Lecture Notes for

Production System Design

Fall 2016

Michael G. Kay
Fitts Department of Industrial and Systems Engineering
North Carolina State University
Raleigh, NC 27695-7906

These notes are provided for the use of the students in ISE 453: Design of Production, Logistics, and Service Systems at North Carolina State University.

Contents

1. Facility Location

1.1 The Supply Chain

Figure 1.1 shows a typical supply chain for a single plant. In the figure, items D and E provided by two different tier-two suppliers are shipped to a tier-one supplier where they are used to produce item B, units of which are then shipped to the plant. Similarly, items F and G are used to produce item C, which, along with B, are the two raw materials used to produce the finished good A. Units of A are shipped to distribution centers (DCs), from which they are delivered to the final customer. The number of units shown along each *lane* (or arc) in the figure is meant to indicate the typical size of each shipment; e.g., one of the reasons that a DC might be used is that it is cheaper to transport larger loads from the plant to the DC and then transport small loads to each customer as needed.

1.1.1 Logistics Network Design

The design of a logistics network involves determining how to supply products to customers at the least cost while providing the desired level of service. The level of service can include both the time required for delivery to the customer and the availability of the product. Given the location of available transshipment points (e.g., ports and DCs), the design problem includes issues like selecting the best mode of transportation between each point in the network, the frequency and quantity of each shipment, and the amount of each product to be stored at each DC. The total cost of transportation and inventory is used to guide the design, subject to service level requirements. Inventory costs are the sum of cycle, in-transit, and safety stock inventory. Generally, there is a trade-off between transportation and inventory costs; e.g., transportation costs are lower for truckload shipments as compared to less-than-truckload, but they can increase cycle inventory costs while product waits for a full truckload to accumulate. In some cases, the design problem includes the need to select the best location for some of the DCs in the network. When location decisions are part of the design, it may be necessary to include the cost of constructing and operating each DC as part of the total cost along with the cost of transportation and inventory.
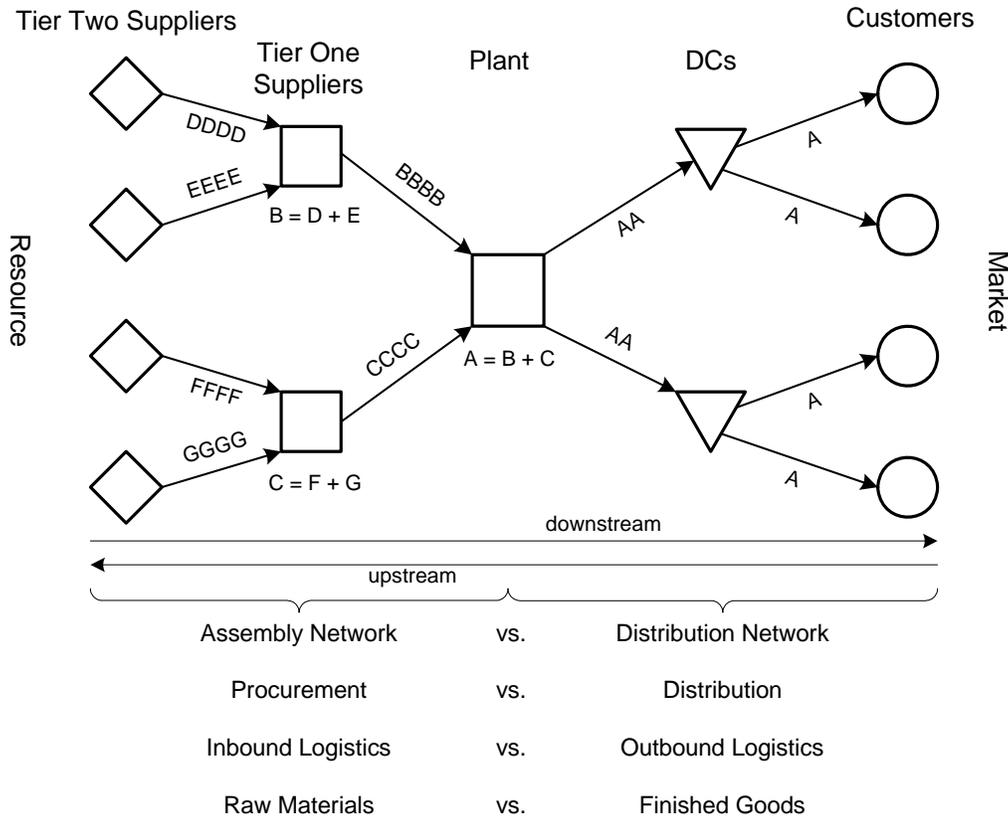
Figure 1.1. Typical logistics network for a plant.

As an example of a network design problem in the furniture industry, assume that a national furniture manufacturer has a large import program from suppliers in China. Currently, all products are shipped through the Panama Canal to a port on the East Coast and then trucked to a single DC. From there product is shipped to retail customers throughout the continental U.S. The manufacturer is considering whether a second DC should be opened at a predetermined location in California and used to receive product through the port of Long Beach. The second DC will decrease transportation costs from the DCs to the customers, but the sum of the inventory held both DCs will need to be greater because of loss of the opportunity at the single DC to pool safety stock inventory. Also, transport costs from China to Long Beach will be lower than the costs to the East Coast, but it may require than full container loads of low-volume products be shipped to each DC as compared to single container loads to the single DC, thereby doubling the cycle inventory levels of these products. Given all these factors, a network design decision can be made by determining the change in total logistics costs (transportation plus inventory costs) associated with opening the second DC.
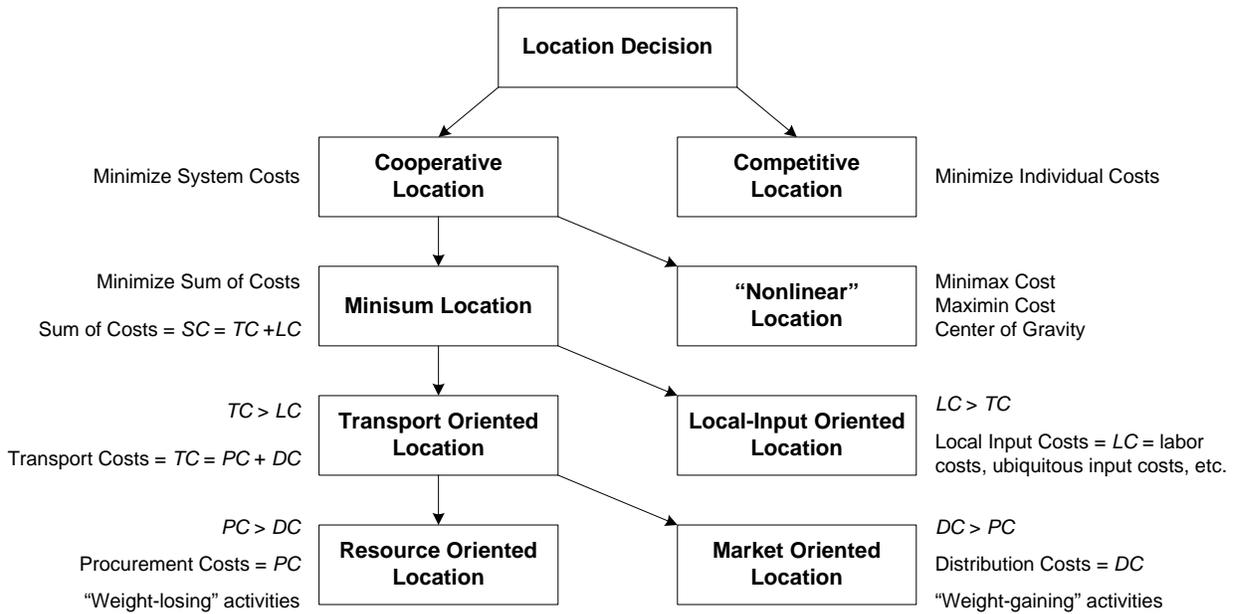
Figure 1.2. Types of location problems.

1.1.2 Location Problems

A taxonomy of the different types of location problems is shown in Figure 1.2. Cooperative location decisions are used to minimize the total system costs of multiple facilities owned by a single firm instead just minimizing the cost of each of the firm's individual facilities. These decisions are possible when the impact of the location of other firms' facilities does not significantly impact the location of your facilities. Cooperative location problems that focus on optimizing something other than the sum of costs (e.g., minimizing the maximum cost) can be considered as "nonlinear" location problems because the location-related costs in these problems are not directly proportional to distance, as is the case in minisum problems. Competitive location decisions are used to minimize the cost of an individual facility with respect to other facilities owned by other firms. These decisions are required when the location of other firms' facilities does impact the location of your facilities, and may result in sub-optimal decisions as compared to cooperative location decisions (cf. Hotelling's law). In what follows, only transport-oriented minisum location problems are considered because these problems are the ones that most benefit from a simple analysis using transport-cost minimization as the sole criterion, and the assumption that costs are directly proportional to distance is usually reasonable; local-input-oriented location problem are typically solved using more complex multi-criteria-based approaches.

1.1.3 Basic Production System

As shown in Figure 1.3, a production system can be considered as a node (or facility) in a logistics network that converts raw materials procured from suppliers into finished goods that are distributed to customers. For most production systems, the material input to the system equals

the material output from the system. *Raw materials* are those inputs that are transported to the production facility; *ubiquitous inputs* (e.g., water) are those available at any location, so that they do not need to be transported. *Finished products* are those outputs transported from the facility; while *scrap* is the output that is disposed of locally (although some outputs termed "scrap" are sometimes transported long distances from the facility for disposal or rework).
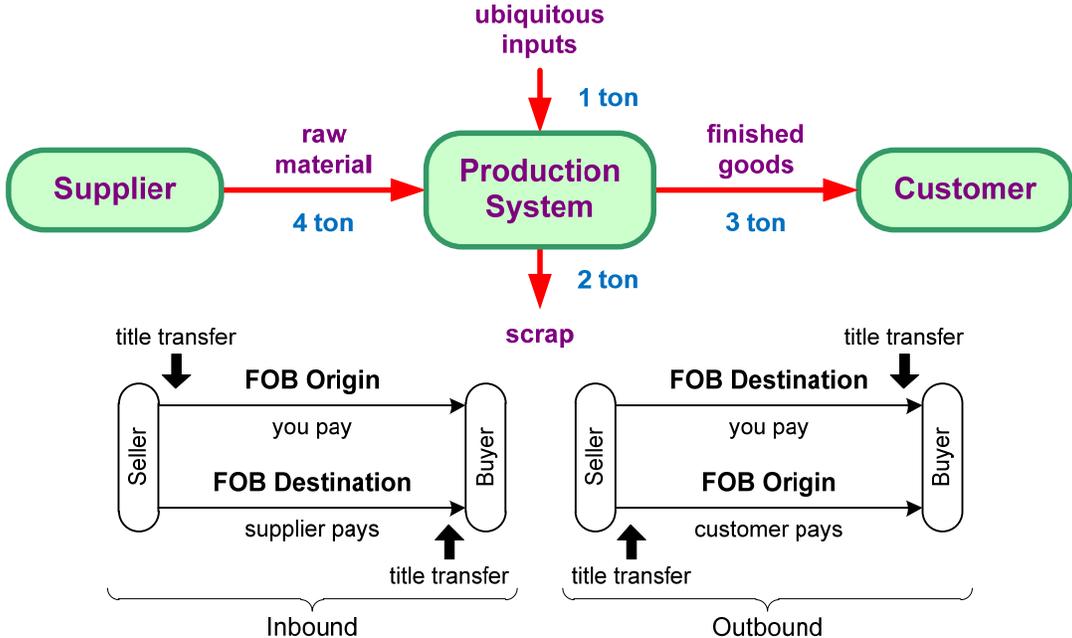
Figure 1.3. Basic production system.

The bottom portion of Figure 1.3 illustrates two different shipping terms that describe when the transfer of title occurs when goods are transported from the seller to the buyer: *FOB Origin* and *FOB Destination*, where FOB stands for free on board.[1] In most cases, the cost of transporting the goods is paid for by whoever is the owner of the goods during the transport. Referring to Figure 1.3, assuming that you represent the production system, the supplier (the seller) would pay for the transport of goods from the supplier's location (the origin) to your (the buyer's) facility (the destination) if the shipping terms were FOB Destination.

10

1.2 Single-Facility Minisum Location

Assuming that local input costs are either the same at every location or are insignificant as compared to transport costs, the minisum transport-oriented single-facility location problem is to locate a new facility (NF) to minimize the sum of weighted distances between NF and m existing facilities EFi, $i = 1, \ldots, m$:

$$\text{Min } TC(X) = \sum_{i=1}^{m} w_i\, d(X, P_i) = \sum_{i=1}^{m} q_i\, \overbrace{\underbrace{r_i\, d(X, P_i)}_{w_i}}^{c_i} \tag{1.1}$$

where

w_i = monetary weight (\$/mile)

q_i = physical weight (tons) or f_i = physical weight rate (tons/year)

r_i = transport rate (\$/ton-mile)

$d(X, P_i)$ = distance between NF at X and EFi at P_i (miles)

c_i = unit cost (\$/ton), used to determine q_i after NF located (transportation problem)

X = location of new facility (NF)

P_i = location of existing facility i (EFi)

m = number of EFs

If physical weight q_i is used, then TC in (1.1) is in units of \$; if, instead, the physical weight rate f_i in units of tons per year is used, then TC is in units of \$/year.

1.2.1 Majority Theorem

The Majority Theorem can be used to determine if one of the EFs has at least half of the total weight (i.e., a majority); if so, then the NF should be located at that EF in order to minimize TC:

$$\textit{Majority Theorem:}\quad \text{Locate NF at EF}j \text{ if } w_j \geq \frac{W}{2}, \quad \text{where } W = \sum_{i=1}^{m} w_i \tag{1.2}$$

The theorem is true for all minisum problems with metric distances, and can used to as a first check before using other means of determining the optimal location for the NF.

1.2.2 Weight-Gaining vs. Weight-Losing Activities

The activity occurring at a NF is considered *weight losing* if the sum of the monetary weights from EFs supplying material to the NF exceeds the sum of the weights of the material sent from the NF to the EFs which it supplies; conversely, if the sum of the monetary weight into the NF is less than the sum of the weight out, then the activity is considered *weight gaining*. In situations

11

where the NF is a distribution center (DC), it is common for the products to be weight gaining because, while the physical weight of the products into the DC is the same as the weight of the products out, more costly modes of transport are used for distribution as compared to procurement, resulting in higher monetary weights on the outbound side of the DC (and thus drawing the location of the DC to the market).

1.2.3 Median Location

Starting from the first EF, a median location is the first EF location at which the cumulative weight of the EFs up to that point is at least half of the total weight of all EFs (i.e., the EF location that splits the total weight into equal halves). The median location is the optimal NF location for all 1-D minisum problems and any 2-D rectilinear distance location problem.

The following procedure can be used to determine the median location:

Median location: For each dimension x of X:

1. Order EFs so that $|x_1| \leq |x_2| \leq \cdots \leq |x_m|$

2. Locate x-dimension of NF at the first EFj where $\sum_{i=1}^{j} w_i \geq \dfrac{W}{2}$,   where $W = \sum_{i=1}^{m} w_i$

If the cumulative weight at EFj exactly equals half of the total weight, then the optimal location along the x-dimension for the NF is any point between and including EFj and EF$(j+1)$. Note that the optimal location for the NF can be determined without knowing the actual distances between the EFs; all that is necessary is to be able to order the EFs along each dimension.

For 2-D minisum problems other than rectilinear distance problems, then an iterative procedure must be used to optimize TC and thus determine the optimal location for a NF.
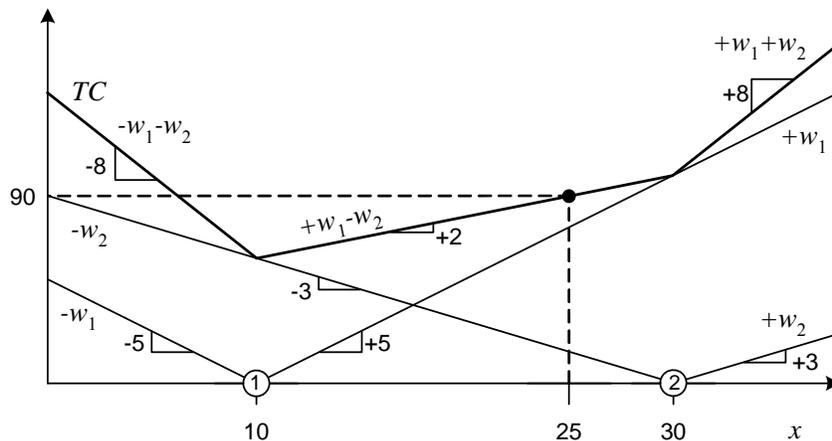
Figure 1.4. Total cost curve for 2 EFs.

12
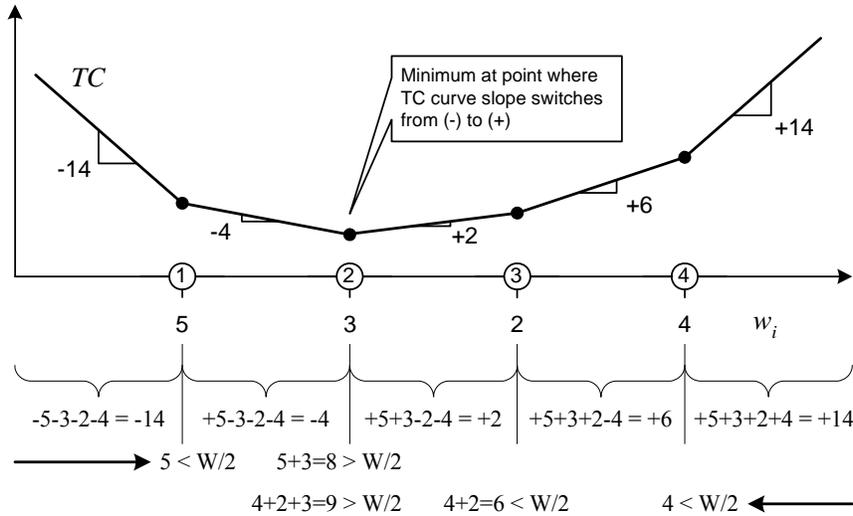
Figure 1.5. Total cost curve for 4 EFs.

Derivation

Figure 1.4 shows an example with two EFs. EF1 is located at $x_1 = 10$ and has a weight of $w_1 = 5$, while EF2 is located at $x_2 = 30$ and has a weight of $w_2 = 3$. The total cost of locating a single NF at x is the sum of the costs for each EF:

$$TC(x) = \beta_1(x - x_1) + \beta_2(x - x_2), \quad \text{where } \beta_i = \begin{cases} w_i, & \text{if } x \geq x_i \\ -w_i, & \text{if } x < x_i \end{cases}$$

For NF at $x = 25$,

$$TC(25) = w_1(25 - 10) + (-w_2)(25 - 30)$$

$$= 5(15) + (-3)(-5) = 90$$

Figure 1.5 shows an example with four EFs, where only the total cost curve is shown. The slope of curve from $-\infty$ to x_1 is

$$\beta_{-\infty,1} = -\sum_{i=1}^{m} w_i = -W$$

and the slope between x_j and x_{j+1} is

$$\beta_{j,j+1} = \sum_{i=1}^{j} w_i - \sum_{i=j+1}^{m} w_i$$

Starting at x_1, the minimum total cost corresponds to the point where the slope of the total cost curve switches from negative to positive (or zero); this is equivalent to finding the first j such that

$$\sum_{i=1}^{j} w_i - \sum_{i=j+1}^{m} w_i \geq 0$$

$$\sum_{i=1}^{j} w_i \geq \sum_{i=j+1}^{m} w_i$$

$$\sum_{i=1}^{j} w_i + \sum_{i=1}^{j} w_i \geq \sum_{i=1}^{j} w_i + \sum_{i=j+1}^{m} w_i$$

$$2\sum_{i=1}^{j} w_i \geq W$$

$$\sum_{i=1}^{j} w_i \geq \frac{W}{2}$$

In Figure 1.5, $W = 14$ and $W/2 = 7$. Starting from either EF1 or EF4, the optimal solution is found at EF2, which corresponds to the location at which the median condition is first satisfied.

1-D Example

Problem: As shown in Figure 1.6, I-40 passes through Asheville, Statesville, Winston-Salem, Greensboro, Durham, Raleigh, and Wilmington. The number of road miles from the beginning of I-40 at the western border of North Carolina to each city is shown below its name. A company wants to build a facility along I-40 to serve customers located in these cities. If the weekly demand in truckloads of the customers in each city is 6, 4, 3, 2, 1, 3, and 5, respectively, determine where the facility should be located to minimize the distance traveled to serve the customers assuming that I-40 will be used for all travel.

Solution: Since $W = 24$, the cumulative weights at the optimal location should equal or exceed $W/2 = 12$, which occurs at Winston-Salem. Note that (1) the same solution is found starting from either Asheville or Wilmington, and (2) the distance between cities is not used to determine the optimal location, just the relative ordering the cities along I-40.
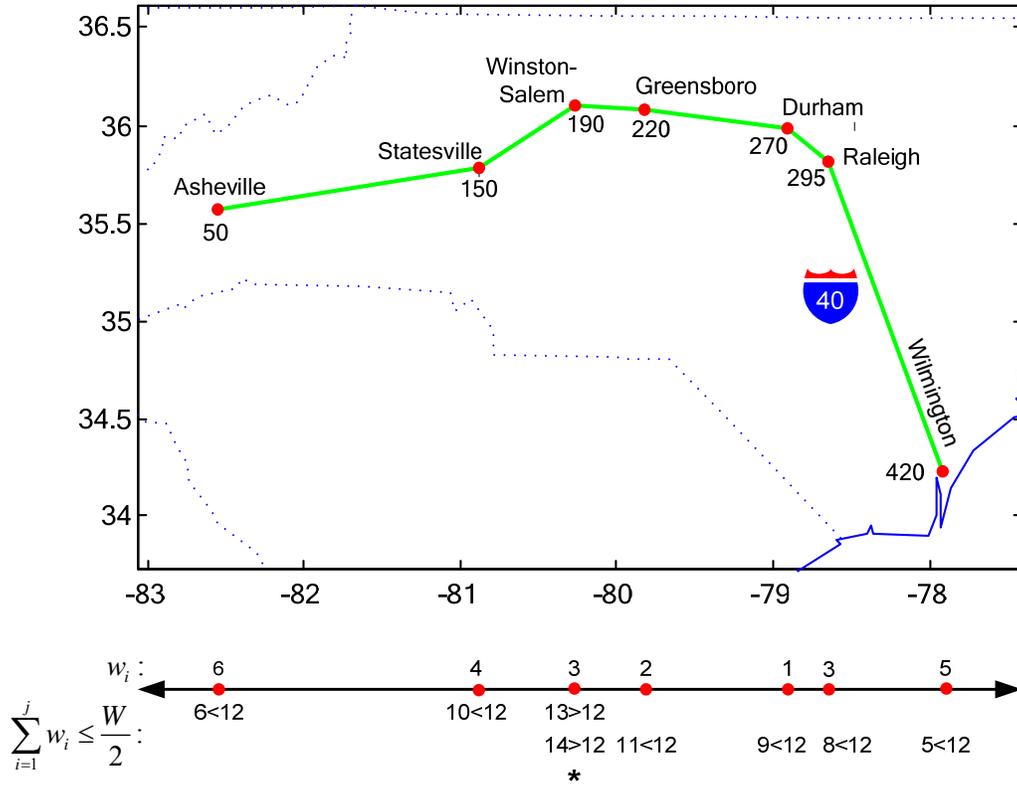
Figure 1.6. 1-D minisum location example.

2-D Example

Problem: A new snack machine is to be located on the floor of a facility (see Figure 1.7). Workers from eight different departments will make 19, 53, 82, 42, 9, 8, 39, and 6 trips per shift to the machine. Assuming rectilinear distance is a reasonable approximation of the actual travel distance, what is the location for the snack machine that will minimize the total distance that the workers have to travel?

Solution: A separate 1-D location problem can be solved for each dimension. Since $W = 258$, the cumulative weights at the optimal location should equal or exceed $W/2 = 129$. Along the bottom (x) dimension, the optimal location is at same x-location as departments 4 and 6, while along the left (y) dimension, the optimal location is anywhere between y-locations of departments 4 and 6. The optimal y-location is not a single point because the cumulative weight exactly equals 129, which corresponds to the total cost curve being flat between points 4 and 6. Note that the weights of departments at the same location are added together along each dimension.
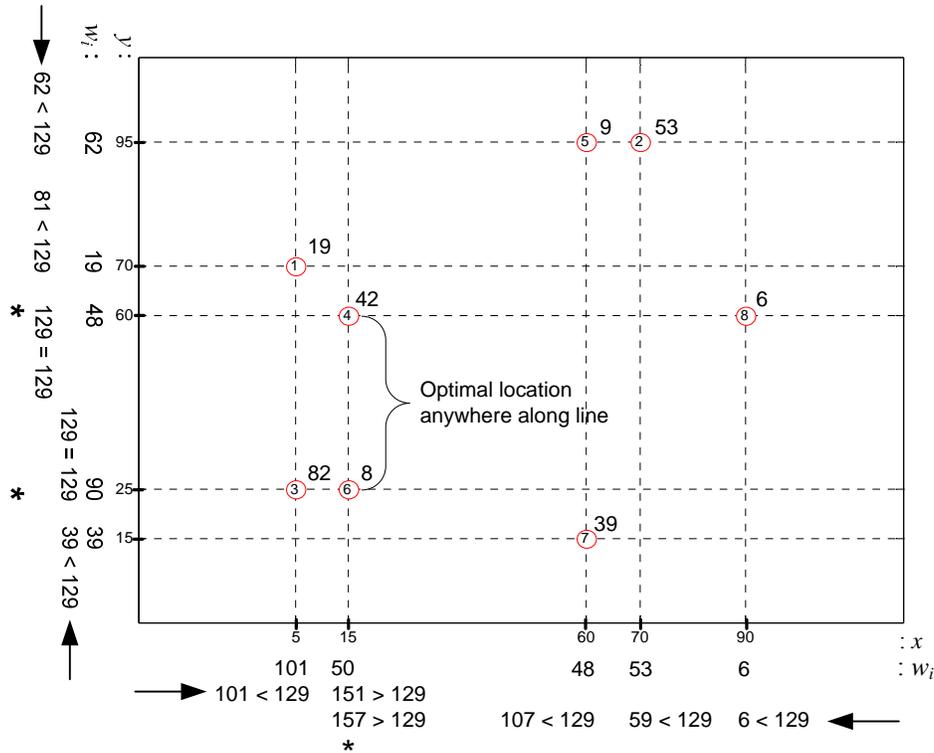
Figure 1.7. 2-D rectilinear-distance minisum location example.

1-D Example with Procurement and Distribution Costs

Problem: A product will be produced at a single plant that will be located along I-40 to serve customers located in the cities shown in Figure 1.6, above. Two tons of raw materials from a supplier in Asheville and a half ton of a raw material from a supplier in Durham are used to produce each ton of finished product that is shipped to customers in Statesville, Winston-Salem, and Wilmington. The demand of these customers is 10, 20, and 30 tons per year, respectively, and it costs $0.33 per ton-mile to ship raw materials to the plant and $1.00 per ton-mile to ship finished goods from the plant to the customers. Determine where the plant should be located so that procurement and distribution costs (i.e., the transportation costs to and from the plant) are minimized.

Solution: As shown in Figure 1.8, total customer demand is 60 tons per year, which translates into a demand of 120 tons per year based on a bill-of material (BOM) ratio of 2 for the supplier in Asheville and 30 tons per year based on a BOM of 0.5 for the supplier in Durham. The resulting monetary weights are used to determine that the plant should be located in Winston-Salem. The plant is (monetary) weight gaining since $\Sigma w_{\text{in}} = 50 < \Sigma w_{\text{out}} = 60$, but the plant is physically weight losing since $\Sigma f_{\text{in}} = 150 > \Sigma f_{\text{out}} = 60$; the difference is due the higher outbound transport rate.

$$r_{\text{in}} = \$0.33/\text{ton-mi} \qquad\qquad r_{\text{out}} = \$1.00/\text{ton-mi}$$

$$f_4 = BOM_4 \sum_{i=1}^{3} f_i = 2(60) = 120, \quad w_4 = f_4 r_{\text{in}} = 40 \quad \text{(4)}$$

$$\text{(1)} \quad f_1 = 10, \quad w_1 = f_1 r_{\text{out}} = 10$$

$$\text{NF} \rightarrow \text{(2)} \quad f_2 = 20, \quad w_2 = f_2 r_{\text{out}} = 20$$

$$f_5 = BOM_5 \sum_{i=1}^{3} f_i = 0.5(60) = 30, \quad w_5 = f_5 r_{\text{in}} = 10 \quad \text{(5)}$$

$$\text{(3)} \quad f_3 = 30, \quad w_3 = f_3 r_{\text{out}} = 30$$

	Asheville		Statesville	Winston-Salem		Durham	Wilmington
w_i :	40		10	20		10	30
	40<55		50<55	70>55			
$\sum_{i=1}^{j} w_i \le \dfrac{W}{2}$:				60>55		40<55	30<55
				*			
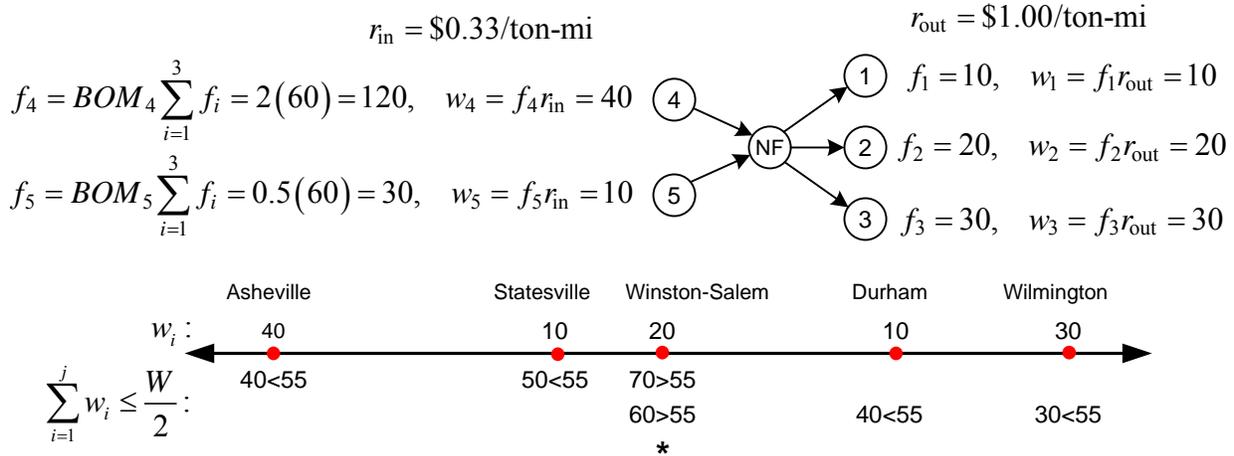
Figure 1.8. 1-D example with procurement and distribution costs.

1.3 Metric Distances

1.3.1 l_p Distances

Given two points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$, a general metric distance function termed the l_p distance or norm can be used to represent several of the most common distances used in practice. The function uses different values of the parameter p to represent these special cases:

$$\text{General } l_p: \quad d_p(P_1, P_2) = \left[|x_1 - x_2|^p + |y_1 - y_2|^p \right]^{\frac{1}{p}}, \quad p \ge 1 \tag{1.3}$$

$$\text{Rectilinear}: \quad d_1(P_1, P_2) = |x_1 - x_2| + |y_1 - y_2| \tag{1.4}$$
$$\text{\small}(p=1)$$

$$\text{Euclidean}: \quad d_2(P_1, P_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1.5}$$
$$\text{\small}(p=2)$$

$$\text{Chebychev}: \quad d_\infty(P_1, P_2) = \max\left\{ |x_1 - x_2|, |y_1 - y_2| \right\} \tag{1.6}$$
$$\text{\small}(p\to\infty)$$

The above distances are defined for two-dimensional points, but they can be extended for points of any dimension. For $0 < p < 1$, the l_p distance is not defined since the triangle inequality does not hold; for $1 < p < 2$, the l_p distance lies between the rectilinear and the Euclidean distances; and for $2 < p < \infty$, the l_p distance lies below the Euclidean distance and decreases as p increases.

Proof of Chebychev Distance

Without loss of generality, let $P_1 = (x, y)$, for $x, y \ge 0$, and $P_2 = (0, 0)$. Then $d_\infty(P_1, P_2) = \max\{x, y\}$ and $d_p(P_1, P_2) = \left[x^p + y^p \right]^{1/p}$.

If $x = y$, then $\lim\limits_{p \to \infty} \left[x^p + y^p \right]^{1/p} = \lim\limits_{p \to \infty} \left[2x^p \right]^{1/p} = \lim\limits_{p \to \infty} \left[2^{1/p} x \right] = x$.

If $x < y$, then $\lim\limits_{p \to \infty} \left[x^p + y^p \right]^{1/p} = \lim\limits_{p \to \infty} \left[\left((x/y)^p + 1 \right) y^p \right]^{1/p} = \lim\limits_{p \to \infty} \left((x/y)^p + 1 \right)^{1/p} y = 1 \cdot y = y$.

A similar argument can be made if $x > y$. ■

1.3.2 Great Circle (Geodesic) Distances

Great circle, or geodesic, distances on the surface of a sphere (e.g., the earth (see Figure 1.9) correspond to the shortest distance between two points on the surface along the circle formed by the intersection of the surface and a plane passing through the center of the sphere (i.e., a "great circle"). The elevations of the points on the surface are usually ignored.
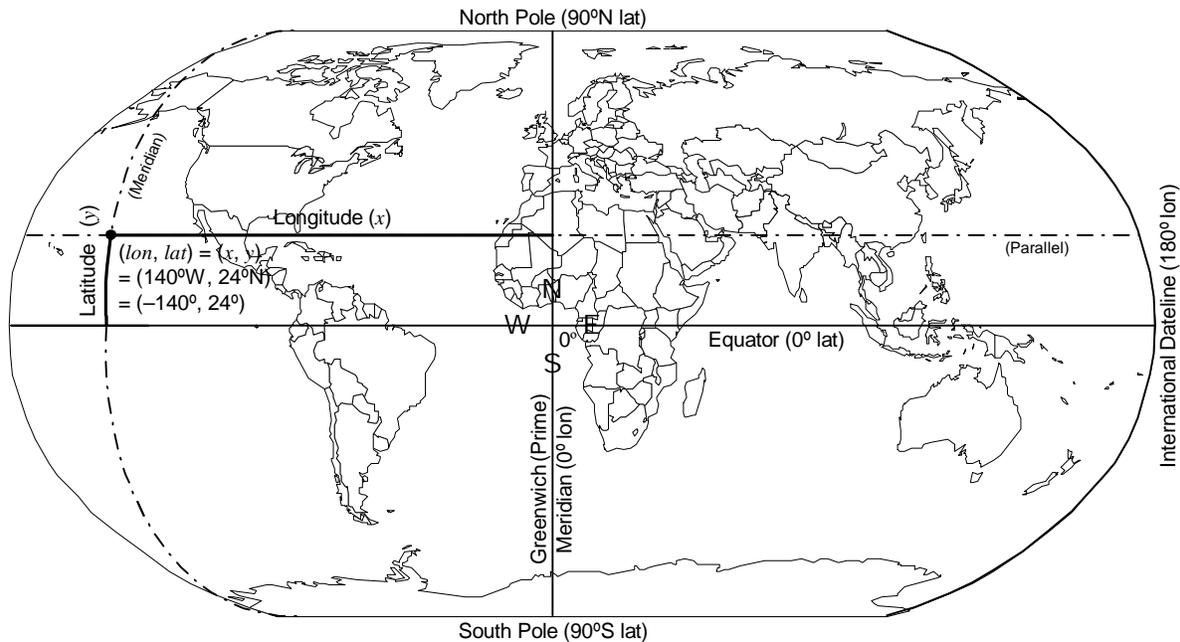
Figure 1.9. Longitude and latitude for points on the surface of the earth.

The great circle distance (d_{GC}) between points 1 and 2 on the surface of the earth, specified by their longitude (*lon*) and latitude (*lat*) angles (in radians), is as follows:

$$(lon_1, lat_1) = (x_1, y_1), \quad (lon_2, lat_2) = (x_2, y_2)$$

$$d_{rad} = (\text{great circle distance in radians of a sphere})$$

$$= \cos^{-1}\left[\sin y_2 \sin y_1 + \cos y_2 \cos y_1 \cos(x_1 - x_2)\right]$$

$$R = (\text{radius of earth at equator}) - (\text{bulge from north pole to equator})$$

$$= 3{,}963.34 - 13.35 \sin\left(\frac{y_1 + y_2}{2}\right) \text{ mi}, \quad = 6{,}378.388 - 21.476 \sin\left(\frac{y_1 + y_2}{2}\right) \text{ km}$$

$$d_{GC} = \text{distance } (x_1, y_1) \text{ to } (x_2, y_2) = \boxed{d_{rad} \cdot R}$$

To convert between decimal degrees and radians: $x_{rad} = \dfrac{x_{deg}}{180}\pi$ and $x_{deg} = \dfrac{x_{rad} \cdot 180}{\pi}$

To convert degrees (DD:MM:SS) to decimal degrees: $x_{deg} = \begin{cases} DD + \dfrac{MM}{60} + \dfrac{SS}{3{,}600}, & \text{if } E \text{ or } N \\[2ex] -DD - \dfrac{MM}{60} - \dfrac{SS}{3{,}600}, & \text{if } W \text{ or } S \end{cases}$

Example

Table 1.1 shows a spreadsheet that calculates the great circle distance from Raleigh, NC to Gainesville, FL, Baghdad, Iraq, and Rio de Janeiro, Brazil. There are typically slight differences in great circle distance calculations due to how accurately the radius of the earth, R, is determined.

Table 1.1. Great Circle Distances from Raleigh

	dd	mm	ss		x (deg)	x (rad)	dd	mm	ss		y (deg)	y (rad)	d(rad)	d (mi)
Raleigh	78	39	32	W	-78.659	-1.373	35	49	19	N	35.8219	0.6252		
Gainesville	82	20	11	W	-82.336	-1.437	29	40	27	N	29.6742	0.5179	0.12	475.0745
Baghdad	44	22		E	44.3667	0.7743	33	14		N	33.2333	0.58	1.62	6407.166
Rio de Janeiro	43	12		W	-43.2	-0.754	22	57		S	-22.95	-0.401	1.181	4679.089

Circuity Factor

Since the great circle distance is usually less than the actual road (or rail) distance between any two locations, the great circle distance can be multiplied by a circuity factor k to approximate the actual road distance:

$$\text{Actual road distance between } P_1 \text{ and } P_2 \quad \approx \quad k \cdot d_{GC}(P_1, P_2)$$

The circuity factor is the average ratio of the actual road distance between two points and the great circle distance between the two points. It can be used to increase the easy-to-determine great circle distance so that it approximates the harder-to-determine actual road distance. You can use the websites *How far is it* or *Google Maps* to determine the great circle distances and the latter to determine road distances.

Circuity factors of 1.15 to 1.25 are typically used for long-distance (> 20 mile) approximations, while factors of 1.25 to 1.50 are typically used for short-distance road approximations and for rail networks because the networks are not as dense. A factor of 1.20 provides a reasonable long-distance road approximation for the continental U.S.

Derivation

In Figure 1.10, a, b, and c are the sides of a spherical triangle and A, B, and C are the corresponding angles. The great circle distance between points 1 and 2 in radians (d_{rad}) corresponds to side c of the triangle.

Given

$$a = 90° - y_2, \quad b = 90° - y_1, \quad C = x_1 - x_2$$

the formula for great circle distances can be derived using the *spherical law of cosines for sides*:

$$\cos(c) = \cos(a)\cos(b) + \sin(a)\sin(b)\cos(C)$$

$$= \cos(90° - y_2)\cos(90° - y_1) + \sin(90° - y_2)\sin(90° - y_1)\cos(x_1 - x_2)$$

$$= \sin(y_2)\sin(y_1) + \cos(y_2)\cos(y_1)\cos(x_1 - x_2)$$
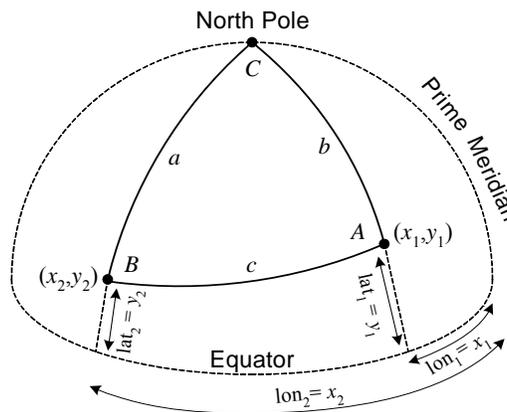
Figure 1.10. Great circle distance derivation.

Solving for c:

$$d_{rad} = c = \cos^{-1}\left[\sin(y_2)\sin(y_1) + \cos(y_2)\cos(y_1)\cos(x_1 - x_2)\right]$$

The above formula can result in round off error if the two points are located at exactly opposite sides of a sphere. Instead, the Haversine formula can be used:[2]

$$d_{rad} = c = 2\sin^{-1}\min\left\{1, \sqrt{\sin^2\left(\frac{y_1 - y_2}{2}\right) + \cos(y_1)\cos(y_2)\sin^2\left(\frac{x_1 - x_2}{2}\right)}\right\}$$

The Haversine formula does not need to be used if the great circle distance is being calculated by hand and the two points are known to not be on opposite sides of the sphere.

1.4 Multifacility Location

In a multifacility location problem, the number of NFs to be located can either be specified or can be determined as part of the location procedure. When the number of NFs is specified, the allocation of EFs to NFs can either be given or determined as part of what is then termed a *location–allocation* problem. If the NF-to-EF allocations are given and there are no interactions between the NFs, then the multifacility problem reduces to a series of single-facility location problems. The location–allocation problem remains difficult even when there are no interactions between the NFs because of the need to determine the allocations.

Determining the best retail warehouse locations is an example of a location–allocation problem, where the EFs are population centroids (e.g., ZIP codes). Table 1.2 lists the best locations for a given number of warehouses. It is assumed that the warehouses serve retail customers located throughout the continental U.S. in proportion to population. Trucks traveling at 400 miles per day are used for all transport. Only outbound transport costs are used in making the location decision; it is reasonable to ignore inbound transport as long as suppliers are located uniformly throughout the country so that the inbound transport costs to each warehouse is approximately the same at any location. As can be seen in the table, the best single warehouse location (Bloomington, Indiana) is not the best location as the number of warehouses increases, while the warehouse in Palmdale, California remains in the best west coast location until a second west coast warehouse in Tacoma, Washington is added as part of the five-warehouse solution.

Table 1.2. Best Retail Warehouse Locations[3]

Number of Locations	Average Transit Time (days)	Warehouse Location		
1	2.20	Bloomington, IN		
2	1.48	Ashland, KY	Palmdale, CA	
3	1.29	Allentown, PA	Palmdale, CA	McKenzie, TN
4	1.20	Edison, NJ Meridian, MS	Palmdale; CA	Chicago, IL
5	1.13	Madison, NJ Dallas, TX	Palmdale, CA Macon, GA	Chicago, IL
6	1.08	Madison, NJ Dallas, TX	Pasadena, CA Macon, GA	Chicago, IL Tacoma, WA
7	1.07	Madison, NJ Dallas, TX Lakeland, FL	Pasadena, CA Gainesville, GA	Chicago, IL Tacoma, WA
8	1.05	Madison, NJ Dallas, TX Lakeland, FL	Pasadena, CA Gainesville, GA Denver, CO	Chicago, IL Tacoma, WA
9	1.04	Madison, NJ Dallas, TX Lakeland. FL	Alhambra, CA Gainesville, GA Denver, CO	Chicago, IL Tacoma, WA Oakland, CA
10	1.04	Newark, NJ Palistine, TX Lakeland, FL Mansfield, OH	Alhambra, CA Gainesville, GA Denver, CO	Rockford, IL Tacoma, WA Oakland. CA

1.4.1 The Uncapacitated Facility Location Problem

Given m EFs and n sites at which NFs can be established, the uncapacitated facility location (UFL) problem can be formulated as the following mixed-integer linear programming (MILP) problem:

$$\text{Min } TC = \sum_{i=1}^{n} k_i y_i + \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} x_{ij} \tag{1.7}$$

subject to

$$\sum_{i=1}^{n} x_{ij} = 1, \quad j = 1, \ldots, m \tag{1.8}$$

$$y_i \geq x_{ij}, \quad i = 1, \ldots, n; \; j = 1, \ldots, m \tag{1.9}$$

$$0 \leq x_{ij} \leq 1, \quad i = 1, \ldots, n; \; j = 1, \ldots, m \tag{1.10}$$

$$y_i \in \{0, 1\}, \quad i = 1, \ldots, n, \tag{1.11}$$

where

k_i = fixed cost of establishing a NF at site i

c_{ij} = variable cost to serve all of EF j's demand from site i

y_i = 1, if NF established at site i; 0, otherwise

x_{ij} = fraction of EF j's demand served from NF at site i.

The UFL problem is a MILP because the y_i's are binary variables and the x_{ij}'s are real variables. In the UFL problem, all x_{ij} are 0 or 1; in the capacitated facility location (CFL) problem, there is a maximum capacity associated with each site, resulting in a x_{ij} value between 0 and 1 whenever not all of an EF j's demand can be served from the NF at site j.

The MILP for the UFL problem is termed the "strong formulation" due to constraints (1.9), which results in an LP relaxation that gives a tight lower bound. When nm constraints (1.9) are replaced with the n constraints

$$m y_i \geq \sum_{j=1}^{m} x_{ij}, \quad i = 1, \ldots, n, \tag{1.12}$$

the formulation is termed "weak" since the lower bound from the LP relaxation is not very tight, resulting in a large branch-and-bound tree.

Example

Given $n = 6$ sites, $m = 6$ EFs, and fixed and variable costs

$$[k_i] = \begin{bmatrix} 8 \\ 8 \\ 10 \\ 8 \\ 9 \\ 8 \end{bmatrix}; \quad [c_{ij}] = \begin{bmatrix} 0 & 3 & 7 & 10 & 6 & 4 \\ 3 & 0 & 4 & 7 & 6 & 7 \\ 7 & 4 & 0 & 3 & 6 & 8 \\ 10 & 7 & 3 & 0 & 7 & 8 \\ 6 & 6 & 6 & 7 & 0 & 2 \\ 4 & 7 & 8 & 8 & 2 & 0 \end{bmatrix},$$

the optimal solution is to establish two NFs at sites 3 and 6 serving EFs 2–4 and EFs 1, 5, and 6, respectively, for a $TC = 31$:

$$[y_i] = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}; \quad [x_{ij}] = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

1.4.2 The *p*-Median Problem

When the number of NFs is specified and all of fixed costs are identical (or not stated), then the fixed costs will have no impact on the location decision and can be set to zero in (1.7) and the following constraint can be added to the UFL problem to formulate the *p-median problem*:

$$\sum_{i=1}^{n} y_i = p, \tag{1.13}$$

where p is the number of NFs to be located. If non-identical fixed are included, then the *p*-median problem generalizes to the *p-UFL problem*, where both *p*-median and UFL are special cases.[4]

1.5 References

The following sources are recommended for further study:

Ahuja, R.K., Magnanti, T.L., and Orlin, J.B., 1993, *Network Flows: Theory, Algorithms, and Applications*, Englewood Cliffs, NJ: Prentice-Hall.

Ballou, R.H., 2004, *Business Logistics/Supply Chain Management*, 5th Ed., Englewood Cliffs, NJ: Pearson.

Daskin, M., 1995, *Network and Discrete Location: Models, Algorithms, and Applications*, New York: Wiley.

Francis, R.L., McGinnis Jr., L.F., and White, J.A., 1992, *Facility Layout and Location: An Analytical Approach*, 2nd Ed., Englewood Cliffs, NJ: Prentice-Hall.

Larson, R.C., and Odoni, A.R., 1981, *Urban Operations Research*, Englewood Cliffs, NJ: Prentice-Hall.

Mirchandani, P.B., and Francis, R.L., Eds., 1990, *Discrete Location Theory*, New York: Wiley.

Simchi-Levi, D., Kaminsky, P., and Simchi-Levi, E., 2008, *Designing and Managing the Supply Chain: Concepts, Strategies, and Case Studies*, 3rd Ed., Boston: McGraw-Hill.

2. Freight Transport

2.1 Logistics Costs

As shown in Table 2.1, the total U.S. logistics costs for 2013 were $1.385 trillion, up 2.3% from 2012. This represents 8.2% of the U.S. Gross Domestic Product[1] (GDP). In 1981, the first full year of trucking deregulation, logistics costs were 16.2% of GDP[2]; in 2000, they were 10.2%,[3] and in 2009 they were 7.7%[4], a record low percentage of GDP. In 1981, transportation and inventory costs represented 45% and 51% of total costs, respectively; in 2013, they represented 62% and 34% of the costs, respectively[5]. In 2013, intercity truck transport alone represents over half of the total transportation costs and 33% of total logistics costs; inventory carrying costs as a percentage of the $2.459 trillion total business inventory were 19.07%.

Table 2.1. Total 2013 U.S. Logistics Costs[6]

Transportation Costs	$ Billion
Intercity Truck	459
Local Truck	204
Rail	74
Maritime	30
Domestic Waterway	7
International Air	13
Domestic Air	20
Forwarders	38
Oil Pipelines	13
Total Transportation Costs	852
Carrying Costs (based on $2,459 billion total inventory)	
Interest	2
Taxes, Obsolescence, Depreciation, Insurance	330
Warehousing	137
Total Carrying Costs	469
Shipper Related Costs	10
Logistics Administration Costs	53
Total Logistics Costs	1,385

2.2 Network Design and Transport Costs

The design of a logistics network has a large impact on transport costs. Raw material or finished goods warehouses can be used for consolidation, cross-docking, or breaking bulk purposes in order to achieve transport economies (see Figure 2.1).
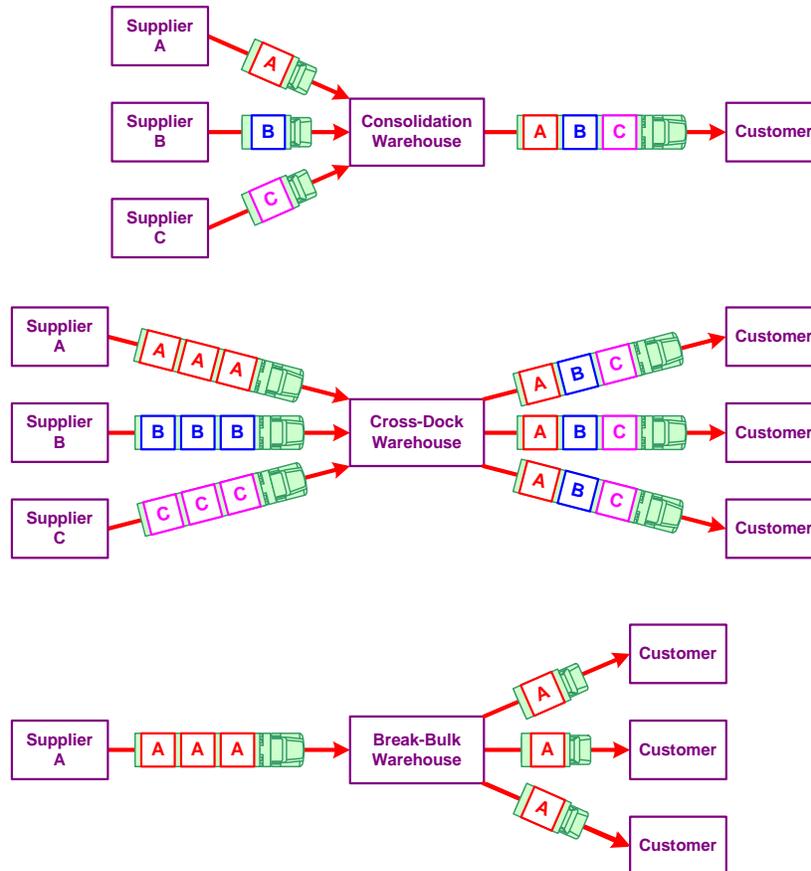
Figure 2.1. Using a warehouse to reduce transport costs.

- *Consolidation warehouse:* A consolidation warehouse is used to combine multiple loads into a single load. Instead of costly LTL or infrequent TL shipments from each supplier directly to the customer, a consolidation warehouse provides less-costly and more frequent TL shipments to the customer. Small delivery vans can be used for short-distance transport between the suppliers and the warehouse, and tractor-trailers for long-distance TL transport to the customer. Consolidation warehouses are typically used by wholesalers.

- *Cross-Dock Warehouse:* A cross-dock warehouse is used to mix freight so that TL shipments can be used for all transport between suppliers and customers. Receiving and shipping are usually coordinated so that no storage is required at the warehouse. In distribution, the ideal of no storage can sometimes be realized using cross docking, where there is a direct flow of material from trucks at the receiving docks to the shipping docks without buffering or storage in-between, but cross docking requires detailed planning and coordination (e.g., implemented using EDI) that in many cases may not be feasible.

- *Break-Bulk Warehouse:* At a break-bulk warehouse, a large long-distance TL shipment from a supplier to broken down into smaller loads that are delivered a short-distance to each customer. Break-bulk warehouses are usually located close to or in each major market served.

2.3 Modes of Transport

2.3.1 Ocean Shipping

Intermodal ISO containers conform to the International Organization for Standardization (ISO) container manufacturing standards. They are used to facilitate loading and unloading at ports and to allow the container to be transported using other modes of transport like trucking and rail. Capacity of ship measured in TEUs (Twenty-foot Equivalent Units), an international standard. Containers were the first "packetized" transport network and were developed in the 1950's by North Carolina trucker Malcolm McLean, who started Sea-Land based in Charlotte.[7]
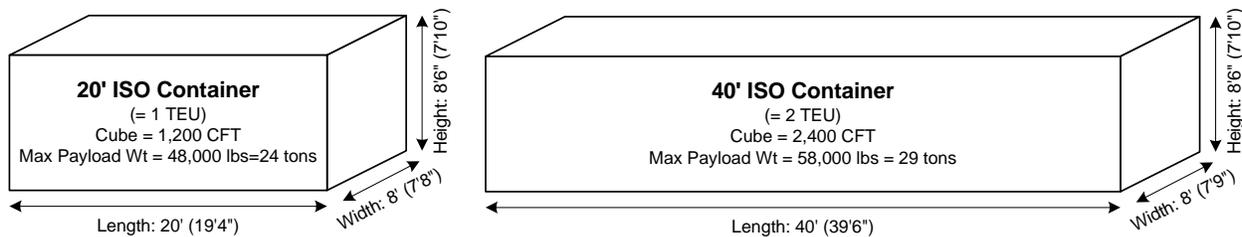
Figure 2.2. Intermodal ISO containers (interior dimensions in parenthesis).

