Measurement of Congestion in Transportation Systems[1]

LOWDON WINGO, JR., Resources for the Future, Inc.

FRAMEWORK OF ANALYSIS

> "Under present conditions congestion is the main obstacle to more traffic. While traffic is not an end in itself, cheap transportation is a contributing factor to the economic division of labor. Ready accessibility of centrally located markets is of particular importance, because it determines the size of markets and the extent to which economies of large-scale production and distribution can be reaped. Congestion, by setting a limit and a premium upon the movement of persons in commodities, restricts the effectiveness with which the functions of centrally located markets can be performed."[2]

● THE EXTENSIVE NETWORK of interaction among urban economic activities depends in large part on the organization and technology of transportation integrating the urban structure. Changes is the conditions of transportation will be reflected in the organization of economic activity throughout the urban space, for these changes make themselves felt as changes in the cost conditions of production and distribution. Where the adjustment of economic activity to the organization of the transportation system is good, one can talk in terms of an efficient organization of the economy and its technology of movement. Where the adjustment is poor, this interaction takes place only under conditions of high and increasing costs. One of the characteristics of poor adjustment has frequently been identified as congestion: the greater the congestion the more inefficient the interaction. Hence, the identification of the congestion element in transportation systems in a matter of interest not only to those who have a responsibility for building and managing a transportation system, such as highway and traffic engineers, but to all of those who have a concern with the economic efficiency of the urban complex. Among these are the students of urban structure and land use, for the external diseconomies associated with transportation systems may well have demonstrable consequences for the structure of land values, for factor payments to labor, as well as for the transfer costs entering into the costs of production. It is in this sense that the phenomenon of congestion in transportation systems commands the attention of the urban economist.

What is congestion? A distinguished mathematical economist, Martin Beckmann, who has analyzed highway traffic with tools of queuing theory and probability models, offers a simple definition: "By congestion we mean here traffic conditions which occur at flows that substantially reduce average speed on the road."[3]

Congestion is a relative phenomenon. It can exist for two vehicles as well as for 20,000; it may last for seconds or for hours. The driver waiting first in line at a toll gate is experiencing the same kind of system phenomenon as if he were the twentieth in line. In short, congestion can exist whenever two or more vehicles seek to occupy the same space simultaneously. In this paper, congestion is defined to be any situa-

---

[1] The theoretical formulation of the ingression function in this paper is a by-product of an investigation of characteristics of the urban space-economy carried on by the author as an endeavor of the Program of Regional Studies at Resources for the Future, Inc. The contributions of other staff members are gratefully acknowledged.

[2] From Beckmann, Martin, McGuire, C. B., and Winsten, C. B., "Studies in the Economics of Transportation." Yale University Press, New Haven, p. 95 (1956).

[3] ibid., p. 49

tion between two or more vehicles such that one or more experience a time loss due to the behavior of the others.

The nature of this time loss has some inherent complexities. All time spent in transportation involves time loss; hence, the time loss associated with congestion is a surplus over and above that which would be lost under "normal" circumstances; that is, traveling the same route in the absence of other vehicles. Thus, the congestion time loss has a highly subjective element, since it is related to the propensities of the individual road user. In this sense congestion losses become extremely amorphous when viewed from the standpoint of the road user.

As against a framework which looks narrowly at the transportation system as a unit with certain time-loss characteristics to be evaluated, a different framework which views the transportation system as an abstract physical system with economic characteristics may offer an objective basis for analysis. Although this paper will discuss transportation systems within this special kind of framework, the conclusions will have relevance in a more general sense, and the technique of analysis a general applicability to many of the operating characteristics of transportation systems.

This paper uses the device of constructing an abstract physical model embodying those characteristics which seem relevant to the problem as a whole. The behavior of such a model should identify certain features of the abstract system which are general for all real cases having the same basic characteristics as the abstract system. The objective is to identify congestion characteristics which are innate in the system and which do not emerge because of the peculiarities of the specific case.

That movement element which in terms of order and relative volume is the most significant in any urban movement system—the journey-to-work—will be abstracted. The volume and distribution of this element is a function of the locally employed labor force, the spatial arrangement of employment centers, and the time-distribution of its working periods. Where the employed labor force is a greater proportion of the population in one case than another, "caeteris paribus," the volume of this movement component will tend to be proportionately greater; likewise, where a locality operates on a single working shift, or where the places of employment are spatially concentrated, relative volumes will tend to be greater than where employment is organized on multiple shifts, where work periods are substantially staggered, or where employment centers are relatively dispersed. Thus, the variables involved in the demand for movement with respect to the journey-to-work are essentially variables arising from the economic organization of production in the community.

There is one important distinction. The volume of movement in this parlance is a quantitative measure based on the "trip" as the basic unit. Movement, for the present purpose, has no spatial significance; a trip which is a block long has the same value as one which is 10 miles long. The volume of transportation relates the volume of movement to a spatial framework. In its most general form a movement system has the following characteristics:

1. It involves a set of spatially distributed origins (analogous to the domiciles of the employed population);
2. It involves an aggregative destination, or assembly point (as in the places of employment of the employed population);
3. It involves a set of paths having specific capacity characteristics, connecting the origin and assembly points, and a set of carriers. (Here a distinction might be made between simple and complex paths, the former being a path on which the carriers cannot change order, as in the case of a single track rail line. Simple and complex carriers may be distinguished in that a simple carrier is associated with a single, discrete origin-destination pair of points. Under the assumption that the system as a whole operates at capacity velocity, complex paths have the same characteristics as simple paths and complex carriers are ruled out.) With respect to carriers and paths a system may be homogeneous, wherein all carriers and paths would have uniform movement characteristics, or heterogeneous, in which case there would be two or more subsystems (each homogeneous within itself) each serving the same movement demand, but each of which has different movement characteristics in terms of paths and movements.

4. Finally, the movement system is rigorously time-ordered in the following respects:
   a. There is a highly ordered schedule of assembly periods which gives the movement system the characteristics of a cyclical character; and
   b. Each cycle is time-constrained such that all units are required to arrive not later than the scheduled time or deadline, and all are released simultaneously in the ebb movement.
5. Time is valuable to all units.

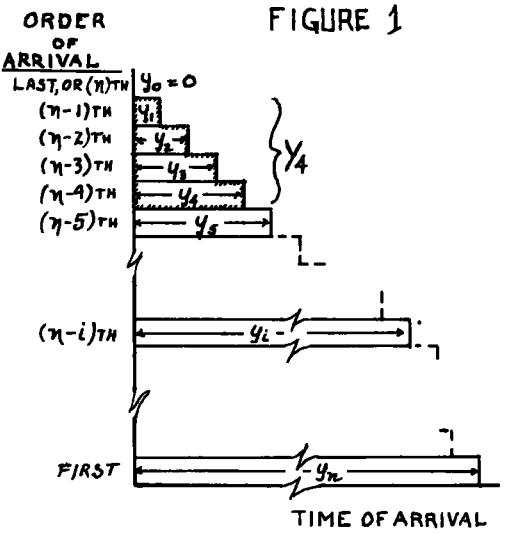

FIGURE 1

Figure 1.

With these few characteristics some of the relevant features of a highly-ordered, high-volume, simple homogeneous system can be appraised. Are there features of a movement system abstracted from spatial considerations which have implications for the analysis of congestion? There are, in fact, and these implications are associated with a phenomenon which has been called the ingression factor.

NATURE OF INGRESSION LOSSES

Ingression is a condition that exists in movement systems when the instantaneous demand on the system exceeds its instantaneous capacity; that is, when the number of units required to arrive at a given point not later than a given time is greater than the ability of the system to admit said units simultaneously. Ingression losses, then, are time losses imposed on units in the system by the inability of the system to handle all demanding units at the same time.

To illustrate how these losses take place, Figure 1 is based on a simple system (with an instantaneous capacity of 1). In this case, there are n units all of which must arrive at the assembly point not later than the deadline of T o'clock. If the last (or nth) arrival is to meet the deadline, then the next-to-last or (n - 1)th must arrive slightly before it. If the capacity of the system is represented as C units per hour, the (n - 1)th unit must arrive 1/C hours befor the nth unit. Thus in Figure 1 the ingression loss experienced by the (n - 1)th unit is measured by $y_1 = 1/C$; that of the (n - 2)th unit by $y_2 = 2/C$; and that of the first unit by $y_n = n/C$. The total ingression loss of the last five units can then be expressed as the sum of the ingression losses for each unit, or

$$Y_i = \Sigma y_i, \quad (i = 0, 1, 2, 3, 4)$$

$$= \frac{0}{C} + \frac{1}{C} + \frac{2}{C} + \frac{3}{C} + \frac{4}{C} = \frac{10}{C} .$$

Letting N stand for the number of units making demands on the system the following can be generalized (see mathematical Appendix B-1):

$$Y = \frac{N(N-1)}{2C} ,$$

in which Y = total ingression loss for the inflow[4] phase of the cycle.

[4]By reversing the demonstration, it can be shown that the ingression loss of the outflow phase is equal to that of the inflow phase. Hence, the term "phase ingression loss" refers to the loss associated with a single phase, whether it be of an inflow or outflow character. "Cycle ingression loss" refers to the total loss of an inflow and an outflow movement.

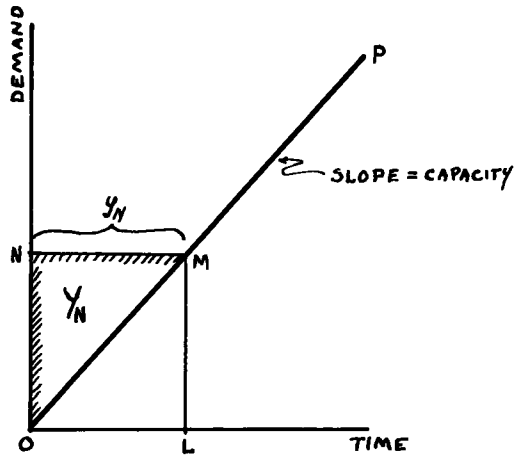

Figure 2.

