + hr_2)/(r_1 + r_2), \\ &\vdots \\ C(j) &= [K + hr_2 + 2hr_3 + \cdots + (j-1)hr_j]/(r_1 + r_2 + \cdots + r_j). \end{aligned}$$

As with the Silver–Meal heuristic, this computation is stopped when $C(j) > C(j-1)$, and the production level is set equal to $r_1 + r_2 + \cdots + r_{j-1}$. The process is then repeated, starting at period j and continuing until the end of the planning horizon is reached.

Example *8.4* Assume the same requirements schedule and costs as given in Example 8.2.

Starting in period 1:

$$\begin{aligned} C(1) &= 80/18 = 4.44, \\ C(2) &= [80 + (2)(30)]/(18 + 30) = 2.92, \\ C(3) &= [80 + (2)(30) + (2)(2)(42)]/(18 + 30 + 42) = 3.42. \end{aligned}$$

Because $C(3) > C(2)$, we stop and set $y_1 = r_1 + r_2 = 48$.

Starting in period 3:

$$\begin{aligned} C(1) &= 80/42 = 1.90, \\ C(2) &= [80 + (2)(5)]/(42 + 5) = 1.92. \end{aligned}$$

Because $C(2) > C(1)$, stop and set $y_3 = r_3 = 42$.

Starting in period 4:

$$\begin{aligned} C(1) &= 80/5 = 16, \\ C(2) &= [80 + (2)(20)]/(5 + 20) = 4.8. \end{aligned}$$

As we have reached the end of the horizon, we set $y_4 = r_4 + r_5 = 5 + 20 = 25$. The solution obtained by the LUC heuristic is $\mathbf{y} = (48, 0, 42, 25, 0)$. It is interesting to note that the policy obtained by this method is different from that for the Silver–Meal heuristic. It turns out that the Silver–Meal method gives the optimal policy, with cost \$310, whereas the LUC gives a suboptimal policy, with cost \$340.

Part Period Balancing

Another approximate method for solving this problem is part period balancing. Although the Silver–Meal technique seems to give better results in a greater number of cases, part period balancing seems to be more popular in practice.

The method is to set the order horizon equal to the number of periods that most closely matches the total holding cost with the setup cost over that period. The order horizon that exactly equates holding and setup costs will rarely be an integer number of periods (hence the origin of the name of the method).

Example *8.5* Again consider Example 8.2. Starting in period 1, we find

Order Horizon	Total Holding Cost
1	0
2	60
3	228

Because 228 exceeds the setup cost of 80, we stop. As 80 is closer to 60 than to 228, the first order horizon is two periods. That is, $y_1 = r_1 + r_2 = 18 + 30 = 48$.

We start the process again in period 3.

Order Horizon	Total Holding Cost
1	0
2	10
3	90

We have exceeded the setup costs of 80, so we stop. Because 90 is closer to 80 than is 10, the order horizon is three periods. Hence $y_3 = r_3 + r_4 + r_5 = 67$. The complete part period balancing solution is $\mathbf{y} = (48, 0, 67, 0, 0)$, which is different from both the Silver–Meal and LUC solutions. This solution is optimal, as it also has a total cost of \$310.

All three of the methods discussed in this section are heuristic methods. That is, they are reasonable methods based on the structure of the problem but don't necessarily give the optimal solution. In Appendix 8–A, we discuss the Wagner–Whitin algorithm that guarantees an optimal solution to the problem of production planning with time-varying demands. While tedious to solve by hand, the Wagner–Whitin algorithm can be implemented easily on a computer and solved quickly and efficiently.

Problems for Section 8.2

10. Perform the MRP calculations for the valves in the example of this section, using the gross requirements schedule that results from EOQ lot sizing for the valve casing assemblies. Use $K = \$150$ and $h = 0.4$.

11. *a.* Determine the planned order release for the motherboards in Problem 4 assuming that one uses the EOQ formula to schedule production. Use $K = \$180$ and $h = 0.40$.
    b. Using the results from part (*a*), determine the gross requirements schedule for the DRAM chips, which are ordered from an outside supplier. The order cost is \$25.00, and the holding cost is \$0.01 per chip per week. What order schedule with the vendor results if the EOQ formula is used to determine the lot size?
    c. Repeat the calculation of part (*b*) for the add-in boards. Use the same value of the setup cost and a holding cost of 28 cents per board per week.

12. *a.* Discuss why the EOQ formula may give poor results for determining planned order releases.
    b. If the forecasted demand for the end item is the same each period, will the EOQ formula result in optimal lot sizing at each level of the product structure? Explain.

13. The problem of lot sizing for the valve casing assembly described for Harmon Music Company in Section 8.2 was solved using the EOQ formula. Determine the lot sizing for the 10 periods using the following methods instead:
    a. Silver–Meal.
    b. Least unit cost.
    c. Part period balancing.
    d. Which lot-sizing method resulted in the lowest cost for the 10 periods?

14. A single inventory item is ordered from an outside supplier. The anticipated demand for this item over the next 12 months is 6, 12, 4, 8, 15, 25, 20, 5, 10, 20, 5, 12. Current inventory of this item is 4, and ending inventory should be 8. Assume a holding cost of \$1 per period and a setup cost of \$40. Determine the order policy for this item based on
    a. Silver–Meal.
    b. Least unit cost.
    c. Part period balancing.
    d. Which lot-sizing method resulted in the lowest cost for the 12 periods?
15. For the counterexample (Example 6.3), which shows that the Silver–Meal heuristic may give a suboptimal solution, do either the least unit cost or the part period balancing heuristics give the optimal solution?
16. Discuss the advantages and disadvantages of the following lot-sizing methods in the context of an MRP scheduling system: lot for lot, EOQ, Silver–Meal, least unit cost, and part period balancing.
17. The time-phased net requirements for the base assembly in a table lamp over the next six weeks are

Week	1	2	3	4	5	6
Requirements	335	200	140	440	300	200

    The setup cost for the construction of the base assembly is \$200, and the holding cost is \$0.30 per assembly per week.
    a. What lot sizing do you obtain from the EOQ formula?
    b. Determine the lot sizes using the Silver–Meal heuristic.
    c. Determine the lot sizes using the least unit cost heuristic.
    d. Determine the lot sizes using part period balancing.
    e. Compare the holding and setup costs obtained over the six periods using the policies found in parts (*a*) through (*d*) with the cost of a lot-for-lot policy.

Problems 18–22 are based on the material appearing in Appendix 8–A.

18. Anticipated demands for a four-period planning horizon are 23, 86, 40, and 12. The setup cost is \$300 and the holding cost is $h = \$3$ per unit per period.
    a. Enumerate all the exact requirements policies, compute the holding and setup costs for each, and find the optimal production plan.
    b. Solve the problem by backward dynamic programming.
19. Anticipated demands for a five-period planning horizon are 14, 3, 0, 26, 15. Current starting inventory is four units, and the inventory manager would like to have eight units on hand at the end of the planning horizon. Assume that $h = 1$ and $K = 30$. Find the optimal production schedule. (Hint: Modify the first and the last period's demands to account for starting and ending inventories.)
20. A small manufacturing firm that produces a line of office furniture requires casters at a fairly constant rate of 75 per week. The MRP system assumes a six-week planning horizon. Assume that it costs \$266 to set up for production of the casters and the holding cost amounts to \$1 per caster per week.
    a. Compute the EOQ and determine the number of periods of demand to which this corresponds by forming the ratio (EOQ)/(demand per period). Let T be

this ratio rounded to the nearest integer. Determine the policy that produces casters once every T periods.

b. Using backward dynamic programming with $N = 6$ and $\mathbf{r} = (75, 75, \ldots, 75)$, find the optimal solution. (Refer to Appendix 8–A.) Does your answer agree with what you obtained in part (*a*)?

21. *a.* Based on the results of Problem 20, suggest an approximate lot-sizing technique. Under what circumstances would you expect this method to give good results?

b. Use this method to solve Example 8A.2 (see Appendix 8–A). By what percentage does the resulting solution differ from the optimal?

22. Solve Problem 17 using the Wagner–Whitin algorithm. (Refer to Appendix 8–A.)

8.3 INCORPORATING LOT-SIZING ALGORITHMS INTO THE EXPLOSION CALCULUS

Example *8.6*

Let us return to Example 8.1 concerning the Harmon Music Company and consider the impact of lot sizing on the explosion calculus. Consider first the valve casing assembly. The time-phased net requirements for the valve casing assembly are

Week	4	5	6	7	8	9	10	11	12	13
Time-phased net requirements	42	42	32	12	26	112	45	14	76	38

The setup cost for the valve casing assembly is \$132, and the holding cost is $h = \$0.60$ per assembly per week. We will determine the lot sizing by the Silver-Meal heuristic.

Starting in week 4:

$$C(1) = 132,$$

$$C(2) = \frac{132 + 0.6(42)}{2} = 78.6,$$

$$C(3) = \frac{132 + 0.6[42 + (2)(32)]}{3} = 65.2,$$

$$C(4) = \frac{132 + 0.6[42 + (2)(32) + (3)(12)]}{4} = 54.3,$$

$$C(5) = \frac{132 + 0.6[42 + (2)(32) + (3)(12) + (4)(26)]}{5} = 55.92.$$

Since $C(5) > C(4)$, we terminate computations and set $y_4 = 42 + 42 + 32 + 12 = 128$.

Starting in week 8:

$$C(1) = 132,$$

$$C(2) = \frac{132 + 0.6(112)}{2} = 99.6,$$

$$C(3) = \frac{132 + 0.6[112 + (2)(45)]}{3} = 84.4,$$

$$C(4) = \frac{132 + 0.6[112 + (2)(45) + (3)(14)]}{4} = 69.6,$$

$$C(5) = \frac{132 + 0.6[112 + (2)(45) + (3)(14) + (4)(76)]}{5} = 92.16,$$

Hence, $y_8 = 26 + 112 + 45 + 14 = 197$.

Production occurs next in week 12. It is easy to show that $y_{12} = 76 + 38 = 114$.

A summary of the MRP calculations using the Silver–Meal (S–M) heuristic to determine lot sizes for the valve casing assembly is as follows:

Week	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Net requirements					42	42	32	12	26	112	45	14	76	38
Time-phased net requirements	42	42	32	12	26	112	45	14	76	38				
Planned order release (S–M)	128	0	0	0	197	0	0	0	114	0				
Planned deliveries					128	0	0	0	197	0	0	0	114	0
Ending inventory					86	44	12	0	171	59	14	0	38	0

It is interesting to compare the holding and setup cost of the policy using the Silver–Meal heuristic with the previous solutions using lot for lot and the EOQ formula. There are exactly three setups in our solution, resulting in a total setup cost of (132)(3) = \$396. The sum of the ending inventories each week is $86 + 44 + \cdots + 38 + 0 = 424$. The total holding cost is (0.6)(424) = \$254.40. Hence, the total cost of the Silver–Meal solution for this assembly amounts to \$650.40. Compare this to the costs of \$1,320 using lot for lot and \$919.80 using the EOQ solution.

It is also interesting to note what would have been the result if we had employed the Wagner–Whitin algorithm to find the true optimal solution. The optimal solution for this problem turns out to be $y_4 = 154$, $y_9 = 171$, and $y_{12} = 114$, with a total cost of \$610.20, which is only a slight improvement over the Silver–Meal heuristic.

We will now consider how the planned order release for the valve casing assembly affects the scheduling for lower-level components. In particular, if we consider the MRP calculations for the valves and assume that the lot sizing for the valves is determined by the Silver–Meal heuristic as well, we obtain

Week	1	2	3	4	5	6	7	8	9	10	11	12	13
Gross requirements				384	0	0	0	591	0	0	0	342	0
Scheduled receipts					96								
On-hand inventory			186	0	96	96	96	0					
Net requirements				198	0	0	0	495	0	0	0	342	0
Time-phased net requirements	198	0	0	0	495	0	0	0	342	0			
Planned order release (S–M)	198	0	0	0	495	0	0	0	342	0			

Note in this calculation summary that the scheduled receipts in period 5 must be held until period 8 before they can be used to offset demand. This results from the zero gross requirements in periods 5 through 8.

The calculation of the planned order release is based on a setup cost of \$80 and a holding cost of \$0.07 per valve per week. It is interesting to note that the Silver–Meal heuristic resulted in a lot-for-lot production rule in this case. This results from the lumpy demand pattern caused by the lot sizing applied at a higher level of the product structure. Both Silver–Meal and Wagner–Whitin give the same results for this example.

Problems for Section 8.3

23. If we were to solve Example 8.6 using the Wagner–Whitin algorithm (described in Appendix 8–A), we would obtain (154, 0, 0, 0, 0, 171, 0, 0, 114, 0) as the planned order release for the valve casing assembly. What are the resulting planned order releases for the valves?
24. Consider the example of Noname Computer Company presented in Problem 4. Suppose that the setup cost for the production of the motherboards is \$180 and the

holding cost is $h = \$0.40$ per motherboard per week. Using part period balancing, determine the planned order release for the motherboards and the resulting gross requirements schedule for the DRAM chips. (Hint: Use the net demand for computers after accounting for starting inventory and returns.)

25. For the example presented in Problem 6, assume that the setup cost for both components A and B is \$100 and that the holding costs are respectively \$0.15 and \$0.25 per component per week. Using the Silver–Meal algorithm, determine the planned order releases for both components A and B and the resulting gross requirements schedules for components C, D, and E.

8.4 LOT SIZING WITH CAPACITY CONSTRAINTS

We consider a variant of the problem treated in Section 8.3. Assume that in addition to known requirements $(r_1, \ldots, r_n)$ in each period, there are also production capacities $(c_1, \ldots, c_n)$. Hence, we now wish to find the optimal production quantities $(y_1, \ldots, y_n)$ subject to the constraints $y_i \leq c_i$, for $1 \leq i \leq n$.

The introduction of capacity constraints clearly makes the problem far more realistic. As lot-sizing algorithms can be incorporated into an MRP planning system, production capacities will be an important part of any realizable solution. However, they also make the problem more complex. The rather neat result that optimal policies always order exact requirements is no longer valid. Determining true optimal policies is difficult and time-consuming, and is probably not practical for most real problems.

Even finding a feasible solution may not be obvious. Consider our simple four-period example with vector $\mathbf{r} = (52, 87, 23, 56)$, but now suppose that the production capacity in each period is $\mathbf{c} = (60, 60, 60, 60)$. First we must determine if the problem is feasible; that is, whether at least one solution exists. On the surface the problem looks solvable, as the total requirement over the four periods is 218 and the total capacity is 240. But this problem is infeasible; the most that can be produced in the first two periods is 120, but the requirements for those periods sum to 139.

We have the following feasibility condition:

$$\sum_{i=1}^{j} c_i \geq \sum_{i=1}^{j} r_i \qquad \text{for } j = 1, \ldots, n.$$

Even when the feasibility condition is satisfied, it is not obvious how to find a feasible solution. Consider the following example.

Example *8.7*

$$\mathbf{r} = (20, 40, 100, 35, 80, 75, 25),$$
$$\mathbf{c} = (60, 60, 60, 60, 60, 60, 60).$$

Checking for feasibility, we have that:

$$r_1 = 20, \qquad c_1 = 60;$$
$$r_1 + r_2 = 60, \qquad c_1 + c_2 = 120;$$
$$r_1 + r_2 + r_3 = 160, \qquad c_1 + c_2 + c_3 = 180;$$
$$r_1 + r_2 + r_3 + r_4 = 195, \qquad c_1 + c_2 + c_3 + c_4 = 240;$$
$$r_1 + r_2 + r_3 + r_4 + r_5 = 275, \qquad c_1 + c_2 + c_3 + c_4 + c_5 = 300;$$
$$r_1 + r_2 + r_3 + r_4 + r_5 + r_6 = 350, \qquad c_1 + c_2 + c_3 + c_4 + c_5 + c_6 = 360;$$
$$r_1 + r_2 + r_3 + r_4 + r_5 + r_6 + r_7 = 375, \qquad c_1 + c_2 + c_3 + c_4 + c_5 + c_6 + c_7 = 420.$$

The feasibility test is satisfied, so we know at least that a feasible solution exists. However, it is far from obvious how we should go about finding one. Scheduling on a lot-for-lot basis is not going to work because of the capacity constraints in periods 3, 5, and 6.

We will present an approximate lot-shifting technique to obtain an initial feasible solution. The method is to back-shift demand from periods in which demand exceeds capacity to prior periods in which there is excess capacity. This process is repeated for each period in which demand exceeds capacity until we construct a new requirements schedule in which lot for lot is feasible. In the example, the first period in which demand exceeds capacity is period 3. We replace r_3 with c_3. The difference of 40 units must now be redistributed back to periods 1 and 2. We consider the first prior period, which is period 2. There are 20 units of excess capacity in period 2, which we absorb. We still have 20 units of demand from period 3 that are not yet accounted for; this is added to the requirement for period 1. Summarizing the results up until this point, we have

$$\begin{array}{rl} & 40\ \ 60\quad 60 \\ \mathbf{r}' = & (\cancel{20}, \cancel{40}, \cancel{100}, 35, 80, 75, 25), \\ \mathbf{c} = & (60, 60,\ \ 60, 60, 60, 60, 60). \end{array}$$

The next period in which demand exceeds capacity is period 5. The excess demand of 20 units can be back-shifted to period 4. Finally, the 15 units of excess demand in period 6 can be back-shifted to periods 4 (5 units) and 1 (10 units). The feasibility condition guarantees that this process leads to a feasible solution.

This leads to

$$\begin{array}{rl} & 50\qquad\quad 60 \\ & \cancel{40}\ \ 60\quad 60\ \ \cancel{55}\ \ 60\ \ 60 \\ \mathbf{r}' = & (\cancel{20}, \cancel{40}, \cancel{100}, \cancel{35}, \cancel{80}, \cancel{75}, 25), \\ \mathbf{c} = & (60, 60,\ \ 60, 60, 60, 60, 60). \end{array}$$

Hence, the modified requirements schedule obtained is

$$\mathbf{r}' = (50, 60, 60, 60, 60, 60, 25).$$

Setting $\mathbf{y} = \mathbf{r}'$ gives a feasible solution to the original problem.

The Improvement Step

We have that lot for lot for the modified requirements schedule $\mathbf{r}'$ is feasible for the original problem. Next we would like to see if we can discover an improvement; that is, another feasible policy that has lower cost. There are a variety of reasonable improvement rules that one can use. We will employ the following one.

For each lot that is scheduled, starting from the last and working backward to the beginning, determine whether it is cheaper to produce the units composing that lot by shifting production to prior periods of excess capacity. By eliminating a lot, one reduces setup cost in that period to zero, but shifting production to prior periods increases the holding cost. The shift is made only if the additional holding cost is less than the setup cost. We illustrate the process with an example.

Example *8.8*

Assume that $K = \$450$ and $h = \$2$.

$$\mathbf{r} = (100, 79, 230, 105, 3, 10, 99, 126, 40),$$
$$\mathbf{c} = (120, 200, 200, 400, 300, 50, 120, 50, 30).$$

Computing the cumulative sum of requirements and capacities for each period makes it easy to see the problem as feasible. However, because requirements exceed capacities in some periods, lot for lot is not feasible. We back-shift excess demand to prior periods in order to obtain the modified requirements schedule $\mathbf{r}' = (100, 109, 200, 105, 28, 50, 120, 50, 30)$.

Lot for lot for the modified requirements schedule **r′** is feasible for the original problem. The initial feasible solution requires nine setups at a total setup cost of $9 \times 450 = \$4{,}050$. The holding cost of the initial policy is $2(0 + 30 + 0 + 0 + 25 + 65 + 86 + 10) = \$432$.

In order to do the improvement step, it is convenient to arrange the data in a table.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
r′	100	109	200	105	28	50	120	50	30
c	120	200	200	400	300	50	120	50	30
y	100	109	200	105	28	50	120	50	30
Excess capacity	20	91	0	295	272	0	0	0	0

Starting from the last period, consider the final lot of 30 units. There is enough excess capacity in prior periods to consider shifting this lot. The latest period that this lot can be scheduled for is period 5. The extra holding cost incurred by making the shift is $2 \times 30 \times 4 = \$240$. As this is cheaper than the setup cost of \$450, we make the shift and increase y_5 from 28 to 58 and reduce the excess capacity in period 5 from 272 to 242.

Now consider the lot of 50 units scheduled in period 8. This lot can also be shifted to period 5 with a resulting additional holding cost of $2 \times 50 \times 3 = \$300$. Again this is cheaper than the setup cost, so we make the shift. At this point we have $y_5 = 108$, and the excess capacity in period 5 is reduced to 192.

The calculations are summarized on our table in the following way:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
r′	100	109	200	105	28	50	120	50	30
c	120	200	200	400	300	50	120	50	30
					108				
					~~58~~			0	0
y	100	109	200	105	~~28~~	50	120	~~50~~	~~30~~
					192				
					~~242~~				
Excess capacity	20	91	0	295	~~272~~	0	0	0	0

Next consider the lot of 120 units scheduled in period 7. At this point, we still have 192 units of excess capacity in period 5. The additional holding cost of shifting the lot of 120 from period 7 to period 5 is $2 \times 120 \times 2 = \$480$. This exceeds the setup cost of \$450, so we do not make the shift.

It is clearly advantageous to shift the lot of 50 units in period 6 to period 5, thus reducing the excess capacity in period 5 to 142 and increasing the lot size in period 5 from 108 to 158. Doing so results in the following:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
r′	100	109	200	105	28	50	120	50	30
c	120	200	200	400	300	50	120	50	30
					158				
					~~108~~				
					~~58~~			0	0
y	100	109	200	105	~~28~~	0	120	~~50~~	~~30~~
					142				
					~~192~~				
					~~242~~				
Excess capacity	20	91	0	295	~~272~~	0	0	0	0

At this point it may seem that we are done. However, there is enough capacity in period 4 to shift the entire lot of 158 units from period 5 to period 4. The additional holding cost of this shift is $2 \times 158 = \$316$. Because this is cheaper than the setup cost, we make the shift.

Summarizing these calculations on the table gives

	1	2	3	4	5	6	7	8	9
r′	100	109	200	105	28	50	120	50	30
c	120	200	200	400	300	50	120	50	30
					0				
					~~158~~				
					~~108~~				
				263	~~58~~	0		0	0
y	100	109	200	~~105~~	~~28~~	~~50~~	120	~~50~~	~~30~~
					300				
					~~142~~				
					~~192~~				
				137	~~242~~	50		50	30
Excess capacity	20	91	0	~~295~~	~~272~~	~~0~~	0	~~0~~	~~0~~

At this point, no additional lot shifting is possible. The solution we have obtained is

$$\mathbf{y} = (100, 109, 200, 263, 0, 0, 120, 0, 0)$$

for the original requirements schedule

$$\mathbf{r} = (100, 79, 230, 105, 3, 10, 99, 126, 40).$$

We will compute the cost of this solution and compare it with that of our initial feasible solution. There are five setups at a total setup cost of $5 \times 450 = \$2{,}250$. The holding cost is $2(0 + 30 + 0 + 158 + 155 + 145 + 166 + 40 + 0) = 2 \times 694 = \$1{,}388$. The total cost of this policy is \$3,638, compared with \$4,482 for our initial feasible policy. For this example the improvement step resulted in a cost reduction of close to 20 percent.

Problems for Section 8.4

26. Consider the example presented in Section 8.2 of scheduling the production of the valve casing assembly.
    a. Suppose that the production capacity in any week is 100 valve casings. Using the algorithm presented in this section, determine the planned order release for the valve casings.
    b. What gross requirements schedule for the valves does the lot sizing you obtained in part (*a*) give?
    c. Suppose that the production capacity for the valves is 200 valves per week. Is the gross requirements schedule from part (*b*) feasible? If not, suggest a modification in the planned order release computed in part (*a*) that would result in a feasible gross requirements schedule for the valves.
27. Solve Problem 14 assuming a maximum order size of 20 per month.
28. *a.* Solve Problem 17 assuming the following production capacities:

Week	1	2	3	4	5	6
Capacity	600	600	600	400	200	200

b. On a percentage basis, how much larger are the total holding and setup costs in the capacitated case than in the solutions obtained from parts (*b*), (*c*), and (*d*) of Problem 17?

29. The method of rescheduling the production of a lot to one or more prior periods if the increase in the holding cost is less than the cost of a setup also can be used when no capacity constraints exist. This method is an alternative heuristic lot scheduling technique for the uncapacitated problem. For Problem 14, start with a lot-for-lot policy and consider shifting lots backward as we have done in this section, starting with the final period and ending with the first period. Compare the total cost of the policy that you obtain with the policies derived in Problem 14.

8.5 SHORTCOMINGS OF MRP

MRP is a closed production system with two major inputs: (1) the master production schedule for the end item and (2) the relationships between the various components, modules, and subassemblies composing the production process for that end item. The method is logical and seemingly sensible for scheduling production lot sizes. However, many of the assumptions made are unrealistic. In this section, we will discuss some of these assumptions, the problems that arise as a result of them, and the means for dealing with these problems.

Uncertainty

Underlying MRP is the assumption that all required information is known with certainty. However, uncertainties do exist. The two key sources of uncertainty are the forecasts for future sales of the end item and the estimation of the production lead times from one level to another. Forecast uncertainty usually means that the realization of demand is likely to be different from the forecast of that demand. In the production planning context, it also could mean that updated forecasts of future demands are different from earlier forecasts of those demands. Forecasts must be revised when new orders are accepted, prior orders are canceled, or new information about the marketplace becomes available. That has two implications in the MRP system. One is that *all* the lot-sizing decisions that were determined in the last run of the system could be incorrect, and, even more problematic, former decisions that are currently being implemented in the production process may be incorrect.

The analysis of stochastic inventory models in Chapter 5 showed that an optimal policy included safety stock to protect against the uncertainty of demand. That is, we would order to a level exceeding expected demand. The same logic can be applied to MRP systems. The manner in which uncertainty transmits itself through a complex multilevel production system is not well understood. For that reason, it is not recommended to include independent safety stock at all levels of the system. Rather, by using the methods in Chapter 5, suitable safety levels can be built into the forecasts for the end item. These will be transmitted automatically down through the system to the lower levels through the explosion calculus.

Example *8.9* Consider Example 8.1 on the Harmon Music Company. Suppose that the firm wishes to incorporate uncertainty into the demand forecasts for weeks 8 through 17. Based on historical records of trumpet sales maintained by the firm, an analyst finds that the ratio of the standard deviation of the forecast error to the mean demand each week is near 0.3.[2] Furthermore, weekly

[2] In symbols, this is written σ/μ and is known as the coefficient of variation.

demand is closely approximated by a normal distribution. Harmon has decided that it would like to produce enough trumpets to meet all the demand each week with probability .90. (In the terminology used in Chapter 5, this means that they are using a Type 1 service level of 90 percent for the trumpets.)

The safety stock is of the form σ_z where z is the appropriate cut-off point from the normal table. Here $z = 1.28$. Incorporating safety stock into the demand forecasts, we obtain

Week	8	9	10	11	12	13	14	15	16	17
Predicted demand (μ)	77	42	38	21	26	112	45	14	76	38
Standard deviation (σ)	23.1	12.6	11.4	6.3	7.8	33.6	13.5	4.2	22.8	11.4
Mean demand plus safety stock ($\mu + \sigma_z$)	107	58	53	29	36	155	62	19	105	53

Of course, this is not the only way to compute safety stock. Alternatives are to employ a Type 2 service criterion or to use a stock-out cost model instead of a service level model. The next step is to net out the scheduled receipts and anticipated on-hand inventory to arrive at a revised MPS for the trumpets. The explosion calculus would now proceed as before, except that the safety stock that is included in the revised MPS would automatically be transmitted to the lower-level assemblies.

Safety lead times are used to compensate for the uncertainty of production lead times in MRP systems. Simply put, this means that the estimates for the time required to complete a production batch at one level and transport it to the next level would be multiplied by some safety factor. If a safety factor of 1.5 were used at Harmon Music Company, the lead times for the components would be revised as follows: bell assembly, 3 weeks; valve casing assembly, 6 weeks; slide assemblies, 3 weeks; valves, 4.5 weeks. Conceptually, safety lead times make sense if the essential uncertainty is in the production times from one level to the next, and safety stocks make sense if the essential uncertainty is in the forecast of the demand for the end item. In practice, both sources of uncertainty are generally present and some mixture of both safety stocks and safety lead times is used.

Capacity Planning

Another important issue that is not treated explicitly by MRP is the capacity of the production facility. The type of capacitated lot-sizing method we discussed earlier will deal with production capacities at one level of the system but will not solve the overall capacity problem. The problem is that even if lot sizes at some level do not exceed the production capacities, there is no guarantee that when these lot sizes are translated to gross requirements at a lower level, these requirements also can be satisfied with the existing capacity. That is, a feasible production schedule at one level may result in an infeasible requirements schedule at a lower level.

Capacity requirements planning (CRP) is the process by which the capacity requirements placed on a work center or group of work centers are computed by using the output of the MRP planned order releases. If the planned order releases result in an infeasible requirements schedule, there are several possible corrective actions. One is to schedule overtime at the bottleneck locations. Another is to revise the MPS so that the planned order releases at lower levels can be achieved with the current system capacity. This is clearly a cumbersome way to solve the problem, requiring an iterative trial-and-error process between the CRP and the MRP.

As an example of CRP, consider the manufacture of the trumpet discussed in Example 8.1 and throughout the rest of this chapter. Suppose that the valves are produced in three work centers: 100, 200, and 300. At work center 100, the molten brass is poured into the form used to shape the valve. At work center 200, the holes are drilled in the appropriate positions in the valves (there are three hole configurations, depending upon

whether the valve is number 1, 2, or 3). Finally, at work center 300, the valve is polished and the surface appropriately graded to ensure that the valve does not stick in operation. A summary of the appropriate information for the work centers is given in the following table.

Work Center	Worker Time Required to Produce One Unit (hours/unit)	Machine Throughput (units/day)
100	0.1	120
200	0.25	100
300	0.15	160

According to this information, there would be a total of six minutes (0.1 hour) of worker time required to produce a single valve at work center 100, and the existing equipment can support a maximum throughput of 120 valves per day. Consider the planned order releases obtained for the valves resulting from the Silver–Meal lot scheduling rule given in Section 8.3:

Week	2	3	4	5	6	7	8	9	10	11
Planned order release (S–M)	198	0	0	0	495	0	0	0	342	0

This planned order release translates to the following capacity requirements at the three work centers:

Week	2	3	4	5	6	7	8	9	10	11
Labor time requirements (hours):										
Work center 100	19.8	0	0	0	49.5	0	0	0	34.2	0
Work center 200	49.5	0	0	0	123.75	0	0	0	85.5	0
Work center 300	29.7	0	0	0	72.25	0	0	0	51.3	0
Machine time requirements (days):										
Work center 100	1.65	0	0	0	4.125	0	0	0	2.85	0
Work center 200	1.98	0	0	0	4.95	0	0	0	3.42	0
Work center 300	1.24	0	0	0	3.09	0	0	0	2.14	0

The capacity requirements show whether the planned order release obtained from the MRP is feasible. For example, suppose that the requirement of 123.75 labor hours in week 6 at work center 200 exceeds the capacity of this work center. This means that the current lot sizing is infeasible and some corrective action is required. One possibility is to split the lot scheduled for week 6 by producing some part of it in a prior week. Another is to adjust the lot sizing for the valve casing assembly at the next higher level of the product structure to accommodate the capacity constraints at the current level. In either case, substantial changes in the initial production plan may be required.

This example suggests an interesting speculation. Would it not perhaps make more sense to determine where the bottlenecks occur *before* attempting to explode the MPS through the various levels of the system? In this way, a feasible production plan could be found that would meet capacity constraints. Additional refinements could then be considered.

Rolling Horizons and System Nervousness

Thus far, our view of MRP is that it is a static system. Given known requirements for the end items over a specified planning horizon, one determines both the timing and the sizes of the production lot sizes for all the lower-level components. In practice,

however, the production planning environment is dynamic. The MRP system may have to be rerun each period and the production decisions reevaluated. Often it is the case that only the lot-sizing decisions for the current planning period need to be implemented. We use the term *rolling horizons* to refer to the situation in which only the first-period decision of an N-period problem is implemented. The full N-period problem is rerun each period to determine a new first-period decision.

When using rolling horizons, the planning horizon should be long enough to guarantee that the first-period decision does not change. Unfortunately, certain demand patterns are such that even for long planning horizons, the first-period decision does not remain constant. Consider the following simple example (from Carlson, Beckman, and Kropp, 1982).

Example *8.10* Suppose that the demand follows the cyclic pattern 190, 210, 190, 210, 190. . . . For a five-period planning horizon, the requirements schedule for periods 1 to 5 is

$$\mathbf{r} = (190, 210, 190, 210, 190).$$

Furthermore, suppose that $h = 1$ and $K = 400$. The optimal solution for this problem obtained from the Wagner–Whitin algorithm is

$$\mathbf{y} = (190, 400, 0, 400, 0).$$

However, suppose that the planning horizon is chosen to be six periods instead of five periods. The requirements schedule for a six-period planning horizon is

$$\mathbf{r} = (190, 210, 190, 210, 190, 210).$$

The optimal solution in this case is

$$\mathbf{y} = (400, 0, 400, 0, 400, 0).$$

That is, the first-period production quantity has changed from 190 to 400. If we go to a seven-period planning horizon, then y_1 will be 190. With an eight-period planning horizon, y_1 again becomes 400. One might think that this cycling of the value of y_1 would continue indefinitely. It turns out that this is *not* the case, however. For planning horizons of $n \geq 21$ periods, the value of y_1 remains fixed at 190.[3] However, even when there is eventual convergence of y_1, as in this example, the cycling for the first 20 periods could be troublesome when using rolling planning horizons.

Another common problem that results when using MRP is "nervousness." The term was coined by Steele (1973), who used it to refer to the changes that can occur in a schedule when the horizon is moved forward one period. Some of the causes of nervousness include unanticipated changes in the MPS because of updated forecasts, late deliveries of raw materials, failure of key equipment, absenteeism of key personnel, and unpredictable yields.

There has been some analytical work on the nervousness problem. Carlson, Jucker, and Kropp (1979 and 1983) use the term *nervousness* specifically to mean that a revised schedule requires a setup in a period in which the prior schedule did not. They have proposed an interesting technique to reduce this particular type of nervousness: Let $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$ be the existing production schedule and $(y_1, y_2, \ldots, y_N)$ be a revised schedule based on new demand information. Suppose that besides the usual costs of holding and setup, there is an additional cost of v if the new schedule $\mathbf{y}$ calls for a setup in a period that the old schedule $\hat{\mathbf{y}}$ did not. This means that there is an additional

[3] I am grateful to Lawrence Robinson of Cornell University for pointing out that this anomaly does not continue for all values of n as was claimed in past editions and by Carlson, Beckman, and Kropp (1982).

setup cost associated with the new schedule in those periods in which no setup was called for in the old schedule. Their method is to increase the setup cost from K to $K + v$ if $\hat{y}_k = 0$ prior to determining the new schedule $\mathbf{y}$. Re-solving the problem with the modified setup costs using any of the lot-sizing algorithms previously discussed now will result in fewer setup revisions. The advantages of the various lot-sizing methods in this context are discussed in the two given references. The revision cost v reflects the relative importance of the cost of nervousness.

Additional Considerations

Although MRP would seem to be the most logical way to schedule production in batch-type operations, as we saw above the basic method has some very serious shortcomings. Other difficulties are considered.

Lead Times Dependent on Lot Sizes

The MRP calculus assumes that the production lead time from one level to the next is a fixed constant independent of the size of the lot. In many contexts this assumption is clearly unreasonable. One would expect the lead time to increase if the lot size increases. Including a dependence between the production lead time and the size of the production run into the explosion calculus seems to be extremely difficult.

MRP II: Manufacturing Resource Planning

As we have noted, MRP is a closed production planning system that converts an MPS into planned order releases. Manufacturing resource planning (MRP II) is a philosophy that attempts to incorporate the other relevant activities of the firm into the production planning process. In particular, the financial, accounting, and marketing functions of the firm are tied to the operations function. As an example of the difference between the perspectives offered by MRP and MRP II, consider the role of the master production schedule. In MRP, the MPS is treated as input information. In MRP II, the MPS would be considered a part of the system and, as such, would be considered a decision variable as well. Hence, the production control manager would work with the marketing manager to determine when the production schedule should be altered to incorporate revisions in the forecast and new order commitments. Ultimately, all divisions of the company would work together to find a production schedule consistent with the overall business plan and long-term financial strategy of the firm.

Another important aspect of MRP II is the incorporation of capacity resource planning (CRP). Capacity considerations are not explicitly accounted for in MRP. MRP II is a closed-loop cycle in which lot sizing and the associated shop floor schedules are compared to capacities and recalculated to meet capacity restrictions. However, capacity issues continue to be an important issue in both MRP and MRP II operating systems.

Obviously, such a global approach to the production scheduling problem is quite ambitious. Whether such a philosophy can be converted to a workable system in a particular operating environment remains to be seen.

Imperfect Production Processes

An implicit assumption made in MRP is that there are no defective items produced. Because requirements for components and subassemblies are computed to exactly satisfy end-item demand forecasts, losses due to defects can seriously upset the balance of the production plan. In some industries, such as semiconductors, yield rates can be as low as 10 to 20 percent for new products. As long as yields are stable, incorporating yield losses into the MRP calculus is not difficult. One computes net demands and lot

Snapshot Application

RAYMOND CORPORATION BUILDS WORLD-CLASS MANUFACTURING WITH MRP II

The Raymond Corporation is a major materials handling equipment manufacturer headquartered in Greene, New York. In order to achieve an estimated payback of 10 times in reduced inventories and organizational efficiencies, the organization underwent a process of implementation of MRP II that spanned over two years. The success of their effort was based on several basic principles. The first was that top management had to be committed to the process of change. For that reason, the first step was the education of the top management utilizing a "canned" educational package available from an outside consulting firm. The CEO facilitated on-site training for the vice presidents, who in turn were responsible for training the middle management implementation team. With top management on board, middle management could not simply ignore the issue and wait for it to go away. Training continued throughout the implementation phase of the project and ultimately included employees at every level of the company.

Once training was far enough along to begin thinking about implementation, a strategy for implementation was laid out. The first step was to take a hard look at the data that would be used as inputs to the system. As a result of this effort, stockroom reporting accuracy went from 66 percent to over 95 percent in about 16 months. Every inventory status report was measured for accuracy and corrected when necessary. Setting up standards and systems that give accurate information can be the lion's share of the benefit from an effort such as this. An early payoff was that reporting accuracy led to reduced need to closely monitor purchasing. Inventory status reports were used for ABC classification, and more time and energy were devoted to managing the "A" items.

In order to determine the effectiveness of new systems, the implementation team would meet weekly to review the progress of internal measures. An attempt was made to determine the cause of problems or lack of progress without pointing fingers and assigning blame. Performance measurements become an important part of the success of the implementation effort. Accurate and up-to-date performance measurements are the cornerstone of any systems change effort, but they must be put in place without threatening workers.

Finally, the author recommends that the *final* step should be the purchase of new software. If the underlying data and measurements systems in place are not sound, the software, no matter how sophisticated, won't help. Many firms believe that purchasing an expensive MRP software system is all that's necessary to achieve reduced inventory, lower materials costs, improved on-time delivery, and so on. However, without employees that are on-board, accurate data, and performance measurements, new software can be more of a hindrance than a help.

Aside from reaching class A status, what benefits did Raymond see from this process? Sheldon (1994) claimed elimination of overtime, elimination of shortages, improved on-time deliveries, reductions in setup times and costs, reductions in rework and scrap rates, and lower inventory and material handling costs. While it is hard to believe that the improvements were as dramatic as this, a carefully planned implementation of MRP II clearly paid off for the Raymond Corporation.

Source: This discription is based on Sheldon (1994).

sizes in the same way, and in the final step divides the planned order release figures by the average yield. For example, if a particular process has a yield rate of 78 percent, one would multiply the planned order releases by $1/0.78 = 1.28$. The problem is much more complex if yields are random and variances are too large to be ignored. Using mean yields would result in substantial stock-outs. One would have to develop a kind of newsboy model that balanced the cost of producing too many and too few and determine an appropriate safety factor. Because of the dependencies of successive levels, this would be a difficult problem to model mathematically. Monte Carlo computer simulation would be a good alternative. To this writer's knowledge, no one has attempted to develop a mathematical model of random yields in the context of MRP systems. (Random yields, however, are discussed by several researchers. For example, Nahmias and Moinzadeh, 1997, consider a mathematical model of random yields in a single-level lot-sizing problem.)

Data Integrity

An MRP system can function effectively only if the numbers representing inventory levels are accurate. It is easy for incorrect data to make their way into the scheduling system. This can occur if a shipment is not recorded or is recorded incorrectly at some level, items entering inventory for rework are not included, scrap rates are higher than anticipated, and so on. In order to ensure the integrity of the data used to determine the size and the timing of lot sizes, physical stock-taking may be required at regular intervals. An alternative to complex physical stock-taking is a technique known as *cycle counting.* Cycle counting simply means directly verifying the on-hand levels of the various inventories comprising the MRP system. For example, are the 45 units of part A557J indicated on the current record the actual count of this part number?

Efficient cycle counting can be achieved in a variety of ways. Stockrooms may have containers that only hold a fixed number or weight of items. Coded shelving systems could be used to more easily identify items with part numbers. Certain areas could be made accessible only to specific personnel. Cycle counting systems can be based on number or on weight. Furthermore, an error in the inventory level must be considered in relative terms. Based on the importance of the item, different percentage errors may be considered acceptable. Different error tolerances should be applied to weigh-counted items versus hand-counted items. If MRP is to have a positive impact on the overall production scheduling problem, the inventory records must be an accurate reflection of the actual state of the system.

Order Pegging

In some complex systems, a single component may be used in more than one item at the next higher level of the system. For example, a company producing many models of toy cars may use the same-sized axle in each of the cars. Gross requirements for axles would be the sum of the gross requirements generated by the MPS for each model of car. Hence, when one component is used in several items, the gross requirements schedule for this component comes from several sources. If a shortage of this component occurs, it is useful for the firm to be able to identify the particular items higher in the tree structure that would be affected. In order to do this, the gross requirements schedule is broken down by the items that generate it and each requirement is "pegged" with the part number of the source of the requirement. Pegging adds considerable complexity to the information storage requirements of the system and should only be considered when the additional information is important in the decision-making process.

Problems for Section 8.5

30. MRP systems have been used with varying degrees of success. Describe under what circumstances MRP might be successful and under what circumstances it would not be successful.
31. Discuss the advantages and disadvantages of including safety stock in MRP lot-sizing calculations. Do you think that a production control manager would be reluctant to build to safety stock if he or she is behind schedule?
32. For what reason is the capacitated lot-sizing method discussed in Section 8.5 not adequate for solving the overall capacity problem?
33. Planned order releases (POR) for three components, A, B, and C, are given below. Suppose that the yields for these components are respectively 84 percent, 92 percent,

and 70 percent. Assuming lot-for-lot scheduling, how should these planned order releases be adjusted to account for the fact that the yields are less than 100 percent?

Week	6	7	8	9	10	11	12	13	14	15	16	17
POR(A)				200	200	80	80	200	400	400	400	
POR(B)			100	100	40	40	100	200	200	200		
POR(C)	200	400	280	100	280	600	800	800	400			

34. Define the terms *rolling horizons* and *system nervousness* in the context of MRP systems.

8.6 JIT FUNDAMENTALS

Just-in-time, lean production, and zero inventories are all names for essentially the same thing: a system of moving material through a plant that requires a minimum of inventory. Some have speculated that the roots of the system go back to the situation in postwar Japan. Devastated by the war, the Japanese firms were cash poor and did not have the luxury of investing cash in excess inventories. Thus, lean production was born from necessity. However, as Japanese cars started gaining in popularity in the United States, it quickly became clear that they were far superior to American- and European-made cars in terms of quality, value, efficiency, and reliability. We know today that JIT and the quality initiatives of the 1950s played important roles in this success.

Two developments were key to the success of this new approach to mass production: the kanban system and SMED (which stands for single minute exchange of dies). Kanban is a Japanese word for *card* or *ticket.* It is a manual information system developed and used by Toyota for implementing just-in-time.

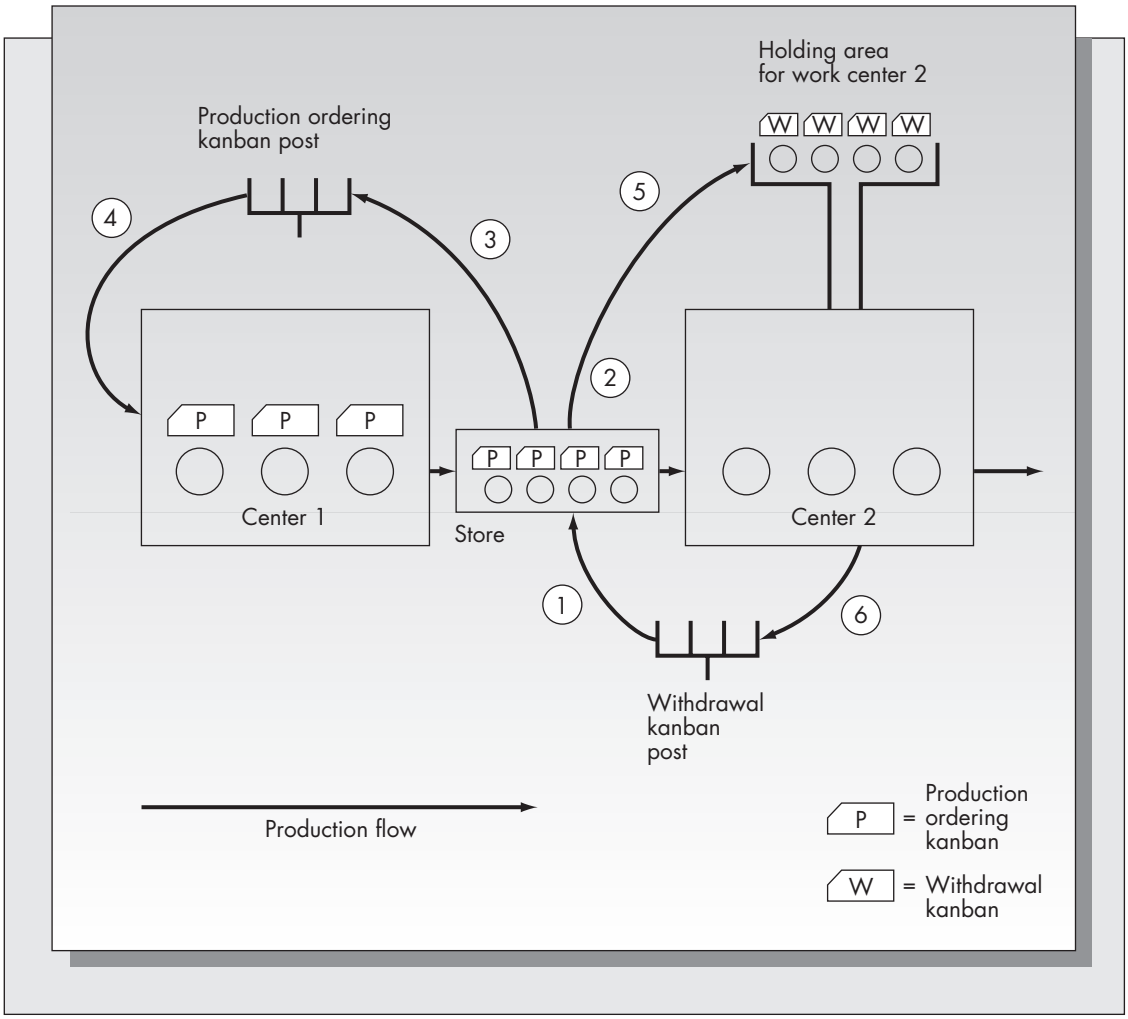
The Mechanics of Kanban

There are a variety of different types of kanban tickets, but two are the most prevalent. These are withdrawal kanbans and production ordering kanbans. A withdrawal kanban is a request for parts to a work center from the prior level of the system. A production ordering kanban is a signal for a work center to produce additional lots. The manner in which these two kanban tickets are used to control the flow of production is depicted in Figure 8–8.

The process is as follows: Parts are produced at work center 1, stored in an intermediate location (known as the store), and subsequently transported to work center 2. Parts are transported in small batches represented by the circles in the figure. Production flows from left to right in the diagram. The detailed steps in the process are as follows (the numbers below appear in the appropriate locations in Figure 8–8):

1. When the number of tickets on the withdrawal kanban reaches a predetermined level, a worker takes these tickets to the store location.
2. If there are enough canisters available at the store, the worker compares the part number on the production ordering kanbans at the store with the part number on the withdrawal kanbans.
3. If the part numbers match, the worker removes the production ordering kanbans from the containers, places them on the production ordering kanban post, and places the withdrawal kanbans in the containers.
4. When a specified number of production ordering kanbans have accumulated, work center 1 proceeds with production.

FIGURE 8–8
Kanban system for two production centers



5. The worker transports parts picked up at the store to work center 2 and places them in a holding area until they are required for production.
6. When the parts enter production at work center 2, the worker removes the withdrawal kanbans and places them on the withdrawal kanban post. (Note that production ordering kanbans for work center 2 are then attached to the parts produced at that work center. These kanban tickets are not shown in Figure 8–8.)

One computes the number of kanban tickets in the system in advance. Toyota uses the following formula (Monden, 1981b):

$$y = \frac{\overline{D}L + w}{a},$$

where

y = Number of kanbans.

$\overline{D}$ = Expected demand per unit of time.

L = Lead time (processing time + waiting time between processes + conveyance time).

w = Policy variable specifying the level of buffer stock, generally around 10 percent of $\overline{DL}$.

a = Container capacity (usually no more than 10 percent of daily demand).

This formula implies that the maximum level of inventory is given by $ay = \overline{DL} + w$. The ideal value of w is zero. However, it is difficult to balance a system so perfectly that buffer stock is eliminated entirely.

As mentioned earlier, the kanban system is a manual information system for implementing just-in-time. JIT systems also can be implemented in other ways, which may be more efficient than the kanban method. More will be said about this later in the section.

Single Minute Exchange of Dies

One of the key components of the success of Toyota's production system was the concept of single minute exchange of dies (SMED), championed by Shigeo Shingo. Shingo is generally credited with developing and implementing SMED at Toyota in 1970, which has become an important part of the overall Toyota production system. The basic theory developed in Chapter 4 tells us that small lot sizes will be optimal only if fixed costs are small (as K decreases in the EOQ formula, so does the value of Q). The most significant component of the cost of setting up for a new operation in a plant is the time required to change over the machinery for that operation, since the production line must be frozen during the changeover operation. This requires changing some set of tools and/or dies required in the process, hence the origin of the term SMED. (A die is a tool used for shaping or impressing an object or material.)

