)enver
952.01
900.05
612.42
422.84
664.63
1466.74
384.57
1536.80
306.70
442.68
1130.05
591.75
852.37
628.55
133.77
999.92
1356.57
590.11
1803.16
1428.37
991.56
473.06
793.91
851.39
1608.87
1440.93
927.53
1071.48
970.99
1091.49
884.15
938.48
1263.84
712.74

2 Cost

Readings for Chapter 2

Daganzo and Newell (1990) describe a model for handling operations, and examine trade-offs among handling transportation and inventory costs. Section 2.4 covers much of the same material. Blumenfeld, Hall, and Jordan (1985), and Horowitz and Daganzo (1986) examine minimum cost shipping strategies, with random demand and travel times, when a fast and expensive transportation mode can be used to forestall shortages. Part of Section 2.5 is devoted to this subject.

2.1 Initial Remarks

This chapter describes how to account for the various costs arising from a logistics operation; it also introduces related terminology and notation. Although this will be done in the context of a single origin producing identical items[1] for a single consuming destination, the formulas and concepts extend to the more general scenarios examined in the latter chapters of this monograph. Any modifications are described in these chapters. This section presents a framework for the classification of logistics cost; specific cost types will be analyzed in the following sections.

In tracing the path of an item from production to consumption, we see that it must be:

(i) carried (handled) from the production area to a storage area,
(ii) held in this area with other items, where they wait for a transportation vehicle,
(iii) loaded into a transportation vehicle,
(iv) transported to the destination, and
(v) unloaded, handled, and held for consumption at the destination.

[1] In this monograph we will often call the indivisible units that move over a logistics system, e.g., persons, letters, parcels, etc., "items." When the logistics system handles an infinitely divisible commodity, such as fluids and grain, the term "item" may also be used; in that context it will denote a fixed, and usually small, quantity of the commodity.

These operations incur costs related to *motion* (i.e., overcoming distance) and cost related to *"holding"* (i.e., overcoming time).

Motion costs are classified as either *handling* costs or *transportation* costs. They are very similar; the main difference being the distances transported and the size of the batches moved together. Handling costs include *packaging* (in step *(i)* above); transportation costs include *loading*. Of course, loading is also a handling activity; and if a clear distinction is desired, one could define as a *handling* cost the portion of loading costs that arise *outside* the *transportation vehicle*, and as a transportation cost, the portion that arises inside the vehicle. It is not really crucial that the cost of the specific action be allocated to a "correct" category. What is important is that in the final analysis all costs are included and none are double counted.

Holding costs include *"rent"* costs and *"waiting"* costs. This is not a generally accepted terminology, but it is useful for our purposes. As the name implies, rent costs include the *rent* for the space, *machinery* needed to store the items in place, plus any *maintenance* costs (such as security, utilities, etc.) directly related to the provision of storage space. Waiting costs are meant to capture the cost of delay to the items, including: the opportunity cost of the capital tied up in storage, any value lost while waiting, etc. For a given set of fixed facilities (machinery and space), thus, the rent costs remain fixed, but the waiting costs depend on how the items are processed; i.e., the rent – unlike the total waiting cost per unit time – does not depend on the amount stored. We will examine these four cost categories one by one, and see how they can be quantified. Our goal is to identify which parameters influence the various costs, and the mathematical form of the relationships.

In analyzing these relationships, it is also important to choose how to present them. For example, one could measure transportation cost as: cost per item transported, cost per year, cost per trip, etc. But not all of these representations are valid for analysis. The cost per item can be converted to cost per year if we multiply it by the number of items produced in a year. The cost per item can be converted to cost per trip if we multiply it by the number of items in the transportation vehicle. Two representations are equivalent if the *conversion factor* is *a constant* that does not depend on the decision variables. For example, if we seek the optimal vehicle dispatching frequency that will maximize the yearly profit for a given production level, the desired solution can be found by minimizing the total cost per year – when price and production levels are constant, yearly profit can be related to yearly costs by a known non-increasing function. The same solution could also be obtained by minimizing the average cost per item because the conversion factor, items produced per year, is a constant. The cost per trip, however, would lead to an erroneous solution. In the remainder of this monograph we will assume that the yearly demand for items does not depend on the decision variables, and, therefore, it will be possible to express cost either as a total per unit time or a prorated average per item.

In our discussion we will usually include all the costs incurred by the items from origin to destination regardless of who pays them (the shipper, the carrier, or somebody else). If ownership of the item changes at some point during transportation (e.g., on arrival at the destination), waiting costs at the origin will be paid by the producer, and inventory costs at the destination by the consumer. While one may feel that costs borne by any entity other than our "client" (i.e., the organization whose operation we are trying to optimize) should be ignored, this is shortsighted. Such an optimization would tend to transfer the burden of the operation to entities other than our client (since their costs are not being considered); and as a result, they may be less willing to participate in the operation. If, for example, a producer ships infrequently (which minimizes its own transportation costs) and, as a result, causes large inventories at the destination, the consumer will be less willing to pay the price – and may expect a discount. Such a discount would obviously have to be included in the optimization of the shipping frequency, but it is difficult to quantify. Our expressions automatically include the quantity that the discount would represent – the increased cost to the consumer. Of course, if this is not desired, appropriate terms can be deleted from the expressions; the techniques remain the same.

Let us now turn our attention to the various cost components. Section 2.2 discuses holding costs, Sec. 2.3 transportation costs and Sec. 2.4 handling costs. Section 2.5 explains how uncertainty and random phenomena influence cost accounting.

2.2 Holding Costs

A sufficiently detailed quantitative description of holding costs can be given in the context of a simple scenario with one origin and one destination. Consider the situation depicted in Fig. 2.1, where items are produced and demanded at a constant rate, D' . The four curves of the figure represent the cumulative number of items to have been: (i) produced, (ii) shipped, (iii) received at the destination, and (iv) consumed. We assume that the ordinates of the curves at time zero (when observation began) have been chosen so as to ensure that the vertical separation between any two curves at that time equals the number of items initially observed between the corresponding stations.

Rarely used in the inventory and queueing literature, cumulative count curves such as those depicted in Fig. 2.1 are particularly useful to trace items through consecutive stages. In our case, they conveniently describe in *one* picture how the number of items in *various*, logistic states (waiting for transportation, being transported, and waiting for consumption) change with time. Notice that the number of items waiting for transportation at *any given time* is the vertical separation between curves (i) and (ii) at the corresponding point on the time axis, the number being transported is the vertical separation between curves (ii) and (iii), and the number waiting for consumption is the vertical separation between curves (iii) and (iv).

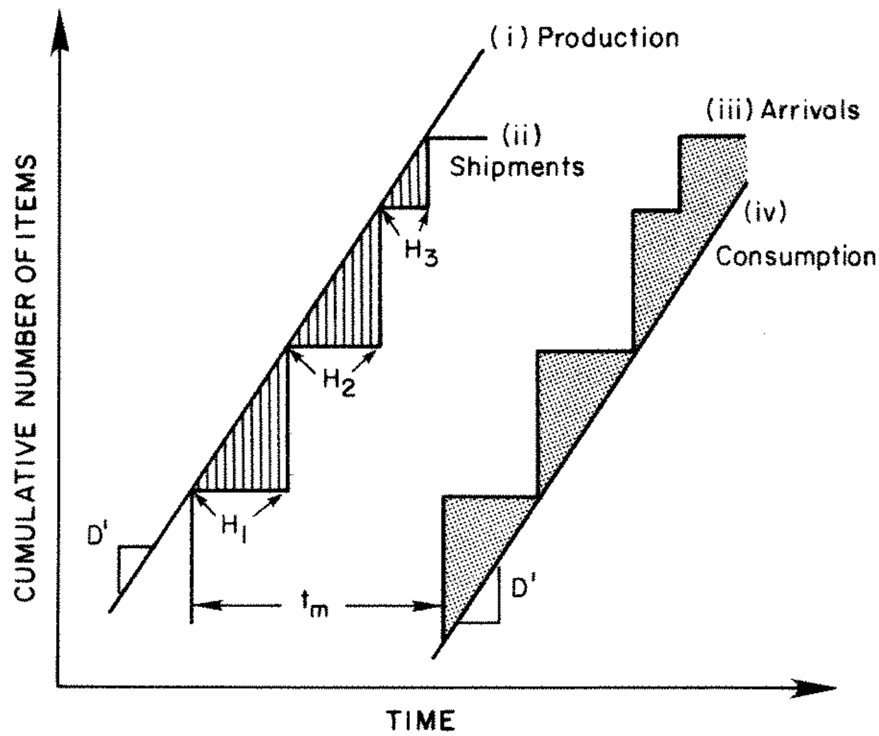

Fig. 2.1 Cumulative item counts at different stages in the logistics operation

Chapter 1 in Newell (1982) shows in detail how various other measures of performance can also be gleaned from these graphs. Of special interest here are horizontal separations between the curves and the intervening areas. When items pass through the system in a "first-in-first-out" order, then the n^{th} item to be counted at each observation station (i, ii, iii, or iv) is the same item; as a result, the horizontal separation between any two curves at ordinate "n" represents the amount of time spent by that item between the corresponding stations. In the figure, thus, t_m represents the transportation time. It should be intuitive that areas between curves represent total amount of wait (in "item-hours") regardless of the order in which items are processed. Thus, the shaded area in the figure represents the number of "item-hours" spent at the origin, and the dotted area represents the number at the destination. It follows that the average horizontal separation between two curves, measured between two points where the curves touch, represents the average time that a typical item spends between the operations represented by the curves. (The average horizontal separation between the curves can be expressed as the ratio of the area, i.e., the total wait, and the vertical separation between the two points, i.e., the number of items processed. Such a ratio is, by definition, the average wait per item.)