Where Figure 1 relates the order of arrival to a time scale, in Figure 2 the total deadline demand on the system n is related to a time scale. Here the slope of the line OP represents the capacity of the system. Thus, for N units of demand on the system the first unit must arrive OL time units before the last if the last is to satisfy the time restraint. The total phase ingression loss for the system, then, is measured by the area of the figure OMN. It can be shown accordingly that for any large value of N, the total phase ingression loss Y is approximately equal to $N^2/2C$. In short, as demand increases in the system the total phase ingression loss tends to increase by the square of the demand (see Fig. 3).

Another important feature emerges in the relationship between the marginal ingression loss Y'_N and the average ingression loss $\bar{y}$. If random entry into the system as far as order is concerned can be assumed, over a large number of cycles the average loss to any one unit will tend to approximate the average loss in the system. The marginal phase ingression loss is the change in the total phase ingression loss effected by the entry of the Nth unit into the system. Mathematically, it can be defined for large values of N and C approximately as follows:

$$Y'_N = \frac{dY}{dN} = \frac{N}{C}$$

the average phase ingression loss can be defined as

$$\bar{y} = \frac{Y}{N} = \frac{N}{2C}$$

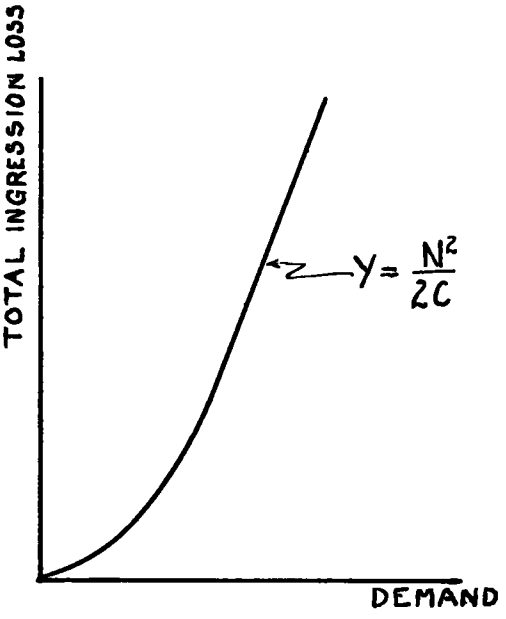

Figure 3.

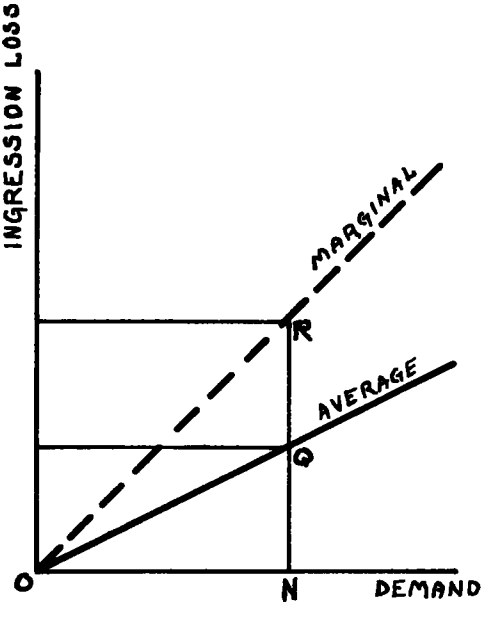

Figure 4.

In Figure 4, the marginal and average phase ingression functions are plotted in terms of total demand N. The entry of the Nth unit into the system, hence, imposes an additional phase ingression loss on the system equivalent to NR, of which the average phase loss QN tends to be borne by the Nth unit over a large number of cycles, while the loss QR is borne by all of the other units in the system. Thus, QR is a pure external diseconomy for all units in the system. Since OR has twice the slope of OQ, then for any value of N, NQ = QR; that is to say that the entry of another unit will tend to impose losses on the remainder of the system equal to the phase ingression loss borne by the marginal unit.

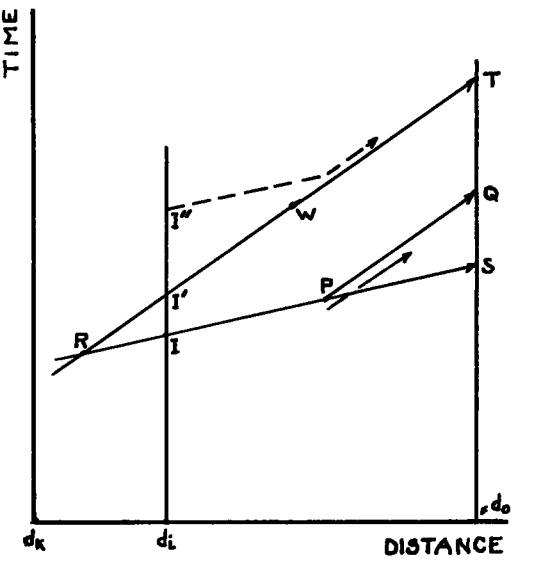

Figure 5.

RELATIONSHIP OF INGRESSION TO CONGESTION

Both ingression and congestion arise from the relationship of the deadline demand upon a system and its capacity. They are thus related phenomena. Ingression is a form of loss which exists in any movement system when instantaneous demand exceeds instantaneous capacity, even when the system is operating at maximum efficiency. In short, given the technology of the system, the ingression function represents the irreducible minimum of time loss resulting from the aggregation of demand.

Congestion losses on the other hand, are time losses which arise from the reduction of the "free-flow," or desired, velocity imposed on a unit because of the behavior of other units in the system. In a "saturated" transportation system where passing opportunities are rare or nonexistent, the movement of all units will approach a uniform flow condition whose velocity is dictated by the slower speeds of the early entrants. This system-dominant velocity may very well be less than the capacity velocity of the system (see Appendix A).

Figure 5 is a time-distance chart based on a simple uniform transportation system of length d_k. T is the deadline at destination d_o. RT is the movement course for any unit moving at system-dominant velocity to arrive at T. A unit domiciled at d_i with a free-flow velocity represented by the slope of RI will determine a point S on the time axis which will represent his deadline if he were to be the first arrival. Hence, IS represents the ideal movement course if conditions permitted him to move at his desired velocity to arrive at d_o first. The course I'T, whose slope represents the system-dominant velocity, is the "latest" course that the unit would tend to follow, for if he began later, say at I", planning to catch up to this terminal travel course by moving initially at his desired velocity, he risks finding the course occupied by a unit entering, say, at W, forcing him to violate his deadline. Under the best of the "safe" conditions, course IS could be realized, in which case the period ST can be defined as the realized ingression loss for this unit for this phase, and there would, hence, be no congestion loss. At the latest I'T would be followed, in which case ST would represent the congestion loss and there would be no actual ingression loss realized. In general, a course such as IPQ would probably be characteristic, where the unit moved at desired velocity until it encountered a slower unit at P, where it was forced to reduce velocity to that of the encountered unit, which might be the end unit of a uniform flow queue. In this case, SQ would represent direct congestion loss, while QT would represent rea-

lized ingression loss.[5] In this example, it can be seen that congestion losses and realized ingression losses tend to be complementary. In conclusion, the maximum potential phase (ST - II') ingression loss can be continued as a measure of the congestion potential in this simple example.[6]

How much congestion will be realized in any one circumstance will be a product of the departure-velocity decisions of all units in a given demand situation. To the extent that these decisions appear to have a random nature within the opportunity limits, the actual emergence of congestion in a transportation system will have certain random characteristics which make it difficult to anticipate in time and space, given the present state of understanding.[7]

It will be noted that the ability of a unit located at d_i to select a course is restricted by the pattern of decisions made by units beyond d_i on the system. Once selected, his ability to realize that decision will be limited by the decisions of units between d_i and d_o (as illustrated by the situation at P in Figure 5, where the path is deflected from PS to PQ because of the behavior of a succeeding unit).

A final point involves the relation of ingression to that form of congestion which results from variations in capacity at points along a system. As in a hydraulic system it is the size of the least pipe that determines the rate of flow at the delivery point, so in a transportation system the flow capacity is a function of the least capacity in the system, and its location with respect to the distribution of demand. Up to this point movement systems have been discussed as though capacity were uniform over the length of the system. This posits a uniform, constant-velocity flow. Since this rate of flow may be less than the desired rate of movement for the individual units, this may be defined as a congested situation, and in this sense one can talk about subjective congestion. Now where capacity at all points along the system is not uniform, in other words, where there are capacity "chokes" in the system, objective congestion loss can be defined as that loss which occurs in a movement system at a point where capacity is less than that in the preceding segment of the system. Here the rate of flow is diminished and units tend to queue up, expending time waiting entry through the constricted part of the system.

The intimate relationship between the concept of ingression and objective congestion loss is vividly illustrated in the following excerpt from the Washington Post of November 18, 1958:

> ". . . AEC has its own daily traffic tie-up which an employee describes as one of the most exasperating periods in each day. Up to 700 cars jockey for position on AEC's lone access road leading to Route 240. The employee said it had taken him twenty minutes the night before to get from the parking lot to Route 240, the usual time was from 12 to 15 minutes . . . ."

This is both an ingression and a congestion phenomenon. The over-all capacity of a system is governed by the least capacity at any point. These capacity chokes give rise to a pure form of congestion loss which occurs at any point where capacity is less than that in the preceding segment of the system. Here the rate of flow is diminished and units queue up waiting entry through the choke.

If the rate of approach of units to the choke could be managed, the congestion loss at the choke could be eliminated. At the other extreme, units may approach the choke

---

[5] Less, of course, the period II', which would not be absorbed in the movement.

[6] The relationship between ingression and congestion is basically a matter of whether or not the loss resulting from the capacity-demand relationship takes place within or outside of this system.

[7] Here the systemic congestion of the above discussion should be distinguished from externally-caused congestion. A traffic accident or a vehicle breakdown can obstruct the flow of units and create a seriously congested situation. Such interruptions are external to this consideration and hence do not enter into the concept of congestion potential.

at the approach capacity. If a system has a capacity choke, the total ingression in the system is computed in terms of the capacity limits set by the choke:

$$Y_c = \frac{N^2}{2C_c}$$

in which

Y_c = total ingression at choke capacity,

C_c = choke capacity.