Die-changing operations are required in automotive plants when switching over the production line from one car model to another. These operations typically required about four hours. The idea behind SMED is that a significant portion of the die-changing operation can be done off-line while the previous dies are still in place and the line continues to operate. According to Shingo (1981), this is accomplished by dividing the die-changing operation into two components: inside exchange of die (IED) and outside exchange of die (OID). The OID operation is to be performed while the line is running in advance of the actual exchange. The goal is to structure the die change so that there are as many steps as possible in the OID portion of the operation.

While this idea sounds simple, it has led to dramatic improvements in throughput rates in many circumstances, both within and outside the automotive industry. Shingo (1981) describes some of the successes:

- At Toyota, an operation that required eight hours for exchange of dies and tools for a bolt maker was reduced over a year's time to only 58 seconds.
- At Mitsubishi Heavy Industry, a tool-changing operation for a boring machine was reduced from 24 hours to only 2 minutes and 40 seconds.
- At H. Weidmann Company of Switzerland, changing the dies for a plastic molding machine was reduced from 2.5 hours to 6 minutes and 35 seconds.
- At Federal-Mogul Company of the United States, the time required to exchange the tools for a milling machine was reduced from 2 hours to 2 minutes.

Of course, SMED cannot be applied in all manufacturing contexts. Even in contexts where it can be applied, the benefits of the die-changing time reduction can be realized only if the process is integrated into a carefully designed and implemented overall manufacturing control system.

Advantages and Disadvantages of the Just-in-Time Philosophy

Champions of just-in-time would have one believe that all other production planning systems are now obsolete. The zeal with which they promote the method is reminiscent of the enthusiasm that heralded MRP in the early 1970s. At that time, some claimed that classical inventory analysis was no longer valid. However, each new production method should be viewed as an addition to, rather than a replacement for, what is already known. Just-in-time can be a useful tool in the right circumstances, but is far from a panacea for the problems facing industry today.

Just-in-time and EOQ are not mutually exclusive. Setup time and, consequently, setup cost reduction result in smaller lot sizes. Smaller lot sizes require increased efficiency and reliability of the production process, but considerably less investment in raw materials, work-in-process inventories, and finished-goods inventories.

Kanban is a manual information system used to support just-in-time inventory control. Just-in-time and kanban are not necessarily wedded. In future years, we expect to see just-in-time systems based on more current and sophisticated information transfer technology. A shortcoming of kanban is the time required to transmit new information through the system. Figure 8–9 considers a schematic of a serial production process with six levels. With the kanban system, the direction of the flow of information is opposite to the direction of the flow of the production. Consider the consequences of a sudden change in the requirements at level 6. This change is transmitted first to level 5, then to level 4, and so on. There could be a substantial time lag from the instant that the change occurs at level 6 until the information reaches level 1.

A centralized information processing system will help to alleviate this problem. If there are sudden changes in the requirements at one end of the system resulting from unplanned changes in demand or breakdowns of key equipment, these changes will be instantly transmitted to the entire system.

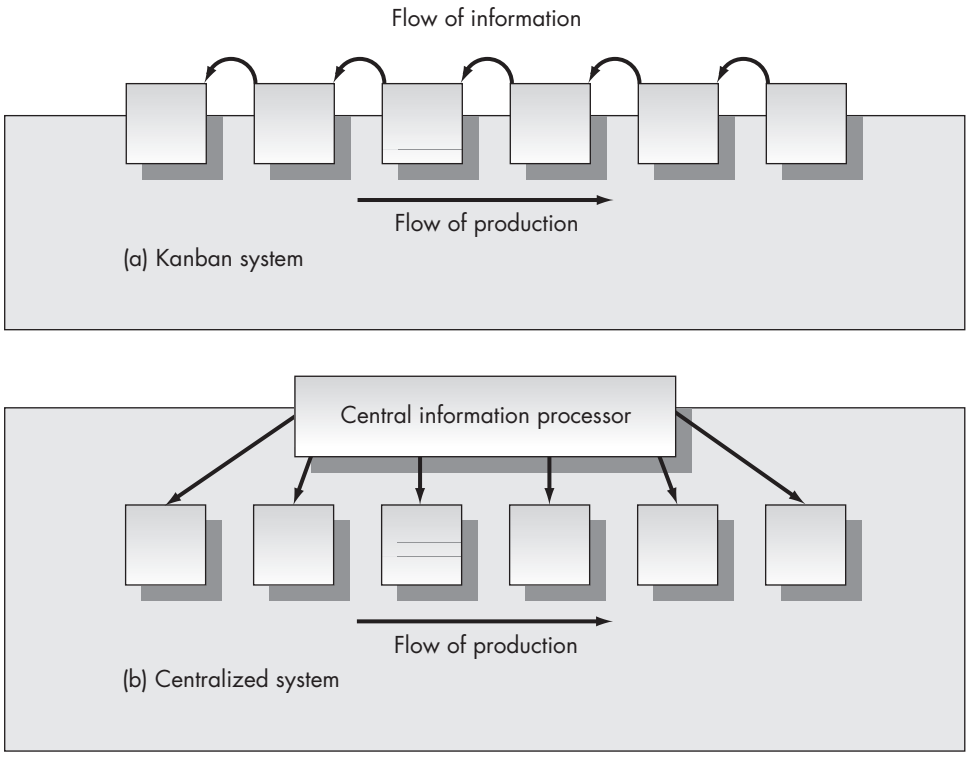


FIGURE 8–9
Kanban information system versus centralized information system

MRP has an important advantage over kanban in this regard. One of the strengths of MRP is its ability to react to forecasted changes in the pattern of demand. The MRP system recomputes production quantities based on these changes and makes this information available to all levels simultaneously. MRP allows planning to take place at all levels in a way that just-in-time, and especially kanban, will not. As Meal (1984) points out:

> Just-in-time production works well when the overall production rate is constant, but it is unsatisfactory for communicating basic changes in production rate to earlier stages in the process. . . . On the other hand using the HPP [hierarchical production planning] approach, plant managers do not rely on their short term signals to establish their early stage production rates.

Just-in-time is most efficient when the pattern of demand is stable and predictable. Changes in the demand may result from predictable causes or random fluctuations or both. MRP makes use of forecasts of anticipated changes in demand and transmits this information to all parts of the productive system. However, neither MRP nor just-in-time is designed to protect against random fluctuations of demand. Both methods could be unstable in the face of high demand variance.

Another potential shortcoming of just-in-time is the idle time that may result when unscheduled breakdowns occur. Part of the Japanese philosophy is that workers should have a familiarity with more than one portion of the productive process. If a breakdown occurs, then workers' attention can be immediately focused on the problem. However, if workers are familiar with only their own operation, there will be significant worker idle time when a breakdown occurs. This is consistent with the trade-off curve presented in Figure 9–4 of Chapter 9 on shop floor control and sequence scheduling. Buffer inventories between successive operations provide a means of smoothing production processes. However, buffer inventories also have disadvantages. They can mask underlying problems. A popular analogy is to compare a production process with a river and the level of inventory with the water level in the river. When the water level is high, the water will cover the rocks. Likewise, when inventory levels are high, problems are masked. However, when the water level (inventory) is low, the rocks (problems) are evident (see Figure 8–10).

Because items are moved through the system in small batches, 100 percent inspection is feasible. Seen in this light, just-in-time can be incorporated easily into an overall quality control strategy. Total quality management (TQM), discussed in Chapter 12, and

FIGURE 8–10
River/inventory analogy illustrating the advantages of just-in-time

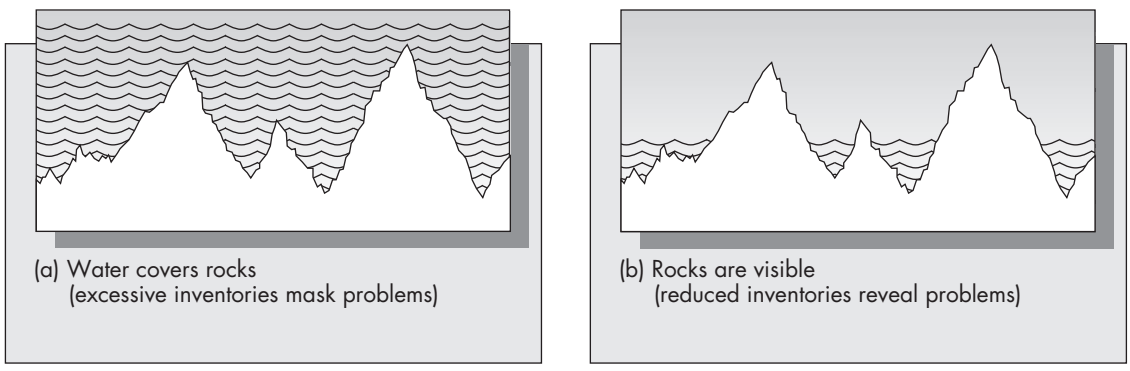


TABLE 8–1
Summary of Advantages and Disadvantages of Just-in-Time and Kanban

Feature	Advantages	Disadvantages
Small work-in-process inventories	1. Decreases inventory costs. 2. Improves production efficiency. 3. Points out quality problems quickly.	1. May result in increased worker idle time. 2. May decrease the production rate.
Kanban information flow system	1. Provides for efficient lot tracking. 2. Inexpensive means of implementing just-in-time. 3. Allows for predetermined level of WIP inventory by presenting number of kanban tickets.	1. Slow to react to changes in demand. 2. Ignores known information about future demand patterns.
Coordinated inventory and purchasing	1. Inventory reduction. 2. Improved coordination of different systems. 3. Improved relationships with vendors.	1. Decreased opportunity for multiple sourcing. 2. Suppliers must react more quickly. 3. Improved reliability required of suppliers

JIT can work together not only to reduce inventory costs, but also to bring about significant improvements in product quality.

As users of just-in-time point out, it is not simply an inventory control system. For just-in-time to work properly, it must be coordinated with the purchasing system and with purchasing strategies. One complaint about just-in-time is that it merely pushes system uncertainty and higher inventories onto the supplier. There is no doubt that greater flexibility on the part of the suppliers is necessary; they must be able to react quickly and provide sufficiently reliable parts to relieve the manufacturer of the necessity to inspect all incoming lots. Furthermore, multiple sourcing becomes difficult under such a system. That is, the firm may be forced to deal with a single supplier in order to develop the close relationship that the system requires. Single sourcing presents risks for both suppliers and manufacturers. The manufacturer faces the risk that the supplier will be unable to supply parts when they are needed, and the supplier faces the risk that the manufacturer will suffer reverses and demand will drop.

Table 8–1 briefly summarizes the primary advantages and disadvantages discussed in this section.

We close by noting that JIT and MRP are certainly not the only ways to approach manufacturing control. In their book, Hopp and Spearman (1996) explore a production planning system based on fixing WIP inventory. If reduction of work-in-process is a goal of JIT, then why not design a system from scratch that forces a desired level of WIP? They have designed just such a system, which they call CONWIP (for *CONstant Work-In-Process*). In a CONWIP system, each time an item is completed (exits the production line), initiation of production of a new item is begun at the start of the line. In this way the WIP stays fixed. One can adjust the size of the WIP based on variability and cost considerations. CONWIP is similar in principle to JIT in that both are pull systems. However, with CONWIP, the information flow is not from a stage in the process

to the previous stage, but from the end of the process to the beginning. The method sounds intriguing, but still needs to stand the test of implementation in the real world.

Implementation of JIT in the United States

A dramatic example of successful implementation of JIT in the United States is the Harley-Davidson Motorcycle Company. Until the early 1980s Harley-Davidson was owned by American Foundry Company (AMF). Harley, well known as a manufacturer of large-displacement motorcycles, faced severe competition from the Japanese. Honda, traditionally a manufacturer of small-displacement motorcycles, was beginning to make inroads into Harley's market. It appeared that motorcycles would become another consumer product produced only by foreign companies.

The first step in Harley's recovery was the purchase of the company from AMF by a group of employees. Shortly after the buyout, top management traveled to Japan to see how its competitors' factories were run. According to Willis (1986),

> The real "secret," the executives discovered, lay not in robotics or high-tech systems but in the intelligent, efficient organization of the company's employees and production systems.

As a result of these visits, Harley-Davidson was completely reorganized. Traditional management structure was replaced by a system in which each employee was given ownership of his or her area of the line. A large proportion of staff jobs were cut, resulting in a much shallower organization chart similar to that in Japanese companies. In this way, problems could not become buried in bureaucracy. The firm also instituted a quality circles program with active participation of the line personnel (see the discussion of quality circles in Chapter 12).

The firm invested in state-of-the-art computer-aided design/computer-aided manufacturing (CAD/CAM) and robotics equipment. The physical layout of the plant was restructured into work cells using a group technology layout (see Chapter 11). Motorcycles were no longer manufactured on a traditional assembly line.

A key change was the manner in which material was moved through the plant. Harley adopted the term MAN (materials as needed) to describe its system, but this was clearly just another name for JIT. To make JIT work, it instituted a program to reduce the number of setups each day. Prior to the restructuring, it produced 13 to 14 different models each day. Under the new system, improved forecasting allowed it to implement two-hour repeating cycles.

In order to make JIT work, Harley needed a commitment from its vendors as well as its own employees. To encourage its vendors to "buy in" to the program, Harley offered training sessions on the principles of JIT. These programs became so popular that Harley started to offer them to firms that were neither vendors nor clients. What is the result of this effort? Today Harley-Davidson's motorcycles are 99 percent defect-free, as compared to only 50 percent in 1982. In addition to achieving dramatic improvements in product quality, Harley developed new employee benefit programs and expanded its product line. All this was accomplished with a simultaneous reduction in costs.

Harley-Davidson is not the only example of successful implementation of JIT in the United States. Jacobson and Hillkirk (1986) describe the turnaround of Xerox, which was beleaguered by competition from Kodak and several Japanese firms (especially Canon) in the early 1980s. Xerox embarked on a self-examination program based on competitive benchmarking and undertook a complete restructuring of its business. Implementation of JIT systems played a major role in its turnaround. For example, it reduced the number of vendors supplying the company from 5,000 to only 400, which included a mix of sources overseas as low-cost suppliers and sources close by to

provide timely deliveries when needed. Its systems allowed it to reduce copier parts inventories by $240 million.

The American auto industry has made significant moves toward embracing JIT, but there is still a long way to go. There are many examples of plants in the United States that have adopted the JIT philosophy. Two General Motors ventures are the NUMMI plant in northern California, a joint venture with Toyota, and the Saturn plant in Tennessee. Ironically, Ford, which pioneered the assembly line for mass production of automobiles, has been the most successful American firm at adopting Japanese manufacturing philosophies. Ford's assembly plant near Atlanta, Georgia, at which the Ford Taurus and the Mercury Sable are manufactured, is impressive on most productivity and quality measures, even when compared with the Japanese. When GM attempted to isolate the factors responsible for Ford's significantly better productivity in this plant as compared to GM's Fairfax, Kansas, plant, which makes the Pontiac Grand Prix, it found that 41 percent of the difference could be attributed to the manufacturability of the car's designs and 48 percent to factory practice (Womack et al., 1990, p. 96).

Although U.S. firms have certainly made significant progress toward adopting JIT principles, the conversion to "lean production" has not been as quick as one would have hoped. This is partly due to an environment that is less conducive to JIT. For example, in the U.S. auto industry, assembly plants and suppliers are often separated by large distances. Japanese firms can have daily and even hourly deliveries from suppliers, because those suppliers are located close to assembly plants. Also, JIT works best when product demand is relatively stable. Sales of automobiles in the United States have been far more cyclic than in Japan (Womack et al., 1990, p. 247), making a pure JIT system more difficult to implement here. Hence, the U.S. auto industry faces a more difficult environment for successful implementation of JIT. Even with these difficulties, both U.S. and European manufacturers will have to make the transition to lean production methods if they hope to remain competitive with the Japanese.

Problems for Section 8.6

35. Discuss the concepts of push versus pull and how they relate to just-in-time.
36. What is the difference between a just-in-time system and a kanban system? Can just-in-time be implemented without kanban?
37. A regional manufacturer of table lamps plans on using a manual kanban information system. On average the firm produces 1,200 lamps monthly. Production lead time is 18 days, and the firm plans to have 15 percent buffer stock. Assume 20 working days per month.
    a. If each container holds 15 lamps, what will be the total number of kanban tickets required? (Use the formula given in this section.)
    b. What is the maximum WIP inventory the company can expect to have using this system?
    c. Suppose that each lamp costs the firm $30 to produce. If carrying costs are based on a 20 percent annual interest rate, what annual carrying cost for WIP inventory is the company incurring? (You may wish to review the discussion of holding costs in Chapter 4.)
38. What is SMED? In what way can die-changing operations be reduced by orders of magnitude? Why is it an advantage to do so?
39. Explain how the dual-card kanban system operates.

40. Discuss the advantages and disadvantages of each of the following features of just-in-time:
    a. Small lot sizes.
    b. Integrated purchasing and inventory control.
    c. Kanban information system.

8.7 A COMPARISON OF MRP AND JIT

MRP and JIT are fundamentally different systems for manufacturing control. As we noted earlier, MRP is a push system and JIT is a pull system. JIT is a reactive system. If a problem develops and the line is shut down, JIT reacts immediately, since requests for new material are discontinued. In this way, one might say that JIT reacts to uncertainties and MRP does not. However, JIT is clearly not going to work well when it is known that demands will vary significantly over time. MRP builds this information into the planning structure while JIT does not.

For most manufacturing environments, implementing a pure JIT system is simply not feasible. Suppliers may not be located in close enough proximity to allow inputs to be delivered according to a rigid schedule. Demands for products may be highly variable, making it impractical to ignore this information in the planning process. It may be difficult to move products in small batches. Implementing SMED may not be possible in some environments. When setup costs are very high, it makes economic sense to produce in large lots and store items rather than change over production processes frequently. With that said, however, major reductions in WIP inventories can be achieved in the vast majority of traditional manufacturing plants. Plants that run lean run better.

Toyota's enormous success in reducing inventory-related costs while producing high-quality products with high throughput rates has formed the basis for much of the support for JIT-based systems. However, it is unclear if JIT is primarily responsible for Toyota's success. Toyota's way of doing business differs from that of American auto makers in many dimensions. Is its success a direct result of JIT methods, or can it be attributed to other factors that might be more difficult to emulate? In order to determine under what circumstances JIT would be most advantageous, Krajewski et al. (1987) developed a large-scale simulator to compare JIT, MRP, and ROP (reorder point) manufacturing environments. Their comparison included 36 distinct factors aggregated into eight major categories, with each factor varied from one to five levels. The eight categories considered and a brief summary of the factors in these categories are listed here:

1. *Customer influence.* Demand forecast error.
2. *Vendor influence.* Vendor reliability (order size received compared to order size requested, and time received compared to time requested).
3. *Buffer mechanisms.* Capacity buffer and safety stock and safety lead times.
4. *Product structure.* Considered pyramid (few end items) and inverted pyramid (many end items) structures.
5. *Facility design.* Routing patterns, shop emphasis, and length of routings.
6. *Process.* Scrap rates, equipment failures, worker flexibility, and capacity imbalances.
7. *Inventory.* Reporting accuracy, lot-sizing rules, and setup times.
8. *Other factors.* Number of items, number of workstations, and several other factors.

The results obtained were very interesting. JIT worked well only in favorable manufacturing environments. This means little or no demand variability, reliable vendors, and small setup times for production. JIT showed poor performance when one or more of these factors became unfavorable. In fact, in a favorable environment both ROP and MRP gave good results as well. This suggests that the primary benefit comes from creating a favorable manufacturing environment, as opposed to simply implementing a JIT system. Perhaps greater benefit would result from carefully evaluating and (where possible) rectifying the key problems affecting manufacturing than by blindly implementing a new production planning system. Alternatively, perhaps it is the *process* of implementing a JIT system from which the greatest benefit is obtained.

New approaches to manufacturing control are quickly tagged with three-letter acronyms and proselytized with the zealousness of a religion. JIT is no exception. At its best, JIT is a set of methods that should be implemented on a continuous improvement basis for reducing inventories at every level of the supply chain. At its worst, JIT is a romantic idea that when applied blindly can be very damaging to worker morale, relationships with suppliers, and ultimately the bottom line. Inventory savings can often be an illusion, as illustrated by Chhikara and Weiss (1995). They show in three case studies that if inventory reductions are not tied to accounting systems and cash flows, inventory reductions do not translate to reduced inventory costs. JIT for its own sake does not make sense. It must be carefully integrated into the entire supply and manufacturing chain in order to realize its benefits.

Zipkin (1991) offers a very thoughtful piece on the "JIT revolution." His main point is that JIT can have real benefits, but if the pragmatism is not distinguished from the romance, the consequence could be disaster. He cites an example of a manager of a computer assembly plant who was ordered to reduce his WIP inventory to zero in six months. This was the result, apparently, of the experiences of the CFO, who had attended a JIT seminar that inspired him to promote massive inventory reductions.

Finally, as noted by Karmarkar (1989), the issue is not to make a choice between MRP and JIT, but to make the best use of both techniques. JIT reacts very slowly to sudden changes in demand, while MRP incorporates demand forecasts into the plan. Does that mean that Toyota, famous for their JIT system, ignores demand forecasts in their manufacturing planning and control? Very unlikely. Understanding what different methodologies offer and their limitations leads to a manufacturing planning and control system that is well designed and efficient. Improvements should be incremental, not revolutionary.

8.8 JIT OR LEAN PRODUCTION?

In recent years the term *lean production* has become commonplace. Is lean production just another term for JIT? Yes and no. Lean production has come to encompass more than JIT, but the goal is the same, namely, to reduce work-in-process inventories to a bare minimum. The term lean production seems to be due to Womack et al. (1990) who used it to describe the Toyota Production System.

One might wonder why this chapter is about JIT and not lean production, considering that lean production (or lean manufacturing) seems to be the favored term these days. JIT is a set of principles for moving materiel through the factory that can be compared directly to MRP. Lean production encompasses all of the concepts of JIT elaborated on in this chapter, but has also been linked to six sigma quality programs (discussed in

Chapter 12), cellular manufacturing systems (discussed in Chapter 11), the focused factory (discussed in Chapter 1), and total productive maintenance (discussed in Chapter 12). Hence, lean production encompasses topics that are treated in depth throughout the book, in addition to JIT principles.

When one reads descriptions of lean production systems from practitioners [such as *The Portal to Lean Production* by Nicholas and Soni (2006)], one is struck by a few things. First, practitioners view lean production in a very broad sense as noted above. Second, there appear to be many success stories of lean production concepts implemented in the United States. We know that the EOQ and EPQ formulas developed in Chapter 4 can be used to determine appropriate run sizes in a factory. Are not these concepts fundamentally at odds with those of lean production? The answer is yes, they are. Run sizes recommended by these formulas could be large depending on costs and usage rates, which is verboten in a lean production system. Which is the better approach?

The answer is that it depends. If the objective is to set run sizes to balance holding and setup costs, the models of Chapter 4 (and Chapter 5 when uncertainty is included) are fine. However, there are many costs and benefits that these models ignore because they are difficult to quantify. How does one quantify the cost of having to rework a large batch of items because a setting on a machine was wrong? How does one quantify the chaos that results from having large stockpiles of work-in-process inventory all over the plant? Thus, many of the benefits of lean production and JIT are hard to incorporate into a model. Simple economic trade-offs tell only a small part of the story. This suggests that modeling the true benefits of lean production is an area of opportunity for researchers. When we have models that take into account all of the costs and benefits of these disparate approaches to running the factory, we can make more intelligent comparisons among different production planning philosophies.

8.9 HISTORICAL NOTES

The specific term *MRP* is relatively new, although the concept of materials planning based on predicted demand and product structures appears to have been around for quite a while. In fact, there is a reference to this approach by Alfred Sloan (1964), who refers to a purely technical calculation by analysts at General Motors in 1921 of the quantity of materials needed for production of a given number of cars. The term *BOM* (bill of materials) *explosion* was commonly used to describe what is now called MRP.

The books by Orlicky (1975) and New (1974) did a great deal to legitimize MRP as a valid and identifiable technique, although the term seems to have been around since the mid-1960s. In addition, the well-known practitioners George Plossl and Oliver Wight also must be given credit for popularizing the method (see, for example, Plossl and Wight, 1971). In Anderson, Schroeder, Tupy, and White (1982) the authors state that the first computerized MRP systems were implemented close to 1970. The number of installed systems has increased since that time at an exponential rate.

Interestingly, much of the work on optimal and suboptimal lot-sizing methods predates the formal recognition of MRP. The seminal paper in this area was by Wagner and Whitin (1958), who first recognized the optimality of an exact requirements policy for periodic-review inventory control systems with time-varying demand, and developed the dynamic programming algorithm described in this chapter. The

Silver–Meal heuristic when opportunities for replenishment are at the start of periods appeared in Silver and Meal (1973). DeMatteis (1968) is generally credited with the part period balancing approach. However, the paper by Gorham (1968) refers to both part period balancing (called the least total cost approach) and least unit cost methods, suggesting that these methods were well known at the time. It is likely that both methods were developed by practitioners before 1968 but were not reported in the literature.

The lot-shifting algorithm for the capacitated problem outlined in Section 8.4 is very similar to one developed by Karni (1981). Similar ideas were explored by Dixon and Silver (1981) as well. Lot sizing is not always used by practitioners in operational MRP systems because of the effect that small errors at a high level of the product structure are telescoped into large errors at a lower level. Furthermore, lot-sizing algorithms require estimation of the holding and setup costs and more calculations than the simple lot-for-lot policy.

The historical references to JIT are contained within the chapter. It is unclear who coined the term JIT, but the concept is clearly derived from Toyota's kanban system. SMED (single minute exchange of dies) has played an important role in the success of the Japanese lean production methods. Shigeo Shingo is generally credited with its development.

8.10 Summary

Materials requirements planning (MRP) is a set of procedures for converting forecast demand for a manufactured product into a requirements schedule for the components, subassemblies, and raw materials comprising that product. A closely related concept is that of the *master production schedule* (MPS), which is a specification of the projected needs of the end product by time period. The *explosion calculus* represents the set of rules and procedures for converting the MPS into the requirements at lower levels. The information required to do the explosion calculus is contained in the *product structure diagram* and the *indented bill-of-materials list.* The two key pieces of information contained in the product structure diagram are the production lead times needed to produce the specific component and the multiplier giving the number of units of the component required to produce one item at the next higher level of the product structure.

Many MRP systems are based on a *lot-for-lot* production schedule. That is, the number of units of a component produced in a period is the same as the requirements for that component in that period. However, if setup and holding costs can be estimated accurately, it is possible to find other lot-sizing rules that are more economical. The optimal lot-sizing procedure is the Wagner–Whitin algorithm. However, the method is rarely used in practice, largely because of the relative complexity of the calculations required (although such calculations take very little time on a computer). We explored three heuristic methods that require fewer calculations than the Wagner–Whitin algorithm, although none of these methods will necessarily give an optimal solution. They are the *Silver–Meal heuristic,* the *least unit cost heuristic,* and *part period balancing.*

We also treated the dynamic lot-sizing problem when capacity constraints exist. One of the limitations of MRP is that capacities are ignored. This is especially important if lot sizing is incorporated into the system. Finding optimal solutions to a capacity-constrained inventory system subject to time-varying demand is an extremely difficult problem. (For those of you familiar with the term, the problem is said to be NP complete, which is a reference to the level of difficulty.) A straightforward heuristic method for obtaining a solution to the capacitated lot-sizing problem was presented. However, incorporating such a method into the MRP system will, in and of itself, *not* solve the complete

capacitated MRP problem, because even if a particular lot-sizing schedule is feasible at one level, there is no guarantee that it will result in a feasible requirements schedule at a lower level.

Truly optimal lot-sizing solutions for an MRP system would require formulating the problem as an integer program in order to determine the optimal decisions for all levels simultaneously. For real assembly systems, which can be as many as 10 to 15 levels deep, this would result in an enormous mathematical programming problem. In view of the host of other issues concerned with implementing MRP, the marginal benefit one might achieve from multilevel optimization would probably not justify the effort involved.

System nervousness is one problem that arises when implementing an MRP system. The term refers to the unanticipated changes in a schedule that result when the planning horizon is rolled forward by one period. Another difficulty is that in many circumstances, production lead times depend on the lot sizes: MRP assumes that production lead times are fixed. Still another problem is that the yields at various levels of the process may not be perfect. If the yield rates can be accurately estimated in advance, these rates can be factored into the calculations in a straightforward manner. However, in many industries the yield rates may be difficult to estimate in advance.

Manufacturing resource planning (MRP II) attempts to deal with some of the problems of implementing MRP by integrating the financial, accounting, and marketing functions into the production-planning function.

In this chapter we also discussed just-in-time. Just-in-time is an inventory control system whose goal is to reduce work-in-process to a bare minimum. The concepts are based on the production control systems used by Toyota in Japan in the 1970s. JIT is a pull system, whereas MRP is a push system. Parts are transferred from one level to the next only when requested. In addition to reduced inventories, this approach allows workers to quickly locate quality control problems. Because parts are moved through the system in small batches, defects can be identified quickly. By the nature of the system, stopping production at one location automatically stops production on the entire line so that the source of the problem can be identified and corrected before defective inventories build up.

An important aspect of JIT is reduction of production lot sizes. To make small lot sizes economical, it is necessary to reduce fixed costs of changeover. This is the goal of single minute exchange of dies (SMED). By dividing the die- or tool-changing operation into portions that can be done off-line and those that must be done on-line, enormous reductions in setup times can be achieved.

Kanban and JIT are closely related. Kanban is a manual information system for implementing JIT that relies on cards or tickets to signal the need for more product. While not very high-tech, the method has worked well in practice.

JIT has several disadvantages vis-à-vis MRP as well as several advantages. JIT will react more quickly if a problem develops. However, JIT systems are very slow to react to changes in the pattern of demand. MRP, on the other hand, builds forecasts directly into the explosion calculus.

Additional Problems for Chapter 8

41. CalcIt produces a line of inexpensive pocket calculators. One model, IT53, is a solar-powered scientific model with a liquid crystal display (LCD). The calculator is pictured in Figure 8–11.

    Each calculator requires four solar cells, 40 buttons, one LCD display, and one main processor. All parts are ordered from outside suppliers, but final assembly is done

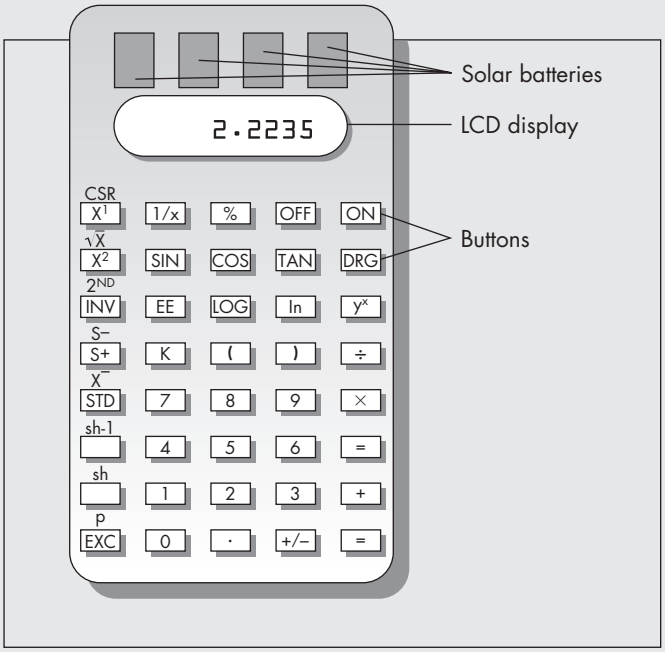


FIGURE 8–11
CalcIt Model IT53 Scientific Calculator (Problem 41)

by CalcIt. The processors must be in stock three weeks before the anticipated completion date of a batch of calculators to allow enough time to set the processor in the casing, connect the appropriate wiring, and allow the setting paste to dry. The buttons must be in stock two weeks in advance and are set by hand into the calculators. The LCD displays and the solar cells are ordered from the same supplier and need to be in stock one week in advance.

Based on firm orders that CalcIt has obtained, the master production schedule for IT53 for a 10-week period starting at week 8 is given by

Week	8	9	10	11	12	13	14	15	16	17
MPS	1,200	1,200	800	1,000	1,000	300	2,200	1,400	1,800	600

Determine the gross requirements schedule for the solar cells, the buttons, the LCD display, and the main processor chips.

42. Consider the example of the CalcIt Company for Problem 41. Suppose that the buttons used in the calculators cost \$0.02 each and the company estimates a fixed cost of \$12 for placing and receiving orders of the buttons from an outside supplier. Assume that holding costs are based on a 24 percent annual interest rate and that there are 48 weeks to a year. Using the gross requirements schedule for the buttons determined in Problem 41, what order policy does the Silver–Meal heuristic recommend for the buttons? (Hint: Express h as a cost per 10,000 units and divide each demand by 10,000.)
43. Solve Problem 42 using part period balancing and least unit cost. Compare the costs of the resulting solutions to the cost of the solution obtained by using the Silver–Meal heuristic.
44. *Work-in-process* (WIP) *inventory* is a term that refers to the inventory of components and subassemblies in a manufacturing process. Assuming lot-for-lot scheduling at all

levels, theoretically the WIP inventory should be zero. Do you think that this is likely to be the case in a real manufacturing environment? Discuss the possible reasons for large WIP inventories occurring even when an MRP system is used.

45. Vivian Lowe is planning a surprise party for her husband's 50th birthday. She has decided to serve shish kabob. The recipe that she is using calls for two pineapple chunks for each shrimp. She plans to size the kabobs so that each has three shrimp. She estimates that a single pineapple will yield about 50 chunks, but from past experience, about 1 out of every 10 pineapples is bad and has to be thrown out. She has invited 200 people and expects that about half will show up. Each person generally eats about 2 kabobs.
    a. How many pineapples should she buy?
    b. Suppose that the number of guests is a random variable having the normal distribution with mean 100 and variance 1,680. If she wants to make enough kabobs to feed all the guests with probability 95 percent, how many pineapples should she buy?
46. In this chapter, we assumed that the "time bucket" was one week. This implies that forecasts are reevaluated and the MRP system rerun on a weekly basis.
    a. Discuss the potential advantages of using a shorter time bucket, such as a day.
    b. Discuss the potential advantages of using a longer time bucket, such as two weeks or a month.
47. Develop a spreadsheet that reproduces the calculations for Example 8.1. As in the example, the spreadsheet should include the net predicted demand for trumpets. The columns should correspond to weeks and should be labeled 1 to 18. Below the net predicted demand for trumpets should be the calculations for the valve casing assembly, and below that the calculations for the valves. For each component, include rows for the following information: (1) gross requirements, (2) scheduled receipts, (3) on-hand inventory, (4) time-phased net requirements, and (5) lot-for-lot planned order release. Your spreadsheet should automatically update all calculations if the net predicted demand for trumpets changes.
48. Two end products, EP1 and EP2, are produced in the Raleigh, North Carolina, plant of a large manufacturer of furniture products located in the Southeast. The product structure diagrams for these products appear in Figure 8–12.

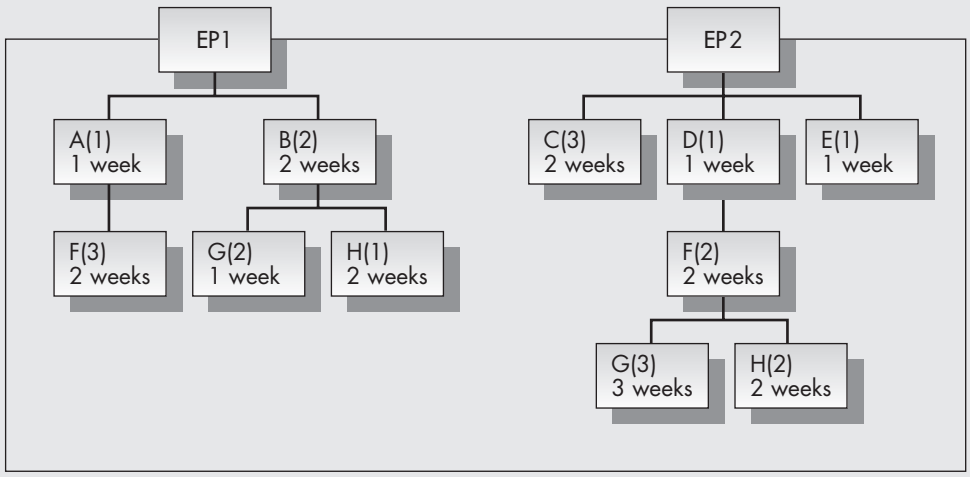


FIGURE 8–12
Product structure diagrams (for Problem 48)

Suppose that the master production schedules for these two products are

Week	18	19	20	21	22	23	24
EP1	120	112	76	22	56	90	210
EP2	62	68	90	77	26	30	54

Assuming lot-for-lot production, determine the planned order releases for components F, G, and H.

49. A component used in a manufacturing facility is ordered from an outside supplier. Because the component is used in a variety of end products, the demand is high. Estimated demand (in thousands) over the next 10 weeks is

Week	1	2	3	4	5	6	7	8	9	10
Demand	22	34	32	12	8	44	54	16	76	30

The components cost 65 cents each and the interest rate used to compute the holding cost is 0.5 percent per week. The fixed order cost is estimated to be \$200. (Hint: Express h as the holding cost per thousand units.)

a. What ordering policy is recommended by the Silver–Meal heuristic?

b. What ordering policy is recommended by the part period balancing heuristic?

c. What ordering policy is recommended by the least unit cost heuristic?

d. Which method resulted in the lowest-cost policy for this problem?

50. A popular heuristic lot-sizing method is known as period order quantity (POQ). The method requires determining the average number of periods spanned by the EOQ and choosing the lot size to equal this fixed period supply. Let λ be the total demand over an N-period planning horizon [$\lambda = (\Sigma\, r_i/n)$] and assume that EOQ is computed as described in Section 8.2. Then $P = \text{EOQ}/\lambda$ rounded to the nearest integer. For the example in Section 8.2, $P = 139/43.9 = 3.17$, which is rounded to 3. The POQ would call for setting the lot size equal to three periods of demand. For the example in Section 8.2, the resulting planned order release would be 116, 0, 0, 150, 0, 0, 135, 0, 0, 38.

a. Compare the cost of the policy obtained by this method for the example in Section 8.2 to that obtained using the EOQ.

b. What are the advantages of this approach over EOQ?

c. Do you think that this method will be more cost effective in general than the heuristic methods discussed in Section 8.2?

d. Solve Problem 17 using this method, and compare the total holding and setup cost with that obtained by the other methods.

e. Solve Problem 49 using this method, and compare the total holding and setup cost with that obtained by the other methods.

51. The campus store at a large Midwestern university sells writing tablets to students, faculty, and staff. They sell more tablets near exam time. During a typical 10-week quarter, the pattern of sales is 2,280, 1,120, 360, 3,340, 1,230, 860, 675, 1,350, 4,600, 1,210. The pads cost the store \$1.20 each, and holding costs are based on a 30 percent annual interest rate. The cost of employee time, paperwork, and handling amounts to \$30 per order. Assume 50 weeks per year.

a. What is the optimal order policy during the 10-week quarter based on the Silver–Meal heuristic? Using this policy, what are the total holding and ordering costs incurred over the 10-week period?

b. The bookstore manager has decided that it would be more economical if the demand were the same each week. In order to even out the demand he limits weekly sales (to the annoyance of his clientele). Hence, assume that the total demand for the 10-week quarter is still the same, but the sales are constant from week to week. Determine the optimal order policy in this case and compare the total holding and ordering cost over the 10 weeks to the answer you obtained in part (*a*). (You may assume continuous time for the purposes of your calculations, so that the optimal lot size is the EOQ.)

c. Based on the results of parts (*a*) and (*b*), do you think that it is more economical in general to face a smooth or a spiky demand pattern?

52. Develop a spreadsheet template for finding Silver–Meal solutions for general lot-sizing problems. Store the holding and setup cost parameters in separate cell locations so that they can be inputted and changed easily. Allow for 30 periods of demand to be inputted in column 2 of the spreadsheet. List the period numbers 1, 2, . . . , 30 in column 1. Work out the logic that gives $C(j)$ in column 3.

    One would use such a spreadsheet in the following way: Input requirements period by period until an increase is observed in column 3. This identifies the first forecast horizon. Now replace entries in column 2 with zeros and input requirements starting at the current period. Continue until the next forecast horizon is identified. Repeat this process until the end of the planning horizon. Use this method to find the Silver–Meal solution for the following production planning problems:

    a. Solve Problem 14 in this manner.

    b. Weekly demands for 2-inch rivets at a division of an aircraft company are predicted to be (in gross)

Week	1	2	3	4	5	6	7	8	9	10	11	12
Demand	240	280	370	880	950	120	135	450	875	500	400	200
Week	13	14	15	16	17	18	19	20	21	22	23	24
Demand	600	650	1,250	250	800	700	750	200	100	900	400	700

    Setup costs for ordering the rivets are estimated to be $200, and holding costs amount to 10 cents per gross per week. Find the lot sizing given by the Silver–Meal method.

53. Along the lines described in Problem 52, construct a spreadsheet for finding the least unit cost lot-sizing rule.

    a. Solve both parts (*a*) and (*b*) of Problem 52.

    b. Which method, least unit cost or Silver–Meal, gave the more cost-effective solution?

Appendix **8–A**

Optimal Lot Sizing for Time-Varying Demand

The techniques considered in Section 8.3 are easy to use and give lot sizes with costs that are generally near the true optimal. This appendix considers how one would go about computing true optimal lot sizes. Optimal in this context means the policy that minimizes the total holding and setup cost over the planning horizon. This appendix shows how optimal policies can

be determined by casting the problem as a shortest-path problem. It also shows how dynamic programming can be used to find the shortest path.

Assume that:

1. Forecasted demands over the next n periods are known and given by the vector $\mathbf{r} = (r_1, \ldots, r_n)$.
2. Costs are charged against holding at \$$h$ per unit per period and \$$K$ per setup. We will assume that the holding cost is charged against ending inventory each period.

In order to get some idea of the potential difficulty of this problem, consider the following simple example.

Example *8A.1*

The forecast demand for an electronic assembly produced at Hi-Tech, a local semiconductor fabrication shop, over the next four weeks is 52, 87, 23, 56. There is only one setup each week for production of these assemblies, and there is no back-ordering of excess demand. Assume that the shop has the capacity to produce any number of the assemblies in a week.

Consider the total number of feasible production policies for Hi-Tech over the four-week period. Let $y_1, \ldots, y_4$ be the order quantities in each of the four weeks. Clearly $y_1 \geq 52$ in order to guarantee that we do not stock out in period 1. If we assume that ending inventory in period 4 is zero (which will be easy to show is optimal), then $y_1 \leq 218$, the sum of all the demands. Hence y_1 can take any one of 167 possible values. Consider y_2. The number of feasible values of y_2 depends upon the value of y_1. As no stock-out is permitted to occur in period 2, $y_1 + y_2 \geq 52 + 87 = 139$. If

$$y_1 \leq 139, \qquad \text{then } 139 - y_1 \leq y_2 \leq 218 - y_1,$$

and if

$$y_1 > 139, \qquad \text{then } 0 \leq y_2 \leq 218 - y_1.$$

With a little effort, one can show that this results in a total of 10,200 different values of just the pair (y_1, y_2). It is thus clear that for even moderately sized problems the number of feasible solutions is enormous.

Searching all the feasible policies is unreasonable. However, an important discovery by Wagner and Whitin reduces considerably the number of policies one must consider as candidates for optimality.

The Wagner–Whitin algorithm is based on the following observation:

Result. An optimal policy has the property that each value of $\mathbf{y}$ is exactly the sum of a set of future demands. (We will call this an exact requirements policy.) That is,

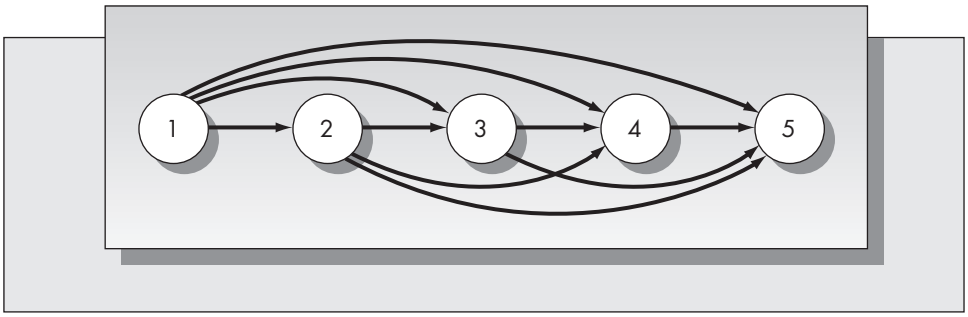
$$\begin{aligned}
&y_1 = r_1, \qquad \text{or } y_1 = r_1 + r_2, \ldots, \qquad \text{or } y_1 = r_1 + r_2 + \cdots + r_n.\\
&y_2 = 0, \qquad \text{or } y_2 = r_2, \qquad \text{or } y_2 = r_2 + r_3, \ldots, \qquad \text{or}\\
&y_2 = r_2 + r_3 + \cdots + r_n\\
&\quad \vdots\\
&y_n = 0 \qquad \text{or } y_n = r_n.
\end{aligned}$$

An exact requirements policy is completely specified by designating the periods in which ordering is to take place. The number of exact requirements policies is much smaller than the total number of feasible policies.

Example *8A.1 (continued)*

We continue with the four-period scheduling problem. Because y_1 must satisfy exact requirements, we see that it can assume values of 52, 139, 162, or 218 only; that is, only four distinct values. Ignoring the value of y_1, y_2 can assume values 0, 87, 110, 166. It is easy to see that every exact requirements policy is completely determined by specifying in what periods ordering should

FIGURE 8A–1
Network representation for lot scheduling (Example 8A.1)



take place. That is, each such policy is the form $(i_1, \ldots, i_n)$, where the values of i_j are either 0 or 1. If $i_j = 1$, then production takes place in period j. Note that $i_1 = 1$ because we must produce in period 1 to avoid stocking out, whereas $i_2, \ldots, i_n$ will each be either 0 or 1. For example, the policy (1, 0, 1, 0) means that production occurs in periods 1 and 3 only. It follows that $\mathbf{y} =$ (139, 0, 79, 0). For this example, there are exactly $2^3 = 8$ distinct exact requirements policies.

A convenient way to look at the problem is as a one-way network with the number of nodes equal to exactly one more than the number of periods. Every path through the network corresponds to a specific exact requirements policy. The network for the four-period problem appears in Figure 8A–1.

For any pair (i, j) with $i < j$, if the arc (i, j) is on the path, it means that ordering takes place in period i and the order size is equal to the sum of the requirements in periods $i, i + 1, \ldots, j - 1$. Period j is the next period of ordering. Note that all paths end at period $n + 1$. The policy of ordering in periods 1 and 3 only would correspond to the path 1–3–5. The path 1–2–4–5 means that ordering is to take place in periods 1, 2, and 4.

The next step is to assign a value to each arc in the network. The value or "length" of the arc (i, j), called c_{ij}, is defined as the setup and holding cost of ordering in period i to meet requirements through period $j - 1$. For example, c_{15} = the cost of ordering in period 1 to satisfy the demands in periods 1 through 4.

Finally, we would like to determine the minimum-cost production schedule, or shortest path through the network. As we will see, *dynamic programming* is one method of solving this problem. However, for a small problem, the optimal policy can be found by simply enumerating the paths through the network and choosing one with minimum cost.

Example *8A.2*

We will solve Example 8A.1 using path enumeration. Recall that $\mathbf{r} = (52, 87, 23, 56)$. In addition, assume that there is a cost of holding of $h = \$1$ per unit per period and a cost of $K = \$75$ per setup.

The first step is to compute c_{ij} for $1 \le i \le 4$ and $i + 1 \le j \le 5$.

$c_{12} = 75$ (setup cost only).

$c_{13} = 75 + 87 = 162$.

$c_{14} = 75 + (23 \times 2) + 87 = 208$.

$c_{15} = 75 + (56 \times 3) + (23 \times 2) + 87 = 376$.

$c_{23} = 75$.

$c_{24} = 75 + 23 = 98$.

$c_{25} = 75 + 23 + (56 \times 2) = 210$.

$c_{34} = 75$.

$c_{35} = 75 + 56 = 131$.

$c_{45} = 75$.

Summarizing these costs in matrix form gives the following:

i \ *j*	1	2	3	4	5
1		75	162	208	376
2			75	98	210
3				75	131
4					75

As there are only eight exact requirements policies, we can solve this problem by enumerating the policies and comparing the costs.

Path	Cost
1–2–3–4–5	$300
1–2–4–5	248
1–2–5	285
1–2–3–5	281
1–3–4–5	312
1–3–5	293
1–4–5	283
1–5	376

It follows that the optimal path is 1–2–4–5 at a cost of $248. This corresponds to ordering in periods 1, 2, and 4 only. The optimal ordering policy is $y_1 = 52$, $y_2 = 110$, $y_3 = 0$, $y_4 = 56$.

*SOLUTION BY DYNAMIC PROGRAMMING

The total number of exact requirements policies for a problem of n periods is 2^{n-1}. As n gets large, total enumeration is not efficient. Dynamic programming is a recursive solution technique that can significantly reduce the number of computations required, although it too can be quite cumbersome.

Dynamic programming is based on the *principle of optimality.* One version of this principle is that if a problem consists of exactly n stages and there are $r < n$ stages remaining, the optimal policy for the remaining stages is independent of policy adopted in the previous stages. Because dynamic programming is not used anywhere else in this text, we will not present a detailed discussion of it. The interested reader should refer to Hillier and Lieberman (1990) for a brief overview or Nemhauser (1966) for a more in-depth treatment at a mathematical level consistent with ours.

Define f_k as the minimum cost starting at node k, assuming that an order is placed in period k. The principle of optimality for this problem results in the following system of equations:

$$f_k = \min_{j>k} (c_{kj} + f_j) \qquad \text{for } k = 1, \ldots, n.$$

The initial condition is $f_{n+1} = 0$.

Example *8A.3* We will solve Example 8A.1 by dynamic programming in order to illustrate the technique. One starts with the initial condition and works backward from period $n + 1$ to period 1.[4] In each period one determines the value of j that achieves the minimum.

[4] Actually, the original Wagner–Whitin (1958) algorithm is based on a forward dynamic programming formulation. Although the forward formulation has some advantages for planning horizon analysis, we feel that backward recursion is more natural and more intuitive.