In our example, the (constant) separation between the production and consumption curves represents the average "waiting" that an item has to do between production and consumption. This is equal to t_m plus the *maximum* interval (or headway) between successive dispatches, $H_1 = \max\{H_i\}$ (see figure):

$$\overline{wait} = H_1 + t_m . \qquad (2.1a)$$

The room needed for storage at any given location should be proportional to the maximum number of items present at the location. This is represented in Fig. 2.1 by the maximum vertical separations between curves. Because the figure assumes that each vehicle carries all the items that have been produced, the storage area required at the origin is proportional to the maximum headway (otherwise the maximum inventory accumulation would be larger); i.e.:

$$maximum\ accumulation = D' H_1 \qquad (2.1b)$$

The maximum accumulation at the destination is the same as it is at the origin (the reader can verify this from the geometry of the figure, remembering that $H_1 = \max\{H_i\}$).

The expressions for average wait and maximum accumulation can be translated into costs *per item* or *per unit time* using cost conversion factors.

2.2.1 Rent Cost

This is the cost of the space and facilities needed to hold the maximum accumulation; for properly designed systems it should be proportional to the maximum accumulation. The proportionality factor will depend on the size of the items, their storage requirements, and the prevailing rents for space. If the facilities are owned (and not leased), then the purchase cost should increase roughly linearly with size. Thus, one can compute an equivalent rent (based on the amortized investment cost over the life of the facilities) which should still be roughly proportional to the maximum accumulation.

Let c_r be the proportionality constant (in $ per item-year); then

$$rent\ cost/year = c_r\ (maxiumum\ accumulation) \tag{2.2a}$$

and if the demand is constant, Eq. (2.1b) allows us to write:

$$rent\ cost/item = c_r\ (max\ accumulation)/D' = c_r H_1 \tag{2.2b}$$

Note that the rent cost per item is independent of flow (the production and consumption rate D') and proportional to the maximum time between dispatches.

2.2.2 Waiting Cost

Waiting cost, also called inventory cost, is the cost associated with delay to the items. As is commonly done in the inventory literature, it will be captured by the product of the total wait done by all items and a constant, c_i , representing the penalty paid for holding one item for one time unit (usually a year). Thus,

$$waiting\ cost/year = c_i (total\ wait\ per\ year)$$

and

$$waiting\ cost/item = c_i (average\ wait/item).$$

Because the above expressions implicitly value all the item-hours equally, caution must be exercised when the penalty depends on: (i) the time of day, week, or year when the wait occurs, and (ii) how long a specific item has already waited. For the example in Fig. 2.1, the waiting cost is:

$$\begin{aligned} \textit{waiting cost/year} &= c_i \left[D'(H_i + t_m)\right] \\ &= (c_i D'H_1) + (c_i D't_m) \end{aligned} \quad , \qquad (2.3a)$$

$$\begin{aligned} \textit{waiting cost/item} &= c_i \left(H_1 + t_m\right) \\ &= (c_i H_1) + (c_i t_m) \end{aligned} \quad . \qquad (2.3b)$$

The left side of Equation (2.3a) assumes that the time unit is one year. The term in brackets represents the average accumulation of inventory in the system (the vertical separation between the production and consumption curves of Fig. 2.1). As we shall see, it is usually convenient to group the terms associated with H_1 in Eqs. (2.2b) and (2.3b), by defining a stationary holding cost per item-day $c_h = c_r + c_i$.

For problems in which the inventory at the destination can be ignored (e.g. for the transportation of people in many cases) the average wait added to t_m should be computed for the shaded area in Fig. 2.1. The result, a value somewhere in between ½ $c_i\bar{H}$ and ½c_iH_1 , is no longer a function of H_1 alone.

If we were shipping people, c_i would represent the "value of time". When shipping freight, this constant would include the opportunity cost of the capital tied up in holding an item for one time unit. (If π denotes the "value" of an item, and i an agreed upon discount rate, then the opportunity cost is πi). For perishable items, and items exposed to loss and damage, c_i should also include any value losses arising from time spent in the system. The constant, c_i is hard to determine precisely. We don't know people's value of time accurately and, as is well known in economics, it is hard to pinpoint "i". Furthermore, in most cases even the value of the items themselves is hard to measure.

Suppose that an item costs π_0 dollars to produce but it is sold for π_1 dollars ($\pi_1 >> \pi_0$). Which of these two values should be used for inventory calculations? The answer depends on market conditions. If the demand is fixed, a reduction in inventory allows the production to be slowed (temporarily only) until the new lower inventory levels are reached (see Fig. 2.2). If the wait is reduced by Δ units, the production of $D'\Delta$ items can be avoided. The resulting one-time savings can be amortized over the life of the operation to yield a cost savings per unit time which is proportional to $D'\Delta\pi_0$. This is the same as saying that c_i is proportional to π_0 .

If on the other hand the market could absorb everything that is produced, one could then sell the extra $D'\Delta$ items in inventory while keeping the production rate constant and the amortized extra revenue per unit time would be proportional to $D'\Delta\pi_1$. This means that c_i would be proportional to π_1.

In practice, one often finds that even π_0 and π_1 are not known; this often happens when the items are components consumed within the firm as part of a multi-plant production process. Accounting systems are typically rigged to track the overall costs of production according to broad categories (e.g., labor, depreciation, etc.) but the costs are not prorated to the different components that are produced.

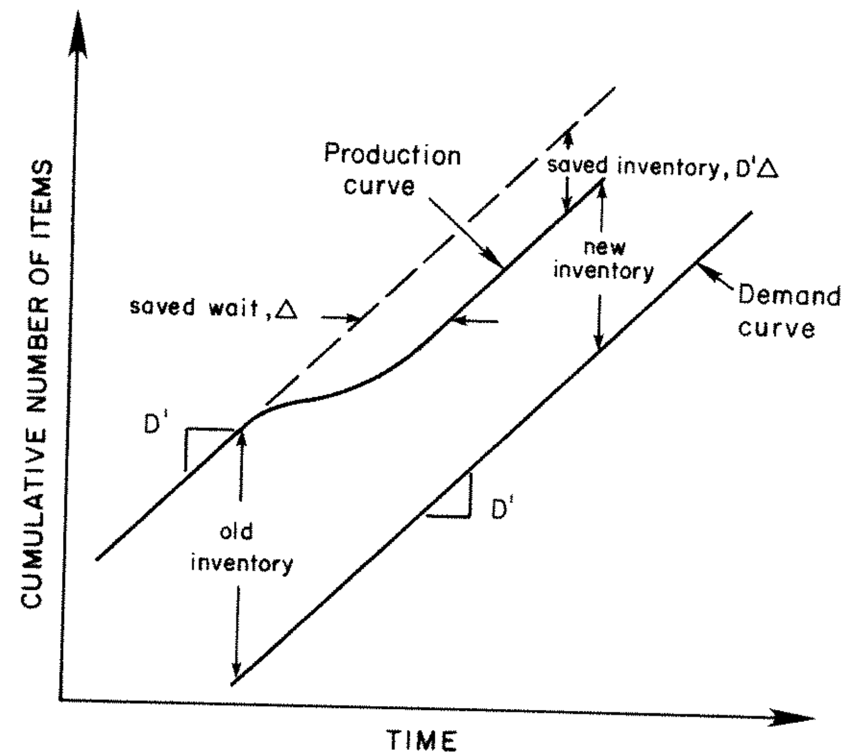

Fig. 2.2 Inventory effect of a temporary reduction in the production rate

In other cases the product can be acquired at different prices from different producers, so π_0 is not fixed. Then, the relevant price used for decision-making is not necessarily the average. For example, if a producer can secure limited supplies of items both for a low price ($\pi_0 = \pi$) and unlimited supply at a higher price ($\pi_0 = \pi' > \pi$), it will try to meet as much of its demand with the cheap items. If the demand rate comfortably exceeds the capacity of the cheap supplier, further increases in the rate would be satisfied at cost π'.Thus, this high value and not an average would be the relevant cost for an analysis of possible market expansions. Clearly, careful consideration is often necessary in determining something as seemingly basic as the "cost of goods sold".

