OPTIMIZATION OF TRAFFIC AND FACILITIES

by William Vickrey

There are probably few areas in modern economic society where conditions are as far from ideal as in the congested traffic and transportation facilities of our great modern metropolitan conurbations. This is equally true in the short run, in terms of making the best use of the facilities we have, and in the longer run, in terms of the appropriateness of the facilities for current and projected traffic needs. This relative inefficiency can be attributed in large measure to the fact that the individual user, faced with alternative ways of achieving his objectives, does not, under existing conditions, receive any obvious indication of the costs which his choice will impose on others, whether by impairment of the quality of service or by the cost of expanding the facilities to the point where this impairment is prevented.

Congestion Costs Caused in relation to Congestion Costs Borne

That these costs are important should be fairly obvious from casual experience, but a few examples will perhaps help towards appreciation of their magnitude. For a fairly wide range of conditions, for example, the relation between the volume of traffic and the time required to cover a given distance can be expressed in an equation of the following form:

\[ t = t_0 + a(q/w)^k \]  

(1)

where \( t_0 \) represents the time required in the absence of any congestion, \( i.e. \) at very low levels of traffic, \( q \) is the flow of traffic in vehicles per hour, \( w \) is a measure of the width or capacity of the facility, and \( a \) and \( k \) are constants. The second term on the right, \( a(q/w)^k \), represents the average number of seconds of delay experienced by each vehicle, at a given level of traffic \( q \), so that we can write for this average delay

\[ z = t - t_0 = a(q/w)^k \]  

(2)

and for the total delay experienced by all vehicles

\[ Z = qz = aw(q/w)^{k+1} \]  

(3)

The increase in total delay resulting from a small increment in traffic can be obtained by differentiating with respect to \( q \):

\[ m = dZ/dq = (k + 1) a(q/w)^k = (k + 1) z \]  

(4)

Thus, in addition to the delay \( z \) experienced by the marginal vehicle, there is a delay of \( kZ \) to other traffic resulting from the addition of this vehicle to the traffic stream. The value of \( k \) will vary somewhat according to circumstances. In very light traffic, where each interaction between two cars is substantially independent of any other interaction, we could expect \( k \) to be close to 1; in data for the Lincoln Tunnel, covering a range of traffic from about 50 per cent of capacity to the point around 90 per cent of capacity where queuing conditions begin to disturb the relation, a value of \( k = 4.5 \) was found to fit very closely. Under conditions comparable to this,
a motorist who was delayed 10 minutes in traffic in making a given trip would be causing in the aggregate 45 minutes of delay to others. If he were required to compensate these others for this delay before being allowed to make the trip, he might well decide on some other alternative, which he will not consider if the only deterrent is the 10 minutes of delay he himself experiences. The alternative might be to use mass transit, to join an auto pool, to make the trip at a less congested time, or to use a less congested but perhaps more circuitous route.

Queuing Costs
Where queuing conditions are set up, the impact on other traffic is likely to increase quite sharply, and can reach very severe proportions. If a given bottleneck starts to back up a queue at 7:30, which is not worked off until 9:30, a car passing the bottleneck at 7:25 before the queue is established causes no increase in the queuing time; a car arriving at 7:35 will find a short queue and experience relatively little delay itself, but will cause an aggregate of 115 minutes of delay to others. Effectively one car is added to the “net queue” from the time of arrival of the car in question until the queue is ultimately worked off. It is assumed, as seems likely, that the flow through the bottleneck will be much the same whether the queue is 500 yards long or two miles long. This result holds even though the queue may be moving at an average speed of 20 miles per hour as against a free-flow speed of 40: under these conditions the addition of one car to the traffic flow will add 2 cars to the observed or “gross queue”, so that, even though only half the time spent in the queue is wasted (the other half being that necessary to cover the distance at the normal speed), twice as many cars are involved. On the other hand, if a car arrives at the end of the queue at 9:00, it may be held up for 20 minutes but cause an increase in the delay to cars behind it of only 10 minutes. And some cars will experience delay in the queue, but leave the queue at a point before the main bottleneck. These will cause no delay to cars routed via the bottleneck; any delay they may cause will be suffered by other branching traffic.

There is also the possibility of a “triggerneck” situation in which cars queuing for one bottleneck block the progress of traffic on other routes. In extreme cases a circular blocking can occur, and it is possible for the addition of a critical increment of traffic to bring traffic movement to a crawl, or even a complete stop, over a wide area.