In the absence of the choke, total ingression would be set by the approach capacity:

$$Y_a = \frac{N^2}{2C_a}$$

in which

Y_a = total ingression at approach capacity,

C_a = approach capacity.

Then the total ingression loss attributable to the choke itself must be the difference between these two ingression levels:

$$X = \frac{N^2}{2C_c} - \frac{N^2}{2C_a}$$

$$= \frac{N^2}{2}\left(\frac{1}{C_c} - \frac{1}{C_a}\right) = \frac{N^2}{2}\left(1 - \frac{C_c}{C_a}\right)$$

in which X = the ingression loss attributable to the choke.

As previously observed, ingression loss at a capacity choke takes place in the form of vehicles queuing up awaiting entry through choke, and hence X is actually a direct measure of a congestion time loss resulting from the capacity impairment. The component $\left(1 - \frac{C_c}{C_a}\right)$ is of interest because it defines the proportion of the total ingression loss of the system which will take place at the choke. Henceforth it will be referred to as the choke coefficient k, and it can be thus shown that the total length of the maximum queue Q at the choke, and the total congestion time loss X at the choke are functions of the choke coefficient q and the deadline demand N:

$$X = \frac{kN^2}{2}$$

and $Q = kN$,

for all large values of N.

To illustrate this point, a path with a uniform capacity of 2,000 vph on which there is a simple deadline demand of 1,000 vehicles may be examined. Assume that the full demand is asserted at the beginning of the path and is destined for the terminus of the path. Now assume, (1) that there is a capacity choke in the path (e. g. a signalized intersection) which will pass vehicles on the path only 75 percent of the time, and (2) that at the terminus vehicles can be evacuated from the path at a maximum rate of 1,200 vph. (As though it were a second capacity choke.) At the most general level this is a simple system with a total capacity of 1,200 vph, developing a total phase ingression computed as follows:

$$Y = \frac{N^2}{2C_c} \tag{1}$$

$$Y_c = 1{,}200 = \frac{1{,}000^2}{2{,}400} = 416.67 \text{ hr} \tag{2}$$

where the capacity for the total system is set by the least capacity of any point. If there were no chokes and the path capacity dominated, the ingression level would be computed thus:

$$Y_a = 2{,}000 = \frac{1{,}000^2}{2 \times 2{,}000} = 250 \text{ hr.} \tag{3}$$

The net result of the two chokes is to increase the total phase ingression by 166.67 hr.

Now congestion time loss X_1 at the first choke can be determined by taking the indicated ingression at the choke capacity and multiplying it by the choke coefficient k_1, where

$$k_1 = (1 - \frac{C_1}{C_a}), \quad C_1 < C_a \tag{4}$$

in which

C_1 = choke capacity of the first choke,
C_a = approach capacity to the first choke.

Then

$$X_1 = k_1 Y_1, \tag{5}$$

or,

$$X_1 = \frac{N^2}{2C_1}(1 - \frac{C_1}{C_a}) = \frac{N^2}{2}(\frac{1}{C_1} - \frac{1}{C_a}),$$

in which

N = deadline demand,
Y_1 = indicated ingression at first choke capacity.

The approach capacity to the second choke is set by the choke capacity of the first. Hence the choke coefficient for this second choke k_2 would be defined thus:

$$k_2 = (1 - \frac{C_2}{C_1}), \quad C_2 < C_1 \tag{6}$$

in which

C_2 = choke capacity of second choke,
C_1 = choke capacity of first choke,

and congestion time loss X_2 defined

$$X_2 = k_2 Y_{C_2}, \tag{7}$$

$$X_2 = \frac{N^2}{2C_2}(1 - \frac{C_2}{C_1}) = \frac{N^2}{2}(\frac{1}{C_2} - \frac{1}{C_1}).$$

Then the total congestion time loss at both chokes X is defined as

$$X = X_1 + X_2$$

$$X = \frac{N^2}{2}[(\frac{1}{C_1} - \frac{1}{C_a}) + (\frac{1}{C_2} - \frac{1}{C_1})] = \frac{N^2}{2}(\frac{1}{C_2} - \frac{1}{C_a}). \tag{8}$$

Coming back to the example

$N = 1{,}000$
$C_a = 2{,}000$
$C_1 = 1{,}500$
$C_2 = 1{,}200$

It can now be computed that the total congestion time loss X imposed by the two chokes is $X = 166.67$ hr or an average of 10 min per vehicle. It can further be computed that this congestion loss is divided between the two choke effects;

$$X_1 = X_2 = 83.33 \text{ hr}$$
or 5 min per vehicle average.

This total congestion loss figure of 166.67 hr is identical with the additional ingression loss computed earlier. Thus the total loss in a system is equal to the computed ingression at path capacity plus the objective congestion loss set by successive restric-

tions in capacity in the direction of movement. This formulation applies to the perfectly managed (maximum efficiency) system. It is the minimum level of loss. Inefficiencies resulting from behavior of units are external to this system, and would hence increase the actual level of loss experienced.

Under this simple formulation, the kind of congestion quantified is that which results from successive restrictions on capacity in the direction of movement.

When capacity is uniform throughout (that is, where $C_a = C_1 = C_2$, etc.), no system congestion loss is defined (the ingression loss is still very real, however).

(Note an interesting conclusion from Eq. 8, a project to eliminate the capacity restraint imposed by the first choke would not change the total quantity of congestion loss in the system; it would merely shift it to the second choke.) In short, the actual objective congestion loss at a capacity choke will lie somewhere between zero and X as defined above, and will vary with the rate at which units approach the choke. Congestion loss at this point will be greatest when the approach flow approximates approach capacity; it will be zero when the approach flow is equal to or less than the choke capacity. Here again, the entry decisions of the individual units will determine the scale of congestion within the limits indicated. Here ingression analysis provides directly a measure of congestion potential.

Thus, two features stand out: (1) in the most ideal of transportation systems the occurrence of congestion resulting from unit behavior may be a sporadic, almost random event, but (2) ingression has a direct relationship to the objective congestion potential in the system. Even though some forms of congestion are sporadic and unpredictable, ingression is a determinable, functional characteristic of a movement system emerging solely from the technology of the system as it influences capacity and from the time-distribution of demands on the system. In short, even though the quantitative prediction of congestion resulting from the chance behavior of units in the system appears unfeasible at this stage of knowledge, the prediction of objective congestion potential as measured by ingression is a comparatively simple matter.

Throughout the remainder of this paper ingression will be considered a feature of an ideal movement system, and in this sense, ingression represents the residual congestion potential in any transportation system having the same characteristics as the movement system.

Ingression as a concept, then, has relevance to the traffic and highway planner as a technique for quantifying the congestion element in traffic systems. For students of urban spatial structure and land use, the analysis of ingression has a more intrinsic relevance as a measure of a fundamental external diseconomy associated with urban growth and change; ingression losses deriving from the innate characteristics of transportation systems will be realized whether or not congestion losses occur.

APPLICATION OF INGRESSION ANALYSIS TO SOME SPECIAL CASES

There are a number of special arrangements in transportation systems which have unique implications in terms of ingression losses. These operate either by modifying the capacity of the systems, as in the case of traffic signals or supplementary arteries, or by affecting the nature and distribution of demand in systems, as in the staggering of employment hours. In what follows, several of these basic arrangements will be analyzed in terms of ingression effects, not so much to provide a rigorous analysis as to demonstrate application of the logic of ingression analysis and the manner in which it can be applied to varying characteristics of transportation systems.

Cyclical Interruption of Flow

A common type of time loss in traffic systems is that resulting from the periodic interruption of flow on one system to permit cross flow. The typical device for such interruption is the fixed interval traffic signal. To analyze the phase ingression effects of such interruption a system such as that schematically illustrated in Figure 6 can be set up. Here is a path OD along which there are located N units of maximum deadline

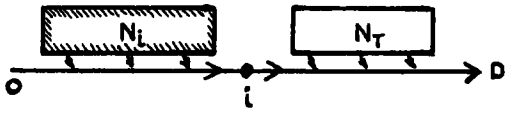

Figure 6. Scheme for interrupted flow system.

demand. At some point I along OD there is a fixed interval interruptor such that a part of the demand N_i must pass through the interruptor in order to arrive at D, while the remainder of the demand N_T is not directly affected by the interruptor. The effects of the interruptor on the flow are shown in Figure 7. It passes pulses of N_i each having a length of p and separated by intervals of length $\bar{p}$. The uninterrupted demand enters the flow by filling in the empty intervals. It is assumed that the first pulse to arrive will be a N_i pulse. There will be some alteration in the ingression values if an N_T pulse is taken as the first arrival, but where N and C are quite large and p and $\bar{p}$ are relatively small the difference is insignificant.

It will be noted that there are two cases. In the first case, the uninterrumpted demand N_T is exhausted before the interrupted demand N_i. The flow pattern thus demonstrates two components—a "saturated" component and an "unsaturated" component which is characterized by unoccupied intervals between pulses of N_i. In the second case the interrupted demand N_i is exhausted first and there is no similar unsaturated segment, since the uninterrupted demand can flow continuously until exhausted after the passage of the last pulse of N_i. It is only in the first case that there are any additional ingression effects over and above what the system would experience in the absence of the interruptor, and these flow from the unoccupied intervals of the unsaturated component.