Most containers are either 20 foot or 40 foot (see Figure 2.2), although high-cube 9.5-ft high and 45-, 48-, or 53-ft long containers are also available. In 2004, the cost of transporting a 40-ft container from China to the U.S. West Coast is around $2,400 (plus fees and duties) and takes 16 to 18 days, and the cost from China to the U.S. East Coast is around $4,000 (plus fees and duties) and takes 26 to 30 days. The cost of transporting a 20-ft container is 70% the cost of a 40-ft container. Containers can also be transport by rail and road. Backhaul is sometimes difficult. The number of containers traveling from the Far East to the U.S. West Coast, 11 million TEUs per year, is 2.5 times the number returning, 4.9 million TEUs.[8]

2.3.2 Trucking

Freight can be transported via private trucking or for-hire trucking. For-hire trucking services include full truckload (TL), less-than-truckload (LTL), and package express (PX) (see Table 2.2). TL is 80% of all trucking.[9] PX can also include transport by rail and air, and it also has a limit on the maximum dimension of a load (e.g., 130 in.) in order to allow automated sortation equipment to be used at terminals. "Parcels" are PX loads, while loads under 2 lbs. are referred to as "packets." Other services include bulk, motor vehicle carrier, refrigerated, and tank car.

Figure 2.3 shows the dimension, cube, and weight capacities of an enclosed van semi-trailer, the most common truck trailer. Refrigerated van semi-trailers are referred to as "reefer" units. The maximum gross weight limit of 80,000 lbs applies to the entire vehicle (i.e., 3-axle tractor and 2-axle 53′ semi-trailer). The maximum payload weight of 50,000 lbs is based on an estimated average "tare" weight for the empty vehicle of approximately 30,000 lbs[10] (13,900 lb tractor and 13,800 lb semi-trailer). Although the *physical cube capacity* of a trailer ranges from 3,332 to 3,968 ft^3 for 48 to 53 ft trailers, respectively, in practice not all of this space can be utilized when different-size items are packed into the trailer, resulting in an *effective cube capacity* from approximately 2,500 to 3,000 ft^3.

Table 2.2. U.S. For-Hire Trucking Services[11]

	TL	LTL	PX
Minimum payload	10,000 lb	150 lb	2 lb
Average payload[12]	30,000 lb	1000 lb	10 lb
Maximum payload	50,000 lb	10,000 lb	70 (UPS) – 150 lb
Average length of haul	294 mi	752 mi	894 mi
Average value	$775/ton	$7002/ton	$37,538/ton
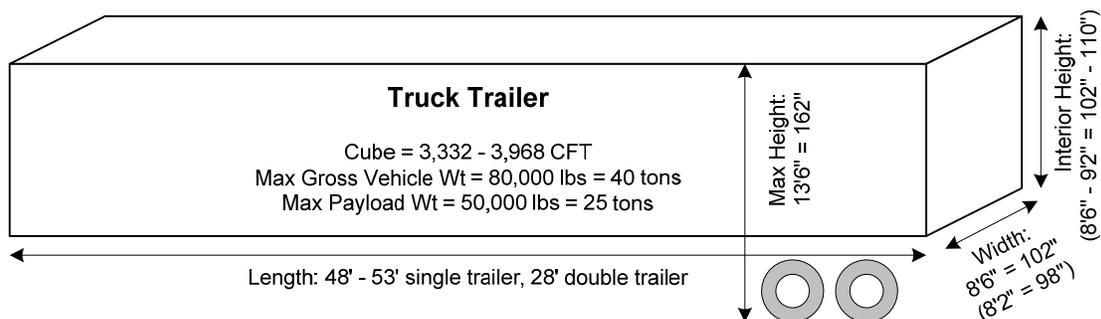
Figure 2.3. Truck enclosed van semi-trailer (interior dimensions in parenthesis).

2.3.3 Rail

Freight can be transported by carload (CL) or less-than-carload (LCL). Truck trailers can be transported on flatcars (TOFC), and ocean containers can be double-stacked on flatcars (COFC). Figure 2.4 shows a typical boxcar. RailInc is a company located in Cary, NC that tracks the movement of railcars throughout the U.S.

Rail Boxcar

Cube = 5,238 - 6,640 CFT
Max Gross Wt = 148,000 - 204,800 lbs = 79 - 102.4 tons
Max Payload Wt = 78,000 - 110,000 lbs = 39 - 55 tons

Height: (10'10" - 13')

Width: (9'6")

Length: (50' - 60')
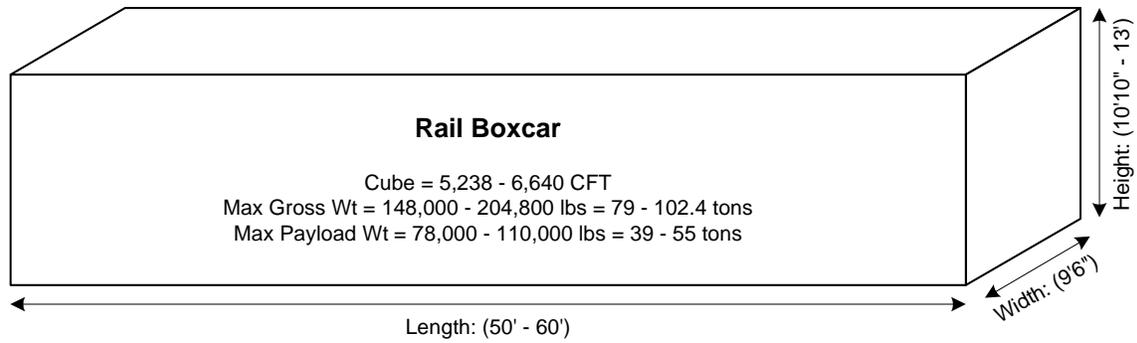
Figure 2.4. Rail boxcar (interior dimensions in parenthesis).

2.4 Trucking Operations

Figure 2.5 shows the variety of different routing alternatives that are available for TL trucking operations. Interleaved routing (Figure 2.5(e)) can be difficult to achieve if the trailer only opens at the rear. The major difference between TL and LTL/PX trucking operations is that the latter requires a network of terminals (see Figure 2.6). In a LTL logistics network, loads in the vicinity of a terminal are collected and delivered to the terminal where they are sorted and loaded onto trucks that provide "linehaul" transport to other terminals. There are fewer firms providing LTL as compared to TL services because of the high cost of constructing a network of terminals. Non-revenue-generating empty (or "deadhead") travel represents approximately 15% of total trucking miles and is used to reposition a tractor-trailer after the final delivery to the next initial pickup point.

Pickup (P) ———— (D) Delivery
(a) Point-to-point (P2P)

(b) Peddling (one-to-many)

(d) Many-to-many

(c) Collecting (many-to-one)

(e) Interleaved
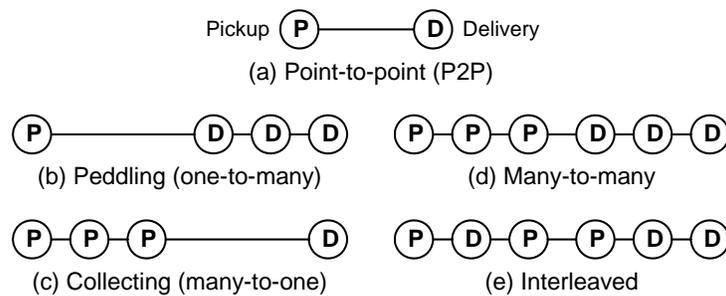
Figure 2.5. TL routing alternatives.

Terminal

Linehaul

Local P/D

Local P/D
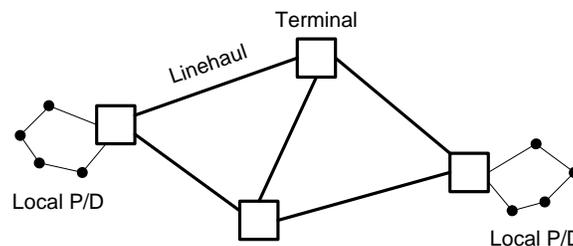
Figure 2.6. Logistics network used for LTL and PX.

As compared to a single shipment transported from its origin to its destination (P2P TL), all of the multi-stop TL routing alternatives shown in Figure 2.5 represent the consolidation of multiple shipments into a single consolidated load. The benefit of consolidation is the potential savings that may accrue from economies of scale in transportation gained by shipping larger loads. As shown in Figure 2.7, consolidated truckloads can be used to transport loads that are smaller than P2P TL and that are of lesser value than LTL.
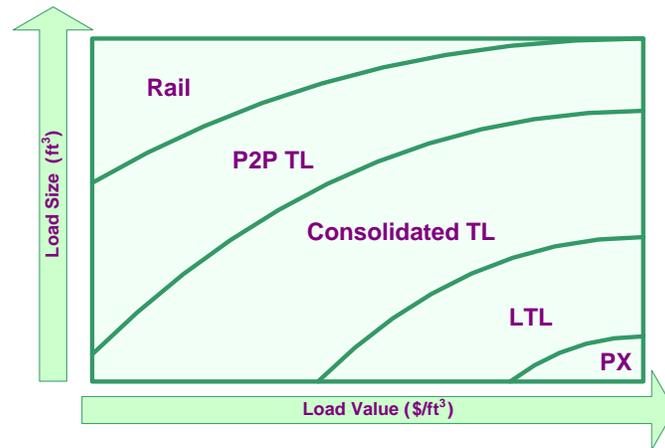
Figure 2.7. Load value versus load size.

2.4.1 HOS Regulations

The Federal Motor Carrier Safety Administration's Hours-of-Service (HOS) regulations provide constraints on the number of hours that a driver can operate a truck: "Drivers may drive up to 11 hours in the 14-hour on-duty window after they come on duty following 10 or more consecutive hours off duty."[13] The HOS regulations effectively limit the total distance traveled by a single driver in a day to around 400 miles. As a result, a DC is limited to serving customers located within a 200 mile radius if it is desired that drivers return to a home location each day. Similar considerations result in a 200-mile maximum separation between LTL terminals. Team drivers can be used to allow almost continual operation, where each driver must rest at least eight consecutive hours in the sleeper berth per HOS regulations.

2.5 One-Time Truck Shipments

The transport mode selected for a particular item is based on the value of item. The cost that is charged to transport single unit of an item via a particular mode (e.g., LCL, LTL) is based on the density of the item. It is more common for a truck enclosed van semi-trailer to *cube out* (i.e., reach its maximum possible cubic volume (e.g., 3500 ft^3 for a truck semi-trailer) than it is to *weigh out* (i.e., reach its maximum weight limit (e.g., 50,000 lb for a truck van semi-trailer): van semi-trailers weigh out only 20% of the time, versus 80% for tank trailers.

The rate paid by a customer for transport service can range anywhere from

- the cost of the service to the provider (the minimum rate) to
- the value of the service to the customer (the maximum rate).

Negotiation is used to set the actual rate within this range based on local market conditions, shipment frequency, etc.

2.5.1 TL Transport Charge

In determining the TL transport charge, revenue per loaded-truck-mile is used instead of the cost per truck-mile because the user of the rate model is assumed to be a shipper (i.e., customer) buying TL service from a carrier on the basis of dollars-per-loaded-mile. The average revenue per mile associated with transporting one loaded trailer one-way is estimated to be \$2.00 for the year 2004.[14] The ratio of an unspecified TL Producer Price Index[15] value (PPI_{TL}) and 102.7, the value for 2004, can be used to adjust the 2004 estimate of \$2.00 to reflect the revenue per loaded truck-mile of the current period:

$$\textit{TL Transport Charge (\$):} \quad c_{TL} = \left\lceil \frac{q}{q_{max}} \right\rceil r\, d \tag{2.1}$$

where

$$\left\lceil \frac{q}{q_{max}} \right\rceil = \text{number of truckloads per shipment}$$

$$r = \text{TL revenue per loaded truck-mile (\$/mi)}$$

$$= \frac{PPI_{TL}}{102.7} \times \$2.00 \,/\, \text{mi} \quad \text{(average estimate)}$$

$$PPI_{TL} = \text{Producer Price Index for TL service } (= 102.7 \text{ in } 2004; 113.5 \text{ in } 2007; 113.3 \text{ in } 2010)$$

$$d = \text{road distance between the O-D pair (mi)}$$

$$q = \text{shipment weight (tons)}$$

$$q_{max} = \text{maximum payload (tons).}$$

In (2.1), it is assumed that any portion of the load exceeding the maximum payload of the truck is still transported TL using additional trucks.

Maximum Payload

The maximum payload for a shipment is the maximum size of each truckload and is determined by whether a trailer is constrained by weight or cubic volume:

$$\textit{Maximum Payload (tons):} \quad q_{max} = \min\left\{ K_{wt}, \frac{s\,K_{cu}}{2000} \right\} \tag{2.2}$$

where

$s =$ item density (lb/ft³)

$K_{wt} =$ weight capacity of truck trailer (tons)

$K_{cu} =$ cube capacity of truck trailer (ft³).

In (2.2), the maximum payload with respect to cubic volume is determined by solving for q in the following equation:

$$K_{cu} = \frac{q}{\left(\dfrac{s}{2000}\right)},$$

where $(s/2000)$ is the density in ton/ft³.

Example: The transport charge in 2005 to ship $q = 2$ tons of $s = 9.72$ lb/ft³ product 532 mi from Raleigh, NC to Gainesville, FL can be estimated as follows: Using the 2005 PPI_{TL} of 108.6, a weight capacity of 25 tons, and, since the length of the trailer has not been determined, an effective cube capacity of 2,750 ft³, which is the midpoint of the effective capacities for 48 to 53 ft trailers.

$$r = (108.6/102.7) \times \$2.00/\text{mi} = \$2.11/\text{mi};$$

$$q_{max} = \min\{25, 9.72\,(2750)/2000\} = 13.37 \text{ tons};$$

$$c_{TL} = \lceil 2/13.37 \rceil (2.12)\,532 = \$1{,}125.13.$$

Aggregate Shipment

When multiple items are shipped together as part of a single load, then it is convenient to view them as a single demand-weighted *aggregate shipment*,[16] where, for m items, the aggregate weight and aggregate density are

$$q_{\text{agg}} = \sum_{i=1}^{m} q_i \tag{2.3}$$

$$s_{\text{agg}} = \frac{\left(\text{aggregate weight, in lb}\right)}{\left(\text{aggregate cube, in ft}^3\right)} = \frac{2000\,q_{\text{agg}}}{\sum_{i=1}^{m} 2000\,q_i/s_i} = \frac{q_{\text{agg}}}{\sum_{i=1}^{m} q_i/s_i}, \tag{2.4}$$

where q_i are in tons and s_i in lb/ft³.

2.5.2 Estimated LTL Transport Charge

The following model was developed from tariff rate tables and provides a general means of estimating rates for LTL transport between origin-destination (O-D) pairs located anywhere within the continental United States.[17] Since it requires only distance, weight, and density as inputs and allows direct comparison of LTL and TL rates, it can be used in the earliest stages of logistics network design when location decisions are being made and when the most appropriate

shipment size for each lane (O-D pair) in the network is being determined. In most commercial transportation management and planning systems, LTL rates are determined using tariff tables (e.g., CzarLite), but this requires the shipper/decision maker to purchase access to the tariff tables and, further, to know what discount to apply to the tariff rates.

The model reflects average industry rates and to allow rate estimates to be adjusted to current economic conditions by using the current Producer Price Index for LTL service:

$$\text{\textit{LTL Rate}} (\$/\text{ton-mi}): \qquad r_{LTL} = PPI_{LTL} \left[\frac{\dfrac{s^2}{8} + 14}{\left(q^{\frac{1}{7}} d^{\frac{15}{29}} - \dfrac{7}{2} \right)(s^2 + 2s + 14)} \right], \qquad (2.5)$$

$$\text{\textit{LTL Transport Charge}} (\$): \qquad c_{LTL} = r_{LTL}\, q\, d \qquad\qquad\qquad\qquad (2.6)$$

where

PPI_{LTL} = Producer Price Index for LTL service[18] (= 104.2 in 2004; 121.0 in 2007; 126.8 in 2010)

and $\dfrac{150}{2,000} \le q \le \dfrac{10,000}{2,000}$ (tons), $37 \le d \le 3354$ (mi), and $2000\, q/s \le 650$ (ft^3), with these conditions representing the range of input data that produced the estimate and also ensuring that the denominator of (2.5) remains positive. The LTL transport rate is shown as a function of q and d. Note that, in actuality, an LTL shipment between any origin and destination is likely to travel a longer distance than the road distance d because it will travel through one or more transshipment terminals along its journey. However, the road distance between O-D pairs is the only readily available measure of distance, and, since the particular network of transshipment terminals is specific to each LTL carrier, it is the only reasonable measure. The model (2.5) is independent of the particular characteristics of any O-D pair and, since the only parameter that distinguishes O-D pairs is road distance, the resulting rate estimate is symmetric with respect to O-D order, with such differences being treated as noise, reflecting the general nature of the model and its use in reasonably general optimization studies.

2.5.3 Minimum Charge

In practice, in addition to the basic P2P TL and LTL charges, there is a minimum charge associated with any shipment that corresponds to the cost of providing the transport service that is not related to the weight of the shipment. There is a separate estimated minimum charge for TL and LTL:

$$\text{\textit{TL Minimum Charge}} (\$): \qquad MC_{TL} = \left(\frac{r}{2} \right) 45 \qquad\qquad\qquad\qquad (2.7)$$

2. FREIGHT TRANSPORT

LECTURE NOTES FOR PRODUCTION SYSTEM DESIGN

$$\text{LTL Minimum Charge (\$):} \quad MC_{LTL} = \left(\frac{PPI_{LTL}}{104.2}\right)\left(45 + \frac{d^{\frac{28}{19}}}{1625}\right), \tag{2.8}$$

where $d > 0$ and $q > 0$; $MC_{TL} = MC_{LTL} = 0$, for $d = 0$ or $q = 0$.

The minimum charge for TL is independent of distance and depends solely on loading and unloading costs at the origin and destination of the shipment, while the charge for LTL is a function of the distance of the shipment because each shipment is loaded and unloaded at each LTL terminal visited in transit and the number of terminals visited increases with the length of the shipment.

2.5.4 Independent Shipment Transport Charge

An *independent shipment* corresponds to either a P2P TL shipment or a LTL shipment. It represents an alternative to a multi-stop *consolidated load*, where multiple shipments are carried on a single truck at the same time and the potential savings associated with transporting multiple shipments on the same truck is offset by an increase in the loaded distance that the truck travels as compared to a P2P TL shipment or the charge associated with shipping via LTL. Such consolidated loads differ from multiple shipments with the same origin and destination (O-D) because the latter can be combined together as an aggregate shipment and treated as a single shipment as long as the shipments occur within the same time period, while with the former there is an additional charge associated with transporting the consolidated load.

Using the TL and LTL transport and minimum charges,

$$\text{Independent Transport Charge (\$):} \quad c_0 = \min\left\{\max\left\{c_{TL}, MC_{TL}\right\}, \max\left\{c_{LTL}, MC_{LTL}\right\}\right\}. \tag{2.9}$$

34
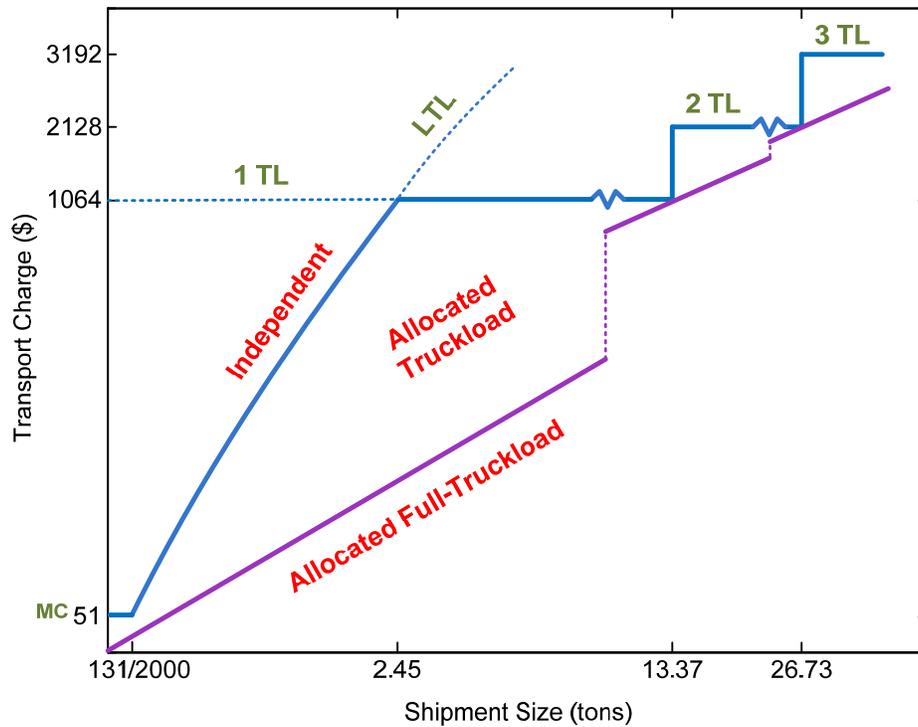
Figure 2.8. Transport charge for a shipment.

An example of transport charges for a range of shipment sizes is shown in Figure 2.8. The independent transport charge is the top solid curve in the figure and represents an upper bound on the charge for a given shipment size (the allocated full-truckload charge shown as the bottom curve represents an effective lower bound on the charge and is discussed in more detail in Section 2.7). For the independent transport charge curve, LTL is used for sizes up to 2.45 tons, a single P2P TL truckload for sizes up to the maximum payload of 13.37 tons, two truckloads for up to two times the maximum payload, etc. A minimum charge of $51.40 is invoked for sizes less than 131 lb. The single independent shipment is from Raleigh, NC to Gainesville, FL and $s = 9.72$ lb/ft^3, $d = 532$ mi, $r = \$2/$mi, $PPI_{LTL} = 104.2$, $K_{wt} = 25$ tons, and, $K_{cu} = 2{,}750$ ft^3.

2.5.5 Comparing TL and LTL Rates

In Figure 2.9, LTL and TL rates are compared for three different load densities for 532-mile shipments between Raleigh and Gainesville, FL as a function of the shipment weight. The LTL rate is (2.5) and the TL rate is $r_{TL} = c_{TL}/(q\,d)$. The shipment weight that yields equal rates between LTL and TL is labeled in each figure. For this example, the volume and weight capacities of the truck trailer are assumed to be $K_{cu} = 3{,}000$ ft^3 and $K_{wt} = 24$ tons, respectively. All of the comparisons are for year 2004, meaning that $PPI_{LTL} = 104.2$ and $r = \$2/$mi.
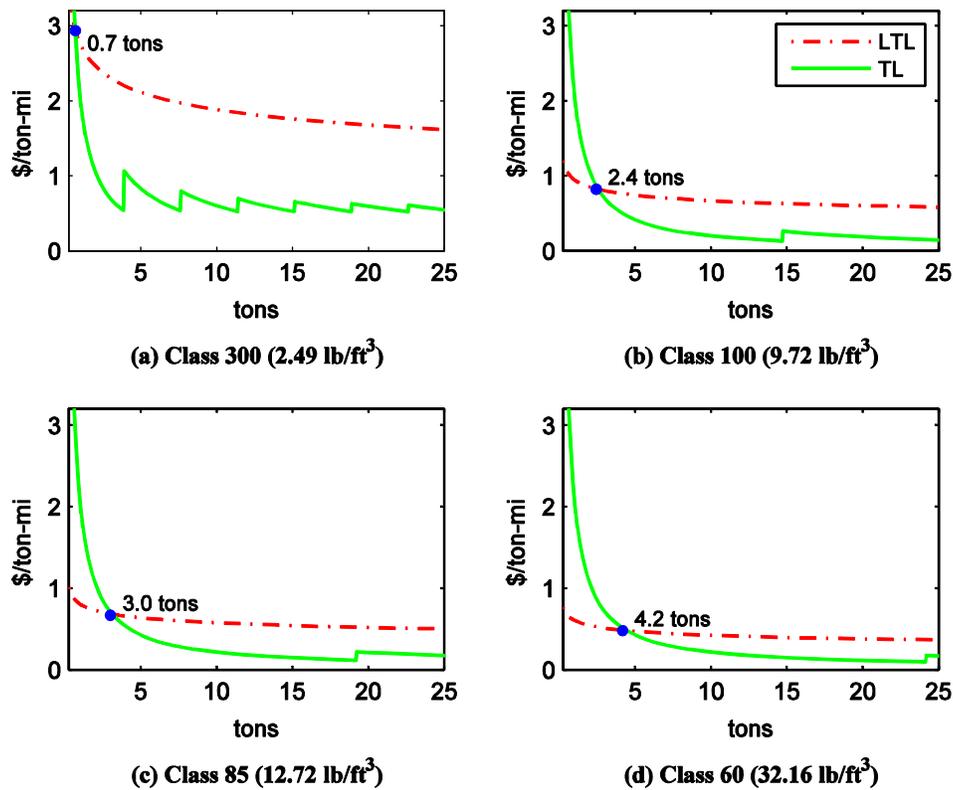
Figure 2.9. TL vs. LTL rates from Raleigh, NC to Gainesville, FL (532 mi).

2.6 LTL Tariff

LTL rates are dependent on a number of factors, prominent among them being the specific origin and destination, the weight of the shipment, and the freight class to which the goods being shipped belong. LTL charges are typically determined from rates quoted in a tariff. A separate table is provided in the tariff for each particular pair of origin and destination (O-D) points (typically zip codes) due to different local market conditions like demand imbalances that can result in an excess of empty trailers at some locations (e.g., more freight is shipped to Florida than from Florida, so that rates to Florida are higher than the rates from Florida).

The specific characteristics of each item that impact its transport cost need to be considered in assigning each item to be transported to a freight class, including the following considerations:

1. Load density (e.g., a large item will cube-out a trailer sooner than a smaller item that has the same weight).

2. Special handling (e.g., fragile loads; hazardous materials; unit load size: it is more costly to handle several small loads that comprise a single shipment that together have the weight as a single large unit load).

3. Stowability (e.g., some items can be nested).

4. Liability (e.g., high value items are more expensive to insure while in transit).

The *National Motor Freight Classification*[19] is typically used to determine the rating of an item. Most LTL carriers have a *Freight All Kinds* (FAK) rate that can be used for any item that cannot be classified. Discounts of up to 15% from the published rates are usually available for a single one-time shipment; when a firm has frequent shipments, discounts of 30–65% can usually be negotiated.

Table 2.3. Class-Density Relationship
(italics indicate value at capacity)

	Load Density (lb/ft^3)		Max Physical	Max Effective
Class	Minimum	Average	Weight (tons)	Cube (ft^3)
500	–	0.52	0.72	*2,750*
400	1	1.49	2.06	*2,750*
300	2	2.49	3.43	*2,750*
250	3	3.49	4.80	*2,750*
200	4	4.49	6.17	*2,750*
175	5	5.49	7.55	*2,750*
150	6	6.49	8.92	*2,750*
125	7	7.49	10.30	*2,750*
110	8	8.49	11.67	*2,750*
100	9	9.72	13.37	*2,750*
92.5	10.5	11.22	15.43	*2,750*
85	12	12.72	17.49	*2,750*
77.5	13.5	14.22	19.55	*2,750*
70	15	18.01	24.76	*2,750*
65	22.5	25.50	*25*	1,961
60	30	32.16	*25*	1,555
55	35	39.68	*25*	1,260
50	50	56.18	*25*	890

Table 2.3 shows the minimum density[20] and average density for each freight class. Also shown in the table is the maximum weight for a shipment assuming a maximum physical payload of 25 tons (50,000 lb) and the maximum cubic volume for a shipment assuming a maximum effective payload of 2,750 ft^3. A *maximum effective cube capacity* of 2,750 ft^3 is approximately 80–70% of the respective 3,332–3968 ft^3 maximum physical cube capacity of a truck trailer (see Figure 2.3), and is used as an estimated of the lost cube capacity associated with the packing of different size loads into a trailer. If same size loads are packed, then the actual maximum cube capacity of a trailer can be calculated. The *maximum physical weight capacity* is used because there is little potential loss of weight capacity associated with packing loads into a trailer.

The *ideal density*[21] ($50,000/2750 = 18.18$ lb/ft^3) is the density at which a full truckload is simultaneously at the tractor-trailer's cubic capacity and its weight capacity. An average freight mix at this density best utilizes both the weight and cube capacities of a tractor-trailer. Shipments weigh-out when their average density is greater than the critical density and cube-out when it is

less. Note that Class 100 represents the "average" load and has an average density of 9.72 lb/ft^3, which is in the range of densities that cube-out a trailer.

Table 2.4 is an example of a tariff table (CzarLite tariff DEMOCZ02 04-01-2000[22]). It is for the O-D pair Raleigh, NC (ZIP code 27606) to Gainesville, FL (ZIP code 32606). As one can see, LTL rates for a given O-D pair are clearly a function of the class of the freight being shipped (first column of Table 2.4), as established by agencies like the National Motor Freight Traffic Association (NMFTA), and the weight of the shipment (bottom row). In the bottom row of the table, the mid-points of the weight ranges, in tons, at which the rates, in $/cwt (where cwt is "hundred weight," or 100 lbs), change, termed *rate breaks*, are provided. The actual road distance spanned by this O-D pair is 532 miles. The minimum charge for this tariff is $95.23. Rates above 10,000 lb (5 tons) are given in the tariff even though LTL shipments usually do not exceed this weight.

Table 2.4. Tariff (in $/cwt) from Raleigh, NC (27606) to Gainesville, FL (32606)
(532 mi, CzarLite DEMOCZ02 04-01-2000, minimum charge = $95.23)

Freight Class	Rate Breaks (i)								
	1	2	3	4	5	6	7	8	9&10
500	341.42	314.14	245.80	201.48	158.60	112.37	55.66	55.66	55.66
400	273.88	251.99	197.19	161.61	127.22	91.12	45.10	45.10	45.10
300	206.34	189.85	148.56	121.76	95.85	69.47	34.43	34.43	34.43
250	172.56	158.77	124.23	101.83	80.15	58.03	28.79	28.79	28.79
200	138.78	127.69	99.92	81.89	64.47	47.19	23.40	23.40	23.40
175	121.37	111.68	87.39	71.62	56.38	41.27	20.39	20.39	20.39
150	104.49	96.13	75.22	61.66	48.53	35.96	17.75	17.75	17.75
125	87.59	80.60	63.07	51.69	40.69	30.24	15.00	15.00	15.00
110	77.57	71.37	55.85	45.77	36.04	28.61	14.40	14.40	14.40
100	71.23	65.55	51.29	42.04	33.09	27.58	14.03	10.80	9.90
92	66.48	61.18	47.88	39.24	30.89	25.75	13.68	10.52	9.66
85	61.74	56.80	44.45	36.43	28.68	23.91	13.20	10.15	9.32
77	56.99	52.44	41.04	33.63	26.48	22.07	12.60	9.68	8.89
70	52.77	48.55	37.99	31.14	24.51	20.43	12.00	9.23	8.47
65	50.07	46.08	36.05	29.56	23.04	19.39	11.87	9.14	8.39
60	47.44	43.64	34.15	28.00	21.82	18.37	11.76	9.04	8.30
55	44.75	41.17	32.22	26.40	20.59	17.32	11.64	8.96	8.22
50	41.57	38.26	29.94	24.54	19.12	16.10	11.52	8.85	8.14
Tons (q_i^B)	0.25	0.5	1	2.5	5	10	15	20	∞

Using the tariff information in Table 2.4, Figure 2.10 shows plots of both the transport charge (labeled Total Cost, TC_{tariff} in the figure) and the rate per ton-mile for a Class 100 shipment from Raleigh to Gainesville. (Values along both axes in Figure 2.10(a) and along the bottom axis in Figure 2.10(b) are plotted on a log-base-10 scale.) The broken lines plotted in the figure show the total charge and rate without applying any minimum charges or *weight breaks*, the latter of which eliminate any incentive for a shipper to over-declare a shipment weight in order to receive a lower rate. Formally, each table can be represented as a matrix, *OD*, containing the freight charge per hundredweight, with the rows in *OD* designating the rate class (*class*) and the columns designating the weight grouping or rate break (*i*), as discussed above.
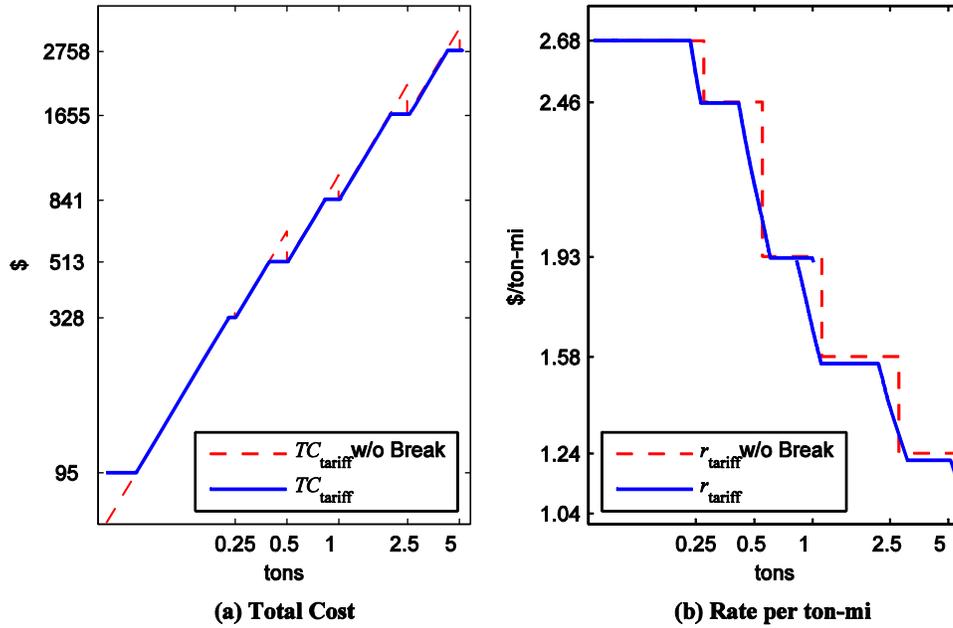
Figure 2.10. LTL tariff from Raleigh, NC to Gainesville, FL for Class 100.

Each rate break i corresponds to a range $\left[q_{i-1}^B, q_i^B\right)$ of shipment weights (in tons) into which a shipment can fall. Weight breaks occur when the charge for the minimum weight of the next rate break, q_i^B, is less than the charge in the current rate break. Given this formalization of the tariff table, the tariff transport charge can be computed as:

$$c_{\text{tariff}} = (1 - disc) \max \left\{ MC, \min \left\{ OD(class, i) \, 20q, OD(class, i+1) \, 20q_i^B \right\} \right\}$$

$$i = \arg \left\{ q_i^B \mid q_{i-1}^B \leq q < q_i^B \right\},$$

(2.10)

where $q_0^B = 0$, MC is the minimum charge, and $disc$ is the discount provided by the carrier, if appropriate. Solving (2.10) for q, the weight break i corresponds to

$$q_i^W = \frac{OD(class, i+1)}{OD(class, i)} q_i^B .$$

(2.11)

For example, using Table 2.4, if one were to ship two tons ($q = 2$) of Class 100 goods ($class = 10$ — i.e., the tenth row of Table 2.4) at a 60% discount ($disc = 0.6$), then $i = 4$ (since $1 = q_3^B \leq q < q_4^B = 2.5$), $OD(class, i) = 42.04$, the total cost is

$$c_{\text{tariff}} = (1 - 0.6) \max \left\{ 95.23, \min \left\{ 42.04 \, (20) \, 2, 33.09 \, (20) \, 2.5 \right\} \right\}$$

$$= 0.4 \max \left\{ 95.23, \min \left\{ 1682, 1655 \right\} \right\} = \$662.$$

The corresponding weight break is $q_4^W = 33.09(2.5)/42.04 = 1.97$ tons, which is just below the shipment size of two tons.

39

2.6.1 Comparison of Estimated LTL Charge

In order to compare the LTL transport charge that was determine using the rate estimate (2.5) to the charge as determined using the tariff, the following three O-D test pairs were used:

- Raleigh, NC (27606) to Gainesville, FL (32606): 532 miles

- Detroit, MI (48234) to Dothan, AL (36302): 926 miles

- Black Mountain, NC (28711) to Salt Lake City, UT (84101): 1938 miles.

These test pairs represent different distances and, in each test pair, a larger population city is paired with a smaller population city so that the rates for that lane reflect a balance of high and low demands, as opposed to, for example, lanes connecting two large cities, which are likely to have more frequent and lower-cost service due to greater competition between carriers serving those cities.
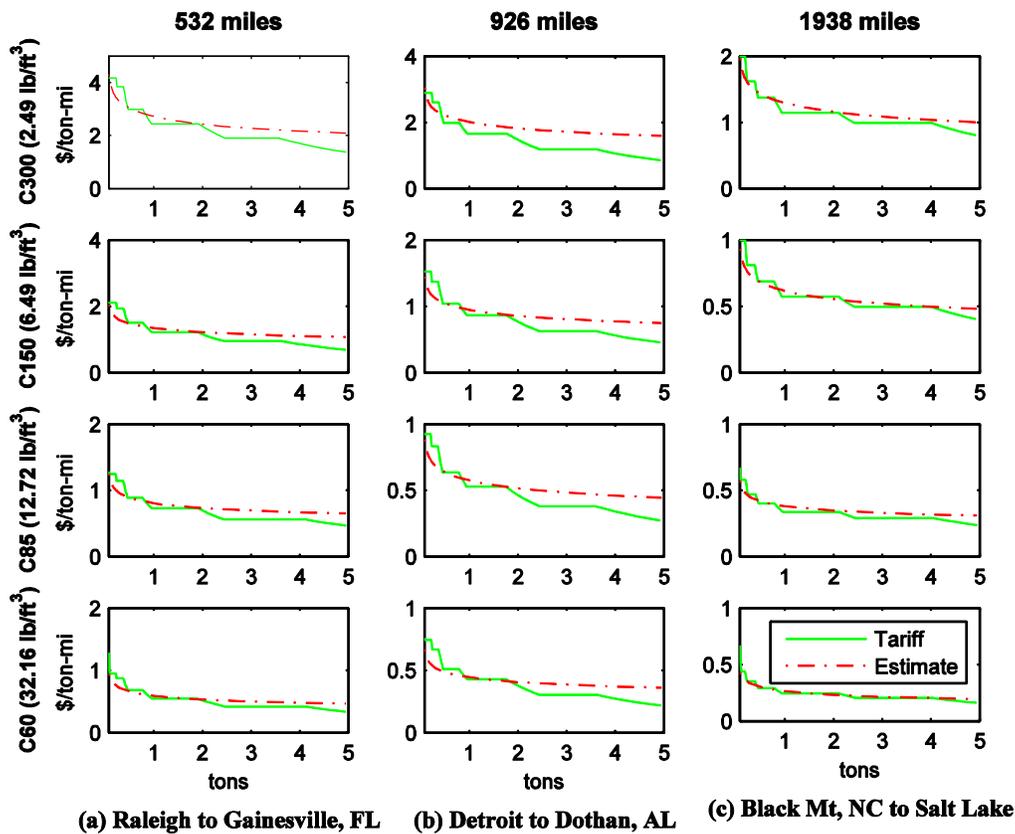

Figure 2.11. Comparison of estimated LTL rate to tariff rate.

Estimate Compared to Tariff

In Figure 2.11, we show a comparison of the estimated LTL rates and tariff rates for four different load densities for each of our O-D test pairs, where the tariff rate is $r_{tariff} = c_{tariff}/(q\,d)$. All of the comparisons are for 2004 (i.e., using $PPI_{LTL} = 104.2$), with all tariff rates discounted by 46.2512% .[23]

**Table 2.5. Comparison of Estimated LTL Charge to One-Time Internet-based Spot Quote
(1000 lb, Class 100)**

| | Raleigh to Gainesville | | | | Detroit to Dothan | | | | Black Mt. to Salt Lake City | | | |
Quote	Serv Day	Insur Liabil	Transport Charge	Incr (%)	Serv Day	Insur Liabil	Transport Charge	Incr (%)	Serv Day	Insur Liabil	Transport Charge	Incr (%)
Est.			$324.25				$405.92				$557.21	
1	2	$2,000	288.69	−11	3	$500	342.90	−16	6	$500	516.30	−7
2	3	500	301.02	–	2	2,570	523.83	29	4	2,570	653.26	17
3	2	10,000	338.27	4	3	5,000	568.63	40	4	5,000	836.09	50
4	2	25,000	362.84	12	3	2,350	595.21	–	3	2,350	875.85	57
5	2	2,570	390.97	–	3	25,000	628.77	55	4	20,000	968.94	74
6	2	2,350	462.36	–	2	2,350	647.13	–	4	25,000	981.35	76
7	1	5,000	464.92	43	3	5,000	777.40	–	5	5,000	1,016.67	–
8	2	20,000	471.38	–					4	2,350	1,062.65	–
9	3	25,000	532.53	–								
10	3	5,000	534.88	–								
11	1	2,350	583.25	–								

Estimate Compared to Spot Quote

In Table 2.5, for each of the three test O-D pairs, the estimated total charge total charge obtained from (2.5) is compared to multiple rate quotes obtained from an internet-based service (Freight101[24]) for a one-time shipment of a 1,000-lb, Class 100 LTL load. The quotes were obtained on August 17, 2006 for an August 28 shipping date, and all shipments were categorized as commercial with no special service or hazmat requirements. For each quote, the number of service days required for transit, the insurance liability for the load, and the total charge are listed. By way of comparison, the first row shows the estimated total charge, c_{LTL}, assuming a value of $PPI_{LTL} = 119.5$ for July 2006, $q = 0.5$ tons, $s = 9.72$ lb/ft^3 (see Table 2.3), and $d = 532$ miles. For each non-dominated quote—i.e., those that are not dominated by another quote with respect to a lower number of service days, greater insurance liability, or a lower total charge— the percentage increase of the quote's total charge over the estimated value c_{LTL} is shown. This increase over the estimated charge represents the premium paid for a one-time shipment through this freight service as compared to an average LTL shipment, most of which operate under longer-term contracts. The average premiums of the non-dominated quotes for each O-D pair are 12%, 27%, and 45%, respectively, and the average of these three averages is 28%. Most of the one-time quotes are greater than the estimated charge. This is not unexpected because the estimated charge reflects an average over all LTL shipments, of which relatively few would be one-time transactions. In general, the lowest quotes correspond to a higher number of service days and a lower insurance liability. With an estimated $7,002 average value per ton (see Table

2.2), the average 1,000-lb LTL shipment used in the comparison would have an estimated value of around \$3,501 and a corresponding average liability of some larger amount.

2.7 Periodic Truck Shipments

When demand for truck transport between O-D points occurs repeatedly over a period of time, multiple demands are usually combined into shipments that occur periodically. Multiple demands are combined into single shipments because of the economies of scale associated with truck transport. Unlike a one-time shipment, where its size and time of occurrence is predetermined, periodic shipments require that the size and interval between shipments be determined. The exact timing and size of each shipment is determined by considering both the total transport costs (*TC*) and the total shipment-related inventory costs (*IC*), or what is termed the total logistics cost:

$$\textit{Total Logistics Cost:} \quad TLC = TC + IC = nc + IC, \tag{2.12}$$

where

$n = \dfrac{f}{q}$ = average shipment frequency (1/yr)

$t = \dfrac{q}{f}$ = average shipment interval (yr)

$f =$ expected annual demand (tons/yr)

$q =$ average shipment size (tons)

$c =$ transport charge (\$).

Since there is no cost savings associated with using more than a single truckload for each periodic shipment, a truck's maximum payload provides an upper bound on the size of each shipment:

$$q \le q_{max}. \tag{2.13}$$

As a result of (2.13), the number of shipments per year equals the number of truckloads. Note that this number is not restricted to integer values, which is reasonable if the demand will continue indefinitely.

Aggregate Periodic Shipment

Since for periodic shipments q is not given and must be determined, f can be used in place of q when determining the aggregate demand and density of multiple items shipped together as part of a single load (cf. (2.3) and (2.4)):

$$f_{agg} = \sum_{i=1}^{m} f_i \tag{2.14}$$

$$s_{agg} = \frac{f_{agg}}{\sum_{i=1}^{m} f_i / s_i} .$$

(2.15)

Also, the aggregate value is the demand-weighted sum of the values of the items:

$$v_{agg} = \sum_{i=1}^{m} \frac{f_i}{f_{agg}} v_i .$$

(2.16)

2.7.1 Allocated Full-Truckload Charge

If inventory costs are small enough relative to transport costs, then a single shipment will always be comprised of an entire full truckload, so that $q = q_{max}$. It other situations, it is also likely that a consolidated load comprised of many different shipments, each with a different cube and weight, will only be transported as a full truckload (e.g., many big-box retailers only transport full truckloads from a DC to a store). In this case, although the different shipments could be combined into a single aggregate shipment, it is often more convenient to determine the transport rate for each shipment per ton-mile assuming that r is allocated to the product based on its maximum payload,

$$\text{FTL Rate (\$/ton-mi):} \quad r_{FTL} = \frac{r}{q_{max}},$$

(2.17)

so that

$$\text{FTL Transport Cost (\$/yr):} \quad TC_{FTL} = f\, r_{FTL}\, d = n\, r\, d = w d .$$

(2.18)

where w is the monetary weight (\$/mi).

When it is reasonable to assume that a shipment will always be transported along with other shipments as part of a full truckload, then using r_{FTL} as the truckload transport rate for the shipment allows its transport cost to be considered in a location decision without having to know the exact mix of other shipments being transported—and the mix, size, and timing can change with each shipment without impacting the decision since only f, the annual demand, is used in the analysis—and its cost per mile for transport is independent of shipment size and distance. The formulation using w is useful in location analysis where (2.18) corresponds to the criterion used for the minisum transport-oriented single-facility location problem.

Another means of indirectly accounting for inventory costs when assuming full-truckload shipments is to place an upper limit on the average shipment interval. If t_{max} is the maximum shipment interval, then the FTL shipment size is the lesser of the following:

$$\text{FTL with Shipment Interval Constraint (\$/ton-mi):} \quad r_{FTL} = \frac{r}{\min\{f\, t_{max}, q_{max}\}} .$$

(2.19)

43

2.7.2 Range of Possible Transport Charges

Figure 2.8 shows the range of possible transport charges for a single shipment. The independent-shipment charge (c_0 via (2.9)) corresponds to the maximum charge, while the full-truckload charge ($c_{FTL} = r_{FTL}\, q\, d$) corresponds to the minimum likely charge. When multiple shipments are carried on a single truck as part of a consolidated load, each shipment's charge corresponds to an allocated portion of the total transport charge and is represented in the figure as any of the charges between by the independent and full-truckload charges. The consolidated charge approaches the full-truckload charge as the load reaches truck capacity. All of the charges are the same at the maximum payload. Note also that an allocated charge can be less the minimum charge for an independent shipment.

2.7.3 Total Logistics Costs

Selecting the size of a shipment on the basis of its *TLC* requires that both inventory carrying and transportation costs be specified as a function of q. Considering only cycle inventory costs,

$$TLC(q) = TC(q) + IC(q) = \frac{f}{q} c(q) + \alpha v h q ,\qquad (2.20)$$

where

> α = average inter-shipment inventory fraction at origin and destination
>
> v = unit value of shipment ($/ton)
>
> h = inventory carrying rate, the cost per dollar of inventory per year (1/yr).

The cost of holding one ton of inventory for one year is $v h$ ($/ton-yr). The inventory carrying rate is expressed as a fraction since the cost per dollar of inventory per year ($/$/yr) reduces to 1/yr. The parameter α denotes the average fraction of the shipment size q that is held, in total, across the origin and the destination. For example, assuming that, on an annual basis, the supply and demand rates between a given O-D pair are constant, then, for a shipment of size q, the expected cycle inventory is $q/2$ at each end,[25] meaning that the total cycle inventory across the origin and the destination is q and, therefore, $\alpha = 1$; also, if supply is not constant (e.g., batch production) but the time of production is not coordinated with the time of shipment, then $\alpha = 1$ (assuming production is equally likely to have occurred at anytime between shipments). If production at the origin is instantaneous, with a constant demand rate at the destination (i.e., the traditional EOQ model), then $\alpha = 0.5$. If, however, both production and consumption are instantaneous, then $\alpha = 0$ (more on this case below). Finally, if the supplier is another firm and there are no negotiations with the supplier—e.g., to share the benefits of using an optimal shipment size that accounts for the supplier's inventory costs—then (from the point of view of the shipper-customer) $\alpha = 0.5$. Thus, annual inventory holding costs are given by $\alpha v h q$.

(a) $3000 value per ton (b) $6000 value per ton
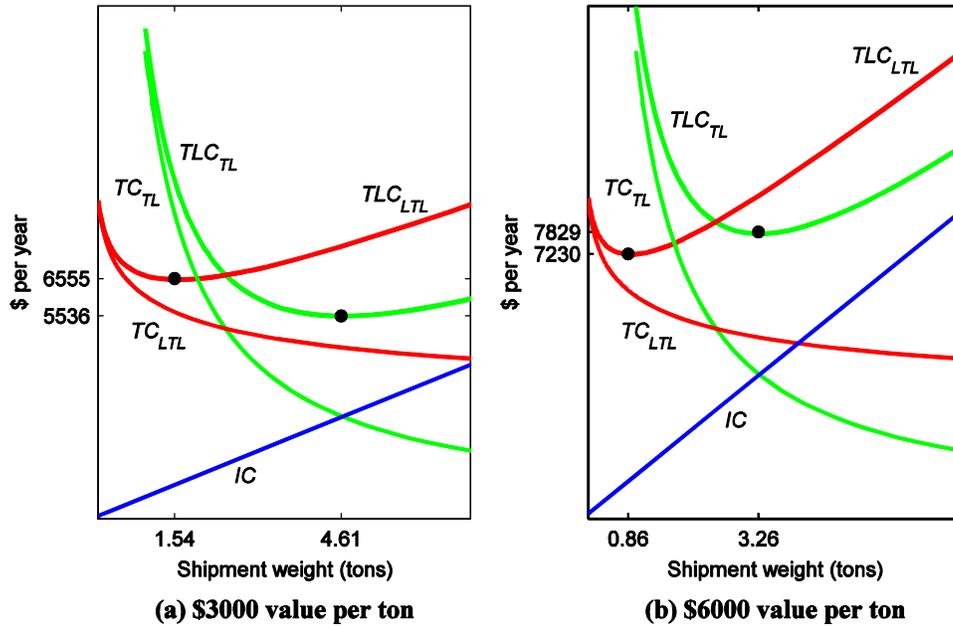
Figure 2.12. Total logistics cost comparison.

The total annual logistics cost for TL shipments is given by the sum of the transportation cost (*TC*) and the cycle inventory costs (*IC*) at the origin and destination, specifically

$$TLC_{TL}(q) = \frac{f}{q}\max\{c_{TL}, MC_{TL}\} + \alpha vhq \approx \frac{f}{q}rd + \alpha vhq. \tag{2.21}$$

Minimizing (2.21) with respect to q yields

$$q_{TL}^* = \min\left\{\sqrt{\frac{f\max\{rd, MC_{TL}\}}{\alpha vh}}, q_{max}\right\} \approx \sqrt{\frac{frd}{\alpha vh}}, \tag{2.22}$$

where the approximation ignores the minimum charge and maximum payload restrictions.

Note that in-transit inventory costs are ignored because, for distances greater than one-day's travel, single-driver TL and LTL transit times are approximately equal (while team drivers can be used for faster TL transit, the cost per mile increases because of the additional labor costs). Also, the increase in value associated with a load reaching its destination, which is at a minimum equal to the cost of transporting the load, is ignored. Finally, for the case discussed above where $\alpha = 0$, total logistics cost consists only of transportation costs, and technically, the value of q that minimizes (2.21) is $q = \infty$; by (2.22), however, this case results in $q_{TL}^* = q_{max}$.

Using a similar approach, for a given shipment density s and road distance d, such that $r_{LTL}(s,q,d)$ can be expressed simply as $r_{LTL}(q)$, the total annual logistics cost for LTL shipments is

$$TLC_{LTL}(q) = \frac{f}{q} \max\{c_{LTL}(q), MC_{LTL}\} + \alpha vhq = f \max\{r_{LTL}(q)d, MC_{LTL}/q\} + \alpha vhq$$

$$. \quad (2.23)$$

$$\approx \frac{f}{q} r_{LTL}(q)qd + \alpha vhq = f r_{LTL}(q)d + \alpha vhq$$

Expression (2.23) cannot be solved in closed-form to determine the optimal LTL shipment size and, instead, must be solved numerically (using, e.g., Solver in Excel or `fminsearch` in MATLAB):

$$q_{LTL}^* = \min\left\{\arg\min_q TLC_{LTL}(q), q_{max}\right\}. \quad (2.24)$$

Overall, therefore, the optimal independent shipment size is

$$q_0^* = \arg\min_q \left\{TLC_{TL}\left(q_{TL}^*\right), TLC_{LTL}\left(q_{LTL}^*\right)\right\}. \quad (2.25)$$

Note from the development leading up to (2.25) that, for a given lane and a given product (density and weight) at a given annual demand, the value of the load, v, becomes the critical factor in determining whether LTL or TL is the preferred mode since cycle inventory costs increase as the load value increases, but transportation costs remain unchanged. This situation is illustrated in Figure 2.12 for shipments during 2004 between Raleigh, NC and Gainesville, FL, with $f = 12$ tons per year, $d = 532$ mi, $s = 9.72$ lb/ft^3, $r = \$2$/mi, $K_{cu} = 3{,}000$ ft^3, $K_{wt} = 24$ tons, $h = 0.2$, and $\alpha = 1$. (In addition to *TLC*, inventory cost and transportation cost for each mode, labeled *IC* and *TC*, respectively, are shown in the figure.) At a value of $v = \$3{,}000$ per ton (Figure 2.12(a)), TL is preferred and the optimal shipment size is 4.61 tons. In contrast, at a value of $v = \$6{,}000$ per ton (Figure 2.12(b)), LTL is preferred and the optimal shipment size is 0.86 tons.