Costs of Expanding Facilities
In the long run, also, costs of providing for additional rush-hour traffic in or near the centre of metropolitan areas can be very high. Construction costs of expressways have been increasing rapidly as the favourable locations are used up and interchanges and interconnections become more frequent and more elaborate, so that it is not unusual to find costs in the range of $3 to $5 million (U.S.) per lane-mile, as against figures of perhaps one-tenth of this amount for mileage constructed in less heavily built-up areas, and still lower costs for roads built to less exacting standards. Five million dollars per lane mile, at 6 per cent per year for interest, amortization and maintenance, spread over 1,500 vehicles per hour, 15 peak hours per week, 50 weeks per year, comes to 27 cents per vehicle mile, if the entire cost be allocated to the peak
traffic. Peak hour transit costs can also be quite high if peaks are sharp: a rail transit car can easily cost $100,000 which again at 6 per cent per year for interest and amortization, 60 seats per car, 250 weekdays per year, comes to 40 cents per seat per weekday, before the car has been moved an inch. In many cases the peak is so sharp in relation to the length of the round trip that a car at the margin makes only one revenue round trip per day, so that the cost of providing the car for the peak passenger is 40 cents per round trip (on an all-seated basis) – to which must be added operating costs, and possibly costs of the right of way.

**Bringing Costs Home to Users**

The remedy that occurs naturally to an economist is to charge prices or tolls which will effectively bring home to the individual user, when he decides how and to what extent he will use the various facilities available to him, just what the costs of the various alternatives are, so that he can choose in the light of a comparison of the benefit to himself and the cost to others. Until recently, to be sure, a proposal to do this for urban streets and highways would have been met with the objection that the cost of setting up a toll-booth on every street corner would be prohibitive, to say nothing of the interference that this in itself would cause to the smooth flow of traffic. But methods have been devised for assessing charges of this sort in reasonably close approximation to congestion costs, without causing undue expense or interference with traffic flow. They range from licence cards with tabs to be torn out to indicate the manner of use, through meters to be kept paid up by the insertion of destructible plastic tokens, to fully automated electronic scanner and response block systems which would permit the registered owner of the vehicle to be billed for his roadway use in much the same manner as he is billed for long-distance telephone calls. The mechanical problem, then, can be taken as solved, at least in principle. It remains to examine what kind of benefit could be expected to flow from such a system.

To begin with it is perhaps worth observing that sometimes a facility becomes worthless precisely because it is free. For example, where a high-speed or short-cut facility of limited capacity has as an alternative a more circuitous or slower route with ample capacity, free operation may mean that a queue builds up during heavy demand periods at the access to the faster facility until the time required for queuing and transit is equal to the transit time by the circuitous route; under these circumstances no one is able to make the trip any faster than if the faster route did not exist. Enlargement of the faster route may be a complete waste of money unless the route is enlarged sufficiently to take care of all the traffic that might offer, though of course an intermediate stage might bring benefits during off-peak periods. More generally, wherever congestion is likely to occur in the absence of pricing, a facility will be worth less as a free facility than if subjected to an appropriate level of toll. Of course, if the toll is excessive, use may easily be curtailed below the optimum level; but, even so, the revenues may be used to good advantage to diminish the adverse impact of other taxes.

Tolls designed to reflect congestion costs can be made to increase the usefulness of urban streets and highways in a number of ways, in addition to promoting a more economical modal split of traffic between private car, bus and rail transit. The result may often be little or no reduction in the actual flow of traffic over given facilities,
and sometimes even an increase, the benefits being through the reduction of delays.

To give a concrete illustration, suppose a bottleneck with a capacity of 4,000 cars per hour attracts, under current conditions, a volume of traffic of 4,400 cars per hour from 4.30 to 6.00, after which the traffic falls off to 2,800 cars per hour. Under these conditions the queue will build up linearly from nil at 4.30 to 600 cars at 6.00, and thereafter decline to zero again at 6.30. The maximum delay, for cars arriving at 6.00, will be 9 minutes, and the average delay 4.5 minutes; the number of cars delayed will be 8,000, or a total delay of 600 vehicle-hours. If evaluated at an average of $1.50 per vehicle-hour, this is a cost due to congestion of $900 per day, or for 250 days per year of $225,000.

The road-builder's way of dealing with this situation is to enlarge the bottleneck, and if this can be done for an outlay of, say, $2 million, the annual saving in congestion costs will constitute a return of better than 11 per cent on this investment—enough, in most circumstances, to justify the project in terms of cost-benefit analysis of the conventional type.