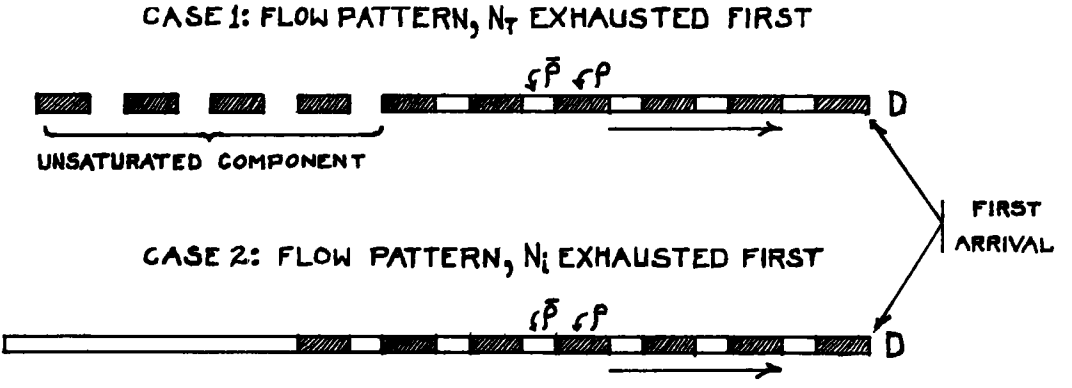

Figure 7.

The number of pulses of N_i can be expressed by $\frac{N_i}{p}$, and the number of pulses of N_T by $\frac{N_T}{\bar{p}}$. Since interruption ingression will only exist where the number of pulses of N_T is less than the number of pulses of N_i, this can be summarized:

$$\text{where } \frac{N_i}{p} > \frac{N_T}{\bar{p}}, \quad \text{then } {}_IY > 0$$

$$\text{where } \frac{N_i}{p} \leqslant \frac{N_T}{\bar{p}}, \quad \text{then } {}_IY = 0$$

or similarly:

where $\frac{N_i}{N_T} > \frac{p}{\bar{p}}$, then ${}_IY > 0$

and where $\frac{N_i}{N_T} < \frac{p}{\bar{p}}$, then ${}_IY = 0$

In this first case it can be shown that the additional phase ingression loss from the operation of the interruptor ${}_IY$ is mathematically defined as follows (see Appendix B-2):

$${}_IY = \frac{(N_i - RN_T)^2}{2RC}, \text{ where } R = \frac{p}{\bar{p}}$$

The ingression loss from interruption can accordingly be managed by the selection of values for p and $\bar{p}$, and there is a set of such values that will yield ${}_IY = 0$.

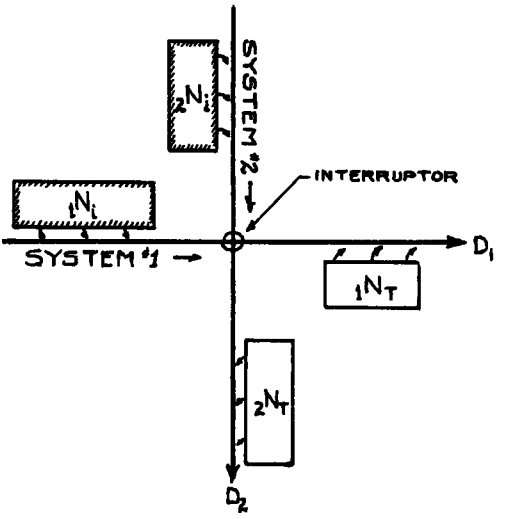

Figure 8.

Cyclical Interruption, the Intersection of Two Systems

The case of two systems with a common deadline so organized in space that part of the flow of each must cross part of the flow of the other at some point I is shown schematically in Figure 8. Instead of 2 demand segments, there now are 4, two in each system: in the first system they are designated as ${}_1N_i$ and ${}_1N_T$, in the second system as ${}_2N_i$ and ${}_2N_T$. To complicate the problem it may be assumed that in each cycle

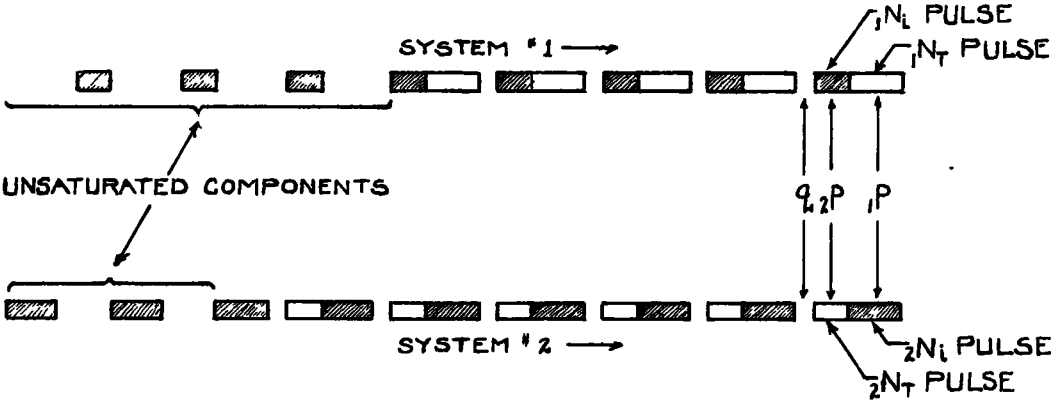

Figure 9. Flow pattern—intersecting systems.

there is a "queuing" time loss indicated by q, which is common to both systems. The pulse pattern is now modified to include 3 elements: the q-loss, the interval during which the interrupted segment of the first system flows through the interruptor ${}_1p$, and the interval during which the interrupted demand segment of the second system flows through the interruptor ${}_2p$. This pattern is indicated in Figure 9.

Some of the characteristics of this simulation are:

1. The introduction of the element q means that there will always be interruption ingression losses in both systems resulting from the effects of q in reducing the capacity of both systems.

2. There will be additional interruption ingression losses in the first system if

$$\frac{{}_1N_i}{{}_1N_T} > \frac{{}_1p}{{}_2p}$$

and in the second system if

$$\frac{{}_2N_i}{{}_2N_T} > \frac{{}_2p}{{}_1p} .$$

3. There is one case in which there will be no interruption ingression losses in either system other than those imposed by q. This case is described in terms of a set of relationships between the four demand segments and the relative lengths of ${}_1p$ and ${}_2p$:

$$\frac{{}_1N_T}{{}_1N_i} = \frac{{}_1p}{{}_2p} = \frac{{}_2N_i}{{}_2N_T} .$$

In all other cases there will be interruption ingression losses over and above the normal system ingression and the q-type losses previously discussed, and such loss may occur in either or both systems.

4. ${}_2p$ can be increased relative to ${}_1p$ with the following consequences: the ingression in the first system will be increased while that in the second system will be decreased. Hence, it is suggested that there is a value of $\frac{{}_1p}{{}_2p}$ which will minimize the total ingression loss in both systems.

Under the assumption that both systems have different basic capacities ${}_1C$ and ${}_2C$, and utilizing the following definitions:

$${}_1N = {}_1N_i + {}_1N_T$$

$${}_2N = {}_2N_i + {}_2N_T$$

$$\text{and} \quad {}_1R = \frac{{}_1p}{{}_2p}$$

the total amount of interruption ingression in both systems (${}_I{}^{Y_{1+2}}$) is defined as follows:

$${}_I{}^{Y_{1+2}} = \frac{[{}_1N - {}_1N_T({}_1R + 1)]^2}{2{}_1C{}_1R} + \frac{[{}_2N({}_1R + q{}_1R + q) - {}_2N_T({}_1R + 1)]^2}{2{}_2C[1 - q({}_1R + 1)][{}_1R + q({}_1R + 1)]}$$

This expression is developed in Appendix B-3. Taking the capacities , demand segments, and q as given, it can be generalized that ${}_I{}^{Y_{1+2}}$ is a function of ${}_1R$ thus:

$${}_I{}^{Y_{1+2}} = f({}_1R)$$

A decision problem might call for the selection of one value of ${}_1R$ (or $\frac{{}_1p}{{}_2p}$) which would minimize ingression losses in the two systems. Quite simply, this is the value of ${}_1R$ which will satisfy

$$\frac{d({}_I{}^{Y_{1+2}})}{d({}_1R)} = 0,$$

where the second derivative is positive.

Staggering Demand

A technique for modifying the demand upon the system involves breaking demand down into smaller segments and giving each segment a different deadline, as in the

staggering of working hours. This situation lends itself to ingression analysis. A total demand N is broken into 4 segments, N_1, N_2, N_3, and N_4. A deliberately unrealistic deadline requirement says that all units of the earliest segment N_1 must arrive at the destination sometime between T_0 and T_1; for the second segment N_2 between T_1 and T_2, etc. Thus all units must arrive at some time prior to T_4, as illustrated in Figure 10.

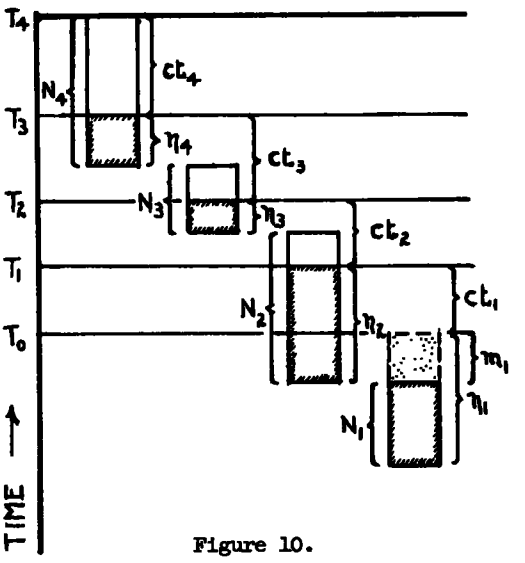

Figure 10.

The system has a common capacity C for all segments of deadline demand, and thus the expression ct_4 represents the total number of units that can arrive during the scheduled interval T_3 to T_4 (or t_4). In Figure 10, then, there is a surplus of units n_4 which cannot satisfy the deadline requirement and which hence develop ingression losses in period t_3. n_4 also has the effect of pushing back the third segment of demand so that it can utilize for arrival only part of its target period, leaving a surplus of demand n_3 to develop ingression losses in the preceding period. Were it not for n_4, all units of N_3 could satisfy the deadline requirement and arrive during t_3 without developing any ingression loss at all. This effect of cumulative excess demand similarly affects N_2 and N_1, and the total ingression loss is the sum of the ingression losses resulting in each situation.