2.8 References

The following sources are recommended for further study:

Ballou, R.H., 2004, *Business Logistics/Supply Chain Management*, 5th Ed., Englewood Cliffs, NJ: Pearson.

Coyle, J.J., Novack, R.A., Gibson, B.J., and Bardi, E.J., 2011, *Transportation*, 7th Ed., South-Western.

Daganzo, C.F., 1999, *Logistics Systems Analysis*, 3rd Ed., Berlin: Springer.

Ravindran, A.R. and Warsing, Jr., D.P., 2013, *Supply Chain Engineering: Models and Applications*, Boca Raton, FL: CRC Press.

3. Economic Analysis

3.1 Introduction

The principle objective of a firm in a market economy is to maximize the profit associated with providing "products" (i.e., goods and/or services) to its customers (i.e., individuals, other firms, or the government). In particular, the objective of the firm is to maximize the difference between the revenue received from the sale of the product to customers and the cost of producing and delivering the product to customers. Profit can be maximized by increasing revenue and/or decreasing cost:

$$Profit(q) = Revenue(q) - Costs(q)$$

Revenue equals the price at which each unit of the product is sold times the quantity of product sold. Price and quantity are inversely related to each other and together constitute customers' demand for a product. It is not possible for a firm to select both price and quantity: at a particular price, the quantity of product that can be sold depends upon customers' demand, and, conversely, to sell a particular quantity of product depends upon the price at which the product is offered for sale. While a firm may be able to influence customers' demand for a product through, for example, advertising, thereby increasing revenue, the typical objective of the production-related activities of the firm is to meet a specified demand for the product at the lowest possible cost. Thus, production contributes to the overall goal of maximizing profits by minimizing the cost of producing the specified demand.

3.2 Costs

3.2.1 Fixed and Variable Costs

For a particular planning horizon, the input factors of the production process have costs that can be considered to be either *fixed* or *variable* depending on whether or not the factors themselves are fixed or variable. When a new production system is being planned, almost all inputs can be varied; once the locations of the facilities are selected, the cost of some of the inputs become

fixed (e.g., transportation costs of raw materials and finished products to/from the facilities, and the prevailing labor rate for the locations); and once the design of each facility is selected, the cost of even more of the inputs become fixed (e.g., material handling equipment costs).

Input factors can become fixed because they cannot be easily and/or economically changed in a short period of time due to the following reasons:

- *Transactions costs*—changing the input would incur too high a transactions cost (e.g., it is expensive to sell a facility if there is a temporary decrease in demand);

- *Specialized resources*—some inputs are specialized to a particular production process and would have little or no value to others (e.g., a custom machine that is only useful as part of a specific production process not used by others to produce the same products, or a worker with special skills);

- *Indivisible resources*—some inputs are "lumpy," that is, the cost of utilizing the input is the same over a range of outputs (e.g., the labor costs to set up a machine are the same whether one or 100 parts are produced).

Due to the creation of an ever greater number of fixed costs at each successive stage of designing a production system, the planning horizon should be long because the cost of making changes to a facility will increase at each stage—although at any particular stage, all previous costs should be considered to be "sunk costs" with respect their impact on current plans.
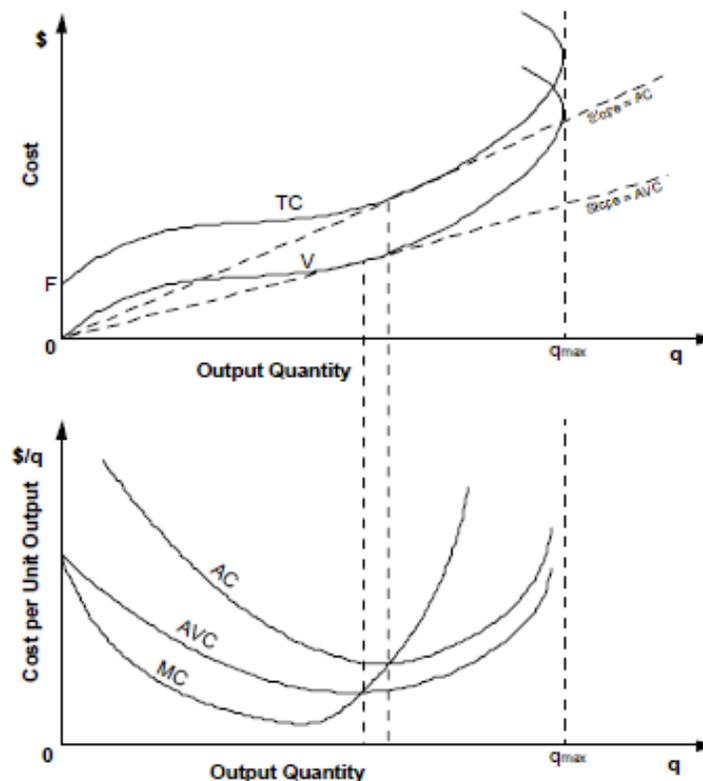

Figure 3.1. Total costs (top) and average costs (bottom) curves.

48

3.2.2 Total and Average Costs

In general, the total cost to produce (and distribute) a particular output quantity of a product is the sum of cost of each of the inputs. Letting F be the sum of fixed costs (of the fixed factors) and $V(q)$ be a function that represents the sum of the variable costs (of the variable factors) for different output quantities q, then the following functions of q represent related costs:

$$\textit{Total Cost:}\quad TC(q) = F + V(q) \cdot q \tag{3.1}$$

$$\textit{Average Cost:}\quad AC(q) = \frac{TC(q)}{q} = \frac{F + V(q) \cdot q}{q} = \frac{F}{q} + V(q) \tag{3.2}$$

$$\textit{Average Variable Cost:}\quad AVC(q) = V(q) \tag{3.3}$$

$$\textit{Marginal Cost:}\quad MC(q) = \frac{\Delta TC(q)}{\Delta q}\,(\text{discrete } q) = \frac{dTC(q)}{dq}\,(\text{continuous } q) \tag{3.4}$$

As shown in Figure 3.1, starting from the level of fixed costs F incurred independently of any output, the total cost function $TC(q)$ rises throughout as output increases, first possibly at a decreasing rate but ultimately at an increasing rate—until the maximum producible quantity q_{max} is attained. When one or more input factors are held fixed, increasing the other inputs will only increase the quantity produced up to a point (q_{max}) due to the "Law of Diminishing Returns," an observed physical law, and increasing these inputs further can even interfere with the production process enough to actually reduce the quantity produced.

Note: the minimum of an average cost curve can be determined geometrically from its corresponding total cost curve by noting that the slope (i.e., rise over run) of the line from the origin to a point along the total cost represents the average cost since $AC(q) = TC(q)/q$.

3.2.3 Long-Run and Short-Run Costs

If the price at which a firm can sell a product is less than the minimum attainable average variable cost AVC, a competitive firm will not produce at all; but at zero output and zero revenues, if the fixed cost F continues the firm will be incurring a loss (and will continue to incur a loss at any price below the minimum AC). The distinction between short-run and long-run costs provides an explanation of how a firm can stay in business when operating at a loss:

> *Long-run costs* $\Leftrightarrow$ all inputs can be varied $\Rightarrow$ only variable costs

> *Short-run costs* $\Leftrightarrow$ some inputs are fixed $\Rightarrow$ both variable and fixed costs

How long is the long run? The long run for a competitive firm is sufficient the time alter all of the inputs to a production process in response to a change in price, including, possibly, having the firm go out of business.

The long-run total cost (*LRTC*) and average cost (*LRAC*) functions represent the total and average cost, respectively, of producing any output q when all inputs are allowed to vary. The short-run total cost (*SRTC*) and average cost (*SRAC*) functions apply when one or more inputs are held fixed at a constant cost F.

The relationships between total short-run and long-run costs are shown in Figure 3.2 (top). The short-run total cost function $SRTC_1$ applies when the fixed factors are held constant at a level appropriate for small-scale production (q_1); similarly, $SRTC_2$ and $SRTC_3$ are associated with the higher levels of fixed cost appropriate for medium-scale (q_2) and large-scale (q_3) production, respectively. The corresponding average cost functions are shown in Figure 3.2 (bottom).
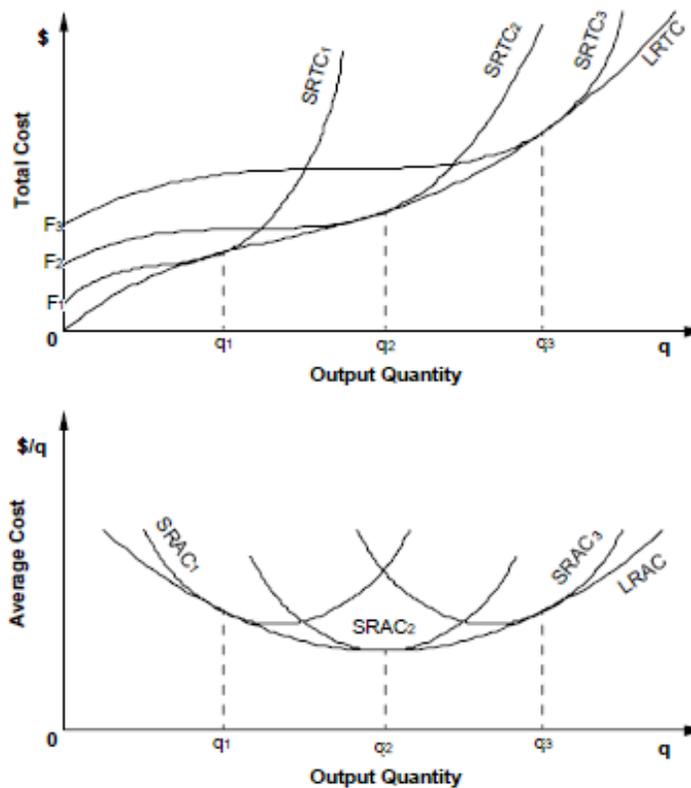

Figure 3.2. Long- and short-run cost curves: Total cost (top) and average cost (bottom).

In the absence of transactions costs, and if only unspecialized and divisible resources were used as inputs, there would be no difference between "long run" and "short run" costs; but, once a facility has been established, the presence of costs associated with the sale and purchase of inputs and the fact that some resources become specialized and/or indivisible results in fixed input factors and a corresponding distinction between the short run and the long run.

Since it is not possible to change fixed inputs immediately (or otherwise they would not be fixed), the response of a firm to a change in demand (and corresponding change in output quantity) or to a change in price depends on whether or not the change is regarded as temporary or permanent:

Temporary change in demand $\Rightarrow$ do not change fixed inputs $\Rightarrow$ stay on current short-run curve

Permanent change in demand $\Rightarrow$ change fixed inputs $\Rightarrow$ move to new short-run curve

Shutdown Decision

A firm makes a profit if a product can be sold at a price that equals or exceeds the cost of producing the product. Assume that a firm is currently making a profit producing a product and that its facility is the only fixed input factor involved in the production; all of the other factors are variable. In response to a change in the demand and/or price of the product in the market, the product can now be produced for *AC* (inclusive of facility costs) and *AVC* (exclusive of facility costs) and can be sold for *P*. If the change is temporary, then the costs and/or price will return to their previous values. Given the new *AC*, *AVC*, and *P*, the firm should make following decisions concerning the production of the product in response to the change:

$$P \geq AC \quad \Rightarrow \quad \textit{Continue} \text{ production}$$

$$AVC \leq P < AC \text{ and } \textit{Temporary} \text{ change} \quad \Rightarrow \quad \textit{Continue} \text{ production}$$

$$P < AVC \text{ and } \textit{Temporary} \text{ change} \quad \Rightarrow \quad \textit{Stop} \text{ production and } \textit{Keep} \text{ facility}$$

$$P < AC \text{ and } \textit{Permanent} \text{ change} \quad \Rightarrow \quad \textit{Stop} \text{ production and } \textit{Sell} \text{ (change) facility}$$

3.2.4 Product Cost

In order for a product to be profitable in the long run (i.e., so that a non-negative net present value is actually realized), a firm needs to recover all of the costs associated with providing the product. The revenues received from the sale of the product are used to recover these costs. The total cost to the firm of providing a product for sale has two principal components, with the majority of the costs of each component being incurred at different stages in the project's life cycle:

- *Investment costs.* Multi-period (typically multi-year) costs associated with acquiring the fixed assets used to provide the product.

  Investment costs, the majority of which are incurred prior to the start of actual production, are associated with the tangible fixed assets (e.g., land, buildings, and equipment) and the intangible fixed assets (e.g., R&D costs for basic research, product design, and production process design) used to provide the product.

- *Operating costs.* Single-period (typically less than a year) costs associated with acquiring the current assets used to produce the product.

  Operating costs, the majority of which are incurred during actual production, are associated with the current assets used to produce the product (e.g., raw materials, purchased components, labor, and facility and general administrative operating costs) and to provide the finished product to its customers (e.g., finished goods inventory, sales and marketing, transport, customer service, and warranty costs).

Although most of the investment costs are incurred prior to the start of the actual production of a product, they can (in most cases) only be recovered from the revenues generated from the final sale of the product to customers. For some products like pharmaceuticals, the costs associated

with R&D and sales and marketing can be many times greater than the product's actual production costs.

3.2.5 Break-Even Analysis

Given a particular production process for a product, the output quantity at which total costs equal total revenues is termed its "break-even point." Break-even analysis can be used to determine the expected profitability of a product at different output quantities so that preliminary decisions can be made concerning, for example, product pricing (which is possible if a firm is not a "price-taker" and has some control over a product's price, e.g., in a contract bid). If the product cost includes the full opportunity costs of all of the input factors used to produce the product, then the firm will breakeven (i.e., earn normal profits) at the break-even point; if all costs are not included (e.g., return on equity), an additional "profit margin" can be added to account for these costs.

Although the short-run total cost function ($SRTC$) for most products will not be linear over the entire range of possible outputs, increasing steeply as q_{max} is approached due to the Law of Diminishing Returns, there typically is some "relevant range" of outputs over which the function is approximately linear. Assuming that over a relevant range of outputs that the total cost (TC) to produce q units of output is $TC = F + V \cdot q$, where F is the fixed cost and V is the variable cost per unit (i.e., a linear average variable cost defined in Eq. (3.3), where $AVC = V(q) = V$), and that the total revenue (TR) received from the sale of q units (also a linear function over the relevant range—price would likely decrease for outputs beyond this range) is $TR = P \cdot q$, where P is per unit price for the product, then the break-even point (q_B) is the output quantity at which $TR = TC$:

$$TR = TC$$

$$P \cdot q = F + V \cdot q$$

$$(P - V)q = F$$

so that

$$\text{Break-Even Point:} \quad q_B = \frac{F}{P - V}. \tag{3.5}$$

In Figure 3.3, the relevant range of outputs is between q_{min} and q_{max}. The break-even points for Alternates 1 and 2, q_{B1} and q_{B2}, are within the relevant range; the break-even point for Alternative 3 is at zero output since the purchase price, V_3, is less than the sale price, P.

3.2.6 Cost Indifference Analysis

Cost indifference analysis can be used to make make-or-buy and how-to-make decisions. Given two different alternatives for providing a product (or a component part of a product), the output quantity at the total costs for the two alternatives are equal is termed the "cost indifference point." *Make-or-buy decisions* arise when one of the alternatives is to purchase (buy) the product and the other alternative is produce (make) the product. If the alternatives are two different

processes for producing the product, then *how-to-make decisions* can be made concerning which process is less costly at different output quantities. How-to-make type decisions arise whenever there is a need to select between alternatives that have different fixed and variable costs (e.g., many material handling equipment selection decisions).
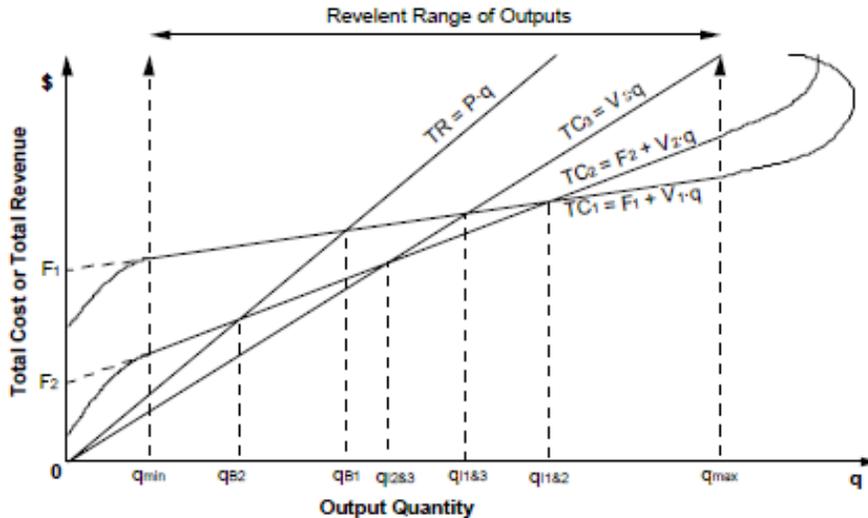

Figure 3.3. Break-even points and cost indifference points for Alternative 1 (make), Alternative 2 (make), and Alternative 3 (buy).

Given fixed costs of F_1 and F_2 and variable costs of V_1 and V_2 for Alternative 1 and Alternative 2, respectively, and assuming total costs are linear over the relevant range, then the cost indifference point is the output quantity ($q_{I1\&2}$) at which $TC_1 = TC_2$:

$$TC_1 = TC_2$$

$$F_1 + V_1 \cdot q = F_2 + V_2 \cdot q$$

$$F_1 - F_2 = (V_2 - V_1) q$$

so that

$$\textit{Cost Indifference Point:} \quad q_{I1\&2} = \frac{F_1 - F_2}{V_2 - V_1}. \qquad (3.6)$$

If one of the alternatives is to purchase the product (Alt. 3 in Figure 3.3), then the fixed cost for that alternative will be zero and its variable cost will be its purchase price. If the variable costs of both alternatives are equal, then the cost indifference point is undefined (division by zero) and the alternative with the lowest fixed cost will have the lowest total cost.

In Figure 3.3, between outputs q_{min} and $q_{I2\&3}$, the product should be purchased (i.e., Alternative 3); between $q_{I2\&3}$ and $q_{I1\&2}$, the product should be produced using the process defined by Alternative 2; and between $q_{I1\&2}$ and q_{max}, the product should be produced using the process defined by Alternative 1.

Operating leverage

Everything else being equal between two alternatives, the alternative with larger fixed costs provides a greater degree of *operating leverage* and associated business risk: an increase in the output quantity beyond the cost indifference point between the two alternatives will result in greater profits for the alternative with greater operating leverage, while a decrease in the output quantity below their cost indifference point will result in a greater reduction in profits (or, possibly, increase in losses).

3.3 Discounting

The discounting process involves the use of the following compound interest formulas. Let

$N =$ number of compounding periods

$i =$ effective interest or discount rate (in decimal form) per compounding period

$FV =$ future value of a cash flow

$PV =$ present value of a cash flow or uniform series of cash flows

$C_t =$ cash flow occurring at the end of period t

Given N periods, the index t can range from 0 to N, where $t = 0$ corresponds to the beginning of the first period (or the end of period 0), $t = 1$ corresponds to the end of the first period (and the beginning of the second period), and $t = N$ corresponds to the end of the Nth period.

3.3.1 Basic Formulas

Future value

For $t = 0$, $FV = PV$; for $t = 1$, $FV = PV(1 + i)$; for $t = 2$, $FV = PV(1 + i)(1 + i) = PV(1 + i)^2$; and, in general, the future value at the end of period t of the present value of a single cash flow (or payment) is

$$\text{Future value of single payment:} \quad FV = PV\left(FV/PV, i\%, t\right) = PV\left(1+i\right)^t. \qquad (3.7)$$

Present value

Using (3.7) and solving for PV, the present value of a single payment occurring at the end of period t is

$$\text{Present value of single payment:} \quad PV = FV\left(PV/FV, i\%, t\right) = FV\left(1+i\right)^{-t}. \qquad (3.8)$$

Net present value

The net present value of a series of cash flows C_t, $t = 0, 1, \ldots, N$, is, using (3.8),

$$\text{Net present value: } NPV = \sum_{t=0}^{N} C_t \left(PV/FV, i\%, t \right) = \sum_{t=0}^{N} C_t \left(1+i \right)^{-t}. \tag{3.9}$$

Uniform series of cash flows

If C is the cash flow per period of a uniform series of N end-of-period cash flows (also termed an "annuity"), where $C = C_t$, $t = 1, \ldots, N$, then, using (3.8),

$$PV = \sum_{t=1}^{N} C \left(PV/FV, i\%, t \right) = C \sum_{t=1}^{N} \left(1+i \right)^{-t}$$

$$= C \left[\left(1+i \right)^{-1} + \cdots + \left(1+i \right)^{-N} \right]. \tag{3.10}$$

Multiplying (3.10) by $(1 + i)$,

$$PV \left(1+i \right) = C \left[\left(1+i \right)^{-1} + \cdots + \left(1+i \right)^{-N} \right] \left(1+i \right)$$

$$= C \left[1 + \left(1+i \right)^{-1} + \cdots + \left(1+i \right)^{-N+1} \right], \tag{3.11}$$

and subtracting (3.10) from (3.11) results in

$$PV \left(1+i \right) - PV = C \left[1 - \left(1+i \right)^{-N} \right],$$

which, assuming $i \neq 0$ and solving for PV, reduces to

$$\text{Present value of uniform series: } PV = C \left(PV/C, i\%, N \right) = C \left[\frac{1 - \left(1+i \right)^{-N}}{i} \right], \quad i \neq 0. \tag{3.12}$$

If $i = 0$, then (3.10) reduces to $C \cdot N$, which can also be derived via (3.12) through the use of l'Hôpital's rule:

$$\lim_{i \to 0} \frac{1 - \left(1+i \right)^{-N}}{i} = \lim_{i \to 0} \frac{N \left(1+i \right)^{-N-1}}{1} = N.$$

The present value of the uniform series of cash flows in (3.12) does not include a cash flow for the beginning of period 0 (i.e., $t = 0$). If the cash flow C_0 occurring at the beginning of the first period is added to the uniform series of N cash flows C occurring at the end of each period, then the net present value is

$$NPV = C_0 + C \left(PV/C, i\%, N \right). \tag{3.13}$$

Using (3.12) and solving for C, the capital recovery cost of a single cash flow (or payment) PV occurring at the beginning of the first period is

$$\begin{array}{l}\textit{Capital recovery cost} \\ \textit{of single payment}\end{array} : \quad C = PV\left(C/PV, i\%, N\right) = PV\left[\dfrac{i}{1-\left(1+i\right)^{-N}}\right], \quad i \neq 0. \qquad (3.14)$$

If $i = 0$, then (3.14) reduces to PV/N.

3.3.2 Effective Cost vs. Capital Recovery Cost

"Effective cost" and "capital recovery cost" represent two alternate ways of determining the full opportunity cost of an investment in fixed assets (see Figure 3.4). They represent, respectively, either the net present value or the uniform per-period cost associated with the loss in value (or depreciation) of the fixed assets and the cost of the capital used to finance the assets (i.e., the interest on long-term debt and the required return on equity). In facilities planning, they can be used to equate the (multi-period) investment costs of fixed assets with the (per-period) operating costs of the assets.
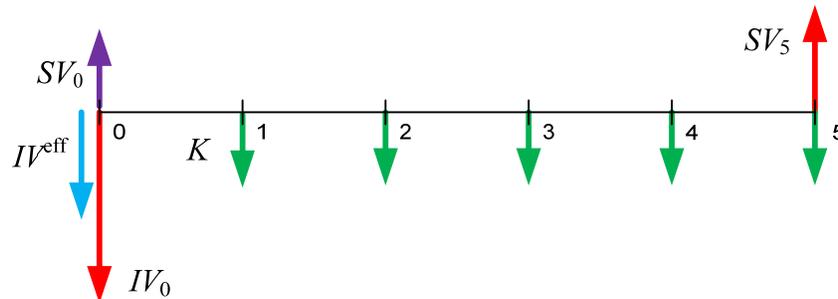

Figure 3.4. Cash-flow diagram comparing IV^{eff} to K.

Given a cost of capital of i per period, the "effective cost" (IV^{eff}) for fixed assets with an initial investment cost of IV and a salvage value of SV at the end of N periods is equal to the initial investment cost IV minus the present value (3.8) of the salvage value SV:

$$\textit{Effective cost:} \quad IV^{\text{eff}} = IV - SV\left(PV/FV, i, N\right) = IV - SV\left(1+i\right)^{-N}. \qquad (3.15)$$

If $SV = 0$, then the effective cost of the fixed assets is equal to their initial investment cost.

Using the capital recovery cost of a single payment (3.14) to convert the effective cost (3.15) to an equivalent uniform series of end-of-period costs, the capital recovery cost per period (K) of the fixed assets is equal to the following:

$$\begin{aligned}\textit{Capital recovery cost:} \quad K = IV^{\text{eff}}\left(C/PV, i, N\right) \quad &= IV^{\text{eff}}\left[\dfrac{i}{1-(1+i)^{-N}}\right] \\ &= \left(IV - SV\right)\left[\dfrac{i}{1-(1+i)^{-N}}\right] + SV \cdot i\end{aligned} \qquad (3.16)$$

where $K = IV^{\text{eff}} \left(C/PV, i, N\right) = \left[IV - SV(1+i)^{-N}\right]\left[\dfrac{i}{1-(1+i)^{-N}}\right]$

$$= IV\left[\dfrac{i}{1-(1+i)^{-N}}\right] - SV\left[\dfrac{i(1+i)^{-N}}{1-(1+i)^{-N}}\right]$$

$$= IV\left[\dfrac{i}{1-(1+i)^{-N}}\right] - SV\left[\dfrac{i-i\left(1-(1+i)^{-N}\right)}{1-(1+i)^{-N}}\right]$$

$$= (IV - SV)\left[\dfrac{i}{1-(1+i)^{-N}}\right] + SV \cdot i$$

$$= (IV - SV)\left(C/PV, i, N\right) + SV \cdot i \,.$$

If $SV = 0$, then $K = C$, the capital recovery cost of the single payment IV.

The capital recovery cost (3.16) provides the following insights concerning investments:

1. If $IV = SV$, then (3.16) reduces to $SV \cdot i$, which is equal to $IV \cdot i$ since $IV = SV$. Since the initial investment is fully recovered and not reduced in value (e.g., working capital, bonds, and most land), the capital recovery cost is just the opportunity cost of the funds invested, as represented by the cost of capital (e.g., the interest payments on a bond, the principal of which is repaid at the bond's maturity date).

2. As $N \to \infty$, then (3.16) reduces to $IV \cdot i$. Since the salvage value will never be realized, the capital recovery cost is equal to the initial investment cost times the capital recovery cost in perpetuity (see Eq. (3.18) in Sec. 3.3.3).

3. If IV represents the amount of an amortized loan (e.g., a home mortgage), where $SV = 0$ from the lender's point of view, then the capital recovery cost is the amount of loan times the capital recovery cost of the single payment (3.14) from the lender to the borrower; it represents the payment required per period from the borrower to the lender to cover principal and interest on the loan.

3.3.3 Further Issues in Discounting

Perpetuity

A "perpetuity" is an infinite series of uniform end-of-period cash flows. If $i > 0$, then $(1+i)^{-N} \to 0$ as $N \to \infty$ so that, using (3.12), the present value of C, the cash flow per period of the perpetuity, is

$$\textit{Present value of perpetuity:} \quad PV = C\left(PV/C, i\%, \infty\right) = \lim_{N \to \infty} C\left[\dfrac{1-(1+i)^{-N}}{i}\right] = \dfrac{C}{i}. \quad (3.17)$$

Likewise, taking the limit of (3.14) as $N \to \infty$ or solving (3.17) for C, the capital recovery cost in perpetuity of PV is

$$\text{Capital recovery cost in perpetuity: } C = PV \left(C/PV, i\%, \infty \right) = PV \cdot i . \qquad (3.18)$$

Nominal interest rates

A "nominal interest rate" is an interest rate specified for a basis period (e.g., a year) that (1) can be different from the compounding period (e.g., a month) and, if different, (2) does not include the effects of inter-compounding-period discounting during the basis period. If i_{nom} is the nominal interest rate per basis period, there are N_{nom} basis periods, and there are M compounding periods during each basis period, then i, the effective interest rate per compounding period, and N, the number of compounding periods, are as follows:

$$i = \frac{i_{\text{nom}}}{M} \quad \text{and} \quad N = M \cdot N_{\text{nom}}$$

The nominal interest rate is equal to the effective interest rate if the basis period is the same as the compounding period.

The most common basis period for specifying nominal interest rates is a year. For example, 12% compounded monthly is usually understood to refer to a nominal interest rate of 12% per year, corresponding to an effective interest rate of 1% per month (or $(1+0.01)^{12} - 1 = 12.68\%$ per year, which is also termed the "annual percentage rate" (APR)).

Nonintegral number of compounding periods

When a cash flow does not occur at the beginning or end of a compounding period, it is possible to use a nonintegral number of compounding periods to discount the cash flow. If N is a nonintegral number of compounding periods such that $N = \lfloor N \rfloor + \lambda$, where $\lfloor N \rfloor$ is the integral portion and λ, $0 < \lambda < 1$, is the fractional portion of N, then N corresponds to $\lfloor N \rfloor$ complete compounding periods and λ of the ($\lfloor N \rfloor + 1$)th period. For example, the future value FV of an amount PV after N periods is the following:

$$FV = PV \left(FV/PV, i\%, \lfloor N \rfloor \right) \left(FV/PV, i\%, \lambda \right) = PV \left(1+i \right)^{\lfloor N \rfloor} \left(1+i \right)^{\lambda}$$

$$= PV \left(1+i \right)^{\lfloor N \rfloor + \lambda} = PV \left(1+i \right)^{N} = PV \left(FV/PV, i\%, N \right).$$

3.4 Capital Project Evaluation

3.4.1 Introduction

Capital project evaluation is the process of planning investments in fixed assets. In facilities planning, it provides a formal means of evaluating alternative facility plans. *Fixed assets* (also termed "capital assets" or "capital goods" or just "capital") are assets whose benefits or returns

are expected to extend beyond one year, while *current assets* (e.g., cash, accounts receivable, and inventories) are assets whose returns are expected within a year—the investment in current assets is termed "working capital." Fixed assets can either be tangible (e.g., the land, buildings, and equipment that comprise a facility) or intangible (e.g., research and development (R&D) expenses for a product).

Since fixed assets provide returns for multiple years into the future, the discounting process is used in capital project evaluation to recognize the time value or opportunity costs of the investment funds used to acquire the fixed assets. Discounting gives explicit consideration to the fact that, all else being equal, a dollar received immediately is preferable to a dollar received at some future date. Discounting is not typically an important issue in current asset analysis, and is usually ignored.

Project interdependence

In capital project evaluation, each separate configuration of fixed assets to be considered for *acceptance* (i.e., selection or acquisition) is termed a *project*. Associated with each project are the costs and benefits (i.e., negative and positive cash flows, respectively) that would occur if the project is accepted. If only a single project is being considered, the project should be accepted only if its benefits meet or exceed its costs; if more than one project is being considered, the degree of interdependence between the projects has to be taken into account in the capital project evaluation. The possible degrees of interdependence between two projects, A and B, are as follows:

- *Prerequisite*—project A is possible only if project B is accepted;

- *Complement*—the benefits (costs) of project A would increase (decrease) if B is accepted;

- *Independent*—project A is not affected by the acceptance of project B, that is, accepting one project does not influence the acceptance of any other project (except with respect to the availability of funds);

- *Substitute*—the benefits (costs) of project A would decrease (increase) if B is accepted;

- *Mutually exclusive*—project A is not possible if project B is accepted, that is, accepting one project implies rejecting the other.

The same type of capital project evaluation can be used to determine whether one or more independent projects should be accepted, or to select a single project from among several mutually exclusive projects (including combinations of prerequisite projects); a more complicated analysis is required for projects that are complements or substitutes of each other.

In facilities planning, capital project evaluation can be used to determine which facilities, each at different locations and/or producing different products, should be built (independent projects), or to select a single design to build a facility at a site from several alternate designs that are available for the facility (mutually exclusive projects). If a project (A) is a prerequisite of another project (B), then both projects together can be considered to be a single project (A&B) that is mutually exclusive with respect to the other (A) by itself. Since in a capital project evaluation it can be difficult to quantitatively determine all of the combinations of costs and benefits between

projects that are complements or substitutes of each other, it may be reasonable to assume that the projects are either independent or mutually exclusive; in this case, a qualitative description of the interdependencies between the projects should, at least, be provided in addition to the quantitative analysis.

Project categories

Facility-related projects are frequently classified into the following categories:

- *Maintenance*—expenditures to maintain the current level of performance of a facility;

- *Cost reduction*—expenditures to lower the cost of producing the current level of output at a facility;

- *Expansion*—expenditures to increase the future level of output of a facility;

- *Construction*—expenditures to establish a new facility;

- *Mandatory*—expenditures required at a facility to comply with, for example, government regulations, insurance requirements, or labor agreements.

Different analysis procedures are often used for the different categories of projects. The cost of capital (i.e., discount rate) and the level of detail of the analysis used for each project category are related to the magnitude of the expenditures and the uncertainty of the costs and benefits associated with the typical projects that occur in each category. Little or no analysis may be required for routine maintenance projects or mandatory projects (that, if avoided, would result in large fines or possible shutdown of the facility); simple procedures like the "payback method" may be used for small-scale cost reduction projects suggested by the engineering staff and approved by the plant manager at the facility level; while formal capital project evaluation procedures may be required for large-scale expansion projects and new facility construction, with final approval at the corporate headquarters level.

Incremental cash flows

In capital project evaluation, a project is evaluated based on the magnitude and timing of the incremental (explicit and implicit) cash flows that would be induced if the project were to be accepted. For a profit-maximizing firm in a market economy, cash outflows and inflows, as signaled through prices, provide the ultimate measure of the costs and benefits (to the firm) associated with engaging in any economic activity. The benefit to the firm from the production of a good or service is the revenue (i.e., cash receipts) received from its sale at a price in the market. The cost to the firm to produce a good or service is the sum of the payments (i.e., cash expenditures) required to attract the input factors or resources necessary for its production. The payments required are the market prices of the resources and are made to the owners of the resources. The market price of a resource represents the amount sufficient for the owner of the resource to forgo the opportunity to receive the benefits associated with the best alternative use of the resource.

3.4.2 Projects with Uniform Cash Flows

When all of the per-period cash flows of each project are uniform, the capital project evaluation can be performed using either the payback method, which does not require discounting any of the cash flows, or a single formula can be used to discount the uniform series cash flows to determine either the one-time net present value (NPV) or per-period net annual value (NAV) (see Figure 3.5). When the cash flows are not uniform, each period's cash flow much be separately discounted and then combined together, a more complex calculation that is usually performed in a spreadsheet.

Payback Method

The payback method is a simple means of evaluating projects. It is typically used for evaluating small cost-reduction projects. The *payback period* of a project is the number of periods it takes to recover the initial investment expenditure from the project's future (undiscounted) positive net cash flows. The payback period of a project can be used in place of determining the net present value of the project. If the project provides a uniform series of positive end-of-period operating profits, then the payback period of the project is

$$Payback\ period = \frac{IV_0}{OP}, \quad \text{for } OP > 0. \tag{3.19}$$

where

$$OP = \begin{cases} OR - OC, & \text{uniform operating profit per period from the project} \\ OC_{\text{current}} - OC_{\text{new}}, & \text{net uniform operating cost } savings \text{ per period} \end{cases}$$

OR = uniform operating revenue per period from the project

OC = uniform operating cost per period of the project

IV_0 = initial net investment expenditure at time 0 for the project.

In many process improvement projects, IV_0 is the net additional investment required to replace the current process with a new process and OP is the net operating cost savings since net operating revenue is zero as long as the sales price is assumed to remain the same; as a result, it is not even necessary to know the actual sales price, which is convenient in many situations since it may be difficult or impossible to assign the output of a single machine to the final sales price to a product that may have several more processing steps at multiple locations before it is sold to a final customer.

The principal advantages of using the payback method to evaluate projects are the following:

1. It is (arguably) easier to understand and explain to others;

2. There is no need to determine a cost of capital;

3. It gives the same relative ranking of mutually exclusive projects as the discounting approach if the projects all have the same OP and all $SV = 0$.

61

For small cost-reduction projects at large firms, there is typically a fixed budget for these types of projects with no implied cost of capital associated with the budget. The principal disadvantages of using the payback method are the following:

1. It ignores discounting (i.e., assumes a 0% cost of capital);

2. It ignores salvage values;

3. It ignores cash flows that occur beyond a project's payback period.

The third disadvantage is not a problem if, as assumed in (3.19), the project provides a uniform series of returns and $SV = 0$. Cost-reduction projects (the bread and butter of most industrial engineers) typically provide uniform returns because, once the project is implemented, their returns (e.g., the cost savings associated with a process improvement) continue to be realized at a constant magnitude until the time at which the production process again changes.

Projects will long payback periods are usually considered risky because more distant cash flows are usually more uncertain. In addition, a firm may not be in business at the point in time at which the returns beyond a project's payback period could be realized. As a rule of thumb:

• small cost-reduction projects with payback periods not exceeding two or three years are usually considered good candidates for acceptance;

• projects with payback periods of a year or less are almost always accepted and given high priority;

• projects with payback periods of greater than two or three years are usually not accepted (or maybe given further consideration only after a more detailed analysis).

(a) Actual cash flows.

(b) Payback method.

(c) Net present value (NPV).

(d) Net annual value (NAV).
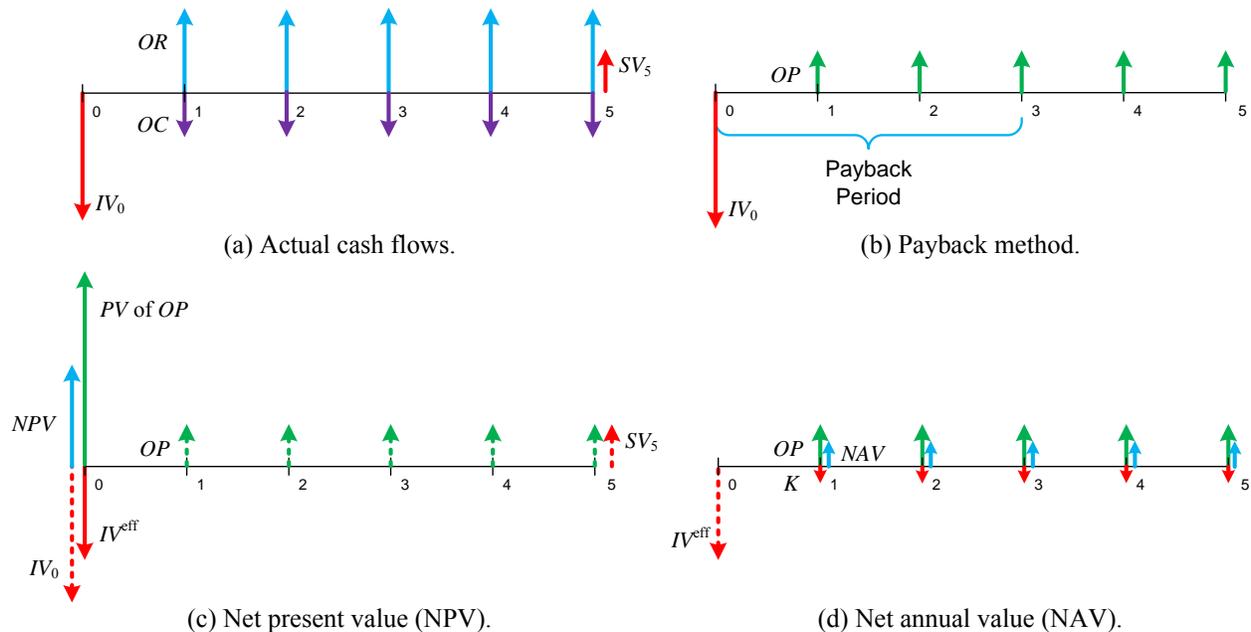
Figure 3.5. Alternative methods of evaluating a project that has uniform per-period cash flows.

Discounting Methods

Net present value (NPV) and net annual (periodic) value (NAV) are two equivalent methods for evaluating projects with uniform cash flows. Each involves discounting all of the cash flows, and differs with respect to whether all of the cash flows are converted to *one-time* cash flows (NPV) or *per-period* cash flows (NAV) (note: since the period used in NAV is typically one year, it is termed net "annual" value—also, net periodic value would have the same abbreviation as NPV.):

$$NPV = PV \text{ of } OP - IV^{\text{eff}}$$

$$\text{Net Present Value:} \qquad = OP\left[\frac{1-(1+i)^{-N}}{i}\right] - IV^{\text{eff}}, \quad i \neq 0 \tag{3.20}$$

$$\text{Net Annual (Periodic) Value:} \quad NAV = OP - K. \tag{3.21}$$

An independent project should be accepted if its $NPV \geq 0$ or its $NAV \geq 0$. If two mutually exclusive projects are being considered, then the project with the largest non-negative value should be selected.

3.4.3 Cost Reduction Projects and Average Cost

If one (current) project already exists and a new project is being considered that might reduce costs, then it is possible to make a decision about whether the new project should replace the current project by using the incremental reduction in operating costs as "revenue" in the economic analysis. This is convenient because the actual revenue does not have to be known, and is valid as long as the new project does not change the actual revenue and that the current project was previously accepted as an independent project.

Note: In evaluating projects, NPV and NAV provide equivalent results, but in determining the average cost AC for a project that involves producing q units of a product per period, $AC = -NAV/q$, assuming $OR = 0$. It is incorrect to divide $-NPV$ by q times N, the life of the project.

Example

A firm has just purchased a widget machine for $2 million. It is expected to have a salvage value of 25% of its initial investment cost at the end of 15 years, and an operating cost of $1.25 per unit. Before it starts operations, the firm has become aware of the availability of a more automated machine that can be purchased for $5 million and is otherwise identical to the current machine except that it will result in an operating cost savings of $0.75 per unit. The current machine can be sold today for $2 million, annual demand is expected to be 500,000 units, and the real cost of capital is 8% compounded annually.

The following questions can be considered in Table 3.1: (a) What is the payback period associated with purchasing the more automated machine? 8 years. (b) Should the more automated machine be purchased? Yes, since the net NPV and the net NAV are non-negative.

(c) What would be the average cost to produce each widget using either the current or the more automated machine? $1.64 and $1.58, respectively. In this example, the operating cost savings becomes the net operating profit. Also, note that the net salvage value associated with selling the current machine and building the new one is $750,000 because the $500,000 salvage value of the current machine after 15 years will not be realized if it is sold today.

Table 3.1. Cost Reduction Example

Common				
Cost of Capital	(i)	8%	8%	
Economic Life	(N, yr)	15	15	
Annual Demand	(q/yr)	500,000	500,000	
Sale Price	($/q)			
Project		**Current**	**New**	**Net**
Investment Cost	(IV, $)	2,000,000	5,000,000	3,000,000
Salvage Percentage		25%	25%	
Salvage Value	(SV, $)	500,000	1,250,000	750,000
Eff. Investment Cost	(IV^{eff}, $)	1,842,379	4,605,948	2,763,569
Cost Cap Recovery	(K, $/yr)	215,244	538,111	322,866
Oper Cost per Unit	($/q)	1.25	0.50	(0.75)
Operating Cost	(OC, $/yr)	625,000	250,000	(375,000)
Operating Revenue	(OR, $/yr)	0	0	0
Operating Profit (OR - OC)	(OP, $/yr)	(625,000)	(250,000)	375,000
Analysis				
Payback Period (IV/OP)	(yr)			8.00
PV of OP	($)	(5,349,674)	(2,139,870)	3,209,805
NPV (PV of OP - IV^{eff})	($)	(7,192,053)	(6,745,818)	446,236
NAV (OP - C^{CR})	($/yr)	(840,244)	(788,111)	52,134
Average Cost ((K + OC)/q)	($/q)	1.68	1.58	

3.4.4 Projects with Nonuniform Cash Flows

The same type of capital project evaluation can be used to determine whether one or more independent projects should be accepted, or to select a single project from among several mutually exclusive projects (including combinations of prerequisite projects); a more complicated analysis is required for projects that are complements or substitutes of each other.

Given M projects, the basic elements used for a capital project evaluation of the projects are the following:

$i =$ weighted average cost of capital (effective interest or discount rate in decimal form) per compounding period at which the cash flows associated with the projects are to be discounted;

N_j = expected life (number of compounding periods) of project j;

IV_j = investment costs, where IV_0 is the initial investment expenditure (negative cash flow) to acquire at beginning of period 0 (or time 0) the fixed assets and working capital associated with project j, and $SV_j = -IV_{N_j}$ is the salvage value (positive or negative cash flow) received at end of period N_j from the disposal of the fixed assets and recovery of working capital associated with project j;

OR_{jt} = operating revenue (or receipts) exclusive of SV_j that would occur during period t if and only if project j is accepted;

OC_{jt} = operating costs (or expenditures) that would occur during period t if and only if project j is accepted, excluding interest expenses of any long-term debt used to finance the investment in the project;

C_{jt} = net cash flow of project j at the end of period t.

$$= OR_{jt} - OC_{jt} - IV_{jt}, \quad t = 0, \ldots, N_j$$

Assumptions

The following assumptions are made to simplify the presentation.

1. The cost of capital i is the same for all projects under consideration.

2. IV_j and OC_{jt} are specified as non-negative values (and, thus, are subtracted in determining the net cash flow).

3. All of the initial investment expenditures associated with each project j are incurred at time 0, the start of the project; if these expenditures are incurred over several periods, then IV_j should represent their present value at time 0.

4. Any and all of the salvage value from each project j is received at the end of period N_j; if the salvage value is received over several periods, then SV_j should represent their future value at the end of period N_j.

5. All projects under consideration are either all independent or all mutually exclusive.

NPV of project

Using Eq. (3.9), the net present value of project j can be determined as follows:

$$NPV_j = \sum_{t=0}^{N_j} C_{jt} \left(PV/FV, i\%, N_j \right) = \sum_{t=0}^{N_j} C_{jt} \left(1 + i \right)^{-t}. \tag{3.22}$$

If project j is accepted, the value of the firm will increase by the amount NPV_j.

Cost of Capital (i) =	0.20					
			Project A			
Period	Oper Rev	Oper Cost	Invest Cost	Net Cash Flow	PV/FV	Present Value
(N)	(OR)	(OC)	(IV)	(C)	$(1/(1+i)^\wedge N)$	$C*(1/(1+i)^\wedge N)$
0	$0	$0	$1,000,000	($1,000,000)	1.0000	($1,000,000)
1	500,000	300,000	0	200,000	0.8333	166,667
2	1,000,000	500,000	0	500,000	0.6944	347,222
3	1,000,000	500,000	0	500,000	0.5787	289,352
4	1,000,000	500,000	0	500,000	0.4823	241,127
5	500,000	375,000	(250,000)	375,000	0.4019	150,704
				Net Present Value (NPV) =		$195,071

Figure 3.6. Use of a spreadsheet to calculate the NPV of a project.

Using a spreadsheet

In Figure 3.6, a spreadsheet is used to calculate NPV_A = $195,071, the net present value of Project A. In the spreadsheet, relative addressing is used for all the calculations except for the cost of capital, 0.20, which is referenced as an absolute address (in MS Excel, it is referenced as C1 (third column, first row)).

Project acceptance criteria

Given M projects that are either all independent or all mutually exclusive, determine, using (3.22), the net present value NPV_j of the cash flows associated with each project j, j = 1, …, M. Projects can then be accepted based on the following criteria.

1. Independent projects: Accept all projects with a non-negative NPV_j; reject all projects with a negative NPV_j.

2. Mutually exclusive projects with equal lives (all N_j equal): Accept the project with the largest non-negative NPV_j; reject all other projects. (Reject all projects if all NPV_j's are negative.)

3. Mutually exclusive projects with different lives (all N_j not equal): Accept the project with the largest non-negative equivalent uniform end-of-period net cash flow C_j, where $C_j = NPV_j \left(C/PV, i\%, N_j \right)$; reject all other projects. (Reject all projects if all C_j's are negative.)

3.4.5 Issues in Capital project evaluation

Opportunity costs

All costs are opportunity costs: the cost of any activity is the loss of the opportunity to receive the benefits associated with the best alternate activity available. If the full opportunity costs of a project are used to determine its net present value NPV, then any amount $NPV > 0$ represents the "economic profit" possible from the project; a $NPV = 0$ represents a project with "normal profit."

For a profit-maximizing firm, the opportunity costs of a project are both

 • the cash outflows that will occur, if and only if the project is accepted, and

- the cash inflows that will not occur, if and only if the project is accepted.

While it is usually not too difficult to identify the "explicit" cash flows associated with a project, it can be much more difficult to identify the "implicit" cash flows that will or will not occur as a result of the project's impact on the existing resources and the future activities of the firm not directly related to the project under consideration. (The term "opportunity costs" is sometimes used to refer to just "implicit costs" in order to distinguish these costs from the "explicit costs" considered in the context of accounting.)

For example, if a project will be using space in a facility that is currently empty, then the opportunity cost of using the space might be zero if the space would otherwise remain empty during the life of the project; but if it possible to rent the space to others or use it for future projects, then the (implicit) opportunity costs induced by accepting the project would be the rental receipts forgone or the costs of expanding the facility to accommodate future projects. If several mutually exclusive projects are being considered that represent all current and future uses of the space (including renting or selling the space), then the opportunity cost of using the space for one particular project is difference between the maximum net present value available from all of the projects and the net present value of the particular project.

Sunk costs

Sunk costs are cash outflows that have either already occurred and cannot be recovered, or have not yet occurred but cannot be avoided. As such, sunk costs should not be considered in capital project evaluation decisions. Cash flows that have already occurred (e.g., past purchases) or cannot be avoided (e.g., future lease payments) have no impact on the incremental cash flows that will occur if a project currently under consideration is accepted. In the short run, the portion of fixed costs that are not recoverable are sunk costs. The only impact that the project can have on the existing fixed assets of a firm is through a change in their salvage value that would occur as a result of the project being accepted. For example, if a project can utilize an existing machine that currently has excess capacity, the opportunity cost of using the machine is any reduction in the future salvage value of the machine; the initial investment expenditures that have already occurred to acquire the machine are sunk costs and should not affect the decision to accept the current project under consideration.

Cost of capital

The cost of capital (also termed the "minimum acceptable rate of return") used in capital project evaluation should be the weighted average of the cost of the different types of funds a firm uses to finance investments (e.g., the after-tax interest rate on new debt and the cost of equity, weighted by the proportion of debt and equity in a firm's capital structure). The cost of capital for a project represents the opportunity cost of the funds invested in the project—if the funds are not invested in the project, they can be invested in other ways to provide revenues at least equal to the cost of capital. The cost of capital can be adjusted upwards or downward to reflect the riskiness of a project.

For example, if a firm's investment funds are obtained using 50% long-term debt, at an after-tax interest rate of 7% per year, and 50% equity, at 13% per year, then the weighted cost of capital is

$(0.5)\cdot 7+ (0.5)\cdot 13= 10\%$. If \$10 of revenue will be received per year from \$100 of investment (composed of \$50 debt and \$50 equity), then $\$50\cdot(0.07)= \$3.50$ is the interest on the debt and $10- 3.50= \$6.50$ is the return on equity.

Timing of cash flows

Many of the cash flows induced by a project may not occur at the beginning or end of the time period used in the capital project evaluation. If cash flows occur a multiple times throughout a period (e.g., monthly sales revenues for a year time period), then it is common practice to use to as an end-of-period amount the simple sum (i.e., without in-period discounting) of the cash flows that occur during the period (e.g., the sum of the monthly sales revenues). If a cash flow occurs at a single point in time, then either the beginning or end of the closest time period, or a nonintegral number of compounding periods, can be used.

Projects with different lives

When two or more mutually exclusive projects have different lives, it is not appropriate, in general, to base the acceptance decision on just NPV because projects with longer lives would have more time to accrue positive cash flows, thereby increasing their net present values and biasing the capital project evaluation decision in their favor. To eliminate this bias, projects with uniform cash flows should evaluated using NAV and, for projects with nonuniform cash flows, the *NPV_j* given in (3.22) should be converted to equivalent uniform end-of-period net cash flows using (3.18).

This approach assumes a "replacement chain" that continues in perpetuity (i.e., an infinite series of replacement projects). The replacement projects are considered to be identical to the initial project in a replacement chain if the timing and magnitude of all of their cash flows are the same as the cash flows of the initial project, except for being shifted a multiple of N_j periods into the future. In some cases, it is not reasonable to assume that all of the cash flows associated with the initial project will be repeated when the project is replaced; for example, the salvage value of existing equipment that, once salvaged at time 0, will not be available for salvage in the future. Such one-time cash flows should not be include in the replacement projects.

Inflation

As long as all of the cash flows associated with a project are expected to increase at the same rate of inflation, the effects of inflation can essentially be ignored in capital project evaluation by both:

1. Stating the future cash flows in terms of "current" (i.e., not including future inflation) dollar amounts, as opposed to the "actual" (i.e., including future inflation) dollar amounts received throughout the project, and

2. Using the "real," as opposed to the actual or "nominal," cost of capital.

If i is the nominal cost of capital and i_{infl} is the rate of inflation, then

$$i_{\text{real}} = \frac{1+i}{1+i_{\text{infl}}} -1 = \frac{1+i -\left(1+i_{\text{infl}}\right)}{1+i_{\text{infl}}} = \frac{i -i_{\text{infl}}}{1+i_{\text{infl}}} \approx i -i_{\text{infl}} \qquad (3.23)$$

is the real cost of capital, where it is assumed that the basis period of the nominal cost of capital or interest rate is the same as the compounding period so that the nominal rate equals the effective rate.