Finally, the value of c_i that one would use in expressions such as Eqs. (2.3) should also reflect any indirect costs of delay to other aspects of the overall operation such as the effect of inventories on quality. These effects may be hard to quantify, but must be considered. Conventional wisdom indicates that large inventories lower quality because their existence reduces the incentive to eliminate defects *at the origin* – after all, items found to be defective can be replaced from the existing stock. Without this incentive, the quality of all the items (even those that are not defective) may suffer.

The value of c_i can change by many orders of magnitude, depending on what is being transported. For people c_i should be on the order of $10 per hour so that a bus load of 30 people would cost between 10^2 and 10^3 dollars per hour. A truck carrying 20,000 lbs of goods costing on the order of $1 per pound (which would be typical of groceries, machinery, etc.) would contain cargo valued at $20,000. Amortized at 10 percent for a (2,000 hour) year, the cargo costs on the order of 10^0 per hour. Cheaper and lighter cargoes can result in even lower costs. These "back-of-the-envelope" calculations illustrate that while it may be difficult to define c_i very precisely in any specific application, it should be possible to estimate its order of magnitude. Fortunately, rough estimates often are all that is needed. As we shall see, the structure of a logistic system depends on the order of magnitude of c_i, but it is not very sensitive to small changes in c_i.

Before turning our attention to motion costs, let us introduce some terminology to identify the two terms, $(c_i H_1)$ and $(c_i t_m)$, of Eq. (2.3b). The first term, which depends on the maximum dispatching headway and arises when the items are stationary, will be called the "stationary inventory cost." The other component $(c_i t_m)$, which arises while the items are moving and is independent of the dispatching headways, will be termed "pipeline inventory cost."

The following two sections discuss motion costs. Transportation costs are addressed first.

2.3 Transportation Costs

We continue with the one-origin/one-destination situation that was depicted in Fig. 2.1. If one uses a public carrier to transport the items from the origin to the destination, the total cost per year will be the sum of the costs of each individual shipment. Published rates increase roughly linearly with shipment size. (The rates increase in steps, but the overall slope is approximately constant for wide ranges of shipment sizes.) The mathematical relationship is:

$$\textit{shipment cost} \approx c_f + c_v v. \tag{2.4a}$$

where v is the shipment size, c_f is a fixed cost per shipment that should include things such as driver wages, and c_v is the rate at which the variable cost per shipment increases size, e.g. due to increased fuel consumption. The cost for shipping a sequence $\{v_i\}$ of n shipments (i = 1, ... ,n) totaling V items ($V = \Sigma_i v_i$) is thus:

$$\textit{cost for n shipments} = \sum_{i=1}^{n} c_f + c_v v_i = c_f n + c_v V. \tag{2.4b}$$

The total cost only depends on the number of shipments, *regardless of what they contain and when they happen*, and the total number of items shipped. The cost per item, thus, decreases with the average shipment size, $\bar{v}$

$$\textit{transportation cost/item} = c_f \left(\frac{n}{V}\right) + c_v = c_f \left(\frac{1}{\bar{v}}\right) + c_v. \tag{2.5a}$$

These economies of scale arise because all the items in a shipment share the fixed cost, c_f.

For our simple problem with one origin and one destination, the only decision variable appearing in Eq. (2.5a) is n (or $\bar{v}$); thus, the variable cost should not influence shipping decisions. We will not eliminate it from our expression, though, because c_v is not a constant for more complicated problems. (As we shall see, c_v depends on distance; and for problems with many origins and destinations, the distance traveled is not fixed.)

2.3.1 Relationship to Headways

Like inventory and holding costs, the cost of transportation depends on the dispatching headways. The relationship is:

$$\textit{Transportation cost/item} = c_f / \left(D'\bar{H}\right) + c_v \tag{2.5b}$$

because $V = \Sigma_i D'H_i = D'\bar{H}\, n$; i.e., $\bar{v} = D'\bar{H}$. The transportation cost decreases with the *average* headway, unlike holding costs which increased with the *maximum* headway. Notice as well that for a given number of shipments, and thus a given average headway, the transportation cost is independent of the specific headways. Hence, *shipments should be spread as regularly as practicable* to reduce the maximum headway and the associated holding cost. If headways can be maintained constant, $H_i = H$, then both holding and transportation costs are functions of H.

2.3.2 Relationship to Distance

As an aside that will become important for multiple origin and/or multiple destination logistic problems, we examine the relationship between transportation cost and distance.

Rate books reveal that c_f and c_v depend mainly on distance; the precise location or origins and destinations also influences these costs but to a lesser extent. The relationships are well approximated by linearly increasing functions of distance, d:

$$c_f = c_s + c_d d \text{ and } c_v = c'_s + c'_d d.$$

The interpretation of these four new constants appearing in the right side of these expressions is easier when the above expressions are substituted for c_f and c_v in Eq. (2.4b). The cost for n shipments totaling V items, when the origin and destination are d distance units apart, can be broken up in four terms as follows:

$$\begin{pmatrix} \text{cost for } n \\ \text{shipments} \end{pmatrix} \approx c_s n + c_d nd \; + \; c'_s V \; + \; c'_d Vd. \qquad (2.5c)$$

The first constant, c_s, is the cost attributable to each trip, regardless of distance and shipment composition; it includes the cost of stopping the vehicle and having it sit idle while it is being loaded and unloaded. Think of it as the fixed cost of *stopping* "c_s", independent of what is being loaded and unloaded. The second constant, c_d, is the cost attributable to each incremental vehicle-mile. It is the vehicle cost (including the driver) for each mile traveled regardless of the vehicle's contents; i.e., the cost of *distance*, "c_d."

The third constant, c'_s, represents the added cost of carrying an extra item. It represents a penalty for delaying the vehicle while loading and unloading the item, as well as the cost of handling the item within the vehicle. (Handling costs outside the vehicle will be considered in Section 2.4.).

The fourth constant is the cost attributable to each incremental item-mile. It can be viewed as the marginal wear and tear and operating cost *per mile* for each extra item carried. This constant, and the fourth term as a whole, should be small compared with the second term (since the cost of a vehicle-mile is relatively independent of a vehicle's contents); it will normally be ignored.

If, instead of a single destination, the vehicle carried the items picked up at the origin to several destinations, making in the process n_s delivery stops, Eq. (2.5c) would likely have to be modified slightly. Logically, rates must reflect the additional delay-cost for the extra stops. However, because not much else changes (the vehicle travels the same distance and carries the same number of items), one would expect only the first term of Eq. (2.5c) to change. Although not verified experimentally, it seems reasonable to expect it will increase proportionately to the number of stops $(1 + n_s)$. Accordingly, if we redefine c_s to be the fixed cost *per* stop, then the cost of making n shipments is

$$\begin{pmatrix} \textit{cost for n} \\ \textit{shipments} \end{pmatrix} \approx c_s (1 + n_s) n + c_d \, n \, d + c'_s \, V, \qquad (2.5d)$$

where the fourth term of Eq. (2.5c) has been neglected.

Whether Eq. (2.5d) matches actual rates when $n_s > 1$ is an open question. Multiple stops, however, are normally made as part of exclusive service agreements between shippers and carriers, which should reflect the carrier's actual operating costs; in that case, Eq. (2.5d) seems justified. That carrier cost (or the shipper cost if it uses its own vehicle fleet) is well approximated by Eq. (2.5d) should be intuitive. Drivers' wages should be proportional to the total vehicle-time for all the trips. Because vehicle depreciation cost (overhead) is proportional to fleet size, i.e., the number of vehicle-years needed per year if the demand for vehicles is not seasonal, overhead can be prorated to the tasks of a year on a total vehicle-time basis. Thus, the sum of overhead and driver wages is proportional to the total vehicle-time for the n shipments. Other vehicle operating costs should be proportional to the total number of *moving* vehicle-hours. Because both the total time and the time in motion are linear functions of the vehicle-miles traveled nd , the number of stops $n(1 + n_s)$, and the total amount of freight hauled V , the total cost should be roughly linear in these variables; i.e., Eq. (2.5d) is a good approximation for the carrier cost.