A qualification is indicated by m_1 in Figure 10. Here the accumulation of n_4, n_3, and n_2 has been so great as to push back N_1, whose deadline period is t_1, to a point where the first unit of N_1 must arrive sometime before T_0. Thus, m_1 represents an element of delay for the entire N_1 segment, during which no N_1 units can arrive because the time is preempted by the demand of later segments.

The ingression loss of the last segment of demand N_4 can be expressed as follows:

$$Y_4 = \frac{n_4}{2C}$$

and similarly Y_3 and Y_2 can be expressed. With Y_1 the case is somewhat different due to the delay element of m_1. The time period represented by m_1 is occupied by no N_1 units and hence must be subtracted out of the ingression expression for N_1. The total ingression losses that would be experienced by m_1 units is expressed as follows:

$$Y_{m_1} = \frac{(m_1)^2}{2C}.$$

Thus, the ingression expression for the deadline demand segment N_1 becomes

$$Y_1 = \frac{n_1^2 - m_1^2}{2C}.$$

In a more general sense, the amount of ingression experienced in any deadline demand segment N_a can be expressed

$$Y_a = \frac{n_a^2 - m_a^2}{2C}$$

or, for the entire system

$$Y = \frac{\Sigma n^2 - \Sigma m^2}{2C}, \quad \text{where } n \geqslant 0, \quad \text{and } m \geqslant 0.$$

In Appendix B-4, this expression is developed in terms of the size of the deadline demand segments, the capacity of the system, and the distribution of the time restraints.

The decision problem suggested here is the arrangement of the demand segments in a system to minimize ingression losses. Brief examination of Figure 10 suggests that if the time periods are chosen so that $ct_a = N_a$ there will be no ingression loss in the system as described. There will likewise be no capacity wasted.

Ingression Consequences of Alternate Routes

As a final case, a situation in which there are two routes of different capacities connecting the demanding units with the assembly point may be considered. The problem suggested is how to distribute demand between the two routes so that phase ingression losses are minimized. The minimum situation will be realized when the marginal ingression losses along both routes are equal, for at this point to shift one unit from one route to another would result in the addition of a greater amount of ingression on the receiving route than would be subtracted from the total on the other. In a simple system the marginal phase ingression loss has been defined as N/C. Thus, to minimize the total phase ingression loss in such a dual system, the total demand would have to be so distributed that

$$\frac{N_1}{C_1} = \frac{N_2}{C_2}$$

or

$$\frac{N_1}{N_2} = \frac{C_1}{C_2}.$$

In short, the ingression losses in such a system will be minimized when the demand is distributed among alternate routes in proportion to their capacities.

These four situations have been discussed here only to demonstrate the manner in which the concept of ingression can be applied to different flow situations as an analytical device to quantify an important social cost emerging from the organization of the basic elements of capacity and demand in movement systems. Each of these cases suggests a decision problem in the management of a system and at the same time suggests a partial "efficiency criterion" for decision-making in such cases. Ingression behaves like a kind of systemic friction; given the coefficient of friction between two surfaces, the amount of friction is a function of the mass of the body moved. This is not unlike ingression, where given the capacity of the system, the amount of ingression loss is a function of the deadline structure of demand. Friction is associated with the dissipation of energy, ingression with the dissipation of socially valuable time. These time losses can be construed as costs (non-money costs at this stage) associated with the organization of movement systems. Considered as costs, they have policy implications within the framework of cost-benefit appraisal of public policy.

SOME POLICY APPLICATIONS OF INGRESSION ANALYSIS

The usefulness of ingression analysis for decision-making stems from the economic characteristics of ingression losses. Ingression involves the expenditure of a socially valuable commodity—time—in the realization of given social and economic purposes. Decisions about urban transportation systems are essentially public investment decisions, and the decision-making criteria must attempt to relate the social benefits from such investment to the social costs of specific investment programs. Ingression is a measure of such a social cost implicit in movement systems, and in a purely engineering sense, average ingression cost is a measure of systemic efficiency. Considering the "trip" to be the basic unit produced by a movement system, average ingression can be construed a unit input, and hence as a technical coefficient that varies with the scale of the "output." Thus, the principle of economizing suggests that, given a level of output and the "prices" of the inputs, one should select that investment program which will minimize the average costs of the system.

This suggests that to be useful for policy purposes it would be necessary to establish a "price," or value, for the time spent in transit by the system user. This intricate

question cannot be investigated here except to say that the economic concept of the marginal value of leisure provides a good theoretical basis for estimating that price, given the structure of wage payments to the user.

The remaining portion of this paper is concerned with some specific illustrations of ingression analysis as a decision-making tool in transportation systems. It will observe some simplifying assumptions:

1. That ingression costs are the only general social costs of immediate concern;
2. That the transportation systems in the examples are simple, homogeneous uniform capacity systems;[8] and
3. That the value of indirect social benefits does not vary substantially among the investment alternatives.

The Role of Ingression in Operation Decisions: Selection of a System-Dominant Velocity

A simple, non-investment type of decision in a transportation system involves the regulation of velocity. This regulation may be conditioned by any of a number of criteria: reduction of the risk factor, increasing rate of flow, vitiating congestion, etc. A situation may be taken where velocity is to be regulated in order to minimize the total time cost of the system, given demand N and the technology of the system, and the distribution of demand along the system. For any unit the time cost of a single trip is equal to the travel time at system-dominant velocity plus the average realized ingression loss, and can be expressed as follows:

$$t = \frac{s}{v} + \frac{N}{2c} = t(v)$$

in which

t = total time-cost of one trip,
s = traveled distance between the origin of the given unit on the system and the system destination,
v = system-dominant velocity (see page 5),
N = total demand on the system, and
c = velocity-specific capacity.

It will be noted that c is some function of v, which will be expressed as c(v), (see Eq. 13 in Appendix A). The total time cost T for one phase of the cycle will be equal to the total travel time plus the total ingression cost of the system:

$$T = \sum_{i=1}^{N} t_i = \frac{N^2}{2\,c(v)} + \frac{1}{v} \sum_{i=1}^{N} s_i$$

In this expression s_i represents the spatial distribution of the demand along the system.

Assuming that c(v) is a more or less conventional function of the form discussed in Appendix A, t(v) will have a minimum value, given s; it will, however, have a different minimum for every value of s. If

$$\sum_{i=1}^{N} s_i = S$$

then

$$T = \frac{N^2}{2\,c(v)} + \frac{S}{v}$$

The decision-problem, then is to select a value for the system-dominant velocity v which will minimize the total time-cost T. This will be the value of v such that

$$\frac{dT}{dv} = 0$$

and the second derivative is positive.

---

[8] Although by extension of the argument the variable capacity system could be comprehended with its objective congestion potential, as discussed earlier.

Assuming no other restraints, such a policy formulation might suggest the setting of a minimum speed limit v to maintain rates of flow which will tend to minimize the total time cost to all units. It is important to note that this value of v is unique to the system defined by N and D. For different values of N and D there will be different values of v which will minimize the time-costs of the specific system; in other words, there is no general value of v which is relevant to all values of N and D. If time-cost minimization is the criterion, velocity regulation must be tailored to the specific, demand-capacity characteristics of the system.

Ingression Factors in Urban Transportation Investment Decisions: Case 1, The Simple System

A characteristic type of system investment problem involves an extant system with a high level of congestion and a set of investment alternatives to reduce the time losses in the system. Take such a system with a high degree of ingression Y_0. Assuming that there are a number of alternative projects which can reduce the value of Y_0, each project can be translated by costs into an investment program requiring a set amount of capital I. Hence, I_1 as the capital required by Alternate 1 which will produce a new capacity in the system of value C_1, and hence a new ingression value Y_1. Let p stand for the value of a unit of ingression loss to the road user. Since Alternate 1 will reduce the amount of ingression per cycle, the value of the ingression reduction can be construed as a kind of "social income," J_1 which can be expressed as follows:

$$J_1 = p(Y_0 - Y_1).$$

Assuming no change in demand N,

$$J_1 = \frac{pN^2}{2}\left(\frac{1}{C_0} - \frac{1}{C_1}\right).$$

Define the annual costs K_1 of Alternate 1 to consist of interest and depreciation on capital I_1 plus the increment to operating costs (if any) resulting from the program thus

$$K_1 = (i + d)I_1 + O_1.$$

Then any program a is "justified" so long as the social income produced by it is greater than the annual social costs of the program:

$$J_a \geq K_a.$$

In short, to maximize the social income,[9] select that investment program K_a which will maximize the benefit-cost ratio

$$(\max)\frac{J_a}{K_a} = (\max)\frac{\frac{pN^2}{2}\left(\frac{1}{C_0} - \frac{1}{C_1}\right)}{(i + d)I_a + O_a}$$

This is a case in which the substitutability of capital costs for ingression costs is the basis of the decision problem. Given the cost of capital i (assuming that depreciation is a technological constant) and the "price" of ingression losses p, an optimum investment program which will maximize the social surplus from the operation of the transportation system can be identified.

Ingression Factors in Urban Transportation Investment Decisions: Case 2, The Hypersystem

At an even higher level are the decisions involving the long-range programing for the development of an urban transportation system. To discuss decisions at this level it is necessary to digress for a moment to define the concept of a "hypersystem." The term "hypersystem" is used here to describe the organization of two or more move-

[9] The fact that K_a represents any one of a finite set of discrete alternatives suggests that a linear programing formulation of the problem might be more versatile.

ment systems into a higher level, integrated system, the basic characteristic of which is an instantaneous capacity greater than 1.

It can be assumed that the movement hypersystem in this framework is homogeneous, that is, consisting of a given number A of movement systems whose technological characteristics are identical in that their non-instantaneous capacities are equal. The total ingression Y_A in such a system varies with the amount and distribution of demand among the components:

$$Y_A = \sum_{i=1}^{A} \frac{N_j^2}{2C_j}$$

in which

Y_A = total ingression loss per cycle in the hypersystem,

C_j = capacity of component "j," and

N_j = demand on component "j."

At an earlier point it was suggested that the total ingression among several systems with a common assembly point would be minimized under the condition that each component bear that proportion of the total demand that its capacity bears to the total capacity of the several systems. A hypersystem meeting these conditions will be referred to as an optimal hypersystem: no other allocation of capacity to demand will produce a lower level of ingression cost.