For example, if the nominal cost of capital is 10% per year (compounded annually) and the inflation rate is expected to be 4% per year, then the net present value of current-dollar cash flows discounted at a real cost of capital of 5.77% ($\approx$ 6%) is the same as the net present value of the actual-dollar cash flows that will occur discounted at the nominal cost of capital 10%. If the cash flows associated with a project are expected to increase at different rates of inflation (which is possible since "the" rate of inflation determined by the government is only a weighted average of a selected number or basket of goods and services in the economy), then their actual dollar amounts and the nominal cost of capital should be used in the capital project evaluation decision.

Taxes

As long as the impact of taxes on the cash flows associated with each project are similar in proportion, then the effects of taxes can be ignored as an initial approximation in capital project evaluation since the relative differences in cash flows will be the same. Taxes should be included in order to estimate absolute cash flows. Unless the goal of tax policy is to change economic decisions, tax effects are usually designed to limit their impact on decision-making.

Depreciation

Depreciation represents the decline in value of a fixed asset as a result of the asset being used for a project—the difference between the value of the asset at the start of a time period and its (salvage) value at the end of the time period. For tax purposes, a certain amount of depreciation can be considered as a cost (i.e., deductible expense) each year to represent the decline in value of fixed assets (except land). During the life of the project, the depreciation of the asset does not generate any cash flows. The only cash flows that may result from the asset being used for the project are expenditures to purchase the asset and any cash flows associated with the disposal of the asset (which may be positive or negative). In capital project evaluation, the only impact of depreciation on each period's cash flow is through the reduction in income tax expense that it provides. The capital recovery cost of a fixed asset provides a per-period cost that is equivalent to the asset's net present value. "Sinking-fund" depreciation plus interest on the initial investment, both at the cost of capital, is the only depreciation method that provides per-period costs that are equivalent to the asset's capital recovery cost ("straight-line" depreciation is equivalent to using a 0% cost of capital to discount the sinking fund).

Cost of capital project evaluation

Since performing a capital project evaluation is itself not a costless operation, it is important to consider the level of detail required for the analysis in comparison to its potential costs and benefits. Once the idea for a project has occurred and prior to starting a formal capital project evaluation of the project, the costs associated with analysis itself are part of the project's costs; once performed, the analysis becomes a sunk cost whether or not the project is accepted.

The principal cost of many cost reduction projects in a facility is the labor of the engineering staff during the projects' planning stages. Because of these considerations, it is useful to perform a preliminary analysis of a project to decide what level of detail will be required to make a decision concerning its acceptance. For example, it may be decided that the "payback period" of a project provides enough information to make a reasonable decision, or it may be judged that the positive net present value of the readily identifiable cash flows associated with a project provide enough of a safety margin to more than outweigh the possible impact of the more difficult to determine opportunity costs (e.g., implicit costs) of the project—the cost of a more detailed analysis would only reduce the project's net benefits without changing the final decision.

Externalities

Although the cash flows associated with a project correspond to the costs and benefits of the project to a firm, they do not necessarily correspond to the costs and benefits of the project to an individual or to society as a whole. The costs and benefits of a project that are external to the firm and are not compensated by the firm are termed "externalities." For example, if as a result of a project an employee of the firm has to work harder at no increase in salary, then the project has increased the personal, non-monetary cost to the employee of his or her job; the firm itself does not incur a cash outflow associated with the increased labor provide by the employee, and, unless the employee owns stock in the firm, he or she will not realize any of the cash benefits resulting from the project. One of the functions of the legal system is to require firms to consider externalities: fines and legal liability serve to attach a cash cost to actions that would otherwise result in externalities. A firm may be required by a regulatory agency of the government to undertake a mandatory project that will not result in any cash inflows; the cash cost to the firm of not accepting the mandatory project would be a fine or possible shutdown.

4. Capacity Planning

4.1 Little's Law

"Factory physics" is a term used by Hopp and Spearman[1] to refer to an approach to analyzing production systems that uses queuing approximation formulas to estimate system performance. With respect to capacity planning, factory physics provides a middle ground between the simple "rough cut" approach traditionally used and the more detailed and time-consuming simulation approach. More details concerning the queueing approximation formulas used in factory physics can be found in Suri et al.[2]

Approximation formulas for a production system's cycle time is related to its throughput via Little's Law:

Little's Law: $$TH = \frac{WIP}{CT} \qquad (4.1)$$

where

> TH = throughput

> > = average output of a production system per time period (e.g., units per hour)

> WIP = work-in-process

> > = average number of units of product in a production system

> CT = cycle time

> > = average time each unit of product is in a production system

For a serial production system or routing, *WIP* is the inventory between the start and end points of the routing and *CT* is average time from release of a job at the beginning of the routing until it reaches an inventory point at the end of the routing.

Figure 4.1 shows a graphical depiction of Little's Law, where the throughput corresponds to the slope of the lines representing the cumulative number of arrivals and departures.
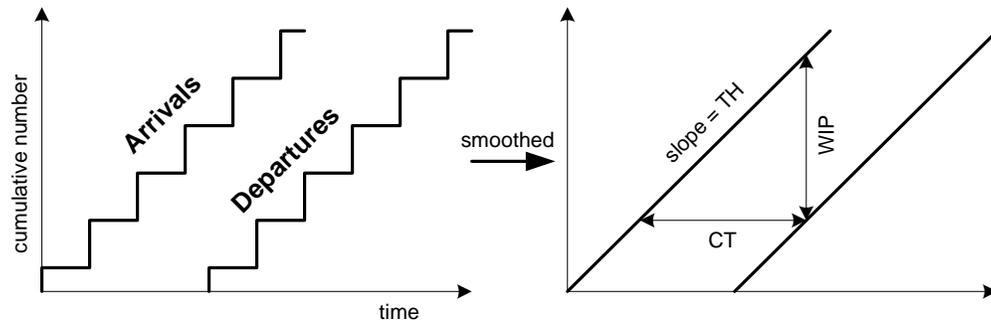
Figure 4.1. Little's Law[3]

4.2 Line Yield

Yield fractions are estimates of the long-run average percentage of nondefective units that are produced and/or identified. During any particular short period of time, the estimate is not likely to be accurate. A large number of defective units are likely immediately following machine setups or when a process is out of control. Factors that can influence the percentage of scrap include: whether the operation is performed using manual or automated equipment, the tolerances specified, the grade of material, and equipment maintenance.

Let y_i = yield fraction of ith operation, i.e., the long-run average percentage of nondefective units produced or identified

$r_{d,i}$ = departure rate (desired output) of nondefective units from ith operation

$r_{a,i}$ = input rate of nondefective units to ith operation.

For each component part, let $r_{a,1}$ be the number of units per period required at the start of its first operation (a leaf node of the operation process chart) and let $r_{a,N}$ be the number of units per period of nondefective final product required (i.e., the product's throughput rate) after the last operation or inspection (the root node of the operation process chart). A total of N operations and inspections are performed on the component part as it is transformed into the final product. Since $r_{a,N}$ is known, and assuming the $y_1, \ldots, y_N$ yield fractions have been estimated, $r_{a,1}$ can be determined as follows:

$$r_{a,1} = \frac{r_{d,N}}{Y_N},\tag{4.2}$$

where

$$Y_i = \prod_{j=1}^{i} y_j \tag{4.3}$$

is the cumulative yield (or *line yield*) from operation 1 to *i*.

4.2.1 Example

A four-workstation production line that has a serial routing is shown in Figure 4.2, where workstation 1 (W/S 1) has three machines (M/C), W/S 2 has six M/C, etc. Because of yield loss, 14.04 parts must be started at W/S 1 to get 10 nondefective parts from W/S 4. Table 4.1 shows the spreadsheet calculations used to determine the line yield, where the shaded cells correspond to input values. Alternatively, the cumulative yield at the last operation can be used to determine the required input to the first operation on W/S 1:

$$r_{a,1} = \frac{r_{d,4}}{Y_4} = \frac{10}{0.85 \cdot 0.9 \cdot 0.95 \cdot 0.98} = \frac{10}{0.71222} = 14.0407 \ . \tag{4.4}$$
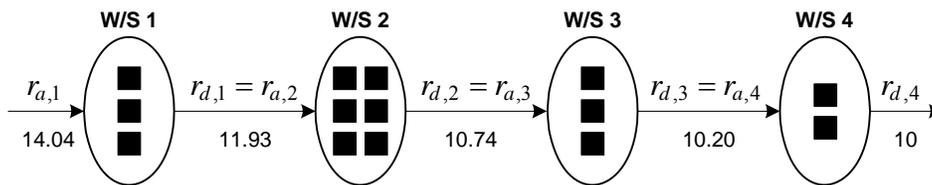
Figure 4.2. Four-workstation production line.

Table 4.1. Line Yield

W/S	1	2	3	4
Arrival Rate (r_a, q/hr)	14.0407	11.9346	10.7411	10.2041
Yield (y)	0.85	0.9	0.95	0.98
Departure Rate (r_d, q/hr)	11.9346	10.7411	10.2041	10

4.2.2 Rework

In addition to final good units, defective units (or scrap) is produced at a rate of $r_{a,i}(1-y_i)$ at each operation i. If significant, this output must somehow be accommodated when planning material flow. Possible uses for scrap include the following:

- reworking it so that it can be reused upstream or downstream in the production process,
- reworking it so that it can be sold as second quality product,
- selling it as is without rework, or paying to have it removed.

4.3 Throughput Feasibility

The throughput requirement of each workstation corresponds to the desired output rate of nondefective units from the workstation

$$\text{throughput from workstation} = TH^{\text{from}} \quad \equiv \quad r_d = \text{departure rate} \ . \tag{4.5}$$

Note: When using Little's Law to determine the WIP at a workstation, the throughput *to* the workstation corresponds to the rate of units input to the workstation

$$\text{throughput to workstation} = TH^{\text{to}} \quad \equiv \quad r_a = \text{arrival rate}. \tag{4.6}$$

Increasing the departure rate to account for yield loss results in the required arrival rate to the station, $r_a = r_d/y$, from which the minimum number of identical machines required at the workstation can be determined so that the effective capacity or service rate of the workstation *strictly exceeds* the arrival rate to the station:

$$\text{arrival rate} = r_a \quad < \quad r_e = \text{service rate}, \tag{4.7}$$

which is equivalent to the requirement that the utilization of the workstation be strictly less than one:

$$\text{utilization} = u = \frac{r_a}{r_e^m} = \frac{r_a t_e}{m} \quad < \quad 1, \tag{4.8}$$

where

r_a = arrival rate to workstation

$r_e^m = \dfrac{m}{t_e}$ = service rate (or effective capacity) of m-machine workstation

m = number of identical machines in workstation ($m \geq \lfloor r_a t_e + 1 \rfloor$)

$t_e = \dfrac{t_0}{A}$ = effective mean process time with failures (preemptive outages)

t_0 = natural mean process time

$A = \dfrac{MTTF}{MTTF + MTTR}$ = availability

$MTTF$ = mean time to failure

$MTTR$ = mean time to repair.

Equation (4.8) can then be used to determine the minimum number of identical machines needed at the workstation in order to satisfy the throughput requirements:

$$m = \lfloor r_a t_e + 1 \rfloor \tag{4.9}$$

(Note: Using $m = \lceil r_a t_e \rceil$ won't work since it is possible that $\lceil r_a t_e \rceil = r_a t_e$, which would result in a utilization of 100%.)

4.3.1 Example

Continuing with the example shown in Table 4.1, Table 4.2 and Table 4.3 show the spreadsheet calculations used to determine the minimum number of machines necessary to produce a throughput of 10 parts per hour from the line shown in Figure 4.1 (i.e., a throughput feasible capacity plan).

Table 4.2. Throughput Feasible Capacity Plan

W/S	1	2	3	4
Arrival Rate (r_a, q/hr)	14.0407	11.9346	10.7411	10.2041
Natural Process Time (t_0, hr)	0.2	0.5	0.25	0.15
MTTF (hr)	40		100	
MTTR (hr)	2	0	5	0
Availability (A)	0.95238	1	0.95238	1
Effective Process Time (t_e, hr)	0.21	0.5	0.2625	0.15
Number of M/C (m)	3	6	3	2
Utilization (u)	0.98285	0.99455	0.93985	0.76531
Yield (y)	0.85	0.9	0.95	0.98
Departure Rate (r_d, q/hr)	11.9346	10.7411	10.2041	10

Table 4.3. Spreadsheet Formulas Used in Table 4.2

	A	B	C	D
1	W/S		1	2
2	Arrival Rate (r_a, q/hr)		=C11/C10	=D11/D10
3	Natural Process Time (t_0, hr)		0.2	0.5
4	MTTF (hr)		40	
5	MTTR (hr)		2	0
6	Availability (A)		=IF(ISBLANK(C4), 1, C4/(C4 + C5))	=IF(ISBLANK(D4), 1, D4/(D4 + D5))
7	Effective Process Time (t_e, hr)		=C3/C6	=D3/D6
8	Number of M/C (m)		=FLOOR(C2*C7 + 1,1)	=FLOOR(D2*D7 + 1,1)
9	Utilization (u)		=C2*C7/C8	=D2*D7/D8
10	Yield (y)		0.85	0.9
11	Departure Rate (r_d, q/hr)		=D2	=E2

4.3.2 Make-to-Stock vs. Make-to-Order

If is possible to hold inventory and the cost of holding inventory is low relative to the cost to purchase a machine, then production can be *make-to-stock*. For dedicated machines producing a single product, the *average* arrival rate over the period during which inventory can be held should be used to determine *m* in (4.9), e.g., the average annual rate is

$$r_a^{\text{avg}} = \frac{\text{annual demand}}{\text{hrs/yr}}.$$

If it is not possible to hold inventory or the cost of holding inventory is high relative to the cost to purchase a machine, then production is usually *make-to-order*. For dedicated machines producing a single product, the *peak* arrival rate over the period during which product must be produced should be used; e.g., if peak demand is likely to be no more than 50% above the average rate, then

$$r_a^{\text{peak}} = 1.5\, r_a^{\text{avg}} .$$

4.3.3 When is Scrap Detected?

During a production operation that results in a product being scrapped, all of the machine time spent producing the defective product is lost. If all defective products could be identified prior to the start of an operation, no time would be lost and thus, in (4.9), the departure rate, r_d, could be used in place of r_a (scrap is still being produced, but no machine time is lost in the process). Using r_a in (4.9) is equivalent to assuming that all defective products are only identified at the end of processing. Although, on average, defective products are likely to be identified about around halfway through processing, it is usually safer to use r_a in (4.9) if no information is available concerning when defects occur during processing.

4.4 Machine Sharing and Setups

If several of the operations used to produce a product use the same type of machine, or several products are produced that have operations that use the same type of machine, then it is possible to reduce the number of machines required by sharing machines among the operations. It is not possible to just add the equipment fractions of each operation sharing the same type of machine together to determine the total number of shared machines required. Unless the parts produced for each operation are almost identical, there is usually some additional time required to reconfigure a machine to produce each batch of different part types. This additional time when the machine is stopped and can not be operating is termed the machine's *internal setup time*; in contrast, the machine's *external setup time* is the time required to prepare to produce a particular part on the machine that does not require the machine to be stopped (e.g., fixture preparation). A machine's "down time" represents the percentage of time the machine is not available for production due to repair or planned maintenance—it does not include internal setup time because the machine is still available for production during this time.

Internal setup time is essentially an additional fixed cost associated with sharing machines among several operations. This additional cost is offset by the savings in equipment costs associated with not having to dedicate underutilized machines to individual operations. In order to spread this fixed cost over a number of units, multiple units of a part type (*production batches*) can be produced on a machine before switching to produce a different part type. Large production batch (or lot) sizes can reduce the average setup cost per unit, but they can cause increases in work-in-process (WIP) inventory, making it difficult to control production, and they

can result in entire batches being scrapped or reworked because of the delay in getting feedback from subsequent operations. One method of reducing batch size without increasing per unit setup cost is to convert internal setup actions to external setup actions (a technique often associated with "Just-In-Time" (JIT) production). Small production batch sizes can increase the average setup cost per unit and can increase equipment costs if additional machines are required due to increases in down time to perform internal setups. In practice, determining the most cost effective batch size is a difficult task.

In many cases, several operations may require many of the same setup actions on the machine. The setup time to switch between these operations is less than the time required for a full machine setup. By grouping these operations together on a shared machine, the time spent on setups can be reduced. The process of grouping together related "families" of operations or products to be performed in "cells" of shared machines is termed "group technology."

4.5 Cycle-Time Estimation

The following queuing approximation formulas can be used to estimate the cycle time of a single machine workstation (G/G/1) or a workstation with m identical machines (G/G/m). When $m = 1$, the equation for G/G/m reduces to the equation for G/G/1. Equation (4.13) is used to estimate the arrival squared coefficient of variation (SCV) of a workstation based on the departure SCV of the previous workstation in the routing, where the arrival SCV of the first workstation in the routing must be initially specified, and Equation (4.14) is used to determine the effective process time SCV with failures:

$$CT = \underbrace{CT_q}_{\text{queuing time}} + \underbrace{t_e}_{\text{process time}} \tag{4.10}$$

$$\textbf{G/G/1:} \quad CT_q = \underbrace{\left(\frac{c_a^2 + c_e^2}{2}\right)}_{\text{variability}} \underbrace{\left(\frac{u}{1-u}\right)}_{\text{utilization}} \underbrace{t_e}_{\text{time}}, \quad \text{if } m = 1 \tag{4.11}$$

$$\textbf{G/G/m:} \quad CT_q = \left(\frac{c_a^2 + c_e^2}{2}\right)\left[\frac{u^{\left(\sqrt{2(m+1)}-1\right)}}{m(1-u)}\right]t_e, \quad \text{if } m \geq 1 \tag{4.12}$$

$$\textbf{Departure SCV:} \quad c_d^2 = \begin{cases} u^2 c_e^2 + \left(1-u^2\right)c_a^2, & \text{if } m = 1 \\ 1 + \left(1-u^2\right)\left(c_a^2 - 1\right) + \dfrac{u^2}{\sqrt{m}}\left(c_e^2 - 1\right), & \text{if } m \geq 1 \end{cases} \tag{4.13}$$

$$\textbf{Eff. Process Time SCV:} \quad c_e^2 = c_0^2 + (1+c_r^2)A(1-A)\frac{MTTR}{t_0} \tag{4.14}$$

where

$$u = \frac{r_a t_e}{m} = \text{utilization}$$

$$m = \text{number of identical machines in workstation } (m \geq \lfloor r_a t_e + 1 \rfloor)$$

$$r_a = \text{arrival rate}$$

$$c_a^2 = \frac{\sigma_a^2}{t_a^2} = \text{arrival SCV (squared coefficient of variation)}$$

$$t_a = \frac{1}{r_a} = \text{mean time between arrivals}$$

$$\sigma_a^2 = \text{variance of arrival time } (= 0, \text{deterministic}; = t_a^2, \text{exponential})$$

$$t_e = \text{effective mean process time with failures (preemptive outages)}$$

$$c_0^2 = \frac{\sigma_0^2}{t_0^2} = \text{natural process time SCV (squared coefficient of variation)}$$

$$\sigma_0^2 = \text{variance of natural process time}$$

$$c_r^2 = \frac{\sigma_r^2}{MTTR^2} = \text{repair time SCV}$$

$$\sigma_r^2 = \text{variance of repair time.}$$

4.5.1 Example

Continuing with the example shown Table 4.2, Table 4.4 shows the calculations used to estimate the cycle time of the line shown in Figure 4.1 and to determine the total cost of the machines used in the line. Note: The arrival rate, r_a, is used to calculate the WIP instead of the throughput, $TH = r_d$, due to the (conservative) assumption that all failures are identified at the end of processing.

Table 4.4. Cycle Time and Total Machine Cost Estimation

W/S		1	2	3	4	Total
Arrival Rate (r_a, q/hr)		14.0407	11.9346	10.7411	10.2041	
Arrival SCV	(c^2_a)	1	1.08758	0.59714	0.90558	
Natural Process Time	(t_0, hr)	0.2	0.5	0.25	0.15	1.1
Natural Process SCV	(c^2_0)	0.25	0	0	0.5	
MTTF	(hr)	40		100		
MTTR	(hr)	2	0	5	0	
Repair Time SCV	(c^2_r)	1	0	0	0	
Availability	(A)	0.95238	1	0.95238	1	
Effective Process Time	(t_e, hr)	0.21	0.5	0.2625	0.15	1.1225
Eff Process Time SCV	(c^2_e)	1.15703	0	0.90703	0.5	
Number of M/C	(m)	3	6	3	2	
Utilization	(u)	0.98285	0.99455	0.93985	0.76531	
Yield	(y)	0.85	0.9	0.95	0.98	
Departure Rate (r_a*y)	(r_d, q/hr)	11.9346	10.7411	10.2041	10	
Departure SCV	(c^2_d)	1.08758	0.59714	0.90558	0.75381	
Cycle Time in Queue (CT_q, hr)		4.26486	8.19096	0.97673	0.15241	13.58496
Cycle Time at W/S (CT, hr)		4.47486	8.69096	1.23923	0.30241	14.70746
WIP in Queue (r_a*CT_q)	(q)	59.8816	97.7558	10.4912	1.55518	169.6839
WIP at W/S	(q)	62.8302	103.723	13.3108	3.08579	182.9499
M/C Cost ($000)		12	18	2	6	
W/S Cost ($000)		36	108	6	12	162

4.6 Basic Capacity Planning Procedure

In traditional "rough cut" capacity planning, only enough capacity is planned at each workstation to meet its throughput (TH^*) requirements; cycle time is ignored. In Table 4.5, both throughput (TH^*) and cycle time (CT^*) requirements are met as part of the design process.

BASIC CAPACITY PLANNING PROCEDURE FOR A SINGLE WORKSTATION

0. Given: $r_a \Rightarrow TH^*$ (throughput), and CT^* (cycle time) requirements.

1. (Throughput feasible) Determine initial number of machines:

$$m = \lfloor r_a t_e + 1 \rfloor \qquad (4.15)$$

2. Calculate CT using (4.11) or (4.12).

3. (Cycle time feasible) If $CT \le CT^*$, then STOP; else, either

    (a) Add machine: $m \Leftarrow m + 1$

    (b) Or make some other change to workstation

    and GOTO Step 2.

**Table 4.5. Cycle Time Feasibility
(Target Cycle Time is < 6 hr)**

This page contains a large rotated spreadsheet titled "Table 4.5. Cycle Time Feasibility." The table is organized into two main sections, "First Change" and "Second Change," each with column blocks labeled "Base," "Add 1 M/C to W/S 1," "Add 1 M/C to W/S 2," "Add 1 M/C to W/S 3," and "Add 1 M/C to W/S 4," with sub-columns 1, 2, 3, 4 and Total.

The row labels (left-hand column) are:

- W/S
- Arrival Rate (r_a, q/hr)
- Arrival SCV (c^2_a)
- Natural Process Time (t_0, hr)
- Natural Process SCV (c^2_0)
- MTTF (hr)
- MTTR (hr)
- Repair Time SCV (c^2_r)
- Availability (A)
- Effective Process Time (t_e, hr)
- Eff Process Time SCV (c^2_e)
- Number of M/C (m)
- Utilization (u)
- Yield (y)
- Departure Rate (r_d*y) (r_d, q/hr)
- Departure SCV (c^2_d)
- Cycle Time in Queue (CT_q, hr)
- Cycle Time at W/S (CT, hr)
- WIP in Queue (r_a*CT_q) (q)
- WIP at W/S (q)
- M/C Cost ($000)
- W/S Cost ($000)
- Cost Increase (Chg - Base) ($000)
- CT Decrease (Base - Chg) (hr)
- Ratio ($000/hr)

4.7 Capacity Planning for Nonstationary Demand

When Eq. (4.12) is used to estimate the cycle time in queue, the estimate assumes that the demand process is stationary and the result refers to a long-term average estimate. When the demand process is not stationary, a simple recursive procedure can be used to estimate the cycle time in queue, CT_q, via Little's law as applied to the total inter-period WIP in queue and the total of the expected arrivals occurring in each period:

$$CT_q = \frac{\text{Total } WIP_q}{\text{Total } TH} = \frac{\displaystyle\sum_{i=1}^{n+1} q_x(i)}{\displaystyle\sum_{i=1}^{n} r_a(i)}, \tag{4.16}$$

where n is the number of periods with arrivals, $r_a(i)$ is the arrival rate for period i, and $q_x(i)$, the excess demand not served during period i, can be calculated for periods 1 to $n + 1$ as follows:

$$q_x(i) = \begin{cases} 0, & i = 0 \\ \max\{0, r_a(i) + q_x(i-1) - r_e\}, & i = 1, \ldots, n \\ (n_x - 1)\left(q_x(n) - \dfrac{n_x r_e}{2}\right), & i = n+1 \end{cases} \tag{4.17}$$

where

$$n_x = \left\lceil \frac{q_x(n)}{r_e} \right\rceil \tag{4.18}$$

is the number of periods required after the n periods of normal operations to complete processing and r_e is the (constant) service rate. (The service rate is assumed to be constant only to simplify the notation; different rates could be used to represent additional capacity that is available during peak periods of demand.)

In estimating CT_q, only the number of periods from the arrival of a demand to the period immediately preceding its processing is used to determine the queuing delay and any queueing that occurs during the period that processing occurs is ignored; thus, if $r_a(i) \le r_e$ for all periods, then CT_q would be zero. The excess demand in period $n + 1$, $q_x(n+1)$, represents the total number of periods of queueing required to complete the processing of the excess demand $q_x(n)$ in the system at the end of normal operations, not including the periods that the actual processing occurs. Since the final excess period is not included, $q_x(n+1)$ is equal to the $(n_x - 1)q_x(n)$ periods of queueing minus the cumulative impact of the r_e units of demand processed in each of the $(n_x - 1)$ periods; or,

$$q_x(n+1) = (n_x - 1)q_x(n) - \left(\sum_{i=1}^{n_x-1} i\right)r_e = (n_x - 1)q_x(n) - \frac{n_x(n_x - 1)}{2}r_e = (n_x - 1)\left(q_x(n) - \frac{n_x r_e}{2}\right).$$

With excess demand, the throughput on a per-period basis is

$$TH = \frac{\displaystyle\sum_{i=1}^{n} r_a(i)}{n + n_x}. \tag{4.19}$$

4.7.1 Example

On an average day, customers are expected to arrive to a barber shop at rates of 5, 7, 20, 32, 6, 15, 45, and 12 per hour for the eight hours that the shop will be open. The arrivals during each hour are totally random and each haircut takes exactly 20 minutes. All customers in the shop at closing will continue to be processed. If there are five barbers, the estimated time that the average customer spends waiting in the barber shop is about 1.2 hours. The estimated average number of customers waiting in the shop during the day is about 15.5 and average number having their hair cut is about 4.3. It takes over two hours to finish processing the 40 customers that are in the shop at closing.

	Total	0	1	2	3	4	5	6	7	8	Excess
r_a	142		5	7	20	32	6	15	45	12	
r_e		15	15	15	15	15	15	15	15	15	
q_x	171	0	0	0	5	22	13	13	43	40	35

$3 = n_x$

$CT_q=$	1.2042
$CT=$	1.5376
$t_e=$	0.3333
$m=$	5
$r_e=$	15
$TH=$	12.909
$WIP_q=$	15.545
$WIP=$	19.848
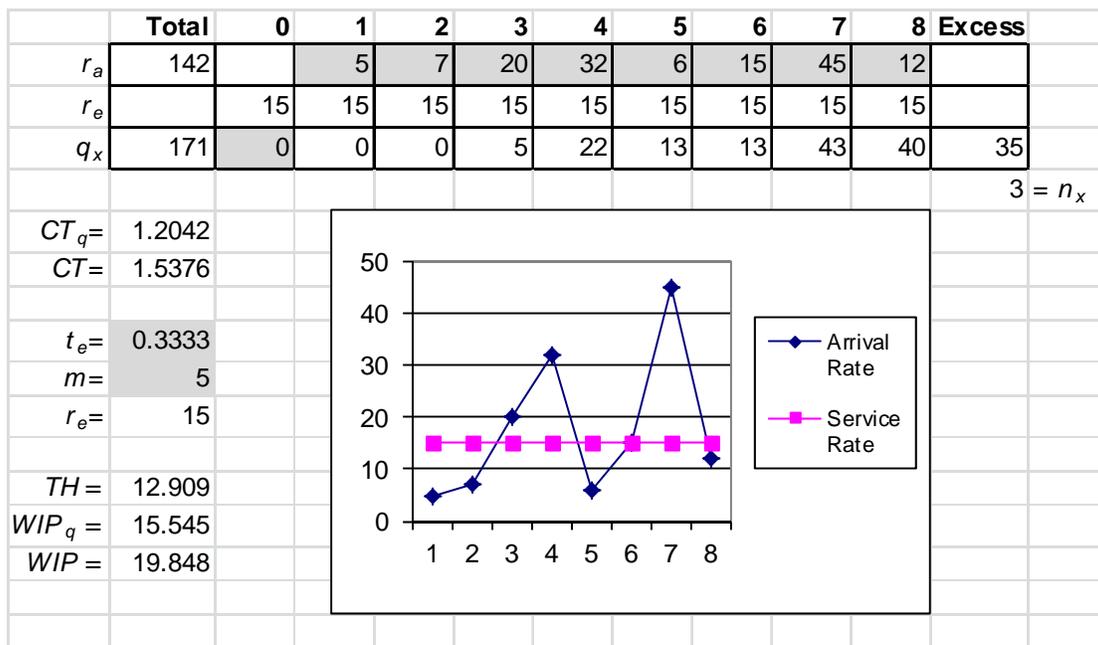
Figure 4.3. Capacity planning for nonstationary demand example.

4.8 References

The following sources are recommended for further study:

Askin, R.G., and Standridge, C.R., 1993, *Modeling and Analysis of Manufacturing Systems*, Wiley.

Cachon, G., and Terwiesch, C., 2006, *Matching supply with demand: An introduction to operations management*, McGraw-Hill.

Hopp, W.J., and Spearman, M.L., 2000, *Factory Physics: Foundations of Manufacturing Management*, 2nd Ed., McGraw-Hill.

Suri, R., Sanders, J.L., and Kamath, M., 1993, "Performance evaluation of production networks," *Handbooks in OR & MS*, vol. 4, Graves, S.C., et al., Eds., pp. 199–286.

5. Facility Layout

Facility layout involves determining the arrangement, and possibly the shape, of space-consuming entities (i.e., activities) in a facility. Examples of activities include machines, workstations, workcenters, production areas, offices, and departments. Activities interact with each other through flows of material, personnel, and/or information. Facility layout can be divided into two areas: machine layout and department layout.

Machine layout involves determining only the arrangement (i.e., position and orientation) of space-consuming entities that have a fixed shape. The term "machine" layout is used because the activities usually considered are the machines and workstations in a manufacturing department, which have a fixed shape.

Department layout involves determining both the arrangement and shape of space-consuming entities. The term "department" layout is used because the activities usually considered are the workcenters, production areas, offices, and departments in a facility, which, unlike machine layout, are not restricted to having a fixed shape. In general, determining a department layout is more "difficult" than determining a machine layout because a large number of different shapes are possible for each activity and the shape of one activity constrains the shapes of other activities.

5.1 Flow Processes

In the design of a production system, material flow related issues include the following:

1. How raw materials and component parts get to their first operation (or inspection).

2. How work-in-process moves from operation to operation—prior to material flow analysis, it is usually assumed that the output of an operation is immediately available as input to the next operation.

3. How finished product is removed from its last operation or inspection.

All of these considerations relate to flow processes:

- Discrete, identifiable items $\Rightarrow$ *discrete flow processes*

- Continuous or bulk, unidentifiable items $\Rightarrow$ *continuous flow processes*

5.1.1 Material Flow System

The material flow system of a facility refers to the flow of items "within" the facility. The type of flow is determined by the activities among which materials flow. The following flow planning hierarchy in Table 5.1 describes the different levels at which material flow can be analyzed. The flow within a workstation or machine will be assumed to be given and will not be considered further with respect to its impact on facility design.

Table 5.1. Flow Planning Hierarchy

Flow Within	Activities	Analysis Level
Workstation or machine	Motions	Motion studies and ergonomics
Department	Machines and workstations	Machine layout
Facility	Departments	Department layout

5.1.2 Total Cost of Material Flow

The following measure can be used to represent the total cost of material flow (TC_{MF}) within a facility. The measure represents the total cost of material movement between physical entities termed "activities." *Material flow* refers to material movement within a facility, and *material transport* refers to material movement to/from a facility. TC_{MF} can be used as a criterion for making facility design decisions, where an activity would correspond to, for example, an individual machine, workcenter, storage area, or department.

$$TC = TC_{\mathrm{MF}} = \sum_{i=1}^{M} \sum_{j=1}^{M} f_{ij} \cdot h_{ij} \cdot d_{ij} \tag{5.1}$$

where

M = number of activities between which material moves

f_{ij} = flow volume between activities i and j, measured in moves per time period

d_{ij} = distance (or move-time) between activities i and j

h_{ij} = cost per move between activities i and j per unit distance (or per unit move-time).

TC measures the total cost of material movement per time period. It can be used as the criterion with which to compare alternate facility designs (TC_{MF}) or alternate facility locations (TC_{tran}).

The "facility design problem"

In facility design, the (initial) flow requirements between activities, the f_{ij}'s, can be determined once the target production rate is determined; the d_{ij}'s and the h_{ij}'s remain unknown, and are only determined during facility design. Once the facility layout is completed so that each activity is

assigned a position in the facility, the distances (or move-times) between activities, the d_{ij}'s, can be determined; once the material handling system design is completed, the handling costs between activities, the h_{ij}'s, can be determined.
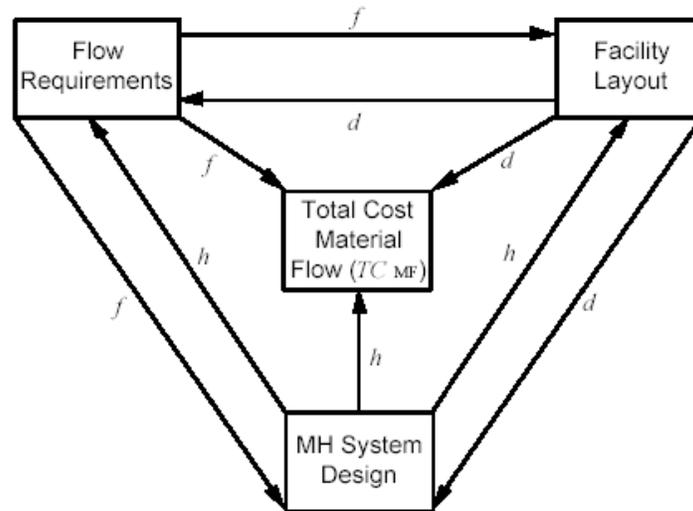
Figure 5.1. Use of total cost of material flow in facility design.

Problem: the "best" layout of the facility (i.e., the layout that minimizes TC_{MF}) depends on knowing the flow volumes and the handling costs, and the "best" material handling system design (i.e., the design that minimizes TC_{MF}) depends on knowing the flow volumes and the distances between activities; thus, facility layout and material handling system design are interdependent (and even PP&S design and facility design are interdependent since the flow requirements may be revised after the initial facility design is completed).

5.2 Machine Layout

Once the initial production system design has been completed for all of the products to be produced in the facility, decisions will have been made concerning the routing to be used to produce each item (or product), the number of each type of machine required for each operation in each workstation, and whether each machine or workstation is dedicated to a single operation or shared among several operations.

For the machine layout problem, the principal relationships between the activities will be assumed to be the quantitative measure of the flow of material between the activities as defined by their flow requirements. Other quantitative and qualitative measures of the relationships between activities are possible (e.g., adjacency) and can be considered in more general facility layout problems.

5.2.1 Flow Requirements

Given a routing for each item k and the total number of each item to be produced, the flow requirements between all pairs i and j of machines can be determined once some measure of the "cost" of each move has been determined.

Unit load sizes

The number of items transferred with each move is termed the *unit load* (or *transportation batch*) *size*. The transportation batch size is usually less than or equal to the production batch size of each operation. Unit load sizes are typically determined during material handling system design.

Equivalence factors

During the initial machine layout, little or no information is usually available concerning the type of material handling to be used to transport unit loads between machines. When this information is not available, it is impossible to determine the actual "cost" to transfer each load; instead, an *equivalence factor* (not an actual cost) can be used for each move to reflect the "relative difficulty" of transferring each unit load—that is, the estimated relative handling-related effort of transferring each load of a move as compared to the effort of transferring the loads of each of the other moves.

If the unit load sizes are not known, then a unit load size of one can be assumed for all moves and equivalence factors can be used to reflect the relative differences in handling-related effort between the items moved; for example, if an item is twice the size of another item, then half the number of items will be able to be placed into any equal sized transfer container (e.g., a tote) and twice the number containers will be required to transfer the same number of items—an equivalence factor for the larger item that is double that of the smaller item will reflect the fact that twice the number of containers are required (and, if a single container is transferred with each move, will translate into twice the number of required moves).

Equivalent flow volumes

The flows from all of the items that move from machine i to machine j can be converted into the following single "weighted" measure of flow, termed the "equivalent flow volume," between machines i and j:

$$\text{Equivalent Flow Volume:} \quad w_{ij} = \sum_{k=1}^{P} f_{ijk} \cdot h_{ijk} \,. \tag{5.2}$$

where P = total number of items (or products) considered

f_{ijk} = number of moves from machine i to machine j for item k during a time period

h_{ijk} = equivalence factor for moves from machine i to machine j for item k

Machine layout criterion

Given M different machines, the flow requirements between the machines are determined once each w_{ij} has been determined. The flow requirements can be represented in an $M \times M$ matrix (or "from/to chart") $\mathbf{W} \equiv \left[w_{ij} \right]$ of equivalent flow volumes. Different layouts of the machines can be compared using the following criterion that represents what could be termed the "total weighted distance of material flow" between the machines:

$$\sum_{i=1}^{M} \sum_{j=1}^{M} w_{ij}\, d_{ij} \,. \tag{5.3}$$

This criterion is similar to the total cost of material flow criterion, TC_{MF}, discussed in Eq. (5.1), except that it does not represent the actual cost of the material flow because only relative equivalence factors are used in place of actual handling costs.

Example

In this example, A, B, and C are $P = 3$ different types of items transferred between $M = 4$ machines shown in Figure 5.2. The total number of each item to be produced is $f_{i,j,A} = f_A = 8$, $f_{i,j,B} = f_B = 5$, and $f_{i,j,C} = f_C = 12$. The routings are:

$$A(8): 1\text{--}2\text{--}3\text{--}4; \quad B(5): 2\text{--}4\text{--}1\text{--}2\text{--}3; \quad C(12): 3\text{--}4\text{--}1\text{--}2\text{--}4$$

If no information is available concerning unit load sizes and, as shown in Figure 5.3, item A is three times the size of item C and item B is twice the size of item C, then equivalence factors of 3, 2, and 1 can be used for items A, B, and C, respectively, to reflect the likely relative handling effort required for each item so that, for example, $h_{i,j,A} = h_A = 3$, $h_{i,j,B} = h_B = 2$, and $h_{i,j,C} = h_C = 1$. If a common unit load container is to be used to transport the items, then, as shown in Figure 5.4, then 2 A's, 3 B's, and 6 C's can be placed on each container and handling effort is likely to be inversely proportional to the number of items that can be placed on the container; for this example, $h_A = \tfrac{1}{2}$, $h_B = \tfrac{1}{3}$, and $h_C = \tfrac{1}{6}$.
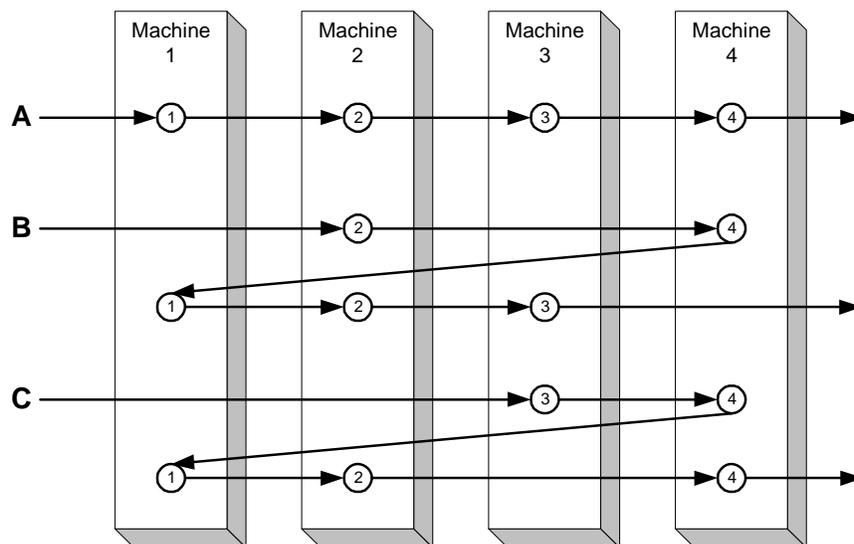

Figure 5.2. Routings for items A, B, and C.

**Figure 5.3. Relative sizes of items A, B, and C.**
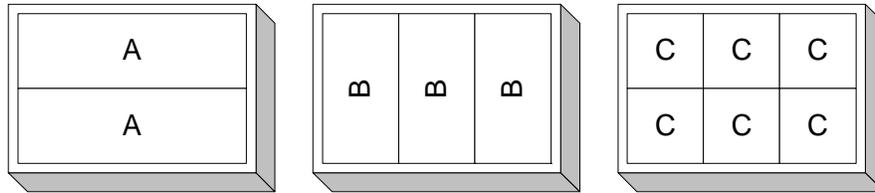
Figure 5.4. Common unit load container for items A, B, and C.

The resulting flow requirements for this example can be represented in the following 4×4 matrix of equivalent flow volumes (with nonzero values shown for the actual moves between machines):

$$\mathbf{W} \equiv \left[w_{ij} \right] = \begin{bmatrix} 0 & w_{1,2} & 0 & 0 \\ 0 & 0 & w_{2,3} & w_{2,4} \\ 0 & 0 & 0 & w_{3,4} \\ w_{4,1} & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 46 & 0 & 0 \\ 0 & 0 & 34 & 22 \\ 0 & 0 & 0 & 36 \\ 22 & 0 & 0 & 0 \end{bmatrix}.$$

5.2.2 Types of Machine Layouts

Four types of machine layouts are typically used in manufacturing departments:

1. *Fixed product layout*—product difficult to move $\Rightarrow$ machines are either at fixed locations adjacent to the product or are brought to product when they are needed

2. *Product layout* (production line)—large stable demand for a small number of products $\Rightarrow$ many dedicated machines and workstations $\Rightarrow$ a few dominate flows exist $\Rightarrow$ most material flow between adjacent machines/workstations $\Rightarrow$ fixed-path material handling possible

3. *Group layout* (or product family, or cellular layout)—several groups of products with related processing requirements and medium total demand for the products in each group $\Rightarrow$ most machines and workstations dedicated to group and shared within group $\Rightarrow$ little material flow between groups $\Rightarrow$ each group can be considered a separate manufacturing department for layout purposes $\Rightarrow$ product layout may be possible within group

4. *Process layout* (job shop)—low demand for a large number of unrelated products $\Rightarrow$ many shared machines and workstations $\Rightarrow$ no dominate flows exist $\Rightarrow$ little material flow between adjacent machines/workstations $\Rightarrow$ variable-path material handling likely

Product layouts can result in the lowest average cost of production if demand is large enough; process layouts can result in increased machine utilization if demand is low, at the cost of increased setup time and material flow; group layouts can realize many of the advantages of both product and process layouts, even when demand is not large, without as high setup and material flow costs.

5.2.3 Flow Patterns within Manufacturing Departments

The flow pattern within a manufacturing department depends on the type of machine layout (see Figure 5.5).
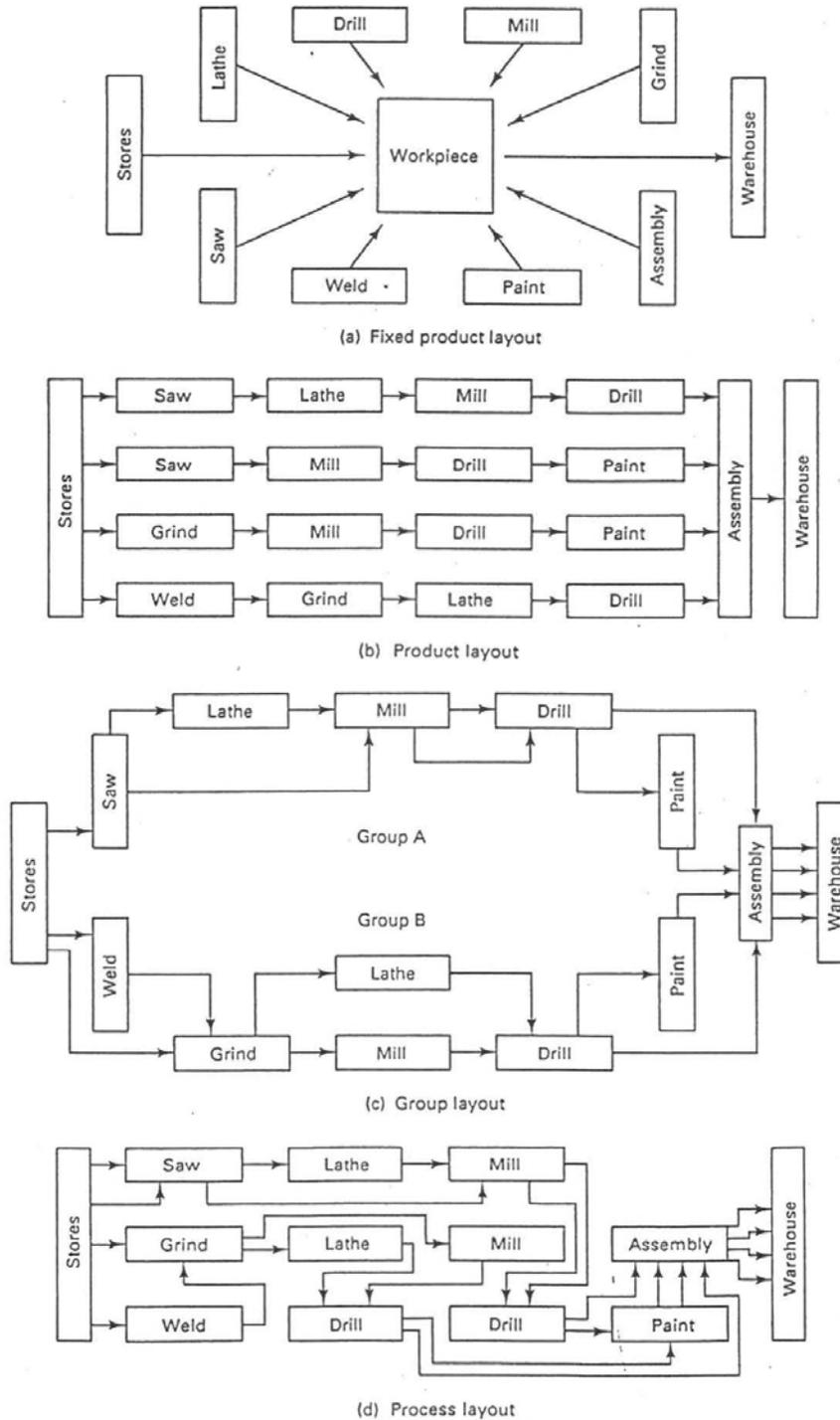

Figure 5.5. Types of machine layouts.[1]

Fixed product layout

Flow of operations to/from the product at a fixed location results in many of the machines being mobile.

Product layout

Most flow of product is between adjacent machines/workstations. Flow patterns depend on whether or not operators are shared between machines/workstations:

Dedicated operators $\Rightarrow$ end-to-end, back-to-back, or odd-angle flow patterns.

Operator shared between two machines/workstations $\Rightarrow$ front-to-front flow patterns.

Operator shared between several machines/workstations $\Rightarrow$ circular or U-shaped flow patterns (U-shaped flow patterns are often used in JIT production systems).

Group layout

Little flow between groups results in process layout flow patterns between groups, and product or process layout flow patterns within each group.

Process layout

The use of variable-path material handling results in most of the flow occurring between machines/workstations and aisles. Flow pattern depends on machine–aisle arrangement:

Two-way aisles $\Rightarrow$ parallel or perpendicular flow patterns

One-way aisles $\Rightarrow$ diagonal flow patterns

5.2.4 QAP Model of Machine Layout

Machine layout can be modeled as a *Quadratic Assignment Problem* (QAP) if each activity (machine) is restricted to occupy one of a finite number of fixed potential locations (*sites*).

The QAP is solved to find the minimum "cost" assignment of M machines to N fixed sites.

$$\textit{QAP for Layout:}\quad \text{Minimize}\quad TC = \sum_{i=1}^{M}\sum_{k=1}^{N}\sum_{j=1}^{M}\sum_{l=1}^{N} c_{ijkl}\, x_{ik}\, x_{jl} \tag{5.4}$$

$$\text{subject to}\quad \sum_{i=1}^{M} x_{ik} = 1, \quad \text{for all sites } k = 1,\ldots,N \tag{5.5}$$

$$\sum_{k=1}^{N} x_{ik} = 1, \quad \text{for all machines } i = 1,\ldots,M \tag{5.6}$$

$$x_{ik} = \{0,1\} \tag{5.7}$$

where

$$x_{ik} = \begin{cases} 1, & \text{if machine } i \text{ is assigned to site } k \\ 0, & \text{otherwise} \end{cases}$$

c_{ijkl} = "cost" of assigning machine i to site k when machine j is assigned to site l.

The objective function (5.4) of the QAP is quadratic because of the $x_{ik} \cdot x_{ik}$ terms. Constraints (5.5) assure that each site is assigned a (possibly dummy) machine. Constraints (5.6) assure each machine is assigned a unique site. Constraints (5.7) make the QAP a 0–1 integer program.

If $M > N$ (i.e., more machines than sites), then the problem is not well defined and has no solution; if $M < N$, the $N - M$ "dummy machines" are added so that $M + (N - M) = N$, each with no interactions with any other machines (i.e., the flow volumes of the dummy machines are all equal to zero). Unless otherwise stated, it will be assumed that $M = N$.

Heuristic (i.e., non-optimal) procedures are typically used to solve QAP's since it is computationally infeasible to find a (globally) optimal solution when M is much greater than 20. With $M = N$, $M!$ possible solutions (i.e., possible assignments) would have to be (at least implicitly) examined to find an optimal solution, which becomes infeasible for $M > 20$ (e.g., even $M = 12 \Rightarrow M! = 12! = 479{,}001{,}600$ possible solutions to examine). $M!$ is the number of possible permutations of the integers 1 to M.

Steepest Descent Pairwise Interchange Heuristic for QAP

The *Steepest Descent Pairwise Interchange* (SDPI) heuristic is a simple procedure for finding "locally optimal" solutions to a QAP. The locally optimal solution *may or may not* be the globally optimal solution.

Let $\mathbf{W} = M \times M$ machine weight matrix (equivalent flow volumes between all pairs of M machines).

$\mathbf{D} = N \times N$ site distance matrix ("distances" between all pairs of the N sites).

$\mathbf{a} = a(1), a(2), \ldots, a(N) = 1 \times N$ machine-to-site assignment (or permutation) vector, where $a(k) = i$ represents the assignment of machine i to site k.

The distance measure used in machine layout is problem-dependent: it may represent the straight line (or Euclidean) distance between two sites if the layout is in a large unobstructed open area; it may represent the rectilinear (or city block) distance if the travel between sites is along aisles; it may represent the actual travel distance a material handling device would travel between sites; or it may represent the time it would take for a material handling device to transport a load between sites.

Each element of $\mathbf{W}$ represents a machine-to-machine weight and each element of $\mathbf{D}$ represents a site-to-site distance. In order to be able to multiply elements of $\mathbf{W}$ with elements of $\mathbf{D}$ to determine the total weighted distance (TC), the elements of $\mathbf{W}$ and $\mathbf{D}$ must be made commensurate either by translating machine-to-machine weights to site-to-site weights, or by translating site-to-site distances to machine-to-machine distances.

Given M machines, each element $w(i, j)$ of **W** represents some measure of the "weight" between machine i and machine j. The assignment vector **a** can be used to translate machine weights to the weights between sites. Given $a(k) = i$ and $a(l) = j$, the weight between site k and site l is equal to the following element of the machine weight matrix **W**: $w(a(k), a(l)) = w(i, j)$.

For a given **a**, **W**, and **D**,

$$TC(\mathbf{a}) = \sum_{k=1}^{N} \sum_{l=1}^{N} w(a(k), a(l)) d_{kl} \tag{5.8}$$

represents the objective function value for **a**.

It is also possible to translate site distances to the distances between machines by using the $1 \times M$ site-to-machine assignment vector $\mathbf{a}^{-1} = \left[a^{-1}(1), a^{-1}(2), \ldots, a^{-1}(M) \right]$, which is the inverse permutation of **a**, where $a^{-1}(a(k)) = k$, for $k = 1, \ldots, N$, so that

$$TC(\mathbf{a}^{-1}) = \sum_{i=1}^{M} \sum_{j=1}^{M} w_{ij} \, d(a^{-1}(i), a^{-1}(j)) \tag{5.9}$$

represents the objective function value for $\mathbf{a}^{-1}$. The only difference between the two representations is that site-to-site weights are compared to site-to-site distances using **a** and machine-to-machine weights are compared to machine-to-machine distances using $\mathbf{a}^{-1}$.

Given an initial machine-to-site assignment vector $\mathbf{a}_0$, a simple heuristic is to determine which pairwise interchange of machines results in the greatest decrease in TC. If no interchange reduces TC, then keep the initial assignment; otherwise, make the pairwise interchange and then continue by looking at pairwise interchanges of the machines in the new assignment vector.

For a given assignment vector **a**, let $\mathbf{a}_{kl}$ be the assignment vector resulting from the interchange of the machines located at sites k and l. There are $\binom{M}{2} = M(M-1)/2$ pairwise interchanges of M machines.

SDPI HEURISTIC

---

0. Given $\mathbf{a}^0$, **W**, and **D**, let $\mathbf{a} \leftarrow \mathbf{a}_0$ and determine $TC(\mathbf{a})$.

1. Determine $\mathbf{a}' = \arg\min\left\{ TC(\mathbf{a}_{kl}) \,\middle|\, k < l \right\}$, where $\mathbf{a}'$ is an assignment that resulted in the minimum $TC(\mathbf{a}_{kl})$ for all $M(M-1)/2$ possible pairwise interchanges of machines in **a**.