On dividing Eq. (2.5d) by V , the average cost per item is obtained:

$$cost/item \approx c_s \frac{1+n_s}{\overline{v}} + c_d \frac{d}{\overline{v}} + c'_s .$$

As a function of the average headway, the costs per item and per unit time are:

$$cost/item \approx c_s \left(\frac{1+n_s}{D'\overline{H}} \right) + c_d \left(\frac{d}{D'\overline{H}} \right) + c'_s . \qquad (2.5e)$$

$$cost/time \approx c_s \left(\frac{1+n_s}{\overline{H}} \right) + c_d \left(\frac{d}{\overline{H}} \right) + c'_s \, D'. \qquad (2.5f)$$

Although Eqs. (2.5e and 2.5f) do not show a dependence on the individual headways, we should recognize that irregular schedules may require slightly larger cost coefficients if the shipper exclusively uses its own private fleet.

This happens because the fleet size needed is dictated by the operation of the system during time periods with the largest numbers of dispatches, with the result that fleet size costs are more closely related to the minimum headway than to the average. An extensive discussion of this issue for a problem with variable demand can be found in Hurdle (1973a) and (1973b); see also Du (1993). Fleet size considerations, thus, provide a *second* incentive to keep transportation schedules as regular as possible.

Finally, note that the in-vehicle time of a typical item, t_m, is also a linear function of distance, d, and number of stops, n_s. This observation will become important later when vehicle routing is a decision variable.

2.3.3 Relationship to Size; Capacity Restrictions

Let us now return to the single origin and single destination situation of Fig. 2.1. So far, we have ignored the possibility of sending *very* large shipments; shipments that would not fit in the largest vehicles on the road. If one were to plot the cost per shipment versus shipment size for a range extending beyond this maximum, v_{max} , for a firm that owns its own vehicles, one would likely find a graph as the one shown in Fig. 2.3. Whenever the shipment size reaches and exceeds a multiple of v_{max} a new vehicle needs to be dispatched with a resulting jump in cost. The steps of Fig. 2.3 should be rather flat (with $c_v v_{max} << c_f$) since the cost of operating a vehicle is rather insensitive to what it contains. Whether or not it is exactly as shown in Fig. 2.3, the transportation cost per shipment function, $f_t(v)$, should be "subadditive;" i.e., it must satisfy: $f_t(x_1 + x_2) \leq f_t(x_1) + f_t(x_2)$ for any x_1, $x_2 \geq 0$. This property is to be expected because one should not be able to reduce the cost of a shipment by shipping it in parts (see Problem 2.3).

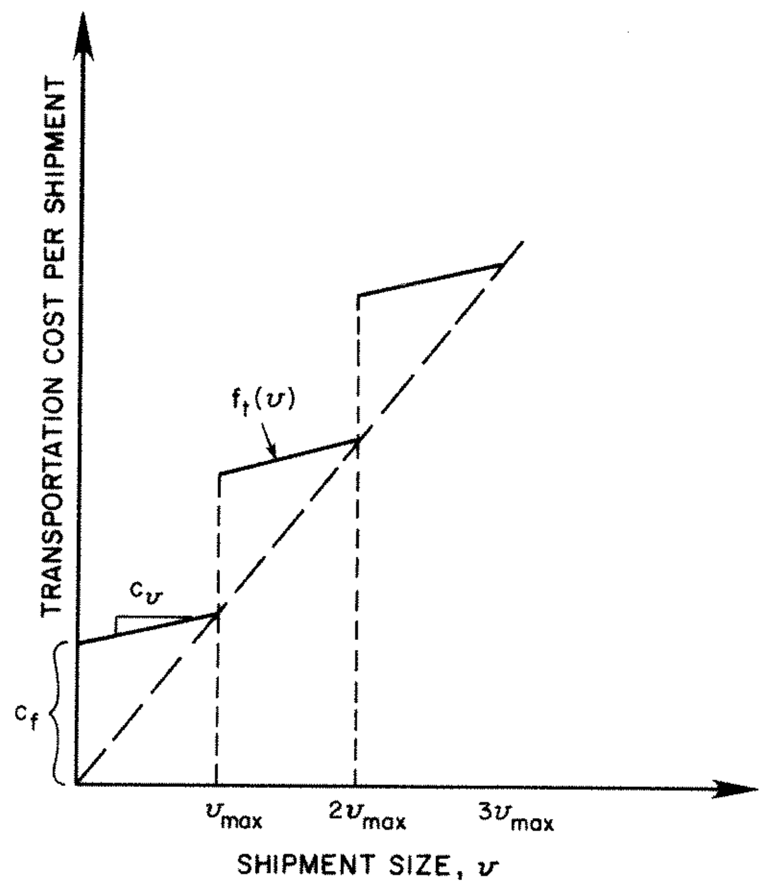

Fig. 2.3 Relationship between transportation cost per shipment and shipment size

For most problems, though, one only needs to consider the linear part of f_t between 0 and v_{max}, as shipments larger than v_{max} are not economical. This can be easily seen if handling costs can be ignored (e.g., if the handling cost per item is a constant, independent of shipment size) by examining the sum of the average *holding* and *motion* costs per item as a function of shipment size. Figure 2.4 plots the average transportation cost per item as would be obtained from Fig. 2.3. The figure also plots *the negative* of the holding costs as a function of shipment size. (We are assuming here that headways are regular, $\bar{H} = H_1 = H$; and we are using Eqs. (2.2b) and (2.3b) with $H_1 = H = v/D'$. Recall that c_h is the stationary holding cost per item-day, $c_h = c_i + c_r$).

The optimal shipment size is the value of v for which the vertical separation between the two curves of Fig. 2.4 is minimum. Clearly, the point can be identified by sliding the "waiting" curve upwards until it first touches the transportation curve. This can only happen either at point P of the figure (where $v = v_{max}$), or else at a point $v < v_{max}$, if the line is sufficiently steep. For most problems, thus, one can ignore the behavior of the transportation curve for $v > v_{max}$, if one remembers to abide by the constraint: $v \leq v_{max}$.

Analytically, the optimal shipment size of Fig. 2.4 is the solution of the following problem:

$$(EOQ):\ \min\left\{ Av + \frac{B}{v} \right\}\ s.t.:\ v \leq v_{max}\ .$$

where

$$A = c_h/D', and\ B = c_f\ .$$

This is the well known "lot size" or "economic order quantity (EOQ)" model of the inventory control literature (Welch, 1956; Arrow et al., 1958) whose roots can be traced to the pioneering work of F.W. Harris in the early part of this century (Harris, 1913a and 1913b). Erlenkotter (1990) describes these works in a historical context.