Thus, in a non-optimal hypersystem a basic kind of investment decision is what might be called an "optimizing" decision; that is, a decision addressed to reducing or eliminating the disparities in average ingression costs among the component systems. This kind of decision is similar to that discussed with respect to simple systems: the selection of an investment program with the highest social benefit-cost ratio, but with the additional restraint that it must tend to reduce the disparities, for only in an optimum hypersystem are the social costs spread equitably among all components of the system (to a certain extent, in a dynamic situation the operation of the urban land market may have an optimizing effect by suppressing development in high ingression sectors and encouraging it in the low ingression sectors).

There is another kind of investment decision associated with the hypersystem. At what level of average ingression should the total system be planned? The standard which has been followed in transportation planning has been to match hourly demand with equal hourly capacities. This implicitly accepts an average ingression of ½ hr per unit per cycle where the time distribution of demand within the hour is highly concentrated. In other words, the matching of hourly capacities to hourly volumes will produce a low volume of ingression loss only where the peak demand upon a system is evenly distributed throughout the hour. There is little to justify such a standard in terms of investment efficiency. If the marginal value of the road user's time-in-transit is high and the cost of capital low, such a standard could be grossly inefficient, for the social costs of ingression would be excessive. At the other extreme, where the value of such time is low and the cost of capital is high, such a standard might well result in the uneconomic allocation of public investment. In short, the efficient level of ingression loss in an optimum hypersystem will be a function of the costs of capital and the costs of ingression.

In the course of this section, this paper has attempted to illustrate the manner in which ingression analysis can be applied to decision-making with respect to urban transportation systems, under the assumption that the proper criteria for investment decisions relates total direct social benefits to the total direct social costs involved in the transportation system. Hopefully, it served to demonstrate that the social cost factor of ingression has relevance to the appraisal of an investment program for the elimination of a grade intersection as well as to the appraisal of a total mass transportation program for a metropolitan area. The major contribution that ingression analysis can make in social decision-making is to assure a broadening of the conventional

benefit-cost concepts to embrace a much larger proportion of the economic effects of changes in urban transportation systems. The traffic engineer and the highway planner play a strategic role in such decision-making; any technique which will permit them to integrate demand-capacity effects into their total framework of analysis and appraisal may make their recommendations more sensitive to the broad array of social objectives and more persuasive to the policy-makers.

CONCLUSION

This paper is an effort to develop a theoretical framework within which the time losses associated with transportation systems can be analyzed. It takes as basic variables, (1) the load imposed upon the system in the form of deadline demand and its distribution in time, and (2) the technology of the system as it affects the capacity of the system. It quantifies the time costs emerging from the ingression features of the system and suggests the limits of objective congestion costs.

Such a theoretical expression has several virtues. It has the merit of generality; it can be applied to any system in which time has value to the participating units, technological capacity can be identified, and in which demand can be quantified with respect to a schedule of deadlines. It has versatility; it is as relevant to the mass transit situation as to the auto-exclusive system, or even to mixed systems. It has a special relevance to conditions of system saturation such as peak-period conditions in urban traffic systems. This versatility has been demonstrated in applying the analysis to several special arrangements of transportation systems. It is simple in that the basic capacity-demand relationship is simple. Complexity in application will vary directly with the degree of precision sought.

Not the least of its virtues are its economic characteristics. The substitution of capital costs for time costs is the basic feature of many investment decisions in transportation systems. Ingression and congestion losses behave as pure external diseconomies such that as the scale of the system increases average time costs likewise increase. It is the economic characteristics of the framework which permits its integration into a broader theoretical framework of the urban space economy.

However, there is much to be done before this framework is empirically useful. It needs direct empirical testing, and this is a problem for transportation research. Simple techniques are needed for identifying the volume of demand and its time distribution in such complex situations as the Central Business District; its stability in the short run under varying conditions must be known. This framework must be given a manageable spatial dimension so that one can talk in terms of the spatial distribution of demand and assembly points as variables in the level of these time costs. Finally, it will take rigorous thought to determine the useful limits of such an analytical technique; to this end, this paper addresses the broad competences of both engineers and economists.

Appendix A

DEFINITION OF CAPACITY

In this hypothetical movement system all units move under "capacity" conditions until demand is exhausted. This assumption is necessary to identify those losses which are innate in the system. Capacity in this sense is not a simple concept.[10]

[10]The complexity of the concept of capacity has given rise to a multiple definition in highway traffic planning:

"Basic capacity: the maximum number of passenger cars that can pass a given point on a lane or roadway during one hour under the most ideal roadway and traffic conditions that can possibly be attained . . . .

"Possible capacity: the maximum number of vehicles that can pass a given point on a lane or roadway in one hour under prevailing roadway and traffic conditions . . .

"Practical capacity: the maximum number of vehicles that can pass a given point

Fundamentally, capacity in a uniform flow movement system derives from the relationship of velocity and the distance interval between successive units.[11]

Here two senses of the term "capacity" must be distinguished.

1. That capacity which exists at a given average velocity in a specified system (either under free-flow or uniform flow conditions), or the velocity specific capacity c, and

2. That capacity in a specified system which is maximum for all values of velocity, or system capacity C.

Unless otherwise specified, the term "capacity" in this paper will refer to system capacity. The velocity at which system capacity is realized will be referred to as capacity velocity.

If the interval remains constant (is independent of velocity) the volume in a system can be increased by increasing the system's velocity, and hence capacity would depend on the technological limits of velocity. In fact, however, the instantaneous capacity[12] of the movement system will always be 1 until velocity is infinite, also, where the system consists of a single path (where the system consists of more than one path, the instantaneous capacity will be equal to the number of paths). Ruling out the case of infinite velocity, as long as instantaneous capacity is less than system demand (which, by virtue of the time-restraints, is instantaneous) there will be ingression potential in the system, even at a system velocity approaching the speed of light.

In a practical sense, however, the interval between units is not independent of velocity. As a rule, it tends to increase with velocity, and the manner in which it tends to increase is related to three basic technological characteristics. Each unit must accomodate its behavior to that of the unit immediately preceding it. The "behavior" of a unit consists of changing its velocity; thus, if the unit preceding changes its velocity negatively, the subject unit must adjust its velocity to avoid collision. Interval is, hence, conditioned by the behavioral characteristics[13] of the units, or more specifically by the risk, feedback, and mechanical characteristics of the system.

The risk factor is a function of the orderliness of the system; it is inversely associated with the ability to anticipate the behavior of preceding units. A completely orderly system would be one in which the behavior of preceding units could be anticipated by at least the time interval inherent in the feedback and mechanical characteristics of the system, wherein interval would be limited only by the physical dimensions of the units themselves. At the other extreme is the completely disorderly system in which behavior of the lead unit is completely unpredictable. In this case, the behavior of the second unit as it accomodates itself to the behavior of the first is completely unpredictable to the third, and so on through the system. Here minimum interval is

---

[10](Continued)

on a roadway or designated lane during one hour without traffic density being so great as to cause unreasonable delay, hazard, or restrictions to maneuver under prevailing roadway and traffic conditions . . ."

Bureau of Public Roads and Highway Research Board, "Highway Capacity Manual." USGPO, Washington, p. 6-8 (1950). See also Institute of Traffic Engineers, "Traffic Engineering Handbook, 2nd Ed, p. 334, New York (1950).

[11]The Traffic Engineering Handbook (p. 331) expresses this as:

$$c = \frac{5280V}{S}$$

in which

c = velocity-specific capacity in vehicles per hour,
V = velocity in mph, and
S = the spacing (interval) between successive vehicles in ft.

[12]That is, the number of units that can occupy any single time-space position on the path.

[13]But not by the actual behavior of the units.

governed by the mechanical and feedback characteristics of the system, and the distance between the units will be governed by the speed with which any one unit can react effectively to unanticipated maximum changes in behavior of the preceding unit. Thus, the extent to which interval is a function of velocity depends on the amount of disorder in the system. Where the system is completely orderly, there will be no association of interval with velocity:[14] where disorder prevails in the system, it is the time of effective response and the velocity that establish interval. It is further obvious that where a system is operating at capacity, disorder may be induced in the system by the unpredictable behavior of a single unit and that this induced disorder will prevail in the system until dissipated by excess capacity.[15]

The time of effective response depends, as has been suggested, on the "feedback" and mechanical characteristics of the system. Feedback involves the sensing of a change in the behavior of the preceding unit and the assertion of controls to accomodate the behavior of the subject unit to it.

A critical relationship exists between the frequency and scope of the changes to be controlled and the time lapse between the detection of change and the assertion of compensating control. Where the sensed changes are comparatively infrequent, the rate of change moderate, and the feedback time very small, a movement system may exhibit a high degree of articulation between units. Where, on the other hand, the system is characterized by comparatively rapid changes with a high rate of change, and where the feedback time is comparatively gross, the degree of articulation may be very low. The crucial "feedback" factor in headway is the time lapse involved from the detection of change in the behavior of one unit to the assertion of compensating control in the succeeding unit.[16] If this is instantaneous, then the only elements to govern interval (assuming a given risk level) are the mechanical characteristics of the system; if comparatively slow, then interval must encompass the feedback time lapse. Assuming a constant level of order in the system, then, slow feedback and high velocity would be associated with a very large interval.

The mechanical characteristics of the system are purely technological. What is the lapsed time (or distance) from the moment that maximum controls are asserted to the time that the unit is decelerated to zero (this being the maximum behavioral response demanded of the unit)? This time (or distance) is a function of the mass of the system and its velocity and an inverse function of the braking power that can be applied by the unit at any one moment. Thus, that component of interval associated with inertial characteristics of the system can be diminished through a reduction of the unit's mass and/or velocity, or through an increase in braking power.[17]

---

[14]Because the behavior of all preceding units would be completely predictable, and there would be no time loss in adjusting to it.

[15]Discussion of the queueing phenomenon in the "Highway Capacity Manual." p. 34.

[16]In a vehicular traffic system, the traffic engineer characterizes this feedback element as "driver perception and reaction time."