2. If $TC(\mathbf{a}') \geq TC(\mathbf{a})$, then **stop**, with **a** as a locally optimal assignment vector; otherwise, $\mathbf{a} \leftarrow \mathbf{a}'$, $TC(\mathbf{a}) \leftarrow TC(\mathbf{a}')$, and **go to** Step 1.

As stated, the heuristic is not deterministic because the $\mathbf{a}'$ determined in Step 1 may be one of several different assignments that have the same minimum TC. One possible rule to make the heuristic deterministic (for computer implementation) is to choose $\mathbf{a}'$ based on lexicographic order so that the $\mathbf{a}'$ chosen is the first such assignment encountered in a traversal from 1 to N of the k and l indices; for example, $\mathbf{a}_{1,2}$ would be chosen over $\mathbf{a}_{1,3}$ and $\mathbf{a}_{2,1}$. In Step 2, $TC(\mathbf{a}')$ is required to be strictly less than $TC(\mathbf{a})$ for an interchange to occur. This requirement eliminates the need to check for the possibility of cycling.

Pairwise interchange graph

In order to illustrate the operation of the SDPI heuristic, a "pairwise interchange graph" can be constructed for small (i.e., less than five machines/sites) problems. In such a graph, each node $\mathbf{a}_x$, $x = 1,\ldots,M!$, represents one of the $M!$ possible assignment vectors and each arc $(\mathbf{a}_x,\mathbf{a}_y)$ represents a pairwise interchange that transforms assignment $\mathbf{a}_x$ into $\mathbf{a}_y$ and assignment $\mathbf{a}_y$ into $\mathbf{a}_x$. Each node is connected via arcs to $M(M-1)/2$ other nodes, representing all of the possible pairwise interchanges from this assignment. The value assigned to each node $\mathbf{a}_x$ is the total cost $TC(\mathbf{a}_x)$. Although the global optimum assignment can be read off the graph for small problems, there would be too many nodes to make this feasible for larger problems. The advantage of the SDPI heuristic for larger problems is that only $O(M^3)$ nodes are considered as compared to all $M!$ nodes.
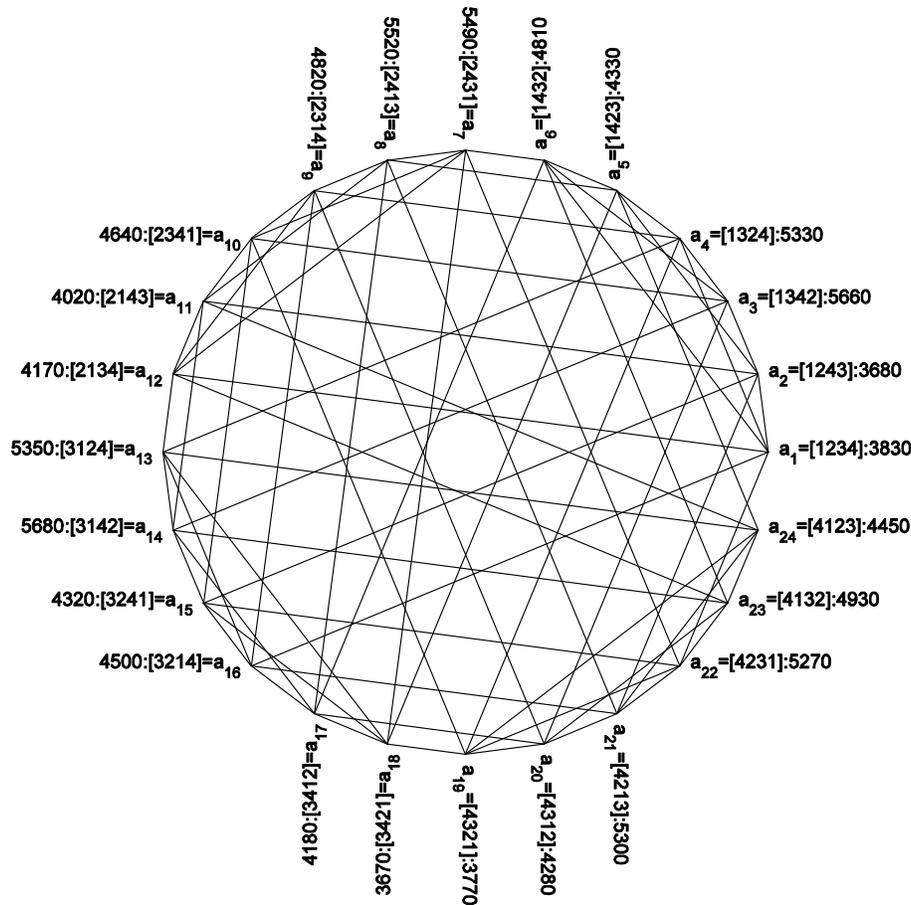
Figure 5.6. Pairwise interchange graph for 4-machine–4-site layout problem.

An example pairwise interchange graph is shown in Figure 5.6 corresponding to an $M = N = 4$ machine layout problem. There are $4! = 24$ nodes in the graph corresponding to all of the possible assignments of 4 machines to 4 sites. The arc from $\mathbf{a}_1$ to $\mathbf{a}_2$ represents $\mathbf{a}_{2,3}$, the interchange of machines 3 and 4 located at sites 3 and 4. The global optimum assignment is $\mathbf{a}_{18}$, corresponding to $TC = 3,670$.

Example 1 of SDPI heuristic: 1-D distances

Continuing the previous example (where the equivalent flow volume matrix $\mathbf{W}$ was determined for items A, B, and C): Let $M = N = 4$,

$$\mathbf{W} = \begin{bmatrix} 0 & 46 & 0 & 0 \\ 0 & 0 & 34 & 22 \\ 0 & 0 & 0 & 36 \\ 22 & 0 & 0 & 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} 0 & 10 & 35 & 50 \\ 10 & 0 & 25 & 40 \\ 35 & 25 & 0 & 15 \\ 50 & 40 & 15 & 0 \end{bmatrix},$$

where the site distance matrix $\mathbf{D}$ is determined from the site locations, represented as circles, as shown in Figure 5.7.

Figure 5.7. Site locations.

In order to use the SDPI heuristic, an initial assignment vector is needed; let

$$\mathbf{a}^0 = [a(1), a(2), a(3), a(4)] = [4, 1, 2, 3],$$

which corresponds to node $\mathbf{a}_{24}$ in the pairwise interchange graph shown in Figure 5.6 and to the initial assignment of machines (squares) to sites (circles) shown in Figure 5.8.

Figure 5.8. Initial machine–site assignment.

Step 0: Given $\mathbf{a}^0$, $\mathbf{W}$, and $\mathbf{D}$, let $\mathbf{a} \leftarrow \mathbf{a}^0$ and

$$TC(\mathbf{a}) = \sum_{k=1}^{4} \sum_{l=1}^{4} w\big(a(k), a(l)\big) \cdot d_{kl} = 4,450$$

Step 1 (1st pass): Of the nodes $\mathbf{a}_5$, $\mathbf{a}_{11}$, $\mathbf{a}_{13}$, $\mathbf{a}_{19}$, $\mathbf{a}_{21}$, and $\mathbf{a}_{23}$ connected to node $\mathbf{a} = \mathbf{a}_{24}$ via arcs in the pairwise interchange graph (Figure 5.6), the assignment vector corresponding to the minimum TC is $\mathbf{a}' = \mathbf{a}_{19} = \arg\min\{TC(\mathbf{a}_5), TC(\mathbf{a}_{11}), TC(\mathbf{a}_{13}), TC(\mathbf{a}_{19}), TC(\mathbf{a}_{21}), TC(\mathbf{a}_{23})\}$.

Step 2 (1st pass): Since $TC(\mathbf{a}') = TC(\mathbf{a}_{19}) = 3,770 < TC(\mathbf{a}) = TC(\mathbf{a}_{24}) = 4,450$, interchange machines at sites 2 and 4 ($\mathbf{a} \leftarrow \mathbf{a}'$), let $TC(\mathbf{a}) \leftarrow TC(\mathbf{a}')$, and GOTO Step 1 (start 2nd pass).

Step 1 (2nd pass): Of the nodes $\mathbf{a}_{20}$, $\mathbf{a}_{22}$, $\mathbf{a}_{24}$, $\mathbf{a}_4$, $\mathbf{a}_{10}$, and $\mathbf{a}_{18}$ connected to node $\mathbf{a} = \mathbf{a}_{19}$ via arcs in the graph, $\mathbf{a}' = \mathbf{a}_{19} = \arg\min\{TC(\mathbf{a}_{20}), TC(\mathbf{a}_{22}), TC(\mathbf{a}_{24}), TC(\mathbf{a}_4), TC(\mathbf{a}_{10}), TC(\mathbf{a}_{18})\}$.

Step 2 (2nd pass): Since $TC(\mathbf{a'}) = TC(\mathbf{a}_{18}) = 3{,}670 < TC(\mathbf{a}) = TC(\mathbf{a}_{19}) = 3{,}770$, interchange machines at sites 1 and 2 ($\mathbf{a} \leftarrow \mathbf{a'}$), let $TC(\mathbf{a}) \leftarrow TC(\mathbf{a'})$, and GOTO Step 1 (start 3rd pass).

Step 1 (3rd pass): Of the nodes $\mathbf{a}_{19}$, $\mathbf{a}_5$, $\mathbf{a}_7$, $\mathbf{a}_{13}$, $\mathbf{a}_{15}$, and $\mathbf{a}_{17}$ connected to node $\mathbf{a} = \mathbf{a}_{18}$ via arcs in the graph, $\mathbf{a'} = \mathbf{a}_{19} = \arg\min\{TC(\mathbf{a}_{19}), TC(\mathbf{a}_5), TC(\mathbf{a}_7), TC(\mathbf{a}_{13}), TC(\mathbf{a}_{15}), TC(\mathbf{a}_{17})\}$.

Step 2 (3rd pass): Since $TC(\mathbf{a'}) = TC(\mathbf{a}_{19}) = 3{,}770 < TC(\mathbf{a}) = TC(\mathbf{a}_{18}) = 3{,}670$, STOP, with

$$\mathbf{a} = \big[a(1), a(2), a(3), a(4)\big] = \big[3, 4, 2, 1\big]$$

as a locally optimal assignment vector, which corresponds to the final assignment of machines (squares) to sites (circles) shown in Figure 3.9, and to node $\mathbf{a}_{18}$ in the pairwise interchange graph (Figure 3.7).


Figure 5.9. Final machine–site assignment.

Inspection of the pairwise interchange graph (Figure 5.6) indicates that the node $\mathbf{a}_{18}$ found by the SDPI heuristic is the only local optima that correspond to the global optimum *TC* of 3,670. If $a^0 = a_1 = [1, 2, 3, 4]$, then the local optima found by the heuristic would have been $a_2 = [1, 2, 3, 4]$, with a *TC* of 3,680, which is greater than the global optimum. Since the SDPI heuristic will not always find the global optimum, the heuristic should be applied to a sequence of several different (randomly generated) initial assignment vectors, picking the best solution found as the final assignment. As a rule of thumb, the best solution found using M^2 different random assignment vectors can be used for an *M*-machine layout problem. Although 3-way interchanges could be used as a heuristic for the machine layout problem, the ease with which initial assignments can be generated reduces their utility.

Example 2 of SDPI heuristic: Four-different 2-D distance scenarios

In this example, the same weight matrix from Example 1 is used, except that a "dummy" fifth machine is added to the matrix since there are five sites and only four machines, thereby creating a 5×5 matrix **W2** by adding a fifth row and fifth column the 4×4 matrix **W** used in Example 1:

$$\mathbf{W2} = \begin{bmatrix} 0 & 46 & 0 & 0 & 0 \\ 0 & 0 & 34 & 22 & 0 \\ 0 & 0 & 0 & 36 & 0 \\ 22 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Four different distance scenarios are considered. In each scenario, sites are indicated as numbered circles. (You can assume that items can use any site as transit points along their paths.)

(a) *Open spaces:* Since there is unobstructed travel between all pairs of sites, Euclidean distances can be used.

(b) *Rectangular grid:* Since travel between all pairs of sites requires a series of 90° turns without any backtracking due to obstructions, rectilinear (a.k.a. city block or Manhattan) distances can be used.

(c) *Circulating conveyor:* Since travel between sites is in a counterclockwise direction along the conveyor, shortest path distances can be used.

(d) *General network:* Since travel between sites is obstructed, making Euclidean and rectangular distances inappropriate, shortest path distances can be used.
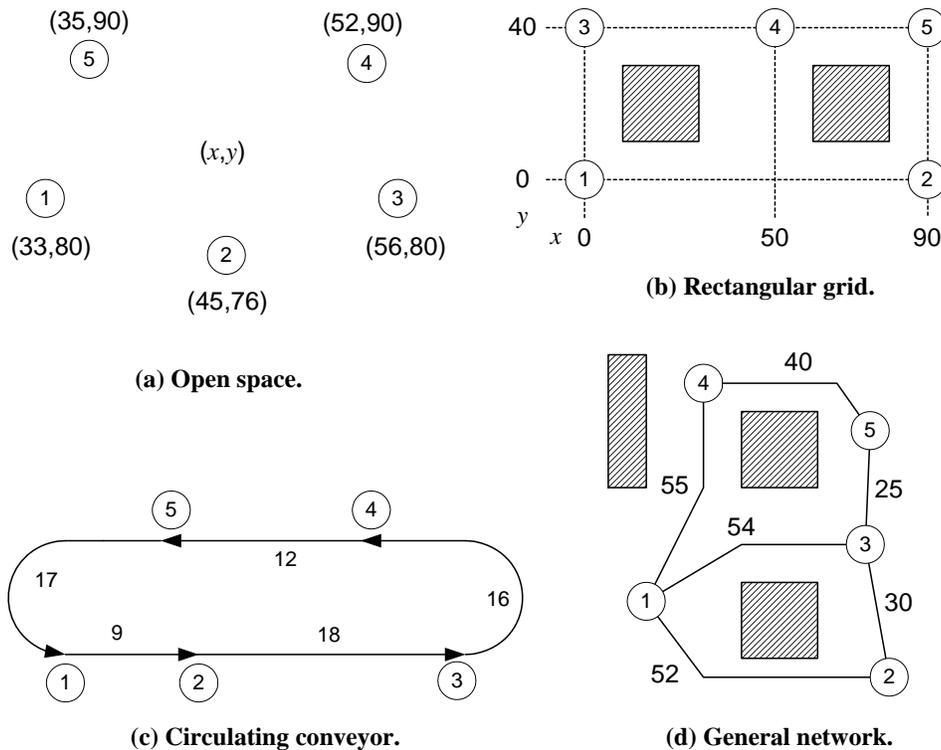
Figure 5.10. Four machine layout scenarios used in Example 2.

5.2.5 QAP Model with Fixed Costs

Up to now, the only costs used in the QAP model of the machine layout problem are the costs associated with the interactions between the machines. The QAP model can be extended to include additional fixed costs associated with assigning machines to particular sites. For each machine, a fixed cost is incurred that depends only on the site to which the machine is assigned; the additional cost is independent of the assignment of the other machines.

As before, the QAP with fixed costs is solved to find the minimum cost assignment of *M* machines to *N* sites.

QAP with Fixed Costs: Minimize $\quad TC = \sum_{i=1}^{M} \sum_{k=1}^{N} c_{ik}\, x_{ik} \quad + \quad \sum_{i=1}^{M} \sum_{k=1}^{N} \sum_{j=1}^{M} \sum_{l=1}^{N} c_{ijkl}\, x_{ik}\, x_{jl}$ (5.10)

$$\text{subject to} \quad \sum_{i=1}^{M} x_{ik} = 1, \quad \text{for all sites } k = 1,\ldots,N \tag{5.11}$$

$$\sum_{k=1}^{N} x_{ik} = 1, \quad \text{for all machines } i = 1,\ldots,M \tag{5.12}$$

$$x_{ik} = \{0,1\} \tag{5.13}$$

where x_{ik} and c_{ijkl} are as before and

$\quad c_{ik} =$ "fixed cost" of assigning machine i to site k.

The objective function (5.10) of the QAP with fixed costs is still quadratic because of the $x_{ik} \cdot x_{ik}$ terms. The constraints are the same as before.

If all of the c_{ijkl}'s are equal to zero or are identical, then the sum of the quadratic terms in the objective function is a nonnegative constant *Constant* and (5.10) reduces to

$$\text{Minimize } TC' = TC - Constant = \sum_{i=1}^{M} \sum_{k=1}^{N} c_{ik}\, x_{ik} \,, \tag{5.14}$$

which is a linear function. The objective function (5.14) together with the constraints (5.11)–(5.13) represent a *Linear Assignment Problem* (LAP). Computationally efficient procedures (e.g., the Hungarian Method) exist for finding (globally) optimal solutions for LAPs.

Likewise, if all of the c_{ik}'s are equal to zero or are identical, then the sum of the linear terms in the objective function (5.10) are either zero or some positive constant and (5.10) reduces to the original QAP objective function (5.4).

5.3 Department Layout

In machine layout, each activity (i.e., machine or workstation) is assumed to have a fixed shape. In a department layout, since the shape of each activity (e.g., department, production area, office, etc.) is not necessarily fixed, both the arrangement and the shapes of the activities are determined.

A *manufacturing department* is a group of machines and/or workstations considered as a single activity for facility layout purposes—although each machine and workstation has a fixed shape, the overall shape of the department is not fixed. *Nonmanufacturing departments* include shipping/receiving areas, storage areas, offices, cafeterias, etc. In a facility, departments do not have to be separated by physical walls or barriers unless there is a need to isolate particular departments from adjacent activities.

5.3.1 Flow Patterns between Departments

The flow between departments determines the overall flow pattern within a facility and typically consists of a combination of the following general flow patterns: straight line, U-shaped, S-shaped, and W-shaped. The beginning and end of the overall flow pattern correspond to the location of the principal receiving (entrance) and shipping (exit) docks, respectively, in the facility. Because of site restrictions, the location of the shipping (S) and receiving (R) docks are often fixed and become a constraint on the flow pattern within the facility.

The following possibilities exist for the location of the S/R docks along the perimeter of the facility:

1. *Single dock for S/R*—dock is used for both S and R (shared) since flow volume is too low to justify separate (dedicated) docks for S and R

2. *Multiple S/R docks at same location*—may increase dock utilization (compared to 3 and 4) since the same docks and equipment can be used for both S and R, and reduces service road requirements to/from the dock apron and truck waiting area (compared to 3 and 4)

3. *Same side but separate locations for S and R docks*—reduces service road requirements to/from the dock aprons and truck waiting areas (compared to 4) and may improve flow pattern within the facility (compared to 1 and 2)

4. *S and R docks on different sides*—increases service road requirements to/from the dock aprons and truck waiting areas (compared to 1–3), but may improve flow pattern within the facility.

The use of single areas in a facility for shipping and receiving is often associated with "centralized storage." It is becoming common in facilities using JIT-type production to have multiple docks around the perimeter of the facility, with docks located close to individual departments. This "decentralized storage" reduces unloading, storage, and travel times within the facility for the frequent deliveries associated with JIT (and may also reduce truck waiting area requirements), but at the possible cost of increases in the number of docks required and the service road requirements to/from the dock aprons. In some cases (termed "third-party logistics" or "3PL"), a logistics company will manage the S/R activities in a facility, up to and off of each production line—the owner of the facility only manages the line.

5.3.2 Block Layout Representations

A *block layout* is an initial, macro-level representation of the basic arrangement, shape, and size of the departments in a facility. It can serve as the basis from which more detailed layouts of the facility can be developed. A more detailed layout (e.g., CAD layout) may include information concerning, for example, the aisles, support columns, doors, and windows in the facility.

Several different representations can be used for block layouts. Each representation places different restrictions on the shape of each department in the layout, which results in different shapes for the overall layout. The following layout representations (see Figure 5.11) form a

hierarchy, starting from the least restrictive (map layout) to the most restrictive (grid and rectangular layouts):

(a) Map layout   (c) Orthogonal layout   (e) Rectangular layout

(b) Polygonal layout   (d) Grid layout
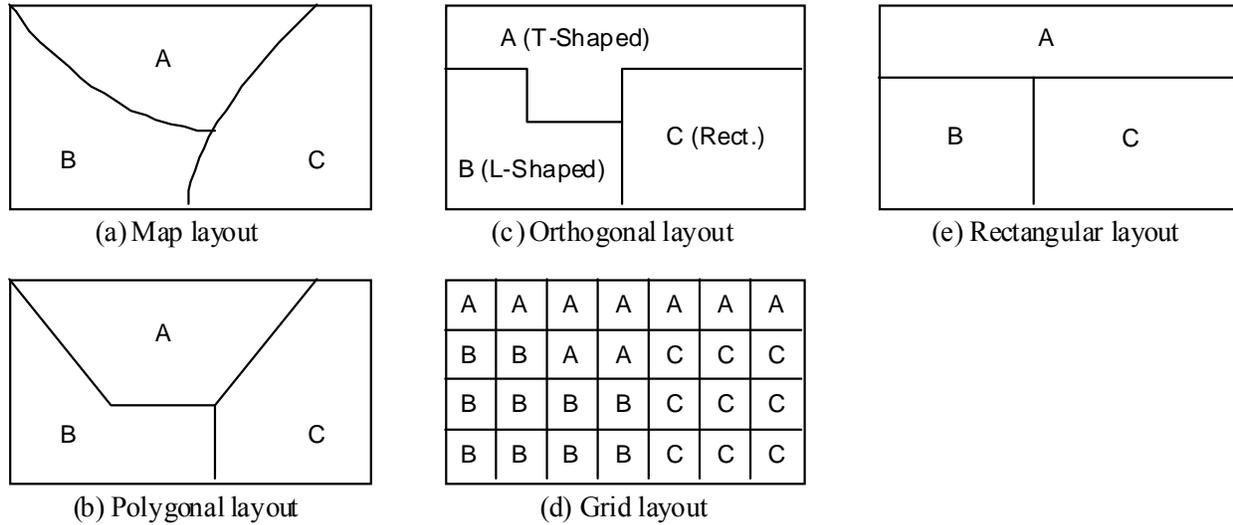
Figure 5.11. Hierarchy of block layout representations.

(a) *Map Layout*—each department is a closed region whose boundary is a curve such that the resulting overall layout is a planar graph or "map" (a graph is planar if it can be drawn so that each edge intersects the other edges only at their endpoints); planar graphs and their duals, which are also planar graphs, are useful for representing adjacency relationships between departments.

(b) *Polygonal Layout*—a map layout in which each department (and the resulting overall layout) is a simple polygon (i.e., a closed region whose boundary is composed of a series of straight lines).

(c) *Orthogonal Layout*—a polygonal layout in which the corners of each department (and the resulting overall layout) form either 90° or 270° angles.

(d) *Grid Layout*—an orthogonal layout in which each department (and the resulting overall layout) is composed of an integral number of equal-sized grid squares; grid layouts are used in CRAFT.

(e) *Rectangular Layout*—an orthogonal layout in which each department is a rectangle; the resulting overall shape of the facility need not be rectangular, although it must be an orthogonal polygon.

In practice, orthogonal layouts with simple-shaped (i.e., rectangular-, L-, or T-shaped) departments inside of a facility with an overall rectangular shape are usually preferred because they correspond to the type of facility that is least costly to build; they correspond closely to the final detailed building plans. Rectangular layouts limit the shape and proximity relationships of departments in the layout. The grid layouts produced by CRAFT are usually translated (by hand) into orthogonal layouts with simple-shaped departments and an overall rectangular shape.

5.3.3 CRAFT

CRAFT (Computerized Relative Allocation of Facilities Technique) is one of the earliest (1963) and best known (it was available for free on mainframes in the 1960s) programs for department layout problems.

CRAFT is based on the SDPI heuristic used for machine layout, but does not require the use of fixed sites at which the activities can be located. CRAFT is not appropriate for machine layout because the shape of activities (i.e., machines) would be altered by the program. Layouts are specified in CRAFT as "grid layouts," where the area and shape of each department is specified by the number and arrangement of unit-sized grid squares. Departments can be fixed in the layout so that their shape and location does not change, and dummy departments can be added to layouts to make the overall shape of the layout, for example, a rectangle with a desired aspect ratio.

In a similar manner as the SDPI heuristic, CRAFT uses the total cost of material flow (TC_{MF}) between departments as its objective. Starting from an initial layout of the departments, interchanges are considered to try and reduce TC_{MF}:

$$TC_{\mathrm{MF}} = \sum_{i=1}^{M} \sum_{j=1}^{M} f_{ij}\, h_{ij}\, d_{ij} \, , \tag{5.15}$$

where M = number of departments

f_{ij} = flow volume between departments i and j, measured in moves per time period

d_{ij} = distance between departments i and j, measured in grid-square units

h_{ij} = cost per move between departments i and j per unit distance.

Input Data

Input data to run CRAFT includes the following:

1. Specification of parameter values (including the type of interchanges to perform and the list of department(s), if any, that are to remain fixed)

2. From-to chart of flow volumes (f_{ij}'s)

3. Move cost matrix (h_{ij}'s)

4. Initial grid layout, where the area and shape of each department is indicated by the number and arrangement of the grid squares used for the department.

Distance

When fixed sites are used (as in machine layout), the distances (d_{ij}'s) can be calculated once; in CRAFT, the distances between departments are different for each different layout considered and must be continually recalculated. The distance between activities in the layout is calculated as the rectilinear distance between the centroids (i.e., centers of mass) of the departments.

Types of Interchanges

Unlike the SDPI heuristic, where only pairwise or 2-way interchanges are made, CRAFT allows the user to specify five different combinations of 2- and 3-way interchanges: 2-way, 3-way, 2-way followed by 3-way, 3-way followed by 2-way, and best of 2- and 3-way.

Allowed Interchanges

Departments can be interchanged if they are not fixed and either

1. they have equal areas (i.e., an equal number of grid squares), or

2. they are adjacent (i.e., they share a common border of positive length, not just meeting at a single point at the corners of two grid squares) and can be interchanged without causing either department to become split.

```
0     1   2   3   4   5   6   7   8
0  1  A   A   A   E   E   C   C   C
0  2  A       A   E   E   C   C   C
0  3  A       A   E   E   D   D   D
0  4  A   A   A   E   E   D       D
0  5  B   B   B   E   E   D       D
0  6  B   B   B   E   E   D   D   D
```

Figure 5.12. Grid layout representation in CRAFT.

5.3.4 Activity Relationships

The layout of activities (e.g., machines in a department or departments in a facility) should be based on the relationships between activities. Activity relationships can be based on quantitative and/or qualitative relationships. In facilities planning, activity relationships are often translated into proximity requirements. A positive relationship between two activities corresponds to the desire that the activities be as close as possible, if not adjacent, in the layout, while a negative relationship corresponds to having the activities as far apart, or at least not adjacent, in the layout. Proximity requirements can often be satisfied by other means than physical proximity; for example, communication networks, walls, or other barriers, etc.

Quantitative Relationships

For manufacturing departments, the principal quantitative measure of activity relationships is usually based on the flow of material between or within the departments. For nonmanufacturing departments, measures of other types of flow may be used; for example, the flow of people or information between activities.

Flow relationships are typically specified in from-to charts. Given M activities, a *from-to chart* is an $M \times M$ table or (square) matrix that represents the $M^2 - M$ relationships *from* each activity *to* each of another activities. The matrix need not be symmetric and the diagonal of the matrix is usually blank or zero. The equivalent flow volume matrix **W** is a from-to chart.

Qualitative Relationships

Qualitative activity relationships correspond to relationships that cannot be directly measured. Examples of possible qualitative relationships between activities include:

- *Material Flow*: flow volumes that are not convenient to directly measure, or, if measured, enables, after conversion to qualitative values, quantitative flow volumes to be combined with other qualitative measures.

- *Personnel Flow*: usually not necessary to determine quantitative measure.

- *Same Resource*: e.g., same support service used by two or more activities (e.g., spare parts).

- *Same Personnel*: e.g., one operator works on two machines.

- *Communication*: e.g., two activities need to be within shouting distance of each other.

- *Safety*: usually want to make sure activities are not located close together, e.g., fumes from one activity may ignite another activity.

- *Noise*: usually want to make sure activities are not located close together, e.g., a machine should not be located close to an office.

- *Structural*: e.g., two activities need to be in a refrigerated area of the facility, but otherwise they do not interact with each other.
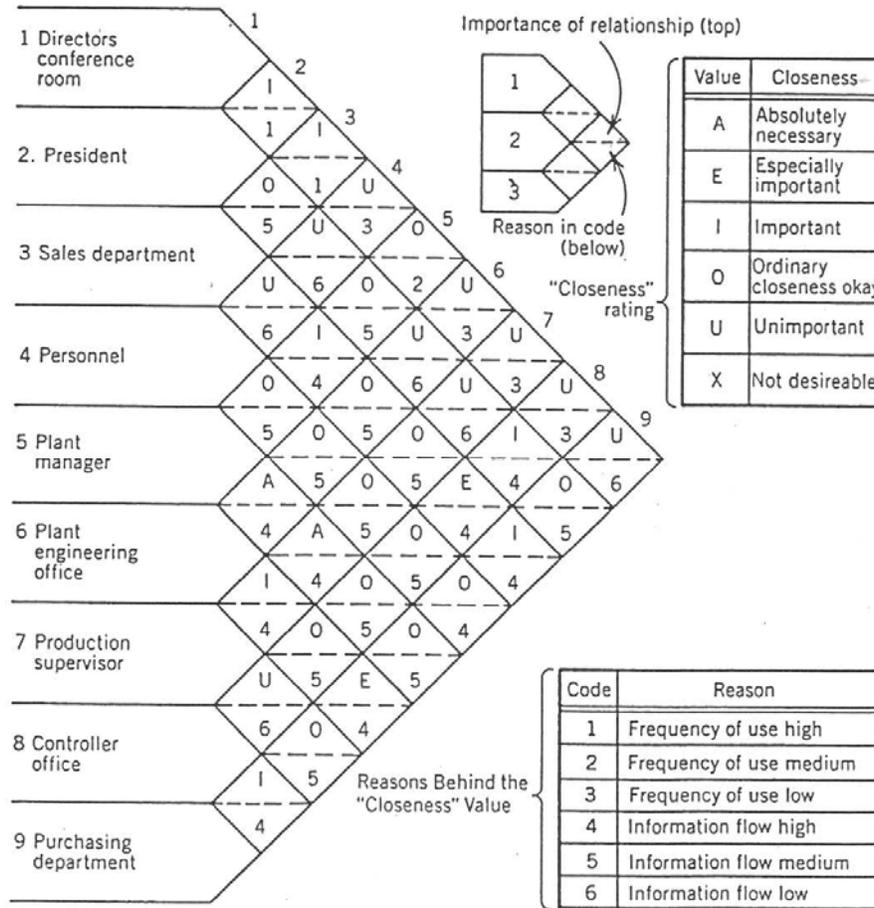
Figure 5.13. Relationship chart.[2]

Closeness Relationship Values

Closeness relationship values can be used to indicate physical proximity requirements between activities that have qualitative and/or quantitative relationships. The values A, E, I, O, U, and X are used to indicate the importance of locating two activities close together in a layout:

A: Must be adjacent

E: Adjacent, if possible

I: Close

O: Near

U: Don't care

X: Must not be adjacent

While flow volumes are "cardinal numbers," closeness relationship values are only "ordinal numbers" (e.g., an A, as compared to an E, value only represents the fact that closeness is more important, but not how much more important). Closeness values can be converted to (arbitrary) cardinal numbers to allow their comparison in algorithmic procedures. Quantitative relationship

measures (e.g., flow) can be converted to closeness relationship values in order to allow consideration of both quantitative and qualitative relationships.

Relationship Chart

Closeness relationship values are specified in a relationship chart (see Figure 5.13). Given M activities, a *relationship chart* is a triangular matrix that represents the $M(M-1)/2$ relationships *between* each pair of activities. A relationship chart can only represent symmetric relationships between activities, as compared to from-to charts, which can represent asymmetric relationships (e.g., $w_{ij} \neq w_{ji}$).

A *relationship graph* provides a spatial representation of a relationship chart. Each activity is a node in the graph and arcs are used to represent activity relationships. Typically, U (unimportant) closeness relationship values are not represented in the graph, and each of A, E, I, O, and X values are represented by arcs of different color, line style, and/or line thickness.

There may be up to $M(M-1)/2$ arcs in a relationship graph. For $M \geq 2$, the $3M - 6$ highest valued arcs (relationships) can be selected to form an *adjacency graph*, a planar subgraph of a relationship graph that shows each relationship satisfied in a layout through adjacency (i.e., activities with a high-valued relationship will be adjacent in the layout). The dual of an adjacency graph is a *map layout* (see Figure 5.11).

5.4 Space and Personnel Requirements

Once PP&S design is completed, the number and type of each machine and workstation is known and the number operators is known. From this information, initial estimates of the space and personnel requirements for each manufacturing department in a facility can be determined and used to develop a more detailed machine layout and an initial material handling system design for the department. Additional analysis is required to estimate the space and personnel requirements for the nonmanufacturing departments in the facility. The initial estimates of the space and personnel requirements of each department can then be used in developing an initial department layout for the facility and the site plan for the facility.

5.4.1 Personnel Requirements

The personnel requirements for each manufacturing department include all machine/workstation operators and all support personnel dedicated to the department (e.g., supervisors, material handling operators, etc.). The personnel requirements for the entire facility includes the personnel requirements for each manufacturing department together with all support personnel not dedicated to manufacturing departments (e.g., administrative staff, maintenance workers, etc.). Machine/workstation operator requirements are directly available from PP&S design; only estimates can be made of the number of additional support personnel required.

An estimate of the personnel requirements for the entire facility is needed in order to determine a number of aspects of the facility and site plan:

- Number of parking spaces needed for employees.

- Number and type of restrooms needed in the facility.

- Capacity of food service areas.

5.4.2 Space Requirements

The total space requirements for the site on which the facility is located is the sum of the space requirements "under-roof" (the facility building) and "outside" (everything except the facility building). Figure 5.15 shows a generic facility. The facility includes low- and high-bay areas, and includes arrows that show a variety of maximum distance limitations. These limitations result in constraints on the size and shape of the facility.

The space requirements for a facility can be developed in two ways:

1. *Calculate:* Starting "from the ground up," calculate the space requirements for each machine/workstation, which leads to the space requirements for each manufacturing department, which, together with the space requirements of the non-manufacturing departments and main aisles, leads to the total space requirements for the entire facility.

2. *Template:* Arrange scaled templates of the major equipment to be used in the facility (e.g., machines, etc.) into a drawing and then design the remainder of the facility "around" the templates. Traditionally, paper cutouts were used as templates; now, software packages like FactoryCAD provide large libraries of templates.

In practice, a combination of both approaches can be used.

In high-bay areas of a facility, mezzanines can be used as an inexpensive means of providing additional storage or office space. A mezzanine (see Figure 5.14) is a stand-alone structure constructed within an existing building to make use of clear space over activities not requiring much headroom (e.g., restrooms, block storage, etc.). At least 14 ft of clear space is needed for a mezzanine.

Figure 5.14. Mezzanine.

107

Frontage Along Main Road

Truck Entrance   Employee/Visitor Entrance

Visitor Parking
(2-5% handicapped)

Visitor Entrance

LOW BAY (15')

Food
Service   Offices

Manufacturing

Emergency Exit

< 150' w/o sprinkler
< 200' w/ sprinkler

< 1,000'

< 500'

Employee Entrance

Employee
Parking

HIGH BAY (30')

Mezzanine

Shipping/Receiving

< 500'

Rest-
room

< 200'

Future Expansion
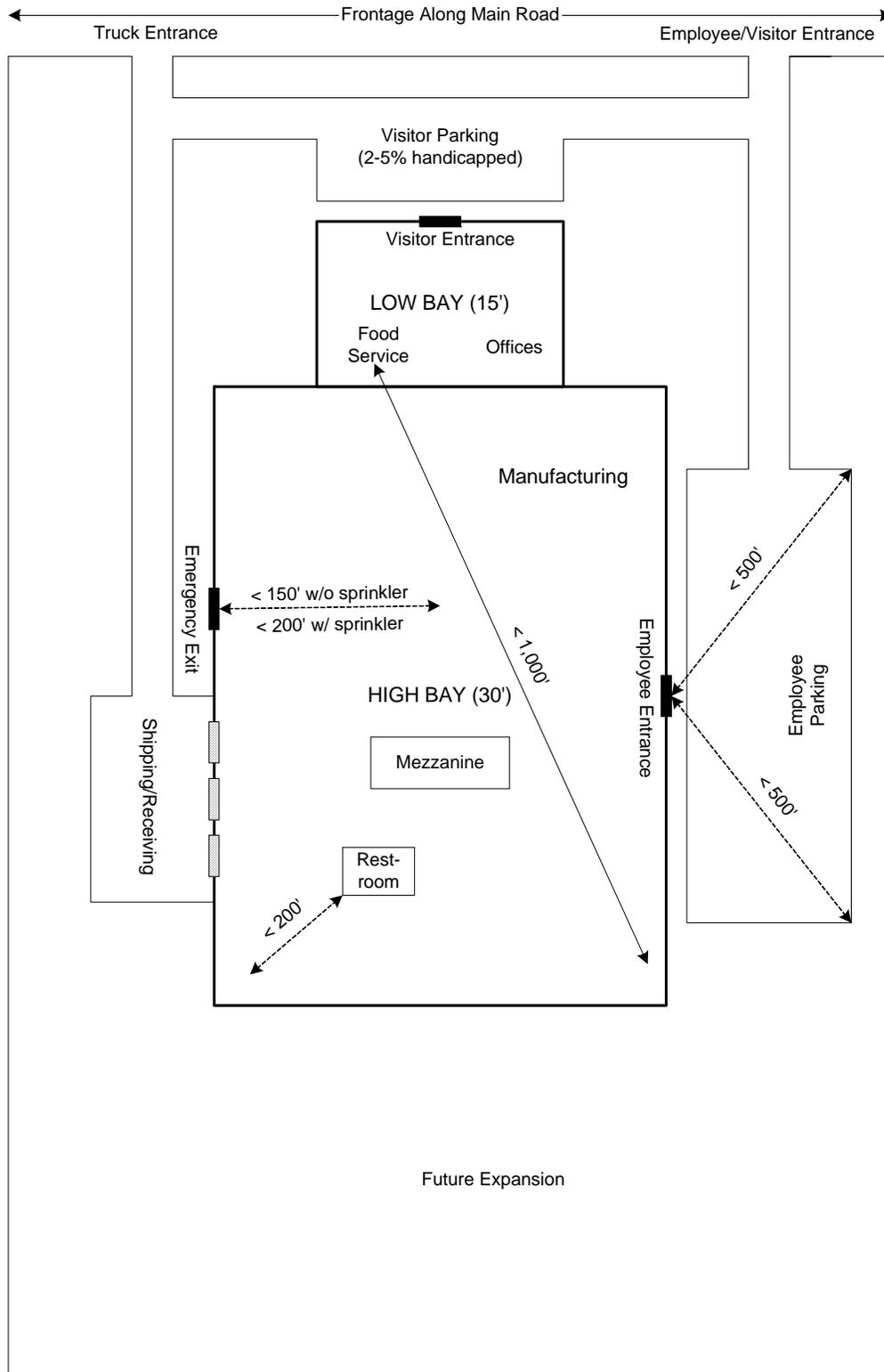
Figure 5.15. Generic facility.

Machines/workstations

In product layouts, production is typically organized around a series of individual machines and/or assembly workstations that may be connected by material handling equipment (e.g., conveyors) that provide fixed-path material flow and the small amount of WIP storage needed. In process layouts, production is typically organized around individual fabrication and/or assembly workstations, with variable-path material flow between the workstations occurring along aisles and high WIP storage requirements at each workstation; each fabrication workstation may include one or more machines together with additional support equipment and storage space for WIP.

The space requirements of a machine/workstation can include the following:

1. *Static machine dimensions*—static width, depth, and height at maximum points.

2. *Machine travel*—maximum travel to the left and right and toward and away from the operator.

3. *Personnel area*—operator at machine together with ingress and egress.

4. *Storage area*—WIP, raw materials, waste and scrap, tools, fixtures, jigs, dies, and maintenance materials.

5. *Support equipment*—bench, tool locker, and maintenance/setup equipment.

The fabrication workstations in a process layout may include 1–5, while the machines and assembly workstations in a product layout may only include 1–3. Machine dimensions and travel are typically available from data sheets.

Manufacturing departments

The total space requirements for a manufacturing department can be estimated as the sum of the space requirements of each machine/workstation together with the space required for departmental aisles and department-wide fixed material handling equipment (MHE). In a process layout, departmental aisles are typically used for both material transport and personnel movement, with little fixed MHE outside of each workstation; in a product layout, departmental aisles may only be used for personnel movement, with fixed MHE connecting each machine/workstation.

Until the material handling system has been designed for the department, an accurate machine layout cannot be determined because material handling costs are not known (e.g., the h_{ij}'s); thus, the actual aisle and MHE space requirements will not be known. Instead, in order to allow an initial department layout to be determined, an aisle/MHE allowance (expressed as a percentage of the total machine/workstation space requirements) can be used. Once the material handling system design is completed, the aisle/MHE allowance can be replaced by their actual space requirements.

Nonmanufacturing departments

The total space requirements for the nonmanufacturing departments in a facility can be estimated once (1) the personnel requirements for the entire facility and (2) the storage and shipping/receiving requirements for materials have been determined.

The types of nonmanufacturing departments in a facility can include the following:

- *Offices*—offices for administrative, support, and engineering staff; meeting/conference rooms; etc.

- *Storage/warehousing*—dedicated areas outside of each manufacturing department for the storage of raw materials, in-process inventories, and finished goods.

- *Shipping/receiving*—dockboard area, maneuvering area for MHE used for loading/unloading, buffer/staging areas, container and trash storage, offices, and trucker's lounge.

- *Restrooms*—located within 200 feet of every permanent workstation.

- *Food services*—located within 1,000 feet of every permanent workstation; possibilities include off-premises dining or a cafeteria with either vending machines, a serving line (over 200 employees), or a full kitchen (over 400 employees); a cafeteria can also be used as a meeting room. The typical space allowance per person in an industrial facility is 12 to 15 square feet.[3]

- Health services, locker rooms, and washrooms.

- *Plant services*—HVAC, pumps, generators, etc.

- *Emergency exits*—an exit must be within 150 ft of any point in a building without sprinklers and within 200 ft of any point in a building with sprinklers.[4]

The space requirements for many of the types of service areas (types 4–7) may be small and should not be included as separate departments when determining an initial block layout, although they should be included in the final detailed facility layout. The requirements of many of the service areas and other aspects a facility are specified in OSHA (Occupational Safety and Health Act) standards.

Total facility

The total space requirements for the entire facility can be estimated as the sum of the space requirements of each department together with the space required for the main interdepartmental aisles and any facility-wide fixed MHE connecting different departments. Until the material handling system for interdepartmental material flow has been designed, an accurate department layout can not be determined; thus, the actual main aisle and MHE space requirements will not be known. Instead, in order to allow an initial block layout to be determined, an aisle/MHE allowance (expressed as a percentage of the total department space requirements) can be used. Once the material handling system design for interdepartmental material flow is completed, the aisle/MHE allowance can be replaced by their actual space requirements.

Total site

The total space requirements for the site on which the facility is to be located can be estimated as the sum of the space requirements under-roof and outside. The total space under-roof is the sum of the space requirements (or "footprint") of each separate building (facility) located on the site. The total space outside is the site's total land area minus the space under-roof. Outside (or yard) areas can include the following areas.

Employee and visitor parking lots

Employee parking lot design requires estimates of the total personnel requirements for each shift; the employee parking lot should accommodate the maximum number of employee vehicles on site at any one time (including shift changes); every employee parking space should be located within 500 feet of an entrance to the facility, and local authorities may regulate the number of parking spaces per employee and the number of handicapped parking spaces; parking lot design involves issues concerning the tradeoffs between accessibility, turnover, and space utilization associated with different parking angles and aisle configurations and the use of random versus dedicated parking space assignment (the issues in parking lot design are similar to those faced in warehouse design).

Dock aprons

Outside space required to access truck docks, including truck turning areas (drive-in docks, where trucks enter the facility, are usually not economical); apron depth and dock bay width depend on the angle of the dock: 90° docks are preferred to "finger docks" (less than 90°) because, even though they maximize apron depth, they minimize bay width (which reduces facility costs).

Truck waiting area

Should be located adjacent to the dock apron and provide space to hold the maximum number of trucks waiting at any one time; blocking in the waiting area makes it difficult to sequence trucks at the docks; trailers parked in the waiting area can be used for short-term storage, subject to "demurrage" (a penalty charge assesses by carriers for detention of their trailers beyond a specified free time (e.g., 30 days)).

Service roads and driveways

Allows access to/from public roads and a site's outside areas (e.g., truck waiting areas, dock aprons, and parking lots); access should be planned so that trucks to not need to back onto the site; two-way service roads should be 24 feet wide and one-way roads should be 12 feet wide; if possible, all truck traffic should circulate counterclockwise because left turns are easier and safer to make than right turns (assuming a left-side steering wheel).

Railroad spur

Allows access to/from off-site rail line and on-site rail docks.

Yard storage

Outside storage of raw materials, finished goods, and waste (can include cooling ponds and waste treatment areas).

Unused site space

Site's total land area minus the space under-roof and outside areas 1–6; the amount of unused space for a site can be influenced by several factors: future expansion possibilities, and minimum land-to-building ratios and minimum building–to–site-border distances decreed by local authorities.

5.5 References

The following sources are recommended for further study:

Heragu, S., 1997, *Facilities Design*, PWS: Boston.

Hanna, S.R., and Konz, S., 2004, *Facility Design and Engineering*, 3rd Ed., Scottsdale, AZ: Holcomb Hathaway.

Sule, D.R., 1994, *Manufacturing Facilities: Location, Planning, and Design*, 2nd Ed., Boston: PWS.

Tompkins, J.A., White, J.A., Bozer, Y.A., and Tanchoco, J.M.A., 2003, *Facilities Planning*, 3rd Ed., Wiley.

6. Warehousing

6.1 The Need for Storage and Warehousing

A warehouse is the point in the supply chain where raw materials, work-in-process (WIP), or finished goods are stored for varying lengths of time. A *public warehouse* is a business that rents storage space to other firms on a month-to-month basis. They are often used by firms to supplement their own private warehouses. Warehouses can be used to add value to a supply chain in two basic ways:

- Storage—allows product to be available where and when it's needed.

- Transport economies—allows product to be collected, sorted, and distributed efficiently.

Warehouses only add value if the benefits of storing products in a warehouse enough to offset the additional cost associated with carrying any inventory. Other potential benefits associated with storage include the following: time bridging, which allows product to be available when it is needed (e.g., storing spare machine parts at the facility); processing, where for some products (e.g., wine), storage can be considered as a processing operation because the product undergoes a required change during storage; and securing, e.g., nuclear waste storage.

In production, ideally, raw material should arrive at a manufacturing facility just when it is needed and then immediately processed, the resulting products should be fabricated and assembled without delay, and the final finished products should be immediately shipped to their customers; in this situation (what could be termed pure "Just-In-Time" or JIT) there is little need for buffering or storing materials. In practice (including real-world JIT), there usually are economic benefits associated with the buffering and/or storage of raw materials, work in process (WIP), and/or finished goods.

In distribution, the ideal of no storage can sometimes be realized using cross docking, where there is a direct flow of material from trucks at the receiving docks to the shipping docks without buffering or storage in-between, but cross docking requires detailed planning and coordination (e.g., implemented using EDI) that in many cases may not be feasible.

In most cases, the benefits associated with buffering and storage are due to the fixed costs associated with the other elements of production and the impact of variability pooling on achieving a target service level. Storing a product allows the other elements of production to

operate more efficiently on a per-unit basis because the fixed costs associated with utilizing the element can be spread over more products; e.g., storing up to a truckload of product in a facility reduces the per-unit costs of shipping, and WIP buffering or storage enables batch production, which reduces the per-unit setup costs.

6.2 Storage System Design

Each distinct type of load is termed an **item** or *stock-keeping unit* (or **SKU**); e.g., each different style, size, and color of a garment would be assigned a unique SKU. Units of each item are stored in **slots** (short for storage location). A slot is a generic term for any of a variety of different types of identifiable storage locations (e.g., racks, bins, marked-off floor areas for block storage). Each slot-item combination has an associated capacity corresponding to the number of units of the item that can be stored in the slot.

Table 6.1. Design Trade-Off

min *Building Costs*	vs.	min *Handling Costs*
⇕		⇕
max *Cube Utilization*	vs.	max *Material Accessibility*
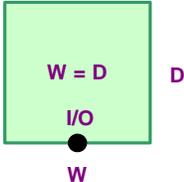
	Shape vs.	
Square shape minimizes perimeter length for a given area, thus minimizing building costs		Aspect ratio of 2 ($W = 2D$) minimizes expected distance from I/O port to slots, thus minimizing handling costs
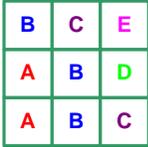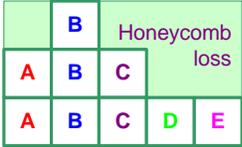
	Storage vs.	
Maximizes cube utilization, but minimizes material accessibility		Making at least one unit of each item accessible decreases cube utilization

The handling costs for the units within a SKU can usually be minimized by always storing and retrieving a unit at the nearest (i.e., least handling effort or cost) available location, or what is termed a *closest open location* (or COL) policy. As long as the inventory levels of each SKU are controlled, a COL policy will result in an approximate uniform rotation of the items; but, if inventory is not controlled, using a COL policy can result in items remaining at far away slots for

a long time. If a strict uniform rotation of the items is required (e.g., due to the items being perishable), then a *first-in, first-out* (or FIFO) policy can be used. In addition, a *last-in, first-out* (or LIFO) policy can be used.

6.2.1 Design Trade-Off

As shown in Table 6.1, warehouse design involves the trade-off between building and handling costs. Handling costs usually dominate building costs when a warehouse is only used for short-term storage, while building costs dominate for longer-term storage.

6.2.2 Storage Locations

Each accessible storage location in a warehouse is assigned a unique address. Multiple units of an item assigned to a single location correspond to the capacity of the location. It is common to alternate between numeric and alphabetic characters in an address to improve readability, and to use even and odd numbers to designate each side of a down aisle.
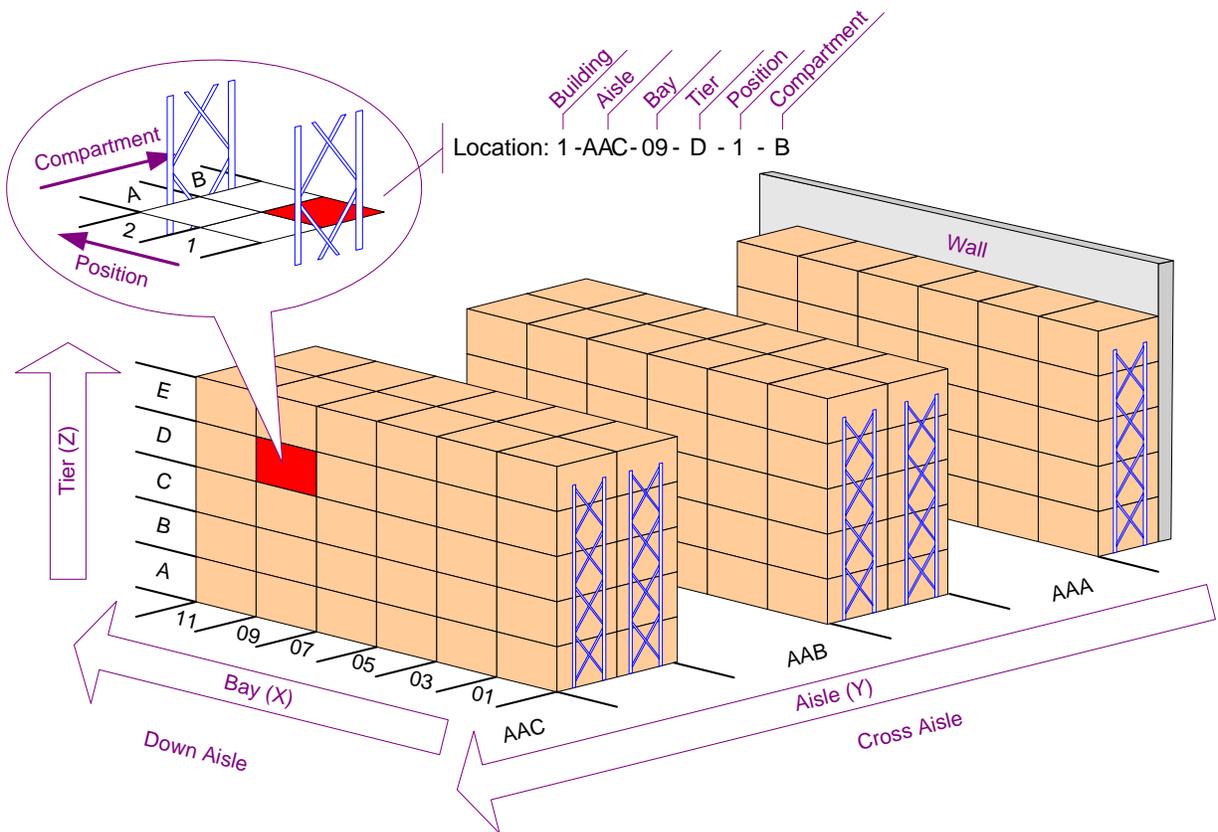

Figure 6.1. Storage locations.

The single address scheme shown in the Figure 6.1 can be used for each different storage medium in the warehouse:

- Pallet racks: Compartment dimension not used since only the front unit of each position is accessible.

- Shelves: All dimensions can be used if compartment dimension is accessible.

- Drawers: Position dimension not used if drawer has odd shaped compartments.

- Block stacking: Only building, aisle, and bay dimensions used to address each lane of storage.