$f_5 = 0.$

$f_4 = \min_{j>4}(c_{4j} + f_j)$

$= 75 \qquad$ at $j = 5$ (the only possible value of j).

$$f_3 = \min_{j>3}(c_{3j} + f_j) = \min\begin{Bmatrix} c_{34} + f_4 \\ c_{35} + f_5 \end{Bmatrix} = \min\begin{Bmatrix} 75 + 75 \\ 131 + 0 \end{Bmatrix} = \min\begin{Bmatrix} 150 \\ 131 \end{Bmatrix}$$

$= 131 \qquad$ at $j = 5$.

$$f_2 = \min_{j>2}(c_{2j} + f_j) = \min\begin{Bmatrix} c_{23} + f_3 \\ c_{24} + f_4 \\ c_{25} + f_5 \end{Bmatrix} = \min\begin{Bmatrix} 75 + 131 \\ 98 + 75 \\ 210 + 0 \end{Bmatrix} = \min\begin{Bmatrix} 206 \\ 173 \\ 210 \end{Bmatrix}$$

$= 173 \qquad$ at $j = 4$.

Finally,

$$f_1 = \min_{j>1}(c_{1j} + f_j) = \min\begin{Bmatrix} c_{12} + f_2 \\ c_{13} + f_3 \\ c_{14} + f_4 \\ c_{15} + f_5 \end{Bmatrix} = \min\begin{Bmatrix} 75 + 173 \\ 162 + 131 \\ 208 + 75 \\ 376 + 0 \end{Bmatrix} = \min\begin{Bmatrix} 248 \\ 293 \\ 283 \\ 376 \end{Bmatrix}$$

$= 248 \qquad$ at $j = 2$.

To determine the optimal order policy, we retrace the solution back from the beginning. In period 1 the optimal value of j is $j = 2$. This means that the production level in period 1 is equal to the demand in period 1, so that $y_1 = r_1 = 52$. The next order period is period 2. The optimal value of j in period 2 is $j = 4$, which implies that the production quantity in period 2 is equal to the sum of the demands in periods 2 and 3, or $y_2 = r_2 + r_3 = 110$. The next period of ordering is period 4. The optimal value of j in period 4 is $j = 5$. This gives $y_4 = r_4 = 56$. Hence, the optimal order policy is $\mathbf{y} = (52, 110, 0, 56)$.

Appendix **8–B**

Glossary of Notation for Chapter 8

$C(T)$ = Average holding and setup cost per period (for Silver–Meal heuristic) or per unit (LUC heuristic) if the current order spans T periods.

c_i = Production capacity in period i.

c_{ij} = Cost associated with arc (i, j) in network representation of lot scheduling problem used for Wagner–Whitin algorithm.

f_j = Minimum cost from period i to the end of the horizon (refer to the dynamic programming algorithm for Wagner–Whitin).

h = Holding cost per unit per time period.

K = Setup cost for initiating an order.

r_i = Requirement for period i.

y_i = Production lot size in period i.

Bibliography

Anderson, J. C.; R. G. Schroeder; S. E. Tupy; and E. M. White. "Material Requirements Planning Systems: The State of the Art." *Production and Inventory Management* 23 (1982), pp. 51–66.

Carlson, R. C.; S. L. Beckman; and D. H. Kropp. "The Effectiveness of Extending the Horizon in Rolling Production Scheduling." *Decision Sciences* 13 (1982), pp. 129–46.

Carlson, R. C.; J. V. Jucker; and D. H. Kropp. "Less Nervous MRP Systems: A Dynamic Economic Lot-Sizing Approach." *Management Science* 25 (1979), pp. 754–61.

Carlson, R. C.; J. V. Jucker; and D. H. Kropp. "Heuristic Lot Sizing Approaches for Dealing with MRP System Nervousness." *Decision Sciences* 14 (1983), pp. 156–69.

Chhikara, J., and E. N. Weiss. "JIT Savings—Myth or Reality?" *Business Horizons* 38 (May–June 1995), pp. 73–78.

DeMatteis, J. J. "An Economic Lot Sizing Technique: The Part-Period Algorithm." *IBM Systems Journal* 7 (1968), pp. 30–38.

Dixon, P. S., and E. A. Silver. "A Heuristic Solution Procedure for the Multi-Item, Single-Level, Limited Capacity Lot Sizing Problem." *Journal of Operations Management* 2 (1981), pp. 23–39.

Gorham, T. "Dynamic Order Quantities." *Production and Inventory Management* 9 (1968), pp. 75–81.

Hillier, F. S., and G. J. Lieberman. *Introduction to Operations Research.* 5th ed. New York: McGraw-Hill, 1990.

Hopp, W., and M. Spearman. *Factory Physics.* Burr Ridge, IL: Richard D. Irwin, 1996.

Jacobson, G., and J. Hillkirk. *Xerox, American Samurai.* New York: MacMillan, 1986.

Karmarkar, U. "Getting Control of Just-In-Time." *Harvard Business Review* 67 (September–October 1989), pp. 122–31.

Karni, R. "Maximum Part Period Gain (MPG)—A Lot Sizing Procedure for Unconstrained and Constrained Requirements Planning Systems." *Production and Inventory Management* 22 (1981), pp. 91–98.

Krajewski, L. J.; B. E. King; L. P. Ritzman; and D. S. Wong. "Kanban, MRP, and Shaping the Manufacturing Environment." *Management Science* 33 (1987), pp. 39–57.

Love, Stephen. *Inventory Control.* New York: McGraw-Hill, 1979.

Meal, H. "Putting Production Decisions Where They Belong." *Harvard Business Review* 62 (1984), pp. 102–11.

Monden, Y. "What Makes the Toyota Production System Really Tick?" *Industrial Engineering* 13, no. 1 (1981a), pp. 36–46.

Monden, Y. "Adaptable Kanban System Helps Toyota Maintain Just-in-Time Production." *Industrial Engineering* 13, no. 5 (1981b), pp. 28–46.

Nahmias, S., and K. Moinzadeh. "Lot Sizing with Randomly Graded Yields." *Operations Research* 46, no. 6 (1997), pp. 974–86.

Nemhauser, G. L. *Introduction to Dynamic Programming.* New York: John Wiley & Sons, 1966.

New, C. *Requirements Planning.* Essex, England: Gower Press, 1974.

Nicholas, J., and A. Soni. *The Portal to Lean Production.* Boca Raton, Auerbach Publications, 2006.

Orlicky, J. *Materials Requirements Planning.* New York: McGraw-Hill, 1975.

Plossl, G., and O. Wight. *Materials Requirements Planning by Computer.* Washington, DC: American Production and Inventory Control Society, 1971.

Sheldon, D. "MRP II Implementation: A Case Study." *Hospital Materiel Management Quarterly* 15, no. 4 (1994), pp. 48–52.

Shingo, S. *Study of "Toyota" Production System from Industrial Engineering Viewpoint.* Tokyo: Japan Management Association, 1981.

Silver, E. A., and H. C. Meal. "A Heuristic for Selecting Lot Size Quantities for the Case of a Deterministic Time-Varying Demand Rate and Discrete Opportunities for Replenishment." *Production and Inventory Management* 14 (1973), pp. 64–74.

Silver, E. A., and R. Peterson. *Decision Systems for Inventory Management and Production Planning.* 2nd ed. New York: John Wiley & Sons, 1985.

Sloan, A. *My Years with General Motors.* Garden City, NY: Doubleday, 1964.

Steele, D. C. "The Nervous MRP System: How to Do Battle." *Production and Inventory Management* 16 (1973), pp. 83–89.

Vollman, T. E.; W. L. Berry; and D. C. Whybark. *Manufacturing and Control Systems.* 3rd ed. New York: McGraw-Hill/Irwin, 1992.

Wagner, H. M., and T. M. Whitin. "Dynamic Version of the Economic Lot Size Model." *Management Science* 5 (1958), pp. 89–96.

Willis, R. "Harley Davidson Comes Roaring Back." *Management Review* 75 (March 1986), pp. 20–27.

Womack, J. P.; D. T. Jones; and D. Roos. *The Machine That Changed the World: The Story of Lean Production.* New York: Harper Perennial, 1990.

Zipkin, P. "Does Manufacturing Need a JIT Revolution?" *Harvard Business Review* 69 (January–February 1991), pp. 40–50.

Chapter **Nine**

Operations Scheduling

"I have an unbelievable assistant who handles all of my scheduling! It's like a Tetris game."
—Neil Patrick Harris

Chapter Overview

Purpose

To gain an understanding of the key methods and results for sequence scheduling in a job shop environment.

Key Points

1. *The job shop scheduling problem.* A job shop is a set of machines and workers who use the machines. Jobs may arrive all at once or randomly throughout the day. For example, consider an automotive repair facility. On any day, one cannot predict in advance exactly what kinds of repairs will come to the shop. Different jobs require different equipment and possibly different personnel. A senior mechanic might be assigned to a complex job, such as a transmission replacement, while a junior mechanic would be assigned to routine maintenance. Suppose the customers bring their cars in first thing in the morning. The shop foreman must determine the sequence in which to schedule the jobs in the shop to make the most efficient use of the resources (both human and machine) available.

   The relevant characteristics of the sequencing problem include
   - The pattern of arrivals.
   - Number and variety of machines.
   - Number and types of workers.
   - Patterns of job flow in the shop.
   - Objectives for evaluating alternative sequencing rules.
2. *Sequencing rules.* The sequencing rules that we consider in this section include.
   - *First come first served (FCFS).* Schedule jobs in the order they arrive to the shop.
   - *Shortest processing time (SPT) first.* Schedule the next job with the shortest processing time.
   - *Earliest due date (EDD).* Schedule the jobs that have the earliest due date first.
   - *Critical ratio (CR) scheduling.* The critical ratio is (due date – current time)/ processing time. Schedule the job with the smallest CR value next.
3. *Sequencing results.* A common criterion for evaluating the effectiveness of sequencing rules is the mean flow time. The flow time of any job is the amount of time that elapses from the point that the job arrives in the shop to the point

that the job is completed. The mean flow time is just the average of all the flow times for all the jobs. The main result of this section is that SPT scheduling minimizes the mean flow time. Another result of interest is that if the objective is to minimize the maximum lateness, then the jobs should be scheduled by EDD. This section also deals with several scheduling algorithms. Moore's algorithm minimizes the number of tardy jobs, and Lawler's algorithm is used when precedence constraints are present (that is, jobs must be done in a certain order).

All the preceding results apply to a single machine or single facility. When scheduling jobs on multiple machines, the problem is much more complex. In this case, there are a few known results. Consider the case of n jobs that must be scheduled on two machines. The main result discovered in this case is that the optimal solution is to sequence the jobs in the same order on both machines (this is known as a permutation schedule). This means that there are a possible $n!$ feasible solutions. This can, of course, be a very large number. However, a procedure discovered by Johnson (1954) efficiently computes the optimal sequence for n jobs on two machines. Essentially the same algorithm can be applied to three machines under very special circumstances. The problem of scheduling two jobs on m machines can be solved efficiently by a graphical procedure.

4. *Sequence scheduling in a stochastic environment.* The problems alluded to previously assume all information is known with certainty. Real problems are more complex in that there is generally some type of uncertainty present. One source of uncertainty could be the job times. In that case, the job times, say $t_1, t_2, \ldots, t_n$, are assumed to be independent random variables with α known distribution function. The optimal sequence for a single machine in this case is very much like scheduling the jobs in SPT order based on expected processing times.

   When scheduling jobs with uncertain processing times on multiple machines, one must assume that the distribution of job times follows an exponential distribution. The exponential distribution is the only one possessing the so-called memoryless property, which turns out to be crucial in the analysis. When the objective is to minimize the expected makespan (that is, the total time to complete all jobs), it turns out that the longest expected processing time (LEPT) first rule is optimal.

   Another source of uncertainty in a job shop is the order in which jobs arrive to the shop. In the automotive job shop example, we assumed that jobs arrived all at once at the beginning of the day. However, in a factory setting, jobs are likely to arrive at random times during the day. In this case, queueing theory can shed some light on how much time elapses from the point a job arrives until its completion. This section outlines several results under assumptions of FCFS, LCFS, and SPT sequencing.

5. *Line balancing.* Another problem that arises in the factory setting is that of balancing an assembly line. While line balancing is not a sequence scheduling problem found in a job shop environment, it is certainly a scheduling problem arising within the plant. Assume we have an item flowing down an assembly line and that a total of n tasks must be completed on the item. The problem is to determine which tasks should be placed where on the line. Typically, an assembly line is broken down into stations and some subset of tasks is assigned to each station. The goal is to balance the time required at each station while taking into account the precedence relationships existing among the individual tasks.

Optimal line balances are difficult to find. We consider one heuristic method, which gives reasonable results in most circumstances.

Scheduling is an important aspect of operations control in both manufacturing and service industries. With increased emphasis on time to market and time to volume as well as improved customer satisfaction, efficient scheduling will gain increasing emphasis in the operations function in the coming years.

In some sense, much of what has been discussed so far in this text can be considered a subset of production scheduling. Aggregate planning, treated in Chapter 3, is aimed at macroscheduling of workforce levels and overall production levels for the firm. Detailed inventory control, discussed in Chapters 4 and 5, concerns methods of scheduling production at the individual item level; Chapter 6 treated vehicle scheduling; and MRP, discussed in Chapter 8, provides production schedules for end items and subassemblies in the product structure.

There are many different types of scheduling problems faced by the firm. A partial list includes

1. *Job shop scheduling.* Job shop scheduling, known more commonly in practice as shop floor control, is the set of activities in the shop that transform inputs (a set of requirements) to outputs (products to meet those requirements). Much of this chapter will be concerned with sequencing issues on the shop floor, and more will be said about this problem in Section 9.1.
2. *Personnel scheduling.* Scheduling personnel is an important problem for both manufacturing and service industries. Although shift scheduling on the factory floor may be considered one of the functions of shop floor control, personnel scheduling is a much larger problem. Scheduling health professionals in hospitals and other health facilities is one example. Determining whether to meet peak demand with overtime shifts, night shifts, or subcontracting is another example of a personnel scheduling problem.
3. *Facilities scheduling.* This problem is particularly important when facilities become a bottleneck resource. Scheduling operating rooms at hospitals is one example. As the need for health care increases, some hospitals and health maintenance organizations (HMOs) find that facilities are strained. A similar problem occurs in colleges and universities in which enrollments have increased without commensurate increases in the size of the physical plant.
4. *Vehicle scheduling.* Manufacturing firms must distribute their products in a cost-efficient and timely manner. Some service operations, such as dial-a-ride systems, involve pick-ups and deliveries of goods and/or people. Vehicle routing is a problem that arises in many contexts. Problems such as scheduling snow removal equipment, postal and bank deliveries, and shipments to customers with varying requirements at different locations are some examples. Vehicle scheduling is discussed in Section 6.6.
5. *Vendor scheduling.* For firms with just-in-time (JIT) systems, scheduling deliveries from vendors is an important logistics issue. Purchasing must be coordinated with the entire product delivery system to ensure that JIT production systems function efficiently. Vollman et al. (1992, p. 191) discuss the application of vendor scheduling to the JIT system at Steelcase. (JIT is discussed in Chapters 1 and 8 of this book.)

6. *Project scheduling.* A project may be broken down into a set of interrelated tasks. Although some tasks can be done concurrently, many tasks cannot be started until others are completed. Complex projects may involve thousands of individual tasks that must be coordinated for the project to be completed on time and within budget. Project scheduling is an important component of the planning function, which we treat in detail in Chapter 10.
7. *Dynamic versus static scheduling.* Most scheduling theory that we review in this chapter views the scheduling problem as a static one. Numerous jobs arrive simultaneously to be processed on a set of machines. In practice, many scheduling problems are dynamic in the sense that jobs arrive continuously over time. One example is the problem faced by an air traffic controller who must schedule runways for arriving planes. The problem is a dynamic one in that planes arrive randomly and runways are freed up and committed randomly over time. Dynamic scheduling problems, treated in Section 9 of this chapter, are analyzed using the tools of queueing theory (discussed in detail in Supplement 2, which follows this chapter).

Scheduling is a complex but extremely important operations function. The purpose of this chapter is to give the reader the flavor of the kinds of results one can obtain using analytical models, and to show how these models can be used to solve certain classes of scheduling problems. Our focus is primarily on job shop scheduling, but we consider several other scheduling problems as well.

9.1 PRODUCTION SCHEDULING AND THE HIERARCHY OF PRODUCTION DECISIONS

Crucial to controlling production operations is the detailed scheduling of various aspects of the production function. We may view the production function in a company as a hierarchical process. First, the firm must forecast demand for aggregate sales over some predetermined planning horizon. These forecasts provide the input for determining the sales and operations planning function discussed in Chapter 3. The production plan then must be translated into the master production schedule (MPS). The MPS results in specific production goals by product and time period.

Materials requirements planning (MRP), treated in detail in Chapter 8, is one method for meeting specific production goals of finished-goods inventory generated by the MPS. The MRP system "explodes" the production levels one obtains from the MPS analysis back in time to obtain production targets at each level of assembly by time period. The result of the MRP analysis is specific planned order releases for final products, subassemblies, and components.

Finally, the planned order releases must be translated into a set of tasks and the due dates associated with those tasks. This level of detailed planning results in the shop floor schedule. Because the MRP or other lot scheduling system usually recommends revisions in the planned order releases, shop floor schedules change frequently. The hierarchy of production planning decisions is shown schematically in Figure 9–1.

Shop floor control means scheduling personnel and equipment in a work center to meet the due dates for a collection of jobs. Often, jobs must be processed through the machines in the work center in a unique order or sequence. Figure 9–2 shows the layout of a typical job shop.

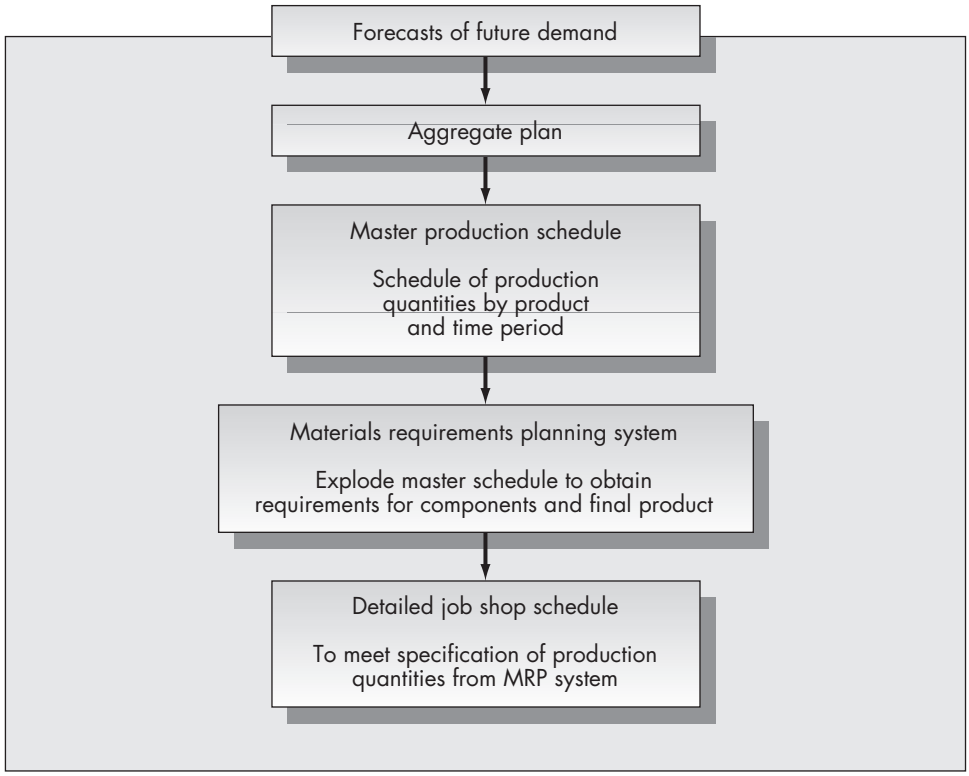


FIGURE 9–1
Hierarchy of production decisions

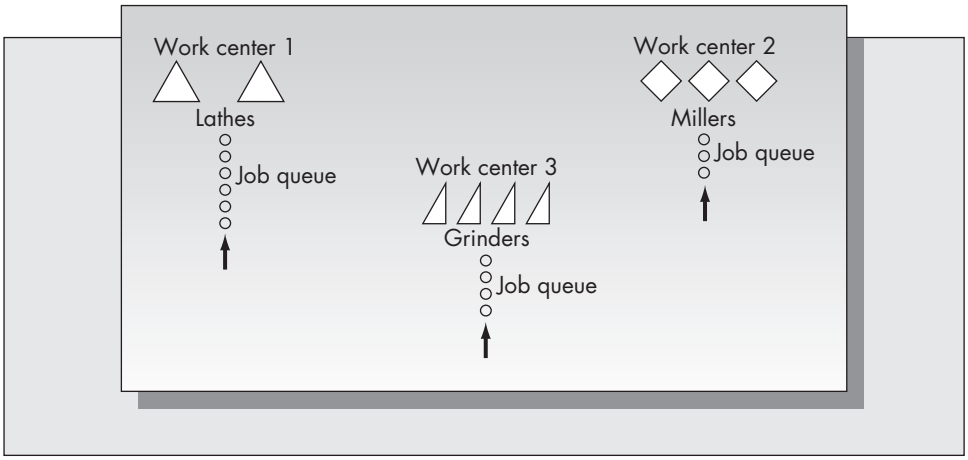


FIGURE 9–2
Typical job shop layout

Both jobs and machines are treated as indivisible. Jobs must wait, or queue up, for processing when machines are busy. This is referred to as discrete processing. Production scheduling in continuous-process industries, such as sugar or oil refining, has a very different character.

Although there are many problems associated with operations scheduling, our concern in this chapter will be job sequencing. Given a collection of jobs remaining to be processed on a collection of machines, the problem is how to sequence these jobs to optimize some specified criterion. Properly choosing the sequencing rule can effect dramatic improvements in the throughput rate of the job shop.

9.2 IMPORTANT CHARACTERISTICS OF JOB SHOP SCHEDULING PROBLEMS

Significant issues for determining optimal or approximately optimal scheduling rules are the following:

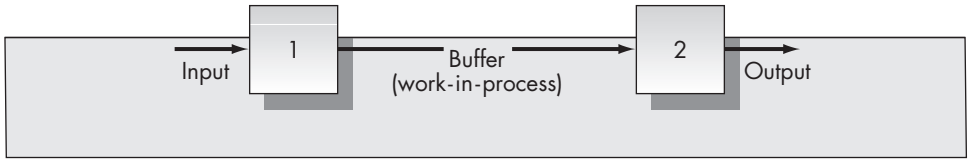
1. *The job arrival pattern.* We often view the job shop problem as a static one in which we take a "snapshot" of the system at a point in time and proceed to solve the problem based on the value of the current state. However, the number of jobs waiting to be processed is constantly changing. Hence, although many of the solution algorithms we consider view the problem as being static, most practical shop scheduling problems are dynamic in nature.
2. *Number and variety of machines in the shop.* A particular job shop may have unique features that could make implementing a solution obtained from a scheduling algorithm difficult. For example, it is generally assumed that all machines of a given type are identical. This is not always the case, however. The throughput rate of a particular machine could depend upon a variety of factors, such as the condition of the machine or the skill of the operator. Depending on the layout of the shop and the nature of the jobs, constraints might exist that would make solutions obtained from an "all-purpose" procedure infeasible.
3. *Number of workers in the shop.* Both the number of workers in the shop and the number and variety of machines in the shop determine the shop's capacity. Capacity planning is an important aspect of production planning. Many control systems, such as traditional MRP discussed in Chapter 8, do not explicitly incorporate capacity considerations. Furthermore, capacity is dynamic. A breakdown of a single critical machine or the loss of a critical employee could result in a bottleneck and a reduction in the shop's capacity.
4. *Particular flow patterns.* The solutions obtained from the scheduling algorithms to be presented in this chapter require that jobs be completed in a fixed order. However, each sequence of jobs through machines results in a pattern of flow of materials through the system. Because materials-handling issues often are treated separately from scheduling issues, infeasible flow patterns may result.
5. *Evaluation of alternative rules.* The choice of objective will determine the suitability and effectiveness of a sequencing rule. It is common for more than one objective to be important, so it may be impossible to determine a unique optimal rule. For example, one may wish to minimize the time required to complete all jobs, but also may wish to limit the maximum lateness of any single job.

Objectives of Job Shop Management

One of the difficulties of scheduling is that many, often conflicting, objectives are present. The goals of different parts of the firm are not always the same. Some of the most common objectives are

1. Meet due dates.
2. Minimize work-in-process (WIP) inventory.
3. Minimize the average flow time through the system.
4. Provide for high machine/worker time utilization. (Minimize machine/worker idle time.)
5. Provide for accurate job status information.
6. Reduce setup times.
7. Minimize production and worker costs.

FIGURE 9–3
A process composed of two operations in series



It is obviously impossible to optimize all seven objectives simultaneously. In particular, (1) and (3) are aimed primarily at providing a high level of customer service, and (2), (4), (6), and (7) are aimed primarily at providing a high level of plant efficiency. Determining the trade-off between cost and quality is one of the most important strategic issues facing a firm today.

Some of these objectives conflict. If the primary objective is to reduce work-in-process inventory (as, for example, with just-in-time inventory control systems, discussed in Chapter 8), it is likely that worker idle time will increase. As the system tightens up by reducing the inventory within and between manufacturing operations, differences in the throughput rate from one part of the system to another may force the faster operations to wait. Although not recommended by those espousing the just-in-time philosophy, buffer inventories between operations can significantly reduce idle time.

As an example, consider the simple system composed of two operations in series, pictured in Figure 9–3. If work-in-process inventory is zero, then the throughput of the system at any point in time is governed by the smaller of the throughputs of the two operations. If operation 1 is temporarily halted by a machine failure, then operation 2 also must remain idle. However, if there is a buffer inventory placed between the operations, then 2 can continue to operate while 1 is undergoing repair or recalibration.

Finding the proper mix between WIP inventory and worker idle time is equivalent to choosing a point on the trade-off curve of these conflicting objectives. (Trade-off, or exchange, curves were discussed in Chapter 5 in the context of multi-item inventory control.) Such a curve is pictured in Figure 9–4*a*. A movement from one point to another along such a curve does not necessarily imply that the system has improved, but rather that different weights are being applied to the two objectives. A true improvement in the overall system would mean that the entire trade-off curve undergoes a downward shift, such as that pictured in Figure 9–4*b*.

9.3 JOB SHOP SCHEDULING TERMINOLOGY

In general, a job shop scheduling problem is one in which n jobs must be processed through m machines. The complexity of the problem depends upon a variety of factors, such as what job sequences are permissible and what optimization criteria are chosen. In this section we define some of the terms that will be used throughout this chapter.

1. *Flow shop.* In a flow shop each of the n jobs must be processed through the m machines in the same order, and each job is processed exactly once on each machine. This is what we typically think of as an assembly line.

2. *Job shop.* A general job shop differs from a flow shop in that not all jobs are assumed to require exactly m operations, and some jobs may require multiple operations on a single machine. Furthermore, in a job shop each job may have a different required sequencing of operations. General job shop problems are extremely complex. All-purpose solution algorithms for solving general job shop problems do not exist.

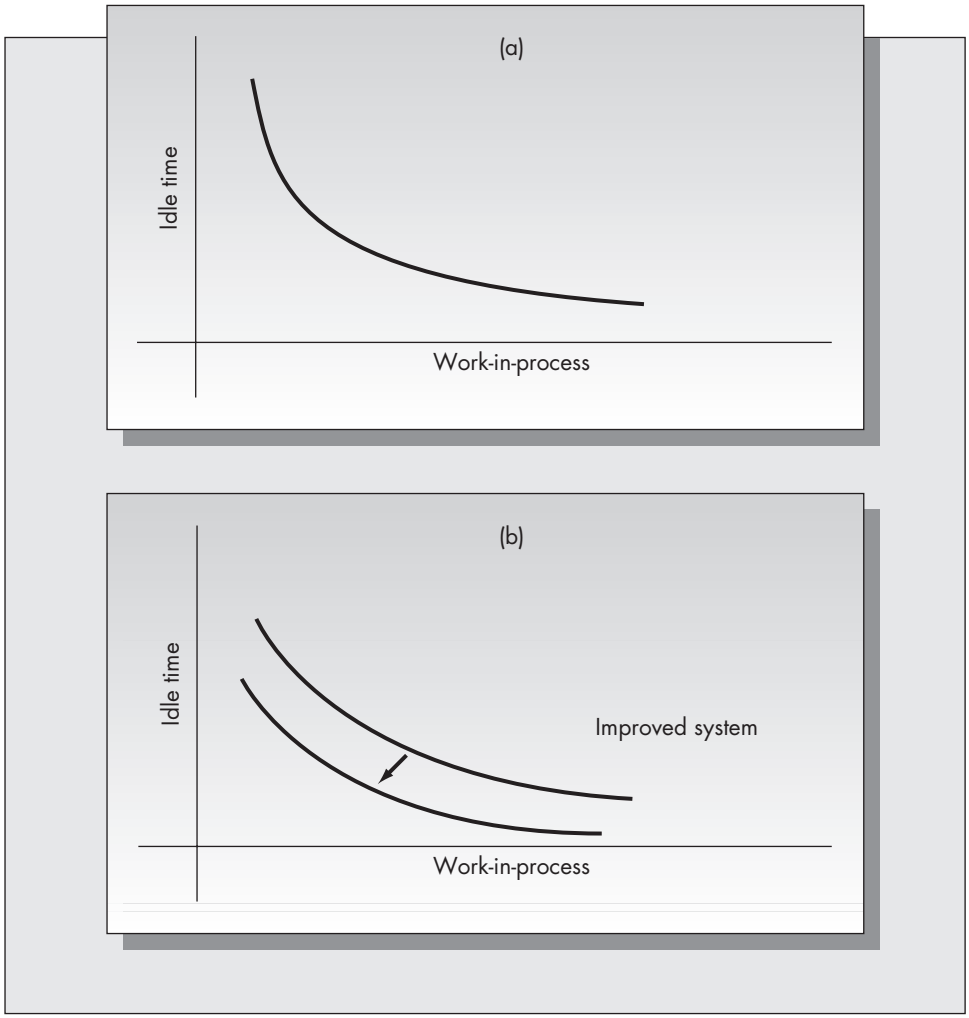


FIGURE 9–4
Conflicting objectives in job shop management

3. *Parallel processing versus sequential processing.* Most of the problems that we will consider involve sequential processing. This means that the m machines are distinguishable, and different operations are performed by different machines. In parallel processing we assume that the machines are identical, and any job can be processed on any machine. An example of parallel processing occurs in a phone switching center, in which calls are processed through the next available server. Parallel processing is discussed in the context of stochastic scheduling in Section 9.8.

4. *Flow time.* The flow time of job i is the time that elapses from the initiation of the first job on the first machine to the completion of job i. Equivalently, it is the amount of time that job i spends in the system. The *mean flow time,* which is a common measure of system performance, is the arithmetic average of the flow times for all n jobs.

5. *Makespan.* The makespan is the flow time of the job that is completed last. It is also the time required to complete all n jobs. Minimizing the makespan is a common objective in multiple-machine sequencing problems.

6. *Tardiness and lateness.* Tardiness is the positive difference between the completion time (flow time) and the due date of a job. A tardy job is one that is completed after

its due date. Lateness refers to the difference between the job completion time and its due date, and differs from tardiness in that lateness can be either positive or negative. Minimizing the average tardiness and the maximum tardiness is also a common scheduling objective.

9.4 A COMPARISON OF SPECIFIC SEQUENCING RULES

In the comparison and evaluation of sequencing rules, we consider the job shop at a fixed point in time. This section will focus on a single machine only. Assume that there is a collection of jobs that must be processed on the machine and that each job has associated with it a processing time and a due date. We compare the performance of four sequencing rules commonly used in practice. The purpose of this section is to illustrate how these sequencing rules affect various measures of system performance.

We compare the following four sequencing rules:

1. *First-come, first-served (FCFS).* Jobs are processed in the sequence in which they entered the shop.
2. *Shortest processing time (SPT).* Jobs are sequenced in increasing order of their processing times. The job with the shortest processing time is first, the job with the next shortest processing time is second, and so on.
3. *Earliest due date (EDD).* Jobs are sequenced in increasing order of their due dates. The job with the earliest due date is first, the job with the next earliest due date is second, and so on.
4. *Critical ratio (CR).* Critical ratio scheduling requires forming the ratio of the processing time of the job, divided by the remaining time until the due date, and scheduling the job with the largest ratio next.

We compare the performance of these four rules for a specific case based on mean flow time, average tardiness, and number of tardy jobs. The purpose of the next example is to help the reader develop an intuition for the mechanics of scheduling before presenting formal results.

Example *9.1* A machining center in a job shop for a local fabrication company has five unprocessed jobs remaining at a particular point in time. The jobs are labeled 1, 2, 3, 4, and 5 in the order that they entered the shop. The respective processing times and due dates are given in the following table.

Job Number	Processing Time	Due Date
1	11	61
2	29	45
3	31	31
4	1	33
5	2	32

First-Come, First-Served

Because the jobs are assumed to have entered the shop in the sequence that they are numbered, FCFS scheduling means that the jobs are scheduled in the order 1, 2, 3, 4, 5.

This results in

Sequence	Completion Time	Due Date	Tardiness
1	11	61	0
2	40	45	0
3	71	31	40
4	72	33	39
5	74	32	42
Totals	268		121

$$\text{Mean flow time} = 268/5 = 53.6.$$
$$\text{Average tardiness} = 121/5 = 24.2.$$
$$\text{Number of tardy jobs} = 3.$$

The tardiness of a job is equal to zero if the job is completed prior to its due date and is equal to the number of days late if the job is completed after its due date.

Shortest Processing Time

Here jobs are sequenced in order of increasing processing time.

Job	Processing Time	Completion Time	Due Date	Tardiness
4	1	1	33	0
5	2	3	32	0
1	11	14	61	0
2	29	43	45	0
3	31	74	31	43
Totals		135		43

$$\text{Mean flow time} = 135/5 = 27.0.$$
$$\text{Average tardiness} = 43/5 = 8.6.$$
$$\text{Number of tardy jobs} = 1.$$

Earliest Due Date

Here jobs are completed in the order of their due dates.

Job	Processing Time	Completion Time	Due Date	Tardiness
3	31	31	31	0
5	2	33	32	1
4	1	34	33	1
2	29	63	45	18
1	11	74	61	13
Totals		235		33

$$\text{Mean flow time} = 235/5 = 47.0.$$
$$\text{Average tardiness} = 33/5 = 6.6.$$
$$\text{Number of tardy jobs} = 4.$$

Critical Ratio Scheduling

After each job has been processed, we compute

$$\frac{\text{Due date} - \text{Current time}}{\text{Processing time}},$$

which is known as the critical ratio, and schedule the next job in order to minimize the value of the critical ratio. The idea behind critical ratio scheduling is to provide a balance between SPT, which only considers processing time, and EDD, which only considers due dates. The ratio will grow smaller as the current time approaches the due date, and more priority will be given to those jobs with longer processing times. One disadvantage of the method is that the critical ratios need to be recalculated each time a job is scheduled.

It is possible that the numerator will be negative for some or all of the remaining jobs. When that occurs it means that the job is late, and we will assume that late jobs are automatically scheduled next. If there is more than one late job, then the late jobs are scheduled in SPT sequence.

First we compute the critical ratios starting at time $t = 0$.

Current time: $t = 0$

Job	Processing Time	Due Date	Critical Ratio
1	11	61	61/11 (5.545)
2	29	45	45/29 (1.552)
3	31	31	31/31 (1.000)
4	1	33	33/1 (33.00)
5	2	32	32/2 (16.00)

The minimum value corresponds to job 3, so job 3 is performed first. As job 3 requires 31 units of time to process, we must update all the critical ratios in order to determine the next job to process. We move the clock to time $t = 31$ and recompute the critical ratios.

Current time: $t = 31$

Job	Processing Time	Due Date − Current Time	Critical Ratio
1	11	30	30/11 (2.727)
2	29	14	14/29 (0.483)
4	1	2	2/1 (2.000)
5	2	1	1/2 (0.500)

The minimum is 0.483, which corresponds to job 2. Hence, job 2 is scheduled next. Since job 2 has a processing time of 29, we update the clock to time $t = 31 + 29 = 60$.

Current time: $t = 60$

Job	Processing Time	Due Date − Current Time	Critical Ratio
1	11	1	1/11 (.0909)
4	1	−27	$-27/1 < 0$
5	2	−28	$-28/2 < 0$

Jobs 4 and 5 are now late, so they are given priority and scheduled next. Since they are scheduled in SPT order, they are done in the sequence job 4, then job 5. Finally, job 1 is scheduled last.

Summary of the Results for Critical Ratio Scheduling

Job	Processing Time	Completion Time	Tardiness
3	31	31	0
2	29	60	15
4	1	61	28
5	2	63	31
1	11	74	13
Totals		289	87

$$\text{Mean flow time} = 289/5 = 57.8.$$
$$\text{Average tardiness} = 87/5 = 17.4.$$
$$\text{Number of tardy jobs} = 4.$$

We summarize the results of this section for all four scheduling rules:

Summary of the Results of Four Scheduling Rules

Rule	Mean Flow Time	Average Tardiness	Number of Tardy Jobs
FCFS	53.6	24.2	3
SPT	27.0	8.6	1
EDD	47.0	6.6	4
CR	57.8	17.4	4

9.5 OBJECTIVES IN JOB SHOP MANAGEMENT: AN EXAMPLE

Example *9.2* An air traffic controller is faced with the problem of scheduling the landing of five aircraft. Based on the position and runway requirements of each plane, she estimates the following landing times:

Plane:	1	2	3	4	5
Time (in minutes):	26	11	19	16	23

Only one plane may land at a time. The problem is essentially the same as that of scheduling five jobs on a single machine, with the planes corresponding to the jobs, the landing times to the processing times, and the runway to the machine.

1. With the given information, two reasonable objectives would be to minimize the total time required to land all planes (i.e., the makespan) or the average time required to land the planes (the mean flow time). The makespan for any sequence is clearly 95 minutes, the sum of the landing times. However, as we saw in Example 9.1, the mean flow time is not sequence independent and the shortest-processing-time rule minimizes the mean flow time. We will show in Section 9.6 that SPT is the optimal sequencing rule for minimizing mean flow time for a single machine in general.

2. An alternative objective might be to land as many people as quickly as possible. In this case we also would need to know the number of passengers on each plane. Suppose that

these numbers are as follows:

Plane	1	2	3	4	5
Landing time	26	11	19	16	23
Number of passengers	180	12	45	75	252

The appropriate objective in this case might be to minimize the *weighted* makespan or the weighted sum of the completion times, where the weights would correspond to the number of passengers in each plane. Notice that the objective function would now be in units of passenger-minutes.

3. An issue that we have not yet addressed is the time that each plane is scheduled to arrive. Assume the following data:

Plane	1	2	3	4	5
Landing time	26	11	19	16	23
Scheduled arrival time	5:30	5:45	5:15	6:00	5:40

Sequencing rules that ignore due dates could give very poor results in terms of meeting the arrival times. Some possible objectives related to due dates include minimizing the average tardiness and minimizing the maximum tardiness.

4. Thus far we have ignored special conditions that favor some planes over others. Suppose that plane number 4 has a critically low fuel level. This would probably result in plane 4 taking precedence. Priority constraints could arise in other ways as well: planes that are scheduled for continuing flights or planes carrying precious or perishable cargo also might be given priority.

The purpose of this section was to demonstrate the difficulties of choosing an objective function for job sequencing problems. The optimal sequence is highly sensitive to the choice of the objective, and the appropriate objective is not always obvious.

Problems for Sections 9.1–9.5

1. Discuss each of the following objectives listed and the relationship each has with job shop performance.
   a. Reduce WIP inventory.
   b. Provide a high level of customer service.
   c. Reduce worker idle time.
   d. Improve factory efficiency.
2. In Problem 1, why are (*a*) and (*c*) conflicting objectives, and why are (*b*) and (*d*) conflicting objectives?
3. Define the following terms:
   a. Flow shop.
   b. Job shop.
   c. Sequential versus parallel processing.
   d. Makespan.
   e. Tardiness.
4. Four trucks, 1, 2, 3, and 4, are waiting on a loading dock at XYZ Company that has only a single service bay. The trucks are labeled in the order that they arrived at the dock. Assume the current time is 1:00 P.M. The times required to unload each truck

and the times that the goods they contain are due in the plant are given in the following table.

Truck	Unloading Time (minutes)	Time Material Is Due
1	20	1:25 P.M.
2	14	1:45 P.M.
3	35	1:50 P.M.
4	10	1:30 P.M.

Determine the schedules that result for each of the rules FCFS, SPT, EDD, and CR. In each case compute the mean flow time, average tardiness, and number of tardy jobs.

5. Five jobs must be scheduled for batch processing on a mainframe computer system. The processing times and the promised times for each of the jobs are listed here.

Job	1	2	3	4	5
Processing time	40 min	2.5 hr	20 min	4 hr	1.5 hr
Promised time	11:00 A.M.	2:00 P.M.	2:00 P.M.	1:00 P.M.	4:00 P.M.

Assume that the current time is 10:00 A.M.

a. If the jobs are scheduled according to SPT, find the tardiness of each job and the mean tardiness of all jobs.

b. Repeat the calculation in part (*a*) for EDD scheduling.

9.6 AN INTRODUCTION TO SEQUENCING THEORY FOR A SINGLE MACHINE

Assume that n jobs are to be processed through one machine. For each job i, define the following quantities:

$$t_i = \text{Processing time for job } i,$$
$$d_i = \text{Due date for job } i,$$
$$W_i = \text{Waiting time for job } i,$$
$$F_i = \text{Flow time for job } i,$$
$$L_i = \text{Lateness of job } i,$$
$$T_i = \text{Tardiness of job } i,$$
$$E_i = \text{Earliness of job } i.$$

The processing time and the due date are constants that are attached to the description of each job. The waiting time for a job is the amount of time that the job must wait before its processing can begin. For the cases that we consider, it is also the sum of the processing times for all the preceding jobs. The flow time is simply the waiting time plus the job processing time ($F_i = W_i + t_i$). The flow time of job i and the completion time of job i are the same. We will define the lateness of job i as $L_i = F_i - d_i$, and assume that lateness can be either a positive or a negative quantity. Tardiness is the positive part of lateness ($T_i = \max[L_i, 0]$), and earliness is the negative part of lateness ($E_i = \max[-L_i, 0]$).

Other related quantities are maximum tardiness $T_{\max}$, given by the formula

$$T_{\max} = \max\{T_1, T_2, \ldots, T_n\},$$

and the mean flow time F', given by the formula

$$F' = \frac{1}{n}\sum_{i=1}^{n} F_i.$$

As we are considering only a single machine, every schedule can be represented by a permutation (that is, ordering) of the integers 1, 2, . . . , *n*. There are exactly $n!$ different permutation schedules [$n! = n(n-1)\cdots(2)(1)$].

Shortest-Processing-Time Scheduling

We have the following result:

Theorem 9.1 **The scheduling rule that minimizes the mean flow time F' is SPT.**

Theorem 9.1 is easy to prove. Let [1], [2], . . . , [*n*] be any permutation of the integers 1, 2, 3, . . . , *n*. The flow time of the job that is scheduled in position *k* is given by

$$F_{[k]} = \sum_{i=1}^{k} t_{[i]}.$$

It follows that the mean flow time is given by

$$F' = \frac{1}{n}\sum_{k=1}^{n} F_{[k]} = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{k} t_{[i]}.$$

The double summation term may be written in a different form. Expanding the double summation, we obtain

$$\begin{aligned} k &= 1\text{: } t_{[1]} \\ k &= 2\text{: } t_{[1]} + t_{[2]} \\ &\vdots \\ k &= n\text{: } t_{[1]} + t_{[2]} + \cdots + t_{[n]}. \end{aligned}$$

By summing down the column rather than across the row, we may rewrite F' in the form

$$nt_{[1]} + (n-1)t_{[2]} + \cdots + t_{[n]},$$

which is clearly minimized by setting

$$t_{[1]} \le t_{[2]} \le \cdots \le t_{[n]},$$

which is exactly the SPT sequencing rule.

We have the following corollary to Theorem 9.1.

Corollary 9.1 **The following measures are equivalent:**

1. **Mean flow time**
2. **Mean waiting time**
3. **Mean lateness**

Taken together, Corollary 9.1 and Theorem 9.1 establish that SPT minimizes mean flow time, mean waiting time, and mean lateness for single-machine sequencing.

Earliest-Due-Date Scheduling

If the objective is to minimize the maximum lateness, then the jobs should be sequenced according to their due dates. That is, $d_{[1]} \leq d_{[2]} \leq \cdots \leq d_{[n]}$. We will not present a proof of this result. The idea behind the proof is to choose some schedule that does not sequence the jobs in order of their due dates; that implies that there is some value of k such that $d_{[k]} > d_{[k+1]}$. One shows that by interchanging the positions of jobs k and $k + 1$, the maximum lateness is reduced.

Minimizing the Number of Tardy Jobs

There are many instances in which the penalty for a late (tardy) job remains the same no matter how late it is. For example, any delay in the completion of all tasks required for preparation of a space launch would cause the launch to be aborted, independent of the length of the delay.

We will describe an algorithm from Moore (1968) that minimizes the number of tardy jobs for the single machine problem.

Step 1. Sequence the jobs according to the earliest due date to obtain the initial solution. That is $d_{[1]} \leq d_{[2]} \leq \cdots \leq d_{[n]}$.

Step 2. Find the first tardy job in the current sequence, say job $[i]$. If none exists, go to step 4.

Step 3. Consider jobs [1], [2], . . . , $[i]$. Reject the job with the largest processing time. Return to step 2.

Step 4. Form an optimal sequence by taking the current sequence and appending to it the rejected jobs. The jobs appended to the current sequence may be scheduled in any order because they constitute the tardy jobs.

Example *9.3*

A machine shop processes custom orders from a variety of clients. One of the machines, a grinder, has six jobs remaining to be processed. The processing times and promised due dates (both in hours) for the six jobs are given here.

Job	1	2	3	4	5	6
Due date	15	6	9	23	20	30
Processing time	10	3	4	8	10	6

The first step is to sequence the jobs according to the EDD rule.

Job	2	3	1	5	4	6
Due date	6	9	15	20	23	30
Processing time	3	4	10	10	8	6
Completion time	3	7	17	27	35	41

We see that the first tardy job is job 1, and there are a total of four tardy jobs. We now consider jobs 2, 3, and 1 and reject the job with the longest processing time. This is clearly job 1. At this point, the new current sequence is

Job	2	3	5	4	6
Due date	6	9	20	23	30
Processing time	3	4	10	8	6
Completion time	3	7	17	25	31

The first tardy job in the current sequence is now job 4. We consider the sequence 2, 3, 5, 4, and reject the job with the longest processing time, which is job 5. The current sequence is now

Job	2	3	4	6
Due date	6	9	23	30
Processing time	3	4	8	6
Completion time	3	7	15	21

Clearly there are no tardy jobs at this stage. The optimal sequence is 2, 3, 4, 6, 5, 1 or 2, 3, 4, 6, 1, 5. In either case the number of tardy jobs is exactly 2.

Precedence Constraints: Lawler's Algorithm

Lawler's algorithm (Lawler, 1973) is a powerful technique for solving a variety of constrained scheduling problems. The objective function is assumed to be of the form

$$\min \max_{1 \le i \le n} g_i(F_i)$$

where g_i is any nondecreasing function of the flow time F_i. Furthermore, the algorithm handles *any* precedence constraints. Precedence constraints occur when certain jobs must be completed before other jobs can begin; they are quite common in scheduling problems. Some examples of functions g_i that one might consider are $g_i(F_i) = F_i - d_i = L_i$, which corresponds to minimizing maximum lateness, or $g_i(F_i) = \max(F_i - d_i, 0)$, which corresponds to minimizing maximum tardiness.

The Algorithm

Lawler's algorithm first schedules the job to be completed last, then the job to be completed next to last, and so on. At each stage one determines the set of jobs not required to precede any other. Call this set V. Among the set V, choose the job k that satisfies

$$g_k(\tau) = \min_{i \in V} (g_i(\tau)),$$

where $\tau = \sum_{i=1}^{n} t_i$ and corresponds to the processing time of the current sequence.

Job k is now scheduled last. Consider the remaining jobs and again determine the set of jobs that are not required to precede any other remaining job. After scheduling job k, this set may have changed. The value of τ is reduced by t_k and the job scheduled next to last is now determined. The process is continued until all jobs are scheduled. Note that as jobs are scheduled, some of the precedence constraints may be relaxed, so the set V is likely to change at each iteration.

Example *9.4*

Tony D'Amato runs a local body shop that does automotive painting and repairs. On a particular Monday morning he has six cars waiting for repair. Three (1, 2, and 3) are from a car rental company and he has agreed to finish these cars in the order of the dates that they were promised. Cars 4, 5, and 6 are from a retail dealer who has requested that car 4 be completed first because a customer is waiting for it. The resulting precedence constraints can be represented as two disconnected networks, as pictured in Figure 9–5.

The times required to repair each of the cars (in days) and the associated promised completion dates are

Job	1	2	3	4	5	6
Processing time	2	3	4	3	2	1
Due date	3	6	9	7	11	7

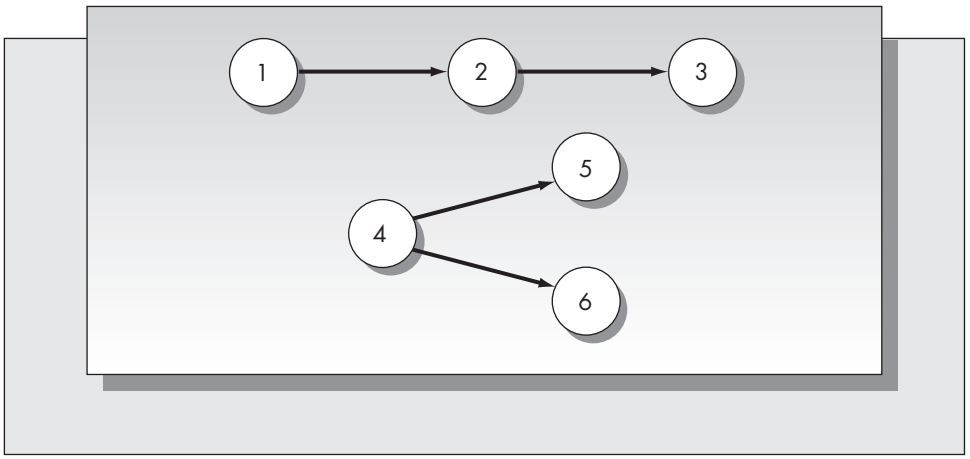


FIGURE 9–5
Precedence constraints for Example 9.4

Determine how the repair of the cars should be scheduled through the shop in order to minimize the maximum tardiness.