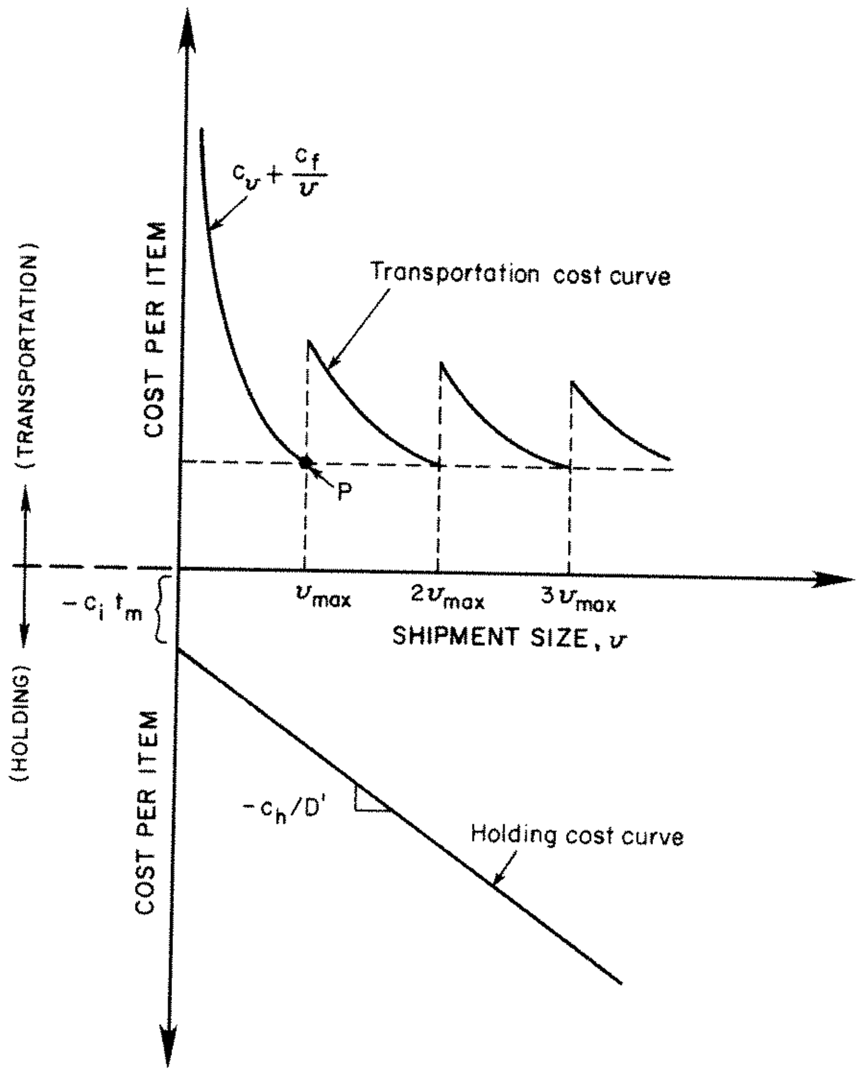

Fig. 2.4 Transportation and holding cost (per item) as functions of shipment size

2.3.4 R

We have
with size
of shipme
varies by
modes.
While
shipment
shows thr
size; as it
higher fix
fixed pipe
for valuab
Fig. 2.
thing by t
If, as sho
lower env
the specif
and conca
duced fur
optimal.
If the i
they exhib
optimal,
breaking
ample, if
0) and (c_f
would be
mal. The
first mode
0.165). If
the modal
curve can
2.4).
If the sl
represent
most ecor
such a ch
shippers d
type is no
be used to

2.3.4 Relationship to Size: Multiple Transportation Modes

We have already seen that shipment cost increases approximately linearly with size (Eq. 2.4), and that this is likely to be true for fairly broad ranges of shipment sizes. This qualification was made because if shipment size varies by a large amount, it may be cost effective to change transportation modes.

While some shipping modes, such as mail, exhibit a low fixed cost per shipment and a high cost per item, others may be the opposite. Fig. 2.5 shows three such curves. Note that the best mode depends on the shipment size; as it grows, one tends to favor the modes with lower variable cost and higher fixed cost. (In comparing modes, the vehicle cost should include the fixed pipeline inventory cost per item, $c_i t_m$; faster modes may be preferred for valuable items.)

Fig. 2.5 displays the transportation cost that results if one ships everything by the cheapest mode – the lower envelope of the three cost curves. If, as shown, cost increases at a decreasing rate for each mode, then the lower envelope also increases at a decreasing rate. Like the cost curves for the specific modes, the shipment cost by the best mode is then increasing and concave, and therefore subadditive; this shows that cost cannot be reduced further by breaking the shipment into parts. The lower envelope is optimal.

If the individual modal component curves are merely subadditive, e.g., they exhibit jumps as in Fig. 2.3, then the lower envelope is not necessarily optimal, or subadditive. In this case, costs can sometimes be reduced by breaking a shipment into parts and sending it by different modes. For example, if the cost parameters of two modes with $v_{max} = 1$ were ($c_f = 1$, $c_v = 0$) and ($c_f = 0$, $c_v = 1.5$), then the single-mode shipment cost for $v = 1.1$ would be either 2 or 1.65; i.e., 1.65 by the best mode. But this is not optimal. The optimum is achieved by sending a one-unit shipment with the first mode (cost = 1) and the remainder with the second mode (cost = 0.165). If shipments can be allocated to the modes in an optimal way and the modal cost curves are subadditive and increasing, then the overall cost curve can itself be shown to be subadditive and increasing (see problem 2.4).

If the shipper operates its own vehicle fleet, the curves of Fig. 2.5 could represent different vehicle types, and the figure would then indicate the most economical vehicle type for the particular shipment size. Because such a choice is not as flexible as a choice of public carriers (i.e. modes), shippers do not change the vehicle fleet often. When the choice of vehicle type is not an issue, then the appropriate (linear) component curve should be used to evaluate transportation cost.

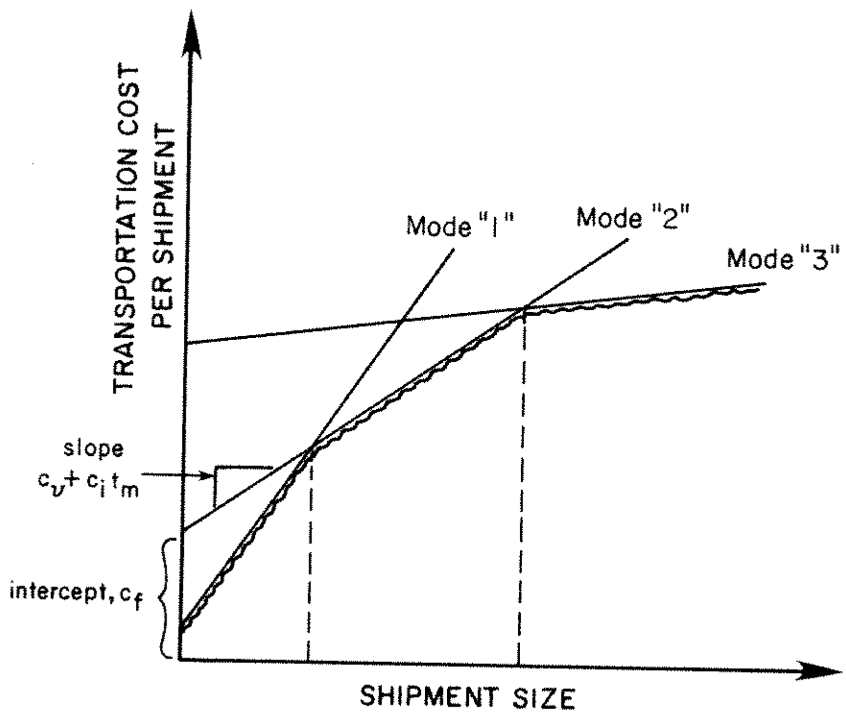

Fig. 2.5 Relationship between shipment cost and size for various transportation modes

2.4 Handling Costs

Handling costs include loading individual items onto a "container", moving the container to the transportation vehicle threshold, and reversing these operations at the destination. The container can be a *box* or a *pallet*, or if the items are large enough, nothing at all. We examine here the cost of handling a shipment of size, v .

If the items are handled individually, the handling cost per shipment should be proportional to v , so that

$$handling\ cost \approx c'_s\, v.$$

If the items are small, it is not economical to move them individually; instead they can be moved on "handling vehicles" such as pallets. Clearly, the handling cost should have a similar form as the transportation function, since items are being transported within a compound. If the batch is smaller than one pallet the cost of handling it should therefore be:

$$\textit{handling cost/batch} \approx c'_f + c'_v\, v. \tag{2.6}$$

The constant c'_f represents the (fixed) cost of moving the pallet regardless of what it contains, including the forklift driver's wages, plus the forklift's depreciation and operating cost. The constant c'_v captures the cost, accounting for both labor and capital, of loading one item on the pallet. If v is larger than the maximum number of items that fit on a pallet, v'_{max}, then the handling cost function per shipment, $f_h(v)$ will still be a scaled down version of the transportation function, as in Fig. 2.6.

At the destination, the handling cost function will be analogous, possibly with different c_f' and c_v' but the same v'_{max}. As a result, the combined handling cost for the shipment at both ends of the trip should still have the form of Fig. 2.6, and should obey Eq. (2.6) if $v \le v'_{max}$.[2]

One could compare the cost of moving items individually and moving them in pallets. But if more than one item fits on a pallet, it will usually be cheaper to move them in pallets.

---

[2] Although we have used the words pallet and forklift repeatedly, we stress here that Eq. (2.6) also applies to other container-filling methodologies that do not use forklifts; e.g., to the "bucket-brigade" method of order-picking using passive conveyors described in Bartholdi and Eisenstein (1996). In these cases, one just needs to make sure that the constants c'_f and c'_v are representative of the actual operation.