- Misc. locations: Receiving, shipping, holding areas, outdoor trailer storage, etc., can all be given unique addresses.


| (a) Dedicated | (b) Randomized | (c) Class-based |
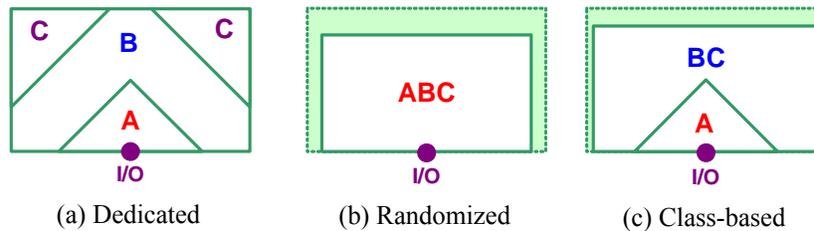
Figure 6.2. Storage area for SKUs, A, B, and C under different policies.

6.2.3 Storage Policies

For multiple SKUs, three types of storage policies (see Figure 6.2) can be used to select storage locations (or *slots*):

1. **Dedicated (or Fixed Slot) Storage**—each SKU has a predetermined number of slots assigned to it.

   The total capacity of the slots assigned to each SKU must equal the storage space corresponding to the maximum on-hand inventory of each *individual* SKU, where the actual storage space might be greater than this due to "honeycomb loss."

   Minimizes handling cost and maximizes building costs. Control is not difficult because each lane can be identified with a permanent label.

2. **Randomized (or Open Slot or Floating Slot) Storage**—each SKU can be stored in any (usually the closest) available slot.

   The total capacity of all the slots must equal the storage space corresponding to the maximum *aggregate* on-hand inventory of all of the SKUs, where the actual storage space might be greater than this due to honeycomb loss.

   Minimizes building cost and maximizes handling costs. Control is more difficult than dedicated storage because the identity of SKU stored at each slot needs to be recorded for retrieval purposes.

3. **Class-based Storage**—a combination of dedicated and randomized storage, where each SKU is assigned to one of several different storage classes.

   Randomized storage is used for each SKU within a class, and dedicated storage is used between classes. Building and handling costs in-between dedicated and randomized.

   Classes can be formed from SKUs whose individual on-hand inventory is negatively correlated (or, at least, uncorrelated).
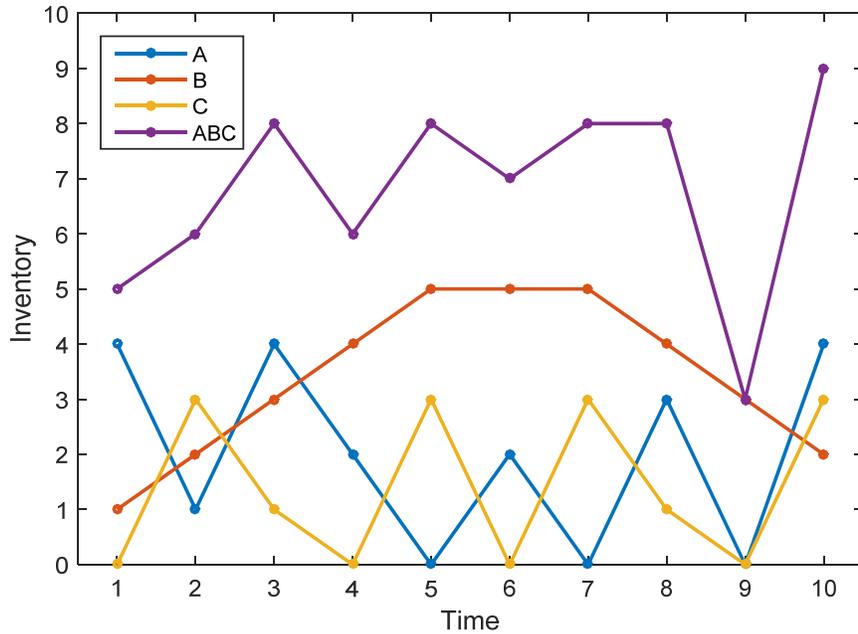

Figure 6.3. Inventory profiles for dedicated and randomized storage policies.

Table 6.2. Inventory and Storage Requirements for Different Storage Policies

	Dedicated			Random	Class-Based		
Time	A	B	C	ABC	AB	AC	BC
1	4	1	0	5	5	4	1
2	1	2	3	6	3	4	5
3	4	3	1	8	7	5	4
4	2	4	0	6	6	2	4
5	0	5	3	8	5	3	8
6	2	5	0	7	7	2	5
7	0	5	3	8	5	3	8
8	3	4	1	8	7	4	5
9	0	3	0	3	3	0	3
10	4	2	3	9	6	7	5
M_i	4	5	3	9	7	7	8

117

In the example shown in Figure 6.3 and Table 6.2, the on-hand inventory over 10 time periods for SKUs A, B, and C and the aggregate inventory for all three SKUs. If a dedicated storage policy is used, then a fixed number of slots must be reserved for each SKU for the entire 10 periods. In this example, SKU A has a peak of 4 in periods 1, 3, and 10; SKU B has a peak of 5 in periods 5 through 7; and SKU C has a peak of 3 in periods 2, 5, 7, and 10. If a randomized policy is used, then the aggregate inventory has a single peak of 9 in period 10. As long as the on-hand inventory of each SKU is not at its maximum at the same time, randomized storage will require a lesser number of slots as compared to dedicated storage and the minimum class-based policy (AB-C).

Table 6.2 shows the same on-hand inventory profiles shown in Figure 6.3 for dedicated and randomized storage. In addition, the three possible class-based storage policies, A+BC, B+AC, and C+AB, are shown, where the single SKU A forms one Class A and the aggregate of levels for SKUs B and C form a second Class BC, etc. (the single SKU classes are not shown because they are the same as the profiles for dedicated storage). Assuming that the storage space required for each unit of each SKU is the same and the capacity of each slot is one unit, the total number of slots for each storage alternative is as follows:

- Dedicated = sum of max SKU levels = 4 + 5 + 3 = 12 slots

- Randomized = max aggregate level = 9 slots.

- Classes C+AB = 3 + 7 = 10 slots (the other two possible class-based require 12 slots).

Based on just storage space requirements, a randomized policy would be preferred; but a dedicated or class-based policy may be preferred because they can sometimes reduce the handling requirements enough compared to randomized to offset their increase in storage requirements (this is an example of the trade-off between building and handling cost). In general, as long as the on-hand inventory of each SKU is not at its maximum at the same time, randomized storage will require a lesser number of slots as compared to dedicated storage.

A combination of dedicated and randomized storage termed "supermarket" storage is used in most less-than-unit-load order picking operations, where randomized storage is used for *reserve stock* and dedicated is used for *forward stock*. Cartons are picked from forward stock (in flow-through racks), and full pallet loads of cartons are taken from reserve stock (in bulk storage) and used to replenish the forward stock.

6.2.4 Cube Utilization and Honeycomb Loss

When storing multiple SKUs in a single region, full utilization of all of the available space is not desirable because it could result in some items not being accessible. *Honeycomb loss*, the price paid for accessibility, is the unusable empty storage space in a lane or stack due to the storage of only a single SKU in each lane or stack since storing items from different SKUs would block access. The empty space associated with partially filled lanes and stacks is termed "horizontal" and "vertical" honeycomb loss, respectively (see Figure 6.4). When a single SKU is stored in a

region, there need not be any honeycomb loss since the depth and height of the region can exactly match the storage space need for the SKU.
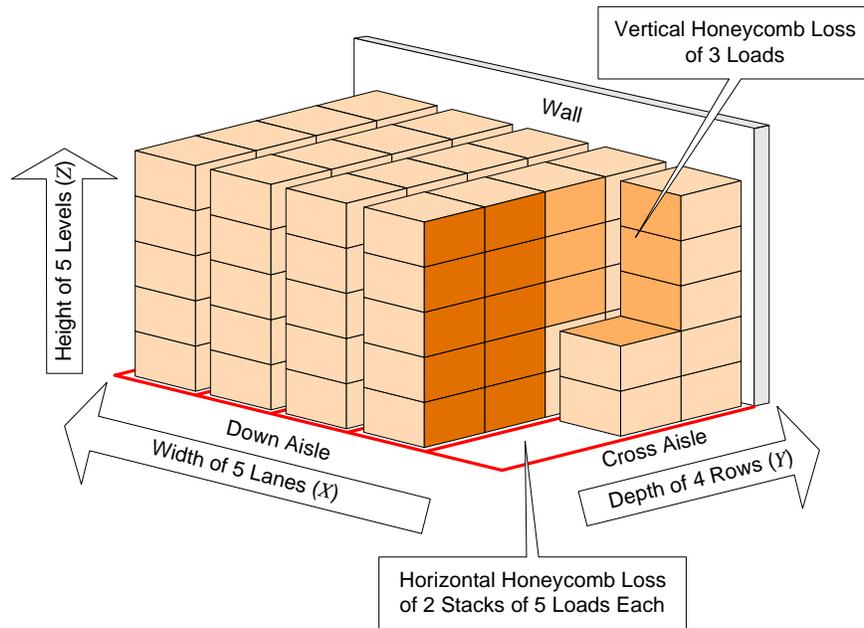

Figure 6.4. Horizontal and vertical honeycomb loss.

Cube utilization is the percentage of the total space (or "cube") required for storage actually occupied by the loads being stored. There is usually a trade-off between cube utilization and material accessibility:

increasing cube utilization $\Rightarrow$ decreased accessibility, and

increasing accessibility $\Rightarrow$ decreased cube utilization.

Bulk storage using block stacking can result in the minimum cost of storage since cube utilization is high and no storage medium is required, but material accessibility is low since only the top of the front stack is accessible and loads at bottom of a stack must not require support. Storage racks are used when support and/or material accessibility is required.

Given a contiguous region where several different SKUs are to be stored, the principal decision variable for deep-lane storage is D, the number of rows of storage for the region. Given a load depth of y, the resulting lane depth is $Y = yD$. Different row values for the region will result in different cube utilizations. Since the space occupied by the items is assumed to be known, cube utilization can be determined once the total space is determined, where the total space is the item space plus honeycomb loss and the space used for access (e.g., down aisles). Given D and assuming *identical size loads for all items*, the cube utilization for dedicated and randomized storage can estimated as follows:

$$\text{Cube utilization} = CU = \frac{\text{item space}}{\text{total space}} = \frac{\text{item space}}{\text{item space} + \left(\begin{array}{c}\text{honeycomb}\\\text{loss}\end{array}\right) + \left(\begin{array}{c}\text{down aisle}\\\text{space}\end{array}\right)}$$

119

$$CU(\text{3-D}) = \begin{cases} \dfrac{x \cdot y \cdot z \cdot \sum_{i=1}^{N} M_i}{TS(D)}, & \text{dedicated} \\[4mm] \dfrac{x \cdot y \cdot z \cdot M}{TS(D)}, & \text{randomized} \end{cases} \qquad (6.1)$$

$$CU(\text{2-D}) = \begin{cases} \dfrac{x \cdot y \cdot \sum_{i=1}^{N} \left\lceil \dfrac{M_i}{H} \right\rceil}{TA(D)}, & \text{dedicated} \\[4mm] \dfrac{x \cdot y \cdot \left\lceil \dfrac{M}{H} \right\rceil}{TA(D)}, & \text{randomized} \end{cases} \qquad (6.2)$$

where

$x =$ lane/unit-load width

$y =$ unit-load depth

$z =$ unit-load height

$M_i =$ maximum number of units of SKU i

$M =$ maximum number of units of all SKUs

$N =$ number of different SKUs

$D =$ number of rows

$TS(D) =$ total 3-D *space* (given D rows of storage).

$TA(D) =$ total 2-D *area* (given D rows of storage).

Although not a cube, 2-D "cube" utilization is often easier to use and is equivalent to 3-D utilization as long as the height of storage is, and will remain, fixed. In (6.2), $CU(\text{2-D}) = (stack\ area)/(total\ area)$, where the stack area is the product of the 2-D footprint of each stack of H items, xy, and the total number of stacks, which is $\lceil M/H \rceil$ for randomized storage. The total area is given by (6.4).

Defining the *effective lane depth* as the depth of the lane plus half of the width of the down aisle in front of the lane, the total space required, as a function of lane depth, is

$$\text{Total space (3-D):} \quad TS(D) = X \cdot \underbrace{\left(Y + \frac{A}{2} \right)}_{\text{Eff. lane depth}} \cdot Z = xL(D) \cdot \left(yD + \frac{A}{2} \right) \cdot zH, \qquad (6.3)$$

where

$X =$ width of storage region (row length)

$Y =$ depth of storage region (lane depth)

Z = height of storage region (stack height)

A = down aisle width

$L(D)$ = number of lanes (given D rows of storage)

H = number of levels.

In most cases, only the 2-D area of a storage region is needed since the height of the region is fixed (see Figure 6.5). To convert the total 3-D space to 2-D area:

$$\text{Total area (2-D): } TA(D) = \frac{TS(D)}{Z} = X \cdot Y^{\text{eff}} = xL(D) \cdot \left(yD + \frac{A}{2} \right) \quad (6.4)$$
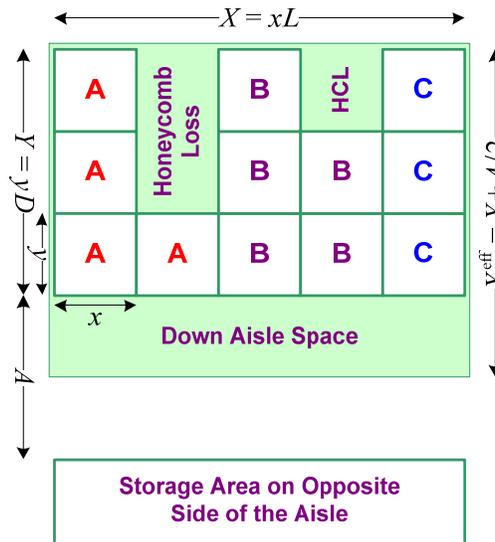

Figure 6.5. Total area of a storage region.

Given D, the total number of lanes required for storage in the region can be estimated as follows:

$$\text{Number of lanes: } L(D) = \begin{cases} \sum_{i=1}^{N} \left\lceil \dfrac{M_i}{DH} \right\rceil, & \text{dedicated} \\[2em] \left\lceil \dfrac{M + NH\left(\dfrac{D-1}{2}\right) + N\left(\dfrac{H-1}{2}\right)}{DH} \right\rceil, & \text{randomized } (N > 1) \end{cases} \quad (6.5)$$

Note: in practice, there should be at least two lanes for each item to facilitate stock rotation so that all of the older units in one lane can be picked even after newer units are stored in the other lane.

For dedicated storage, the honeycomb loss can be directly determined for each item via the ceiling operation $\lceil \bullet \rceil$ in (6.5), which then determines the corresponding number of lanes required; for randomized storage, since only the total maximum number of units of items, M, is known and not the specific the number of each SKU that comprise this total at the exact time that

the total reaches its maximum (unless the SKU's inventory levels are not perfectly correlated), the honeycomb loss can only be estimated by assuming that, at the maximum inventory level, the number of items in the partially filled lane and/or stack for each SKU is equally likely (see Figure 6.6).


$$\text{Expected Loss:} \frac{1}{D}\big[(D-2)+(D-1)\big]=\frac{1}{D}(1+2)=\frac{1}{D}\sum_{i=1}^{D-1}i=\frac{1}{D}\left(\frac{(D-1)D}{2}\right)=\frac{D-1}{2}=1$$

Figure 6.6. Expected honeycomb loss for dedicated storage.

If the SKUs' inventory levels are uncorrelated and items are either stored or retrieved at a constant rate so that, on average, half of the M_i items of SKU i are likely to be present at any given time, then

$$M = \left\lfloor \sum_{i=1}^{N} \frac{M_i}{2} + \frac{1}{2} \right\rfloor. \quad \text{(uncorrelated, constant demand)} \tag{6.6}$$

This estimate can be increased to include safety stock for each item, SS_i. For example, if the order size of each of three different products is 50 units and 5 units of each item are held as safety stock, then

$$M = \left\lfloor \sum_{i=1}^{N}\left(\frac{M_i - SS_i}{2} + SS_i\right) + \frac{1}{2} \right\rfloor = \left\lfloor 3\left(\frac{50}{2}+5\right)+\frac{1}{2}\right\rfloor = 90. \tag{6.7}$$

Given the number of lanes L, the (3-D) honeycomb loss is:

$$\text{Honeycomb loss} = \begin{cases} xyz\left(L(D)DH - \sum_{i=1}^{N}M_i\right), & \text{dedicated} \\ xyz\left(L(D)DH - M\right), & \text{randomized} \end{cases}. \tag{6.8}$$

An estimate of (3-D) honeycomb loss for randomized storage that is used to determine the expected number of lanes in (6.5) is

$$\text{Expected honeycomb loss} = Nxyz\left[\underbrace{H\left(\frac{D-1}{2}\right)}_{\text{Partial lane loss}} + \underbrace{\left(\frac{H-1}{2}\right)}_{\text{Partial stack loss}}\right] \tag{6.9}$$

Given the number of lanes of storage, the corresponding (3-D) down aisle space is

$$\text{Down aisle space} = xL(D)\cdot\frac{A}{2}\cdot zH\ . \tag{6.10}$$

Optimal Lane Depth

The lane depth that maximizes cube utilization corresponds to best compromise between honeycomb loss (6.8) and down-aisle space loss (6.10) (see, also, Figure 6.7).

Table 6.3. Cube Utilization for Dedicated Storage

Storage Area at Different Lane Depths	Item Area	Lanes	Total Area	Cube Util.

	Item Area	Lanes	Total Area	Cube Util.
$D = 1$ / $A/2 = 1$	12	12	24	50%
$D = 2$ / $A/2 = 1$	12	7	21	57%
$D = 3$ / $A/2 = 1$	12	5	20	60%

Example: Dedicated Storage

The optimal value for dedicated storage can be determined by calculating the utilization associated with each stack using for D ranging from 1 to $\max\{M_i'\}$. In this example for SKUs A, B, and C (see Table 6.3), $x = 1$, $y = 1$, $z = 1$, $M_A = 4$, $M_B = 5$, $M_C = 3$, $N = 3$, $A = 2$, and $H = 1$. Starting with a lane depth of $D = 1$, which results in 12 lanes of storage and a cube utilization of 50%, the value used for D is increased until either the cube utilization starts to decrease or D reaches the maximum number of units required for any of the SKUs, $\max\{M_i\}$ (at which point there would be a single lane for each SKU). In the example, the cube utilization is still increasing

at $D = 3$, so $D = 4$ would need to be considered next. At $D = 3$, the honeycomb loss is 3 units and the down-aisle space loss is 5 units, for a total loss of $8 = 20 - 12$ units.

Example: Randomized Storage

Unlike dedicated storage, where the optimal lane depth corresponding to the maximum cube utilization is determined by value checking each different value of D, the optimal lane depth, D^*, for randomized storage can be determined by direct calculation using an analytical approximation formula. Since the item space is constant in (6.2), cube utilization can be maximized by minimizing total space. Minimizing (6.3) (ignoring the ceiling operation in (6.5)) by solving for D in $dTS(D)/dD = 0$ results in the following expression to determine D^*, the lane depth (in rows) that maximizes cube utilization:

$$\text{Optimal lane depth for randomized storage (in rows):} \quad D^* = \left\lfloor \sqrt{\frac{A(2M-N)}{2NyH}} + \frac{1}{2} \right\rfloor. \quad (6.11)$$

Taking the floor of $\lfloor \bullet + 0.5 \rfloor$ in (6.11) forces the result to the nearest integer. Equation (6.11) provides only an approximation of the optimal lane depth because the ceiling operation in (6.5) is ignored; to calculate the optimal depth, actual $TS(D)$ values should be directly calculated for several D values close to D^*.

	1	2	3	4	5	6	7	8	9	10
Item Space	24,000	24,000	24,000	24,000	24,000	24,000	24,000	24,000	24,000	24,000
Honeycomb Loss	1,536	3,648	5,376	7,488	9,600	11,712	13,632	15,936	17,472	20,160
Aisle Space	38,304	20,736	14,688	11,808	10,080	8,928	8,064	7,488	6,912	6,624
Total Space	63,840	48,384	44,064	43,296	43,680	44,640	45,696	47,424	48,384	50,784

Lane Depth (in Rows)
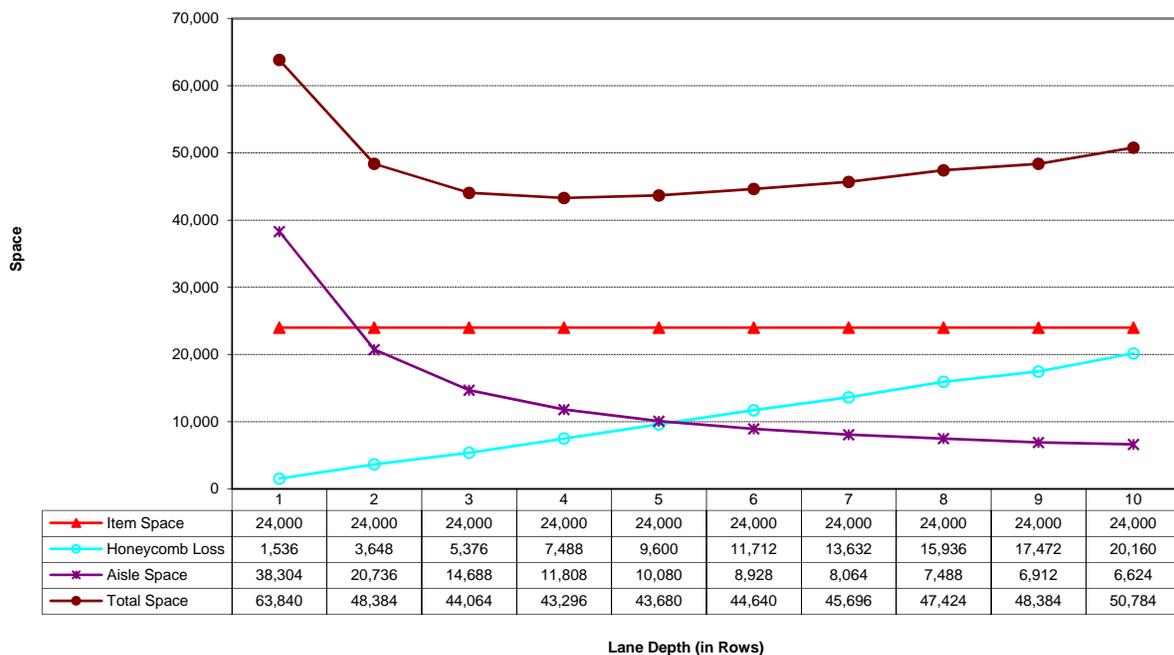
**Figure 6.7. Total space associated with different lane depths
for deep-lane randomized storage.**

In this example, $x = 4$, $y = 4$, $z = 3$, $M = 500$, $N = 20$, $A = 12$, and $H = 4$. Figure 6.7 shows the total space associated with D ranging from 1 to 10. Also shown are the components of total space: item space, honeycomb loss, and down-aisle space. Using (6.11) or finding the minimum

total space in Figure 6.7, the optimal lane depth is $D^* = 4$. D^* is then used in (6.5) to determine the number lanes, $L(4) = 41$, which is then used in (6.3) to determine the total space, $TS(4) = 43,296$. The corresponding maximum cube utilization is as follows:

$$\text{Max cube utilization} = \frac{\text{item space}}{\text{total space } (D = 4)} = \frac{x \cdot y \cdot z \cdot M}{TS(4)} = \frac{4 \cdot 4 \cdot 3 \cdot 500}{43,296} = 0.5543 = 55.43\%.$$

6.2.5 Estimating Handling Costs

Minimizing handling costs usually increases building costs, where the cost of racks, etc., are included as part of the building costs. Warehouse design involves determining the best compromise between these issues. Handling costs can be estimated by determining:

1. Expected time required for each move based on an average of the time required to reach each slot in the region.

2. Number of vehicles needed to handle a target peak demand for moves, e.g., moves per hour.

3. Operating costs per hour of vehicle operation, e.g., labor, fuel.

4. Annual operating costs based on annual demand for moves.

5. Total handling costs as the sum of the annual capital recovery costs for the vehicles and the annual operating costs.

Storage and Retrieval Cycle

A storage and retrieval (S/R) cycle is one complete roundtrip from an I/O port to slot(s) and back to the I/O. The type of cycle depends on load carrying ability of the material handling device. Most fork trucks can carry only one pallet load at a time, while a cart used for piece order picking can carry multiple loads at the same time.

- Carrying one load at-a-time:

  Single command

  Storage: carry one load to slot for storage and return empty back to I/O port, or

  Retrieval: travel empty to slot to retrieve load and return with it back to I/O port.
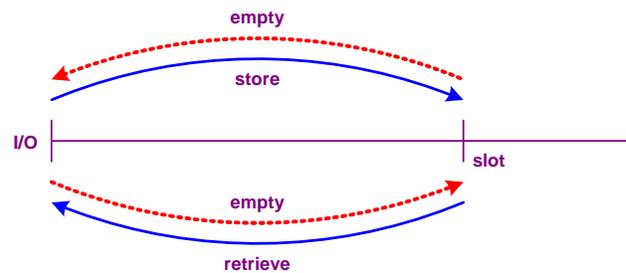

Figure 6.8. Single-command S/R cycle.

  Dual command

  Combine storage with a retrieval: Store load in slot 1, travel empty to slot 2 to retrieve load. Can reduce travel

distance by a third. Also termed task "interleaving."

Figure 6.9. Dual-command S/R cycle.

- Carrying multiple loads:

Multiple command

Multiple loads can be carried at the same time. Used in case and piece order picking.

Figure 6.10. Multiple command S/R cycle.

Expected Time per S/R Cycle

The expected time for each single-command (SC) S/R cycle is

$$\text{Single-command:} \quad t_{SC} = \frac{d_{SC}}{v} + t_L + t_U = \frac{d_{SC}}{v} + 2t_{L/U}, \tag{6.12}$$

where

d_{SC} = expected distance per SC cycle

v = average travel speed (e.g.: 2 mph = 176 fpm walking; 7 mph = 616 fpm riding)

t_L = loading time

t_U = unloading time

$t_{L/U}$ = loading/unloading time, if same value

The expected time for each dual-command (DC) S/R cycle is

$$\text{Dual-command:} \quad t_{DC} = \frac{d_{DC}}{v} + 2t_L + 2t_U = \frac{d_{DC}}{v} + 4t_{L/U} \tag{6.13}$$

Estimating Expected Distance

It is helpful to consider determining the expected distance for a storage region consisting of just a single row of slots (i.e., a 1-D region as shown in Figure 6.11) because the result for a 2-D region, assuming rectilinear distances, is the same as the 1-D result for each dimension. The following results assume (1) all S/R cycles are single-command, (2) rectilinear distances, and (3) each slot is region used with equal frequency (i.e., randomized storage). For dedicated or class-based storage, the expected distance for each SKU or class would be determined separately. Similar formulae can be developed for dual-command S/R cycles.

Figure 6.11. 1-D expected distance.
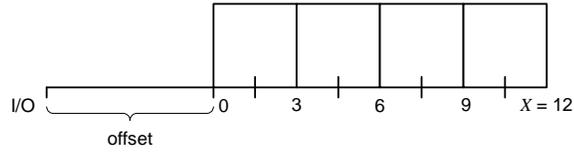
Figure 6.12. Off-set from I/O port.

1-D Expected Distance

In the 1-D region shown in Figure 6.11, the expected distance is the average distance from I/O port to midpoint of each slot; e.g., $[2(1.5) + 2(4.5) + 2(6.5) + 2(10.5)]/4 = 12$, which corresponds to $d_{SC} = X = 12$ as determined in (6.14).

$$TD_{1\text{-way}} = \sum_{i=1}^{L}\left(i\frac{X}{L} - \frac{X}{2L}\right) = \frac{X}{L}\left(\sum_{i=1}^{L} i - \frac{1}{2}\sum_{i=1}^{L}(1)\right)$$

$$= \frac{X}{L}\left(\frac{L(L+1)}{2} - \frac{L}{2}\right) = \frac{XL + X - X}{2} = \frac{XL}{2}$$

$$ED_{1\text{-way}} = \frac{TD_{1\text{-way}}}{L} = \frac{X}{2}$$

$$d_{SC} = 2(ED_{1\text{-way}}) = X \tag{6.14}$$

In the handling cost example shown in Figure 6.16, below, the I/O point was located along the perimeter of the storage region. In many cases, the I/O point is not adjacent to the storage region and each move involves travel between the I/O point and the perimeter of the storage region (see Figure 6.12). The area between the I/O and the storage region may be a different storage region. If the I/O port is off-set from the storage region, then 2 times the distance of the offset is added the expected distance within the slots:

$$d_{SC} = 2(offset) + X \tag{6.15}$$

2-D Expected Distance

Since dimensions X and Y are independent of each other for rectilinear distances, the expected distance for a 2-D rectangular region with the I/O port in a corner is just the sum of the distance in X and in Y:

$$d_{SC}^{rect} = X + Y \tag{6.16}$$

For a triangular region with the I/O port in the corner (see Figure 6.13), let $X = Y$ and $L = D$, so that

$$TD_{1\text{-way}} = \sum_{i=1}^{L} \sum_{j=1}^{L-i+1} \left[\left(i\frac{X}{L} - \frac{X}{2L} \right) + \left(j\frac{X}{L} - \frac{X}{2L} \right) \right] = \quad \cdots$$

$$= \frac{X}{6}\left(2L^2 + 3L + 1\right)$$
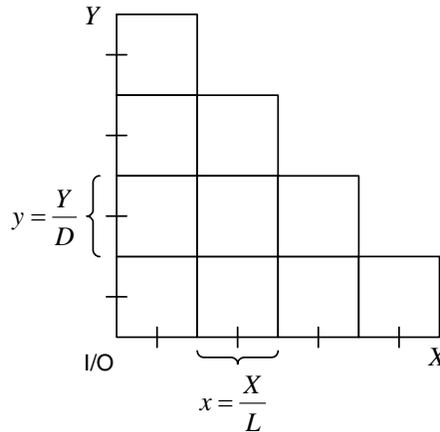
Figure 6.13. 2-D triangular region.

The expected distance result for a triangular region is an approximation that becomes exact as the number of slots in the region increases:

$$ED_{1\text{-way}} = \frac{TD_{1\text{-way}}}{\dfrac{L(L+1)}{2}} = \frac{2}{3}X + \frac{X}{3L} = \frac{2}{3}X, \quad \text{as } L \to \infty$$

$$d_{SC}^{tri} = 2\left(\frac{2}{3}X\right) = 2\left(\frac{1}{3}X + \frac{1}{3}Y\right) = \frac{2}{3}(X + Y). \tag{6.17}$$

The expected distance for two region shapes, rectangular and triangular, and two different I/O point configurations will be considered. In Figure 6.14, the I/O point is assumed to be located off to a side of the region; in Figure 6.15, it is assumed to be offset from the middle of a side of the region. The expected distances for each configuration are as follows:

(a) Rectangular   (b) Triangular
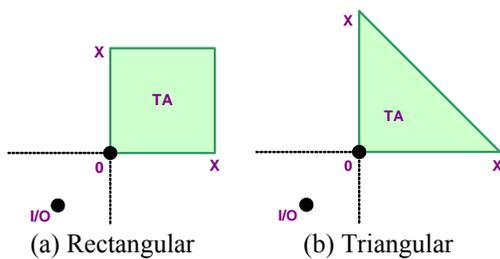
Figure 6.14. I/O-to-side configurations.

(a) Rectangular   (b) Triangular
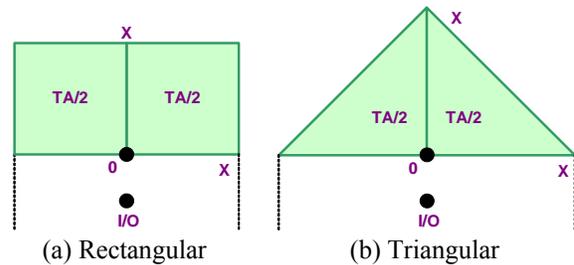
Figure 6.15. I/O-at-Middle configurations.

$$\textit{Rect. I/O-to-Side: } TA = X^2 \Rightarrow X = \sqrt{TA} \Rightarrow d_{SC} = \boxed{2\sqrt{TA}} \tag{6.18}$$

128

$$\textit{Tri. I/O-to-Side: } TA = \frac{1}{2}X^2 \Rightarrow X = \sqrt{2TA} = \sqrt{2}\sqrt{TA} \Rightarrow d_{SC} = \frac{4}{3}\sqrt{2}\sqrt{TA} = \boxed{1.886\sqrt{TA}} \quad (6.19)$$

$$\textit{Rect. I/O-in-Middle: } \frac{TA}{2} = X^2 \Rightarrow X = \sqrt{\frac{TA}{2}} = \frac{\sqrt{TA}}{\sqrt{2}} \Rightarrow d_{SC} = \sqrt{2}\sqrt{TA} = \boxed{1.414\sqrt{TA}} \quad (6.20)$$

$$\textit{Tri. I/O- in-Middle: } \frac{TA}{2} = \frac{1}{2}X^2 \Rightarrow X = \sqrt{TA} \Rightarrow d_{SC} = \frac{4}{3}\sqrt{TA} = \boxed{1.333\sqrt{TA}} \quad (6.21)$$

Given the opportunity to select a shape for a storage region where the I/O point is to be located off to a side of the region (Figure 6.14), a square and an isosceles right triangle are the shapes that minimize the expected distance for rectangular and triangular regions, respectively. In both cases, $X = Y$, and the expected distance can be determined in terms of the total storage area, TA. The distance does not include I/O offset. Thus, given the same TA for both the rectangular and triangular regions, the triangular region provides a $(2 − 1.886)/2 = 5.7\%$ reduction in expected distance as compared to a rectangular region. If the region is not a square or an isosceles right triangle, then the formulae on the previous slide can be used.

Given the opportunity to select a shape for a storage region where the I/O point is assumed to be offset from the middle of a side of a region (Figure 6.15), a rectangle with as aspect ratio of 2 (side-by-side squares) and side-by-side isosceles right triangles are the shapes that minimize the expected distance for rectangular and triangular regions, respectively. In both cases, $X = Y$, and the expected distance can be determined in terms of the total storage area, TA. The distance does not include I/O offset. Given a square rectangular region (as on the previous slide) with the same TA as a rectangular region with as aspect ratio of 2, the latter region provides a $(2 − 1.414)/2 = 29.3\%$ reduction in expected distance as compared to the square region.

Handling Cost Example

The storage region in this example has a total area of $TA = 20,000$ square feet and thousands of slots (see Figure 6.16). The expected distance for a single-command S/R cycle from the I/O point (e.g., loading dock) to all of the slots is 200 feet. The formula used to determine the expected distance depends on the shape of the storage region and the location for the I/O point relative to the region. The expected time per S/R cycle is determined by converting the expected distance to time, assuming a travel speed of 200 fpm, and adding the time required for loading (30 s) and unloading (30 m) for the single-command cycle.

In this example, the peak demand is 75 moves per hour. If the warehouse operates for 2,000 hours per year, then the annual demand of 100,000 moves corresponds to 50 moves per hour, thus the peak demand is 50% greater than the average demand. The investment related cost of $7,500 per year for all of the trucks is proportional to the peak demand, while the operating cost of $33,333 per year for labor is proportional to the average demand. The only operating cost considered is labor cost ($10 per hour), which is usually the largest such cost (fuel is typically one-tenth the cost of labor). Labor cost is determined based only on the total hours spend performing the required moves. If the three truck operators needed for the peak periods of

demand were solely dedicated to performing these moves and each was available for 2,000 hours per year, then the labor cost would be $3 \times 10 \times 2{,}000 = \$60{,}000$ per year. The lower cost (\$33,333) assumes that it is possible for the operators to perform other tasks during the off-peak periods.

TA = 20,000

I/O
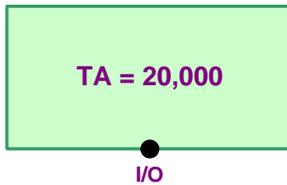
Figure 6.16. Handling cost example.

Expected Distance: $d_{SC} = \sqrt{2}\sqrt{TA} = \sqrt{2}\sqrt{20{,}000} = 200$ ft

Expected Time: $t_{SC} = \dfrac{d_{SC}}{v} + 2t_{L/U}$

$= \dfrac{200 \text{ ft}}{200 \text{ fpm}} + 2(0.5 \text{ min}) = 2$ min per move

Peak Demand: $r_{\text{peak}} = 75$ moves per hour

Annual Demand: $r_{\text{year}} = 100{,}000$ moves per year

Number of Trucks: $m = \left\lfloor r_{\text{peak}} \dfrac{t_{SC}}{60} + 1 \right\rfloor = \lfloor 3.5 \rfloor = 3$ trucks

Handling Cost: $TC_{\text{hand}} = mC_{\text{truck}} + r_{\text{year}} \dfrac{t_{SC}}{60} C_{\text{labor}}$

$= 3(\$2{,}500 / \text{truck}) + 100{,}000 \dfrac{2}{60} (\$10 / \text{hr})$

$= \$7{,}500 + \$33{,}333 = \$40{,}833$ per year

Estimating AS/RS

An automated storage/retrieval system (AS/RS) consists of an integrated computer-controlled system that implements the storage/warehousing elements (e.g., storage medium, transport mechanism, and controls) with various levels of automation for fast and accurate random storage of products and materials. One of the unique aspects of an AS/RS with respect to its design is the mode of operation of the S/R machines. In the design of most storage systems, rectilinear distances can be used to represent the movement of the transport mechanisms; in an AS/RS, the S/R machines can move a load in the horizontal direction along an aisle and lift the load in the vertical direction simultaneously (and, typically, at different speeds), so that the use of rectilinear distances would overestimate the distance (or time) the load travels.

Letting v_x and v_z be the horizontal (X) and vertical (Z) speeds, respectively, of an S/R machine, then the time required for the machine to move from (x_0, z_0) to (x, z), assuming instantaneous acceleration, can be represented by the Chebychev "distance":

$$\max\left\{ \frac{|x - x_0|}{v_x}, \frac{|y - y_0|}{v_y} \right\} \tag{6.22}$$

For each aisle of an AS/RS, the I/O port for the aisle is typically at the end the aisle and at the bottom level of the racks in the aisle; thus, assuming $(x_0, z_0) = (0, 0)$ as the location of the I/O port (and ignoring the horizontal movement (Y) of the S/R machine's shuttle into the racks), the time required to travel from the I/O port to location (x, z) is

$$\max\left\{\frac{x}{v_x}, \frac{y}{v_y}\right\}. \tag{6.23}$$

6.2.6 Dedicated Storage Assignment Problem (DSAP)

In this section, items are assigned to the slots so that the total cost of material flow is minimized.

Given a layout with N items, the following Dedicated Storage Assignment Problem (DSAP) can be used to determine slot assignments:

<div align="center">

DSAP SOLUTION PROCEDURE

</div>

1. *Order Slots:* Compute the expected cost for each slot and then put into nondecreasing order.

2. *Order Items:* Put the flow density (flow per unit of volume) for each item i into nonincreasing order

$$\frac{f_{[1]}}{M_{[1]}s_{[1]}} \geq \frac{f_{[2]}}{M_{[2]}s_{[2]}} \geq \cdots \geq \frac{f_{[N]}}{M_{[N]}s_{[N]}}$$

3. *Assign Items to Slots:* For $i = 1, \ldots, N$, assign item $[i]$ to the first slots with a total volume of at least $M_{[i]}s_{[i]}$.

where,

$f_i =$ flow (i.e., moves per period) of item i

$s_i =$ storage space per unit of item i

$M_i =$ maximum number of units of item i

The *volume* of storage space needed for an item i is $M_i s_i$. If the handling costs are identical for all moves between slots and I/O ports, then slot cost can be viewed as the expected distance traveled between the slot and all of the I/O ports. The cube-per-order index (COI), which is the reciprocal of the flow per unit volume, is sometimes used instead, and items are then stored in nondecreasing order.

Assumptions

The following assumptions must be satisfied in order to be able to use the DSAP procedure:

1. All storage/retrieval (S/R) operations are performed as single-command cycles.

2. For item i, the probability of a move to/from each slot assigned to the item is the same.

3. The *factoring assumption*:

    (a) Handling costs and distances (or times) are identical for all items

    (b) The percent of S/R moves of an item stored at slot j to/from I/O port k is identical for all items.

If the factoring assumption is not satisfied and the storage space per unit of each item is the same, then the DSAP can be solved as a Transportation Problem. Transportation problems can be solved relatively easily using commercial software packages.

Due to Assumptions 1 and 2, the slots do not interact with each other; if some of the S/R operations were dual command or part of case or piece order picking (see the slotting discussion in Section 6.5.1), then the cost of assigning a slot to an item would depend on what items were assigned to the other slots. If the probabilities of using slots for an item were not all equal (e.g., if the slots that are nearer an I/O port had a higher probability of being used), then the cost for a slot would depend on what other slots are assigned to the item.

Assumption 2 would be valid if, for example, both a FIFO retrieval policy is used for all items, and the slot assigned to item i that has remained empty the longest is always the next slot used for storage. In practice, these conditions would be approximately satisfied if all storages (retrievals) took place in a short time period (e.g., receiving (shipping) of truck loads of material) and the slots were emptied (filled) before the next storages (retrievals) took place.

Assumption 3 is termed the *factoring assumption* because it allows the total cost to be factored into the product of two terms, one based only on the slot cost and one based only on the cube per order. In practice, Assumption 3(a) would be satisfied if, for example, the same MHE is used for all items and the handling characteristics (including loading/unloading times) are the same for all items. Assumption 3(b) would be valid if, for example, there is only one I/O port, or there are two ports and one is used only for input and the other port is used only for output and the ratio of flow into a slot to flow out of a slot is identical for all items; the assumption would need to be verified in other situations.

1-D DSAP Example

The DSAP procedure is used to assign items A, B, and C to dedicated 1-D storage regions:

Step 1. The cost of each slot is its distance from the I/O point; thus, from left to right, slots are in nondecreasing order.

Step 2. Items are ordered C-A-B, which corresponds to ranking their flow density values in nonincreasing order 7.00, 6.00, and 1.40.

Step 3. Item C is assigned to the 3 leftmost slots; item A to the next 4 leftmost slots; and item B to the next 5 leftmost slots.

The data for the example is given in Table 6.4, and is a continuation of the example given in Table 6.2. The assignment C-A-B minimizes the total distance (436) required to complete the 24, 7, and 21 single-command S/R cycles for items A, B, and C, respectively. The total distance is calculated by summing together the product of the expected distance and flow for each item. The expected distance is calculated as $d_{sc} = 2(offset) + X$. The storage region for item A is offset 3 units from the I/O point because of the slots occupied by item C, and item B is offset 7 units because of the slots occupied by items C and A.

Table 6.4. Data for 1-D DSAP Example

			Dedicated			Random	Class-Based		
		A	B	C		ABC	AB	AC	BC
Max units	M	4	5	3		9	7	7	8
Space/unit	s	1	1	1		1	1	1	1
Flow	f	24	7	21		52	31	45	28
Flow Density	$f/(M \times s)$	6.00	1.40	7.00		5.78	4.43	6.43	3.50

Table 6.5. 1-D DSAP Example

The optimal assignment C-A-B, which corresponds to ranking the flow density values in nonincreasing order, results in the minimum total distance as compared to all other possible dedicated slot assignment. One possible alternative is to rank the just the flow values in nonincreasing order (24, 21, and 7), which results in the assignment A-C-B with a total distance greater than the optimal assignment (460 vs. 436). The optimal class-based assignment (C-AB) and randomized storage (ABC) both have greater total distances than dedicated storage but

require a less space, illustrating the trade-off between building costs and handling costs in warehouse design.

Table 6.6. Comparison of Different Storage Policies

1-D Slot Assignments	Total Distance	Total Space
Dedicated (flow density) — I/O: C C C A A A A B B B B B	436	12
Dedicated (flow only) — I/O: A A A A C C C B B B B B	460	12
Class-based — I/O: C C C AB AB AB AB AB AB AB	466	10
Randomized — I/O: ABC ABC ABC ABC ABC ABC ABC ABC ABC	468	9

2-D DSAP Example

This example is the same as the 1-D Slotting example except that the slots are ordered based on their 2-D rectilinear distance from the I/O port as shown in Figure 6.17 (a). Distances are determined from the center of the I/O square to the center of each slot, and each slot is assumed to include a portion of aisle space. The contours of equal distance slots have a triangular shape because distances are rectilinear, as opposed to the circular-shaped contours that would be formed is distances were Euclidean (i.e., straight-line). In the optimal assignment shown in Figure 6.17 (c), item C is assigned to the slots of the contour closest to the I/O.
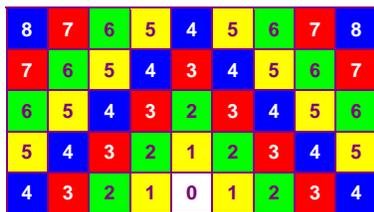
(a) Distance from I/O to Slot   (b) Original Assignment (TD = 215)   (c) Optimal Assignment (TD = 177)
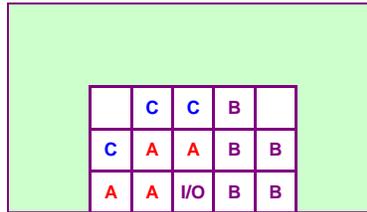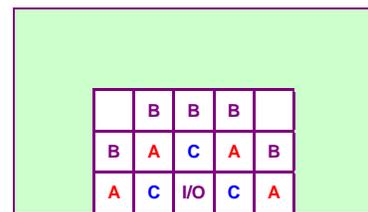
Figure 6.17. 2-D DSAP example.

6.3 Storage Equipment

Storage equipment is used for holding or buffering materials over a period of time. Some storage equipment may include the transport of materials (e.g., the S/R machines of an AS/RS, or storage carousels). If materials are block stacked directly on the floor, then no storage equipment is required. Storage racks are used to provide support to a load and/or to make the load accessible.

Table 6.7. Storage Equipment

1. Block stacking (no equipment)	9. Stacking frame
2. Selective pallet rack	10. Bin shelving
(a) Single-deep rack	11. Storage drawers
(b) Double-deep rack	12. Storage carousel
3. Drive-in rack	13. Vertical lift module
4. Drive-through rack	14. A-frame
5. Push-back rack	15. Automatic storage/retrieval system (AS/RS)
6. Flow-through rack	(a) Unit load AS/RS
(a) Carton-flow rack	(b) Miniload AS/RS
(b) Pallet-flow rack	(c) Man-on-board AS/RS
7. Sliding rack	(d) Deep-lane AS/RS
8. Cantilever rack	

1. Block stacking (no equipment)

Block stacking is the storage of loads on top of each other in stacks placed in lanes on the floor (a.k.a. *floor storage*)

Adv: Easy to implement and very flexible. Low investment cost since no storage medium is required.

Disadv: Honeycomb loss can be significant since, on average, half of one lane will be empty for each item. In most cases, only LIFO retrieval is possible in each lane unless there is an aisle at the back. Damage to loads might cause instable stacks.

Two to ten rows of storage typically used

Width of each lane limited by vehicle width, not load width

Storage racks are used when support and/or material accessibility is required

2. Selective pallet rack

Pallets are supported between load-supporting beams. Most popular type of storage rack. Special attachments and decking can be used to make the racks capable of supporting other types of unit loads besides pallets (e.g., coils, drums, skids)

Load-on-beam racks are used to provide clearance for straddles; *load-on-floor* racks can be used when it is not necessary to use straddles.

2(a) Single-deep rack

Single position (slot) per position

Adv: Provides complete and fast accessibility to all loads with no honeycomb loss

Disadv: Can result in low cube utilization because of aisle space requirements, which can be influenced by the lift truck used (e.g., a turret truck would increase and a standard counterbalanced would decrease utilization)
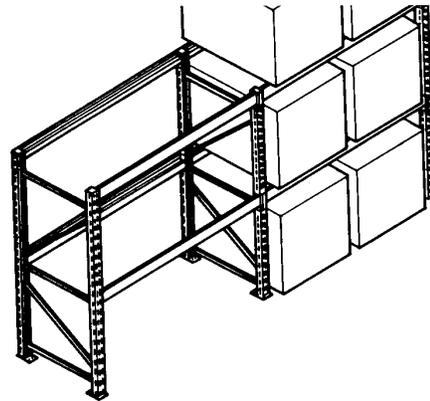
2(b) Double-deep rack

Two pallets stored per position

Adv: Provides greater cube utilization than single-deep racks because more loads can be accessed from the same side of the rack

Disadv: In order to access rear load in rack, an extended reaching mechanism is required on the lift truck
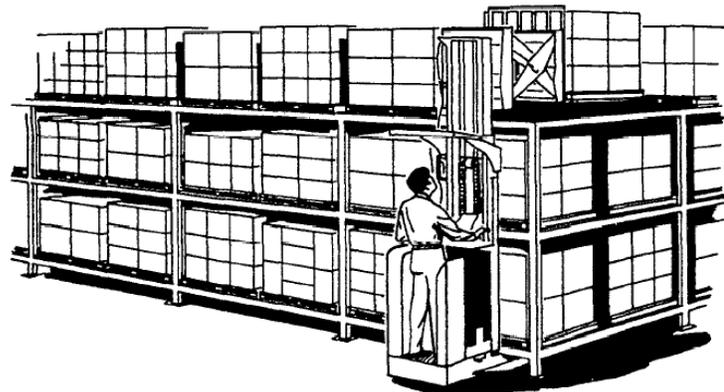
[NAVSUP Pub 529]

Typically used when the inventory level for an item is at least five or when loads are stored and picked in multiples of two pallets[1]

3. Drive-in rack

Loads are supported by rails attached to the upright beams. Lift trucks are driven between the uprights beams.

Adv: Provides high density pallet storage.

Disadv: Requires uniform-size loads. Lengthy storage and retrieval times due to care required by driver inside of the rack.

Closed at one end, allowing entry
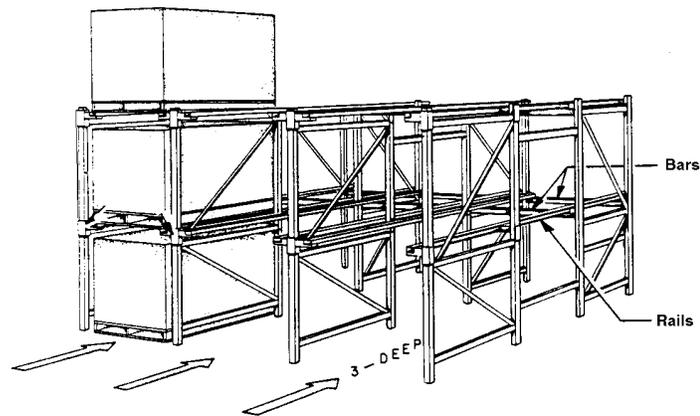
from one end (LIFO)

---

4. Drive-through rack

Similar to drive-in rack, except open at
  both ends, allowing access from
  both ends (FIFO)

Used for staging loads in a flow-thru
  fashion [Frazelle, WC WH and MH]

---
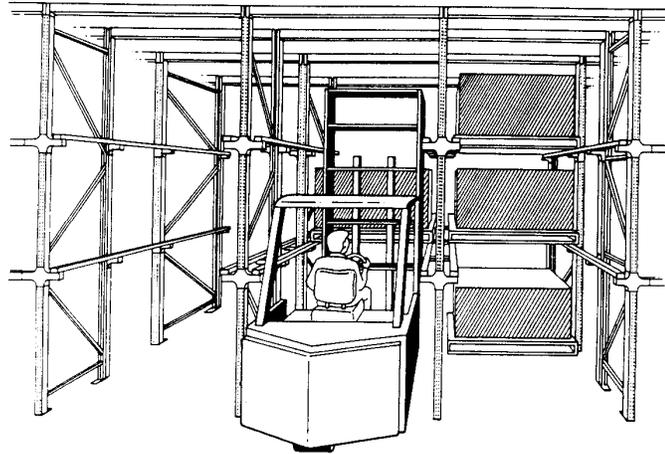
5. Push-back rack

Loads are supported on an incline to enable
  gravity-based movement of the loads within
  the rack via roller conveyor. Used to
  provide highly accessible pallet storage.

Provides LIFO storage in each lane: Loaded
  and unloaded at the lower end and closed at
  the higher end

Adv: Can be used to enable deep-reach
  storage without the need for extended reach
  mechanisms for loading/unloading

Disadv: Rack investment costs are greater
  than for double-deep racks

Maximum depth is 5 loads

Rails

Carts
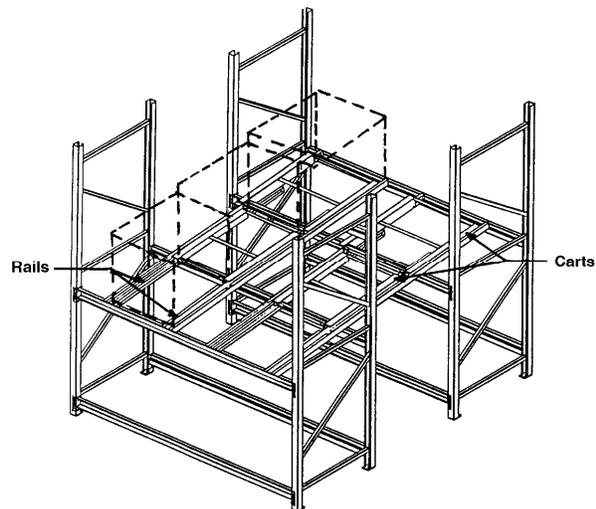
6. Flow-through rack

Similar to push-back rack in terms of
   storage density, except greater
   storage depth is possible. Rack is
   loaded at higher end and unloaded at
   lower end, providing FIFO storage
   in each lane.

Termed *pallet-flow rack* and *carton-
   flow rack* (pictured) when pallets
   and cartons used, respectively

6(a) Carton-flow rack

Adv: Allows a large cubic volume of
   product to be accessible from a
   small pick face area, supporting
   relatively high pick rates. Replenishment does not interfere with picking.

Disadv: More expensive than bin shelving.

Can have LED displays attached to shelf beam for "pick-to-light" operations (see Section 6.5.5)

6(b) Pallet-flow rack

Adv: Replenishment does not interfere with picking.

Disadv: Requires twice as much aisle space as push-back racks, but overall storage density could
   be higher because of greater storage depth.