Solution

1. First we find the job scheduled last (sixth). Among the candidates for the last position are those jobs that are not predecessors of other jobs. These are 3, 5, and 6. The total processing time of all jobs is $2 + 3 + 4 + 3 + 2 + 1 = 15$. (This is the current value of τ.) As the objective is to minimize the maximum tardiness, we compare the tardiness of these three jobs and pick the one with the smallest value. We obtain $\min\{15 - 9, 15 - 11, 15 - 7\} = \min\{6, 4, 8\} = 4$, corresponding to job 5. Hence job 5 is scheduled last (position 6).

2. Next we find the job scheduled fifth. The candidates are jobs 3 and 6 only. At this point the value of τ is $15 - 2 = 13$. Hence, we find $\min\{13 - 9, 13 - 7\} = \min\{4, 6\} = 4$, which corresponds to job 3. Hence, job 3 is scheduled in the fifth position.

3. Find the job scheduled fourth. Because job 3 is no longer on the list, job 2 now becomes a candidate. The current value of $\tau = 13 - 4 = 9$. Hence, we compare $\min\{9 - 6, 9 - 7\} = \min\{3, 2\} = 2$, which corresponds to job 6. Schedule job 6 in the fourth position.

4. Find the job scheduled third. Job 6 has been scheduled, so job 4 now becomes a candidate along with job 2, and $\tau = 9 - 1 = 8$. Hence, we look for $\min\{8 - 6, 8 - 7\} = \min\{2, 1\} = 1$, which occurs at job 4.

5. At this point we would find the job scheduled second. However, we are left with only jobs 1 and 2, which, because of the precedence constraints, must be scheduled in the order 1–2.

Summarizing the results, the optimal sequence to repair the cars is 1–2–4–6–3–5.

In order to determine the value of the objective function, the maximum tardiness, we compute the flow time for each job and compare it to the due date. We have

Job	Processing Time	Flow Time	Due Date	Tardiness
1	2	2	3	0
2	3	5	6	0
4	3	8	7	1
6	1	9	7	2
3	4	13	9	4
5	2	15	11	4

Hence, the maximum tardiness is four days. The reader should convince him- or herself that any other sequence results in a maximum tardiness of at least four days.

Snapshot Application

MILLIONS SAVED WITH SCHEDULING SYSTEM FOR FRACTIONAL AIRCRAFT OPERATORS

Celebrities, corporate executives, and sports professionals are a large part of the group that uses private planes for travel. For many of these people, it doesn't make economic sense to purchase planes. An attractive alternative is fractional ownership, especially for those that have only occasional need of a plane. Fractional ownership of private planes provides owners with the flexibility to fly to over 5,000 destinations (as opposed to about 500 for the commercial airlines). Other advantages include privacy, personalized service, fewer delays, and the ability to conduct business on the plane.

The concept of a fractional aircraft program is similar to that of a time-share condominium, except that the aircraft owners are guaranteed access at any time with as little as four hours notice. The fees are based on the number of flight hours the owner will require: one-eighth share owners are allotted 100 hours of annual flying time, one-quarter share owners 200 hours, and so forth. The entire system is coordinated by fractional management company (FMC). Clearly, the problem of scheduling the planes and crews can become quite complex.

When scheduling planes and crews, the FMC must determine schedules that (1) meet customer requests on time, (2) satisfy maintenance and crew restrictions, and (3) allow for specific aircraft trip assignments and requests. The profitability of the FMC will depend upon how efficiently they perform these tasks. A group of consultants attacked this problem and developed a scheduling system known as ScheduleMiser.[1] The inputs to this system are trip requests, aircraft availability, and aircraft restrictions over a specified planning horizon. Note that even though owners are guaranteed service with only four hours notice, the vast majority of trips are booked at least three days or more in advance. This gives the FMC a reliable profile of demand over a two- to three-day planning horizon. Note that aircraft schedules must be coordinated with crew schedules, as crew work rules cannot be violated.

ScheduleMiser is the underlying engine that drives the larger planning system known as Flight Ops. ScheduleMiser is based on a mixed-integer mathematical formulation of the problem. The objective function consists of five terms delineating the various costs in the system. Several sets of constraints are included to ensure that demands are filled, crews are properly scheduled, and planes are not overbooked. This system was adopted and implemented by Raytheon Travel Air in November of 2000 (now Flight Options) for scheduling their fleet of over 100 aircraft. Raytheon reported a savings of over $4.4 million in the first year of implementation of this system. This is only one example of many mathematical-based scheduling systems that have been implemented in the airline industry.

[1] Martin, C., D. Jones, and P. Keskinocak. "Optimizing On-Demand Aircraft Schedules for Fractional Aircraft Operators," *Interfaces,* 33, no. 5, September–October 2003, pp. 22–35.

Problems for Section 9.6

6. Consider the information given in Problem 4. Determine the sequence that the trucks should be unloaded in order to minimize
   a. Mean flow time.
   b. Maximum lateness.
   c. Number of tardy jobs.
7. On May 1, a lazy MBA student suddenly realizes that he has done nothing on seven different homework assignments and projects that are due in various courses. He estimates the time required to complete each project (in days) and also notes their due dates:

Project	1	2	3	4	5	6	7
Time (days)	4	8	10	4	3	7	14
Due date	4/20	5/17	5/28	5/28	5/12	5/7	5/15

Because projects 1, 3, and 5 are from the same class, he decides to do those in the sequence that they are due. Furthermore, project 7 requires results from projects 2 and 3, so 7 must be done after 2 and 3 are completed. Determine the sequence in which he should do the projects in order to minimize the maximum lateness.

8. Eight jobs are to be processed through a single machine. The processing times and due dates are given here.

Job	1	2	3	4	5	6	7	8
Processing time	2	3	2	1	4	3	2	2
Due date	5	4	13	6	12	10	15	19

Furthermore, assume that the following precedence relationships must be satisfied:

$$2 \rightarrow 6 \rightarrow 3.$$

$$1 \rightarrow 4 \rightarrow 7 \rightarrow 8.$$

Determine the sequence in which these jobs should be done in order to minimize the maximum lateness subject to the precedence restrictions.

9. Jane Reed bakes breads and cakes in her home for parties and other affairs on a contract basis. Jane has only one oven for baking. One particular Monday morning she finds that she has agreed to complete five jobs for that day. Her husband John will make the deliveries, which require about 15 minutes each. Suppose that she begins baking at 8:00 A.M.

Job	Time Required	Promised Time
1	1.2 hr	11:30 A.M
2	40 min	10:00 A.M.
3	2.2 hr	11:00 A.M.
4	30 min	1:00 P.M.
5	3.1 hr	12:00 NOON
6	25 min	2:00 P.M.

Determine the sequence in which she should perform the jobs in order to minimize

a. Mean flow time.
b. Number of tardy jobs.
c. Maximum lateness.

10. Seven jobs are to be processed through a single machine. The processing times and due dates are given here.

Job	1	2	3	4	5	6	7
Processing time	3	6	8	4	2	1	7
Due date	4	8	12	15	11	25	21

Determine the sequence of the jobs in order to minimize

a. Mean flow time.
b. Number of tardy jobs.
c. Maximum lateness.
d. What is the makespan for any sequence?

9.7 SEQUENCING ALGORITHMS FOR MULTIPLE MACHINES

We now extend the analysis of Section 9.6 to the case in which several jobs must be processed on more than one machine. Assume that n jobs are to be processed through m machines. The number of possible schedules is staggering, even for moderate values of both n and m. For each machine, there are $n!$ different orderings of the jobs. If the jobs may be processed on the machines in any order, it follows that there are a total of $(n!)^m$ possible schedules. For example, for $n = 5$, $m = 5$, there are $24{,}883 \times 10^{10}$, or about 25 billion, possible schedules. Even with the availability of inexpensive computing today, enumerating all feasible schedules for even moderate-sized problems is impossible or, at best, impractical.

In this section we will present some known results for scheduling jobs on more than one machine. A convenient way to represent a schedule is via a Gantt chart. As an example, suppose that two jobs, I and J, are to be scheduled on two machines, 1 and 2. The processing times are

	Machine 1	Machine 2
Job I	4	1
Job J	1	4

Assume that both jobs must be processed first on machine 1 and then on machine 2. The possible schedules appear in four Gantt charts in Figure 9–6. The first two schedules are known as permutation schedules. That means that the jobs are processed in the same sequence on both machines. Clearly, for this example, the permutation schedules provide better system performance in terms of both total and average flow time.

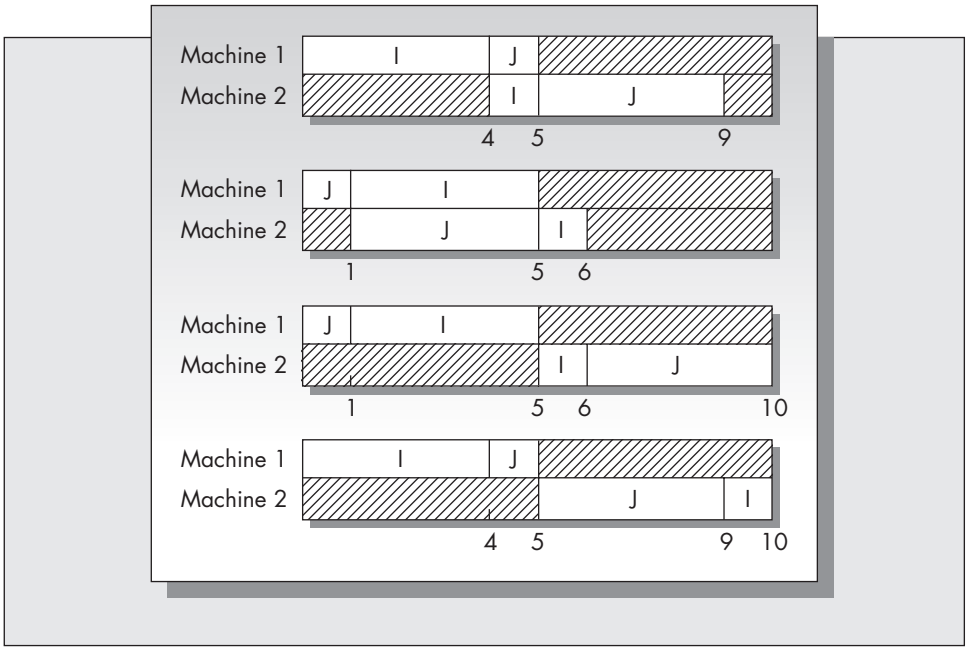


FIGURE 9–6
All possible schedules for two jobs on two machines

Recall that the total flow time (or makespan) is the total elapsed time from the initiation of the first job on the first machine until the completion of the last job on the last machine. For the given schedules, the makespans (total flow time) are respectively 9, 6, 10, and 10.

The mean flow time is also used as a measure of system performance. For the first schedule in the example, the mean flow time is $(5 + 9)/2 = 7$. For the second schedule, it is $(5 + 6)/2 = 5.5$, and so on.

A third possible objective is minimization of the mean idle time in the system. The mean idle time is the arithmetic average of the idle times for each machine. In schedule 1, we see that machine 1 is idle for 4 units of time (between times 5 and 9) and machine 2 is idle for 4 units of time as well (between times 0 and 4). Hence the mean idle time for schedule 1 is 4. In schedule 2, both machines 1 and 2 are idle for 1 unit of time, giving a mean idle time of 1. The mean idle times for schedules 3 and 4 are 5 units of time.

Scheduling *n* Jobs on Two Machines

Assume that n jobs must be processed through two machines and that each job must be processed in the order machine 1 then machine 2. Furthermore, assume that the optimization criterion is to minimize the makespan. The problem of scheduling on two machines turns out to have a relatively simple solution.

Theorem 9.2 **The optimal solution for scheduling *n* jobs on two machines is always a permutation schedule.**

Theorem 9.2 means that one can restrict attention to schedules in which the sequence of jobs is the same on both machines. This result can be demonstrated as follows. Consider a schedule for n jobs on two machines in which the sequencing of the jobs on the two machines is different. That is, the schedule looks as follows:

By reversing the position of these jobs on either machine, the flow time decreases. By scheduling the jobs in the order I–J on machine 2 the pair (I, J) on machine 2 may begin after I is completed on machine 1, rather than having to wait until J is completed on machine 1.

Because the total number of permutation schedules is exactly $n!$, determining optimal schedules for two machines is roughly of the same level of difficulty as determining optimal schedules for one machine.

A very efficient algorithm for solving the two-machine problem was discovered by Johnson (1954). Following Johnson's notation, denote the machines by A and B. It is assumed that the jobs must be processed first on machine A and then on machine B. Suppose that the jobs are labeled i, for $1 \leq i \leq n$, and define

$A_i =$ Processing time of job i on machine A.

$B_i =$ Processing time of job i on machine B.

Johnson's result is that the following rule is optimal for determining an order in which to process the jobs on the two machines.

Rule: Job i precedes job $i + 1$ if $\min(A_i, B_{i+1}) < \min(A_{i+1}, B_i)$.

An easy way to implement this rule is as follows:

1. List the values of A_i and B_i in two columns.
2. Find the smallest remaining element in the two columns. If it appears in column A, then schedule that job next. If it appears in column B, then schedule that job last.
3. Cross off the jobs as they are scheduled. Stop when all jobs have been scheduled.

Example *9.5* Five jobs are to be scheduled on two machines. The processing times are

Job	Machine A	Machine B
1	5	2
2	1	6
3	9	7
4	3	8
5	10	4

The first step is to identify the minimum job time. It is 1, for job 2 on machine A. Because it appears in column A, job 2 is scheduled first and row 2 is crossed out. The next smallest processing time is 2, for job 1 on machine B. This appears in the B column, so job 1 is scheduled last. The next smallest processing time is 3, corresponding to job 4 in column A, so that job 4 is scheduled next. Continuing in this fashion, we obtain the optimal sequence

2–4–3–5–1.

The Gantt chart for the optimal schedule is pictured in Figure 9–7. Note that there is no idle time between jobs on machine A. This is a feature of all optimal schedules.

Extension to Three Machines

The problem of scheduling jobs on three machines is considerably more complex. If we restrict attention to total flow time only, it is still true that a permutation schedule is optimal (this is not necessarily the case for average flow time). Label the machines A, B, and C. The three-machine problem can be reduced to (essentially) a two-machine problem if the following condition is satisfied:

$$\min A_i \geq \max B_i \qquad \text{or} \qquad \min C_i \geq \max B_i$$

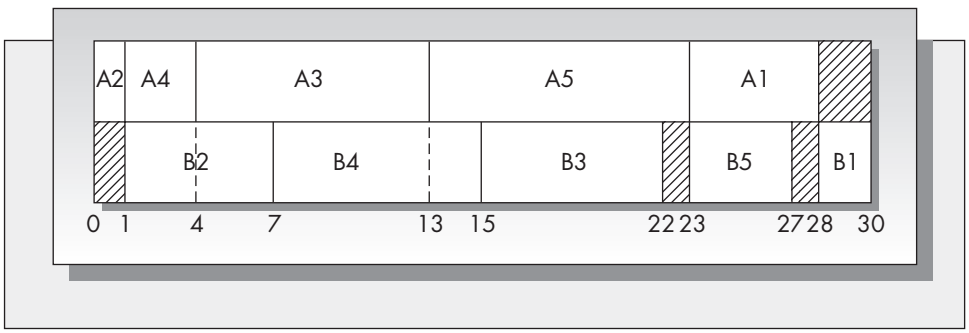


FIGURE 9–7
Gantt chart for the optimal schedule for Example 9.5

It is only necessary that *either one* of these conditions be satisfied. If that is the case, then the problem is reduced to a two-machine problem in the following way.

Define $A_i' = A_i + B_i$, and define $B_i' = B_i + C_i$. Now solve the problem using the rules described for two machines, treating A_i' and B_i' as the processing times. The resulting permutation schedule will be optimal for the three-machine problem.

Example *9.6* Consider the following job times for a three-machine problem. Assume that the jobs are processed in the sequence A–B–C.

	Machine		
Job	**A**	**B**	**C**
1	4	5	8
2	9	6	10
3	8	2	6
4	6	3	7
5	5	4	11

Checking the conditions, we find

$$\min A_i = 4,$$
$$\max B_i = 6,$$
$$\min C_i = 6,$$

so that the required condition is satisfied. We now form the two columns A′ and B′.

	Machine	
Job	**A′**	**B′**
1	9	13
2	15	16
3	10	8
4	9	10
5	9	15

The problem is now solved using the two-machine algorithm. The optimal solution is

1–4–5–2–3.

Note that because of ties in column A, the optimal solution is not unique.

If the conditions for reducing a three-machine problem to a two-machine problem are not satisfied, this method will usually give reasonable, but possibly suboptimal, results. As long as the objective is to minimize the makespan or total flow time, a permutation schedule is optimal for scheduling on three machines. (It is not necessarily true, however, that a permutation schedule is optimal for three machines when using an average flow time criterion.)

Note that we assume that the machines are different and the processing proceeds sequentially: all jobs are assumed to be processed first on machine 1, then on machine 2. For example, machine 1 might be a drill press and machine 2 a lathe. A related problem that we discuss in the context of stochastic scheduling is that of parallel processing on identical machines. In this case the machines are assumed to perform the same function, and any job may be assigned to any machine. For example, a collection of 10 jobs might

require processing on either one of two drill presses. The results for parallel processing suggest that SPT is an effective rule for minimizing mean flow time, but longest processing time first (LPT) is often more effective for minimizing total flow time or makespan. We will discuss parallel processing in the context of random job times in Section 9.8.

The Two-Job Flow Shop Problem

Assume that two jobs are to be processed through m machines. Each job must be processed by the machines in a particular order, but the sequences for the two jobs need not be the same. We present a graphical procedure for solving this problem developed by Akers (1956).

1. Draw a Cartesian coordinate system with the processing times corresponding to the first job on the horizontal axis and the processing times corresponding to the second job on the vertical axis. On each axis, mark off the operation times in the order in which the operations must be performed for that job.
2. Block out areas corresponding to each machine at the intersection of the intervals marked for that machine on the two axes.
3. Determine a path from the origin to the end of the final block that does not intersect any of the blocks and that minimizes the vertical movement. Movement is allowed only in three directions: horizontal, vertical, and 45-degree diagonal. The path with minimum vertical distance will indicate the optimal solution. Note that this will be the same as the path with minimum horizontal distance.

   This procedure is best illustrated by an example.

Example *9.7* A regional manufacturing firm produces a variety of household products. One is a wooden desk lamp. Prior to packing, the lamps must be sanded, lacquered, and polished. Each operation requires a different machine. There are currently shipments of two models awaiting processing. The times required for the three operations for each of the two shipments are

Job 1		Job 2	
Operation	**Time**	**Operation**	**Time**
Sanding (A)	3	A	2
Lacquering (B)	4	B	5
Polishing (C)	5	C	3

The first step is to block out the job times on each of the axes. Refer to Figure 9–8 for this step. Every feasible schedule is represented by a line connecting the origin to the tip of block C, with the condition that the line not go through a block. Only three types of movement are allowed: horizontal, vertical, and 45-degree diagonal. Horizontal movement implies that only job 1 is being processed, vertical movement implies that only job 2 is being processed, and diagonal movement implies that both jobs are being processed. Minimizing the flow time is the same as maximizing the time that both jobs are being processed. This is equivalent to finding the path from the origin to the end of block C that maximizes the diagonal movement and therefore minimizes either the horizontal or the vertical movement.

Two feasible schedules for this problem are represented in Figure 9–8. The total time required by any feasible schedule can be obtained in two ways: it is either the total time represented on the horizontal axis (12 in this case) plus the total vertical movement (4 and 3 respectively), or the total time on the vertical axis (10 in this case) plus the total horizontal movement (6 and 5 respectively). Schedule 1 has total time 16, and schedule 2 has total time 15. Schedule 2 turns out to be optimal for this problem.

The Gantt chart for the optimal schedule appears in Figure 9–9.

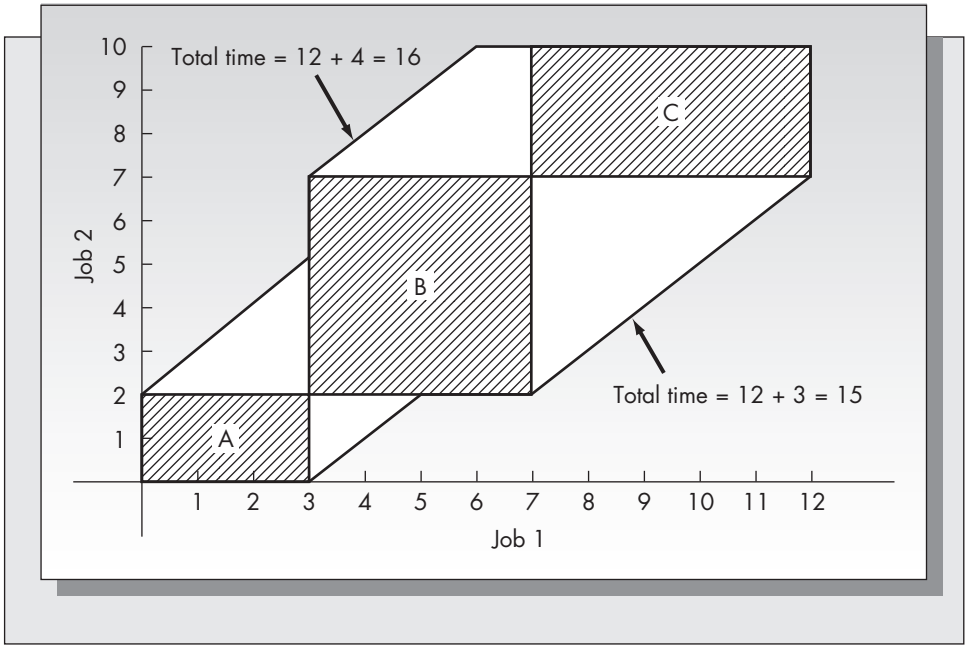


FIGURE 9–8
Graphical solution of Example 9.7

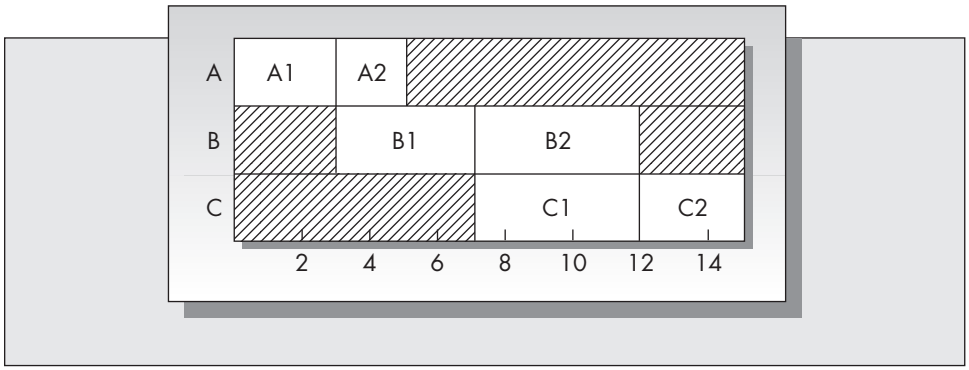


FIGURE 9–9
Gantt chart for optimal solution to Example 9.7

We should note that this method does *not* require the two jobs to be processed in the same sequence on the machines. We present another example to illustrate the case in which the sequence of the jobs is different.

Example *9.8*

Reggie Sigal and Bob Robinson are roommates who enjoy spending Sunday mornings reading the Sunday newspaper. Reggie likes to read the main section first, followed by the sports section, then the comics, and finally the classifieds. Bob also starts with the main section, but then goes directly to the classifieds, followed by the sports section and finally the comics. The times required (in tenths of an hour) for each to read the various sections are

Reggie		Bob	
Required Sequence	**Time**	**Required Sequence**	**Time**
Main section (A)	6	Main section (A)	4
Sports (B)	1	Classifieds (D)	3
Comics (C)	5	Sports (B)	2
Classifieds (D)	4	Comics (C)	5

The goal is to determine the order of the sections read for each to minimize the total time required to complete reading the paper. In this problem we identify Reggie as job 1 and Bob as job 2. The sections of the paper correspond to the machines.

Solution

In order to obtain the optimal solution, we first block out the processing times for each of the jobs. Assume that job 1 (Reggie) is blocked out on the *x* axis and job 2 (Bob) on the *y* axis. The processing times are sequenced on each axis in the order stated. The graphical representation for this problem is given in Figure 9–10. In the figure, two different feasible schedules are represented as paths from the origin to the point (16, 14). The top path represents a schedule that calls for Bob to begin reading the main section first (job 2 is processed first on machine A), and the lower path calls for Reggie to begin first. The lower path turns out to be optimal for this problem with a processing time of 20.

The optimal processing time for this problem is 20. As the time units are in tenths of an hour, this is exactly two hours. One converts the lower path in Figure 9–10 to a Gantt chart in the following way: From time 0 to 6, Reggie reads A and Bob is idle. Between times 6 and 12, the 45-degree line indicates that both Bob and Reggie are reading. Reggie reads B for 1 time unit and C for 5 time units, and Bob reads A for 4 time units and D for 3 time units. Reggie is now idle for one time unit because the path is vertical at this point, and begins reading D at time 13. When Reggie completes D, he is done. Starting at time 15, Bob reads C and completes his reading at time 20. Figure 9–11 shows the Gantt chart indicating the optimal solution.

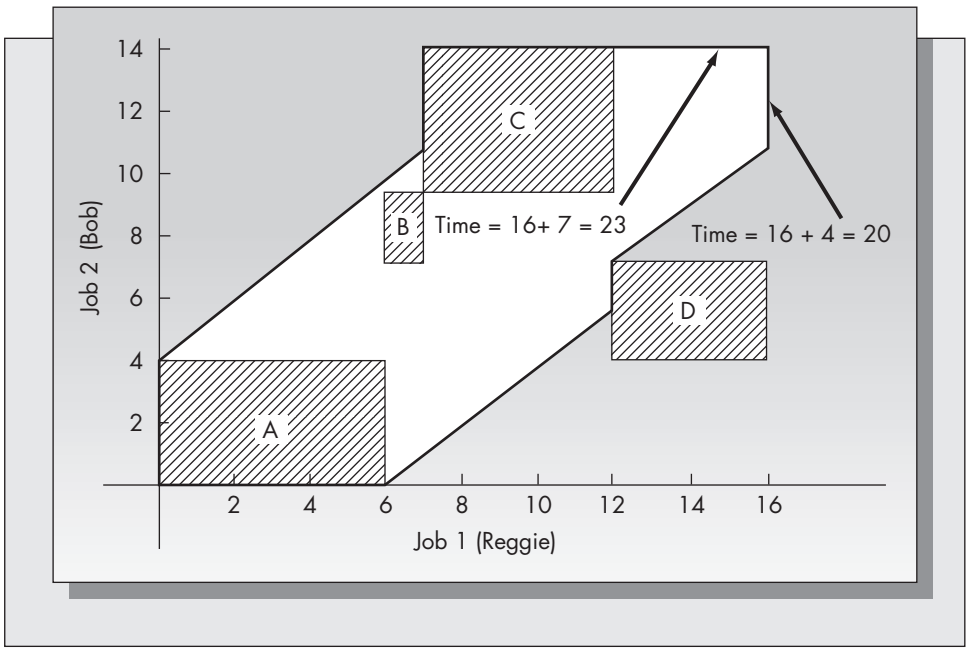


FIGURE 9–10
Graphical solution of Example 9.8

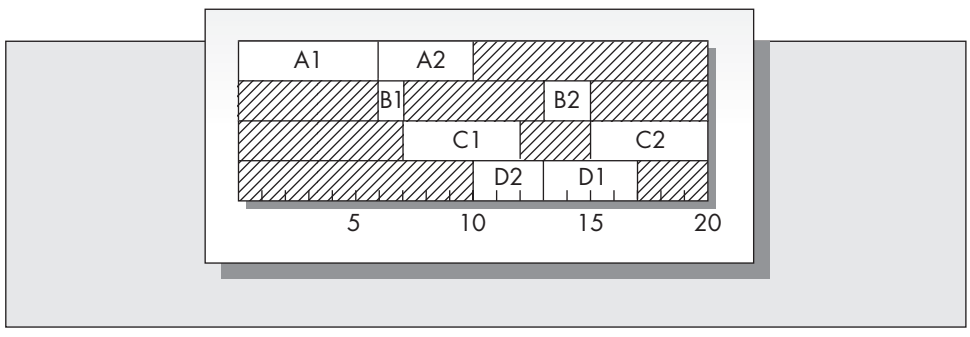


FIGURE 9–11
Gantt chart for optimal solution to Example 9.8

Problems for Section 9.7

11. Consider Example 9.6, illustrating the use of Johnson's algorithm for three machines. List all optimal solutions for this example.
12. Suppose that 12 jobs must be processed through six machines. If the jobs may be processed in any order, how many different possible schedules are there? If you were to run a computer program that could evaluate 100 schedules every second, how much time would the program require to evaluate all feasible schedules?
13. Two law students, John and Marsha, are planning an all-nighter to prepare for their law boards the following day. Between them they have one set of materials in the following five subjects: contracts, torts, civil law, corporate law, and patents. Based on their previous experience, they estimate that they will need the following amount of time (in hours) with each set of materials:

	Contracts	Torts	Civil	Corporate	Patents
John	1.2	2.2	0.7	0.5	1.5
Marsha	1.8	0.8	3.1	1.1	2.3

They agree that Marsha will get the opportunity to see each set of notes before John. Assume that they start their studying at 8:00 P.M. Determine the exact times that each will begin and end studying each subject in order to minimize the total time required for both to complete studying all five subjects.

14. The following four jobs must be processed through a three-machine flow shop.

	Machine		
Job	**A**	**B**	**C**
1	4	2	6
2	2	3	7
3	6	5	6
4	3	4	8

Find the optimal sequencing of the jobs in order to minimize the makespan. What is the makespan at the optimal solution? Draw a Gantt chart illustrating your solution.

15. Mary and Marcia Brown are two sisters currently attending university together. Each requires advising in five subjects: history, English, mathematics, science, and religion. They estimate that the time (in minutes) that each will require for advising is

	Mary	Marcia
Math	40	20
History	15	30
English	25	10
Science	15	35
Religion	20	25

They think that the five advisers will be available all day. Mary would like to visit the advisers in the order given in the table, and Marcia would prefer to see

them in the order math, religion, English, science, and history. At what times should each plan to see the advisers in order to minimize the total time for both to complete their advising?

16. Two jobs must be processed through four machines in the same order. The processing times in the required sequence are

Job 1		Job 2	
Machine	**Time**	**Machine**	**Time**
A	5	A	2
B	4	B	4
C	6	C	3
D	3	D	5

Determine how the two jobs should be scheduled in order to minimize the total makespan, and draw the Gantt chart indicating the optimal schedule.

17. Peter Minn is planning to go to the Department of Motor Vehicles to have his driver's license renewed. His friend, Patricia, who is accompanying him, is applying for a new license. In both cases there are five steps that are required: (A) having a picture taken, (B) signing a signature verification form, (C) passing a written test, (D) passing an eye test, and (E) passing a driving test.

For renewals the steps are performed in the order A, B, C, D, and E with average times required, respectively, of 0.2, 0.1, 0.3, 0.2, and 0.6 hour. In the case of new applications, the steps are performed in the sequence D, B, C, E, and A, with average times required of 0.3, 0.2, 0.7, 1.1, and 0.2 hour, respectively. Peter and Pat go on a day when the department is essentially empty. How should they plan their schedule in order to minimize the time required for both to complete all five steps?

9.8 STOCHASTIC SCHEDULING: STATIC ANALYSIS

Single Machine

An issue we have not yet addressed is uncertainty of the processing times. In practice it is possible and even likely that the exact completion time of one or more jobs may not be predictable. It is of interest to know whether or not there are some results concerning the optimal sequencing rules when processing times are uncertain. We assume that processing times are independent of one another.

In the case of processing on a single machine, most of the results are quite similar to those discussed earlier for the deterministic case. Suppose that n jobs are to be processed through a single machine. Assume that the job times, $t_1, t_2, \ldots, t_n$, are random variables with known distribution functions. The goal is to minimize the *expected* average weighted flow time; that is,

$$\text{Minimize } E\left(\frac{1}{n}\sum_{i=1}^{n} u_i F_i\right),$$

where u_i are the weights and F_i is the (random) flow time of job i.

Rothkopf (1966) has shown that the optimal solution is to sequence the jobs so that job i precedes job $i + 1$ if

$$\frac{E(t_i)}{u_i} < \frac{E(t_{i+1})}{u_{i+1}}.$$

Notice that if we set all the weights $u_i = 1$, then this rule is simply to order the jobs according to the minimum expected processing time; that is, it is essentially the same as SPT.

In the case of due date scheduling with random processing times, results are also similar to the deterministic case. Banerjee (1965) shows that if the objective is to minimize the maximum over all jobs of the probability that a job is late, then the optimal schedule is to order the jobs according to earliest due date (or according to earliest expected due date when due dates are themselves random).

Multiple Machines

Somewhat more interesting results exist for scheduling jobs with random job times on multiple machines. In fact, there are results available for this case that are *not* true for the deterministic problem.

An assumption that is usually made for the multiple-machine problem is that the distribution of job times is exponential. This assumption is needed because the exponential distribution is the only one having the *memoryless property.* (This property is discussed in detail in Chapter 13 on reliability and maintenance.) The requirement that job times be exponentially distributed is severe in the context of scheduling. In most job shops, if processing times cannot be predicted accurately in advance, it is unlikely that they would be exponentially distributed. Why? The memoryless property tells us that the probability that a job is completed in the next instant of time is independent of the length of time already elapsed in processing the job. Certain applications, such as telephone systems and shared computer applications, may be accurately modeled in this manner, but by and large the exponential law would not accurately describe processing times for most manufacturing job shops.

With these caveats in mind, we present several results for scheduling jobs on multiple machines with random job times. Consider the following problem: n jobs are to be processed through two identical parallel machines. Each job needs to be processed only once on either machine. The objective is to minimize the expected time that elapses from time zero until the last job has completed processing. This is known as the expected makespan. We assume that the n jobs have processing times $t_1, t_2, \ldots, t_n$, which are exponential random variables with rates $\mu_1, \mu_2, \ldots, \mu_n$. This means that the expected time required to complete job i, $E(t_i)$, is $1/\mu_i$.

Parallel processing is different from flow shop processing. In flow shop processing, the jobs are processed first on machine 1 and then on machine 2. In parallel processing, the jobs need to be processed on only one machine, and any job can be processed on either machine. Assume that at time $t = 0$ machine 1 is occupied with a prior job, job 0, and the remaining processing time of job 0 is t_0, which could be a random or deterministic variable. The remaining jobs are processed as follows: let [1], [2], . . . , [n] be a permutation of the n jobs. Job [1] is scheduled on the vacant machine. Job [2] follows either job 0 on machine 1 or job [1] on machine 2, depending on which is completed first. Each successive job is then scheduled on the next available machine.

Let $T_0 \leq T_1 \leq \cdots \leq T_n$ be the completion times of the successive jobs. The makespan is the time of completion of the last job, which is T_n. The expected value of the makespan is minimized by using the longest-expected-processing-time-first rule (LEPT).[1] Note that this is exactly the opposite of the SPT rule for a single machine. The optimality of scheduling the jobs in decreasing order of their expected size rather than in increasing order (as the SPT rule does) is more likely a result of parallel processing than of randomness of the job times.

[1] However, if we consider flow time, then SEPT (shortest expected processing time first) minimizes the expected flow time on two machines.

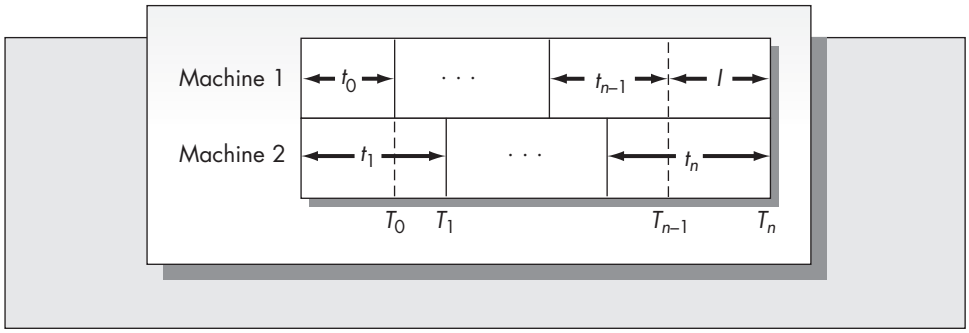


FIGURE 9–12
Realization of parallel processing on two machines with random job times

We intuitively show the optimality of LEPT for this problem as follows: Consider the schematic diagram in Figure 9–12, which gives a particular realization of the processing times for an arbitrary sequencing of the jobs. In Figure 9–12 the random variable I corresponds to the idle time of the machine that does not process the job completed last. Intuitively, we would like to make I as small as possible in order to minimize the expected makespan. This can be shown more rigorously as follows.

From the picture, it is clear that

$$T_n + T_{n-1} = \sum_{i=0}^{n} t_i$$

and

$$T_n = T_{n-1} - I.$$

Solving for T_{n-1} in the second equation and substituting into the first gives

$$T_n + T_n - I = \sum_{i=0}^{n} t_i$$

or

$$2T_n = \sum_{i=0}^{n} t_i + I.$$

As Σt_i is fixed independent of the processing sequence, it follows that minimizing $E(T_n)$ is equivalent to minimizing $E(I)$. Because I is minimized when the processing time of the last job is minimized, we schedule the jobs in order of decreasing expected processing time. Note that this result does *not* necessarily carry over to the case of parallel processing on two machines with deterministic processing times. However, in the deterministic case, scheduling the longest job first will generally give good results when minimizing total makespan. SPT is superior for minimizing mean flow time.

A class of problems we will not discuss, but for which there are several interesting results, are problems in which jobs are to be processed through m nonidentical processors and the processing time does not depend on the job. See Righter (1988) for the characterization of the optimal scheduling rule for this problem.

The Two-Machine Flow Shop Case

An interesting question is whether there is a stochastic analog to Johnson's algorithm for scheduling n jobs on two machines in a flow shop setting; that is, where each job must be processed through machine 1 first, then through machine 2. Johnson's algorithm tells us that job i precedes job $i + 1$ if

$$\min(A_i, B_{i+1}) < \min(A_{i+1}, B_i)$$

in order to minimize the makespan.

Now suppose that $A_1, A_2, \ldots, A_n$ and $B_1, B_2, \ldots, B_n$ are exponential random variables with respective rates $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$. We now wish to minimize the expected value of the makespan. As the minimum of two exponential random variables has a rate equal to the sum of the rates, it follows that

$$E[\min(A_i, B_{i+1})] = \frac{1}{a_i + b_{i+1}}.$$

$$E[\min(A_{i+1}, B_i)] = \frac{1}{a_{i+1} + b_i}.$$

It follows that Johnson's condition translates in the stochastic case to the condition that

$$a_i - b_i \geq a_{i+1} - b_{i+1},$$

so that the jobs should be scheduled in the order of decreasing values of the difference in the rates.

Example *9.9*

Consider Example 9.5, used to illustrate Johnson's algorithm, but let us assume that the job times are random variables having the exponential distribution with mean times given in the example. Hence, we have the following:

	Expected Times		Rates		
Job	**A**	**B**	**A**	**B**	**Differences**
1	5	2	0.20	0.500	−0.30
2	1	6	1.00	0.170	0.83
3	9	7	0.11	0.140	−0.03
4	3	8	0.33	0.125	0.21
5	10	4	0.10	0.250	−0.15

Ordering the jobs according to decreasing values of the differences in the final column results in the sequence

2–4–3–5–1.

which is exactly the same sequence we found in the deterministic case using Johnson's algorithm.

This section considered several solution procedures when job times are random variables. Even when job times are known with certainty, randomness resulting from other sources may still be present. For example, when considering scheduling as a dynamic problem, one must determine the pattern of arrivals to the system. It is common for jobs to arrive according to some random process and queue up for service. Queueing theory and simulation are useful tools for dealing with randomness of this type. Conway, Maxwell, and Miller (1967) discuss the application of both simulation and queueing to operations scheduling problems.

Problems for Section 9.8

18. Consider Example 9.2 in Section 9.5 on determining the optimal sequence to land the planes. Suppose that the landing times are random variables with standard deviation equal to one-third of the mean in each case.

a. In what sequence should the planes be landed in order to minimize the expected average weighted flow time, if the weights to be used are the reciprocals of the number of passengers on each plane?

b. For the sequence you found in part (*a*), what is the probability that all planes are landed within 100 minutes? Assume that the landing times are independent normally distributed random variables. Will your answer change if the planes are landed in a different sequence?

19. A computer center has two identical computers for batch processing. The computers are used as parallel processors. Job times are estimated by the user, but experience has shown that an exponential distribution gives an accurate description of the actual job times. Suppose that at a point in time there are eight jobs remaining to be processed with the following expected job times (expressed in minutes):

Job	1	2	3	4	5	6	7	8
Expected time	4	8	1	50	1	30	20	6

a. In what sequence should the jobs be processed in order to minimize the expected completion time of all eight jobs (i.e., the makespan)?

b. Assume that computer A is occupied with a job that has exactly two minutes of processing time remaining and computer B is idle. If job times are deterministic, show the start and end times of each job on each computer using the sequence derived in part (*a*).

20. Six ships are docked in a harbor awaiting unloading. The times required to unload the ships are random variables with respective means of 0.6, 1.2, 2.5, 3.5, 0.4, and 1.8 hours. The ships are given a priority weighting based on tonnage. The respective tonnages are 12, 18, 9, 14, 4, and 10. In what sequence should the ships be unloaded in order to minimize the expected weighted time?

21. Solve Problem 13 assuming that the times required by John and Marsha are exponentially distributed random variables with expected times given in Problem 13.

22. Five sorority sisters plan to attend a social function. Each requires hair styling and fitting for a gown. Assume that the times required are independent exponentially distributed random variables with the mean times for the fittings of 0.6, 1.2, 1.5, 0.8, and 1.1 hours, respectively, and mean times for the stylings of 0.8, 1.6, 1.0, 0.7, and 1.3 hours, respectively. Assume that the fittings are done before the stylings and that there is only a single hair stylist and a single seamstress available. In what sequence should they be scheduled in order to minimize the total expected time required for fittings and stylings?

9.9 STOCHASTIC SCHEDULING: DYNAMIC ANALYSIS

The scheduling algorithms discussed thus far in this chapter are based on the assumption that all jobs arrive for processing simultaneously. In practice, however, scheduling jobs on machines is a dynamic problem. We use the term *dynamic* here to mean that jobs are arriving randomly over time, and decisions must be made on an ongoing basis as to how to schedule those jobs.

Queueing theory provides a means of modeling some dynamic scheduling problems. Chapter 7 and Supplement 2, provide a review of basic queueing theory. In this section familiarity with the results presented in Supplement 2 is assumed.

Consider the following problem. Jobs arrive completely at random to a single machine. This means that the arrival process is a *Poisson* process. Assume that the mean arrival rate is λ. We initially will assume that processing times are exponentially distributed with mean $1/\mu$. This means that the average processing rate is μ and processing times are independent identically distributed exponential random variables. Finally, we assume that jobs are processed on a first-come, first-served (FCFS) basis. In queueing terminology, we are assuming an M/M/1/FCFS queue. Other processing sequences also will be considered.

Basic queueing theory answers several questions about the performance characteristics of this scheduling problem. First, the probability distribution of the number of jobs in the system (the number waiting to be processed plus the number being processed) in the steady state is known to be geometric with parameter $\rho = \lambda/\mu$. That is, if L is the number of jobs in the system in steady state, then

$$P\{L = i\} = \rho^i(1 - \rho) \qquad \text{for } i = 0, 1, 2, 3, \ldots$$

The expected number of jobs in the system is $\rho/(1 - \rho)$. This implies that a solution exists only for $\rho < 1$. Intuitively this makes sense: the rate at which jobs arrive in the system must be less than the rate at which they are processed to guarantee that the queue does not grow without bound.[2]

Minimizing mean flow time is a common objective not only in static scheduling, but also in dynamic scheduling. The flow time of a job begins the instant the job joins the queue of unprocessed jobs and continues until its processing is completed. For the dynamic scheduling problem, the flow time of a job is a random variable; it depends on the realization of the processing times of preceding jobs as well as its own processing time. The queueing term for the flow time of a job is the waiting time in the system and is denoted by the symbol W. Supplement 2 shows that the distribution of the flow time for the M/M/1/FCFS queue is exponential with parameter $\mu - \lambda$. That is,

$$P\{W > t\} = e^{-(\mu-\lambda)t} \qquad \text{for all } t > 0.$$

Also derived are the distribution for the waiting time in the queue and the expected number of jobs in the queue waiting to be processed. We can see the application of these formulas to the dynamic scheduling problem in the following example.

Example *9.10* A student computer laboratory has a single laser printer. Jobs queue up on the printer from the network server in the lab and are completed on a first-come, first-served basis. The average printing job requires four minutes, but the times vary considerably. Experience has shown that the distribution of times closely follows an exponential distribution. At peak times, about 12 students per hour require use of the printer, but the arrival of jobs to the printer can be assumed to occur completely at random.

Assuming a peak traffic period, determine the following:

a. The average number of jobs in the printer queue.

b. The average flow time of a job.

[2] What is not obvious, however, is what happens at the boundary when $\lambda = \mu$ or $\rho = 1$. It turns out that when the processing and the arrival rates are equal (and interarrival times and job processing times are random), the queue still grows without bound. The reason for this is certainly not obvious, but it appears to be a consequence of the randomness of the arrivals and processing times.

c. The probability that a job will wait more than 30 minutes before it begins processing.

d. The probability that there are more than six jobs in the system.

Solution

First we must determine the service and arrival rates. As each printing job requires an average of 4 minutes, it follows that the service *rate* is $\mu = 1/4$ per minute or 15 per hour. The arrival rate is given as $\lambda = 12$ per hour. The traffic intensity $\rho = 12/15 = .8$.

a. The average number of jobs in the queue is l_q which is given by (refer to Supplement 2)

$$l_q = \frac{\rho^2}{1 - \rho} = \frac{.64}{.2} = 3.2.$$

b. The average flow time of a job is the same as the waiting time in the system. From Chapter 7,

$$w = \frac{\rho}{\lambda(1 - \rho)} = \frac{.8}{(12)(.2)} = 0.3333 \text{ hour (20 minutes)}.$$

c. Here we are interested in $P\{W_q > 0.5\}$. The distribution of W_q is essentially exponential with parameter $\mu - \lambda$, but with positive mass at zero.

$$P\{W_q > t\} = \rho e^{-(\mu-\lambda)t} = .8e^{-(3)(0.5)} = .1785.$$

d. Here we wish to determine $P\{L > 6\}$. From the solution of Example 7.5 in Chapter 7, we showed that

$$P\{L > k\} = \rho^{k+1} = .8^7 = .2097.$$

Selection Disciplines Independent of Job Processing Times

Although our sense of fair play says that the service discipline should be FCFS, there are many occasions when jobs are processed in other sequences. For example, consider a manufacturing process in which parts are stacked as they are completed. The next stage in the process may simply take the parts from the top of the stack, resulting in a last-come, first-served (LCFS) discipline. Similarly, one could envision a situation in which completed parts are thrown into a bin and taken out at random, resulting in service occurring in a random order.

Queueing theory tells us that as long as the selection discipline *does not depend on the processing times,* the mean flow times (and hence mean numbers in the system and mean queue lengths) are the same.[3] However, the **variance** of the flow times does depend on the selection discipline. The flow time variance is greatest with LCFS and least with FCFS among the three disciplines FCFS, LCFS, and random.[4] The second moments of the flow time are given by

$$E_{\text{LCFS}}(W^2) = \frac{1}{1 - \rho} E_{\text{FCFS}}(W^2),$$

$$E_{\text{RANDOM}}(W^2) = \frac{1}{1 - \rho/2} E_{\text{FCFS}}(W^2).$$

(The second equation has been proved for exponential processing times only, but it has been conjectured that it holds in general.) Recall that the variance of a random

[3] Many of the results quoted in this section are discussed more fully in Chapter 9 of Conway, Maxwell, and Miller (1967). In fact, even though this book is over 40 years old, it still provides the most comprehensive treatment of the application of queueing theory to scheduling problems.

[4] This is strictly true only if we eliminate the possibility of preemption. Preemption means that a newly arriving job is allowed to interrupt the service of a job already in progress. We will not treat preemptive disciplines here.

variable is the second moment minus the mean squared, so we can obtain the variance of W directly from these formulas.

Example *9.10 (continued)*

Determine the variance of the flow times for the laser printer during peak hours assuming (1) FCFS, (2) LCFS, and (3) random selection disciplines.

Solution

Because under FCFS the flow time is exponentially distributed with parameter $\mu - \lambda$, it follows that the mean flow time is $1/(\mu - \lambda)$ and the variance of the flow time is $1/(\mu - \lambda)^2$.

$$E_{\text{FCFS}}(W) = \frac{1}{\mu - \lambda} = \frac{1}{15 - 12} = \frac{1}{3} \text{ hour.}$$

$$\text{Var}_{\text{FCFS}}(W) = \frac{1}{(3)^2} = \frac{1}{9}.$$

Because $\text{Var}(W) = E(W)^2 - (E(W))^2$, it follows that

$$E_{\text{FCFS}}(W)^2 = \text{Var}_{\text{FCFS}}(W) + (E_{\text{FCFS}}(W))^2 = \frac{1}{9} + \left(\frac{1}{3}\right)^2 = \frac{2}{9}.$$

From these results, we obtain

$$E_{\text{LCFS}}(W)^2 = \left(\frac{1}{1 - .8}\right)\left(\frac{2}{9}\right) = \frac{10}{9}, \quad \text{giving}$$

$$\text{Var}_{\text{LCFS}}(W) = \frac{10}{9} - \left(\frac{1}{3}\right)^2 = 1.0.$$

Similarly,

$$E_{\text{RANDOM}}(W)^2 = \left(\frac{1}{1 - .4}\right)\left(\frac{2}{9}\right) = 0.3704, \quad \text{giving}$$

$$\text{Var}_{\text{RANDOM}}(W) = 0.3704 - \left(\frac{1}{3}\right)^2 = 0.2593.$$

Hence, we see that if the jobs are processed on the printer on an LCFS basis, the variance of the flow time is 1.0, as compared to 2/9 for FCFS. Processing jobs in a random order also increases the variance of the flow time over FCFS, but by a much smaller degree. Intuitively, the variance is larger for LCFS than for FCFS, because when jobs are processed in the opposite order of their arrival, it is likely that a job that has been in the queue for a while will continue to be "bumped" by newly arriving jobs. The result will be a very long flow time. A similar phenomenon occurs in the random case, but the effect is not as severe.