[17]This relationship has been generalized for automobiles in the "Traffic Engineering Handbook," p. 72, as follows:

$$B = \frac{0.0105\ V^{\text{exp } 7/3}}{f}$$

in which

B = vehicle braking distance in ft,
V = velocity in mph, and
f = coefficient of friction (at a velocity of 20 mph).

Which can be restated

$$B = \frac{0.0105\ NV^{\text{exp } 7/3}}{F}$$

in which

F = the minimum force necessary for deceleration, and
N = the mass of the carrier.

In summary, the velocity-specific capacity c of the system is a direct function of the velocity V of the system and an inverse function of the interval I among units:

$$c = \frac{V}{I} \tag{9}$$

Interval contains three components: the length of the unit L, the interval component resulting from feedback time lapse i_f, and the interval component associated with the mechanical characteristics of the system i_m, the latter two being modified by a disorder function, henceforth expressed as a "risk coefficient" r as follows:

$$I = L + r\,(i_f + i_m). \tag{10}$$

Here the risk coefficient r will vary from zero (where the system is completely orderly) to 1 (where the system is completely disorderly).[18]

The relation of the risk coefficient to some measure of orderliness will depend on the scope of potential loss, the "negative payoff," associated with malarticulation of the system. It would be a misleading conclusion to assume that the risk coefficient should always be 1 as long as there is any negative payoff. However, the size of the risk coefficient has a direct effect on the capacity of the system; hence the real problem of valuation of the relationship of disorder to the risk coefficient resides in the relating of the marginal change in payoff to the marginal gain in capacity of the system. The gain in capacity can be measured in terms of the reduction of ingression in the system, while the valuation of the increase in negative payoff is perhaps somewhat more complex.

Mathematical expression can now be given to the relationship between capacity and the factors associated with it.

$$i_f = Vt_c \tag{11}$$

in which

t_c = time lapse associated with the feedback organization.

$$i_m = \frac{aMV^b}{F} \tag{12}$$

in which a and b are technological constants,

M = mass of carrier,

F = minimum braking power to decelerate a carrier of mass M moving at velocity V to V = 0 in distance i_m. (See footnote 17.)

Then

$$c = \frac{V}{rV\left[t_c + \frac{aMV^{(b-1)}}{F}\right] + L} \tag{13}$$

(Substituting Eq. 11 and Eq. 12 in Eq. 10 and Eq. 10 in Eq. 9)

---

[17](Continued)

This can be associated with the more general statement in Eq. 12.

[18]Thus, with respect to vehicular traffic, safe (velocity-specific) capacity might be defined as that relationship between velocity and interval that exists when the spacing of units is governed by "driver stopping distance." Since auto traffic has the characteristics of a highly disordered system, it would have a risk coefficient approaching one. Thus, the traffic engineer's definition of driver stopping distance is analogous to the concept of interval with a risk coeeficient of 1. Driver stopping distance is defined in the "Traffic Engineering Handbook" as having three components (p. 71): (1) Driver perception-reaction time (which is analogous to the feedback component), (2) Brake-lag distance, and (3) Vehicle braking distance (the latter two being subsumed in the mechanical component of the author's formulation).

and

$$C = c(\max) \quad \text{where } dc/dv = 0 \text{ and the second derivative} < 0 \tag{14}$$

in which C = system capacity

The capacity of the movement system, then, will vary with system velocity, but in a fashion determined by the level of system disorder, by the nature of the control system, and by the technological relationship of mass and braking power of the units of the system.[19]

[19] "The Highway Capacity Manual" (Table I, p. 3) presents some 23 expressions for "safe following distance" in terms of velocity. All of these expressions have the general form

$$I = xV^y + zV + L$$

if

$$x = \frac{aMr}{F},$$

$y = b$, and
$z = r\,t_c$.

This general form is converted to the author's expression, which is the denominator in Eq. 13. Thus, "safe following distance" corresponds in mathematical expression with the concept of interval, especially if r = 1, as seems indicated.

If the five expressions which accept x = 0 (which appear to assume no braking distance) and the three expressions which take z = 0 (thus assuming no driver reaction time) are disregarded, 15 expressions are left which are "full." These 15 equations describe a set of curves which are convex upward and which are heavily skewed in the direction of the vertical (or capacity) axis. (See Figure 1, p. 2 of the "Highway Capacity Manual.") Their maximums fall between 8.3 mph and 28.3 mph, and they range in volume between 1,050 and 2,520 vehicles per hour for a single traffic lane. The values for t_c ranged from 0.50 to 1.00 second (with 9 accepting the high value); 13 of the 15 accepted y = 2, while 2 assumed y = 2.3 (this is the decimal equivalent of 7/3 (see footnote 17); the values for x when divided by the conversion factor $\frac{(22)^2}{15}$ to compensate for expression in terms of ft/sec ranged from 0.0116 to 0.157 with 12 using values less than 0.05. The important point is that as long as x, y, and z, have positive values greater than 1, there will always be a "maximum" capacity associated with a system, and that capacity will be defined by the velocity. It is this "maximum" capacity that is implied in the use of the term "system capacity."

It should be acknowledged that where feedback is instantaneous, and where the mechanical characteristics of all carriers are uniform, theoretically the most persuasive expression is that which takes the coefficient x = 0, such that

$$I = zV + L$$

A little reflection will justify this: if the feedback time is instantaneous two vehicles could be moving in the system "bumper to bumper" (I = L) and any change in the behavior of the first would be simultaneously compensated for by the second; here both x and z = 0. Where the braking distance is uniform for all units, it cancels out as a spacing element, leaving only the independent variables of velocity and feedback time.

A capacity expression based on such a spacing expression yields a C that increases with V to some asymptotic value. This value is, in fact z^{exp-1} at a capacity velocity of infinity.

For purposes of simplicity, this paper has assumed that x has a positive value greater than 0 because this appears to conform more closely to empirical observations.

See also discussion in Beckmann, et al., op. cit., p. 19.

Appendix B

MATHEMATICS OF INGRESSION

B-1: System Ingression—(Refer to Figures 1, 2, 3, and 4)

The quantity of ingression loss experienced by the (n - i)th unit is

$$y_{n-i} = \frac{i}{C}. \tag{15}$$

For convenience of notation this expression can be modified:

$$y_i = \frac{n-i}{C} \tag{16}$$

in which

y_i = ingression loss experienced by the ith arrival,

n = total number of arrivals, and

C = capacity of the movement system.

Expressing the total number of arrivals n as system demand N, the mathematical expression for total ingression loss Y in such a system can be derived as follows:

$$Y = \sum_{i=1}^{N} y_i = \frac{1}{C} \sum_{i=1}^{N} (N - i) = \frac{N^2 - N}{2C}, = \frac{N(N-1)}{2C}. \tag{17}$$

The average ingression loss $\bar{y}$, or ingression loss per unit demand is then,

$$\bar{y} = \frac{N-1}{2C}. \tag{18}$$

The marginal ingression loss in terms of demand y'_N is $\frac{dY}{dN}$, or

$$y'_N = \frac{2N-1}{2C} \tag{19}$$

and the marginal ingression loss in terms of capacity y'_C is $\frac{dY}{dC}$ or

$$y'_C = \frac{N - N^2}{2C}. \tag{20}$$

Where both N and C are quite large, the following can be approximated:

$$Y = \frac{N^2}{2C} \tag{21}$$

$$\bar{y} = \frac{N}{2C}$$

$$y'_N = \frac{N}{C}$$

$$y'_C = -(\frac{N}{C})^2.$$

Ingression loss being a time loss, it is expressed in units of time established by the capacity statement, which may be in terms of units per second, minute, hour, etc. Where risk factors and demand on the system are fixed, the level of ingression loss in the system can be decreased only through an increase in the capacity of the system, and this can be effected in only two ways: (1) by increasing the number of paths (and thus increasing the instantaneous capacity of the system) and (2) by modifying the technological features of the system (braking power, feedback time, mass). (See Appendix A.)

B-2: Ingression Loss in a Simple, Interrupted System—(Refer to Figures 6 and 7)

The problem is to define interruption ingression loss Y' in terms of the length of the system T, the position of an interruptor i, the distribution of units along the system N_i and N_T, the basic capacity of the system C, and the pulse characteristics of the interruptor p and $\bar{p}$.

(a) N = total demand in the system;
(b) N_i = first element (or interrupted) demand;
(c) N_T = second element (or uninterrupted) demand;
(d) R = pulse-interval ratio = $\frac{p}{\bar{p}}$, when p = the time length of the period in which flow takes place through the interruptor, and $\bar{p}$ = the period of no flow.
k = the proportion of N_i moving in the saturated component; or the number of pulses of interrupted demand matched by pulses of uninterrupted demand divided by the total number of pulses of interrupted demand, thus
(e) $R = \frac{kN_i}{N_T}$, and $k = \frac{RN_T}{N_i}$
(f) r = ratio of interruption, or relative pulse = $\frac{p}{p + \bar{p}}$ length (also referred to as the interruption restraint).

$$Y_2' = \frac{(N_T + kN_i)^2}{2C}, \tag{22}$$

from the general ingression equation $Y = \frac{N^2}{2C}$. This is the simple ingression loss generated in the saturated sector.

$$Y'' = \frac{(N_T + kN_i)}{C}(1 - k)\,N_i, \tag{23}$$

Y'' is the ingression imposed on the nonsaturated component by the time necessary to exhaust $(N_T + kN_i)$ which is the demand in the saturated component.

$$Y_1' = \frac{[(1 - k)N_i]^2}{2rC}, \tag{24}$$

Y_1' is the simple ingression loss deriving from the nonsaturated segment, from the general ingression equation (see Eq. 22). Here $(1 - k)N_i$ is the demand in the unsaturated component and r is the interruption restraint on the capacity of the unsaturated component. (See description of problem, pp. 9, 10.)

$$\bar{Y} = Y_1' + Y_2' + Y'' \tag{25}$$

by definition, $\bar{Y}$ being the total ingression loss in the system.

$$\bar{Y} = \frac{(N_T + kN_i)^2}{2C} + \frac{(N_T + kN_I)\,(1 - k)N_i}{C} + \frac{[(1 - k)N_i]^2}{2rC} \tag{26}$$

$$N_T = N - N_i, \tag{27}$$

by definition.

$$\bar{Y} = \frac{(N - N_i + kN_i)^2}{2C} + \frac{(N - N_i + kN_i)\,(1 - k)N_i}{C} + \frac{[(1 - k)N_i]^2}{2rC} \tag{28}$$

but $$(N - N_i + kN_i) = [N - (1 - k)N_i] \tag{29}$$

Let $$(1 - k)N_i = a \tag{30}$$

by definition.