Storage depth of 40 to 50 possible for pallet-flow racks

---
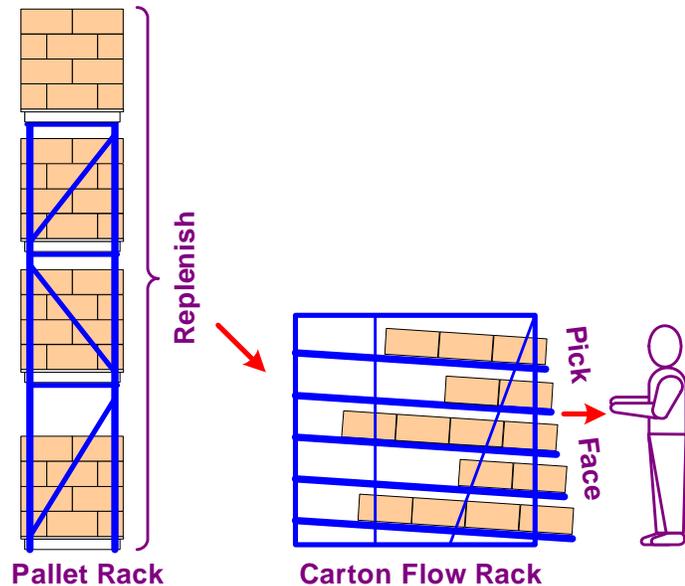
7. Sliding rack

Location of the aisle is changed by
   sliding rows of racks along guide
   rails in floor (a.k.a. *mobile rack*)

Used when only single-deep storage is
   possible and space is very limited or
   expensive

Adv: High cube utilization and
   complete accessibility to all loads

Disadv: More expensive compared to
   other storage racks. Lengthy storage
   and retrieval times because one can
   only pick in one lane at a time.

138
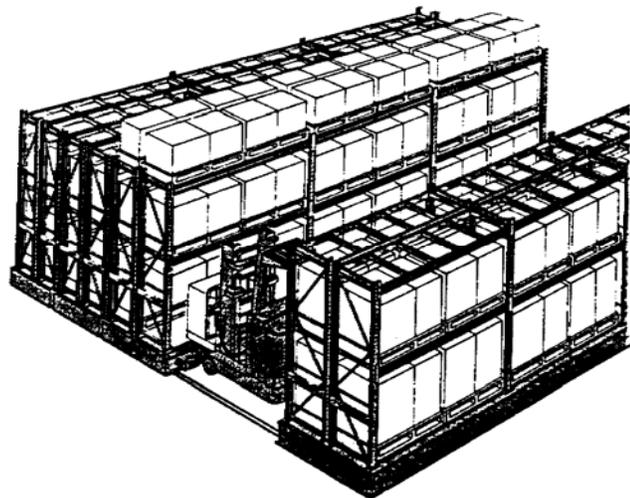
Relies on having a reliable power source available.

Provides increased security for items compared to other racks

Typically found in library stacks, vaults, and climate-controlled (e.g., refrigerated) storage rooms

---

8. Cantilever rack

Loads are supported by two or more cantilevered "arms" (i.e., horizontal beams supported at only one end)

Similar to pallet racks, except the front upright and front shelf beams are eliminated

Used when there is a need for a full clear shelf that can be loaded from the front without obstruction from rack support uprights

Typically used to store long loads (e.g., bar stock, pipes, lumber)

---
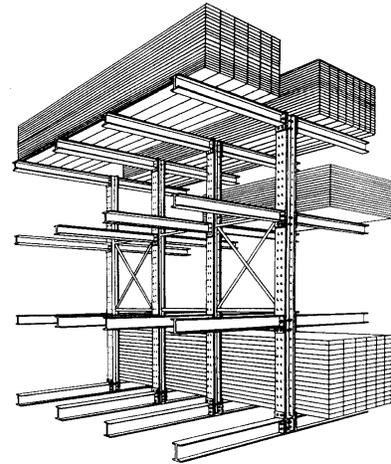
9. Stacking frame

Interlocking units that enable stacking of a load so that crushing does not occur

Can be disassembled and stored compactly when not in use

Pallet frames can be used to enable multilevel block stacking

Cost per frame: $100–$300 (can be leased for short-term increases in inventory)

---
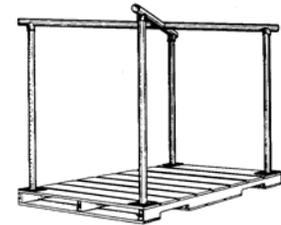
10. Bin shelving

Alternative to racks to store small, loose, nonpalletized items. Pieces placed either directly on shelves or in bins or cartons.

Adv: Low cost.

Disadv: Can result in excessive travel for picker. Difficult to pick from top shelf depending on the height of the picker and the weight of the unit. Replenishment can interfere with picking.

Several levels of shelves (and storage drawers) on a mezzanine can be used to allow multi-level picking (max 4 levels). A lift truck or vertical reciprocating conveyor, e.g., can be used to help with removing units from the top level.

11. Storage drawers

Drawers provide an alternative to bin shelving to store small, loose items

Adv: Drawers can provide increased security compared to bin shelving and is most important when the demand for a specific item is low and infrequent. Easy to install at point of use.

Disadv: Space is frequently underutilized unless there is an ongoing disciplined approach to managing the location and usage of the units being stored. Replenishment can interfere with picking. Cannot see inside the drawers, making labeling an important issue.

12. Storage carousel

Carousel consists of a set of horizontally (pictured) or vertically revolving storage baskets or bins.

Adv: Allows a large number of items to be picked at a high rate.

Disadv: Replenishment cannot occur during picking operations (typically, replenishment takes place during a separate shift or is interleaved between peak picking periods).

One operator picks from 2 to 4 carousels (termed a "pod") in order to minimize waiting time while other carousels are moving.

Each level of the carousel can rotate independently in a clockwise or counter-clockwise direction

Control ranges from manually activated push buttons to automated computer controlled systems

Pick Cart
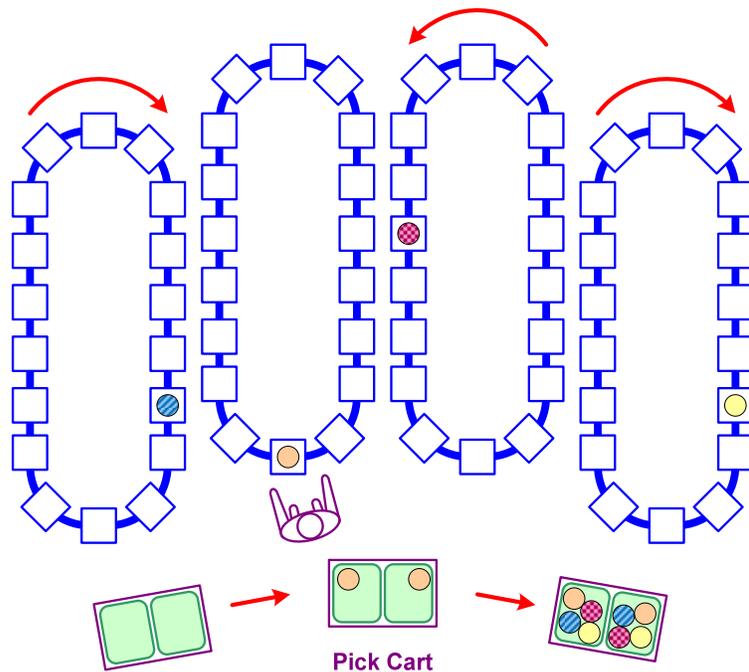
13. Vertical lift module

Pieces stored on trays inside a multi-bay enclosure that are delivered to the
   opening of a bay for picking by a servo-driven lift carriage.

Adv: Provides high-bay storage, and dense storage since height of trays can
   vary. All picking occurs at a user-adjustable waist height. Can provide
   even greater security compared to bin shelving when the operation of the
   module is often under computerized control, which can increase cost.

Disadv: High cost. Requires reliable power source.
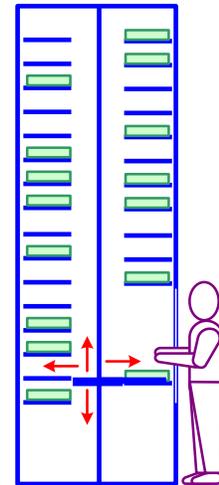
14. A-frame

Units are dispensed from parallel
   arrays of vertical angled channels
   onto a belt conveyor that carries
   them into a container.

Adv: Very high pick rate.

Disadv: Only feasible for small, rigid
   items of uniform shape that are not
   fragile. Requires manual
   replenishment..
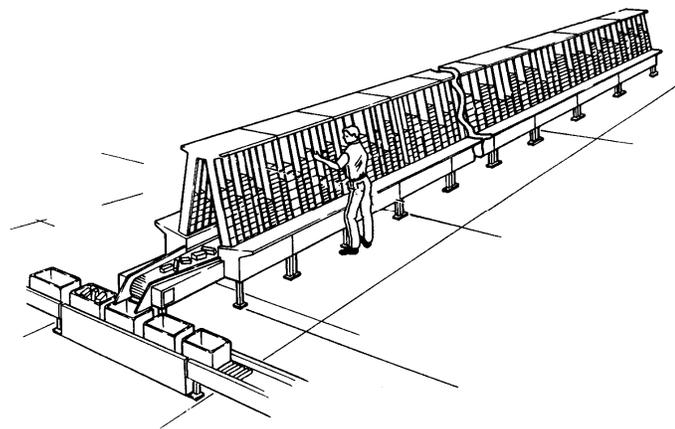
Enables fully automated
   piece picking, with
   manual replenishment.
   Popular within
   pharmaceutical
   distribution centers

Inverted A-frame used for
   flat items

Other types of fully
   automated piece picking
   systems include robotic
   based systems that are
   similar in construction
   to robotic pick and
   place palletizers

**Parallel Arrays of
Vertical Channels**

Replenish

**Carton Flow
Racks**

Takeaway Conveyor
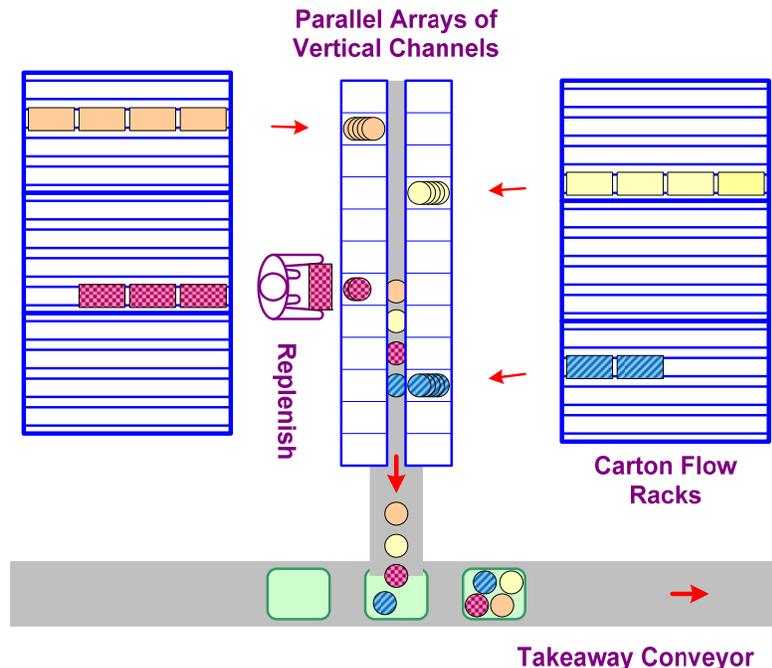
15. Automatic storage/retrieval systems (AS/RS)

Consists of integrated computer-controlled system that combines storage medium, transport
mechanism, and controls with various levels of automation for fast and accurate random
storage of products and materials

Storage/retrieval (S/R) machine in an AS/RS operates in narrow aisle, serving rack slots on both
sides of aisle; can travel in horizontal (along the aisle) and vertical (up and down a rack)
directions at same time

Adv: Fewer material handlers, better material control (including security), and more efficient use
of storage space

Disadv: Typically, high capital and maintenance costs, and more difficult to modify

Although AS/RS were originally developed for warehousing and distribution operations, they are
now also being used for in-process storage as part of an automated job shop. In an automated
job-shop, an AS/RS can be combined with an automatic identification system and an
automatic transportation system (e.g., automatic conveyors and/or an AGV system) to provide
real-time material control capabilities. The material stored in the AS/RS can include both
finished goods and work in process and even production tools and jigs.

Components:

Racks: A typical AS/RS utilizes high-rise storage racks, ranging in height between 40 and 80
feet or higher, for random storage. High-rise racks require tight rack tolerances and level
floors, all of which increase the cost of the racks as compared to a basic storage rack. The
racks in an AS/RS can be freestanding or uses to support the building (RSS—rack-supported
structure).

S/R Machine: An S/R machine in an AS/RS operates in a narrow aisle, serving rack slots on both
sides of the aisle. The machine can travel in the horizontal (along the aisle) and vertical
(up and down a rack) directions at the same time. Often the machine is captive to one aisle,
although, if throughput requirements do not justify dedicating a machine to each aisle, a
transfer car can be provided to move the machine from the end of one aisle to another, thus
enabling the machine to operate in more than one aisle. The machine is a structural single- or
multiple-mast frame that rides on one or two floor-mounted wheel rails. A carriage carrying a
load-supporting mechanism (or shuttle) operates within the frame. The shuttle is used to
store/retrieve loads at the racks and, at the end of the aisle, to transfer loads onto or away from
conveyors, vehicles, or pick-up and delivery (P/D) stations or transfer stations. Deep-reach
"mole" S/R machines can detach and run into a lane.

Control: The operation of an AS/RS can be controlled by an operator working from a console,
but in many cases, the control system is under complete computer control. Typically,
distributed control, where each S/R machine is controlled by a dedicated computer with
interfaces with a central computer, is used to increase system reliability.

15(a) Unit load AS/RS

Used to store/retrieve loads that are palletized or
  unitized and weigh over 500 lbs.

Stacking heights up to 130 ft. high, with most ranging
  from 60 to 85 ft. high; 5 to 6 ft. wide aisles; single-
 or double-deep storage racks
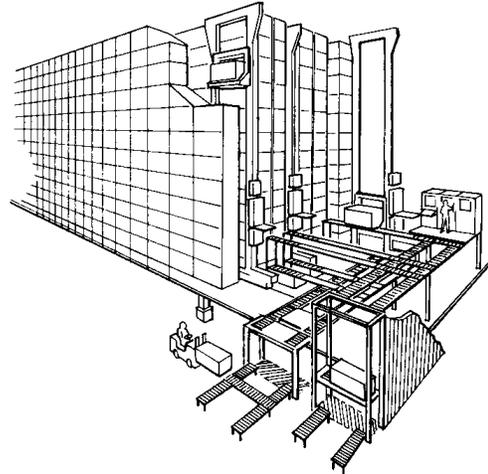
15(b) Miniload AS/RS

Used to store/retrieve small parts and tools that
  can be stored in a storage bin or drawer

End-of-aisle picking and replenishment

Stacking heights range from 12 to 20 ft.; bin
  capacities range from 100 to 750 lbs.

Not typically used for order picking because of
  long cycle times and high cost

Termed a "microload AS/RS" when capacity is
  less than 100 lbs (used in assembly, kitting, and
  testing operations to deliver small containers of
  parts to individual workstations)

Workstations are typically located on the sides of a pair of racks and the S/R machine operates
  between the racks to move containers to openings in the racks (storage lanes) located next to
  each station

143
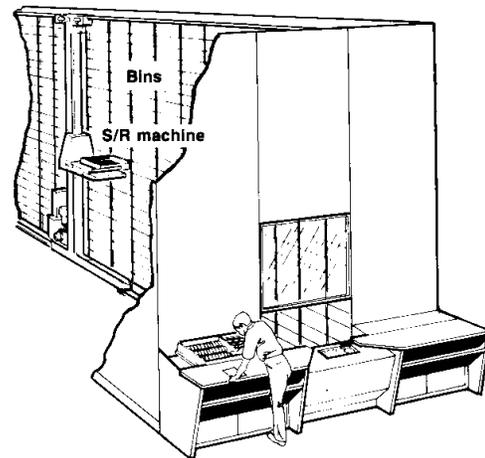
15(c) Man-on-board AS/RS

Used for in-aisle picking; operator picks from shelves, bins, or drawers within
the storage structure

Manual or automatic control

S/R machine is similar to an order picker or turret truck and can sometimes
operate as an industrial truck when outside an aisle, except the S/R is guided
along a rail when operating in an aisle
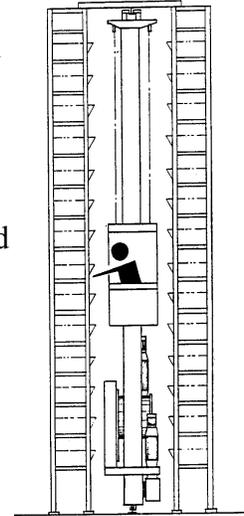
15(d) Deep-lane AS/RS

Similar to unit load AS/RS, except loads can be stored to greater
depths in the storage rack

A rack-entry vehicle is used to carry loads into the racks from
the S/R machine, and is controlled by the S/R machine

Termed an "automated item retrieval system" when
used to automatically retrieve individual items or
cases, with replenishment (storage) taking place
manually from the rear of a flow-through storage
lane and items are pushed forward with a rear-
mounted pusher bar for automatic picking from the
front of the storage lane
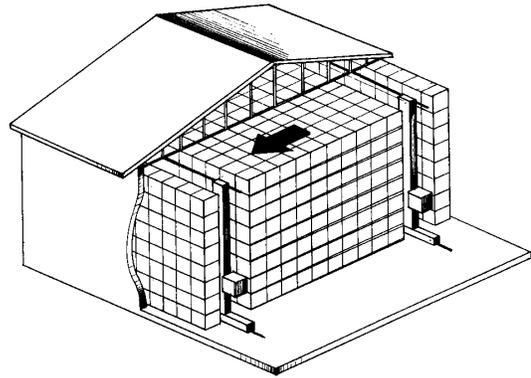
Table 6.8 provides a comparison of the different type of storage equipment that can be used for
pallet picking. Note that there are approximately two lanes of storage per item:

$$\text{Lanes per item} = \frac{\text{Inventory level per item}}{\text{Storage depth per lane}} \approx 2$$

Table 6.8. Pallet Storage Comparison

Type of Rack	Block Stacking	Single Deep	Double Deep	Drive In	Drive Through	Push Back	Pallet Flow	Sliding Racks
Investment cost per position*	–	1	1.1	1.4	1.4	3	5	7
Storage depth per lane	2–10	1	2	5–10	5–10	2–5	2–5	1
Inventory level per SKU	≥ 5	< 3	≥ 5	≥ 20	≥ 20	3–10	3–10	< 3
All loads accessible	No	Yes	No	No	No	No	No	Yes
S/R throughput	Good	Good	Average	Average	Average	Good	Excellent	Poor
FIFO retrieval possible	Yes	Yes	No	No	Yes	No	Yes	Yes
Vertical honeycomb loss	Yes	No	No	Yes	Yes	No	No	No

* Ratio relative to Single Deep cost (see Gross & Associates' "Rules of Thumb" for current cost estimates).

6.4 Warehouse Operations

6.4.1 Typical Warehousing Functions

In most warehouses (see Figure 6.18 and Figure 6.19), products are received and, if they cannot be cross-docked and immediately shipped out, are putaway into storage until they are needed to fill a customer's order, at which time they are picked from storage, packed, sorted, and unitized, if necessary, and then shipped to customers. A separate forward picking storage area can be used to enable more efficient order picking. It is replenished from a reserve storage area. Periodically, partially filled storage locations containing the same type of item are consolidated into a single location to improve space utilization, items are moved to different storage locations to improve handling efficiency in a process termed rewarehousing, and the contents of storage locations are counted in order to verify the accuracy of inventory records in a process termed *cycle counting*. Storage for pallet and case picking occupies the majority of space in a typical warehouse.
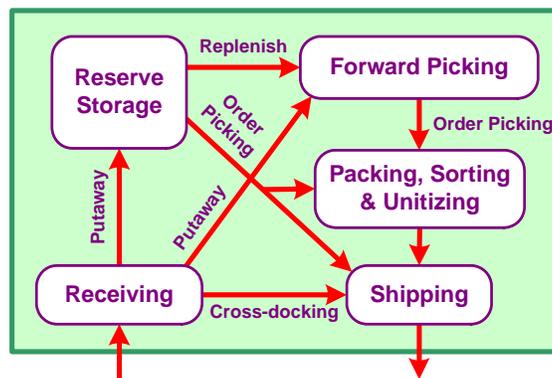


Figure 6.18. Typical warehousing functions.
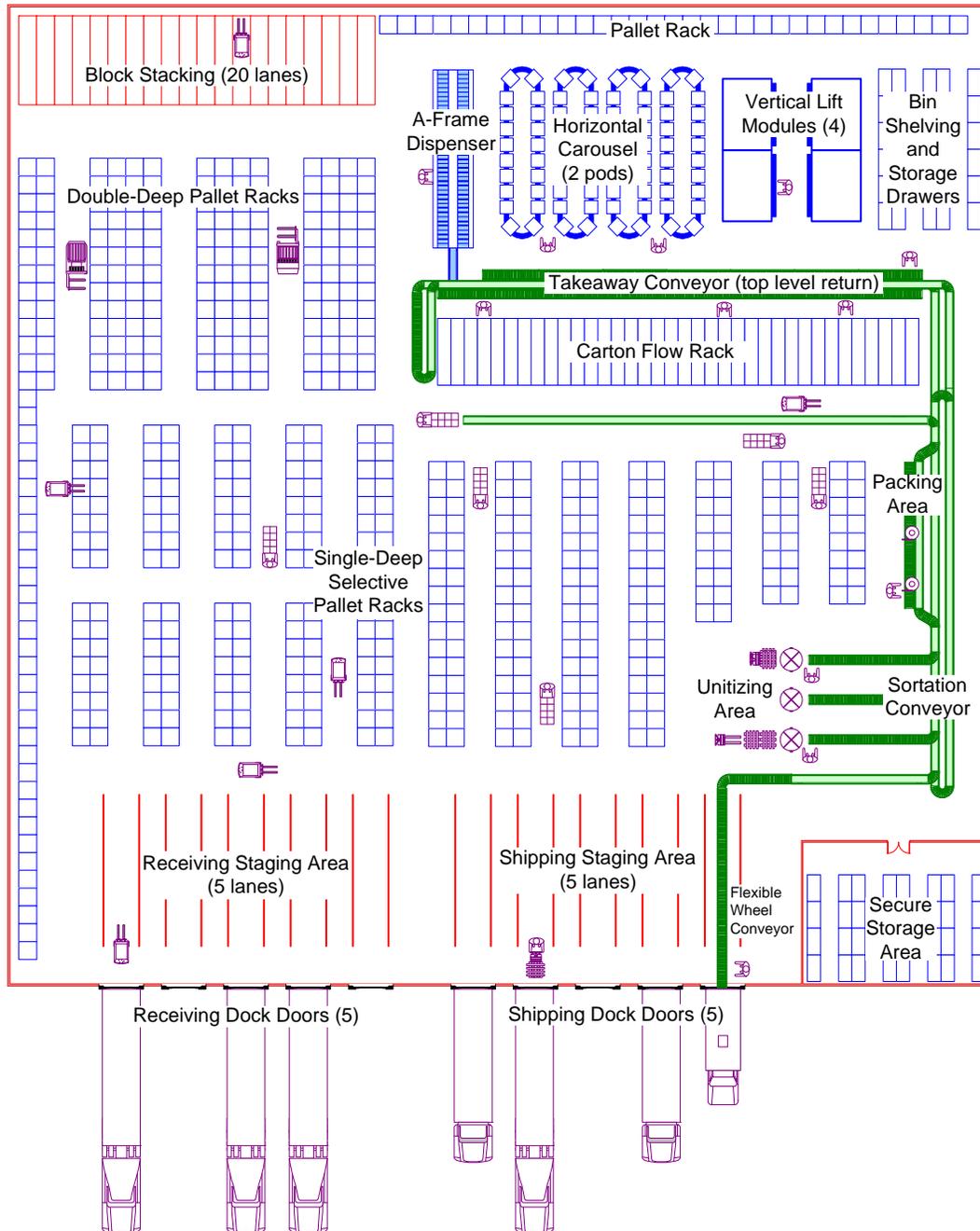
Figure 6.19. Example distribution center.

6.4.2 Warehouse Control

Warehouse control involves the interplay between inventory control and location management. The warehouse management system (WMS) is the software system that enables real-time, paperless control of warehouse operations. As shown in Figure 6.20, the WMS of a single warehouse interfaces with the corporation's enterprise resource planning (ERP) software where item, carrier, and customer master files common to all of the firm's warehouses reside. This

information is used to create and maintain an inventory master file and a location master file. The WMS uses these files along with control logic to execute the required warehouse operations, which include interfacing with the various automated material handling equipment subsystems and generating pick lists for order picking. Advance shipping notices (ASNs) are sent to the WMS from suppliers as part of the receiving function, and the WMS sends ASNs to customers as part of the shipping function of a warehouse. A separate transportation management system (TMS) is typically used to determine shipping details.
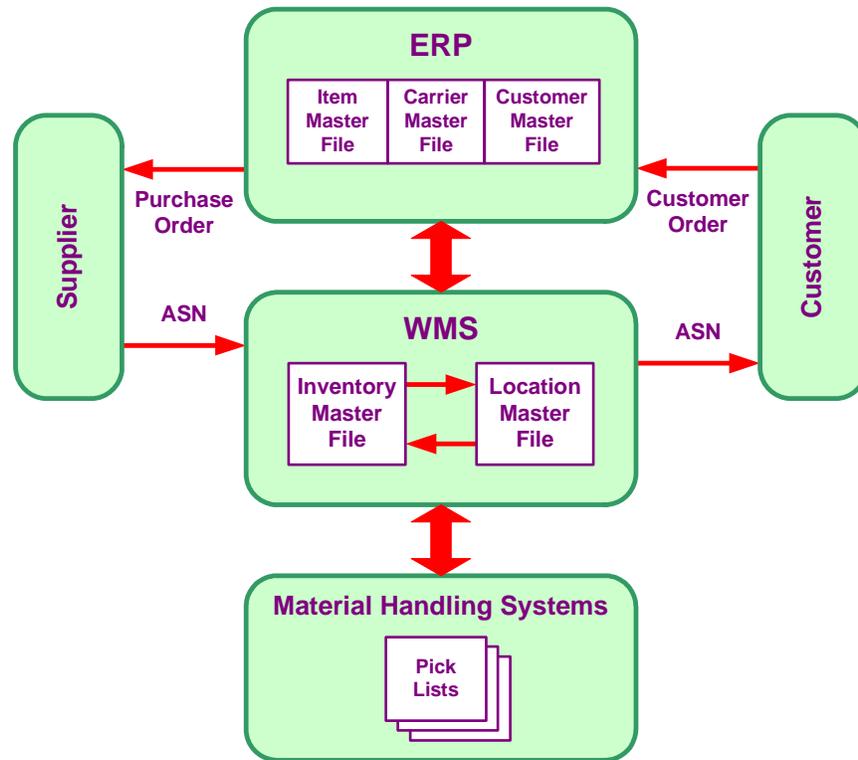

Figure 6.20. Warehouse management system.

At its lowest level, warehouse operations involve the storage of an object at a location or the movement of an object between locations. Inventory is the quantity of each item stored in the warehouse, and the inventory master file acts as the repository for all inventory in the warehouse. It contains the total quantity and storage locations of each item stored in the warehouse and is used together with the location master file to control material transport operations. The location master file provides the link between the WMS's logical representation of the warehouse and the physical layout of the warehouse. The item master file is used to identify valid items that are handled in the warehouse, and includes information about the item that is need for picking purposes (see Figure 6.26 for an example). The carrier master file includes transportation-related information (e.g., rate schedules) that is used for shipping completed orders, and the customer master file is used to store customer preferences for how orders are to be shipped so that it does not need to be included in each order.
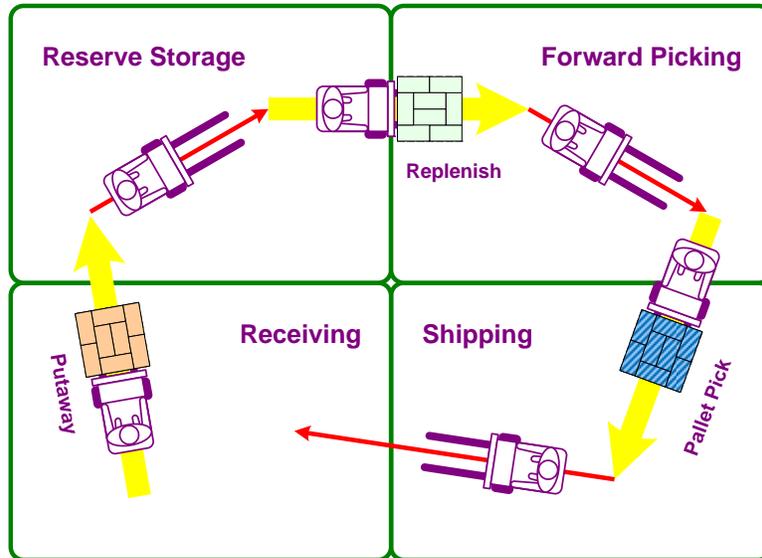
Figure 6.21. Task interleaving.

Inventory Master File

Item	On-Hand Balance	In-Transit Qty.	Locations
A	2	1	11,21
B	4	0	12,22

Location Master File

| Location	Item	On-Hand Balance	In-Transit Qty.
11	A	1	0
12	B	3	0
21	A	1	1
22	B	1	0
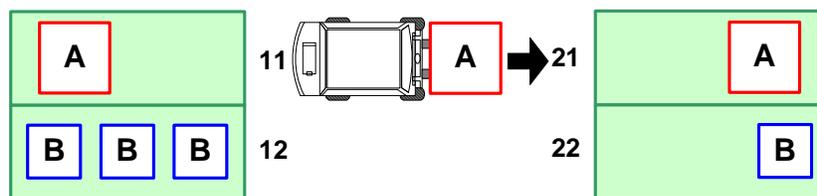
Figure 6.22. Validation.

In order to improve the efficiency of transport operations in a warehouse, a WMS and RF communications can be used to dispatch material handlers from one task to another in real time based on their proximity, resulting in the interleaving of putaway, replenishment, and picking operations (see Figure 6.21).

Validation

Validation is the verification that an inventory movement was performed correctly. Independent data is collected concerning the identity of the movable unit and the beginning and ending

locations of the move. This is then matched to the record of the move maintained in the WMS. Any discrepancies are singled out for immediate correction in order to maintain the accuracy of the inventory records. In Figure 6.22, although three units of item A are in the warehouse, one unit is in transit from location 11 to location 21 and will not be available for picking until the WMS is notified that it has been delivered at location 21, at which time the on-hand balance at location 21 will increase to 2 and the in-transit quantity will decrease to 0.

Logistics-related Codes

Table 6.9 lists the three major categories of codes that are used in logistics-related activities. In warehousing, item-level SKU codes are used for inventory control, while unit-level RFID tags are just starting to be used to track each individual unit of an item in a warehouse, thereby facilitating FIFO stock rotation, for example. The use of a globally unique code allows products to move through the supply chain from firm to firm without the need to apply firm-specific codes when product is received at the warehouse door. The use of commodity codes is most useful for procurement activities where, for example, a similar item from multiple vendors is being sought and each vendor has a different SKU number.

Table 6.9. Logistics-Related Codes

	Commodity Code	**Item Code**	**Unit Code**
Level	Category	Class	Instance
Description	Grouping of similar objects	Grouping of identical objects	Unique physical object
Function	Product classification	Inventory control	Object tracking
Names	—	Item number, Part number, SKU, SKU + Lot number	Serial number, License plate
Codes	UNSPSC, GPC	GTIN, UPC, ISBN, NDC	EPC, SSCC, GLN

Common codes used in logistics include the following:

- UNSPSC (United Nations Standard Products and Services Code). A hierarchical 8-digit code that can used to classify all products and services at the segment, family, class, and commodity levels.

- GPC (Global Product Catalogue). A directory of product attributes that allows independent data repositories to be synchronized for global, multi-industry supply chain messaging and reporting.

- GTIN (Global Trade Item Number). All-numeric system for assigning globally unique codes to trade items (products and services). GTIN includes UPC, ISBN, and NDC.

- UPC (Universal Product Code). The standard bar code for retail items in North America.

- ISBN (International Standard Book Numbering). An international standard code for books.

- NDC (National Drug Code). A code maintained by the FDA for identifying prescription drugs and some over the counter drugs.

- EPC (Electronic Product Code). A globally unique serial number for physical objects identified using RFID tags.

- SSCC (Serial Shipping Container Code). A globally unique serial number for identifying movable units (carton, pallet, trailer, etc.).

- GLN (Global Location Number). A globally unique serial number for identifying physical places and legal entities. A location can refer to a physical place, such as a building or a storage area within a building (including a specific shelf in a warehouse), or a legal entity such as a company or a division of a company.[2]

6.4.3 Receiving

Receiving introduces inventory into the warehouse and prepares it for storage or customer order fulfillment. It is the process of unloading, verifying, inspecting, and staging of material transported to a warehouse in preparation for putaway or cross-docking, sometimes including sorting and repackaging of the material. Purchase orders (including shop orders and return authorizations) are sent to suppliers to authorize the shipment of material to the warehouse. In response to a purchase order, a supplier sends an advance shipment notice (ASN) to the WMS. The ASN contains information about the material contained in each movable unit (case, carton, pallet, etc.) contained in the shipment. Using SSCC, unique serial numbers can be assigned to each component at the time of its manufacture so that re-labeling is not required as part of the receiving function even when the material is not stored in the top-level container in which it is received.

The basic steps in receiving are the following:

1. Unloading of material from trailer.

2. Identify supplier with ASN, and associate material with each moveable unit listed in ASN.

3. Assign inventory attributes to movable unit from the item master file, possibly including repackaging material into new movable unit and assigning new serial number.

4. Inspect material to ensure that specifications are satisfied, possibly including holding some or all of the material for testing, and report any variances.

5. Stage units in preparation for putaway.

6. Update item balance in inventory master and assign units to a receiving area in location master.

7. Create receipt confirmation record.

8. Add units to putaway queue.

6.4.4 Putaway

Putaway is the process of moving material from the receiving area to a storage location or, in the case of cross-docking, directly to the shipping area. A putaway algorithm is used in the WMS to search for and validate locations where each movable unit in the putaway queue can be stored. The efficiency of all subsequent warehouse operations depends on performance of the putaway algorithm. The following inventory and location attributes are used in the algorithm to make the selection:

- *Environment*—used to restrict the locations where an item can be stored; e.g., refrigerated storage, caged area for high-value or controlled substances, quarantine area for units being held for inspection.

- *Container type* (pallet, case, or piece)—location can hold container type matching unit's type; a piece can be stored in a case or pallet location, and a case can be stored in a pallet location if necessary.

- *Product processing type*—specifies locations in processing area (e.g., floor, conveyable, nonconveyable) best suited for picking item.

- *Velocity* (A, B, or C)—matches turnover of item (A, fast; B, medium; C, slow) with the ease of storage and retrieval to/from location.

- *Preferred putaway zone*—item should be stored in location in the same zone as related items in order to, for example, improve picking efficiency.

An example putaway algorithm is as follows:

1. If moveable unit already allocated to a customer order, then it is moved (cross-docked) to shipping area.

2. If unit is being held for inspection, then it is moved to a location whose environment attribute is designated for quarantine storage; if no location found, then keep unit at its current location.

3. If unit not being cross-docked or held for inspection, then:

   (a) Search for available location that matches unit's environment, container type, product processing type, velocity, and preferred putaway zone attributes.

   (b) If no location found, drop preferred putaway zone attribute and repeat search.

   (c) Until location found, use next best velocity value and repeat search.

   (d) If no location found, restore original velocity value and, until location found, use next best product processing type and repeat search.

   (e) If no location found, report exception to operator.

6.5 Order Picking[3]

6.5.1 Basic Concepts

Order picking is the process of removing material from storage in response to specific customer orders or shop orders. Order picking is at the intersection of warehousing and order processing (see Figure 6.23): it includes the physical material handling processes associated with retrieving (or picking) items efficiently, and the information processing associated with searching and updating inventory records as orders are filled.
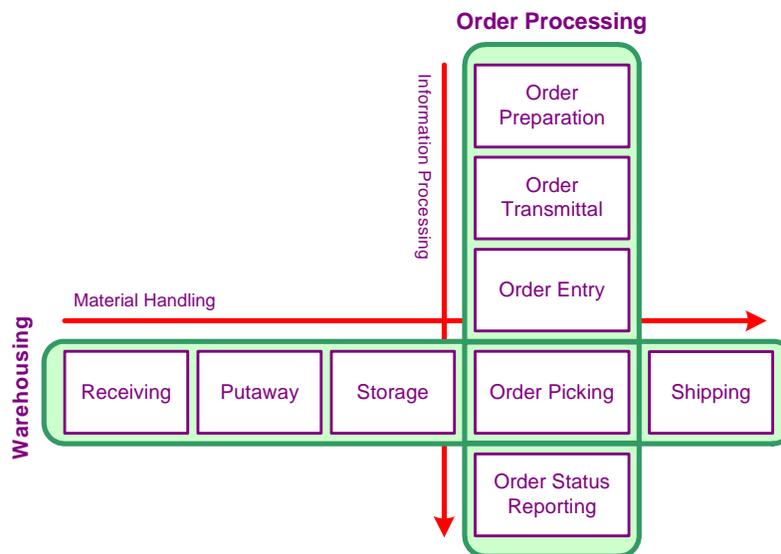

Figure 6.23. Order picking in relation to warehousing and order processing.

Order picking is the most critical activity in most distribution operations because it is the point at which customer expectations are actually filled. While the process of placing the fewer large unit loads into a warehouse is usually mechanized, the process of picking the many small items from a warehouse is often very labor intensive. This makes order picking the most costly warehousing activity, representing 55% of all operating costs in a typical warehouse.

An *order* indicates the type and quantity of items required. Each distinct type of item is termed a *SKU*. A *unit* is an instance of a SKU. Each SKU-quantity pair in an order is termed a *line*. A *pick list* indicates the sequence at which the storage locations of SKUs are to be visited along with the number of units to be picked from each location for one or more orders. Groups of orders are picked during planning periods termed *waves*. There can be one or more waves during each shift. Multiple waves are used to coordinate picking with other material flows in the facility and shipping schedules.

Table 6.10. Methods of Order Picking

Method	Pickers per Order	Orders per Picker
Discrete	Single	Single
Zone	Multiple	Single
Batch	Single	Multiple
Zone-Batch	Multiple	Multiple

Methods of Order Picking

There are three basic methods for order picking (see Table 6.10), with zone-batch picking being a combination method:

- *Discrete picking*. A single picker picks all of the items for a single order. Although an entire order can be packed while it's being picked, with no need for sortation and consolidation, travel time can be excessive if there is a low number of picks per order and congestion in aisles can occur if there are a large number of orders being picked.

- *Zone picking*. Each picker picks only the items of an order that are located in an assigned zone. This allows different techniques and equipment to be used in each different zone, and can reduce travel time as long as fast moving SKUs are located in the most accessible locations, but can be difficult to balance the amount of work in each zone (the "bucket-brigade" technique[4] can be used to dynamically balance each zone). Two variations of zone picking are *simultaneous picking*, where items for an order are picked simultaneously in each zone and then consolidated, making it possible to minimize the total picking time required for an order (which is useful if there are multiple waves per shift); and *progressive assembly*, where an order is passed from one zone to the next, eliminating the need to consolidate the order but increasing its total picking time (a.k.a. pick-and-pass).

- *Batch picking*. A single picker picks all of the items for multiple orders. This can reduce travel time (as long as the batched orders have items located in close proximity) and can reduce search time if multiple orders visit common locations, but items must be sorted into individual orders (sorting can occur during or after picking), and it might take a long time to accumulated enough orders that have items that are located in close proximity.

- *Zone-batch picking*. Combination of zone and batch picking, where multiple pickers each pick portions of multiple orders. This provides more opportunities for batching since items in the same zone are in close proximity and more orders with larger size items can be batched since picker does not carry full orders, but it requires the highest degree of coordination (e.g., can require both consolidation and sortation).

Levels of Order Picking

There are three major levels of order picking based on the size of the unit being picked (see Figure 6.24):

- *Pallet picking*, where full pallets of cartons or layers of cartons are retrieved (a.k.a. *unit-load picking*).

- *Case picking*, where full cartons of items are retrieved (termed *split-case picking* if inner packs of items from cartons are retrieved).

- *Piece picking*, where the individual units of issue to the customer of an item are retrieved (a.k.a. *broken-case picking*).
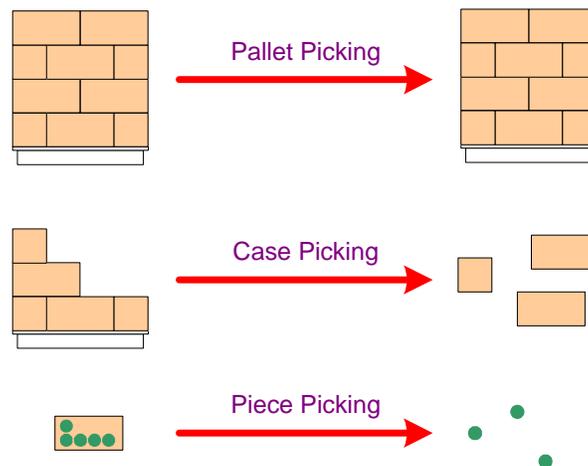
Figure 6.24. Levels of order picking.
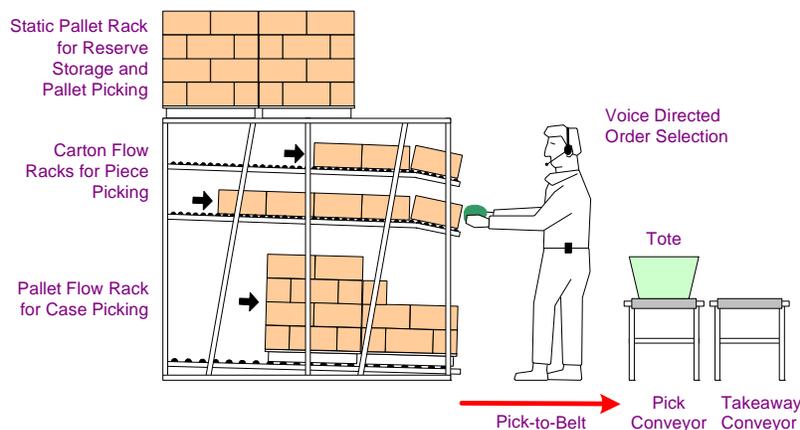
Figure 6.25. Example order picking operation.

Storage for pallet and case picking occupies the majority of space in a typical warehouse (see Figure 6.19); while piece picking is the most labor intensive and is typically the largest component of total order picking operating costs. One goal in designing an efficient picking operation is to try to pick the largest unit load size possible that will fulfill a customer's order;

154

picking, for example, one case instead of several pieces or one pallet instead of several cases. Figure 6.25 illustrates an order picking operation where all three levels of order picking occur in a single rack structure: using hands-free verbal transmission of pick instructions and pick confirmations, pieces are picked from inside a case on carton flow rack into tote on pick conveyor, and single cases are picked from pallet bottom-level flow rack directly onto the conveyor. Carton and pallet flow racks provide pick storage areas that minimize operator travel, and the takeaway conveyor transports completed orders and full totes.

Activity Profiling

Activity profiling is the systematic analysis of the items and orders handled in a warehouse in order to improve its design and operation. In the design of an order picking system, a representative set of customer orders are used together with the item master file to generate parameters that are used for a variety of different warehousing decisions, including equipment and method selection and slotting. If available, the previous three months to one year of customer orders provide a reasonable representative set of orders.

Item Master

SKU	Length	Width	Depth	Cube	Weight
A	5	3	2	30	1.25
B	3	2	4	24	4.75
C	8	6	5	180	9.65
D	4	4	3	32	6.35
E	6	4	5	120	8.20

Customer Orders

Order: 1	
SKU	Qty
A	5
B	3
C	2
D	6

Order: 2	
SKU	Qty
A	4
C	1

Order: 3	
SKU	Qty
A	2

Order: 4	
SKU	Qty
B	2

Order: 5	
SKU	Qty
C	1
D	12
E	6

Total Lines = 11

Lines per Order = 11/5 = 2.2

Cube per Order = 493.2

SKU	Flow per Item	Lines per Item	Cube Movement
A	11	3	330
B	5	2	120
C	4	3	720
D	18	2	576
E	6	1	720

Demand Correlation Distibution

| SKU	A	B	C	D	E
A		0.2	0.4	0.2	0.0
B			0.2	0.2	0.0
C				0.4	0.2
D					0.2
E					

Figure 6.26. Activity profiling example.

155

An example of activity profiling is presented in Figure 6.26. Each customer order in composed of one or more lines, where each line represents an item–quantity pair (e.g., Order 1 has five lines, with the first line indicating five units of item A have been ordered). In this example, the five customer orders along with the item master file are used to create the following warehouse design parameters:

- *Total Lines*—the total number of lines for all items in all orders over some period of time (representative of total picking activity); used to select piece-picking methods.

- *Lines per Order*—the average number of different items (i.e., lines or SKUs) in an order; used to select piece-picking method.

- *Cube per Order*—the average total cubic volume of all of the units (i.e., pieces) in an order; used to select piece-picking methods.

- *Flow per Item*—the total number of storage and retrieval operations performed for the item over some period of time; used to select pallet-picking equipment.

- *Lines per Item* (a.k.a. *popularity*)—the total number of lines for the item in all orders over some period of time (representative of picking activity for item); used to select case- and piece-picking equipment and for slotting.

- *Cube Movement*—the total unit demand of the item over some period of time times the cubic volume of each unit (representative of the cube in storage for the item); used to select pallet-, case-, and piece-picking equipment.

- *Demand Correlation*—the percent of orders in which both items appear; used for zoning and slotting (see below), but not for batching because that is an operational decision.

In Figure 6.26, the item master file includes the dimensions, cubic volume, and weight of each item. Note that an item's cube can be less than the product of its dimensions (e.g., items C and D). In addition, the unit of measure (UOM) is typically included. The UOM is a description of whether the quantity of inventory for an item refers to individual units (eaches or pieces), cases, or pallets. A conversion ratio is used whenever multiple units of measure are used for the same item.

Slotting

Slotting refers to the assignment of items to storage locations so that subsequent picking operations are best supported. Slotting for pallet picking differs from case and piece picking because only a single pallet can be picked at a single location. As a result, the DSAP covered in Section 0 is only appropriate for (single-command) pallet picking. For case and piece picking, the basic idea behind the DSAP, which is to use assign frequently picked (fast) items to the most convenient pick locations, is combined with demand correlation information in order to assign items that are picked together to locations that are close together (using). Flow per item measures pick frequency for pallets, and lines per item measures the number of times multiple units of item will be picked from locations where cases and pieces are stored. As shown in Figure 6.27, the most convenient pick locations are those that are in the "golden zone" and those that are close to the takeaway conveyor.
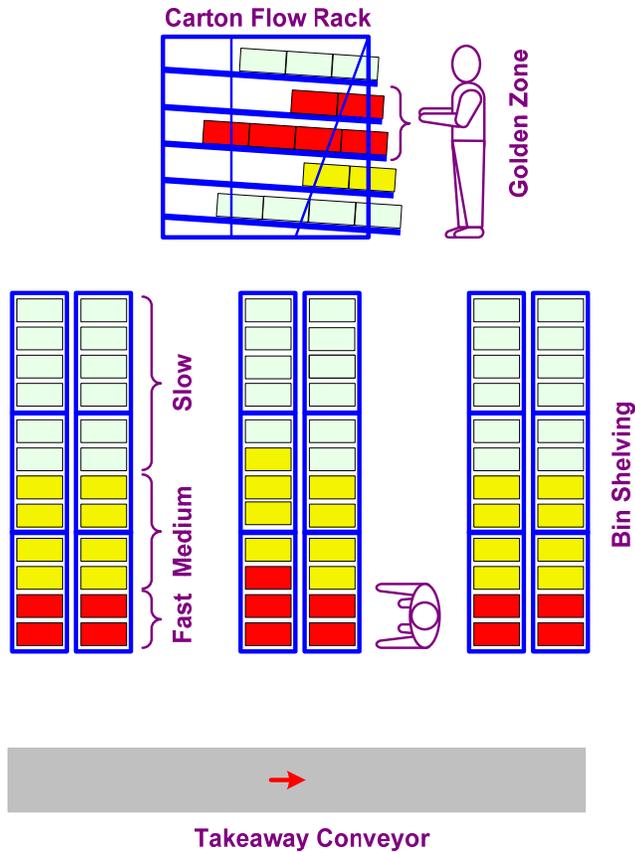
Figure 6.27. Slotting.
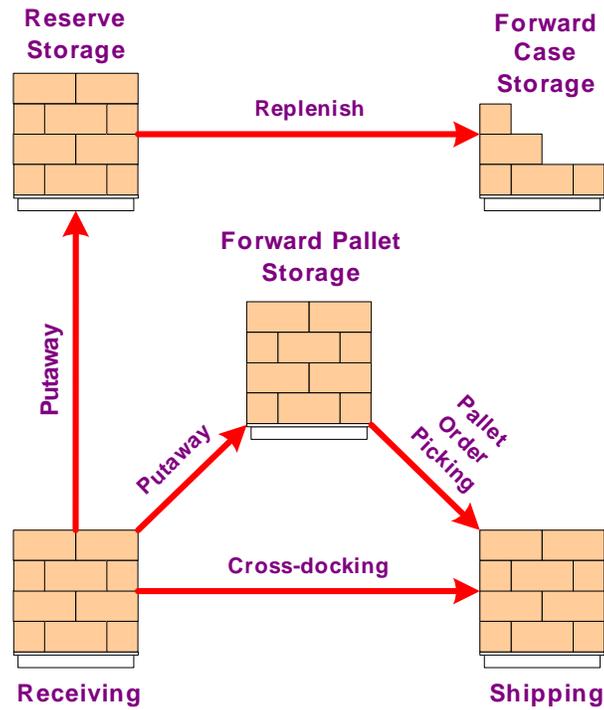

Figure 6.28. Pallet picking.
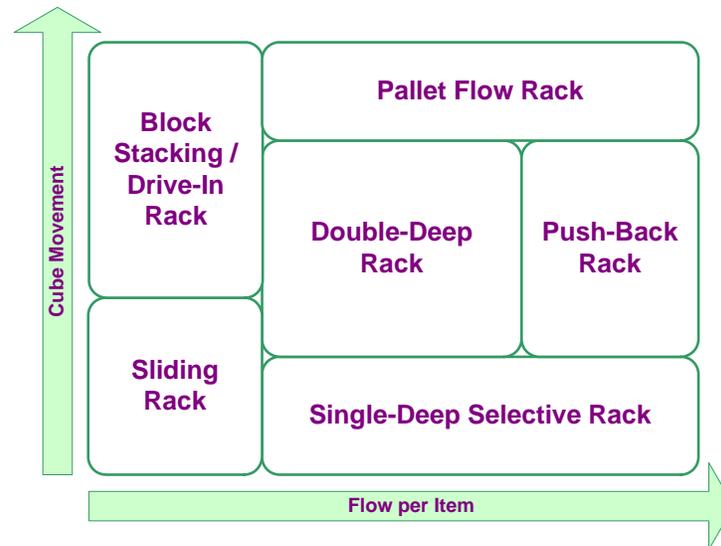
Figure 6.29. Pallet storage equipment.[5]

6.5.2 Pallet Picking

The transport of full pallet loads is used for putaway, replenishment, and cross-docking operations in a warehouse in addition to pallet order picking (see Figure 6.28). As shown in Figure 6.29, cube movement and flow per item can be used to select an appropriate type of equipment to store pallet loads of an item. Additionally, high labor costs can favor the use of a unit-load AS/RS.

6.5.3 Case Picking

Case order picking involves the retrieval of full carton loads of each item or inner packs of items from cartons. As shown in Figure 6.30, pallet loads of cases of the same item are transformed into pallet loads of mixed items through case order picking, sortation, and unitizing. Although much of the same storage equipment is used, case picking is more complex than pallet picking. As a result, a greater variety of case-picking techniques are available. As shown in Figure 6.31, cube movement and flow per item can be used to select an appropriate type of equipment for case picking.

Both manual and automated case picking is used. Pallet racks are used for all types of manual case picking, while automated picking uses specialized equipment. Also, carton flow racks can sometimes be used for case picking. The methods of manual case picking include pick to pallet, pick to belt, and pallet pick with sort.

Automated case picking equipment can provide high picking rates with no cost for labor. Replenishment and unitizing may or may not be fully automated. Two types of automated equipment are available: *flow delivery lanes*, where cases of each item are pushed from parallel merge chutes onto a sortation conveyor from which they are consolidated into individual orders and unitized, providing a high pick rate for fast moving items; and, for smaller cases, *case dispensers*, where cases are dispensed into larger containers that form each load.
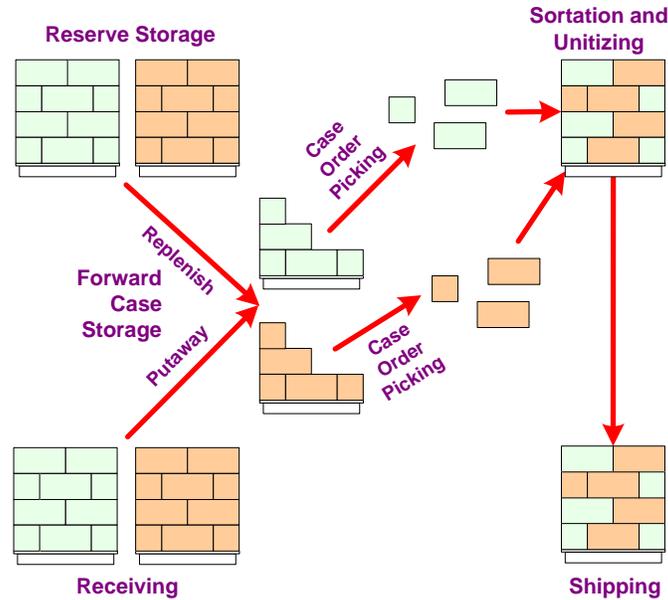
Figure 6.30. Case picking.
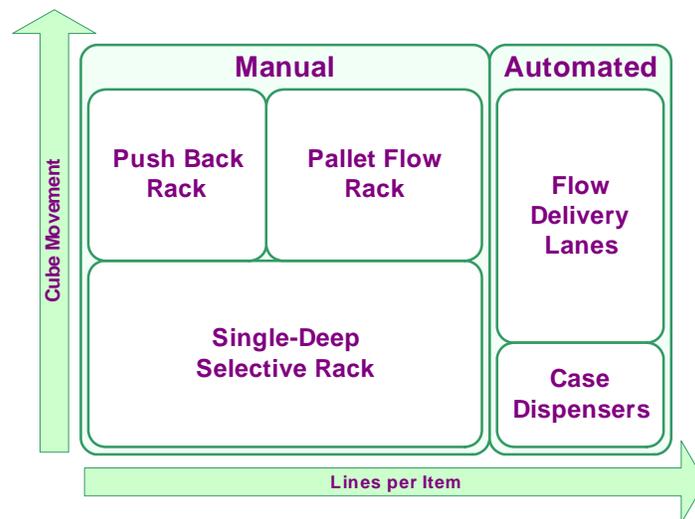
Figure 6.31. Case picking equipment.