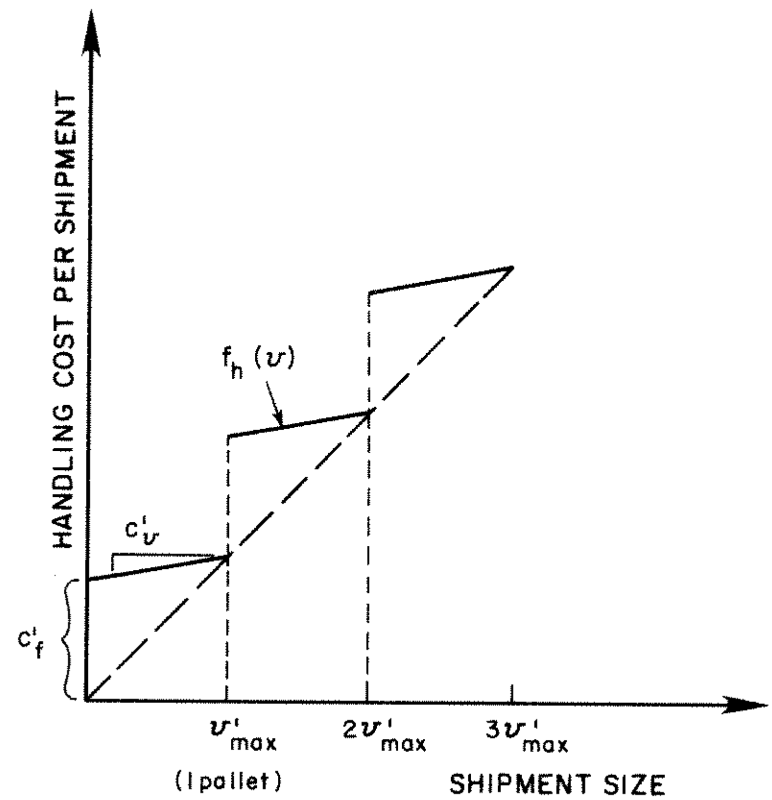

Fig. 2.6 Handling cost per shipment as a function of shipment size

2.4.1 Motion cost

Figure 2.7 depicts the sum of transportation plus handling costs for $v'_{max} << v_{max}$. The function, $f_m = f_t + f_h$, is still subadditive and increasing. (See problem 2.4.)

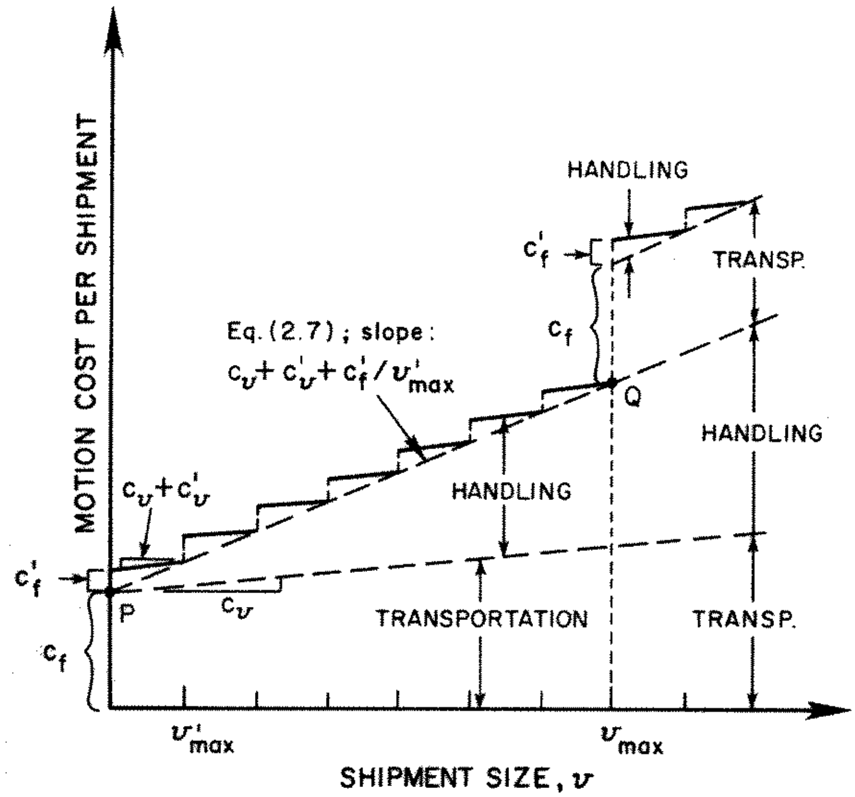

Fig. 2.7 Relationship between shipment size and the combined cost per shipment of transportation and handling

Note that to within an error of c'_f, the motion cost per shipment, $f_m(v)$, can be approximated by line $\overline{PQ}$ of the figure, which is a lower bound:

$$f_m(v) \approx c_f + \left(c_v + c'_v + \frac{c'_f}{v'_{\max}} \right) v . \tag{2.7}$$

This indicates that handling costs can be subsumed in the transportation cost function, Eq. (2.4a), with a suitable definition for the fixed and variable cost: [3]

[3] If $v < v'_{\max}$ it is better to use: $c''_f = c_f + c'_f$ and $c''_v = c_v + c'_v$

$$c''_f = c_f \ \ and \ \ c''_v = c_v + c'_v + \frac{c'_f}{v'_{max}}.$$

The expression for the variable motion cost per item, c''_v, is intuitive; in addition to the variable transportation and handling costs per item, c_v and c'_v, it includes each item's prorated share of the fixed cost per pallet, c'_f / v'_{max}.

2.4.2 The Lot Size Trade-Off with Handling Costs

If we prorate the cost of a shipment to the items that it contains, we can construct a figure, analogous to Fig. 2.4, which can be used to determine the optimal size of the shipments. Figure 2.8 is not extended beyond v_{max}, since larger shipments continue to be undesirable. Note from the figure that if the waiting cost curve is pushed upwards, the first point of contact is either $v < v'_{max}$ (if the waiting cost curve is very steep), or else it is likely to be an integer multiple of v'_{max}. (This is not always the case, but very little is lost by assuming that it is – Daganzo and Newell, 1987). Because the lower bound from Eq. 2.7 is exact when v is an integer multiple of v'_{max}, one could use it instead of the exact (scalloped) curve while restricting v to be a multiple of v'_{max}. Except for the variable cost coefficient, c''_v, this equation matches Eq. (2.4a), and we saw already that variable costs do not influence the optimal shipment size. Thus, if shipment size is restricted to be an integer multiple of v'_{max}, *the optimal shipment size is independent of handling costs.*

We now examine the consequences of relaxing this restriction. If the optimal shipment size, v^*, is greater than one pallet, we see from Fig. 2.8 that allowing v to differ from a multiple of a pallet cannot improve things appreciably. In the most favorable case the cost savings can be shown to be about one tenth of c_f/v'_{max}, with much smaller savings in other cases; see problem 2.5. Thus, even without the restriction, one can safely ignore handling costs in determining shipment size.

If, on the other hand, v^* is smaller than one pallet, then handling costs should be considered; there may be a significant difference between $f_m(v)$ and its lower bound (see Fig. 2.8 and the previous footnote). If $c'_f >> c_f$, then the optimal shipment size may be noticeably larger than if handling costs had been ignored.