Then $$\bar{Y} = \frac{(N - a)^2}{2C} + \frac{(N - a)a}{C} + \frac{a^2}{2rC} . \tag{31}$$

$$\frac{\bar{Y} = rN^2 - 2raN + ra^2 + 2raN - 2ra^2 + a^2}{2rC} = \frac{a^2(1 - r) + rN^2}{2rC} \tag{32}$$

$$_iY = \bar{Y} - Y, \text{ and } Y = \frac{N^2}{2C}, \tag{33}$$

by definition, interruption ingression loss $_iY$ being the difference between the aggregate ingression loss of the interrupted system $\bar{Y}$ and its simple ingression loss Y.

$$_iY = \frac{a^2(1 - r)}{2rC} \tag{34}$$

but $$R = \frac{r}{1 - r} \tag{35}$$

from (d) and (f).

$$_iY = \frac{a^2}{2RC} \tag{36}$$

but $$a = (1 - k)N_i, \quad k = \frac{RN_T}{N_i}, \tag{37}$$

from Eq. 30 and (f).

$$a = N_i - RN_T \tag{38}$$

$$_iY = \frac{(N_i - RN_T)^2}{2RC}, \tag{39}$$

substituting Eq. 38 in Eq. 36.

Where $R = \frac{p}{\bar{p}}$, and $a \geqslant 0$ under the considerations previously posited, $0<k<1$. Thus interruption ingression loss varies with the location of the interruptor (N_i / N_T) and the pulse-interval ratio. Given a density distribution in such a system, the amount of interruption ingression depends essentially on the positioning of the interruptor, increasing as it approaches the assembly point. It depends, in the second place on the value of R.

B-3: Interruption-Ingression, Effect on Compound System—(Refer to Figures 8 and 9)

A realistic restraint on the value of R emerges from the intersection of two simple systems, much as in the manner of the intersection of two arterial streets, where a traffic signal regulates the alternate flows of traffic.

(a) $_IY_1$ and $_IY_2$ = interruption ingression loss on systems 1 and 2, respectively;
(b) $_IY_{1+2}$ = interruption ingression loss in both systems;
(c) $_1N$ = total demand on system 1;
(d) $_2N$ = total demand on system 2;
(e) $_1N_T$ and $_2N_T$ = uninterrupted segment of demand on both systems;
(f) $_1R$ and $_2R$ = pulse-interval ratios on systems 1 and 2, respectively;
(g) $_1C$ and $_2C$ = respective capacities of the two systems; and
(h) q = queuing loss constant.

To find: The total interruption ingression loss in both systems in terms of the pulse-interval ratios, the capacities of the systems, and the distribution of demand on both systems with respect to the point of intersection.

$$ {}_{I}Y_{1+2} = {}_{I}Y_{1} + {}_{I}Y_{2} \tag{40} $$

by definition.

Let

$$ {}_{1}N_{i} = {}_{1}N - {}_{1}N_{T} \tag{41} $$

and

$$ {}_{2}N_{i} = {}_{2}N - {}_{2}N_{T} $$

by definition.

$$ {}_{I}Y_{1} = \frac{({}_{1}N - {}_{1}N_{T} - {}_{1}R_{1}N_{T})^{2}}{2_{1}C_{1}R} \text{ and } {}_{I}Y_{2} = \frac{({}_{2}N - {}_{2}N_{T} - {}_{2}R_{2}N_{T})^{2}}{2_{2}C_{2}R} , \tag{42} $$

see Eq. 39.

$$ {}_{1}R = \frac{p}{\bar{p}} = \frac{{}_{1}p}{{}_{2}p + q} \tag{43} $$

and

$$ {}_{2}R = \frac{p}{\bar{p}} = \frac{{}_{2}p}{{}_{1}p + q} , $$

see definition (d) in Subsection B-2.

$$ {}_{1}p + {}_{2}p + q = 1 $$

Then

$$ {}_{2}R = \frac{1 - q({}_{1}R + 1)}{{}_{1}R + q({}_{1}R + 1)} \tag{44} $$

$$ {}_{i}Y_{1+2} = \frac{[{}_{1}N - {}_{1}N_{T}({}_{1}R + 1)]^{2}}{2_{1}C_{1}R} + \frac{[{}_{2}N({}_{1}R + q_{1}R + q) - {}_{2}N_{T}({}_{1}R + 1)]^{2}}{2\,{}_{2}C[1 - q({}_{1}R + 1)][{}_{1}R + q({}_{1}R + 1)]} \tag{45} $$

substituting Eq. 44 into Eq. 42 and Eq. 42 into Eq. 40.

Eq. 45 answers the requirements. Given two intersecting systems, their capacities, their distribution of demand with respect to the point of intersection and given the technological constant q, the total interruption ingression loss in the two systems is a direct function of the pulse-interval ratio R (expressed in terms of the first system ${}_{1}R$). Examination of this equation indicates that there is a case, given the demand on each system and the value of ${}_{1}R$, in which there is a value for ${}_{1}N_{T}$ and ${}_{2}N_{T}$ such that the aggregate interruption ingression loss will be zero. In all other cases, however, there is a positive interruption load loss in the system. Given all other factors except ${}_{1}R$, there is a value for ${}_{1}R$ that will minimize this interruption ingression loss. There is a different value for ${}_{1}R$ that will equate the average ingression loss in the two systems. In the special case where $\frac{{}_{2}N_{T}}{{}_{2}N} = \frac{{}_{1}N_{T}}{{}_{1}N}$ the ${}_{1}R$ which will equate the average values will also provide the minimum interruption ingression loss in the systems.

B-4: The Ingression Consequences of Complex Time Restraints

Assume that the arrival time restraint requires that all units be present at the destination at some time during the time period $T_1 - T_2$ (the time-space fixed arrival restraint hitherto has required that all units be present at the destination at the fixed time T). So long as the demand on the system does not exceed the capacity of the system during this period, there is no ingression loss, that is to say, an arrival time can be selected for each unit such that it will fall within $T_1 - T_2$, and associated with a departure time, such that the only time loss is that of ideal travel time. Where demand

exceeds capacity, however, it is necessary to satisfy the restraint that the excess demand arrive at the destination in advance of the period, and there wait until the period begins. Thus ingression loss exists in this system only when demand exceeds the derived capacity of the period and can be expressed as follows:

$$Y = \frac{(N - Ct)^2}{2C}, \text{ where } N > Ct; \text{ for all other values of } N,\ Y = 0. \tag{46}$$

in which

N = total demand in the system,
C = capacity of the system, and
t = elapsed time between the limits of the arrival time restraint.

An extension of this argument will govern the case in which the total demand is broken up into segments, each having a different arrival time restraint, but using the same system. This is analogous to the staggering of working hours in the CBD and can be expressed as follows:

(a) $\veebar$ = an operation symbol, viz. a $\veebar$ b: "a or b, whichever is larger;"
(b) n'_a = cumulative excess demand not satisfied by capacity of periods a to x;
(c) N_a = actual demand associated with period a;
(d) C = capacity per unit of time;
(e) n_a = excess of demand for period a over capacity for period a;
(f) t_a = time length of period a;
(g) x = total number of periods;
(h) a = 1, 2, 3, . . . , x;
(i) m_a = impairment of volume in period a resulting from cumulative excess volume in period a + 2 and later; and
(j) Y_a = ingression loss developed by excess of demand over capacity in period a.

$$n'_a = [(N_x \veebar Ct_x) - Ct_x] + [(N_{x-1} \veebar Ct_{x-1}) - Ct_{x-1}] + \ldots + [(N_{x-(a-1)} \veebar Ct_{x-(a-1)}) - Ct_{x-(a-1)}] \tag{47}$$

Let
$$[(N_a \veebar Ct_a) - Ct_a] = n_a \tag{48}$$

Then
$$n'_a = n_x + n_{x-1} + \ldots + n_{x-(a-1)} = \sum_{b=x-(a-1)}^{x} N_b \tag{49}$$

$$m_a = n_x + n_{x-1} + \ldots + n_{x-(a-2)} - ct_{x-(a-1)} = \sum_{b=x-(a-2)}^{x} n_b - ct_{x-(a-1)} \tag{50}$$

$$Y_a = \frac{n'_a(n'_a - 1) - m_a(m_a - 1)}{2C} \tag{51}$$

or, in high volume systems
$$\frac{Y_a = n'^2_a - m^2_a}{2C}$$

$$Y = \sum_{a=1}^{x} Y_a, \tag{52}$$

in which Y = total ingression loss over x periods.

Two conclusions are immediately generated from this expression: the maximum ingression loss results when all of the t's are equal to zero, this being the case of the fixed arrival time restraint. When this is the case, the value of this expression becomes $\frac{N^2}{2C}$. A minimum value for this expression exists for any set of t values such that for every pair Ct_i, N_i, $Ct_1 \geqslant N_i$.

This is in essence the case of the total ingression of N broken down into k independent systems, each having identical capacity restraints, and the value of the expression becomes:

$$\frac{N_1^2 + N_2^2 + N_3^2 + \ldots + N_k^2}{2C}.$$

Where neither of the special conditions above exists,

$$\frac{(N_1 + N_2 + N_3 + \ldots + N_k)^2}{2C} \geqslant Y \geqslant \frac{N_1^2 + N_1^2 + N_3^2 + \ldots + N_k^2}{2C}.$$

Thus the expression to the right can be modified to meet the conditions specified where the time-of-arrival restraints are defined as a time period, as mentioned earlier.