Selection Disciplines Dependent on Job Processing Times

One of the goals of research into dynamic queueing models is to discover optimal selection disciplines. Previously we stated that the average measures of performance (w, w_q, l, and l_q) are independent of the selection discipline as long as the selection discipline does not depend on the job processing times. However, it was shown in Chapter 7 that selection based on shortest processing times can make a large difference to average waiting time. Furthermore, consider the case in which job processing times are realized at the instant the job joins the queue. This assumption is reasonable for most industrial scheduling applications. For machine processing problems, the work content is likely to be a multiple of the number of parts that have to be processed, so the processing time is known at the instant a job joins the queue. An example in which this does not hold

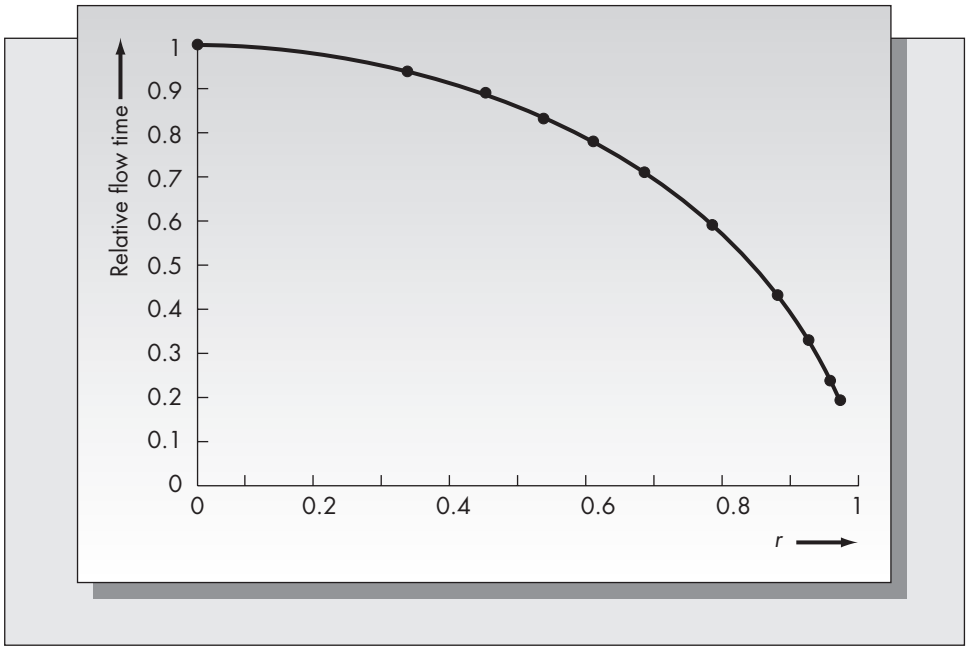


FIGURE 9–13
Relative flow times: SPT versus FCFS

Source: Adapted from tabled results in Conway, Maxwell, and Miller (1967), p. 184.

would be a bank. It is generally not possible to tell how long a customer will require for service until he or she enters service and reveals the nature of the transaction.

When job processing times are realized when a job joins the queue, it is possible to use a selection discipline that is dependent on job times. One such discipline, discussed at length earlier in this chapter, is the SPT rule: The next job processed is the one with the shortest processing time. It turns out that the SPT rule is effective for the dynamic problem as well as the static problem.

SPT scheduling can significantly reduce the size of the queue in the dynamic scheduling problem. In Figure 9–13 we show just how dramatic this effect can be. This figure assumes an M/M/1 queue. Define the relative flow time as the ratio

$$\frac{E_{\text{SPT}}(\text{Flow Time})}{E_{\text{FCFS}}(\text{Flow Time})}.$$

We see that as the traffic intensity increases, the advantage of SPT at reducing flow time (and hence the number in the system and the queue length) improves. The queue is reduced by "cleaning out" the shorter jobs. For values of ρ near 1, this ratio could be as low as 0.2. Because of Little's formula (see Chapter 7), the expected number in the system and the flow time are proportional, so this curve also represents the ratio of the expected numbers in the system for the respective selection disciplines.

An interesting question is what happens to the variance of the performance measures under each selection discipline. Again assuming exponential service times, Table 9–1 gives the variances of the flow times for SPT and FCFS as a function of the traffic intensity, ρ.

What we see from this table is that for low values of the traffic intensity (less than about .7), SPT has slightly lower variance than FCFS. However, as the traffic intensity approaches 1, the variance under SPT increases dramatically. For $\rho = .99$, the variance of the flow time under SPT is more than 16 times that for FCFS. Hence, the reduction in the mean flow time achieved by SPT comes at the cost of possibly increasing the variance.

TABLE 9–1
Variance of the Flow Time under FCFS and SPT

Source: Conway, Maxwell, and Miller (1967), p. 189.

ρ	FCFS	SPT
.1	1.2345	1.179
.2	1.5625	1.482
.3	2.0408	1.896
.4	2.6666	2.563
.5	4.0000	3.601
.6	6.2500	5.713
.7	11.1111	12.297
.8	25	32.316
.9	100	222.2
.95	400	1,596.5
.98	2,500	22,096
.99	10,000	161,874

A large portion of the research in dynamic scheduling considers priority disciplines, which means that incoming jobs are classified into groups and priority is given to certain groups over others. Priority scheduling is common in hospital emergency rooms. It also occurs in scheduling batch jobs on a mainframe computer, as certain users may be given priority over others. Priorities may be preemptive or nonpreemptive. Preemptive priority means that a newly arriving job with a higher priority than the job currently being processed is permitted to interrupt the service of the job in process. The interrupted job may continue where it left off at a later time, or it may have to start processing again from scratch. We will not pursue this complex area, but note that SPT is quite robust; it is optimal for a large class of priority scheduling problems.

The $c\mu$ Rule

Consider the following scheduling problem. Jobs arrive randomly to a single machine with exponential processing times. We allow the jobs in the queue to have different service rates μ_i. That is, at any moment, suppose that there are n jobs waiting to be processed. We index the jobs 1, 2, 3, . . . , n and assume that the time required to complete job i has the exponential distribution with mean $1/\mu_i$. In addition, suppose that there is a return of c_i if job i is completed by some fixed time t. The issue is, what is the best choice for the next job to be processed if the objective is to maximize the total expected earnings?

Derman et al. (1978) showed that the optimal policy is to choose the job with the largest value of $c_i\mu_i$. Notice that if the weights are set equal to 1, the $c\mu$ rule is exactly the same as SPT in expectation. Hence, this can be considered to be a type of weighted SPT scheduling rule. It turns out that the $c\mu$ rule is optimal for several other versions of the stochastic scheduling problem. We refer the interested reader to Pinedo (1983) and the references listed there.

Problems for Section 9.9

23. A computer system queues up batch jobs and processes them on an FCFS basis. Between 2 and 5 P.M., jobs arrive at an average rate of 30 per hour and require an average of 1.2 minutes of computer time. Assume the arrival process is Poisson and the processing times are exponentially distributed.

    a. What is the expected number of jobs in the system and in the queue in the steady state?

b. What are the expected flow time and the time in the queue in the steady state?
c. What is the probability that the system is empty?
d. What is the probability that the queue is empty?
e. What is the probability that the flow time of a job exceeds 10 minutes?

24. Consider the computer system of Problem 23.
    a. Compute the variance of the flow times assuming FCFS, LCFS, and random selection disciplines.
    b. Using a normal approximation of the flow time distribution for the LCFS and random cases, estimate the probability that the flow time in the system exceeds 10 minutes in each case.
25. A critical resource in a manufacturing operation experiences a very high traffic intensity during the shop's busiest periods. During these periods the arrival rate is approximately 57 jobs per hour. Job processing times are approximately exponentially distributed with mean 1 minute.
    a. Compute the expected flow time in the system assuming an FCFS processing discipline, and the expected flow time under SPT using Figure 9–13.
    b. Compute the probability that a job waits more than 30 minutes for processing under FCFS.
    c. Using a normal approximation, estimate the probability that a job waits more than 30 minutes for processing under LCFS.

9.10 ASSEMBLY LINE BALANCING

The problem of balancing an assembly line is a classic industrial engineering problem. Even though much of the work in the area goes back to the mid-1950s and early 1960s, the basic structure of the problem is relevant to the design of production systems today, even in automated plants. The problem is characterized by a set of n distinct tasks that must be completed on each item. The time required to complete task i is a known constant t_i. The goal is to organize the tasks into groups, with each group of tasks being performed at a single workstation. In most cases, the amount of time allotted to each workstation is determined in advance, based on the desired rate of production of the assembly line. This is known as the cycle time and is denoted by C. A schematic of a typical assembly line is given in Figure 9–14. In the figure, circles represent tasks to be done at the corresponding stations.

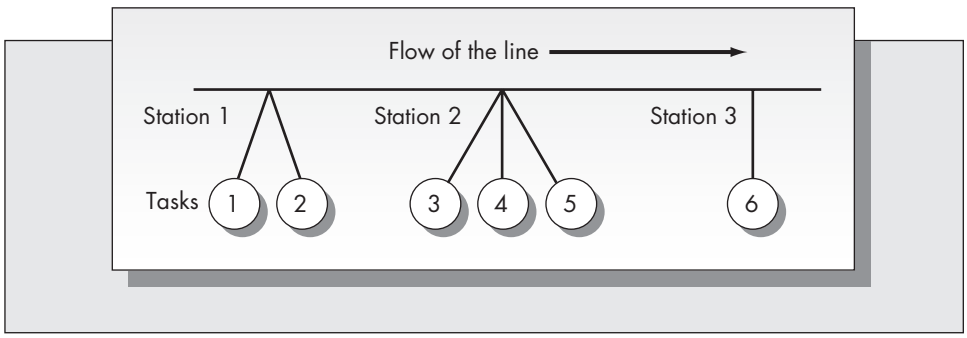


FIGURE 9–14
Schematic of a typical assembly line

Assembly line balancing is traditionally thought of as a facilities design and layout problem, and one might argue that it would be more appropriately part of Chapter 11. Assigning tasks to workstations has traditionally been a one-shot decision made at the time the plant is constructed and the line is set up. However, the nature of the modern factory is changing. New plants are being designed with flexibility in mind, allowing new lines to be brought up and old ones restructured on a continuous basis. In such an environment, line balancing is more like a dynamic scheduling problem than a one-shot facilities layout problem.

There are a variety of factors that contribute to the difficulty of the problem. First, there are precedence constraints: some tasks may have to be completed in a particular sequence. Another problem is that some tasks cannot be performed at the same workstation. For example, it might not be possible to work on the front end and the back end of a large object such as an automobile at the same workstation. This is known as a *zoning restriction.* Still other complications might arise. For example, certain tasks may have to be completed at the same workstation, and other tasks may require more than one worker.

Finding the optimal balance of an assembly line is a difficult combinatorial problem even when the problems previously described are not present. Several relatively simple heuristics have been suggested for determining an approximate balance. Many of these methods require few calculations and make it possible to solve large problems by hand.

Let $t_1, t_2, \ldots, t_n$ be the time required to complete the respective tasks. The total work content associated with the production of an item, say T, is given by

$$T = \sum_{i=1}^{n} t_i.$$

For a cycle time of C, the minimum number of workstations possible is $[T/C]$, where the brackets indicate that the value of T/C is to be rounded to the next larger integer. Because of the discrete and indivisible nature of the tasks and the precedence constraints, it is often true that more stations are required than this ideal minimum value. If there is leeway in the choice of the cycle time, it is advisable to experiment with different values of C to see if a more efficient balance can be obtained.

We will present one heuristic method from Helgeson and Birnie (1961) known as the *ranked positional weight technique.* The method places a weight on each task based on the total time required by all the succeeding tasks. Tasks are assigned sequentially to stations based on these weights. We illustrate the method by example.

Example *9.11* The final assembly of Noname personal computers, a generic mail-order PC clone, requires a total of 12 tasks. The assembly is done at the Lubbock, Texas, plant using various components imported from the Far East. The tasks required for the assembly operations are

1. Drill holes in the metal casing and mount the brackets to hold disk drives.
2. Attach the motherboard to the casing.
3. Mount the power supply and attach it to the motherboard.
4. Place the main processor and memory chips on the motherboard.
5. Plug in the graphics card.
6. Mount the DVD burner. Attach the controller and the power supply.
7. Mount the hard disk drive. Attach the hard disk controller and the power supply to the hard drive.
8. Set switch settings on the motherboard for the specific configuration of the system.
9. Attach the monitor to the graphics board prior to running system diagnostics.
10. Run the system diagnostics.

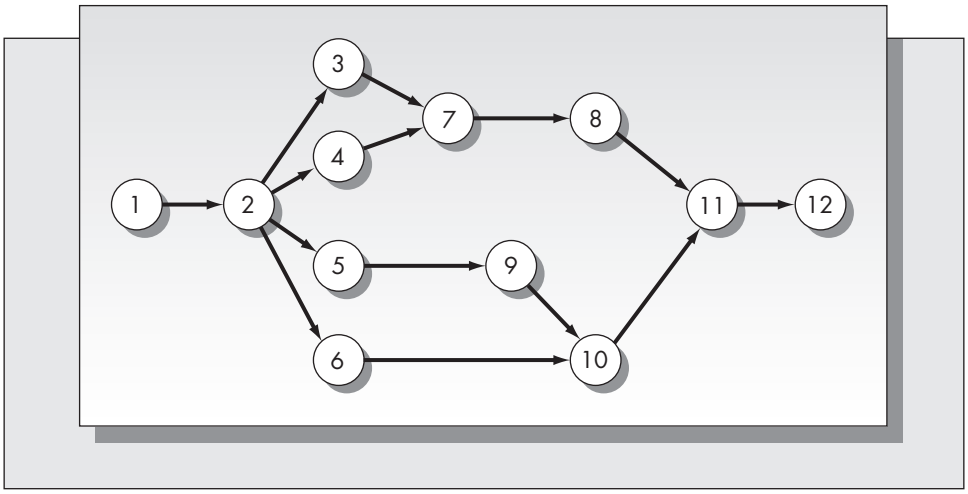


FIGURE 9–15
Precedence constraints for Noname computer (Example 9.11)

11. Seal the casing.
12. Attach the company logo and pack the system for shipping.

The holes must be drilled and the motherboard attached to the casing prior to any other operations. Once the motherboard has been mounted, the power supply, memory, processor chips, graphics card, and disk controllers can be installed. The floppy drives are placed in the unit prior to the hard drive and require that the power supply and controller be in place first. Based on the memory configuration and the choice of graphics adapter, the switch settings on the motherboard are determined and set. The monitor must be attached to the graphics board so that the results of the diagnostic tests can be read. Finally, after all other tasks are completed, the diagnostics are run and the system is packed for shipping. The job times and precedence relationships for this problem are summarized in the following table. The network representation of this particular problem is given in Figure 9–15.

Task	Immediate Predecessors	Time
1	—	12
2	1	6
3	2	6
4	2	2
5	2	2
6	2	12
7	3, 4	7
8	7	5
9	5	1
10	9, 6	4
11	8, 10	6
12	11	7

Suppose that the company is willing to hire enough workers to produce one assembled machine every 15 minutes. The sum of the task times is 70, which means that the minimum number of workstations is the ratio 70/15 = 4.67 rounded to the next larger integer, which is 5. This does not mean that a five-station balance necessarily exists.

The solution procedure requires determining the positional weight of each task. The positional weight of task *i* is defined as the time required to perform task *i* plus the times required to perform all tasks having task *i* as a predecessor. As task 1 must precede all other tasks, its positional weight is simply the sum of the task times, which is 70. Task 2 has positional weight 58. From Figure 9–15

TABLE 9–2
Positional Weights for Example 9.11

Task	Positional Weight
1	70
2	58
3	31
4	27
5	20
6	29
7	25
8	18
9	18
10	17
11	13
12	7

we see that task 3 must precede tasks 7, 8, 11, and 12, so that the positional weight of task 3 is $t_3 + t_7 + t_8 + t_{11} + t_{12} = 31$. The other positional weights are computed similarly. The positional weights are listed in Table 9–2.

The next step is to rank the tasks in the order of decreasing positional weight. The ranking for this case is 1, 2, 3, 6, 4, 7, 5, 8, 9, 10, 11, 12. Finally, the tasks are assigned sequentially to stations in the order of the ranking, and assignments are made only as long as the precedence constraints are not violated.

Let us now consider the balance obtained using this technique assuming a cycle time of 15 minutes. Task 1 is assigned to station 1. That leaves a slack of three minutes at this station. However, because task 2 must be assigned next, in order not to violate the precedence constraints, and the sum $t_1 + t_2$ exceeds 15, we close station 1. Tasks 2, 3, and 4 are then assigned to station 2, resulting in an idle time of only one minute at this station. Continuing in this manner, we obtain the following balance for this problem:

Station	1	2	3	4	5	6
Tasks	1	2, 3, 4	5, 6, 9	7, 8	10, 11	12
Idle time	3	1	0	3	5	8

Notice that although the minimum possible number of stations for this problem is five, the ranked positional weight technique results in a six-station balance. As the method is only a heuristic, it is possible that there is a solution with five stations. In this case, however, the optimal balance requires six stations when $C = 15$ minutes.

The head of the firm assembling Noname computers is interested in determining the minimum cycle time that would result in a five-station balance. If we increase the cycle time from $C = 15$ to $C = 16$, then the balance obtained is

Station	1	2	3	4	5
Tasks	1	2, 3, 4, 5	6, 9	7, 8, 10	11, 12
Idle time	4	0	3	0	3

This is clearly a much more efficient balance: The total idle time has been cut from 20 minutes per unit to only 10 minutes per unit. The number of stations decreases by about 16 percent, while the cycle time increases by only about 7 percent. Assuming that a production day is seven hours, a value of $C = 15$ minutes would result in a daily production level of 28 units per assembly operation and a value of $C = 16$ minutes would result in a daily production level of 26.25 units per assembly operation. Management would have to determine whether the decline in the production rate of 1.75 units per day per operation is justified by the savings realized with five rather than six stations.

An alternative choice is to stay with the six stations, but see if a six-station balance can be obtained with a cycle time less than 15 minutes. It turns out that for values of the cycle time of both 14 minutes and 13 minutes, the ranked positional weight method will give six-station balances. The $C = 13$ solution is

Station	1	2	3	4	5	6
Tasks	1	2, 3	6	4, 5, 7, 9	8, 10	11, 12
Idle time	1	1	1	1	4	0

Thirteen minutes appear to be the minimum cycle time with six stations. The total idle time of eight minutes resulting from the balance above is two minutes less than that achieved with five stations when $C = 16$. The production rate with six stations and $C = 13$ would be 32.3 units per day per operation. Increasing the number of stations from five to six results in a substantial improvement in the throughput rate.

In this section we presented the ranked positional weight heuristic for solving the assembly line balancing problem. Other heuristic methods exist as well. One is COMSOAL, a computer-based heuristic developed by Arcus (1966). The method is efficient for large problems involving many tasks and workstations. Kilbridge and Wester (1961) suggest a method similar to the ranked positional weight technique.

There are optimal procedures for solving the line balancing problem, but the calculations are complex and time-consuming, requiring either dynamic programming (Held et al., 1963) or integer programming (Thangavelu and Shetty, 1971). More recent interest in the line balancing problem has focused on issues relating to uncertainty in the performance times for the individual tasks. (See Hillier and Boling, 1986, and the references contained there.)

Virtually all assembly line balancing procedures assume that the objective is to minimize the total idle time at all workstations. However, as we saw in this section, an optimal balance for a fixed cycle time may not be optimal in a global sense. Carlson and Rosenblatt (1985) suggest that most assembly line balancing procedures are based on an incorrect objective. The authors claim that maximizing profit (rather than minimizing idle time) would give a different solution to most assembly line balancing problems, and they present several models in which both numbers of stations and cycle time are decision variables.

Problems for Section 9.10

26. Consider the example of Noname computers presented in this section.
    a. What is the minimum cycle time that is possible? What is the minimum number of stations that would theoretically be required to achieve this cycle time?
    b. Based on the ranked positional weight technique, how many stations are actually required for the cycle time indicated in part (*a*)?
    c. Suppose that the owner of the company that sells Noname computers finds that he is receiving orders for approximately 100 computers per day. How many separate assembly lines are required assuming (*i*) the best five-station balance, (*ii*) the best six-station balance (both determined in the text), and (*iii*) the balance you obtained in part (*b*)? Discuss the trade-offs involved with each choice.
27. A production facility assembles inexpensive telephones on a production line. The assembly requires 15 tasks with precedence relationships and activity times as shown in Figure 9–16. The activity times appear next to the node numbers in the network.

FIGURE 9–16
(For Problem 27)

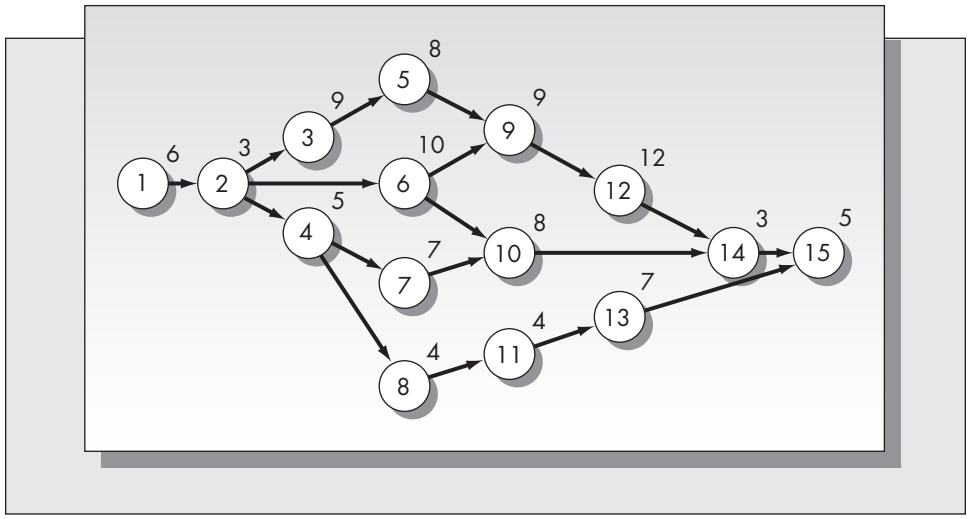
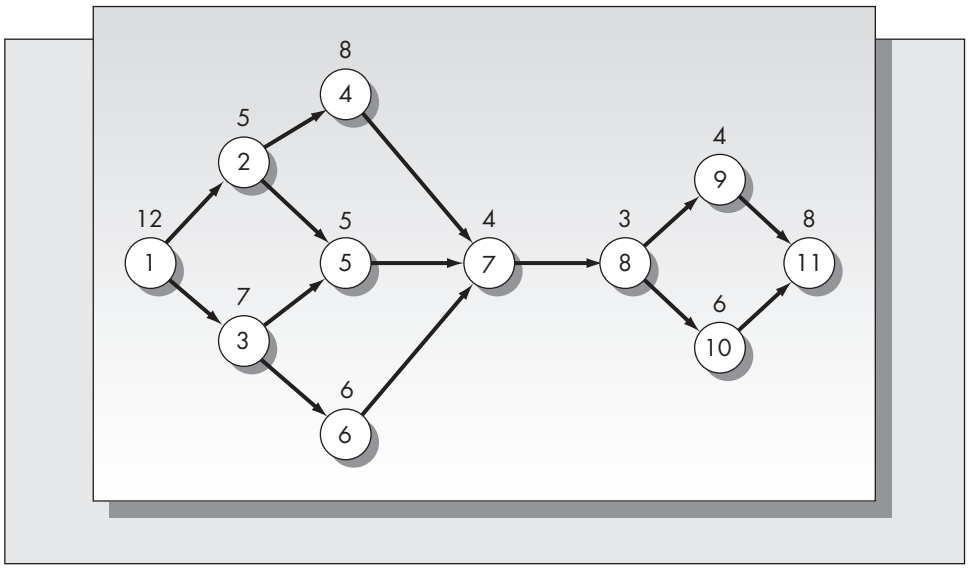


FIGURE 9–17
(For Problem 29)



a. Determine positional weights for each of the activities.

b. For a cycle time of 30 units, what is the minimum number of stations that could be achieved? Find the $C = 30$ balance obtained using the ranked positional weight technique.

c. Is there a solution with the same number of stations you found in part (*b*) but with a lower cycle time? In particular, what appears to be the minimum cycle time that gives a balance with the same number of stations you found in part (*b*)?

28. For the data given in Problem 27, determine by experimentation the minimum cycle time for a three-station balance.

29. Consider the assembly line balancing problem represented by the network in Figure 9–17. The performance times are shown above the nodes.

a. Determine a balance for $C = 15$.

b. Determine a balance for $C = 20$.

Snapshot Application

MANUFACTURING DIVISIONS REALIZE SAVINGS WITH SCHEDULING SOFTWARE

Motorola Goes for MOOPI

Motorola Corporation has adopted MOOPI from Berclain Group of Canada to schedule the remanufacture of 1,000 engine controllers each month in the Automotive and Industrial Electronics Group in its Seguin, Texas, plant. Utilization of used engine controllers has improved from 35 percent to over 70 percent and the time required for scheduling has reduced from 100 to only 8 hours per month as a consequence of implementing the software. The problem of scheduling cores remanufacture is a very complex one, requiring five or six employees when it was done manually. According to Eileen Svoboda, a process improvement manager:

> Remanufacturing in itself is a unique and complex process because the raw material—the core—arrives randomly through time, and the material required to remanufacture it is unknown until after the core is inspected. Right now in our business we have 7,000 Motorola core model numbers that can be remanufactured into approximately 800 different customer models. To keep all this straight in the old manual system we relied on paper trails, as well as the memory and expertise of a few key individuals in Texas, Michigan, and Illinois.

The software allows the group to schedule one month's worth of production in detail, as well as some additional planning functions for a 26-week time horizon. Motorola considered four different scheduling systems. MOOPI was chosen because it could both handle complex material assignment needs and easily run what-if simulations. Although the simulation feature played an important part in the decision, MOOPI is basically an optimization package, and, according to Berclain, Motorola's primary goal was optimization of materials utilization.

The system's database incorporates manufacturing data, including sequence of operations or routings, bills of materials, and setup sequencing information. It distinguishes raw materials, subassembly, and finished goods inventory levels. It can exchange data with other inventory control systems, and it details production orders for all work centers.

H.P. Implements FastMAN for Materials Scheduling

The Hewlett-Packard Corporation of Palo Alto, California, has experienced enormous growth in the highly competitive PC business. Part of this success is attributable to careful management of the materials in the PC product plants.

H.P.'s scheduling system is PC-based and is used in conjunction with its MRP II system. One post-MRP scheduler used by H.P. is FastMAN, produced by the company with the same name based in Chicago, Illinois. Today, H.P. has eight installations of FastMAN and three with APS, a workstation-based system recently made available on PCs. According to Dr. Lee Seaton, a senior industry analyst with H.P.:

> The materials content of a personal computer constitutes over 90 percent of its cost. Unfortunately, we frequently had too much of one thing and too little of something else. And we knew it was only going to get worse, since product life cycles continue to accelerate, recently down from 15 months to under one year. We had to reduce product write-offs and have a smoother cut-off for a given product set. Too often, we'd launch a successful product, but either not have enough to sell or at the end of the cycle have excessive material write-offs.

While material control is normally the domain of MRP II, H.P. was not having the desired level of success with MRP II. Dr. Seaton claims that 30 percent of the MRP runs were considered useless. Updates of inventory levels or bills of materials were not entered into the system in a timely fashion, making the results out of date. The new scheduling tools allowed H.P. employees to verify the validity of MRP runs with desktop systems. Also, the software provided the capability for determining the consequences of business decisions in a spreadsheet or graphical form. Engineering changes could more easily be incorporated into the materials plan, and excess material could be utilized more effectively by constructing "matched sets." The key to H.P.'s success was "getting information into the hands of the decision makers," says Seaton.

Source: These two applications are discussed in Parker (1995).

c. What appears to be the minimum cycle time that can be achieved using the number of stations you obtained in part (*a*)?

d. What appears to be the minimum cycle time that can be achieved using the number of stations you obtained in part (*b*)?

9.11 HISTORICAL NOTES

Interest in studying the effects of different sequencing strategies in the job shop is relatively recent. One of the earliest monographs in the open literature that considered sequencing issues is by Salveson (1952). The first significant published work to present analytical results for optimal sequencing strategies is Johnson (1954). Bellman's (1956) review article discusses sequence scheduling (among other topics) and presents an interesting proof of the optimality of Johnson's algorithm based on dynamic programming arguments. Bellman appears to be the first to have recognized the optimality of a permutation schedule for scheduling n jobs on three machines.

Johnson's work spawned considerable interest in scheduling. The excellent monograph by Conway, Maxwell, and Miller (1967), for example, lists over 200 publications up until 1967. It is likely that considerably more than 200 papers have been published since that time. Much of the recent research in sequence scheduling has focused on stochastic scheduling. Weiss (1982) provides an interesting synopsis of work in this area. That jobs should be scheduled in order of decreasing expected processing time when scheduling is on two parallel processors and the objective is to minimize the expected makespan appears to have been discovered at about the same time by Bruno and Downey (1977) and Pinedo and Weiss (1979). The elegant method of proof we present is from Pinedo and Weiss (1979).

Much of the work on dynamic scheduling under uncertainty appears in the Conway, Maxwell, and Miller (1967) monograph. The $c\mu$ rule, apparently discovered by Derman et al. (1978) in an entirely different context, led to several extensions. See, for example, Righter and Shanthikumar (1989).

9.12 Summary

Scheduling problems arise in many contexts in managing operations. One of the areas explored in this chapter was optimal sequencing rules for scheduling jobs on machines, a problem that arises in *shop floor control.* We also considered dynamic scheduling problems, which are more typical of the kinds of scheduling problems that one encounters in management of service systems.

Much of the emphasis of the chapter was on determining efficient *sequencing rules.* The form of the optimal sequencing rule depends on several factors, including the pattern of arrivals of jobs, the configuration of the job shop, constraints, and the optimization objectives. Typical objectives in job shop management include meeting due dates, minimizing work-in-process, minimizing the average flow time, and minimizing machine idle time.

We discussed four sequencing rules: FCFS (first-come, first-served), SPT (shortest processing time first), EDD (earliest due date), and CR (critical ratio). For sequencing jobs on a single machine, we showed that *SPT* optimized several objectives, including mean flow time. However, SPT sequencing could be problematic. Long jobs would be constantly pushed to the rear of the job queue and might never be processed. For that reason, pure SPT scheduling is rarely used in practice. *Critical ratio scheduling* attempts to balance the importance placed on processing time and the remaining time until the due date. However, there is little evidence to suggest that critical ratio scheduling performs well relative to the common optimization criteria such as mean flow time. As one would expect, *earliest-due-date* scheduling performs best when the goal is to minimize the maximum tardiness.

We considered a variety of algorithms for *single-machine sequencing,* including Moore's (1968) algorithm for minimizing the number of tardy jobs and Lawler's (1973) algorithm for minimizing any nondecreasing function of the flow subject to precedence constraints. An excellent summary of algorithms for sequence scheduling can be found in French (1982). Several techniques for *multiple-machine scheduling* also were presented in this chapter. Johnson's (1954) classic work on scheduling *n jobs through two machines* is discussed, as well as Akers's (1956) graphical method for *two jobs through m machines.*

Stochastic scheduling problems were treated in two contexts: static and dynamic. Static problems are stochastic counterparts to the problems treated earlier in the chapter, except that job completion times are random variables. Most of the simple results for static stochastic scheduling problems require the assumption that job times have the exponential distribution. Because of the memoryless property of the exponential distribution, this assumption is probably not satisfied in most manufacturing job shops. The dynamic case occurs when jobs arrive randomly over time. *Queueing theory,* discussed in detail in Chapter 7 and Supplement 2, is the means for analyzing such problems.

The *assembly line balancing problem* is one in which a collection of tasks must be performed on each item. Furthermore, the tasks must be performed in a specified sequence. The problem is to allot the tasks to workstations on an assembly line. The quality of the balance is measured by the idle time remaining at each station. Determining the best mix of cycle time (amount of time allotted to each station) and number of stations is an extremely difficult analytical problem. We discuss one very simple heuristic solution method from Helgeson and Birnie (1961), known as the ranked positional weight technique, which provides an efficient balance quickly.

While most of this chapter concerns analytical methods for developing schedules, it is important to keep in mind that many real scheduling problems are too complex to be modeled mathematically. In such cases, *simulation* can be a very valuable tool. A computer-based simulation is a model of a real process expressed as a computer program. By running simulations under different scenarios, consequences of alternative strategies can be evaluated easily. Simulations are particularly effective when significant randomness or variation exists.

Job shop *scheduling software* is a growing business. Programs designed to run on PCs and workstations are available from several vendors. Most of these programs provide convenient interfaces to existing enterprise and MRP systems. While the market for the post-MRP schedulers is much smaller than the market for full-blown integrated MRP II systems, it is growing and probably exceeds $100 million in the United States alone. These programs are designed to run in conjunction with an MRP system, and generally incorporate some combination of optimization and simulation. Many companies have been successful at implementing these programs on the factory floor.

There are a number of excellent texts on scheduling. For further reading in this area, we would suggest the books by Baker (1974), French (1982), and Conway, Maxwell, and Miller (1967). Pinedo (1995) provides comprehensive coverage.

Additional Problems on Scheduling

30. Mike's Auto Body Shop has five cars waiting to be repaired. The shop is quite small, so only one car can be repaired at a time. The number of days required to repair each car and the promised date for each are given in the following table.

Cars	Repair Time (days)	Promised Date
1	3	5
2	2	6
3	1	9
4	4	11
5	5	8

Mike has agreed to provide a rental car to each customer whose car is not repaired on time. Compare the performance of the four sequencing rules FCFS, SPT, EDD, and CR relative to minimizing average tardiness.

31. For each of the problems listed, indicate precisely what or who would correspond to jobs and who or what would correspond to machines. In each case discuss what objectives might be appropriate and special priorities that might exist.

    a. Treating patients in a hospital emergency room.

    b. Unloading cargo from ships at port.

    c. Serving users on a time-shared computer system.

    d. Transferring long-distance phone calls from one city to another.

32. Six patients are waiting in a hospital emergency room to receive care for various complaints, ranging from a sprained ankle to a gunshot wound. The patients are numbered in the sequence that they arrived and are given a priority weighting based on the severity of the problem. There is only one doctor available.

Patient	1	2	3	4	5	6
Time required	20 min	2 hr	30 min	10 min	40 min	1 hr
Priority	1	10	3	5	2	2

    a. Suppose that the patients are scheduled on an FCFS basis. Compute the mean flow time and the *weighted* mean flow time, where the weight is the priority number associated with each patient.

    b. Perform the calculations in part (*a*) for SPT sequencing.

    c. Determine a sequence different from those in parts (*a*) and (*b*) that achieves a lower value of the weighted mean flow time.

33. Consider Mike's Auto Body Shop mentioned in Problem 30. What sequencing of the jobs minimizes the

    a. Mean flow time?

    b. Maximum lateness?

    c. Number of tardy jobs?

34. Consider the situation of the emergency room mentioned in Problem 32. Determine the sequence in which patients should be treated in order to minimize the *weighted* value of the mean flow time.

35. Barbara and Jenny's Ice Cream Company produces four different flavors of ice cream: vanilla, chocolate, strawberry, and peanut fudge. Each batch of ice cream is produced in the same large vat, which must be cleaned prior to switching flavors. One day is required for the cleaning process.

    At the current time they have the following outstanding orders of ice cream:

Flavor	**Order Size (gallons)**	**Due Date**
Vanilla	385	3
Chocolate	440	8
Strawberry	200	6
Peanut fudge	180	12

It takes one day to produce the ice cream and a maximum of 100 gallons can be produced at one time. Cleaning is required only when flavors are switched. The production for one flavor will be completed prior to beginning production for another flavor. Cleaning is always started at the beginning of a day.

Treating each ice cream flavor as a different job, find the following:

a. The sequence in which the ice cream flavors should be produced in order to minimize the mean flow time for all of the flavors.

b. The optimal sequence to produce the flavors in order to minimize the number of flavors that are late.

36. Consider Barbara and Jenny's Ice Cream Company, mentioned in Problem 35. Suppose that if vanilla or strawberry is produced after chocolate or peanut fudge, an extra day of cleaning is required. For that reason they decide that the vanilla and strawberry will be produced before the chocolate and peanut fudge.

a. Find the optimal sequencing of the flavors to minimize the maximum lateness using Lawler's algorithm.

b. Enumerate all the feasible sequences and determine the sequence that minimizes the maximum lateness by evaluating and comparing the objective function value for each case.

37. Irving Bonner, an independent computer programming consultant, has contracted to complete eight computer programming jobs. Some jobs must be completed in a certain sequence because they involve program modules that will be linked.

Job	Time Required (days)	Due Date
1	4	June 8
2	10	June 15
3	2	June 10
4	1	June 12
5	8	July 1
6	3	July 6
7	2	June 25
8	6	June 29

Precedence restrictions:

$$1 \rightarrow 2 \rightarrow 5 \rightarrow 6.$$
$$4 \rightarrow 7 \rightarrow 8.$$

Assume that the current date is Monday, June 1, and that Bonner does not work on weekends. Using Lawler's algorithm, find the sequence in which he should be performing the jobs in order to minimize maximum lateness subject to the precedence constraints.

38. William Beebe owns a small shoe store. He has 10 pairs of shoes that require resoling and polishing. He has a machine that can resole one pair of shoes at a time, and the time required for the operation varies with the type and condition of the shoe and the type of sole that is used. Shoes are polished on a machine dedi-cated to this purpose as well, and polishing is always done after resoling. His assistant generally does the polishing while Mr. Beebe does the resoling. The resoling and polishing times (in minutes) are

Shoes	Resoling Time	Polishing Time
1	14	3
2	28	1
3	12	2
4	6	5
5	10	10
6	14	6
7	4	12
8	25	8
9	15	5
10	10	5

In what order should the shoes be repaired in order to minimize the total makespan for these 10 jobs?

39. A leatherworks factory has two punch presses for punching holes in the various leather goods produced in the factory prior to sewing. Suppose that 12 different jobs must be processed on one or the other punch press (i.e., parallel processing). The processing times (in minutes) for these 12 jobs are given in the following table.

Job	1	2	3	4	5	6	7	8	9	10	11	12
Time	26	12	8	42	35	30	29	21	25	15	4	75

Assume that both presses are initially idle. Compare the performance of SPT and LPT (longest processing time) rules for this example. (In parallel processing the next job is simply scheduled on the next available machine.)

40. An independent accountant is planning to prepare tax returns for six of her clients. Prior to her actually preparing each return, her secretary checks the client's file to be sure all the necessary documentation is there and obtains all the tax forms needed for the preparation of the return. Based on past experience with the clients, her secretary estimates that the following times (in hours) are required for preparation of the return and for the accountant to complete the necessary paperwork prior to filing each return:

Client	Secretary Time	Accountant Time
1	1.2	2.5
2	1.6	4.5
3	2.0	2.0
4	1.5	6.0
5	3.1	5.0
6	0.5	1.5

In what order should the work be completed in order to minimize the total time required for all six clients?

41. Five Hong Kong tailors—Simon, Pat, Choon, Paul, and Wu—must complete alterations on a suit for the duke and a dress for the duchess as quickly as possible. On the dress, Choon must first spend 45 minutes cutting the fabric, then Pat will spend 75 minutes sewing the bodice, Simon will need 30 minutes stitching the sleeves, Paul 2 hours lacing the hem, and finally Wu will need 80 minutes for finishing touches. As far as the suit is concerned, Pat begins with shortening the sleeves, which requires 100 minutes. He is followed by Paul, who sews in the lining in 1.75 hours; Wu, who spends 90 minutes sewing on the buttons and narrowing the lapels; and finally Choon, who presses and cleans the suit in 30 minutes.

    Determine precisely when each tailor should be performing each task in order to minimize the total time required to complete the dress and the suit. Assume that the tailors start working at 9 A.M. and take no breaks. Draw the Gantt chart indicating your solution.

42. The assembly of a transistorized clock radio requires a total of 11 tasks. The task times and predecessor relationships are given in the following table.

Task	Time (seconds)	Immediate Predecessors
1	4	
2	38	
3	45	
4	12	1, 2
5	10	2
6	8	4
7	12	5
8	10	6
9	2	7
10	10	8, 9
11	34	3, 10

a. Develop a network for this assembly operation.

b. What is the minimum cycle time that could be considered for this operation? What is the minimum number of stations that could be used with this cycle time?

c. Using the ranked positional weight technique, determine the resulting balance using a cycle time of 45 seconds.

d. Determine by experimentation the minimum cycle time that results in a four-station balance.

e. What is the daily production rate for this product if the company adopts the balance you determined in part (*c*)? (Assume a six-hour day for your calculations.) What would have to be done if the company wanted a higher rate of production?

43. Suppose in Problem 42 that additional constraints arise from the fact that certain tasks cannot be performed at the same station. In particular suppose that the tasks are zoned in the following manner:

Zone 1	Tasks 2, 3, 1, 4, 6
Zone 2	Tasks 5, 8, 7, 9
Zone 3	Tasks 10, 11

Assuming that only tasks in the same zone category can be performed at the same station, determine the resulting line balance for Problem 42 based on a 45-second cycle time.

44. The Southeastern Sports Company produces golf clubs on an assembly line in its plant in Marietta, Georgia. The final assembly of the woods requires the eight operations given in the following table.

Task	Time Required (min.)	Immediate Predecessors
1. Polish shaft	12	
2. Grind the shaft end	14	
3. Polish club head	6	
4. Imprint number	4	3
5. Connect wood to shaft	6	1, 2, 4
6. Place and secure connecting pin	3	5
7. Place glue on other end of shaft	3	1
8. Set in grips and balance	12	6, 7

a. Draw a network to represent the assembly operation.

b. What is the minimum cycle time that can be considered? Determine the balance that results from the ranked positional weight technique for this cycle time.

c. By experimentation, determine the minimum cycle time that can be achieved with a three-station balance.

45. Develop a template that computes several measures of performance for first-come, first-served job sequencing. Allow for up to 20 jobs so that column 1 holds the numbers 1, 2, . . . , 20. Column 2 should be the processing times to be inputted by the user, and column 3 the due dates also inputted by the user. Column 4 should be the tardiness and column 5 the flow time. Develop the logic to compute the mean flow time, the average tardiness, and the number of tardy jobs. (When computing the average of a column, be sure that your spreadsheet does not treat blanks as zeros.)

a. Use your template to find the mean flow time, average tardiness, and number of tardy jobs for Problem 7 (Section 8.6), assuming FCFS sequencing.

b. Find the mean flow time, average tardiness, and number of tardy jobs for Problem 8, assuming FCFS sequencing.

c. Find the mean flow time, average tardiness, and number of tardy jobs assuming FCFS sequencing for the following 20-job problem:

Job	Processing Time	Due Date	Job	Processing Time	Due Date
1	10	34	11	17	140
2	24	38	12	8	120
3	16	60	13	23	110
4	8	52	14	25	160
5	14	25	15	40	180
6	19	95	16	19	140
7	26	92	17	6	130
8	24	61	18	23	190
9	4	42	19	25	220
10	12	170	20	14	110

46. *a*. Solve Problem 45(*a*) assuming SPT sequencing. To do this you may use the spreadsheet developed in Problem 45 and simply sort the data in the first three columns, using column 2 (the processing time) as a sort key.

b. Solve Problem 45(*b*) assuming SPT sequencing.

c. Solve Problem 45(*c*) assuming SPT sequencing.

47. *a*. Solve Problem 45(*a*) assuming EDD sequencing. In this case one sorts on the due date column.

b. Solve Problem 45(*b*) assuming EDD sequencing.

c. Solve Problem 45(*c*) assuming EDD sequencing.

Bibliography

Akers, S. B. "A Graphical Approach to Production Scheduling Problems." *Operations Research* 4 (1956), pp. 244–45.

Arcus, A. L. "COMSOAL: A Computer Method for Sequencing Operations for Assembly Lines." *International Journal of Production Research* 4 (1966), pp. 259–77.

Baker, K. R. *Introduction to Sequencing and Scheduling.* New York: John Wiley & Sons, 1974.

Banerjee, B. P. "Single Facility Sequencing with Random Execution Times." *Operations Research* 13 (1965), pp. 358–64.

Bellman, R. E. "Mathematical Aspects of Scheduling Theory." *SIAM Journal of Applied Mathematics* 4 (1956), pp. 168–205.

Bruno, J., and P. Downey. "Sequencing Tasks with Exponential Service Times on Two Machines." Technical Report, Department of Electrical Engineering and Computer Science, University of California at Santa Barbara, 1977.

Carlson, R., and M. Rosenblatt. "Designing a Production Line to Maximize Profit." *IIE Transactions* 17 (1985), pp. 117–22.

Conway, R. W.; W. L. Maxwell; and L. W. Miller. *Theory of Scheduling.* Reading, MA: Addison Wesley, 1967.

Derman, C.; G. L. Lieberman; and S. M. Ross. "A Renewal Decision Problem." *Management Science* 24 (1978), pp. 554–61.

Fisher, M.; A. J. Greenfield; R. Jaikumar; and J. T. Uster III. "A Computerized Vehicle Routing Application." *Interfaces* 12 (1982), pp. 42–52.

Fishman, G. S. *Concepts and Methods in Discrete Event Digital Simulation.* New York: John Wiley & Sons, 1973.

French, S. *Sequencing and Scheduling: An Introduction to the Mathematics of the Job Shop.* Chichester, England: Ellis Horwood Limited, 1982.

Graves, S. C. "A Review of Production Scheduling." *Operations Research* 29 (1981), pp. 646–75.

Held, M.; R. M. Karp; and R. Shareshian. "Assembly Line Balancing—Dynamic Programming with Precedence Constraints." *Operations Research* 11 (1963), pp. 442–59.

Helgeson, W. P., and D. P. Birnie. "Assembly Line Balancing Using the Ranked Positional Weight Technique." *Journal of Industrial Engineering* 12 (1961), pp. 394–98.

Hillier, F. S., and R. W. Boling. "On the Optimal Allocation of Work in Symmetrically Unbalanced Production Line Systems with Variable Operation Times." *Management Science* 25 (1986), pp. 721–28.

Hutchison, J.; G. Leong; and P. T. Ward. "Improving Delivery Performance in Gear Manufacturing at Jeffrey Division of Dresser Industries." *Interfaces* 23, no. 2 (March–April 1993), pp. 69–83.

Johnson, S. M. "Optimal Two and Three Stage Production Schedules with Setup Times Included." *Naval Research Logistics Quarterly* 1 (1954), pp. 61–68.

Kilbridge, M. D., and L. Wester. "A Heuristic Method of Line Balancing." *Journal of Industrial Engineering* 12 (1961), pp. 292–98.

Lawler, E. L. "Optimal Sequencing of a Single Machine Subject to Precedence Constraints." *Management Science* 19 (1973), pp. 544–46.

Moore, J. M. "An *n*-job, One Machine Sequencing Algorithm for Minimizing the Number of Late Jobs." *Management Science* 15 (1968), pp. 102–109.

Nichols, J. C. "Planning for Real World Production." *Production* 106, no. 8 (August 1994), pp. 18–20.

Parker, K. "What New Tools Will Best Tame Time." *Manufacturing Systems* 12, no. 1 (January 1994), pp. 16–22.

Parker, K. "Dynamism and Decision Support." *Manufacturing Systems* 13, no. 4 (April 1995), pp. 12–24.

Pinedo, M. "Stochastic Scheduling with Release Dates and Due Dates." *Operations Research* 31 (1983), pp. 554–72.

Pinedo, M. *Scheduling, Theory, Algorithms and Systems*. Englewood Cliffs, NJ: Prentice Hall, 1995.

Pinedo, M., and G. Weiss. "Scheduling Stochastic Tasks on Two Parallel Processors." *Naval Research Logistics Quarterly* 26 (1979), pp. 527–36.

Randhawa, S. U., and R. Shroff. "Simulation-Based Design Evaluation of Unit Load Automated Storage/Retrieval Systems." *Computers and Industrial Engineering* 28, no. 1 (January 1995), pp. 71–79.

Righter, R. "Job Scheduling to Minimize Weighted Flowtime on Uniform Processors." *Systems and Control Letters* 10 (1988), pp. 211–16.

Righter, R., and J. G. Shanthikumar. "Scheduling Multiclass Single-Server Queuing Systems to Stochastically Maximize the Number of Successful Departures." *Probability in the Engineering and Informational Sciences* 3 (1989), pp. 323–33.

Rothkopf, M. S. "Scheduling with Random Service Times." *Management Science* 12 (1966), pp. 707–13.

Salveson, M. E. "Production Planning and Scheduling." *Econometrica* 20 (1952), pp. 554–90.

Swain, J. J. "Flexible Tools for Modeling." *OR/MS Today* 20, no. 6 (1993), pp. 62–78.

Thangavelu, S. R., and C. M. Shetty. "Assembly Line Balancing by Zero One Integer Programming." *AIIE Transactions* 3 (1971), pp. 61–68.

Vasilash, G. "Scheduling for the Shop Floor." *Production* 107, no. 6 (June 1995), pp. 46–47.

Vollman, T. E.; W. L. Berry; and D. C. Whybark. *Manufacturing Planning and Control Systems.* 3rd ed. Homewood, IL: Richard D. Irwin, 1992.

Chapter **Ten**

Project Scheduling

"Man does not plan to fail, he just fails to plan." -Anonymous

Chapter Overview

Purpose

To understand how mathematical and graphical techniques are used to assist with the task of scheduling complex projects in an organization.

Key Points

1. *Project representation and critical path identification.* There are two convenient graphical techniques for representing a project. One is a Gantt chart. The Gantt chart was used in Chapter 9 to represent sequence schedules on multiple machines. However, representing a project as a Gantt chart has one significant drawback. Precedence relationships (that is, specifying which activities must precede other activities) are not displayed. To overcome this inadequacy, we represent a project as a network rather than a Gantt chart. A network is a set of nodes and directed arcs. Nodes correspond to milestones in the project (completion of some subset of activities), and arcs to specific activities.