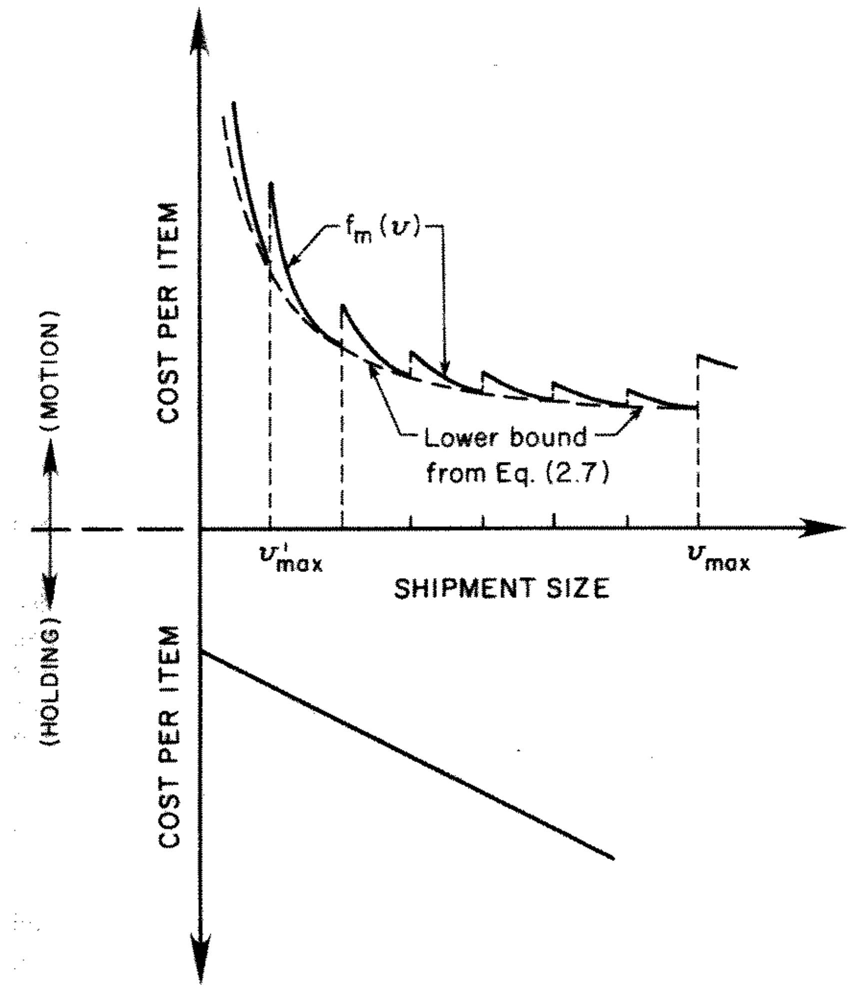

Fig. 2.8 Motion and holding costs (per item) as functions of shipment size

In summary, the following simple recipe can be used: If economic shipment sizes are likely to be larger than a pallet, ignore handling costs in the decision; but if shipment sizes are smaller than a pallet, then include the fixed cost of handling a pallet as part of the fixed cost per shipment and select the shipment size which is the minimum of problem "EOQ" with $A = c_h/D'$ and $B = c''_f$.

More complicated motion curves would arise if items had to be put into boxes, which could be put onto pallets, which would then travel on trucks. Because the relationship of boxes to pallets is analogous to the relationship between pallets and trucks, the additional handling step would be reflected by a second set of scallops on Fig. 2.8. The selection of an optimal shipment size would be affected by this second set of scales in a similar way: *the cost of moving and filling boxes can be ignored if the optimal shipment size is larger than a box*; otherwise, the fixed cost per shipment, c''_f , should include the fixed cost of moving one box including opening and closing it, but not the cost of filling it.

With a properly defined c''_f, the optimal shipment size should still follow from the solution of the "EOQ" problem:

$$(EOQ)\min\left\{ Av+\frac{B}{v}+C \right\} ;\, v \le v_{\max} \ , \qquad (2.8a)$$

where:

$$A = c_h/D' \ , \ B = c''_f \ , and \ \ C = c_i\, t_m + c''_v \ . \qquad (2.8b)$$

Note that c''_v should include any handling costs (per item) not included in c''_f, and that the minimum of Eq. (2.8a) is unaffected by c''_v since C is an additive constant in Eq. (2.8a). Also remember that if the minimum of (2.8a) is greater than one box (or pallet) the shipment size should then be rounded to the nearest box (or pallet); the cost, however, remains close to the minimum of (2.8a) without rounding.

2.5 Stochastic Effects

We have assumed in our discussion of cost that the transportation travel time and the production and consumption rates are constant. These assumptions can be violated in two ways. The production and demand rates (and the travel time perhaps as well) may vary over time in a predictable manner, and also unpredictably. Predictable variations such as seasonal trends and day of the week effects will be examined in Chapter 3; optimal decisions can be found because costs can be predicted.

Unpredictable variations are another matter and are examined here; they require additional inventories, and may also increase transportation cost. Continuing with the single origin and single destination model, for the rest of this section we assume that production is driven by consumption. That is, the destination requests deliveries so that its inventory level can sustain at all times the demand that is anticipated. With inherently unpredictable demand and travel times, however, it is no longer possible to time the shipments so they arrive just as the stock at the destination is running out, as with the first two shipments of Figure 2.1. Stochastic variations are the subject of much attention in the inventory control literature, where the objective is to determine optional levels of "safety stock" and reorder "trigger points" (see Peterson and Silver (1979) or Zipkin (2000) for example).

These stochastic phenomena complicate matters, but in many cases the *added* holding plus motion costs (per item) that arise due to randomness can be shown to be known linear functions of either v or 1/v, and in other cases completely independent of v . This is fortunate because the added costs can then be captured by a deterministic EOQ model, Eq. (2.8a), where some of the constants ("A", "B", and "C") have been increased.

2.5.1 Stochastic Effects Using Public Carriers

Newell, in some unpublished notes, has pointed out that if transportation is reliable enough to ensure that shipments arrive at the destination in the order in which they were requested, then the added cost due to randomness is constant. As a result, the demand and travel time uncertainty should influence neither the frequency of dispatching nor the average lot size.

A common ordering strategy uses a trigger point v_0 as follows: whenever the inventory on hand *plus* the number of items on back order equals v_0 , a shipment of size v is requested.[4] The reorder headways for this strategy vary because the demand varies, but the shipment sizes remain constant.

Let us assume that the demand arrival process can be approximated by a diffusion process with rate D' (items per unit time) and index of dispersion γ (items).[5] The index of dispersion represents the variance to mean ratio of the number of items to have arrived in one time unit. (Note that if items are measured by a physical quantity such as tons, cubic feet, etc., γ shares these units.) A suitable choice of γ approximates most of the processes examined in the inventory literature. Let us also assume that the lead time, T_ℓ, (the time between order placement and receiving) has mean t_ℓ and standard deviation σ_ℓ. (The lead time should be close to the average transportation time, $t_{m'}$, if the origin can keep up with the requests; but this assumption is not needed here.)

[4] These strategies are called "(s,S)" in the inventory literature; see e.g., Peterson and Silver (1979); Zipkin (2000).

[5] For a diffusion process, the number of arrivals in any time interval is a normal random variable, with mean and variance proportional to the duration of the interval, and independent of the arrivals in non-overlapping intervals. Newell (1982) proposes to approximate queuing and inventory phenomena with diffusion processes.

If one desires to avoid stock-outs, the trigger point, v_0 , should be large enough to ensure that no stock-out occurs immediately before the arrival of an order. The best way of exercising this policy can be found with the help of the three curves in Fig. 2.9, relating time to the cumulative number of items that have been: (i) ordered, (ii) received, and (iii) consumed at the destination. The dashed lines in the figure represent the portion of the curves that is not yet known at time "NOW". A request for a shipment is depicted immediately after time "NOW" since at that time the sum of the inventory on hand and the back orders is shown to be v_0 . Because all the back orders are sure to have arrived before the new order, it is clear from the figure that a stock-out will be averted immediately before the new order arrives if the future consumption until the new order arrives (segment $\overline{PQ}$ in Fig. 2.9) does not exceed the inventory currently on hand plus the back orders, v_0 .

This condition can be expressed probabilistically if we recognize that, conditional on the lead time, T_ℓ , $\overline{PQ}$ is normal with mean $D'T_\ell$ and variance $D'\gamma T_\ell$. The unconditional first two moments of $\overline{PQ}$ are thus:

$$E\left(\overline{PQ}\right) = D' t_\ell$$
$$\text{var}\left(\overline{PQ}\right) = D'^2 \sigma_\ell^2 + D'\gamma t_\ell .$$

If the trigger point, v_0 , is chosen several standard deviations greater than $D't_\ell$, stock-outs will be rare. The precise value of v_0 is not important for our analysis (it is a function of D', γ, t_ℓ, σ_ℓ , and nothing else); what is important is that, as the figure clearly indicates, the contribution of v toward the maximum and average accumulation is insensitive to v_0 . This would, in fact, be the case even if v_0 were chosen in a more involved manner (e.g. recognizing the distribution of T_ℓ). Existing methods for selecting trigger points and shipment sizes (Peterson and Silver, 1977, Zipkin, 2000) exploit this insensitivity.

In order to choose the optimal v , the motion and inventory costs must be balanced, as shown in prior sections. In the long run, the motion costs with and without stochastic effects are the same because the same number of shipments are sent in both cases (D'/v shipments per unit time), but the holding costs are larger with stochastic phenomena. The maximum number of items present at the destination will certainly occur after the arrival of an order. As shown in the figure, for a typical order, this number is:

$$v_0 + v - \overline{PQ},$$

which is largest when $\overline{PQ}$ is as small as possible. The term (v_0 - $\overline{PQ}$) represents the contribution of randomness toward higher inventories; but the term is not dependent on our decision variable, v . Thus, except for an additive constant, the holding costs are as in the deterministic case; the optimal shipment size remains the same.