Pick to Pallet

Pick to pallet is the simplest and most common type of case picking technique and it has the lowest equipment costs. Several variations are possible: floor- vs. multi-level, and discrete vs. batch.

Floor- vs. Multi-level Pick to Pallet

In floor-level picking (see Figure 6.32), only the bottom level of the pallet rack is used for picking and is replenished from pallets stored at higher levels. It is used for picking fast moving items since it is quicker to pick from the floor and replenish from above, and it is easy to combine pallet picking with case picking.

159

In multi-level picking (see Figure 6.32), all levels of the pallet rack are used for picking and are replenished from a separate reserve storage area. It is used to pick a large number of different slow moving items since entire rack is used for pick storage, and requires an order picker truck to lift picker to pallet (unless mezzanines are provided at each level, in which case pick to belt can be used on each level).
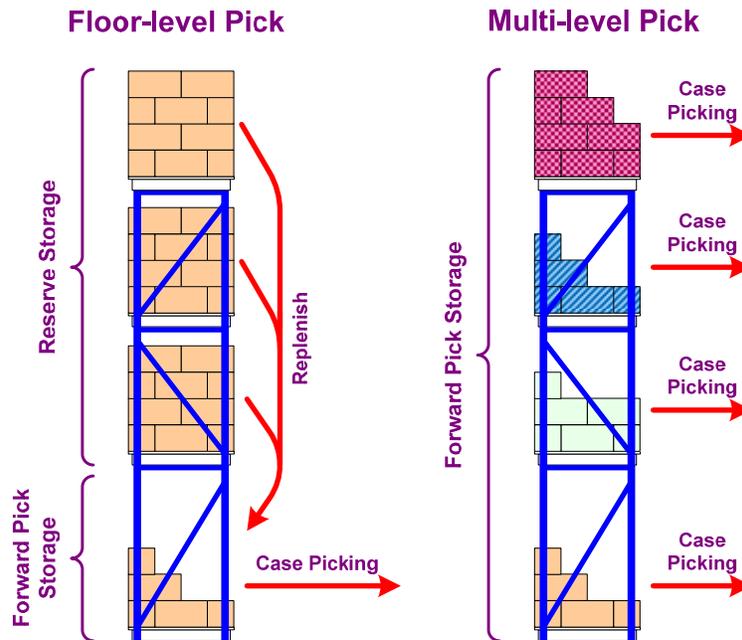
Figure 6.32. Floor- vs. multi-level pick to pallet.

Discrete vs. Batch Pick to Pallet

In discrete pick to pallet (see Figure 6.33), all of the cases in an order are picked onto the same pallet in which they will be shipped, thereby combining picking with unitizing. The picking can be floor-level or multi-level.

In pure batch picking, the cases needed for several orders are picked onto a single pallet and then sorted into individual orders. It is used when each order is much less than a pallet load.

In zone-batch picking (see Figure 6.34), cases are simultaneously picked to several pallets and then inducted onto a sortation conveyor from which they are consolidated into individual orders and unitized. This provides a high pick rate, but requires both a sortation conveyor and that each case has a label so that it can be scanned during subsequent sortation. The recirculation loop of the conveyor is used to control loading sequence onto shipping pallet.

Pick to Belt

In pick to belt picking (see Figure 6.35), cases for a batch of orders are simultaneously picked and inducted onto a sortation conveyor from which they are consolidated into individual orders and unitized. This provides a higher pick rate than pick to pallet because picking is combined with induction and it does not require the picker to travel. Replenishment does not interfere with

160

picking when using pallet flow racks are used, but it does require a sortation conveyor and labeled cases.
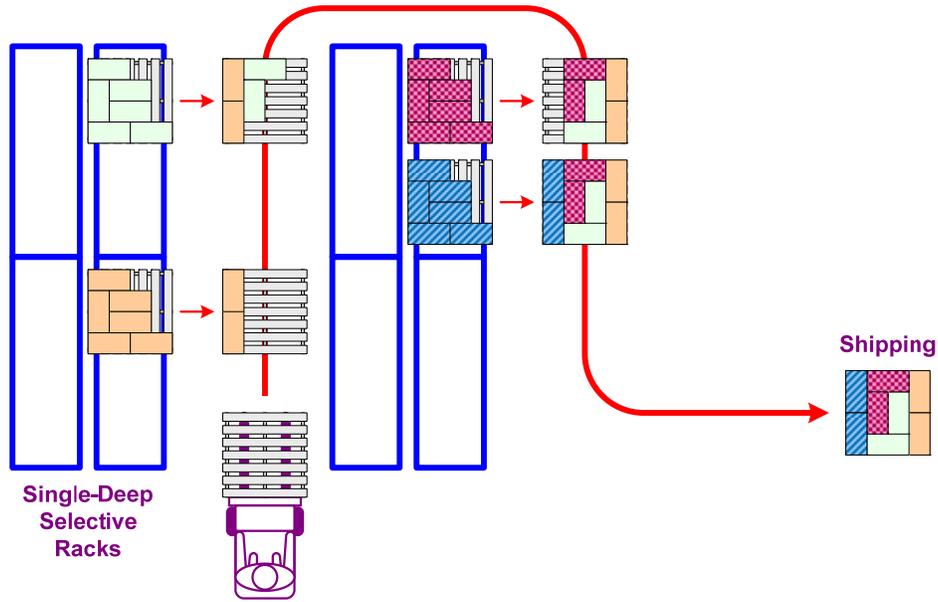

Figure 6.33. Discrete pick to pallet.
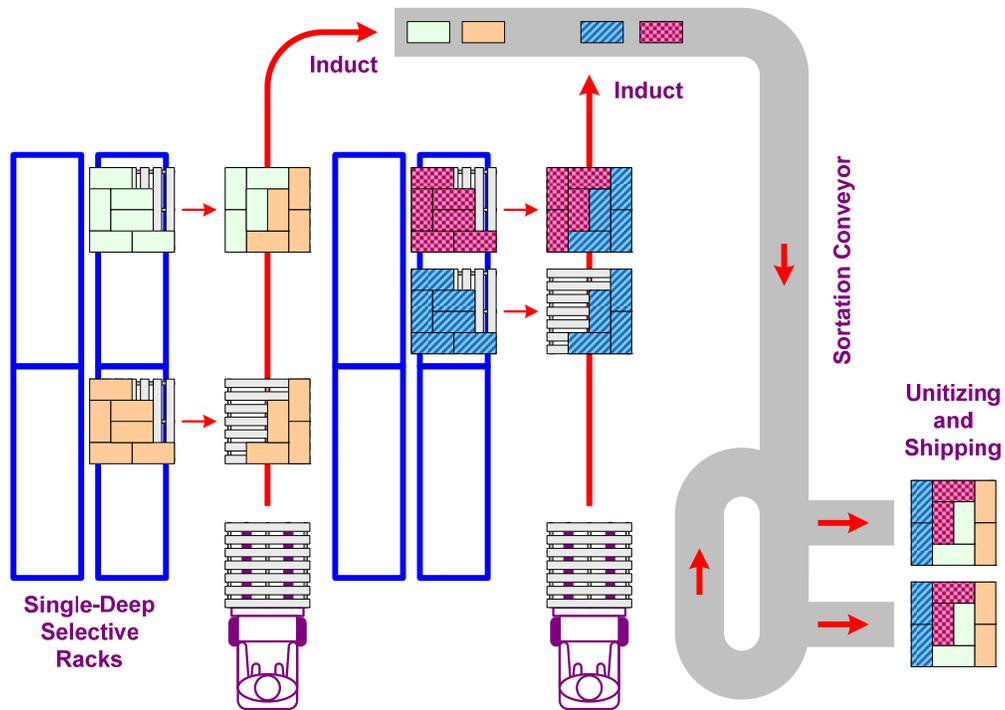
**Figure 6.34. Zone-batch pick to pallet.**
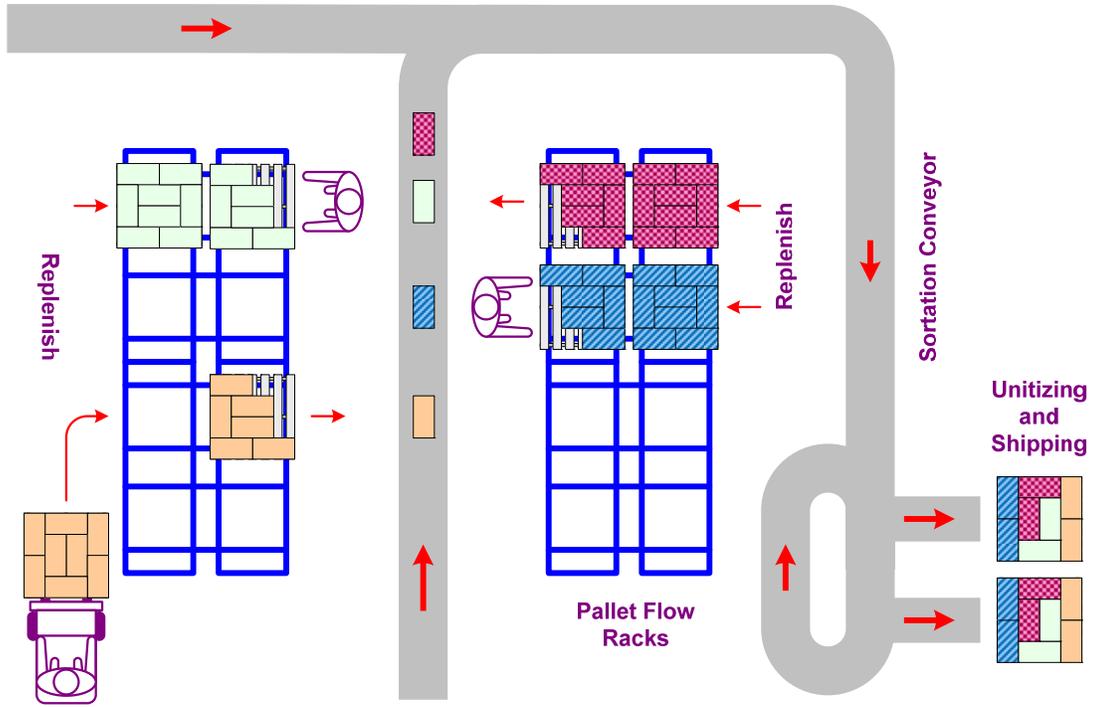
Figure 6.35. Pick to belt.
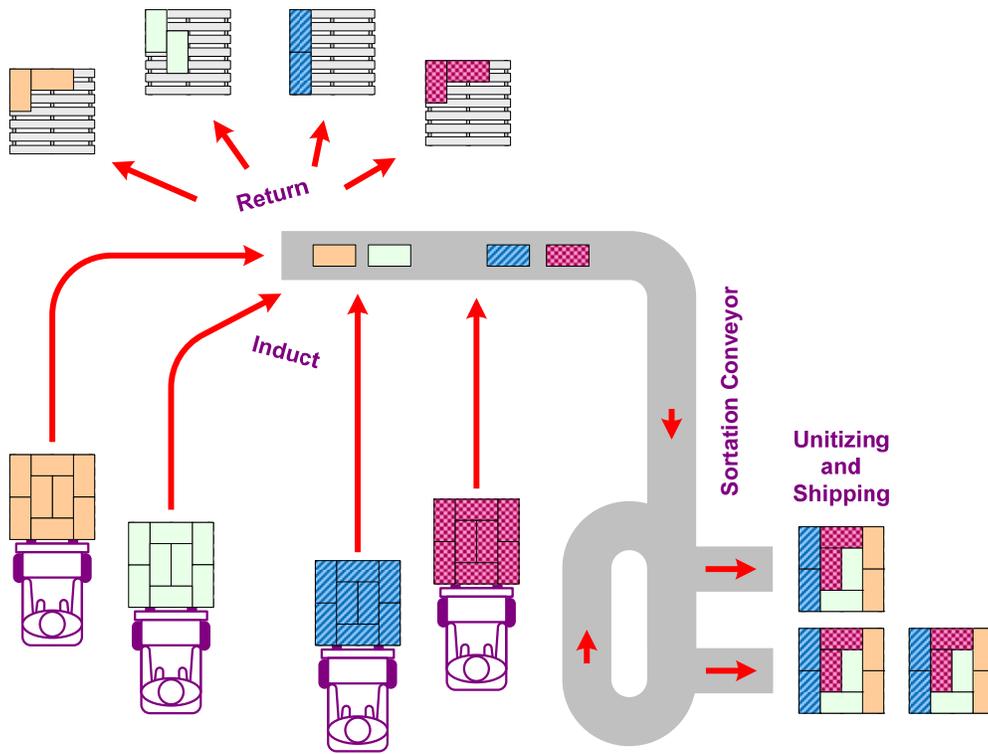
Figure 6.36. Pallet pick with sort.

Pallet Pick with Sort

In pallet pick with sort case picking (see Figure 6.36), full pallet loads of the items needed for a batch of orders are brought to a sortation conveyor where a portion of the cases from each pallet are inducted onto the conveyor and consolidated into individual orders and unitized. The partially full pallets are then returned to storage. This provides a high pick rate for fast moving items because pallets can be picked from reserve storage (i.e., no replenishment) , but it does require a sortation conveyor and labeled cases. Pallet pick with sort is similar to *putting*, and can be combined with pick to pallet for slower moving items.
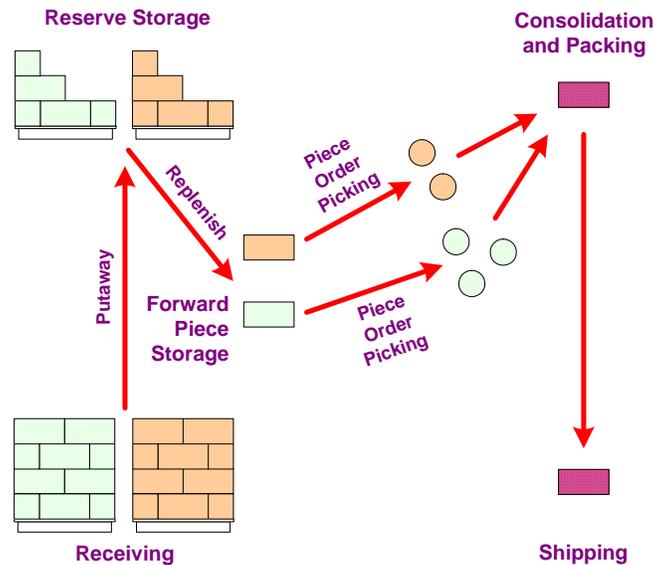

Figure 6.37. Piece picking.

6.5.4 Piece Picking

Piece order picking involves the retrieval of individual units of an item, where each piece picked (a.k.a. an *each*) is the unit of issue to the final customer. As shown in Figure 6.37, for each order, pieces are picked from cases of the same item and then packed into a container that is shipped to the customer. Piece picking is more complex than either case or pallet picking and, as a result, a variety of specialized piece-picking techniques and equipment have been developed. are available. As shown in Figure 6.38, cube movement and flow per item can be used to select an appropriate type of equipment for piece picking. Mini-load storage and retrieval machines can be used to enable fully automated replenishment of carton flow racks. Automation allows more frequent replenishment to better support picking operations and allows the item mix to be easily changed.
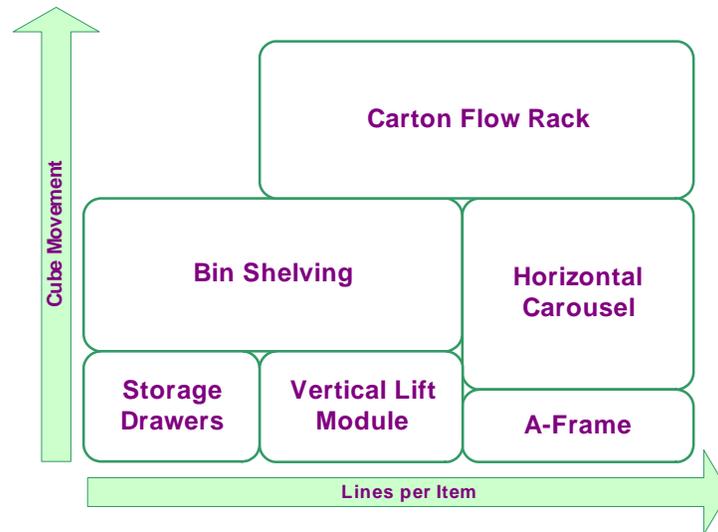
Figure 6.38. Piece picking equipment.[6]

As shown in Figure 6.39, an appropriate method for piece picking can be selected using the parameters lines per order, cube per order, and total lines. Discrete picking is infrequently used pick pieces (or cases). In the following, an example of how each method can be implemented for piece picking is presented.
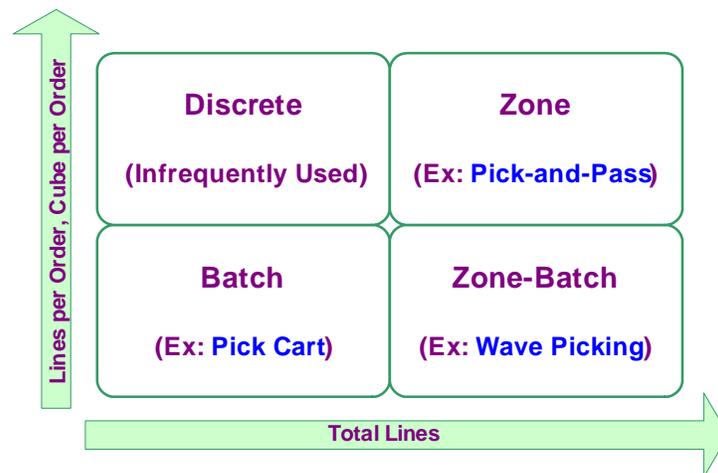
Figure 6.39. Methods of piece picking.

Batch Example: Pick Cart

In pick-cart batch piece picking, separate cartons or totes for each order are placed on a cart. Packing can occur either during or after picking:

- *Pick-and-pack*—shipping carton used during picking
- *Pick-then-pack*—reusable tote sent to packing station after picking (see Figure 6.40)

Although the use of a pick cart provides a low-cost, low-tech means for piece picking, poor pick-tour generation can result in excessive travel for picker. Pick carts can also be used for pick-and-pass zone-batch picking, where carts are passed from zone to zone.
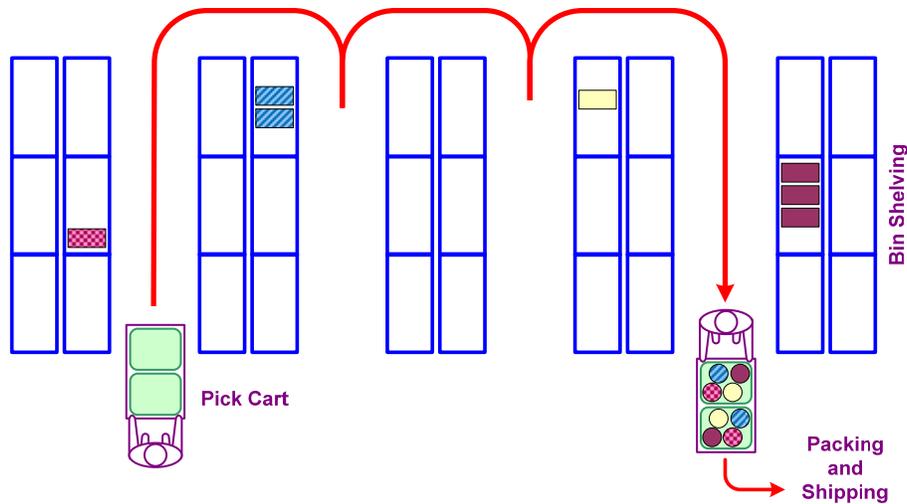

Figure 6.40. Pick-cart batch piece picking.

Zone Example: Pick-and-Pass

Items from each zone picked to tote and then tote is passed to the next zone. Cartons can be used instead of totes to allow pick-and-pack. Two different configurations used:

- *All zones visited*—if pick conveyor is attached to rack (see Figure 6.41), then tote is scanned at first zone and then visits each zone (totes maintain fixed sequence).

- *Skip zones*—if pick and takeaway conveyors are offset from racks, then takeaway conveyor can be used to move totes only to the zones with picks (requires tote scanning at each zone).
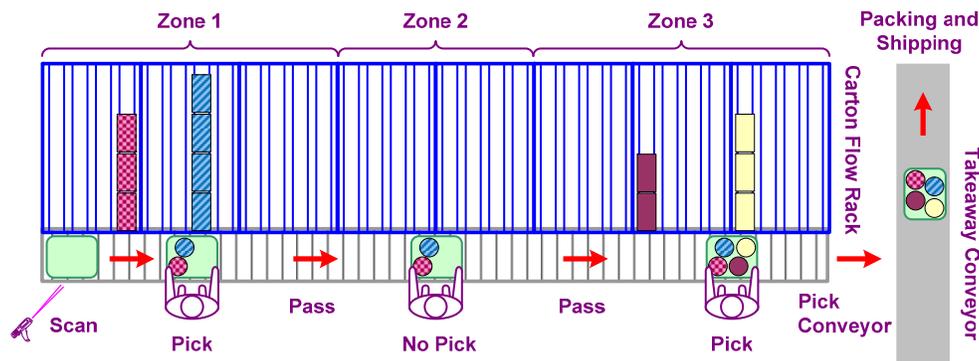

**Figure 6.41. Pick-and-pass zone piece picking.**
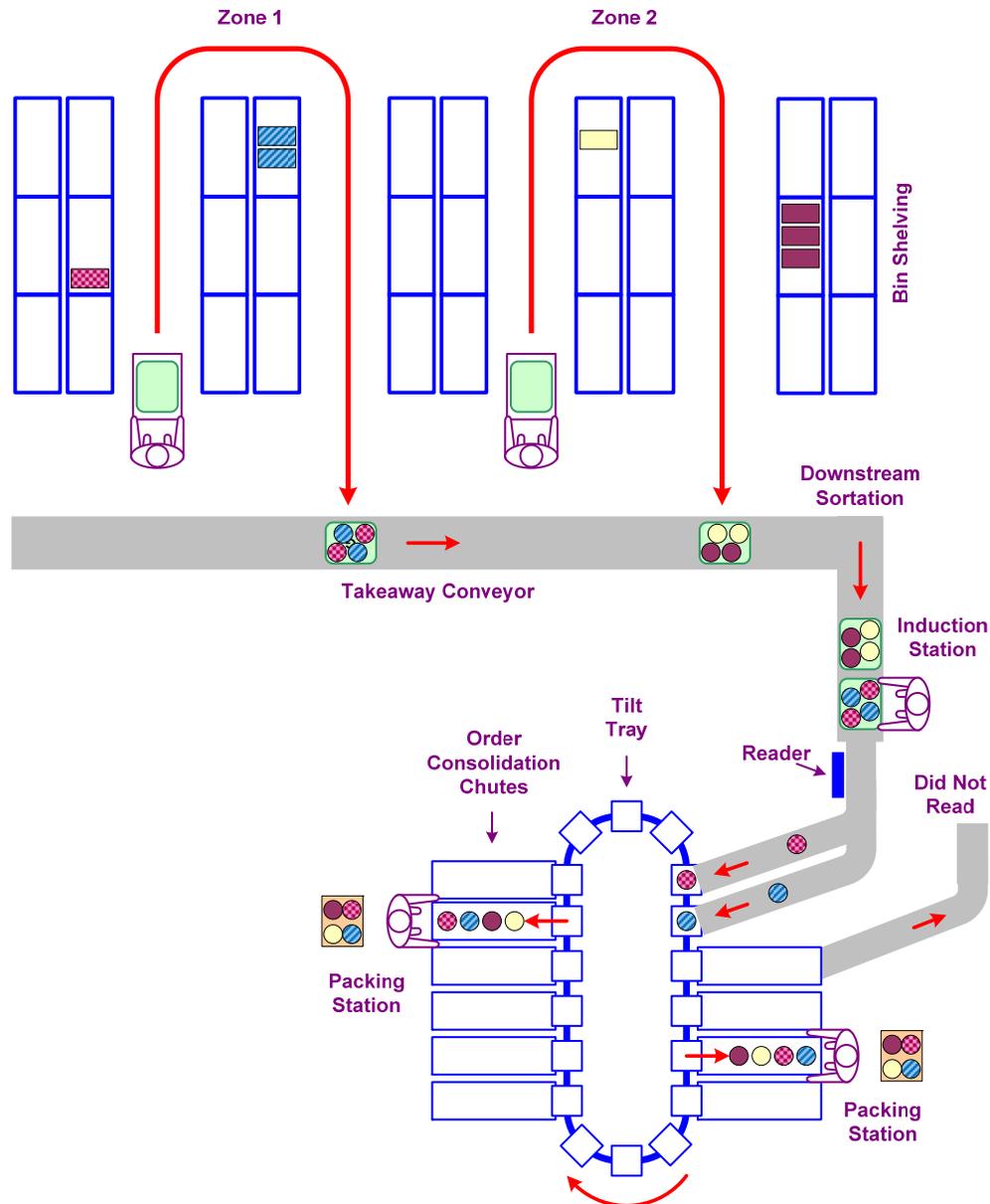
Figure 6.42. Wave zone-batch piece picking, including downstream tilt-tray-based sortation.

Zone-Batch Example: Wave Picking

Groups of orders are picked during a short period of time (a wave) in order to coordinate picking with shipping schedules or because downstream sortation has limited order capacity. Although wave picking enables a very high pick rate, it requires each piece picked to have a label so that it can be scanned during subsequent sortation. As shown in Figure 6.42, in each zone, all of the units of each item needed for all of the orders in the wave are picking into a tote. The totes from each zone are then sent downstream for sortation and packing. Pieces are then unloaded from totes, scanned, and then inducted onto sortation conveyor from which they are consolidated into individual orders and packed. Two different types of sorters are typically used:

166

- *Tilt-tray sorter* (pictured)—each piece slides onto a tray, which subsequently tilts to send the piece into its consolidation chute; sorting rates of over 10,000 units per hour are possible, but items cannot be fragile; the number of consolidation chutes limits the number of orders in each wave, and one operator can be used to pack multiple chutes.

- *Cross-belt sorter*—similar to a tilt-tray sorter except that a short belt conveyor is used in place of a tray; sorting rates of over 20,000 units per hour are possible.


(a) Picking Operation   (b) Putting Operation
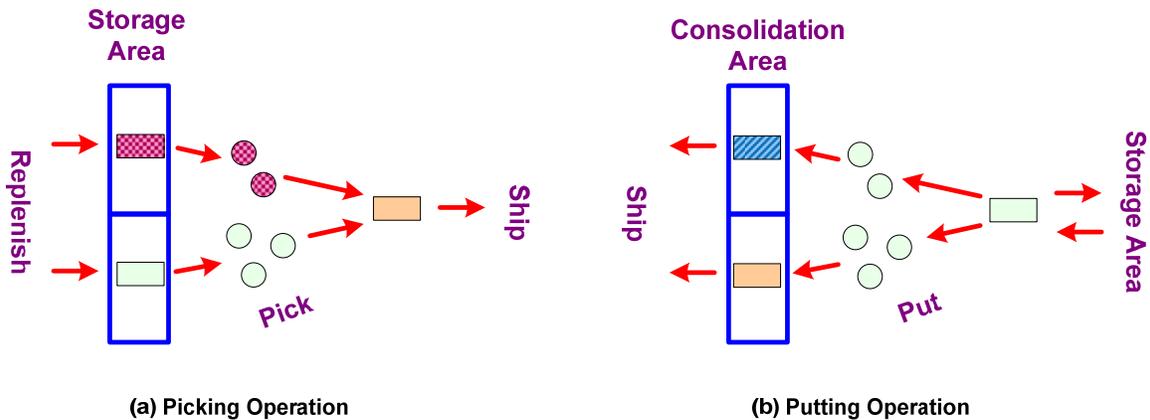
Figure 6.43. Picking vs. putting.

Picking vs. Putting

In some situations, it is more efficient to put instead of pick. In putting (see Figure 6.43(b)), a single carton or pallet load of an item is brought to a consolidation area and used to fill many orders at the same time. Picking consolidates many items into one order, while putting distributes (multiple units of) one item to many orders. Putting reverses the typical picking process:

- *Pick*—many items to one order

- *Put*—one item to many orders

Putting can be used to efficiently pick a large number of orders. In putting operations, shelves, carts, or carousels can be used for consolidation.

6.5.5 Picking Process

Most order picking processes involve the following basic steps:

1. Identifying the location of each pick.

2. Confirming the pick.

3. Indicating any shortage of product.

A variety of different identification and communications equipment can be used to implement the picking process. A communications link between the WMS and the pickers enables real-time rebalancing of the pick line during the picking process.

Line	Loc	SKU	Qty	UOM
1	E4	A	4	ea
2	D6	B	3	ea
3	C5	C	1	ea
4	B3	D	12	ea

**(a) Paper pick list**

(b) Picking tour
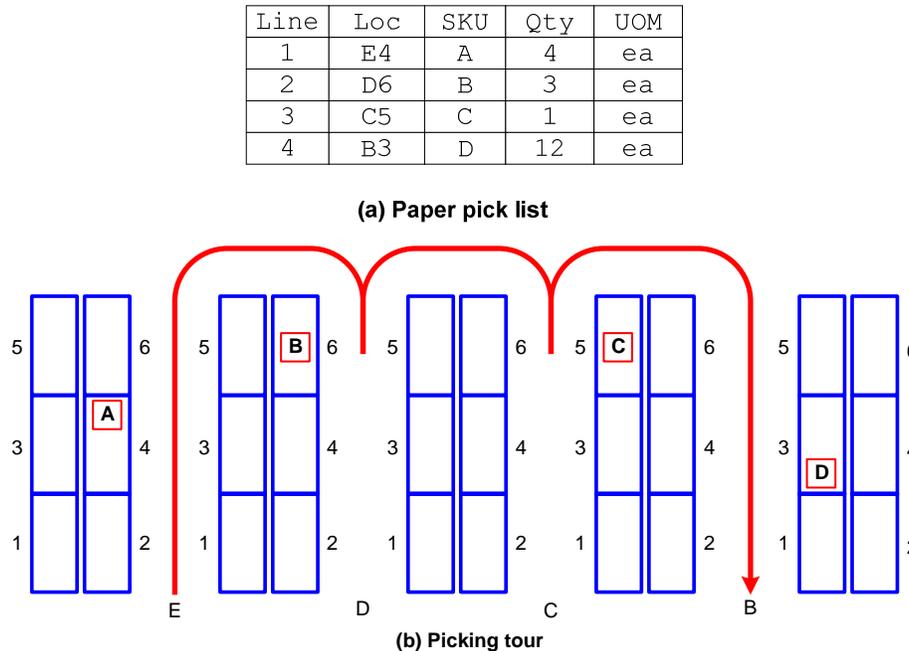
Figure 6.44. Pick-to-paper.

Pick-to-Paper

The basic pick-to-paper process is as follows (see Figure 6.44):

1. Paper pick list given to picker. List includes the location, SKU ID, quantity, and units of measure (UOM) of all items and in the sequence that they should be picked.

2. Weight scale on a pick cart might be used to (indirectly) confirm each item as it is picked, and any shortage can be noted on pick list.

The WMS generates the pick tour. Sometimes, location address sequence is used to determine tour, resulting travel up or down each aisle. Better algorithms (or picker modification) can generate shorter distance tours. The advantages of pick-to-paper are that it is reasonably fast, low-cost, low-tech, and an experienced picker can see entire tour and can often modify inefficient tours that have generated by the WMS tour-generation algorithms. The disadvantages are that the paper list held by the picker can interfere with picking, resulting in slower pick rates, and there is no direct pick confirmation. Also, because of the lack of a communications link to the WMS, it is not possible to perform real-time rebalancing of the pick line and shortages are only communicated at end of tour, thereby delaying the updating of the WMS.

Bar Code Scanning

The basic bar code scanning picking process is as follows:

1. Location, quantity, and SKU ID of an item to pick are presented to the picker on the display of the portable data terminal.

2. Picker then scans or keys-in the check digit to confirm the location.

3. Picker scans the unit or keys-in confirmation of the pick, noting any shortage.

4. Return to 1 if more picks.

Although scanning provides real-time pick confirmation and shortage indication, and the bar code labels and readers are low cost, it can slow down picking and the portable data terminal sometimes interferes with picking; also, the entire pick tour is usually not displayed to picker, which makes it difficult for the picker to modify the tour to improve its efficiency (cf. pick-to-paper).
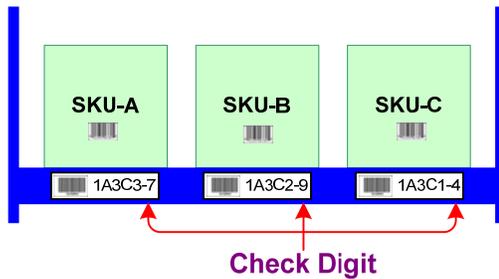

Figure 6.45. Location labels for bar code scanning.   **Figure 6.46. Portable data terminal.**
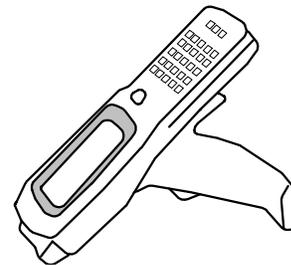
As shown in Figure 6.45, each location label includes a bar code and a printed address that includes a check digit. The check digit is typed using the keypad on the portable data terminal and provides a fast means of location identification in situations where it is not feasible to scan the location label (e.g., from a long distance). The portable data terminal can be handheld (see Figure 6.46), arm-mounted, or vehicle-mounted and is used for scanning bar code labels and communicates with the WMS via a radio frequency (RF) link.

Figure 6.47. Bar codes used in warehousing.[7]

A variety of bar codes are used in warehousing. Numeric bar codes are smaller in size than alphanumeric codes and are used when space is at a premium. The most common codes are the following (see Figure 6.47):

- Interleaved 2 of 5—used in warehouse for pick location labels.

- Code 39—used in warehouse to mark storage locations.

169

- Code 128—used in warehousing for shipping labels since it can represent all 128 ASCII characters instead of Code 39's 43.

- UPC/EAN—used for retail merchandise marking.
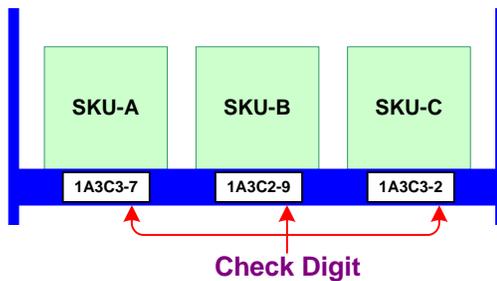

Figure 6.48. Location labels for pick-to-voice.
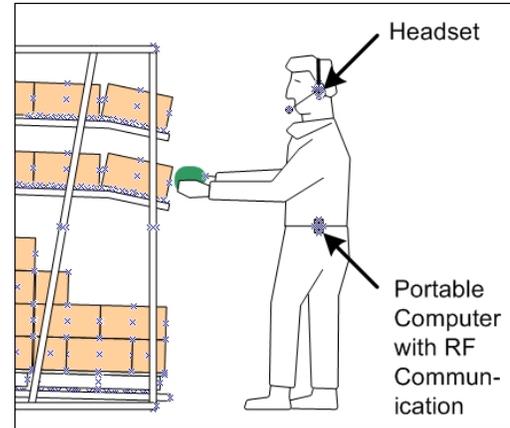
Figure 6.49. Portable computer and headset used in pick-to-voice.

Pick-to-Voice

The basic pick-to-voice process is as follows:

1. SKU ID and quantity of an item to pick are spoken to picker through a headset.

2. Picker then says the check digit to confirm the location (see Figure 6.48).

3. Picker says quantity picked followed by the word "picked" to confirm the pick, indirectly noting any shortage.

4. Return to 1 if more picks.

Although voice provides hands-free real-time pick confirmation and shortage indication, and location labels are low cost, speaking may slow down picking and it difficult for the picker to modify a pick tour to improve its efficiency since the entire tour is not known to the picker (cf. pick-to-paper). As shown in Figure 6.49, a portable computer used for voice processing and RF communications. Speaker-dependant and speaker-independent voice recognition available.
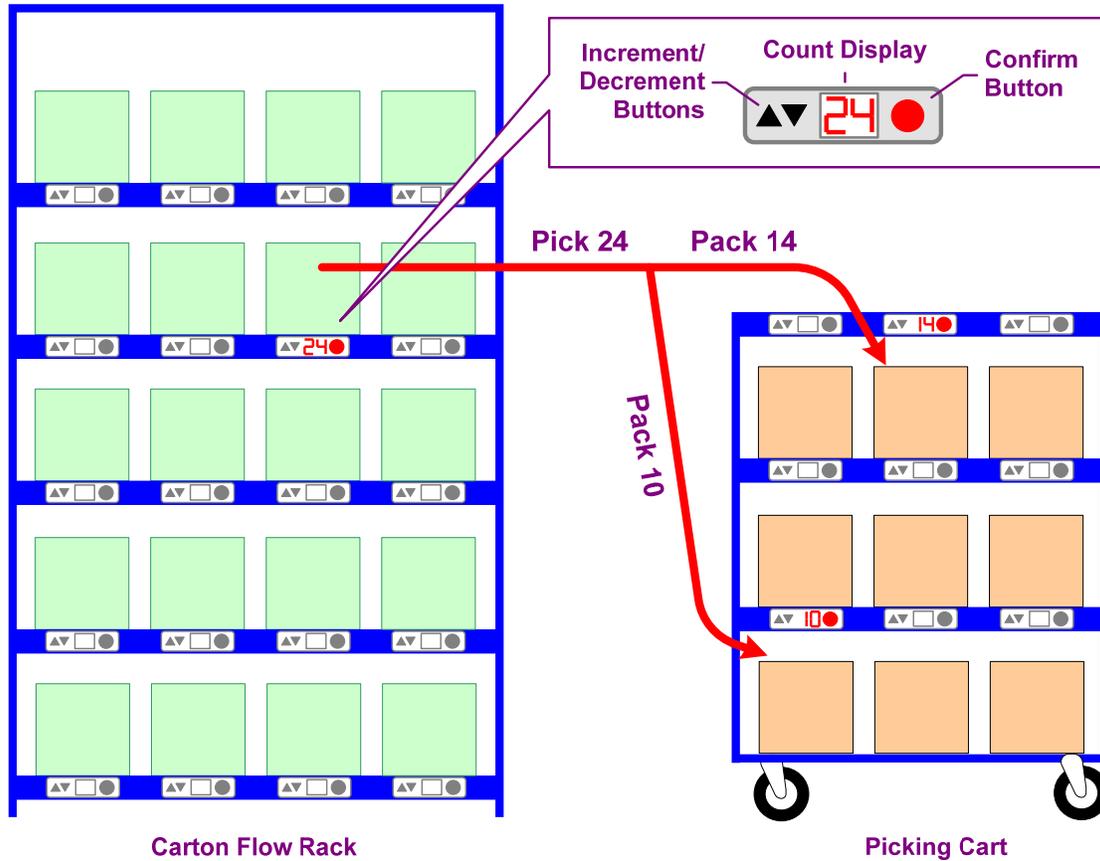
Figure 6.50. Pick-to-light.

Pick-to-Light

The basic pick-to-light process is as follows (see Figure 6.50):

1. Quantity of pick indicated by LED on display at the pick location.

2. Picker then hits button on display to the confirm location and pick, using decrement button to note shortage.

3. In batch picking, displays can also be used to indicate and confirm packing.

4. Return to 1 if more picks.

The main advantage of pick-to-light is that it can enable very fast picking and packing. Also, it provides real-time pick confirmation and shortage indication. Its main disadvantage is that display cost is proportional to the number of pick locations, as compared to the portable data terminal and voice recognition equipment used in bar code scanning and pick-to-voice, respectively, that are proportional to number of pickers (which is much less than the number of pick locations). The increment button on the display is only used for cycle counting. Displays communicate with the WMS via a wire network in the rack.

6.6 Warehousing Glossary

Activity profiling. The systematic analysis of the items and orders handled in a warehouse in order to improve its design and operation.

Advance shipping notice (ASN). Electronic information concerning a single shipment of movable units sent to a WMS from suppliers and sent from a WMS to customers.

Batch picking. An order picking method where a single picker picks all of the items for multiple orders.

Broken-case picking. Alternate term for piece picking.

Case picking. Retrieval of full carton loads of each item or inner packs of items from cartons (the latter a.k.a. split-case picking).

Consolidation. The process of (a) combining material from several partially filled storage locations containing the same item into a single location, (b) combining several orders into a single shipment, or (c) combining several portions of an order at a single location.

Cross-docking. The process of moving material from a receiving area directly to a shipping area without long-term storage of the material.

Customer order. Request that indicates the type and quantity of SKUs to be shipped to a customer; each SKU–quantity pair in the order is termed a line (cf. purchase order and shop order).

Cycle counting. The process of counting the contents of storage locations in order to verify the accuracy of inventory records.

Discrete picking. An order picking method where a single picker picks all of the items for a single order.

Each. An individual unit picked during piece picking.

ERP (Enterprise resource planning) system. Software system that control the entire operations of a firm. The item, carrier, and customer master files are maintained by an ERP system and are used as a common data source for orders and ASNs.

Forward picking. A storage area designed for efficient piece and case order picking that is usually replenished from reserve storage but sometimes directly from receiving.

Inner pack. Package used inside of a carton to allow more efficient split-case picking instead of individual piece picking when a less-than-carton-size number of units are to be picked.

Inventory master file. File maintained by a WMS that contains the total quantity and storage locations of each item stored in the warehouse. Used together with the location master file to control material transport operations.

Inventory. The number of units of each item stored in a warehouse.

Item. Grouping of identical objects, i.e., a class or collection of units; inventoried items are usually referred to as stock-keeping units (or SKUs).

Item master file. File that includes the dimensions, cubic volume, weight, and unit of measure of each item; used along with representative customer orders in activity profiling.

Line. An item–quantity pair in an order.

Location master file. File maintained by a WMS that contains the quantity of the item available at each storage location in the warehouse. Used together with the inventory master file to control material transport operations.

Movable unit. A single identifiable unit load (e.g., carton, pallet, trailer, etc.) that is moved between and stored at a location.

Order picking. The process of removing material from storage in response to specific customer orders or shop orders (cf. putting).

Order. Request to ship, receive, or transport material as indicated in a customer order, purchase order, or shop order, respectively.

Packing. The process of preparing a container for shipment.

Pallet picking. Retrieval of full pallets of cartons, or layers of cartons from a pallet (a.k.a. unit-load picking).

Pick conveyor. A non-powered conveyor (e.g., wheel or roller) used in piece picking to support a tote or other container while it is being filled.

Picking. Short for order picking (cf. putting).

Pick list. Request to a picker that indicates the sequence in which the storage locations of SKUs are to be visited along with the quantity of units to be picked from each location for one or more orders.

Pick-and-pass. Alternate term for progressive assembly picking.

Piece picking. Retrieval of individual units (or "eaches") of an item, where each piece picked is the unit of issue to the final customer (a.k.a. broken-case picking).

Progressive assembly picking. Variation of zone picking where an order is passed from one zone to the next, eliminating the need to consolidate the order but increasing its total picking time (a.k.a. pick-and-pass).

Purchase order. Request that indicates the type and quantity of items to be received from a vendor; each item–quantity pair in the order is termed a line (cf. customer order and shop order).

Putaway. The process of moving material from a receiving area to a storage location.

Putting. Putting reverses the typical picking process: in picking, units of many items are picked into one order; in putting, units of one item are put into many orders.

Receiving. The process of unloading, verifying, inspecting, and staging of material transported to a warehouse in preparation for putaway or cross-docking, sometimes including sorting and repackaging of the material.

Replenishment. The process of moving material from reserve storage to a forward picking area.

Reserve storage. An area intended for the storage of material in full pallet load sizes from which both forward picking areas are replenished and pallet orders and some case orders are picked.

Rewarehousing. The process of moving items to different storage locations to improve handling efficiency.

Shipping. The process of staging, verifying, and loading orders to be transported from a warehouse.

Shop order. Request that indicates the type and quantity of SKUs to be transported from a warehouse to a production area; each SKU–quantity pair in the order is termed a line (cf. customer order and shop order).

Simultaneous picking. Variation of zone picking where the items for an order are picked simultaneously in each zone and then consolidated, making it possible to minimize the total picking time required for an order (which is useful if there are multiple waves per shift).

Slot. Alternate term for a storage location.

Sortation. The process of merging, identifying, inducting, and separating material to be conveyed to specific destinations.

Split-case picking. Variation of case picking where inner packs of items from cartons are retrieved.

SSCC (Serial Shipping Container Code). A globally unique serial number for identifying a movable unit (e.g., a pallet).

SKU (Stock-keeping unit). An inventoried item.

Storage location. An identifiable location in a warehouse assigned a unique address and used to store a single item, where the capacity of the location corresponds to the maximum number of units of the item that can be stored at the location (a.k.a. slot).

Takeaway conveyor. A powered conveyor (e.g., belt or live roller) used in piece and case picking to transport completed orders.

Unit load. Either a single unit of an item, or multiple units so arranged or restricted that they can be handled as a single entity and maintain their integrity.

Unit. Instance of an item, i.e., a unique physical object.

Unitizing. The process of combining multiple smaller containers into a larger container that can be handled as a single unit load.

Unit-load picking. Alternate term for pallet picking.

Unit of measure (UOM). A description of whether the quantity of inventory for an item refers to individual units (eaches or pieces), cases, or pallets. A conversion ratio is used whenever multiple units of measure are used for the same item.

Wave. A planning period for picking groups of orders that can be used to coordinate picking with shipping schedules or because downstream sortation has limited order capacity; there can be multiple waves during each shift.

WMS (Warehouse management system). Software system that enables real-time, paperless control of the operations of a single warehouse.

Zone picking. An order picking method where each picker only picks the items of an order that are located in the portion of the storage area assigned to the picker for picking; simultaneous picking and progressive assembly picking are two variations of zone picking.

Zone-batch picking. A combination of zone and batch picking, where multiple pickers each pick portions of multiple orders.

6.7 References

The following sources are recommended for further study:

Adams, N.D., et al., 1996, *Warehouse & Distribution Automation Handbook*, McGraw-Hill.

Bartholdi, III, J.J., and Hackman, S.T., 2014, *Warehouse & Distribution Science*, version 0.96, www2.isye.gatech.edu/~jjb/wh/book/editions/wh-sci-0.96.pdf.

Frazelle, E., 2002, *World-Class Warehousing and Material Handling*, McGraw-Hill.

Mulcahy, D.E., 1994, *Warehouse Distribution & Operations Handbook*, McGraw-Hill.

Mulcahy, D.E., and Dieltz, J., 2004, "Order-Fulfillment Systems," in *Order-Fulfillment and Across-the-Dock Concepts, Design, and Operations Handbook*, St. Lucie Press, pp. 111–140.

Piasecki, D., 2005, "Order Picking: Methods and Equipment for Piece Pick, Case Pick, and Pallet Pick Operations," InventoryOps.com, www.inventoryops.com/order_picking.htm.

Notes

1 Facility Location

[1] *What are FOB shipping terms?* http://simplestudies.com/what-are-fob-shipping-terms.html (accessed Jan 2012).

[2] Sinnott, R.W., 1984, "Virtues of the Haversine", *Sky and Telescope* 68(2):159, as reported in http://www.census.gov/cgi-bin/geo/gisfaq?Q5.1

[3] Table is adapted from *Inbound Logistics*, Oct. 2004, p. 50.

[4] Krarup, J., and Pruzan, P.M., "Ingredients of locational analysis," in Mirchandani, P.B., and Francis, R.L., Eds., 1990, *Discrete Location Theory*, New York: Wiley, p. 20.

2 Freight Transport

[1] 25th Annual State of Logistics Report State of Logistics Report, Nov 2014, http://info.tranzact.com/hs-fs/hub/329734/file-2168640880-pdf/2mw-email/25th_Annual_SOL_Report_Update_11_10_14.pdf, p. 2

[2] Schulz, J.D., "State of Logistics 2010: Business logistics costs fall 18.2 percent," *Modern Material Handling*, Jun. 9, 2010.

[3] Trunick, P.A., 2005, "Can we hold the line on logistics costs?" *Logistics TODAY*, Penton Media.

[4] 21st Annual State of Logistics Report State of Logistics Report, as reported in *Inbound Logistics*, July 2010, p. 28.

[5] Logistics Costs and U.S. Gross Domestic Product, Federal Highway Administration, U.S. Dept. of Transportation, Aug. 25, 2005, http://ops.fhwa.dot.gov/Freight/freight_analysis/econ_methods/lcdp_rep/index.htm.

[6] 25th Annual State of Logistics Report, p. 13

[7] Taggart, S., 1999, "The 20-Ton Packet," *Wired*, Oct.

[8] Economist, 2004, "Boxed in and clogged up," *Economist*, Oct. 14.

[9] Matheson, B., 2003, *Inbound Logistics*, Dec., p. 42.

[10] SCAG, 2002, *SCAG Goods Movement Truck Count Study*, Sep. 25.

[11] LTL average value taken from: National Transportation Statistics, Table 1-53: Value, Tons, and Ton-Miles of Freight Shipments within the United States by Domestic Establishments, Bureau of Transportation Statistics, Washington, DC (Updated April 2006).

[12] PX column: Scherck, T.R., "A view of the future," Colography Group.

[13] Revised Hours-of-Service Regulations—2005, Federal Motor Carrier Safety Administration, http://www.fmcsa.dot.gov/rules-regulations/truck/driver/hos/brochure2005.htm.

[14] M.G. Kay and D.P. Warsing, "Modeling truck rates using publicly available empirical data," *International Journal of Logistics Research and Applications*, 12(3):165–193, 2009.

[15] Producer price index industry data: General freight trucking, long-distance, truckload. Series ID=PCU4841214841212 (Dec. 2003 = 100 base). Bureau of Labor Statistics, U.S. Department of Labor. http://www.bls.gov/ppi/home.htm.

[16] Termed "composite product" in Blumenfeld, D.E., Burns, L.D., Daganzo, C.F., Frick, M.C., and Hall, R.W., 1987, "Reducing logistics costs at General Motors," *Interfaces* 17(1):26–47.

[17] See [14] for details.

[18] Producer price index industry data: General freight trucking, long-distance, LTL. Series ID=PCU484122484122. Bureau of Labor Statistics, U.S. Department of Labor. http://www.bls.gov/ppi/home.htm.

[19] Pugh, W.W., 2005, *National Motor Freight Classification*, STB NMF 100-AF, National Motor Freight Traffic Association, Inc., Alexandria, VA.

[20] As specified in *National Classification Committee Density Guidelines*, National Motor Freight Traffic Association, Inc., http://www.nmfta.org/Portals/0/Documents/NCC Density Value Guidelines 2005 0809.pdf.

[21] Popken, D.A., 1996, "An analytical framework for routing multiattribute multicommodity freight," *Trans. Res.-B* 30(2):133–145.

[22] CzarLite Online, http://www.smc3.com/Applications/WebCzarLite/Entry.asp.

[23] See [14] for details.

[24] Freight101.com, Instant Quote. https://www.freight101.com

[25] Hall, R.W., 1985, "Dependence between shipment size and mode in freight transportation," *Transportation Science* 19(4):436–444.

4 Capacity Planning

[1] Hopp, W.J., and Spearman, M.L., 2000, Factory Physics: Foundations of Manufacturing Management, 2nd Ed., Boston: Irwin McGraw-Hill.

[2] Suri, R., J.L. Sanders, and M. Kamath (1993). "Performance evaluation of production networks," *Handbooks in OR & MS*, Vol. 4, S.C. Graves et al., Eds., pp. 199–286.

[3] Adapted from Fig. 1.1 in Taylor and Karlin (1998)

5 Facility Layout

[1] Fig. 2.18 in Francis, R.L., McGinnis Jr., L.F., and White, J.A., 1992, *Facility Layout and Location: An Analytical Approach*, 2nd Ed., Englewood Cliffs, NJ: Prentice-Hall.

[2] Fig. 3.34 in Tompkins, J.A., et al., 2003, *Facilities Planning*, 3rd Ed., Wiley.

[3] Tompkins, *Facilities Planning*, Table 4.4.

[4] Tompkins, *Facilities Planning*, p. 514–515.

6 Storage and Warehousing

[1] Frazelle, E., 2002, *World-Class Warehousing and Material Handling*, McGraw-Hill.

[2] GS1, 2011, "The value and benefits of the GS1 system of standards, http://www.gs1.org/docs/GS1_System_of_ Standards.pdf (accessed March 2011).

[3] Much of the material in this section is based on web-based order picking training modules developed for the Material Handling Industry of America, www.mhia.org/et/et_mhi_elessons_home.cfm.

[4] Bartholdi, J., and Eisenstein, D., 1996, "A production line that balances itself," *Operations Research*, 44(1):21–34.

[5] Adapted from Fig. 5-11 in Frazelle, *World-Class Warehousing*.

[6] Adapted from Fig. 2-12 in Frazelle, *World-Class Warehousing*.

[7] Barcodes made using the online Barcode Label Printer, http://www.barcodesinc.com/generator/.