   In the network representation, the goal is to identify the critical, or longest, path. In the spirit of "a chain is only as strong as its weakest link," a project cannot be completed until all the activities along the critical path are completed. The length of the critical path gives the earliest completion time of the project. Activities not along the critical path (noncritical activities) have slack time—that is, they can be delayed without necessarily delaying the project. In Section 10.2, we present an algorithm for identifying the critical path in a network. (This is only one of several solution methods.)
2. *Time costing methods.* Consider a construction project. Each additional day that elapses results in higher costs. These costs include direct labor costs for the personnel involved in the project, costs associated with equipment and material usage, and overhead costs. Let us suppose one has the option of decreasing the time of selected activities, but also at some cost. As the times required for activities along the critical path are decreased, the expediting costs increase but the costs proportional to the project time decrease. Hence, there is some optimal time for the project that balances these two competing costs. The problem of cost-optimizing the time of a project can be solved manually or via linear programming.
3. *Project scheduling with uncertain activity times.* In some projects, such as construction projects, the time required to do specific tasks can be predicted

accurately in advance. In most cases, past experience can be used as an accurate guide, even for novel projects. However, this is not the case with research projects. When undertaking the solution of an unsolved problem, or designing an entirely new piece of equipment, it is difficult, if not impossible, to predict activity times accurately in advance. A more reasonable assumption is that activity times are random variables with some specified distribution.

A method that explicitly allows for uncertain activity times is the project evaluation and review technique (PERT). This technique was developed by the Navy to assist with planning the Polaris submarine project in 1958. The PERT approach is to assume that planners specify for each activity a minimum time, a, a maximum time, b, and a most likely time, m, for each activity. These estimates are then used to construct a beta distribution for each activity time. The PERT assumption is that the critical path will be the path with the longest expected completion time (which is not necessarily the case), and the total project time will be the sum of the times along the critical path. Assuming activity times are independent random variables, one computes the mean and variance along the critical path by summing the means and variances of the activity times. The central limit theorem is then used to justify the assumption that the project completion time has a normal distribution with mean and variance computed as previously described. Note that this is only an approximation, since there is no guarantee that the path with the longest expected completion time will turn out to be the critical path. Determining the true distribution of project completion time appears to be a very difficult problem in general. However, PERT provides a reasonable approximation and is certainly an improvement over the deterministic critical path method (CPM).

4. *Resource considerations.* Consider a department within a firm in which several projects are simultaneously ongoing. Suppose that each member of the department is working on more than one project at a time. Since the time of each worker is limited, each project manager is competing for a limited resource, namely, the time of the workers. One could imagine other cases where the limited resource might be a piece of equipment, such as a single supercomputer in a company. In these cases, incorporating resource constraints into the project planning function can be quite a challenge. We present an example of balancing resources, but know of no general-purpose method for solving this problem.

Rome wasn't built in a day, and neither were the pyramids of Egypt, the Empire State Building, the Golden Gate Bridge, or the Eiffel Tower. These were all complex projects that required careful planning and coordinating. How does one organize and monitor such massive projects? Effective project management could make or break a project. While many different skills are required to be an effective project manager, quantitative techniques can be an enormous help. These techniques are the subject of this chapter.

What are the consequences of poor project management? One is cost overruns. How often have we heard members of Congress express dismay at the cost overruns in military projects? In some cases these overruns could not be avoided: Unforeseen obstacles arose or technological problems could not be solved as easily as anticipated. In many cases, however, these problems were a consequence of poor project scheduling and management.

Large complex projects involving governments in partnership with business are perhaps the most vulnerable to delays and overruns. A case in point is the Trans-Alaska Pipeline System, designed to transport large quantities of oil from Prudhoe Bay on Alaska's north slope to Port Valdez on the Gulf of Alaska. Goodman (1988) describes it as one of the most complex and massive design and construction projects of recent times. Political roadblocks, environmental concerns, and contract disputes plagued the project from the start. The list of important players changed several times, resulting in some firms (Bechtel, in particular) not having enough time to do an adequate job of project planning. The Alyeska Pipeline Service Company, which was responsible for much of the actual building of the pipeline, incurred excessive cost overruns, partly due to poor project management. In retrospect, it is clear that many of the problems with the project were a consequence of the fact that there was never a single project team to oversee the entire integrated project cycle.

Project scheduling and project management methods have been an important part of doing business for many firms. For example, the Lockheed-Martin Missiles and Space Company of Mountain View, California, uses project scheduling methods not only for monitoring and control of projects, but also for the process of preparing bids and developing proposals. In fact, Lockheed was part of the team that developed PERT, a technique considered in detail in this chapter.

This chapter reviews analytical techniques for project management. Effective people management can be just as important a factor in getting projects done on time and within budget. The firm must create a structure and environment conducive to properly motivating employees. Poor organizational design and incentive structures can be just as serious a problem as poor project planning.

The project management methods considered in this chapter have been used to plan long-term projects such as launching new products, organizing research projects, and building new production facilities. The methods also have been used for smaller projects such as building of residential homes. Two techniques that are treated in detail are the *critical path method* (CPM) and the *project evaluation and review technique* (PERT). Both methods were developed almost simultaneously for solving very different project management problems in the late 1950s. Although the two labels are used interchangeably today, we will retain the terminology consistent with the original intent of the methods. That is, CPM deals with purely deterministic problems, whereas PERT allows randomness in the activity times.

The primary elements of critical path analysis are:

1. *Project definition*. This is a clear statement of the project, the goals of the project, and the resources and personnel that the project requires.

2. *Activity definitions*. The project must be broken down into a set of indivisible tasks or activities. The project planner must specify the work content of each activity and estimate the activity times. Often the most difficult part of project planning is finding the best way to break down the project into a collection of distinct activities.

3. *Activity relationships*. An important part of the project planning methodology is a specification of the interrelationships among the activities. Known as *precedence constraints,* these describe the logical sequence to complete the activities comprising the project.

4. *Project scheduling*. A project schedule is a specification of the starting and ending times of *all* activities comprising the project. Using the techniques of this chapter, we will show how specification of the activity times and precedence constraints yields an efficient schedule.

5. *Project monitoring*. Once the activities have been suitably defined and a schedule determined, the proper controls must be put in place to ensure that project milestones are met. The project manager must be prepared to revise existing schedules if unforeseen problems arise.

Because of the level of detail and precision required, critical path analysis is most effective when the project can be easily expressed as a well-defined group of activities. Construction projects fall into this category. Precedence constraints are straightforward, and activity times are not difficult to estimate. For this reason, CPM has found wide acceptance in the construction industry. However, there are case studies reported in the literature of successful applications in a variety of environments, including military, government, and nonprofit.

10.1 REPRESENTING A PROJECT AS A NETWORK

As we saw in Chapter 9 on shop floor scheduling and control, a *Gantt chart* is a convenient graphical means of picturing a schedule. A Gantt chart is a horizontal bar graph on which each activity corresponds to a separate bar. Consider the following simple example.

Example *10.1* Suppose that a project consists of five activities, A, B, C, D, and E. Figure 10–1 shows a schedule for completing this project, in the form of a Gantt chart. According to the figure, task A starts at 12:00 and finishes at 1:30. Tasks B and C start at 1:30, B finishes at 2:30, C finishes at 3:30, and so on. The project is finally completed at 6:00.

Notice that Figure 10–1 gives no information about the relationships among the activities. For example, the figure implies that E cannot start until D is completed. However, suppose that E is permitted to start any time after A, B, and C finish. Then E can start at 3:30, and the project can be completed at 5:00 rather than at 6:00.

Although the Gantt chart is a useful means of representing the schedule, it is not a very useful planning tool because it does not show the precedence constraints. For this reason, one graphically represents the project as a network. Networks, unlike Gantt charts, explicitly show the precedence constraints.

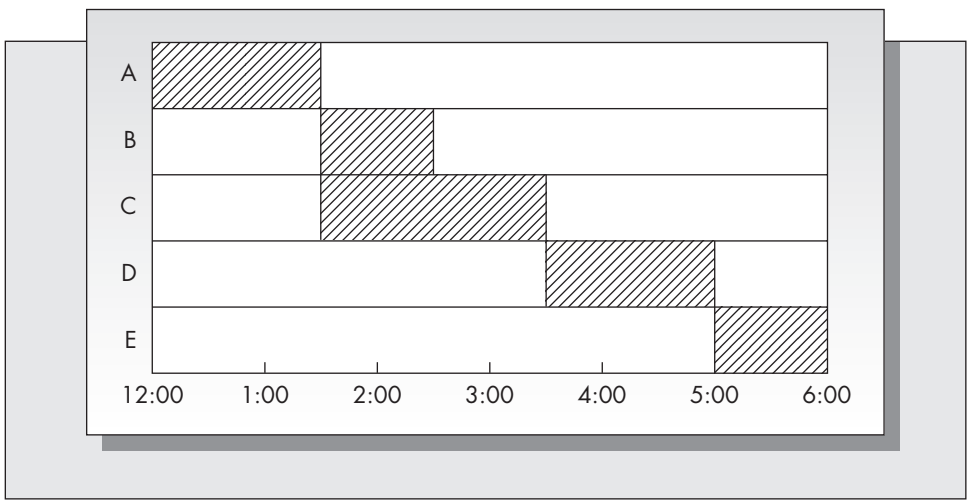


FIGURE 10–1
Gantt chart for five activities (refer to Example 10.1)

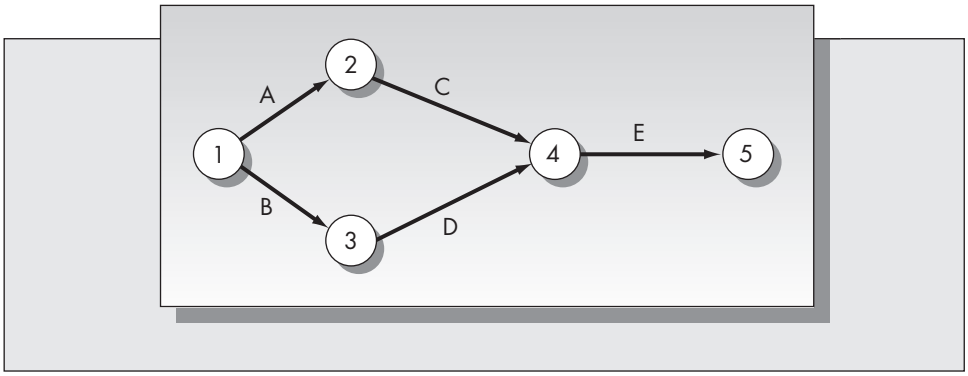


FIGURE 10–2
Project network for Example 10.2

A network is a collection of nodes and directed arcs. In the traditional network representation of a project, a node corresponds to an event and an arc to an activity or task. Events may be (1) the start of a project, (2) the completion of a project, or (3) the completion of some collection of activities. This method of representation is known as activity-on-arrow and is historically the most common means of representing a project as a network. An alternative method of representation is activity-on-node. Although the latter method has some advantages, it is rarely used in practice. The activity-on-node method will be illustrated in Section 10.2.

Example *10.2* Suppose a project consists of the five activities A, B, C, D, and E that satisfy the following precedence relationships:

1. Neither A nor B has any immediate predecessors.
2. A is an immediate predecessor of C.
3. B is an immediate predecessor of D.
4. C and D are immediate predecessors of E.

The network for this project appears in Figure 10–2. Node 1 is always the initiation of the project. All activities having no immediate predecessors emanate from node 1. As A is an immediate predecessor of C, we must have a node representing the completion of A. Similarly, as B is an immediate predecessor of D, there must be a node representing the completion of B. Notice that although both A and B must also be completed before E, they are not *immediate* predecessors of E.

Representing a project by a network is not always this straightforward.

Example *10.3* Consider Example 10.2, except now replace (3) with (3′):

3′. A and B are immediate predecessors of D.

Try to find a network representation of this project. One might think that the correct representation is the one in Figure 10–3, as this network implies that D must wait for both A and B. However, this representation is incorrect, because it also shows that C must wait for both A and B as well.

The set of precedent relations (1), (2), (3′), and (4) require that we introduce a pseudo activity between nodes 2 and 3. The correct representation of the system is given in Figure 10–4. The pseudo activity, labeled P, is a directed arc from node 2 to node 3 and is represented by a broken line. Note that the direction of the arrow is important. With the pseudo activity, node 3 corresponds to the completion of both activities A and B, whereas node 2 still corresponds to the completion of only activity A.

FIGURE 10–3
Incorrect network representation for Example 10.3

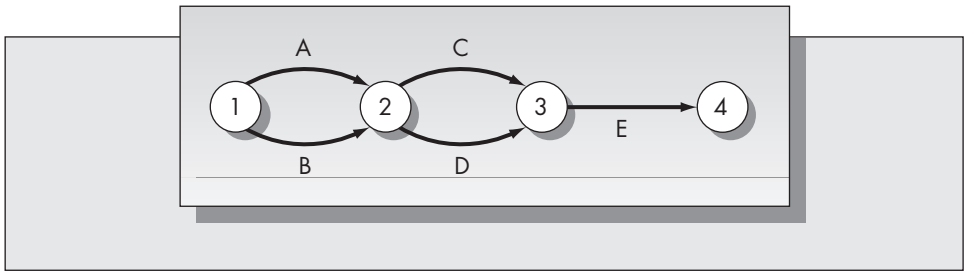
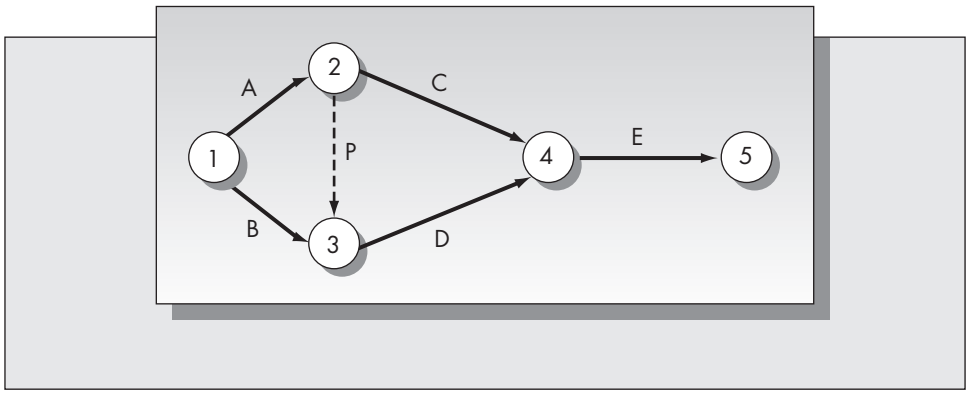


FIGURE 10–4
Correct network representation for Example 10.3



An important aspect of critical path analysis is defining an appropriate set of tasks. A problem that may arise is that one defines the tasks too broadly. As an example, suppose in Example 10.3 that E could start when only half of C was completed. This means that C would have to be further broken down into two other activities, one representing the first half of C and one representing the second half of C. Conversely, it is possible to define activities too narrowly. This can result in an overly complicated network representation, with portions of the network resembling Figure 10–3.

10.2 CRITICAL PATH ANALYSIS

Once having represented the project as a network, we can begin to explore the methods available for answering a variety of questions. For example,

1. What is the minimum time required to complete the project?
2. What are the starting and ending times for each of the activities?
3. Which activities can be delayed without delaying the project?

A path through the network is a sequence of activities from node 1 to the final node. One labels a path by the sequence of nodes visited or by the sequence of arcs (activities) traversed. In the network pictured in Figure 10–4 there are exactly three distinct paths: 1–2–4–5 (A–C–E), 1–2–3–4–5 (A–P–D–E), and 1–3–4–5 (B–D–E). Each path is a sequence of activities satisfying the precedence constraints. Because the project is completed when all activities are completed, it follows that *all* paths from the initial node to the final node must be traversed. Hence, the minimum time to complete the project must be the same as the length of the *longest* path in the network.

Consider Example 10.3 and suppose that activity times are the same as those indicated in Figure 10–1. The activity times and the precedence constraints are

Activity	Time (hours)	Immediate Predecessors
A	1.5	
B	1.0	
C	2.0	A
D	1.5	A and B
E	1.0	C and D

The lengths of the three paths are

Path	Time Required to Complete (hours)
A–C–E	4.5
A–P–D–E	4.0
B–D–E	3.5

To complete the project, all three paths must be completed. The longest path for this example is obviously A–C–E. This is known as the *critical path*. The length of the critical path is the *minimum* completion time of the project. The activities that lie along the critical path are known as the critical activities. A delay in a critical activity results in a delay in the project. However, activities that do not lie along the critical path have some slack. A delay in a noncritical activity does not necessarily delay the project.

Enumerating all paths is, in general, not an efficient way to find the critical path. Later, we consider a general procedure, not requiring path enumeration, for identifying the critical path and the start and finish times for all activities comprising the project. First we consider the following case study.

Example *10.4*

Simon North and Irving Bonner, computer consultants, are considering embarking on a joint project that will involve development of a relatively small commercial software package for personal computers. The program involves scientific calculations for a specialized portion of the engineering market. North and Bonner have broken down the project into nine tasks.

The first task is to undertake a market survey in order to determine exactly what the potential clientele will require and what features of the software are likely to be the most attractive. Once this stage is completed, the actual development of the programs can begin. The programming requirements fall into two broad categories: graphics and source code. Because the system will be interactive and icon driven, the first task is to identify and design the icons. After the programmers have completed the icon designs, they can proceed with the second part of the graphics development, design of the input/output screens. These include the various menus and report generators required in the system.

The second part of the project is coding the modules that do the scientific calculations. The first step is to develop a detailed flowchart of the system. After they complete the flowchart, the programmers can begin work on the modules. There are a total of four modules. The work on modules 1 and 2 can begin immediately after completion of the flowchart. Module 3 requires parts of module 1, so the work on module 3 cannot begin until module 1 is finished. The programming of module 4 cannot start until both modules 1 and 2 are completed. Once

the graphics portion of the project and the modules are completed, the two separate phases of the system must be merged and the entire system tested and debugged.

North has managed to obtain some funding for the project, but his source requires that the program be completed and ready to market in 25 weeks. In order to determine whether this is feasible, the two programmers have divided the project into nine indivisible tasks and have estimated the time required for each task. The list of these tasks, the times required, and the precedence relationships are given below.

Task	Time Required (weeks)	Immediate Predecessors
A. Perform market survey	3	
B. Design graphic icons	4	A
C. Develop flowchart	2	A
D. Design input/output screens	6	B, C
E. Module 1 coding	5	C
F. Module 2 coding	3	C
G. Module 3 coding	7	E
H. Module 4 coding	5	E, F
I. Merge modules and graphics and test program	8	D, G, H

The total of the task times is 43 weeks. Based on this, the programmers conclude that it would be impossible to complete their project in the required 25 weeks. Fortunately, they see the error in their thinking before breaking off relations with their funding source. Since some of the activities can be done concurrently, the project should take fewer than 43 weeks. The network representation for this project is given in Figure 10–5. In order to satisfy the precedence constraints we require two pseudo activities, P_1 and P_2. We will determine the critical path *without* enumerating all paths through the network. The critical path calculations yield both the critical path *and* the allowable slack for each activity as well.

For illustrative purposes, we represent the project network using the activity-on-node format in Figure 10–6. One advantage of this format is that pseudo activities are not required, although for certain networks dummy starting and ending nodes may be needed. Given a list of activities and immediate predecessors, it should be easy to find the network representation in either format.

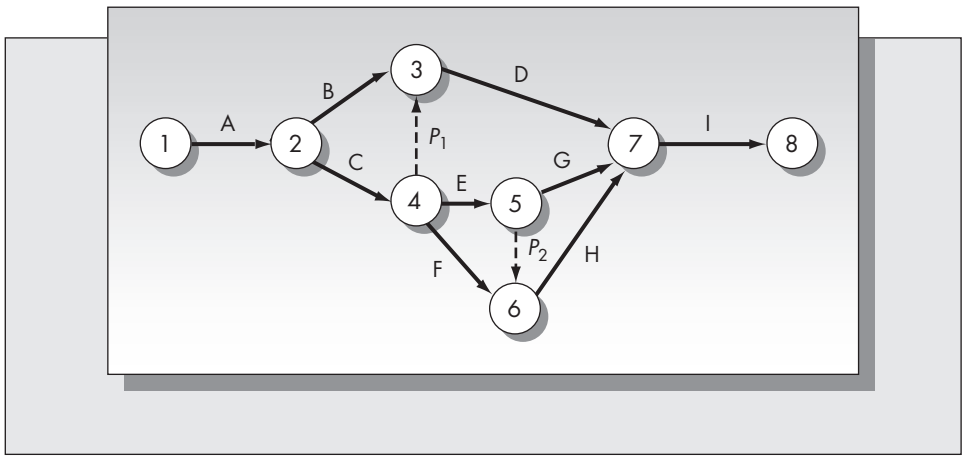


FIGURE 10–5
Network for Example 10.4

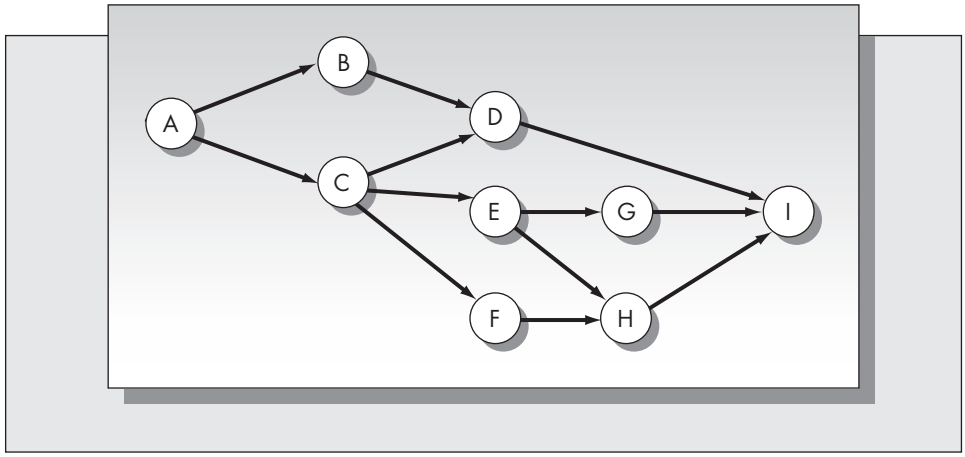


FIGURE 10–6
Network for Example 10.4: Activity-on-node representation

Practitioners prefer the activity-on-arrow method, believing that it is a more intuitive way to represent the project. We will use it exclusively for the remainder of the chapter.

Finding the Critical Path

We compute four quantities for each activity:

ES_i = Earliest starting time for activity i,

EF_i = Earliest finishing time for activity i,

LF_i = Latest finishing time for activity i (without delaying the project),

LS_i = Latest starting time for activity i (without delaying the project).

Suppose that t_i represents the time required to complete activity i. Then it is easy to see that

$$EF_i = ES_i + t_i$$

and

$$LS_i = LF_i - t_i.$$

The steps in the process are

1. *Compute the earliest times for each activity.* One computes the earliest times by a forward pass through the network; that is, the computations proceed from node 1 to the final node.
2. *Compute the latest times for each activity*. One computes the latest times by a backward pass through the network; that is, the computations proceed from the final node to node 1.

We will illustrate the computations using Example 10.4. The first step is to set the earliest starting times for all activities emanating from node 1 to zero and add the activity times to obtain the earliest finishing times. In this case, the only activity emanating from node 1 is activity A. Because B has A as an immediate predecessor, it follows that $ES_B = EF_A$. Similarly for C, $ES_C = EF_A$. Compute the earliest finishing times for

B and C by adding the activity times to the earliest starting times: $EF_B = ES_B + t_B$, and $EF_C = ES_C + t_C$. Summarizing the calculations up to this point, we have

Activity	Time	Immediate Predecessor	ES	EF
A	3	—	0	3
B	4	A	3	7
C	2	A	3	5

Calculating the earliest starting time for D is more difficult. Activity D has two immediate predecessors, B and C. That means that D cannot start until *both* B and C have been completed. It follows that the earliest starting time for D is the *later* of the earliest finishing times for B and C. That is,

$$ES_D = \max(EF_B, EF_C) = \max(7, 5) = 7.$$

General Rule: The earliest starting time of an activity is the *maximum* of the earliest finishing times of its immediate predecessors.

Continuing in this manner, we obtain the following earliest times for all remaining activities:

Activity	Time	Immediate Predecessor	ES	EF
A	3	—	0	3
B	4	A	3	7
C	2	A	3	5
D	6	B, C	7	13
E	5	C	5	10
F	3	C	5	8
G	7	E	10	17
H	5	E, F	10	15
I	8	D, G, H	17	25

At this point we have actually determined the length of the critical path. It is the maximum of the earliest finish times, or 25 weeks in this case. However, we must find the latest times before we can identify the critical activities.

The computation of the latest times proceeds by working backward through the network and applying essentially the dual of the procedure for the earliest times. The first step is to set the latest finishing time of all the activities that enter the final node to the maximum value of the earliest finishing times. In this case there is only a single ending activity, so we set

$$LF_I = 25.$$

The latest starting time is obtained by subtracting the activity time, so

$$LS_I = 25 - 8 = 17.$$

Next we must determine the latest finishing time for all the activities that enter node 7, which are D, G, and H. Because these activities end when I begins, we have

$$LF_D = LF_G = LF_H = LS_I = 17.$$

One finds the latest starting times for D, G, and H by simply subtracting the activity times. Summarizing the calculations up to this point, we have

Activity	Time	Immediate Predecessor	ES	EF	LS	LF
A	3	—	0	3		
B	4	A	3	7		
C	2	A	3	5		
D	6	B, C	7	13	11	17
E	5	C	5	10		
F	3	C	5	8		
G	7	E	10	17	10	17
H	5	E, F	10	15	12	17
I	8	D, G, H	17	25	17	25

Because F ends when H begins, $LF_F = LS_H = 12$, and $LS_F = 12 - 3 = 9$. Now consider activity E. From Figure 10–5, E has both G and H as immediate successors. This means that E must end prior to the time that both G and H start. Hence the latest finishing time for E is the *earlier* of the latest start times for G and H. That is, $LF_E = \min(LS_G, LS_H) = \min(10, 12) = 10$.

General Rule: The latest finishing time of an activity is the *minimum* of the latest start times of its immediate successors.

According to the network diagram of Figure 10–5, C has both E and F as immediate successors. Hence $LF_C = \min(LS_E, LS_F) = \min(5, 9) = 5$, and $LS_C = 5 - 2 = 3$. Because B has only D as an immediate successor, $LF_B = LS_D = 11$, and $LS_B = 11 - 4 = 7$. Finally, A has both B and C as immediate successors, so $LF_A = \min(LS_B, LS_C) = \min(7, 3) = 3$, and $LS_A = 3 - 3 = 0$.

The complete summary of the calculations for this example is

Activity	Time	Immediate Predecessors	ES	EF	LS	LF	Slack
A	3	—	0	3	0	3	0
B	4	A	3	7	7	11	4
C	2	A	3	5	3	5	0
D	6	B, C	7	13	11	17	4
E	5	C	5	10	5	10	0
F	3	C	5	8	9	12	4
G	7	E	10	17	10	17	0
H	5	E, F	10	15	12	17	2
I	8	D, G, H	17	25	17	25	0

We have added a column labeled "slack." This is the difference between the columns LS and ES (it is also the difference between the columns LF and EF). The slack is the amount of time that an activity can be delayed without delaying the project. The activities with zero slack—A, C, E, G, and I—are the critical activities and constitute the critical path.

Figure 10–7 is a Gantt chart that shows the starting and the ending times for each activity. The noncritical activities are shown starting at the earliest times, although they can be scheduled at any time within the period marked by the slack.

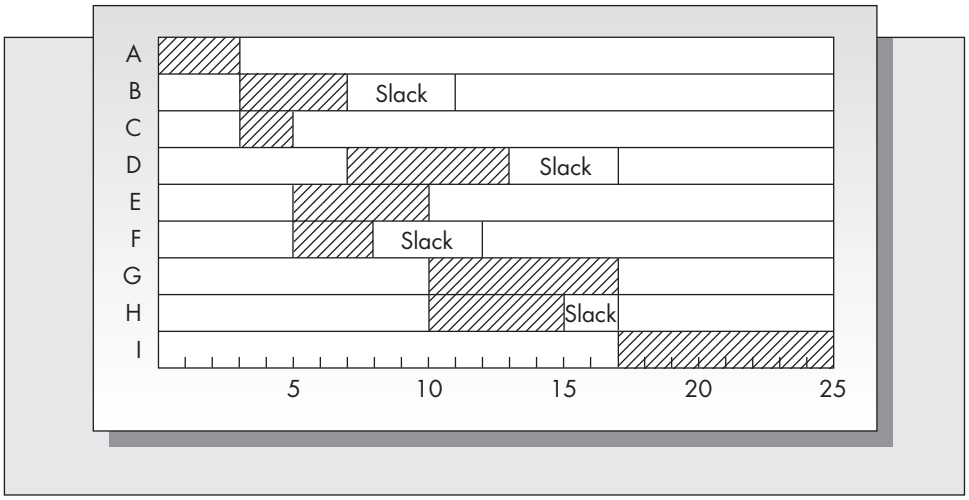


FIGURE 10–7
Gantt chart for software development project of Example 10.4

Example *10.4 (continued)*

Suppose that the two programmers begin the project on June 1. Consider the following questions:

a. On what date should the merging of the graphics and the program modules begin in order to guarantee that the programmers complete the project on time?

b. What is the latest time that the screen development can be completed without delaying the project?

c. Suppose that North discovers a bug in a coding of module 2, so that the time required to complete module 2 is longer than anticipated. Will this necessarily delay the project?

d. Suppose that a similar problem is discovered in module 1. Will the project necessarily be delayed in this case?

e. If Bonner is responsible for the coding of the modules and North is responsible for the screen development, and all time estimates are accurate, will the programmers be able to complete the project within 25 weeks?

Solution

a. The merging of the graphics and the program modules is activity I. This is a critical activity, so it must be started no later than its earliest start time, which is week 17. The calendar date would be September 14.

b. The screen development is activity D. This is not a critical activity and has a latest finishing time of 17. The calendar date is again September 14.

c. Module 2 is activity F. Because F is not critical, a delay of up to 4 weeks is permitted.

d. Module 1 is activity E. This is a critical activity, so a delay will necessarily delay the project.

e. The answer is no. The problem is that, assuming that Bonner can work only on a single module at one time, the current schedule requires concurrent programming of modules 1 and 2 and modules 3 and 4. The programmers would have to take in another partner or subcontract some of the module development in order to complete the project within 25 weeks.

Problems for Sections 10.1 and 10.2

1. Answer the following questions for Example 10.4.

   a. What event is represented by node 6 in Figure 10–5?

   b. What group of activities will have to be completed by week 16 in order to guarantee that the project will not be delayed?

c. Suppose that module 4 is coded by a subcontractor who delivers the module after 7 weeks. How much will it delay the project if the subcontractor does not start until week 11?

2. For Example 10.4, suppose that the programmers choose not to obtain additional help to complete the project; that is, Bonner will code the modules without outside assistance.
   a. In what way will this alter the project network and the precedence constraints?
   b. Find the critical path of the project network you obtained in part (*a*). How much longer is it than the one determined in Example 10.4?
3. A project consisting of eight activities satisfies the following precedence constraints:

Activity	Time (weeks)	Immediate Predecessors
A	3	
B	5	A
C	1	A
D	4	B, C
E	3	B
F	3	E, D
G	2	F
H	4	E, D

   a. Construct a network for this project. (You should need only one pseudo activity.)
   b. Compute the earliest and the latest starting and finishing times for each activity and identify the critical path.
   c. Draw a Gantt chart of the schedule for this project based on earliest starting times.
4. For the project network pictured in Figure 10–8,
   a. List the immediate predecessors of each activity.
   b. Try to determine the critical path by enumerating all paths from node 1 to node 10.
   c. Compute the earliest starting and finishing times for all activities and identify the critical activities.

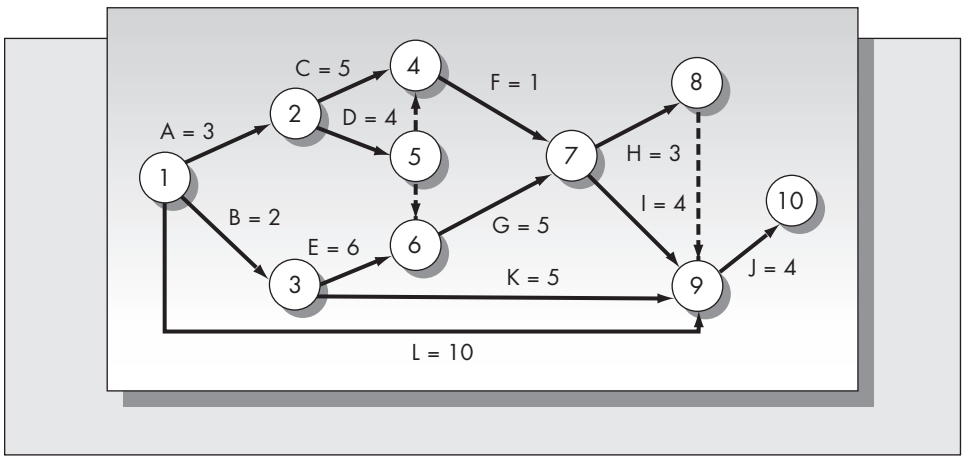
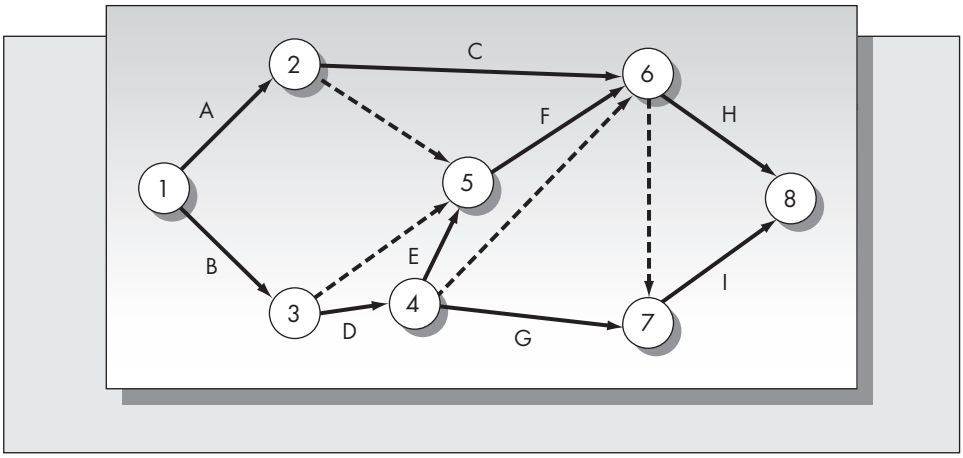


FIGURE 10–8
Project network for Problem 4

FIGURE 10–9
Network for Problem 5



5. Consider the network pictured in Figure 10–9.
   a. Determine the immediate predecessors of each activity from the network representation. (Hint: Be certain that you consider only *immediate* predecessors.)
   b. Redraw the network based on the results of part (*a*) with only two pseudo activities.
6. A project consists of seven activities. The precedence relationships are given in the following table.

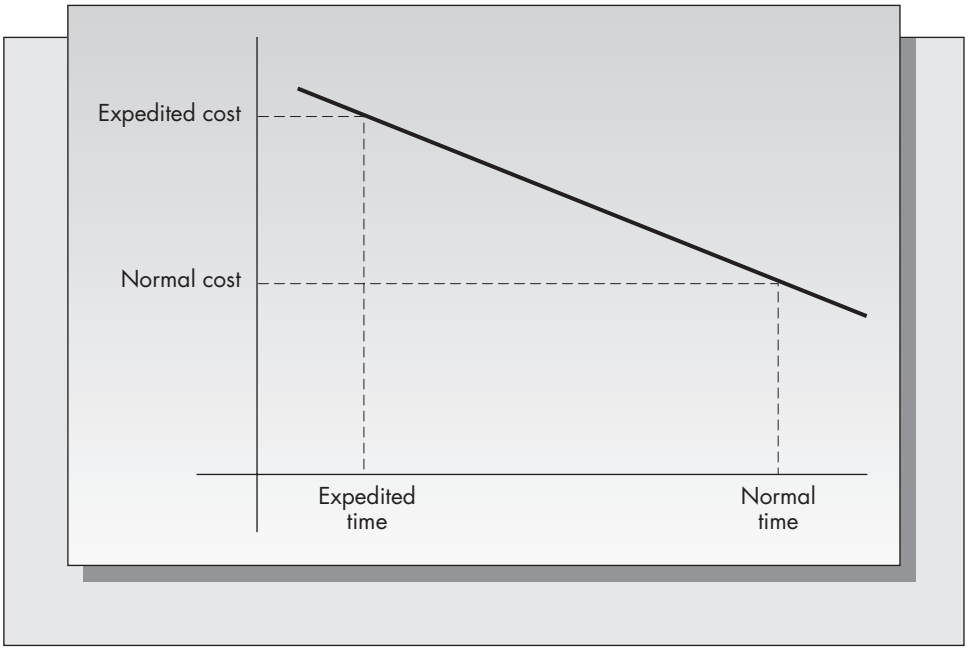
Activity	Time (days)	Immediate Predecessors
A	32	
B	21	
C	30	
D	45	A
E	26	A, B
F	28	C
G	20	E, F

   a. Construct a network for this project.
   b. Compute the earliest and the latest starting and finishing times for each activity and identify the critical path.
   c. Draw a Gantt chart of the schedule for this project based on earliest starting times.
   d. What group of activities will have to be completed by day 60 in order to guarantee that the project will not be delayed?

10.3 TIME COSTING METHODS

Besides assisting with the scheduling of large projects, CPM is a useful tool for project costing and comparing alternative schedules based on cost. This section will consider the costs of expediting activities and how one incorporates expediting costs into the project management framework.

FIGURE 10–10
The CPM cost–time linear model



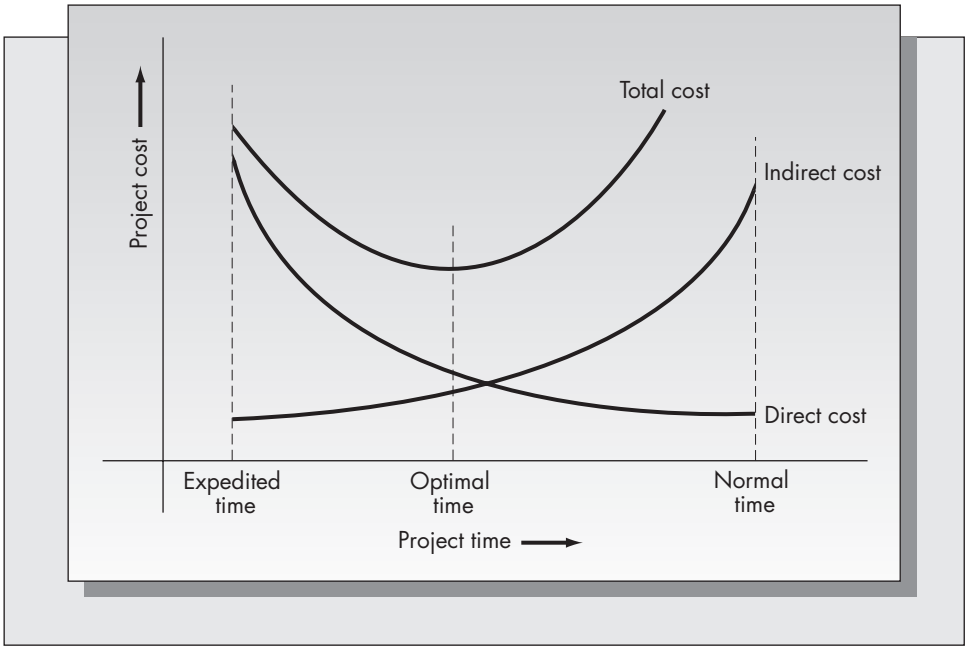
In Section 10.2, we assumed that the time required to complete each activity is known and fixed. In this section we will assume that activity times can be reduced at additional cost. Assume that the time specified for completing an activity is the *normal time*. The minimum possible time required is defined as the *expedited time*. Furthermore, assume that the costs of completing the activity in each of these times are known. Then the CPM assumption is that the costs of completing an activity at times between the normal and the expedited times lie along a straight line, as pictured in Figure 10–10. Assuming that the expediting cost function is linear should be reasonable in most circumstances.

There are two types of costs in most projects: direct and indirect costs. Direct costs include costs of labor, material, equipment, and so on. Indirect costs include costs of overhead such as rents, interest, utilities, and any other costs that increase with the length of the project.

Indirect costs and direct costs are respectively increasing and decreasing functions of the project completion time. When these functions are convex, the total cost function, which is their sum, also will be convex. This means that there will be a value of the project time between the normal and the expedited times that is optimal in the sense of minimizing the total cost. Convex cost functions are pictured in Figure 10–11. Figure 10–10 shows the direct cost as a function of time for a given activity.

The general approach is to successively reduce the project time by one week (or in whatever unit of time activities are measured) until no further reductions are possible or until an optimal solution is identified. At each reduction, one computes the resulting additional direct cost. One continues this process until the minimum total cost solution is identified. A difficulty that arises is that as particular activities are expedited, it is possible that new paths will become critical. The procedure can be illustrated using Example 10.4 introduced in Section 10.2.

FIGURE 10–11
Optimal project completion time



Example *10.5*

Let us return to Example 10.4 concerning the two computer consultants, Irving Bonner and Simon North. In Section 10.2, we saw that it is possible for them (with some additional help) to complete their project in 25 weeks. Once they complete the project and place the program on the market, the consultants anticipate that they will receive an average of $1,500 per week for the first three years that the product is available. By completing the project earlier, they hope to be able to realize this income earlier.

They carefully consider each activity and the possibility of reducing the activity time and the associated costs. They estimate that the normal costs are $500 per week for activities that the consultants themselves do, and either more or less for activities that they contract out. Most of the activities can be expedited by subcontracting parts of the programming. Expediting costs vary based on the nature of the task.

They obtain the following estimates:

Activity	Normal Time (weeks)	Expedited Time (weeks)	Normal Cost	Expedited Cost	Cost per Week
A	3	1	$1,000	$ 3,000	$1,000
B	4	3	4,000	6,000	2,000
C	2	2	2,000	2,000	
D	6	4	3,000	6,000	1,500
E	5	4	2,500	3,800	1,300
F	3	2	1,500	3,000	1,500
G	7	4	4,500	8,100	1,200
H	5	4	3,000	3,600	600
I	8	5	8,000	12,800	1,600

The final column, cost per week, shows the slope of the cost curve pictured in Figure 10–10. One computes the slope from the formula

$$\text{Cost per week} = \frac{\text{Expedited cost} - \text{Normal cost}}{\text{Normal time} - \text{Expedited time}}.$$

One finds the total cost of performing the project in the normal time of 25 weeks by summing the normal costs for each of the activities. This sum is \$29,500. If all activities are expedited, the total cost of the project increases to \$48,300. Is this additional cost worth the \$1,500 per week in additional revenue they expect to realize?

If we replace the normal times with the expedited times, then using the methods of Section 10.2 (or just inspecting the network), it is easy to show that there will be two critical paths based on expedited times: A–C–E–G–I and A–C–E–H–I. The project completion time is reduced to 16 weeks. The additional income that the programmers realize by reducing the project completion time from 25 weeks to 16 weeks is (9)(1,500) = \$13,500, but the additional cost of the reduction is \$48,300 − \$29,500 = \$18,800. Hence, it is not economical to reduce all the activities to their expedited times. (Expediting all activities means that both critical and noncritical activities are expedited. However, there is clearly no economy in expediting noncritical activities. Hence, it is likely that there is a solution with a project completion time of 16 weeks that costs less than \$48,300. The preliminary analysis is used to get a ballpark estimate of the cost of reducing the project to its minimum time.)

It is likely that there is a project time between 16 and 25 weeks that is optimal. To determine the optimal project time, we will find the increase in the direct cost resulting from successive reductions of one week. If the cost increase is less than \$1,500, then the reduction is clearly economical, and additional reductions should be considered. If the cost of the reduction exceeds \$1,500, then further reductions are not economical and the process should be terminated.

The key point to note is that in order to reduce the time required to complete the project by one week, it is necessary to reduce the time of an activity *along the current critical path,* or activities along the current critical paths if more than one path is critical. Reducing the time of a noncritical activity will not reduce the project time.

Initially, we have the following:

Project Time	Critical Path(s)	Critical Activities	Current Time	Expedited Time	Cost to Reduce by One Week
25	A–C–E–G–I	A	3	1	\$1,000
		C	2	2	
		E	5	4	1,300
		G	7	4	1,200
		I	8	5	1,600

The least expensive activity to reduce is A. We can reduce activity A to 1 week without introducing any new critical paths. Because the cost of each weekly reduction is less than \$1,500, it is economical to reduce A to its minimum time, which is one week. At that point we have the following:

Project Time	Critical Path(s)	Critical Activities	Current Time	Expedited Time	Cost to Reduce by One Week
23	A–C–E–G–I	A	1	1	
		C	2	2	
		E	5	4	\$1,300
		G	7	4	1,200
		I	8	5	1,600

The next cheapest activity to reduce is G. The critical path will remain the same until G is reduced to five weeks. (Consider reducing G by one week at a time to be certain that no additional paths become critical.) When G is reduced to five weeks, both paths A–C–E–G–I and

A–C–E–H–I are critical. Reducing G from seven weeks to five weeks results in the following:

Project Time	Critical Path(s)	Critical Activities	Current Time	Expedited Time	Cost to Reduce by One Week
21	A–C–E–G–I				
	A–C–E–H–I	A	1	1	
		C	2	2	
		E	5	4	$1,300
		G	5	4	1,200
		H	5	4	600
		I	8	5	1,600

In order to further reduce the project time to 20 weeks, it is necessary to be certain that we make the reduction along *both* critical paths. At first it would seem that we should reduce H because its marginal cost is least. However, this is not the case. If we reduce H to four weeks, then only the critical path A–C–E–H–I is reduced and not the path A–C–E–G–I. If we reduce both H and G, the increase in the direct cost is $1,800, which is not economical. Does this necessarily mean that it is not worth reducing the project time any further? The answer is no.

Note that the activities A, C, E, and I lie simultaneously along *both* critical paths. Hence, a reduction in the activity time of any one of these four activities will result in a reduction of the project time. Among these activities, E can be reduced from five to four weeks for under $1,500. Making this reduction, we obtain

Project Time	Critical Path(s)	Critical Activities	Current Time	Expedited Time	Cost to Reduce by One Week
20	A–C–E–G–I				
	A–C–E–H–I	A	1	1	
		C	2	2	
		E	4	4	
		G	5	4	$1,200
		H	5	4	600
		I	8	5	1,600

At this point the cost of reducing the project by an additional week exceeds $1,500, so we have reached the optimal solution. The reduction from 25 weeks to 20 weeks costs a total of (1,000)(2) + (1,200)(2) + 1,300 = $5,700 in additional direct costs, and results in a return of (1,500)(5) = $7,500 in additional revenue. If all cost and time estimates are correct, the programmers have realized a savings of $1,800 by taking costing into account.

Problems for Section 10.3

7. Consider Example 10.5 presented in this section in which the two programmers are to develop a commercial software package.
   a. What is the minimum project time and the total direct cost of completing the project in that time?
   b. Suppose that all the activity times are reduced to their minimum values. Explain why the total direct cost obtained in Example 10.5 is different from the cost you obtained in part (*a*).

8. Consider the project described in Problem 3 with the normal and the expedited costs and times given in the following table.

Activity	Immediate Predecessor	Normal Time	Expedited Time	Normal Cost	Expedited Cost
A	—	3	2	$200	$250
B	A	5	3	600	850
C	A	1	1	100	100
D	B, C	4	2	650	900
E	B	3	2	450	500
F	E, D	3	2	500	620
G	F	2	1	500	600
H	E, D	4	2	600	900

   a. Consider successive reductions in the project time of one week and find the direct cost of the project after each reduction.
   b. Suppose that indirect costs are $150 per week. Find the optimal project completion time and the optimal total project cost.

9. Discuss the assumption that the cost–time curve is linear. What shape might be more realistic in practice?

10. Consider Problem 6. Suppose that the normal and the expedited costs and times are as given in the following table.

Activity	Normal Time	Expedited Time	Normal Cost	Expedited Cost
A	32	26	$ 200	$ 500
B	21	20	300	375
C	30	30	200	200
D	45	40	500	800
E	26	20	700	1,360
F	28	24	1,000	1,160
G	20	18	400	550

If indirect costs amount to $100 per day, determine the optimal time to complete the project and the optimal project completion cost.

10.4 SOLVING CRITICAL PATH PROBLEMS WITH LINEAR PROGRAMMING

Finding critical paths in project networks and finding minimum cost schedules can be accomplished by using either dedicated project scheduling software or linear programming. If one expects to solve project scheduling problems on a continuing basis, it is worth the investment in a dedicated software product. However, linear programming is a useful tool for solving a moderately sized problem on an occasional basis. Many linear programming packages, such as Excel's Solver, are widely available and easy to learn and use. This section shows how to formulate and solve critical path problems as linear programs. (In what follows, this chapter assumes that the reader is

familiar with formulating problems as linear programs and interpreting computer output. Supplement 1, which follows Chapter 3, provides a discussion of linear programming and how linear programs can be solved with Excel.)

Based on the choice of objective functions, the linear programming solution will give either earliest or latest start times for each node in the project network. We first formulate the earliest start time problem. Assume that the network representation of the project consists of nodes 1 through m, with node 1 representing the starting time of the project and node m representing the ending time of the project.

Let

$$x_i = \text{Earliest start time for node } i.$$
$$t_{ij} = \text{Time required to complete activity } (i, j).^{1}$$

Then the minimum project completion time is the solution to the following linear program:

$$\min \sum_{i=1}^{m} x_i$$

subject to $x_j - x_i \geq t_{ij}$ for all pairs of nodes corresponding to activity (i, j).

$$x_i \geq 0 \qquad \text{for } 1 \leq i \leq m.$$

The constraints guarantee that there is sufficient time separating each node to account for the activity times represented by the arc between the nodes. As the objective function minimizes the x_i values, the linear programming solution will give the earliest start times. The latest start times can be found by replacing the objective function with

$$\min \left\{ m x_m - \sum_{i=1}^{m-1} x_i \right\}.$$

As x_m is the project completion time, we still wish to find the smallest allowable value of x_m, so its sign must remain positive. Because each $x_i \leq x_m$, the multiplier m for the term x_m guarantees that the objective function is positive to ensure that we obtain a bounded solution. We reverse the sign of the remaining node variables so that the minimization will seek their largest values, but will still seek the minimum value for x_m. This means that the variable values will be the latest start times for the activities emanating from each of the nodes. Once the earliest and the latest start times for all the nodes have been determined, it is easy to translate this into earliest and latest start times for the activities using the network representation of the project. The resulting value of x_m, the minimum project completion time, is the same for both formulations.