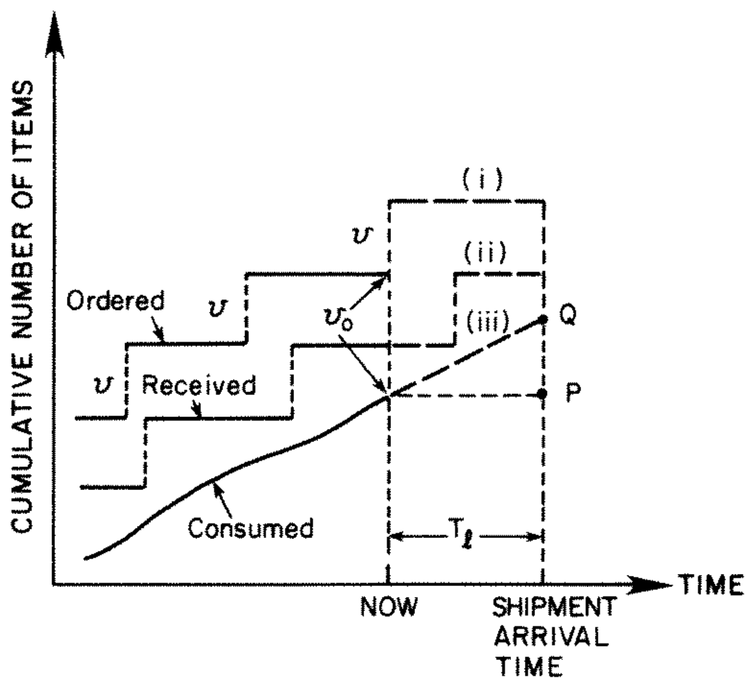

Fig. 2.9 Evolution over time of the cumulative number of items ordered, received and consumed for a simple trigger point strategy

For clarity, the inventory at the origin was ignored in the foregoing discussion. Yet, the irregular way in which orders are placed will undoubtedly raise inventory and production costs at the origin. These effects, however, are shown below to be largely independent of v (their actual magnitude depends on how frequently the production is adjusted) and, thus, should not influence shipping decisions.

If, as is usual, there is an incentive to maintain a steady production rate, then one would set it at a value D'_p , slightly greater than D' to ensure that the overall demand can be met in the long run. Although inventories at the origin would then tend to grow with time, every once in a while (every many reorders, presumably) the production process could be interrupted for a while to allow the demand to catch up with the cumulative number of items produced. The frequency of these stoppages would depend on production and inventory cost considerations.

A simple strategy would stop production whenever the inventory at the origin (after a shipment) reaches a critical value, v_1 , and would resume it (also after a shipment) when the inventory dips below another value, v_2; see Fig. 2.10. The maximum inventory is therefore: $v_1 + v$, and the average inventory: $1/2(v_1 + v_2 + v)$. The cost of production should be a function only of D'_p and the duration of the on and off periods. The on and off periods, however, only depend on v_1 , v_2 and on the statistical properties of the smooth curve tangent to the crests of the orders sent curve (see Fig. 2.10). On a scale large compared with v , this curve shares the statistical properties of the demand curve *which do not depend on v*. Therefore, the optimal production decisions (i.e., the choices of v_1, v_2 , and D'_p) do not depend on v.

As before, the inventory (maximum and average) can be decomposed into a portion that is proportional to v (represented by the shaded area in Fig. 2.10) and independent of the production strategy, and a remaining portion which is influenced by the production scheme and is independent of v . Thus, the extra production and inventory costs arising at both the origin and the destination due to the unpredictability of demand are largely independent of v . They can be ignored when determining the optimal shipment size.

The foregoing discussion is not an exception; stochastic effects can be captured within the scope of a deterministic EOQ model in other situations as well. Problem 2.6 discusses the use of a private vehicle fleet, and the following subsection considers an operation where two different transportation modes are used.

2.5.2 Stochastic Effects Using Two Shipping Modes

It has been assumed so far that stock-outs are avoided by holding inventories large enough to absorb fluctuations in demand and in the transportation lead time. In some instances, if a second, much more expensive, shipping mode is available for expediting shipments, the total costs may be reduced by expediting small shipments at critical times. In these instances the optimal lot size v is also the result of an EOQ trade-off, although the trigger point decision is no longer independent of the shipment size decision. For the following discussion it is assumed that the expedite mode is so fast that its lead time can be ignored.

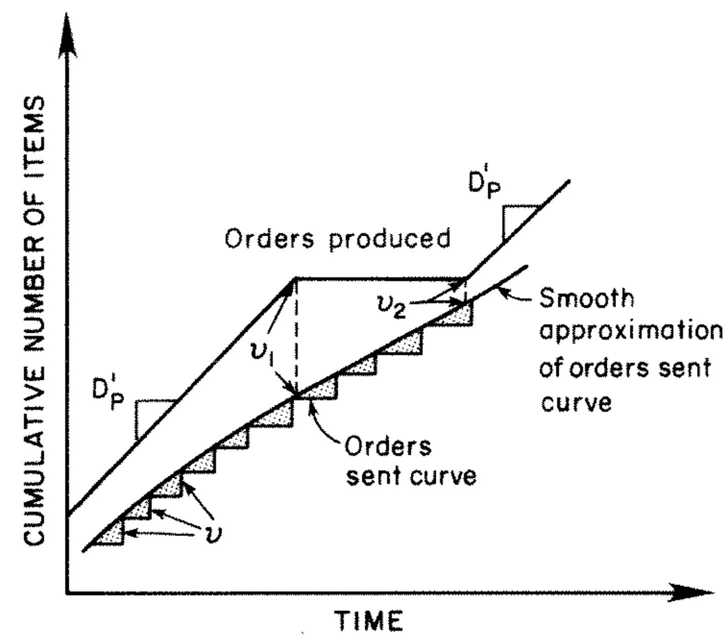

Fig. 2.10 Inventory effect of production and transportation decisions

Most of the time the expedite mode lies in wait, and the system operates as if the primary mode was the only mode (see Fig. 2.9). The trigger point v_0, however, does not have to be chosen as conservatively as before, because when a stock-out is imminent enough items can be sent by the premium mode to avoid it.

The analysis is simple. If, as is commonly the case, the time between reorders is large compared with the primary mode's lead time (i.e., so that when the trigger point, v_0, is reached there aren't any unfilled orders) then the probability that some items have to be expedited in the time between ordering and receiving a lot (of size v) does not depend on v . It is a decreasing function of v_0 , approaching zero when $(v_0 - E(\overline{PQ}))^2 >> \text{var}(\overline{PQ})$.

The exact form of the expected amount expedited per *regular* shipment will depend on the strategy used for choosing the expedited lot sizes. (Although these could be fixed, if possible they should be chosen just large enough to meet demand until the regular order arrives). In any case, the expected amount expedited per regular shipment will also be a decreasing function of v_0 , $f(v_0)$. Assuming that the cost per item expedited is a constant, c_e , we find that the expected expediting cost per regular shipment is: $c_e f(v_0)$. The moving cost per item is as a result:

$$\begin{pmatrix} \textit{moving cost} \\ \textit{per item} \end{pmatrix} = \frac{c_f + c_v v + c_e f(v_0)}{v + f(v_0)} = \frac{c_f + (c_e - c_v) f(v_0)}{v + f(v_0)} + c_v \,.$$

The maximum inventory still occurs when $\overline{PQ}$ is as small as possible, and remains: $v + v_0 - \overline{PQ}$; the total cost per item is thus:

$$\textit{cost/item} = c_v + \frac{c_f + (c_e - c_v) f(v_0)}{v + f(v_0)} + c_h \left(v + v_0 - \overline{PQ}\right)/D' \,.$$

For a *given* v_0 , if we think of the expected amount shipped by both modes with every regular shipment, $v' = [v + f(v_0)]$ as the "lot size," the equation is still of the EOQ form (2.8a), where the fixed moving cost has been increased to include the expected cost of expediting, $(c_e - c_v) f(v_0)$:

$$\textit{cost/item} = \left\{ c_v + c_h \left[v_0 - \overline{PQ} - f(v_0)\right]/D' \right\} + \frac{c_f + (c_e - c_v) f(v_0)}{v'} + c_h \, v'/D' \,.$$

Unlike in the previous case, though, the trigger point v_0 should not be chosen independently of v. If v is large so that shipments are infrequent, expediting a significant amount of freight with the average shipment only increases the moving costs marginally. But if v is small, the penalty for expediting is paid more often; it may be more efficient to increase v_0.