Example *10.4 (continued)*

We will solve Example 10.4 using linear programming. From the network representation of the project given in Figure 10–5, we see that there are a total of eight nodes. The formulation of the problem that gives the earliest start times is

$$\min \sum_{i=1}^{8} x_i$$

[1] We used letters to represent activities in the earlier sections, but for most computer-based systems activities are represented in the form (i, j), where i is the origination node and j is the destination node.

subject to the following:

$$
\begin{aligned}
x_2 - x_1 &\geq 3 && \text{(A)}\\
x_3 - x_2 &\geq 4 && \text{(B)}\\
x_4 - x_2 &\geq 2 && \text{(C)}\\
x_3 - x_4 &\geq 0 && (P_1)\\
x_5 - x_4 &\geq 5 && \text{(E)}\\
x_6 - x_4 &\geq 3 && \text{(F)}\\
x_6 - x_5 &\geq 0 && (P_2)\\
x_7 - x_3 &\geq 6 && \text{(D)}\\
x_7 - x_6 &\geq 5 && \text{(H)}\\
x_7 - x_5 &\geq 7 && \text{(G)}\\
x_8 - x_7 &\geq 8 && \text{(I)}\\
x_i &\geq 0 \qquad \text{for } 1 \leq i \leq 8.
\end{aligned}
$$

The activity giving rise to each constraint is shown in parentheses. Constraints for the pseudo activities must be included as well. Also, the constraint corresponding to P_1 is $x_3 - x_4 \geq 0$, and not vice versa, because P_1 corresponds to the directed arc from node 4 to node 3. The relevant portion of the Excel output is given here.

```
Target Cell (Min)
     Cell       Name    Original Value  Final Value
     $M$7    Min  Value               0           77

Adjustable Cells
     Cell       Name    Original Value  Final Value
     $B$5     x1                    0            0
     $C$5     x2                    0            3
     $D$5     x3                    0            7
     $E$5     x4                    0            5
     $F$5     x5                    0           10
     $G$5     x6                    0           10
     $H$5     x7                    0           17
     $I$5     x8                    0           25
```

The minimum project completion time is the value of x_8, which is 25 weeks. The earliest times correspond to nodes rather than activities, so these times must be converted to activity times. This is easy to do by referring to the network representation in Figure 10–5. For example, because both B and C emanate from node 2, they would have earliest start times of 3 (the value of x_2). You should satisfy yourself that the earliest start times for the other activities obtained in this way agree with the solution we obtained in Section 10.2.

Changing the objective function to $8x_8 - \Sigma x_i$ and rerunning Excel gives the following output:

```
Target Cell (Min)
     Cell       Name    Original Value  Final Value
     $M$7    Min  Value             148          142

Adjustable Cells
     Cell       Name    Original Value  Final Value
     $B$5     x1                    0            0
     $C$5     x2                    3            3
     $D$5     x3                    7           11
     $E$5     x4                    5            5
     $F$5     x5                   10           10
     $G$5     x6                   10           12
     $H$5     x7                   17           17
     $I$5     x8                   25           25
```

This now gives the latest start times for all the activities. The noncritical activities are those with the latest start times that differ from the earliest start times; the magnitude of the difference is the slack time. Again, you should assure yourself that these results agree with those obtained in Section 10.2.

Linear Programming Formulation of the Cost–Time Problem

Linear programming also can be used to find the optimal completion time when expediting costs are included. The formulation can result in a large linear program even for moderately sized problems. We again will refer to activities by the pair (i, j), where i is the origination node and j is the destination node. Define M_{ij} as the expedited time for activity (i, j) and N_{ij} as the normal time for activity (i, j). Suppose that the cost–time function (Figure 10–10) has the representation $a_{ij} - b_{ij}t_{ij}$, where a_{ij} is the y intercept, b_{ij} the slope, and t_{ij} the activity time. Let C be the indirect cost per day. Then the linear programming formulation of the problem of finding the optimal project completion time is

$$\min \sum_{\text{all}(i,j)} [a_{ij} - b_{ij}t_{ij}] + Cx_m$$

subject to

$$\begin{aligned} x_j - x_i &\geq t_{ij} && \text{for all activities } (i, j), \\ t_{ij} &\leq N_{ij}, \\ t_{ij} &\geq M_{ij}, \\ x_i &\geq 0 && \text{for } 1 \leq i \leq m, \\ t_{ij} &\geq 0 && \text{for all activities } (i, j). \end{aligned}$$

Note that the a_{ij} terms in the objective function are constants and can be eliminated without altering the solution. These constants can be added later to find the optimal value of the objective function.

The activity times, t_{ij}, are now problem variables rather than given constants. This linear programming problem is considerably larger than the one in Section 10.3. There is now a variable for each node in the network *and* a variable for each activity. However, the added burden is still less work than finding the minimum cost solution manually.

Example *10.5 (continued)*

We will solve Example 10.5 using linear programming. The relevant data for developing the linear programming formulation are

Activity	Node Representation	a_{ij}	b_{ij}	M_{ij}	N_{ij}
A	(1, 2)	4,000	1,000	1	3
B	(2, 3)	12,000	2,000	3	4
C	(2, 4)	—	—	2	2
D	(3, 7)	12,000	1,500	4	6
E	(4, 5)	9,000	1,300	4	5
F	(4, 6)	6,000	1,500	2	3
G	(5, 7)	12,900	1,200	4	7
H	(6, 7)	6,000	600	4	5
I	(7, 8)	20,800	1,600	5	8

The resulting linear program is

$$\begin{aligned} \text{Min}\{&-1{,}000t_{12} - 2{,}000t_{23} - 1{,}500t_{37} - 1{,}300t_{45} - 1{,}500t_{46} \\ &- 1{,}200t_{57} - 600t_{67} - 1{,}600t_{78} + 1{,}500x_8\} \end{aligned}$$

subject to

$$
\begin{aligned}
x_2 - x_1 - t_{12} &\geq 0,\\
x_3 - x_2 - t_{23} &\geq 0,\\
x_4 - x_2 - t_{24} &\geq 0,\\
x_3 - x_4 &\geq 0,\\
x_5 - x_4 - t_{45} &\geq 0,\\
x_6 - x_4 - t_{46} &\geq 0,\\
x_6 - x_5 &\geq 0,\\
x_7 - x_3 - t_{37} &\geq 0,\\
x_7 - x_6 - t_{67} &\geq 0,\\
x_7 - x_5 - t_{57} &\geq 0,\\
x_8 - x_7 - t_{78} &\geq 0,\\
1 \leq t_{12} &\leq 3,\\
3 \leq t_{23} &\leq 4,\\
4 \leq t_{45} &\leq 5,\\
2 \leq t_{46} &\leq 3,\\
4 \leq t_{57} &\leq 7,\\
4 \leq t_{67} &\leq 5,\\
5 \leq t_{78} &\leq 8,\\
4 \leq t_{37} &\leq 6,\\
t_{24} &= 2,\\
x_i \geq 0 \quad &\text{for } 1 \leq i \leq 8,\\
t_{ij} \geq 0 \quad &\text{for all activities } (i, j).
\end{aligned}
$$

The upper and the lower bound constraints on the t_{ij} variables would have to be entered as two separate constraints into the computer, giving a total of 28 constraints.

The Excel output for this problem is

```
Target Cell (Min)
     Cell       Name   Original Value  Final Value
     $T$7    Min  Value               0       -19500

Adjustable Cells
     Cell       Name   Original Value  Final Value
     $B$5     x1                     0            0
     $C$5     x2                     0            1
     $D$5     x3                     0            6
     $E$5     x4                     0            3
     $F$5     x5                     0            7
     $G$5     x6                     0            7
     $H$5     x7                     0           12
     $I$5     x8                     0           20
     $J$5     t12                    0            1
     $K$5     t23                    0            4
     $L$5     t24                    0            2
     $M$5     t37                    0            6
     $N$5     t45                    0            4
     $O$5     t46                    0            3
     $P$5     t57                    0            5
     $Q$5     t67                    0            5
     $R$5     t78                    0            8
```

The optimal length of the project is given by the value of $x_8 = 20$, and the optimal activity times are the values of t_{ij}. Notice that these times agree with the activity times we found to be optimal in Section 10.3. The total cost of this solution is found by adding Σa_j, which is \$82,700, and the cost for activity C, which was not included in the objective function. (We could have included the cost of activity C in the objective function by adding the term $1{,}000t_{24}$, as t_{24} is fixed at 2.) The resulting value for the total cost of the project at 20 days is \$82,700 + \$2,000 − \$19,500 = \$65,200.

In general, linear programming is not an efficient way to solve large network problems. Algorithms that exploit the network structure of the problem are far more efficient. Many commercial software products based on such algorithms are available for project scheduling. Even PC-based software products are capable of solving large projects. Linear programming is a useful tool for solving moderately sized problems. However, for those with a large project scheduling problem, or with an ongoing need for project scheduling, we recommend a dedicated project scheduling program. See Section 10.9 for a comprehensive discussion of PC-based project management software.

Problems for Section 10.4

11. Solve Problem 3 by linear programming.
12. Solve Problem 4 by linear programming.
13. Formulate Problem 5 as a linear program.
14. Solve Problem 6 by linear programming.
15. Solve Problem 8 by linear programming.
16. Solve Problem 10 by linear programming.

10.5 PERT: PROJECT EVALUATION AND REVIEW TECHNIQUE

PERT is a generalization of CPM to allow uncertainty in the activity times. When activity times are difficult to predict, PERT can provide estimates of the effect of this uncertainty on the project completion time. However, for reasons that will be given in detail, the results of the analysis are only approximate. Let T_i be the time required to complete activity i. In this section we will assume that T_i is a random variable. Furthermore, we assume that the collection of random variables $T_1, T_2, \ldots, T_n$ is mutually independent. The first issue addressed is the appropriate form of the distribution of these random variables. Define the following quantities:

a = Minimum activity time,

b = Maximum activity time,

m = Most likely activity time.

As an example, suppose that $a = 5$ days, $b = 20$ days, and $m = 17$ days. Then the probability density function of the activity time is as pictured in Figure 10–12. The density function should be zero for values less than 5 and more than 20, and should be a maximum at $t = 17$ days. The point where the density is a maximum is known as the *mode* in probability theory. The *beta distribution* is a type of probability distribution defined on a finite interval that may have its modal value anywhere on the interval. For this reason, the beta distribution is usually used to describe the distribution of individual activity times. The assumption of beta-distributed activity times is used to justify simple approximation formulas for the mean and the variance, but is rarely used to make probabilistic statements concerning individual activities.

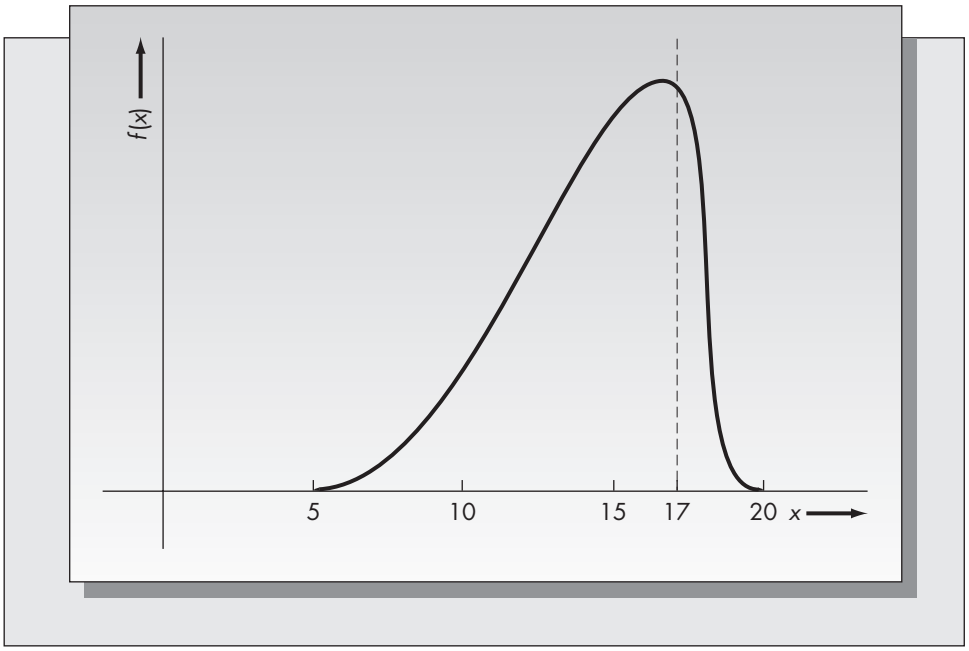


FIGURE 10–12
Probability density of activity time

The beta distribution assumption is used to justify the approximations of the mean μ and the standard deviation σ of each activity time. The traditional PERT method is to estimate μ and σ from a, b, and m using the following formulas:

$$\mu = \frac{a + 4m + b}{6}, \qquad \sigma = \frac{b - a}{6}.$$

The formula for the standard deviation seems to be based on the following property of the normal distribution: limits at distance 3σ to either side of the mean for a normal variate include all the population with probability exceeding .99. In view of this property, it is assumed that there are six standard deviations from a to b. The formula for the mean is obtained by assuming the variance approximation as well as the beta distribution for the activity time. Given the variance, a, b, and m, the mean is determined by solving a cubic equation. Calculating μ for various values of the other parameters and developing the best fit by linear regression results in the one-four-one weighting scheme. (See Archibald and Villoria, 1967, p. 449.)

Squaring both sides in the formula for σ gives

$$\sigma^2 = \frac{(b - a)^2}{36}$$

where σ^2 is the variance.

The uniform distribution is a special case of the beta distribution. If the job time has a uniform distribution from a to b, the density function would be rectangular, as pictured in Figure 10–13. The variance of a uniform variate is $(b - a)^2/12$. Because we would expect the variance to be less for a peaked distribution such as the one pictured in Figure 10–12, the approximation for σ^2 recommended earlier seems reasonable. Note that the one-four-one weighting scheme gives the correct value of the mean, $(a + b)/2$, if one substitutes $m = (a + b)/2$ as the mode.

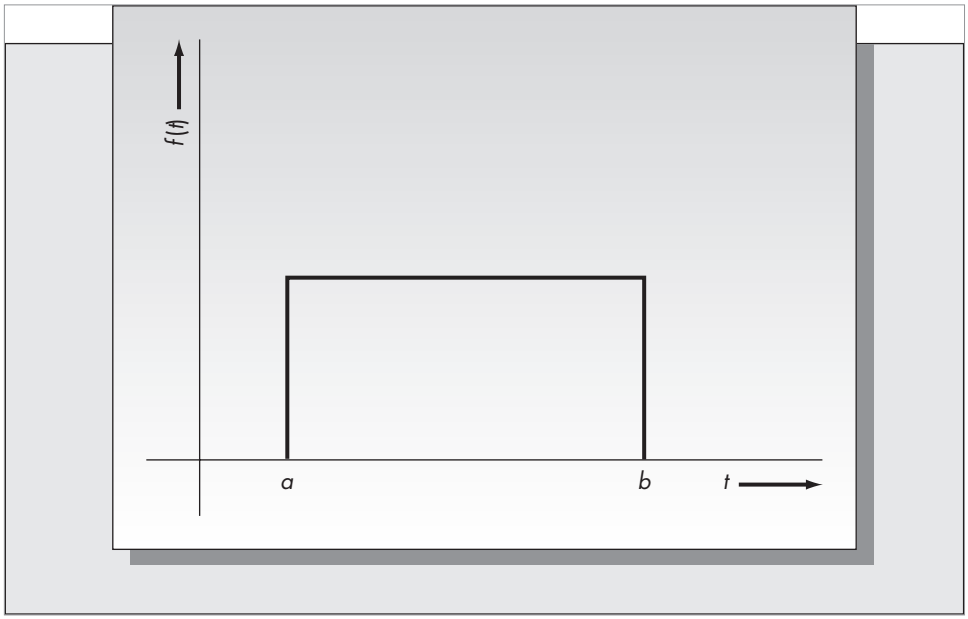


FIGURE 10–13
Uniform density of activity times

In PERT, one assumes that the distribution of the total project time is normal. The *central limit theorem* is used to justify this assumption. Roughly, the central limit theorem says that the distribution of the sum of independent random variables is approximately normal as the number of terms in the sum grows large. In most cases, convergence occurs quickly. Hence, because the total project time is the sum of the times of the activities along the critical path, it should be approximately normally distributed as long as activity times are independent.

Suppose that the critical activity times are $T_1, T_2, \ldots, T_k$. Then the total project time T is

$$T = T_1 + T_2 + \cdots + T_k.$$

It follows that the mean project time, $E(T)$, and the variance of the project time, $\text{Var}(T)$, are given by

$$E(T) = \mu_1 + \mu_2 + \cdots + \mu_k,$$
$$\text{Var}(T) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_k^2.$$

These formulas are based on the following facts from probability theory: the expected value of the sum of *any* set of random variables is the sum of the expected values, and the variance of a sum of *independent* random variables is the sum of the variances. Independence of activity times is required in order to easily obtain the variance of the project time and to justify the application of the central limit theorem. One could modify the variance formula to incorporate correlations among the activities, but the assumption of normality of the project completion time might no longer be accurate. For these reasons, explicit treatment of the dependencies among activity times is rare in practice.

Summarizing the PERT method.

1. For each activity obtain estimates of a, b, and m. These estimates should be supplied by the project manager or by someone familiar with similar projects.
2. Using these estimates, compute the mean and the variance of each of the activity times from the given formulas.

3. Based on the mean activity times, use the methods of Section 10.4 to determine the critical path.
4. Once the critical activities are identified, add the means and the variances of the critical activities to find the mean and the variance of the total project time.
5. The total project time is assumed to be normally distributed with the mean and the variance determined in step 4.

Using the assumption that the project time is normally distributed, we can address a variety of issues. An example will illustrate the method.

Example *10.6* Consider the case study of Example 10.4 of the two computer consultants developing a software project. Before embarking on the project, they decide that it is important to consider the uncertainty of the times required for certain tasks. As with any software project, unanticipated bugs can surface and cause significant delays. Based on their past experience, the programmers decide that the values of a, b, and m are as follows:

Activity	Min (a)	Most Likely (m)	Max (b)	$\mu = \frac{a + 4m + b}{6}$	$\sigma^2 = \frac{(b - a)^2}{36}$
A	2	3	4	3	0.11
B	2	4	10	4.67	1.78
C	2	2	2	2	0
D	4	6	12	6.67	1.78
E	2	5	8	5	1.00
F	2	3	8	3.67	1.00
G	3	7	10	6.83	1.36
H	3	5	9	5.33	1.00
I	5	8	18	9.17	4.69

Using these values, we have computed the means and the variances of each of the activity times. Activity C, the design of the flowchart, requires precisely two days. Hence, the variance of this activity time is zero.

The traditional PERT method is to compute the critical path based on the mean activity times. In this example, the introduction of uncertainty does not alter the critical path. It is still A–C–E–G–I. (However, it is possible that in other cases the critical path based on the mean times will *not* be the same as the critical path based on the most likely times.) The expected project completion time, $E(T)$, is simply the sum of the mean activity times along the critical path. In this example,

$$E(T) = 3 + 2 + 5 + 6.83 + 9.17 = 26 \text{ weeks.}$$

Similarly, the variance of the project completion time, $\text{Var}(T)$, is the sum of the variances of the activities along the critical path. In this case,

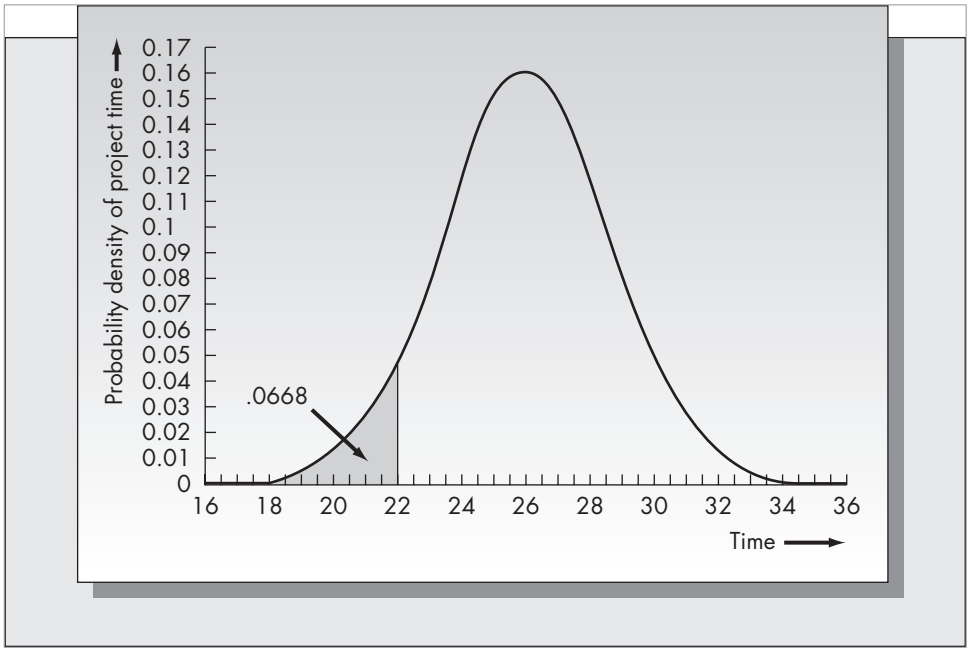
$$\text{Var}(T) = 0.11 + 0 + 1.0 + 1.36 + 4.69 = 7.16.$$

The assumption made is that the total project completion time T is a normal random variable with mean $\mu = 26$ and standard deviation $\sigma = \sqrt{7.16} = 2.68$. We can now answer a variety of specific questions concerning this project.

Example *10.7* Answer the following questions about the project scheduling problem described in Example 10.6.

1. What is the probability that the project can be completed in under 22 weeks?
2. What is the probability that the project requires more than 28 weeks?
3. Find the number of weeks required to complete the project with probability .90.

FIGURE 10–14
Answer to Example 10.7, Part 1



Solution

1. We wish to compute $P\{T < 22\}$.

$$P\{T < 22\} = P\left\{\frac{T - \mu}{\sigma} < \frac{22 - \mu}{\sigma}\right\} = P\left\{Z < \frac{22 - 26}{2.68}\right\}$$
$$= P\{Z < -1.5\} = .0668$$

Z is the standard normal variate. The probability is from Table A–1 of the standard normal distribution in Appendix A. The solution is pictured in Figure 10–14.

2. $$P\{T > 28\} = P\left\{\frac{T - \mu}{\sigma} > \frac{28 - \mu}{\sigma}\right\} = P\left\{Z > \frac{28 - 26}{2.68}\right\}$$
$$= P\{Z > .75\} = .2266.$$

The solution to this problem is pictured in Figure 10–15.

3. Here we wish to find the value of t such that $P\{T \le t\} = .90$:

$$.90 = P\{T \le t\} = P\left\{Z < \frac{t - \mu}{\sigma}\right\}.$$

It follows that

$$\frac{t - \mu}{\sigma} = z_{.90},$$

or $t = \mu + \sigma z_{.90} = 26 + (2.68)(1.28) = 29.43$ weeks.

The notation $z_{.90}$ stands for the 90th percentile of the standard normal distribution. That is, $P\{Z \le z_{.90}\} = .90$. Its value, 1.28, is found from Table A–1 in Appendix A at the back of the book. The solution is pictured in Figure 10–16.

FIGURE 10–15
Answer to Example 10.7, Part 2

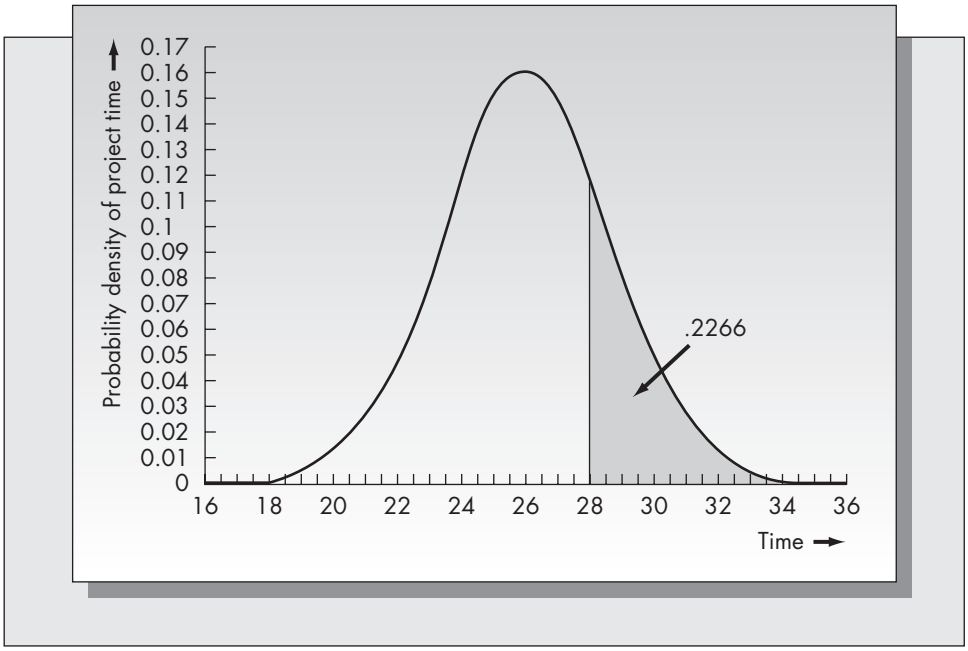
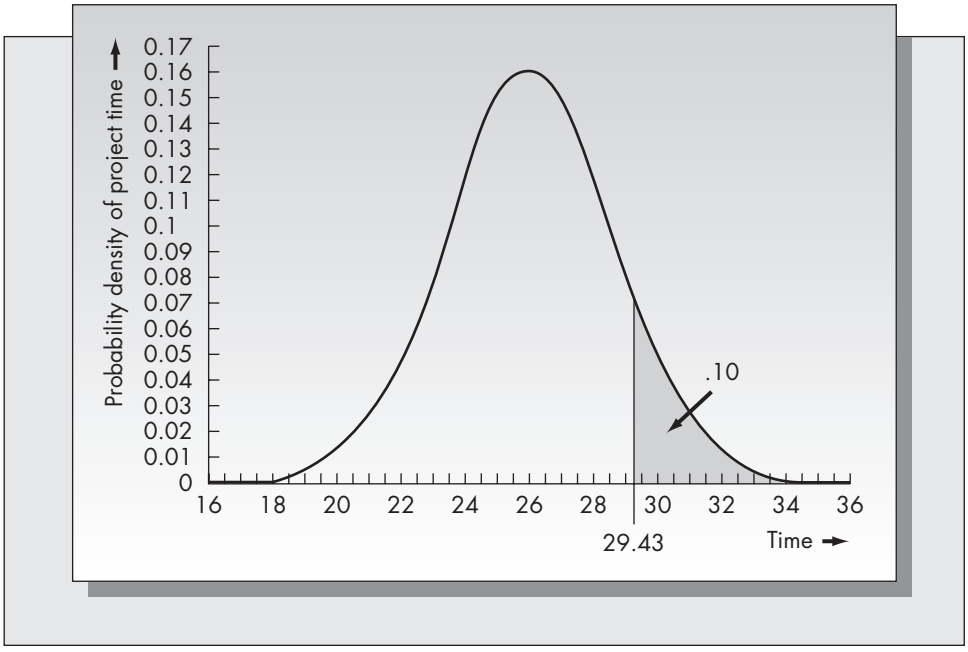


FIGURE 10–16
Answer to Example 10.7, Part 3



Path Independence

A serious limitation of the PERT method is the assumption that the path with the longest expected completion time is necessarily the critical path. In most networks there is a positive probability that a path other than the one with the longest expected completion time, in fact, will be critical. When this is the case, the PERT calculations just presented could be very misleading.

Consider Example 10.7. Suppose that the project has been completed and the following realizations of the activity times observed:

Activity	Actual Time Required to Complete	Activity	Actual Time Required to Complete
A	3.4	F	5.0
B	4.0	G	6.2
C	2.0	H	7.2
D	7.0	I	13.0
E	3.5		

With these activity times, the critical path is no longer A–C–E–G–I, but is now A–C–F–H–I. Hence, the PERT assumption that a fixed path is critical is not accurate.

Unfortunately, determining the exact distribution of the critical path is difficult. In general, this distribution is not known. The difficulty is that, because different paths include the same activities, the times required to complete different paths are *dependent* random variables.

Depending on the particular configuration of the project network, assuming independence of two or more paths may be more accurate than assuming a single critical path. Consider the following example.

Example *10.8*

Consider the project pictured in Figure 10–17. In the figure we have included the means and the variances of the activity times. Assume that these times are expressed in weeks. There are exactly two paths from node 1 to node 9: A–C–E–G–I and B–D–F–H–I. The expected times of these paths are 41 and 40 weeks, respectively. Using PERT, we would assume that the critical path is A–C–E–G–I. However, there is almost an equal likelihood that path B–D–F–H–I will turn out to be critical after the activity times are realized.

In this example, the two paths have only activity I in common. Hence, the paths are almost statistically independent. Performing the calculations assuming two independent paths could give very different results from assuming a unique critical path.

Suppose that we wish to determine the probability that the project is completed within 43 weeks. Let T_1 be the time required to complete path A–C–E–G–I and T_2 the time required

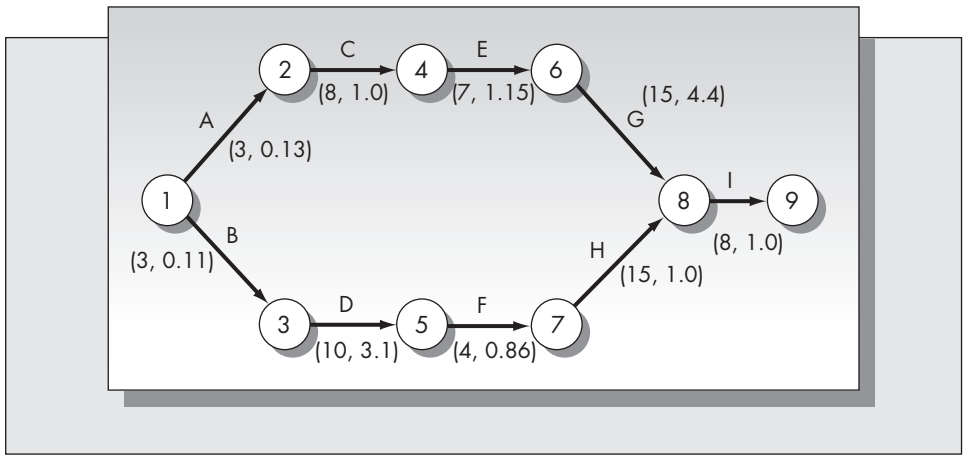


FIGURE 10–17
Network for PERT Example 10.8

to complete path B–D–F–H–I. Using the methods of Section 10.4, we conclude that T_1 and T_2 are approximately normally distributed, with

$$E(T_1) = 41,$$
$$\text{Var}(T_1) = 0.13 + 1 + 1.15 + 4.4 + 1 = 7.68,$$

and

$$E(T_2) = 40,$$
$$\text{Var}(T_2) = 0.11 + 3.1 + 0.86 + 3.0 + 1 = 8.07.$$

Let T be the total project time. Clearly, $T = \max(T_1, T_2)$. It follows that

$$P\{T < 43\} = P\{\max(T_1, T_2) < 43\} = P\{T_1 < 43, T_2 < 43\}$$

(the last equality follows from the fact that if the maximum of two quantities is less than a constant, then both quantities must be less than that constant)

$$= P\{T_1 < 43\}\, P\{T_2 < 43\}$$

(which follows from the assumption of path independence)

$$= P\left\{Z < \frac{43 - 41}{\sqrt{7.68}}\right\} P\left\{Z < \frac{43 - 40}{\sqrt{8.07}}\right\}$$
$$= P\{Z < 0.72\} P\{Z < 1.05\}$$
$$= (.7642)(.8413) = .6429.$$

The methods of Section 10.4 would have given the estimate of completing the project within 43 weeks as .7642. For this network, .6429 is far more accurate.

Certain calculations are more complex if we assume path independence. For example, suppose that we wanted to know the number of weeks required to complete the project with probability .90. Then we wish to find t to satisfy

$$P\{T_1 < t\}\, P\{T_2 < t\} = .90$$

or

$$P\left\{Z < \frac{t - 41}{2.77}\right\} P\left\{Z < \frac{t - 40}{2.84}\right\} = .90.$$

One calculates t by trial and error. Because the probabilities are likely to be close, a good starting guess for the value of each probability is $\sqrt{.90} \approx .95$, which gives $t = 45.6$ for the first term and $t = 44.7$ for the second term. The correct value is approximately $t = 45.2$ weeks, which results in a value of .904 for the product of the two probabilities.

Because the two paths for this project have only one activity in common, the assumption of path independence is quite reasonable and the answers obtained in this manner far more accurate than those found by assuming a unique critical path. In most networks, however, paths may have many activities in common, and the assumption of path independence may be inaccurate. Consider the network for Example 10.4 pictured in Figure 10–5. In this example, there are a total of five paths:

A–B–D–I,
A–C–P_1–D–I,
A–C–E–G–I,
A–C–E–P_2–H–I,
A–C–F–H–I.

The expected lengths of the five paths are, respectively, 23.51, 20.84, 26, 23, and 23.17. Unfortunately, path A–C–F–H–I contains three activities in common with the longest expected

path, A–C–E–G–I. In such a case, it is not clear which choice will give more accurate results: including this path and assuming path independence or excluding it from consideration.

We will compute the probability that the project can be completed in under 22 weeks assuming path independence.

Path	Expected Completion Time	Variance of Completion Time
A–B–D–I	23.5	8.36
A–C–P_1–D–I	20.8	6.58
A–C–E–G–I	26.0	7.16
A–C–E–P_2–H–I	23.0	6.80
A–C–F–H–I	23.2	6.80

If T is the project completion time and $T_1, \ldots, T_5$ are the times required to complete each of the five paths listed in the table, above, then

$$T = \max(T_1, \ldots, T_5).$$

It follows that

$$\begin{aligned} P\{T < 22\} &= P\{T_1 < 22, T_2 < 22, T_3 < 22, T_4 < 22, T_5 < 22\} \\ &\approx P\{T_1 < 22\}\, P\{T_2 < 22\}\, P\{T_3 < 22\}\, P\{T_4 < 22\}\, P\{T_5 < 22\}. \end{aligned}$$

Again, assuming a normal distribution for each of the path completion times, we have

$$P\{T_1 < 22\} = P\left\{Z < \frac{22 - 23.5}{\sqrt{8.36}}\right\} = P\{Z < -0.52\} = .3015,$$

$$P\{T_2 < 22\} = P\left\{Z < \frac{22 - 20.8}{\sqrt{6.58}}\right\} = P\{Z < -0.47\} = .6808,$$

$P\{T_3 < 22\} = .0668$ (from Example 9.7, Part 1).

$$P\{T_4 < 22\} = P\left\{Z < \frac{22 - 23}{\sqrt{6.8}}\right\} = P\{Z < -0.38\} = .3520,$$

$$P\{T_5 < 22\} = P\left\{Z < \frac{22 - 23.2}{\sqrt{6.8}}\right\} = P\{Z < -0.46\} = .3228.$$

Hence, it follows that

$$P\{T < 22\} \approx (.3015)(.6808)(.0668)(.3520)(.3228) = .0016.$$

The true value of the probability will fall somewhere between .0016 and the probability computed by traditional PERT methods, .0668. Assuming path independence can have a very significant effect on the probabilities. For this example, it is safe to say that the likelihood that the consultants complete the project in less than 22 weeks is far less than 6.68 percent and is probably well under 1 percent.

Problems for Section 10.5

17. Referring to Example 10.4 and Figure 10–5, what is the probability that node 6 (i.e., the completion of activities A, C, E, and F) is reached before 12 weeks have elapsed?
18. With reference to Example 10.4 and Figure 10–5, what is the conditional probability that the project is completed by the end of week 25, given that activities A through H (node 7) are completed at the end of week 15? Assume that the time required to complete activity I is normally distributed for your calculation.

19. Consider the following PERT time estimates:

Activity	Immediate Predecessors		*a*	*m*	*b*
	A	—	2	5	9
	B	A	1	6	8
	C	A	3	5	12
	D	B	2	4	12
	E	B, C	4	6	8
	F	B	6	7	8
	G	D, E	1	2	6
	H	F, G	4	6	16

a. Draw a network for this project and determine the critical path based on the *most likely times* by inspection.

b. Assuming that the critical path is the one you identified in part (*a*), what is the probability that the project will be completed before 28 weeks? Before 32 weeks?

c. Assuming that the critical path is the one you identified in part (*a*), how many weeks are required to complete the project with probability .95?

20. Consider the project network pictured in Figure 10–18. Assume that the times attached to each node are the means and the variances of the project completion times, respectively.

a. Identify all paths from nodes 1 to 6.

b. Which path is critical based on the expected completion times?

c. Determine the probability that the project is completed within 20 weeks assuming that the path identified in part (*b*) is critical.

d. Using independence of paths A–C–E and B–F–G only, recompute the answer to part (*c*).

e. Recompute the answer to part (*c*) assuming independence of the paths identified in part (*a*).

f. Which answer, (*c*), (*d*), or (*e*), is probably the most accurate?

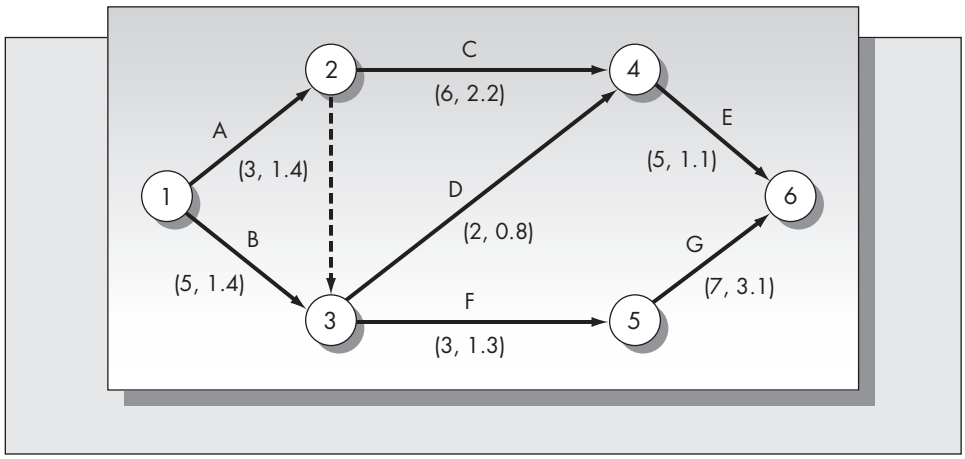


FIGURE 10–18
Project network (for Problem 20)

Snapshot Application

WARNER ROBINS STREAMLINES AIRCRAFT MAINTENANCE WITH CCPM PROJECT MANAGEMENT

Annual spending on aircraft maintenance and repair amounts to close to $100 billion annually worldwide, about half of which is accounted for by the military. Of this amount, about $20 billion is spent by the U.S. military alone. Warner Robins Air Logistics Center near Macon, Georgia, provides maintenance, repair, and overhaul for its customer, the U.S. Air Force. Warner Robins supports several aircraft including the C-5 Galaxy, C-17 Globemaster, C-130 Hercules transport, and the F-15 Eagle fighter jet.

A team of researchers considered the task of scheduling operations for the C-5. The C-5 line has 24 frontline supervisors and about 460 mechanics organized into several skill groups. The line operates in two shifts. Scheduling repair is a challenge for several reasons. Repairs due to damage are quite unpredictable, and even scheduled maintenance may have unforeseen problems crop up. Typical program depot maintenance may require as much as 40,000 to 50,000 worker-hours, but this figure could be as much as 10,000 hours more than anticipated. These long lead times and inherent uncertainties make due date scheduling extremely difficult.

To help ameliorate these problems, the research team implemented a new approach to project scheduling. The approach is based on concepts developed by Elihu Goldratt (the developer of OPT, a defunct manufacturing scheduling tool). Goldratt labeled his method critical chain project management (CCPM), an alternative to PERT for scheduling projects with uncertain activity times. In PERT, one considers the distribution of each activity time, and effectively builds in a buffer for each activity. CCPM considers a buffer for the entire project only. One then identifies those activities that consume the buffer after the fact. As an example, in the case of the C-5 line, the managers were able to identify floorboard replacement as an activity that consistently consumed the buffer. By focusing attention on this activity, management was able to reduce the time for this activity by 45 percent, thus resulting in substantial improvements in overall project completion times. Other activities were also identified as being trouble spots, and the overall efficiency of the C-5 schedule significantly improved. The research team estimated overall revenue savings of almost $50 million annually as a result of improved scheduling of the C-5 maintenance, repair, and overhaul operation. Similar scheduling methods were also being considered for the other aircraft programs at the base.

Source: M. M., Srinivasan, W. D. Best, and S. Chandrasekaran. "Warner Robins Air Logistics Center Streamlines Aircraft Repair and Overhaul," *Interfaces* 37(2007), pp. 7–21.

21. Consider the project described in Problem 3. Suppose that the activity times are random variables with a constant *coefficient of variation* of 0.2. (The coefficient of variation is the ratio σ/μ.) Assume that the times given in Problem 3 are the mean activity times.

    a. Compute the standard deviation of each of the activity times.

    b. Find the mean and the variance of the path with the longest expected completion time.

    c. Using the results of part (*b*), estimate the probability that the project is completed within 20 weeks.

    d. Using the results of part (*b*), estimate the number of weeks required to complete the project with probability .85.

10.6 RESOURCE CONSIDERATIONS

Resource Constraints for Single-Project Scheduling

An implicit assumption made throughout this chapter is that sufficient resources are available and only the *technological constraints* (precedence relationships) are important for setting schedules. In most environments, however, resource constraints cannot

be ignored. Examples of limited resources that would affect project schedules are workers, raw materials, and equipment. Because traditional CPM ignores resource considerations, one manager described it as a "feasible procedure for producing a nonfeasible schedule."

Determining optimal schedules for complex project networks subject to resource limitations is an extremely difficult combinatorial problem. A single project may require a variety of different resources, or many different projects may compete for one or more resources. Heuristic (approximate) methods are generally used to modify schedules obtained by more conventional means.

Allocation of scarce resources among competing activities has been an area of considerable interest in recent years. This section considers the allocation process using the example of the programming project (Example 10.4) discussed earlier in the chapter.

Example *10.9* Consider Example 10.4 about the two programmers developing a software package. An analysis based only on precedence considerations showed that the project could be completed in 25 weeks without expediting any activities. This analysis, however, is based on the assumption of unlimited resources. In fact, as we noted earlier, it is not possible for the programmers to complete the project within 25 weeks without obtaining additional help. What is the minimum time required for them to complete the project if they do not obtain additional assistance?

Assume that North is responsible for activities B (design of the graphic icons) and D (design of the input/output screens), and Bonner has the expertise required for the development of modules 1 and 2 (activities E and F). Furthermore, assume that either of the programmers can perform activities G and H (modules 3 and 4), but both must work on the final merging and testing. Consider the Gantt chart pictured in Figure 10–7. It shows an infeasible schedule because activities D, E, and H must be done simultaneously between weeks 7 and 8, and activities D, G, and H must be done simultaneously between weeks 10 and 13.

If only Bonner and North are to do the programming, they must reorder the activities so as not to schedule more than two activities simultaneously. Furthermore, they must be certain that these are not two activities that can only be done by one of them. First, they try to find a feasible schedule without delaying the project by rescheduling the noncritical activities within the allowable slack. From Figure 10–7 it is clear that no matter how one rearranges the noncritical activities within the available slack, there is no way to avoid scheduling three activities simultaneously at some point.

This means that they cannot complete the project within 25 weeks without additional help. We determine a feasible schedule by considering the sequence of activities performed by each programmer. At week 3 North begins work on B and Bonner on C. At week 5 Bonner begins work on E, and at week 7 North begins work on D. Because F is module 2 coding, which must be done by Bonner, F must now follow E, so that F starts at week 10. Both G and H can begin at week 13 when both programmers are free. Finally, I starts at week 20 and ends at week 28. Hence, the project time has increased from 25 weeks to 28 weeks as a result of resource considerations.

Figure 10–19 shows the Gantt chart for the modified schedule. This schedule is feasible because there are never more than two activities scheduled at any point in time. North begins work at week 3 and works on the activities B, D, G, and I, while Bonner, who also begins working at week 3, performs activities C, E, F, H, and I. Note that activities G and H are interchangeable.

Including resource considerations has a number of interesting consequences. For one, the critical path is no longer the same. Activities B, D, and F no longer have any slack, so they are now critical. Furthermore, there is considerably less slack time overall.

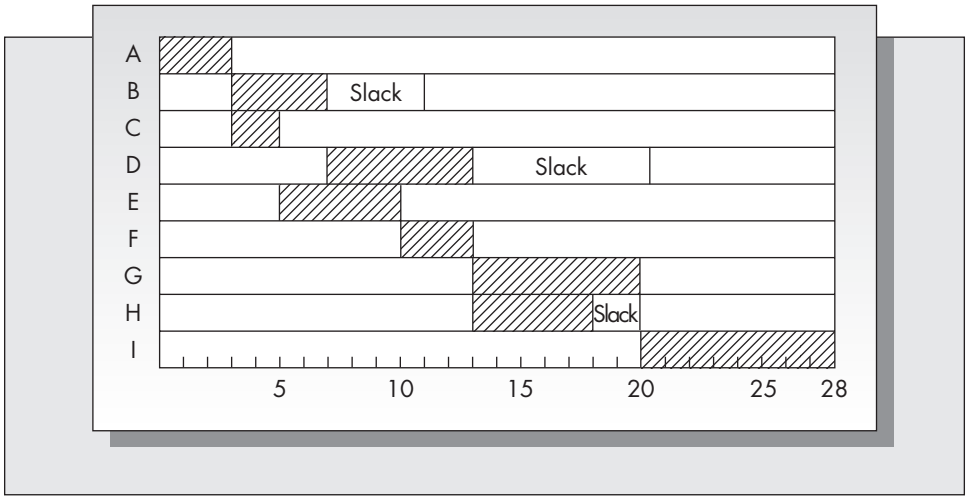


FIGURE 10–19
Gantt chart for modified schedule for Example 10.9

In general, the inclusion of resource constraints has the following effects:

1. The total amount of scheduled slack is reduced.
2. The critical path may be altered. Furthermore, the zero-slack activities may not necessarily lie along one or more critical paths.
3. Earliest- and latest-start schedules may not be unique. They depend on the particular rules that are used to resolve resource limitations.

Several heuristic methods for solving this problem exist. Most involve ranking the activities according to some criterion and resolving resource conflicts according to the sequence of the ranking. Some of the ranking rules that have been suggested include the following:

1. *Minimum job slack*. Priority is given to activities with the smallest slack.

2. *Latest finishing times*. When resource conflicts exist, this rule assigns priority to the activity with the minimum latest finishing time.

3. *Greatest resource demand*. This rule assigns priority on the basis of the total resource requirements of all types, giving highest priority to the activities having greatest resource demand. The rationale behind this method is to give priority to potential bottleneck activities.

4. *Greatest resource utilization*. This rule gives priority to that combination of activities that results in the maximum resource utilization (minimum idle time) in any scheduling interval.

These rules (among others) were compared by Davis and Patterson (1975). Their results indicated that the first two methods on the list tended to be the best performers.

Optimal solution methods for project scheduling under resource constraints exist as well. The formulation may be an integer program requiring some type of branch and bound procedure, or may be solved by a technique known as implicit enumeration. For large networks these methods require substantial computer time, but can provide efficient solutions for moderately sized networks with few resource limitations. Patterson (1984) compares and contrasts three optimal solution methods.

Resource Constraints for Multiproject Scheduling

Dealing with resource constraints in single-product scheduling problems can be difficult, but the difficulties are magnified significantly when a common pool of

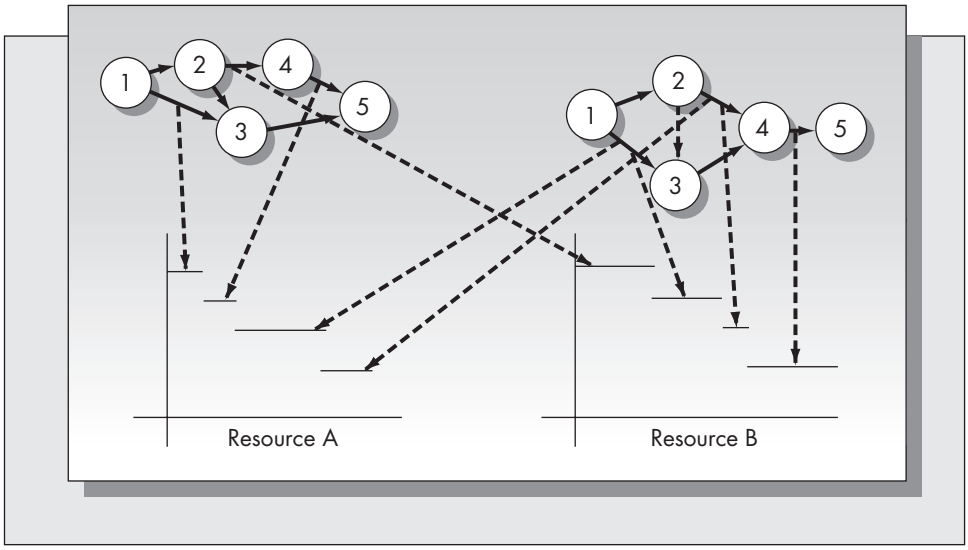


FIGURE 10–20
Two projects sharing two resources

resources is shared by a number of otherwise independent projects. Figure 10–20 shows this type of problem. Two projects require resources A and B. Delaying activities in order to resolve resource conflicts can have far-reaching consequences for all projects requiring the same resources. Commercial computer systems exist that have the capability of dealing with tens of projects and resource types, and possibly thousands of activities.

Resource Loading Profiles

Project planning is a useful means for generating schedules of interrelated activities. One significant kind of planning result involves the loading profiles of the required resources. A loading profile is a representation over time of the resources needed. As long as the requirements associated with each activity are known, one can easily obtain the resulting loading profiles for all required resources.

Example *10.10*

We illustrate the procedure with the case study first introduced in Example 10.4. Let us suppose that the two programmers are developing the program on a single multiuser computer system. The system can segment the random access memory (RAM) between the two users and can segment the permanent storage on the hard disk as well. Both RAM and permanent storage are measured in gigabytes (GB). The requirements for RAM memory and permanent memory for each of the activities comprising the project are given in the following table.

Activity	RAM Required (GB)	Permanent Storage Required (GB)
A	0	0
B	1	60
C	0.5	5
D	2	30
E	1	10
F	3	5
G	1.5	15
H	2	10
I	4	80

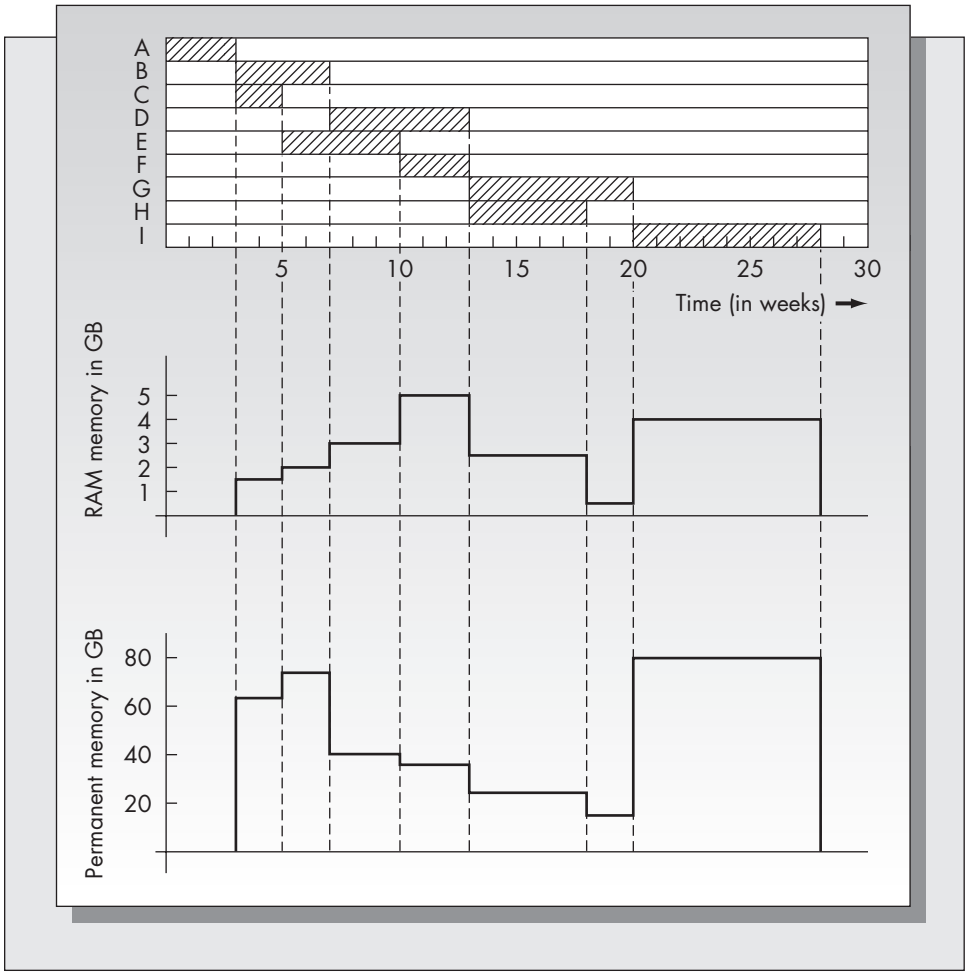


FIGURE 10–21
Load profiles for RAM and permanent memory (refer to Example 10.10)

Assume that the activities comprising the project follow the starting and ending times pictured in the Gantt chart of Figure 10–19. Figure 10–21 shows the resulting load profiles of RAM and permanent memory. According to these profiles, the system will require at least 5 GB of RAM and 80 GB of permanent storage. Notice that we are assuming that these requirements are *not* cumulative. Once a portion of the project is completed, the results can be stored on tape or disk and retrieved at a later time.

Resources are either consumable or nonconsumable. In Example 10.10 the resources were nonconsumable. The workforce is another example of a nonconsumable resource. Typical consumable resources are cash or fuels. An issue that arises with consumable resources is the cumulative amount of the resource consumed at each point in time. We will explore load profiles for consumable resources in the problems at the end of this section.

A desirable feature of load profiles is that they be as smooth as possible. Large variations in resource requirements make planning difficult and may result in exceeding resource availability at some point in time. The idea behind *resource leveling* is to reschedule noncritical activities within the available slack in order to smooth out the pattern of resource usage. Often it is possible to do this rescheduling by inspection. For larger networks a systematic method is desirable. Burgess and Killebrew (1962)

describe a technique for leveling resource profiles that is based on reducing the value of the sum of squares of the resource profile curve by rescheduling activities within the available slack. We will not review the procedure here.

In summary, resource loading profiles provide an important means of determining the requirements imposed by any particular schedule. Given a project schedule, one can usually construct the loading profiles without the aid of a computer. Rather than a strict forecast of requirements, the profiles are often more useful as rough planning guides when significant variations in activity time are anticipated. According to Moder, Phillips, and Davis (1983), they are probably more widely used than any other resource analysis technique.

Problems for Section 10.6

22. For the case problem discussed in this chapter (Example 10.10), suppose that the requirements for RAM and permanent storage are given by

Activity	RAM Required (GB)	Permanent Storage Required (GB)
A	0	0
B	1.5	30
C	2.5	20
D	0.5	10
E	2.0	40
F	1.5	15
G	2.0	25
H	1.5	20
I	4.0	50

a. Determine the load profiles for RAM and permanent storage assuming that the activities are scheduled according to the Gantt chart of Figure 10–7.

b. Determine the load profiles for RAM and permanent storage assuming that the activities are scheduled according to the Gantt chart of Figure 10–19.

23. Consider the project described in Problem 3. Three machines, M1, M2, and M3, are required to complete the project. The requirements for each activity are as follows:

Activity	Machines Required
A	M1, M2
B	M1, M3
C	M2
D	M1, M2
E	M2
F	M1, M3
G	M2, M3
H	M1

Determine the minimum time required to complete the project if there is only one machine of each type available. How many weeks are added to the project time when resource requirements are considered?

24. Consider the project described in Problem 6. The tasks require both welders and pipe fitters. The requirements are

Activity	Number of Welders Required	Number of Pipe Fitters Required
A	6	10
B	3	15
C	8	8
D	0	20
E	10	6
F	10	9
G	4	14

a. Determine the load profiles for welders and pipe fitters assuming an early-start schedule.

b. Determine the load profiles assuming a late-start schedule.

25. Consider the project network pictured in Figure 10–22. Activity times, measured in days, are shown in parentheses next to each activity label.

a. Determine the earliest and the latest starting and finishing times for all activities. Draw a Gantt chart based on earliest starting times, but indicate activity slack where appropriate. How many days are required to complete the project?

b. A single critical piece of equipment is required in order to complete the following activities: A, B, C, D, G, and H. Determine a feasible schedule for the project assuming that none of these activities can be done simultaneously.

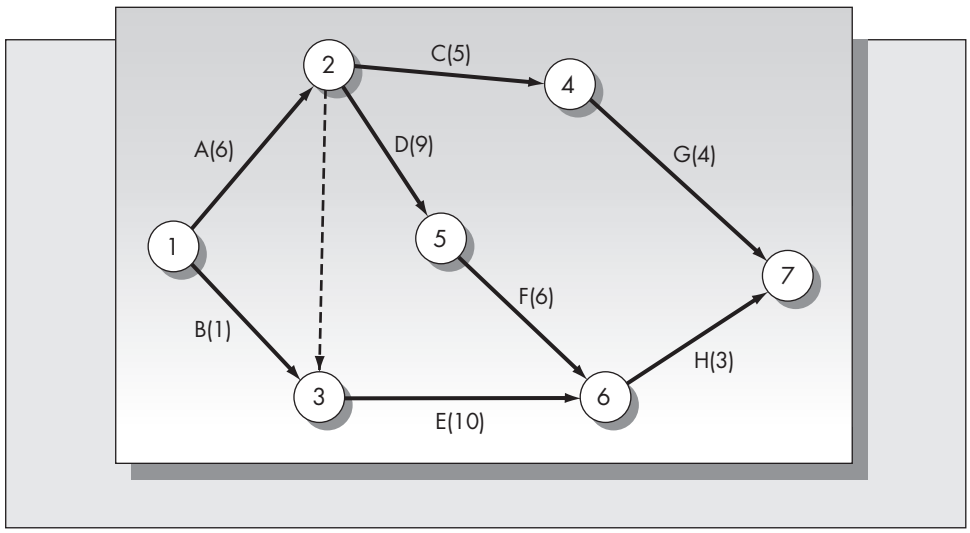


FIGURE 10–22
Project network (for Problem 25)

c. Two resources, R1 and R2, are used for each activity. Assume that these are both consumable resources with daily requirements as follows:

Activity	Daily Requirement of R1	Daily Requirement of R2
A	4	0
B	8	6
C	10	9
D	18	4
E	12	3
F	5	12
G	3	2
H	0	6

Determine resource loading profiles based on the schedule found in part (*b*).

d. Based on the results of part (*c*), determine the *cumulative* amounts of resources R1 and R2 consumed if the schedule found in part (*b*) is used.

10.7 ORGANIZATIONAL ISSUES IN PROJECT MANAGEMENT

This chapter has been concerned with reviewing techniques for assisting with the project management function. Successful project management also depends on effective people management. How the organization is structured can be an important factor in whether or not a project succeeds.

The classic structure of an organization is a line organization. That means that there is a clear pyramid structure: Vice presidents report to the president, directors report to vice presidents, middle managers report to directors, and so forth. In a line organization, usually one person at the bottom is assigned to coordinate several employees who may be in other departments at the same level of the organization. The individual responsible will be given the title of project leader or project officer. The line organization is probably the weakest organizational structure for interdepartmental project management.

At the other end of the organizational spectrum is the divisional project organization. In this setting, employees are completely freed up to the project organization for the life of the project. They report only to the project leader. This is the strongest organizational structure from the point of view of promoting successful project completion. However, the divisional project organization requires frequent shifting of employees among projects, which can be disruptive for the employees and expensive for the company.

More recently, firms have been experimenting with the matrix organization. In the matrix organization, the firm is organized both horizontally and vertically. The vertical structure is the same as in the traditional line organization. The horizontal direction corresponds to individual projects that may span several functional departments. Each employee reports vertically to his or her functional superior and horizontally to his or her project leader.

The matrix organization is a compromise between the pure line organization and the project organization. However, when project teams span functional departmental boundaries, problems arise. Typically, employees' first loyalty will be to their direct functional superior. Dual subordination causes conflicts when demands are made on an employee from two directions at once. Hence, in order for the matrix concept to work, the project leader

must be empowered to set priorities and provide incentives for outstanding performance from project members. In addition, there must be a shared sense of responsibility for the project among project team members. Managing split loyalties among team members is no easy task, however. Texas Instruments Corporation, for example, long advocated the matrix organization, but found that ambiguous lines of authority were causing problems.

The structure of the project team varies with the application. For example, Vollmann et al. (1992) recommend that a project team assigned the task of implementing a manufacturing planning and control system consist of five to eight employees. It would ideally be comprised of one representative from each of marketing, engineering, production planning, line manufacturing, and management information systems departments. A team member from finance also might be desirable. They recommend that the project team be freed from other responsibilities and physically isolated from their functional departments during the project's course (meaning that they advocate a project organization). The best team is comprised of experienced employees from within the company who are familiar with the business. The project team leader should be someone who will ultimately be a user of the new system, such as a professional in production planning or manufacturing.

10.8 HISTORICAL NOTES

It is generally recognized that the two network-based scheduling systems discussed in this chapter, CPM and PERT, were developed almost simultaneously in the United States in the late 1950s. CPM was a result of a joint effort by the Du Pont Company and Remington Rand (Walker and Sayer, 1959). The technique resulted from a study aimed at reducing the time required to perform plant overhaul, maintenance, and construction. The CPM methodology outlined in this report included the cost–time trade-off of indirect and direct costs discussed in Section 9.3, as well as the methods for developing the project network and identifying the critical path. That linear programming could be used to solve project scheduling problems appears to have been first discovered by Charnes and Cooper (1962).

PERT was developed as a result of the Polaris Weapons System program undertaken by the Navy in 1958 (Department of the Navy, 1958). PERT was the result of a joint effort by Lockheed Aircraft Corporation, the Navy Special Projects Office, and the consulting firm of Booz, Allen, and Hamilton. Although the PERT system shared many features with the CPM system, the PERT project focused on activity time uncertainty rather than on project costs. An interesting issue related to the PERT approach is the justification for the approximation formulas for the mean and variance. Both Grubbs (1962) and Sasieni (1986) raised this issue. It appears that the formula for the variance is assumed (probably based on properties of the normal distribution) and the formula for the mean was obtained as a consequence of the assumption of the beta distribution and the formula assumed for the variance (see Archibald and Villoria, 1967, p. 449).

Moder, Phillips, and Davis (1983) discuss a little-known reference that precedes the development of PERT and CPM by almost 30 years and may have been the true genesis of project planning methodology. In 1931 a Polish scientist named Karol Adamiecki developed and published a planning device known as the Harmonygraph (Adamiecki, 1931). The Harmonygraph is basically a vertical Gantt chart modified to include immediate predecessors. The technique requires the use of sliding tabs for each activity. (See Moder, Phillips, and Davis, 1983, pp. 10–12, for a more comprehensive discussion of the Harmonygraph method.)

Research into project networks continues. One area of continuing interest is resource-constrained networks (Patterson, 1984). Another is networks with random

activity times. As we indicated in Section 10.5, the PERT methodology gives only an approximation of the distribution of the project completion time. Recent interest has focused on developing efficient simulations (Sullivan, Hayya, and Schaul, 1982) or determining the exact distribution of the project completion time assuming other than a beta distribution for the time required for each activity (Kulkarni and Adlakha, 1986).

10.9 PROJECT MANAGEMENT SOFTWARE FOR THE PC

As the installed base of personal computers increases, the demand for software also increases. Project management software was available shortly after mainframe computers were sold to business. Most of the major early computer manufacturers (IBM, Honeywell, Control Data, among others) marketed some type of project management software as early as the 1950s. With the spread of personal computers to the business community in the 1980s, software providers realized that there was a significant market for PC-based project management software. Project management tools, along with word processing and spreadsheets, became an important component of the software suite available for personal computers. While many open source and proprietary software products are available, Microsoft® Project continues to be a popular seller.

One of the newer developments for project management is Web-based software. In the period from 1998–2000 it appeared that the Web would take over the world. A company that mentioned a Web site in its promotional literature would find its stock price soaring on the basis of that announcement alone. It was estimated that in order for Amazon.com to generate enough revenue to justify its share price at the high in 2000, they would have to sell books to everyone in the world. The market capitalization of Cisco Corporation, a manufacturer of routers based in Northern California, exceeded that of companies ten times its size or more such as IBM, General Motors, and AT&T. In fact, at its high, Cisco had the largest market cap of *any* American company. Of course, just as the wild speculation in tulip bulbs in Holland in the 17th century came to an abrupt end, the wild speculation in Internet stocks suffered the same fate. Twenty-somethings all over the country became instant millionaires, as the stock prices of their start-ups soared, and lost it all just as quickly as these prices collapsed. In fact, the vast majority of Internet start-ups (including many reasonably successful firms like Webvan and Pets.com) went out of business.

But the dot-com bust of 2001 didn't mean the end of the growing influence of the Internet. The utilization of the Internet continued to increase, and many of the companies left standing when the dust cleared went on to be quite successful. More and more software applications that were formerly available only on a stand-alone basis were migrated to the Web. This is true of project management software as well. That a Web-based project management package can be used on a PDA is a boon to portability. The number of wireless hotspots is increasing daily, with additions such as university campuses, airports, and even Starbucks. This means that professionals have access to shared files anywhere in the world on devices weighing a few ounces and having battery lives measured in days rather than hours.

Many project management packages that are dedicated to specific applications. Several are available for helping legal departments manage their workloads. Both Web-based and standalone packages are available for construction and building applications. These are only two examples. The demand for both special-purpose and general-purpose project management software continues to grow.

Snapshot Applications

PROJECT MANAGEMENT HELPS UNITED STAY ON SCHEDULE

United Airlines has been a long-time user of project management methods and software to help keep its business on track. United owns 549 aircraft and makes 2,000 flights every day. For an operation of this magnitude to work properly takes both a commitment from its employees and quality project management tools. Richard Gleason, MIS project office manager at United, chose Project Workbench, a PC-based software package from Advanced Business Technology Corporation in New York City, for the task.

United uses this package to integrate more than 80 mainframe applications in diverse areas such as architecture development and technical support under one operational due date. In addition, the software assists United with managing its worldwide fleet maintenance program.

The software allows United to link dependencies among an unlimited number of distinct projects and calculate schedules for these projects. The software is more expensive than most of the general-purpose software aimed at the mass market (see the discussion of project management software in Section 10.9). However, United feels that the additional cost is more than offset by the flexibility in scheduling multiple projects with dependent activity links (Ouelette, 1994).

THOMAS BROTHERS PLANS STAFFING WITH PROJECT MANAGEMENT SOFTWARE

Thomas Brothers Maps, Inc., based in Irvine, California, is one of the country's premier makers of road maps. The firm supplies street guides and maps that cover many West Coast cities and counties. They have 230 workers and annual revenues exceeding $20 million. Bob Foster, the president of the firm, began using client–server project management software to plan a schedule for hiring and training cartographers. Foster chose PlanView, a Windows-based package, from the company with the same name based in Austin, Texas. The software handles planning for the 250 projects the company undertook in 1995. The project management package allows the president to track exactly how each worker is deployed.

One of the features of PlanView that attracted Foster and other personnel at Thomas Brothers was the ability of the software to track resources as well as schedules. PlanView looks at the labor pool in terms of resource overload rather than as a succession of tasks and potential bottlenecks. The firm can track its progress on a large number of multiple projects by maintaining a running schedule of the workload for each cartographer. According to Foster, "We've never been able to see so far ahead so clearly."

Another feature of PlanView is that the package uses a standard SQL database and can interface easily with other database products (Oracle 7 in this case). The project management information supplied by the system can be used to keep track of other business functions as well, such as tracking the cost of sales.

FLORIDA POWER AND LIGHT TAKES PROJECT MANAGEMENT SERIOUSLY

Florida Power and Light (FPL), a major utility company, is responsible for managing two nuclear power—generating facilities at Turkey Point and St. Lucie. To handle the management of these facilities as well as several other functions within the firm, FPL established an independent project management department. At its peak, the department had 17 project control personnel to support more than 600 engineers and analysts. Separate dedicated groups were established for each plant.

To deal effectively with contractors, the group established project control and execution reporting requirements for all major contracted work. This process required FPL to identify their major contractors and begin negotiations with those contractors to implement the system. It turned out that four contractors were responsible for 80 percent of the workload. Within a year, each of these contractors had implemented a satisfactory project control system.

To make the system more user friendly, FPL abandoned its traditional mainframe software and replaced it with Welcom's PC-based software. While the PC-based system was not as powerful as the mainframe package for handling resource modeling issues, the feeling was that local ownership of the process afforded by PCs would compensate for the new software's limitations.

Self-assessment of FPL's project management function showed that customers were very satisfied with the project management function and scheduling support afforded by the project management group. In fact, FPL was awarded the Deming Prize several years ago for its commitment to continuous improvement and quality of service. The success of its project management initiatives played an important role in this achievement (Cooprider, 1994).

10.10 Summary

When large projects consist of many interrelated activities that must be completed in a specific sequence, project management techniques provide useful tools for preparing and administering schedules for these projects. Project planning takes place at many levels of an organization, for projects lasting from a few days to months or even years.

This chapter focused on the critical path method (CPM) and its extensions. *Networks* are a convenient means of representing a project. There are two ways of using networks to represent projects: activity-on-arrow and activity-on-node. Using the activity-on-arrow method, the nodes of the network correspond to completion of some subset of activities. When nodes rather than arrows are used to represent activities, pseudo activities are not required. However, activity-on-arrow is far more common in practice and, in general, more intuitively appealing.

The *critical path* is the *longest* path or chain through the network. The length of the critical path is the minimum project completion time, and the activities that lie along the critical path are known as the critical activities. Delay in a critical activity delays the project. Noncritical activities have *slack,* which means that they can be delayed without necessarily delaying the project. Although for small networks the critical path can be identified readily by inspection, for larger networks an algorithmic procedure is required. This chapter presented a method, involving both forward and backward passes through the network, that specified the earliest and the latest starting and ending times for all activities.

One of the goals of the early development work on CPM was to consider the effect of *project costing*. We assume that costs are either direct or indirect. Direct costs include labor, material, and equipment; these costs increase if the project time decreases. Indirect costs include costs of rents, interest, and utilities; these costs increase if the project time increases. The goal of the analysis is to determine the optimal time to perform the project that minimizes the sum of indirect and direct costs.

Linear programming is one means of solving project scheduling problems. The linear programming formulations considered here solve both the CPM problem and the CPM problem with cost–time trade-offs. Although not treated in this chapter, linear programming also can be used to solve some cost–time problems when the direct costs are nonlinear functions of the activity duration times.

PERT is an extension of critical path analysis to incorporate uncertainty in the activity times. For each activity, three time estimates are required: (1) the minimum activity time (called a), (2) the maximum activity time (called b), and (3) the most likely activity time (called m). Based on these estimates, one approximates the mean and the standard deviation of the activity time. The project time is assumed to follow a normal distribution with mean equal to the sum of the means along the path with the longest expected completion time and the variance equal to the sum of the variances along this same path. Depending on the configuration of the network, assuming path independence of two or more paths could give more accurate results. Although the PERT approach is only approximate, it does provide a measure of the effect of uncertainty of activity times on the total project completion time.

Resource considerations also were considered. Traditional CPM and PERT methods ignore the fact that schedules may be infeasible because of insufficient resources. Typical examples of scarce resources that might give rise to an infeasible project schedule are workforce, raw materials, and equipment. When schedules are infeasible, noncritical activities should be rescheduled within the available slack if possible. If not, critical activities may have to be delayed and the project completion date moved ahead. Resource loading profiles are a useful tool for determining the requirements placed on resources by any schedule. These profiles can be used as rough planning guides as major projects evolve.

We also discuss *organizational design for effective project management*. From a project management perspective, the traditional line organization is the weakest organizational structure. On the other end of the spectrum is the project organization. Here, employees are freed up from their usual responsibilities for the life of the project. Some firms have experimented with matrix organizations, which is a hybrid of the two designs. Most companies, however, have retained the traditional line structure.

The chapter concluded with a brief overview of the software available for project management. The explosion of the personal computer has been accompanied by an explosion of software. Project management is no exception. There is an entire range of software products available. Most of the programs available are designed to run on a PC for a single user. Although many of the PC-based software can handle multiple projects and resources, very large systems require more powerful tools. There are programs available for mainframe computers and client–server systems that allow for multiple users, projects, and resources. These packages can interface with large databases and other parts of the firm's financial and production systems.

Additional Problems on Project Scheduling

26. Two brothers have purchased a small lot, in the center of town, where they intend to build a gas station. The station will have two pumps, a service area for water and tire maintenance, and a main building with restrooms, office, and cash register area. Before they begin excavating the site, the local authorities must approve the location for a gasoline station and be certain that the placement of the storage tanks will not interfere with water, gas, and electric lines that are already in place.

    Once the site has been approved, the excavation can begin. After excavation, the three primary parts of the construction can begin: laying in the gasoline tanks, building the water and tire service area (including installation of the air compressor), and constructing the main building. The surfacing can begin after all building is completed. After surfacing, the site must be cleaned and the station's signs erected. However, before the station can open for business, the air compressor must be inspected, tested, and approved.

    The activities and the time required for each of them are as follows:

Activity	Time Required (weeks)
A: Obtain site approval	4
B: Begin site excavation	2
C: Place and secure gasoline tanks	3
D: Install gasoline pumps	1
E: Connect and test gasoline pumps	1
F: Construct service area	2
G: Install and connect water and air compressor	3
H: Test compressor	1
I: Construct main building including the restrooms, the office, and the cash register area	5
J: Install plumbing and electrical connections in the main building	3
K: Cover tanks and surface the area	4
L: Clean site	2
M: Erect station signs	1

a. Based on the description of the project, determine the activity precedence relationships and develop a network for the project.

b. Determine the critical path and the earliest and the latest starting and finishing times for each activity.

c. Draw the Gantt chart for this project based on earliest times.

d. Suppose that the air compressor fails to function correctly and must be replaced. It takes two weeks to obtain another compressor and test it. Will the project necessarily be delayed as a result?

e. List the activities that must be completed by the end of the 15th week in order to guarantee that the project is not delayed.

f. Solve this problem using linear programming.

27. A scene is being shot for a film. A total of 11 distinct activities have been identified for the filming. First, the script must be verified for continuity, the set erected and decorated, and the makeup applied to the actors. After the set is completed, the lighting is set in place. After the makeup is applied, the actors get into costume. When these five activities are completed, the first rehearsal of the scene commences, which is followed by the second scene rehearsal with the cameraperson. While the rehearsals are going on, verifications of the audio and the video equipment are made. After both rehearsals and verifications are completed, the scene is shot. Afterward, it is viewed by the director to determine if it needs to be reshot.

    The list of the activities and the activity times is

Activity	Time Required (days)
A: Check script for story continuity	2.0
B: Decorate set; place necessary props	4.5
C: Check lighting of scene	1.0
D: Apply makeup to actors	0.5
E: Costumes for actors	1.5
F: First rehearsal (actors only)	2.5
G: Video verification	2.0
H: Sound verification	2.0
I: Second rehearsal (with camera and lights)	2.0
J: Shoot scene	3.5
K: Director's OK of scene	1.5

a. Develop a network for the filming of the scene.

b. Compute the earliest and the latest finishing and starting times for each of the activities, and identify the critical path.

c. Draw the Gantt chart for this project. Assume that activities with slack are scheduled so that there is equal slack before and after the activities.

d. Suppose that the video verification (G) shows that the equipment is faulty. Four additional days are required to obtain and test new equipment. How much of a delay in the total time required to film the project will result?

e. One of the costumes is damaged as it is being fitted (activity E). How much extra time is available for repair without delaying the project?

f. What kind of delays can you envision as a result of the uncertainty in the time of activity K?

g. Solve this problem by linear programming.

28. A guidance and detection system is being built as part of a large defense project. The detection portion consists of radar and sonar subsystems. Separate equipment is required for each of the subsystems. In each case, the equipment must be calibrated prior to production. After production, each subsystem is tested independently. The radar and the sonar are combined to form the detection system, which also must be tested prior to integration with the guidance system. The final test of the entire system requires complex equipment. The activities and the activity times are

Activity	Time Required (days)
A: Calibrate machine 1 (for radar)	2.0
B: Calibrate machine 2 (for sonar)	3.5
C: Calibrate machine 3 (for guidance)	1.5
D: Assemble and prepare final test gear	7.0
E: Make radar subsystem	4.5
F: Make sonar subsystem	5.0
G: Make guidance subsystem	4.5
H: Test radar subsystem	2.0
I: Test sonar subsystem	3.0
J: Test guidance subsystem	2.0
K: Assemble detection subsystem (radar and sonar)	1.5
L: Test detection subsystem	2.5
M: Final assembly of three systems	2.5
N: Testing of final assembly	3.5

a. Construct a network for this project.

b. Determine the earliest and the latest starting and finishing times for all the activities, and identify the critical path.

c. Draw a Gantt chart for this project based on the earliest times.

d. How much time is available for assembling and calibrating the final test gear without delaying the project?

e. What are the activities that must be completed by the end of 10 days to guarantee that the project is not delayed?

f. What are the activities that must be started by the end of 10 days to guarantee that the project is not delayed?

g. Solve this problem using linear programming.

29. Consider the filming of the scene described in Problem 27. Based on past experience, the director is not very confident about the time estimates for some of the activities. The director's estimates for the minimum, most likely, and maximum times for these activities are

Activity	*a*	*m*	*b*
F	2	3	12
I	1	2	8
J	3	4	10
K	1	2	7

a. Including these PERT time estimates, how long is the filming expected to take?

b. What is the probability that the number of days required to complete the filming of the scene is at least 30 percent larger than the answer you found in part (*a*)?

c. For how many days should the director plan in order to be 95 percent confident that the filming of the scene is completed?

30. Consider the following project time and cost data:

Activity	Immediate Predecessors	Normal Time	Expedited Time	Normal Cost	Expedited Cost
A	—	6	6	$ 200	$ 200
B	A	10	4	600	1,000
C	A	12	9	625	1,000
D	B	6	5	700	800
E	B	9	7	200	500
F	C, D	9	5	400	840
G	E	14	10	1,000	1,440
H	E, F	10	8	1,100	1,460

a. Develop a network for this project.

b. Compute the earliest and the latest starting and finishing times for each of the activities. Find the slack time for each activity and identify the critical path.

c. Suppose that the indirect costs of the project amount to $200 per day. Find the optimal number of days to perform the project by expediting one day at a time. What is the total project cost at the optimal solution? What savings have been realized by expediting the project?

d. Solve this problem using linear programming.

Appendix 10–A

Glossary of Notation for Chapter 10

a = Minimum activity time for PERT.

b = Maximum activity time for PERT.

EF_i = Earliest finishing time for activity i.

ES_i = Earliest starting time for activity i.

LF_i = Latest finishing time for activity i.

LS_i = Latest starting time for activity i.

m = Most likely activity time for PERT.

M_{ij} = Expedited time for activity (ij).

μ = Expected activity time estimate for PERT.

N_{ij} = Normal time for activity (ij).

σ = Estimate of the standard deviation of the activity time for PERT.

T = Project completion time for PERT; T is a random variable.

t_{ij} = Time required for activity (ij); t_{ij} is a constant in the standard formulation and a variable in the cost–time formulation.

x_i = Earliest start time for node i (linear programming formulation).

Bibliography

Adamiecki, Karol. "Harmonygraph." *Polish Journal of Organizational Review,* 1931. (In Polish.)

Archibald, R. D., and R. L. Villoria. *Network-Based Management Systems (PERT/CPM)*. New York: John Wiley & Sons. 1967.

Burgess, A. R., and J. B. Killebrew. "Variational Activity Level on a Cyclic Arrow Diagram." *Journal of Industrial Engineering* 13 (1962), pp. 76–83.

Charnes, A., and W. W. Cooper. "A Network Interpretation and a Directed Subdual Algorithm for Critical Path Scheduling." *Journal of Industrial Engineering* 13 (1962), pp. 213–19.

Cooprider, D. H. "Overview of Implementing a Project Control System in the Nuclear Utility Industry." *Cost Engineering* 36, no. 3 (March 1994), pp. 21–24.

Davis, E. W., and J. H. Patterson. "A Comparison of Heuristic and Optimum Solutions in Resource-Constrained Project Scheduling." *Management Science* 21 (1975), pp. 944–55.

Department of the Navy, Special Projects Office, Bureau of Ordnance. "PERT: Program Evaluation Research Task." Phase I Summary Report. Washington, D.C., July 1958.

Goodman, L. J. *Project Planning and Management*. New York: Van Nostrand Reinhold, 1988.

Grubbs, F. E. "Attempts to Validate Certain PERT Statistics, or 'Picking on Pert.'" *Operations Research* 10 (1962), pp. 912–15.

Kulkarni, V. G., and V. G. Adlakha. "Markov and Markov-Regenerative PERT Networks." *Naval Research Logistics Quarterly* 34 (1986), pp. 769–81.

Lockyer, K. G. *Critical Path Analysis, Problems and Solutions*. London: Isaac Pitman and Sons, Ltd., 1966.

Moder, J. J.; C. R. Phillips; and E. W. Davis. *Project Management with CPM, PERT, and Precedence Diagramming*. 3rd ed. New York: Van Nostrand Reinhold, 1983.

Ouelette, T. "Project Management Helps Airline Stick to Schedule." *Computerworld* 28, no. 45 (November 7, 1994), p. 79.

Patterson, J. H. "A Comparison of Exact Approaches for Solving the Multiple Constrained Resource Project Scheduling Problem." *Management Science* 30 (1984), pp. 854–67.

Sasieni, M. "A Note on PERT Times." *Management Science* 32 (1986), pp. 1652–53.

Sullivan, R. S.; J. C. Hayya; and R. Schaul. "Efficiency of the Antithetic Variate Method for Simulating Stochastic Networks." *Management Science* 28 (1982), pp. 563–72.

Vollmann, T. E.; W. L. Berry; and D. C. Whybark. *Manufacturing Planning and Control Systems*. 3rd ed. Homewood, IL: Richard D. Irwin, 1992.

Walker, M. R., and J. S. Sayer. "Project Planning and Scheduling." Report 6959. Wilmington, DE: E. I. du Pont de Nemours & Co., Inc., March 1959.

Wiest, J. D., and F. K. Levy. A *Management Guide to PERT/CPM*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1977.

Chapter **Eleven**

Facilities Layout and Location

"When you come to a fork in the road, take it."
—Yogi Berra

Chapter Overview

Purpose

To understand the major issues faced by a firm when designing and locating new facilities, and to learn the quantitative techniques for assisting with the decision-making process.

Key Points

1. *Fundamentals.* Before deciding on the appropriate layout for a new facility, whether it be a factory, hospital, theme park, or anything else, one must first study the patterns of flow. The simplest flow pattern is straight-line flow, as might be encountered on an assembly line. Other patterns include U flow, L flow, serpentine flow, circular flow, and S flow. Another issue is desirability or undesirability of locating operations near each other. For example, in a hospital, the emergency room must be near the hospital entrance, and the maternity ward should be close to the area where premature babies are cared for. A graphical technique for representing the relative desirability of locating two facilities near each other is the activity relationship chart (or rel chart). From-to charts give the distances between activities, which can be used to compute costs associated with various layouts.
2. *Types of layouts.* In the factory setting, the appropriate type of layout depends on the manufacturing environment and the characteristics of the product. A *fixed position layout* is appropriate when building large items such as planes or ships that are difficult and costly to move. Workstations are located around the object, which remains stationary. More typical is the *product layout* where machines or workstations are organized around the sequence of operations required to produce the product. Product layouts are most typical for mass production. In the case of small- to medium-sized companies, a *process layout* makes more sense. Here one groups similar machines or similar processes together. Finally, *layouts based on group technology* might be appropriate. In this case, machines might be grouped into machine cells where each cell corresponds to a part family or group of part families.

3. *Computerized layout techniques.* For large complex factories or service facilities, determining the best layout manually is impractical. There are several computerized layout techniques available to assist with this function. They include CRAFT, COFAD, ALDEP, CORELAP, and PLANET. All of these methods are intended for the factory setting and share the objective of minimizing materials handling costs. Both CRAFT and COFAD are based on the principle of improvement. This means that the user must specify an initial layout. From there, one considers pairwise interchanges of departments and chooses the one with the largest improvement.

   Both ALDEP, CORELAP, and PLANET are construction routines rather than improvement routines. Layouts are determined from scratch, and there is no requirement that the user specify an initial layout. There is some controversy regarding whether human planners or computer programs produce better layouts. In one study where groups of 20 chosen from 74 people trained in layout techniques were compared with computerized layouts, the humans fared much better. Others criticized this study on the grounds that most layout departments are not that well staffed.

4. *Flexible manufacturing systems.* A flexible manufacturing system (FMS) is a collection of numerically controlled machines connected by a computer-controlled materials flow system. Typical flexible manufacturing systems are used for metal cutting and forming operations and certain assembly operations. Because the machines can be programmed, the same system can be used to produce a variety of different parts. Flexible manufacturing systems tend to be extremely expensive (some costing upwards of $10 million). As a result, the added flexibility may not be worth the cost. While the FMS can have many advantages (reducing work-in-process inventory, increased machine utilization, flexibility), these advantages are rarely justified by the high cost of these systems. An alternative that is more popular is flexible manufacturing cells. These are smaller than full-blown systems, but still provide more flexibility than single-function equipment.

5. *Locating new facilities.* Where to locate a new facility is a complex and strategically important problem. Hospitals need to be close to high-density population centers, and airports need to be near large cities, but not too near because of noise pollution. New factories are often located outside the United States to take advantage of the lower labor costs overseas. But these savings might come at a high price. Political instability, unfavorable exchange rates, infrastructure deficiencies, and long lead times are a few of the problems that arise from locating facilities abroad. Often such decisions are more strategic than tactical and require careful weighing of the advantages and disadvantages at the level of top management.

   However, in cases where the primary objective is to locate a facility to be closest to its customer base, quantitative methods can be very useful. In these cases, one must specify how distance is measured. Straight-line distance (also known as Euclidean distance) measures the shortest distance between two points. However, straight-line distance is not always the most appropriate measure. For example, when locating a firehouse, one must take into account the layout of streets. Using rectilinear distance (as measured by only horizontal and vertical movements) would make more sense in this context. Another consideration is

that not all customers are of equal size. For example, a bakery would make larger deliveries to a supermarket or warehouse store than to a convenience store. Here one would use a weighted distance criterion. In the remainder of this chapter, we review several quantitative techniques for finding the best location of a single facility under various objectives.

Where to locate facilities and the efficient design of those facilities are important strategic issues for business as well as the military, nonprofit institutions, and government. During the 1980s, Northern California's Silicon Valley experienced tremendous growth in microelectronics and related industries. One can get some idea of how significant this growth was from the American Electronics Association Member Directory: more than 50 percent of the listings are in the Bay Area. Most of these firms were "start-ups," adapting a new technology to a specialized segment of the marketplace. With the surge in demand for microcomputers came support industries producing hard disk drives, floppy disks, semiconductor manufacturing equipment, local area networks, and a host of other related products. In many cases, large investments in capital equipment were required before the first unit could be sold. A typical example is the Read-Rite Corporation of Milpitas, California, the largest independent producer of thin film heads for reading from and writing to Winchester hard drives in the world. When the firm was founded in 1983, an initial capital investment of $40 million was required. The venture capital investors bore the risk that the firm would survive. Read-Rite is typical of hundreds of high-tech start-ups.

What equipment should be purchased, how facilities should be organized, and where the facilities should be located are fundamental strategy issues facing any manufacturing organization. Service industries also are faced with the problem of finding effective layouts. Achieving efficient patient flows in hospitals is one example. Another example is the obvious care used in laying out the many theme parks across the United States. Anyone who has visited the San Diego Zoo or Disney World in Orlando, Florida, can appreciate the importance of effective layout and efficient management of people in service industries.

Tompkins and White (1984) estimated that 8 percent of the U.S. gross national product has been spent on new facilities annually since 1955. This does not include the cost of modification of existing facilities. If the authors' estimates are correct, we are spending in excess of $500 billion annually on construction and modification of facilities. The authors go on to claim that from 20 to 50 percent of the total operating expenses in manufacturing are attributed to materials handling costs. Effective facilities planning could reduce these costs by 10 to 30 percent annually, the authors claim.

An important part of the enormous success of Japanese companies in achieving manufacturing dominance in several key industries is efficient production. Efficient production includes efficient design of the product, employee involvement, lean inventory material management systems, and intelligent layout and organization of facilities.

Chapter 1 touched on some of the qualitative issues that management must consider when deciding on the location of new facilities. This chapter treats in greater depth the issues associated with locating new facilities and the methods for determining the best layout of operations in those facilities. In addition to exploring the qualitative factors that should be taken into account, we will also consider how a

manager can employ computers and quantitative techniques to assist with these complex decisions.

How a plant or workplace should be laid out is, in a sense, a special version of the location problem. Determining a suitable layout means finding the locations of departments within some specified boundary. When designing new facilities, the planner also must decide on the size and shape of the facility as well as the configuration of the departments within it.

Quantitative techniques are most useful when the goal is to minimize or maximize a single dimensional objective such as cost or profit. The objective function used in location problems generally involves either Euclidean or rectilinear distance (these terms will be defined later in Section 11.8). However, minimizing total distance traveled may not make sense in all cases. As an extreme example, consider the problem of locating a school. A location that requires 100 students to travel 10 miles each is clearly not equally desirable to one that requires 99 students to travel 1 mile and one student to travel 901 miles.

For layout problems, the most common objective used in mathematical models is to minimize the cost of materials handling. Furthermore, such models invariably assume that the number of materials handling trips from every work center to every other work center is known with certainty. In most real environments such assumptions are naive at best. Another deficiency of mathematical models is that they ignore such factors as plant safety, flexibility of the layout for future design changes, noise, and aesthetics.

This does not imply that mathematical and computer models are not useful for solving layout and location problems. What it does mean is that quantitative solutions must not be taken on blind faith. They must be carefully considered in the context of the problem. Used properly, the results of mathematical and computer models can reduce significantly the number of alternatives that the analyst must consider.

11.1 THE FACILITIES LAYOUT PROBLEM

Determining the best layout for a facility is a classical industrial engineering problem. Early industrial engineers often were known as efficiency experts and were interested in determining layouts to optimize some measure of production efficiency. For the most part, this view continues today, especially in plant layout. However, layout problems occur in many environments outside the plant. Some of these include

1. Hospitals
2. Warehouses
3. Schools
4. Offices
5. Workstations
6. Banks
7. Shopping centers
8. Airports
9. Industrial plants

Each of these layout problems has unique characteristics. Our focus in this chapter will be on techniques for finding layouts of industrial plants, although many of the methods can be applied to other facilities as well. The objectives in a plant layout study might include one or more of the following:

1. Minimize the investment required in new equipment.
2. Minimize the time required for production.
3. Utilize existing space most efficiently.
4. Provide for the convenience, safety, and comfort of the employees.
5. Maintain a flexible arrangement.
6. Minimize the materials handling cost.
7. Facilitate the manufacturing process.
8. Facilitate the organizational structure.

Vollmann and Buffa (1966) suggest nine steps as a guide for the analysis of layout problems:

1. Determine the compatibility of the materials handling layout models with the problem under study. Find all factors that can be modeled as materials flow.
2. Determine the basic subunits for analysis. Determine the appropriate definition of a department or subunit.
3. If a mathematical or computer model is to be used, determine the compatibility of the nature of costs in the problem and in the model. That is, if the model assumes that materials handling costs are linear and incremental (as most do), determine whether or not these assumptions are realistic.
4. How sensitive is the solution to the flow data assumptions? What is the impact of random changes in these data?
5. Recognize model idiosyncrasies and attempt to find improvements.
6. Examine the long-run issues associated with the problem and the long-run implications of the proposed solution.
7. Consider the layout problem as a systems problem.
8. Weigh the importance of qualitative factors.
9. Select the appropriate tools for analysis.

11.2 PATTERNS OF FLOW

As we noted in Section 11.1, the objective used most frequently for quantitative analysis of layout problems is to minimize the materials handling cost. When minimizing the materials handling cost is the primary objective, a flow analysis of the facility is necessary. Flow patterns can be classified as horizontal or vertical. A horizontal flow pattern is appropriate when all operations are located on the same floor, and a vertical pattern is appropriate when operations are in multistory structures. Francis and White (1974) give six horizontal flow patterns and six vertical flow patterns. The six horizontal patterns appear in Figure 11–1.

The simplest pattern is (a), which is straight-line flow. The main disadvantage of this pattern is that separate docks and personnel are required for receiving and

shipping goods. The L shape is used to replace straight-line flow when the configuration of the building or the line requires it. The U shape has the advantage over the straight-line configuration of allowing shipping and receiving to be at the same location. The circular pattern is similar to the U shape. The remaining two patterns are used when the space required for production operations is too great to allow use of the other patterns.

Two charts that supply useful information regarding flows are (1) the activity relationship chart and (2) the from-to chart.

Activity Relationship Chart

An activity relationship chart (also called a rel chart for short) is a graphical means of representing the desirability of locating pairs of operations near each other. The following letter codes have been suggested for determining a "closeness" rating.

A	Absolutely necessary. Because two operations may use the same equipment or facilities, they must be located near each other.
E	Especially important. The facilities may require the same personnel or records, for example.
I	Important. The activities may be located in sequence in the normal work flow.
O	Ordinary importance. It would be convenient to have the facilities near each other, but it is not essential.
U	Unimportant. It does not matter whether the facilities are located near each other or not.
X	Undesirable. Locating a welding department near one that uses flammable liquids would be an example of this category.

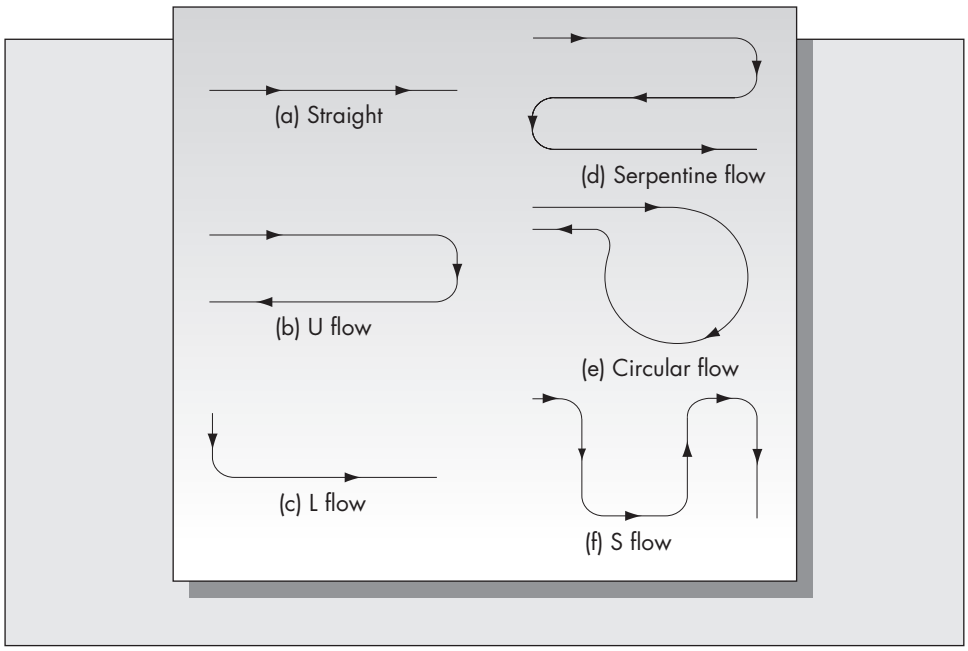


FIGURE 11–1
Six horizontal flow patterns

The closeness ratings are represented in an activity relationship chart that specifies the appropriate rating for each pair of departments. Consider the following example.

Example *11.1*

Meat Me, Inc., is a franchised chain of fast-food hamburger restaurants. A new restaurant is being located in a growing suburban community near Reston, Virginia. Each restaurant has the following departments:

1. Cooking burgers.
2. Cooking fries.
3. Packing and storing burgers.
4. Drink dispensers.
5. Counter servers.
6. Drive-up server.

The burgers are cooked on a large grill, and the fries are deep fried in hot oil. For safety reasons the company requires that these cooking areas not be located near each other. All hamburgers are individually wrapped after cooking and stored near the counter. The service counter can accommodate six servers, and the site has an area reserved for a drive-up window.

An activity relationship chart for this facility appears in Figure 11–2. In the chart, each pair of activities is given one of the letter designations A, E, I, O, U, or X. Once a final layout is determined, the proximity of the various departments can be compared to the closeness ratings in the chart. Note that Figure 11–2 gives only the closeness rating for each pair of departments. In the original conception of the chart, a number giving the reason for each closeness rating is also included in every cell. These numbers do not appear in our example.

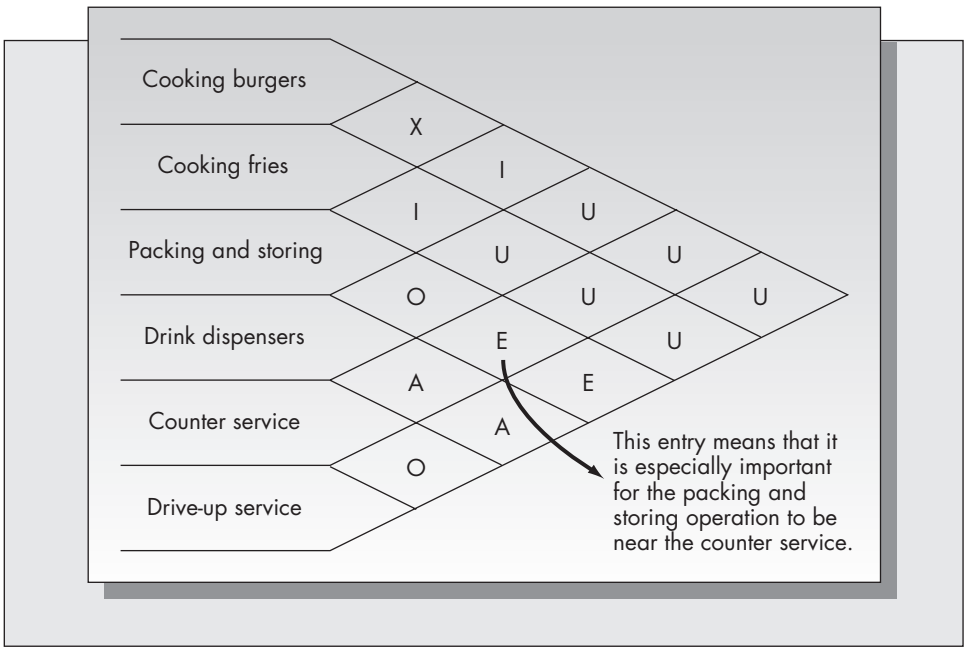


FIGURE 11–2
Activity relationship chart for Meat Me fast-food restaurant

From-To Chart

A from-to chart is similar to the mileage chart that appears at the bottom of many road maps and gives the mileage between selected pairs of cities. From-to charts are used to analyze the flow of materials between departments. The two most common forms are charts that show the distances between departments and charts that show the number of materials handling trips per day between departments. A from-to chart differs from an activity relationship chart in that the from-to chart is based on a specific layout. It is a convenient means of summarizing the flow data corresponding to a given layout.

Example *11.2*

A machine shop has six work centers. The work centers contain one or more of the following types of machines:

1. Saws
2. Milling machines
3. Punch presses
4. Drills
5. Lathes
6. Sanders

The from-to chart in Figure 11–3 shows the distance in feet between centers of the six departments. Note that the chart in the example is symmetric; that is, the travel distance between A and B is the same as the travel distance between B and A. This is not necessarily always the case, however. There may be one-way lanes, which allow material flow in one direction only, or an automated materials handling system that moves pallets in one direction only.

Figure 11–4 shows a from-to chart that gives the number of materials handling trips per day. These figures could be based on a specific product mix produced in the shop or simply be representative of an average day.

FIGURE 11–3
From-to chart showing distances between six department centers (measured in feet)

From \ To	Saws	Milling	Punch press	Drills	Lathes	Sanders
Saws		18	40	30	65	24
Milling	18		38	75	16	30
Punch press	40	38		22	38	12
Drills	30	75	22		50	46
Lathes	65	16	38	50		60
Sanders	24	30	12	46	60	