There are likely, however, to be more economical ways of meeting the situation. For example, if it were possible to select, from among those who would be a part of the peak traffic if left to their own devices, those who have alternatives that they regard as not very much inferior to the use of the congested facility, and to offer them a bonus for shifting to these alternatives, it might be possible to eliminate the queuing by paying a bonus of, say, 25 cents per vehicle trip to 600 drivers—a total cost of $150. If there were available an uncongested but more circuitous route taking 10 minutes longer, this would not be attractive under the original conditions in competition with the 9 minutes' maximum delay in traffic; but, for a person valuing his time at slightly less than the average figure of $1.50 per hour used above, it should be attractive, even as against the direct route without delay, if coupled with the 25 cents bonus. Other alternatives that might be thus encouraged would be the use of car pools, mass transit, shifting of the time of travel to before 4.30 or after 6.00, and even travel to other destinations. Thus payment of the bonus would leave everyone at least as well satisfied as if the bottleneck had been enlarged, at only a fraction of the cost of enlarging the bottleneck.

In practice, of course, the payment of such a bonus is not practical, if only because of the difficulty of ascertaining who would and who would not actually have used the bottleneck facility during the critical period in the absence of both bonus and congestion. But an alternative method exists for obtaining the same efficient pattern of utilization of resources, namely, the collection of a toll of 25 cents for each vehicle actually using the bottleneck between 4.30 and 6.00. The collection of a toll would indeed have the additional advantage over payment of the bonus that the revenue obtained would permit reductions in the rates of other taxes, such as the gasoline tax or even income taxes, and thereby reduce the harmful impact of these taxes on the efficient allocation of resources in other areas.

It is also possible that in some cases the cost of the bribe or bonus that would have to be paid to eliminate the queuing might be so great as to make the enlargement of the bottleneck the preferable alternative. Even here, however, the optimum solution will almost always involve some combination of toll and enlargement of the facility. Even where the enlargement, if it is to be made at all, must for technical reasons be made in a large enough increment to take care of all current traffic, a marginal
adaptation remains available in the selection of the time, relative to the secular growth of traffic, at which the improvement is to be made.

**Tolls for Greatest Efficiency in Queuing Situations**

In the idealized bottleneck-and-queue situation, where the capacity of the bottleneck is sharply defined, the ideal toll is clearly that which will just eliminate the queue. If traffic flows are stable and predictable from day to day, there will be some pattern of tolls over time that will accomplish this: the toll is zero whenever traffic is below the capacity of the bottleneck, and, whenever traffic at zero toll would exceed capacity, the toll is just sufficient to keep traffic at the capacity level without the formation of a queue. The situation is in effect that which is approached as a limit as $k$ goes to infinity in equation (1). To impose a toll at times when traffic is less than capacity would restrict the use of the facility to no purpose, assuming that below capacity no significant delay occurs. To allow a queue to develop causes wasteful delay without producing any increased flow through the bottleneck or enhancing the usefulness of the facility in any way. Thus the appropriate pattern of toll can be developed by trial and error for any given facility; the levels of toll so arrived at will in turn provide information essential for estimating when and to what extent expansion of the facility is justified.

In practice, of course, traffic is subject to random fluctuations, so that it is impossible to meet the above conditions completely (at least in the absence of some form of advance reservation system, which might be appropriate for long-distance plane travel but is clearly too costly for the relatively small decision-units involved in urban traffic). Therefore the toll must be set in some compromise manner, the costs involved in allowing occasional brief queues to develop being balanced against the costs of occasional under-utilization. The loss from under-utilization can be estimated fairly closely, using the effective toll at the moment as a benchmark, the assumption being that the forgone uses were valued by the potential users at slightly less than the toll. The queuing cost, however, cannot be estimated by any such benchmark, since vehicles caught in the queue will be a random sample of all traffic, ranging from unemployed and retired persons with time on their hands to busy professionals and executives, and busloads of miscellaneous people. The two sides of the balance are asymmetrical not only in the basis for measurement, but also in the pattern of variation, the loss from under-utilization containing a strong linear element, while the loss from queuing would tend to vary more nearly as the square of the amount of excess traffic. But, even if full allowance is made for uncertainty of this kind, trial and error should provide a practical procedure for arriving at the correct toll to within a very narrow margin.

**Tolls and Cost Allocation for Non-Queue Congestion**

Useful as the bottleneck-and-queue case is for purposes of illustration, the more general case is likely to be that in which the capacity limit is not a sharp corner, but in which the costs of delay mount gradually, though rapidly, as capacity flow is approached. Here it is a fairly straightforward matter, given a speed-flow curve for the facility in question, to compute the marginal increment of delay in vehicle min-
utes per increment of traffic in vehicle miles; the problem is one of finding an appropriate conversion factor to convert vehicle minutes of delay into a dollar figure. The problem is essentially no different, however, from that involved in any cost-benefit analysis undertaken for the purpose of determining whether or not a proposed construction project is justified.

Indeed, if one were to make the assumption of constant returns to scale in the provision of highway facilities (meaning that the cost of providing the facilities to make possible a given pattern of traffic flow at a given speed is proportional to the volume of traffic, so long as the volume varies without changing the proportions in which the traffic divides among various routes and times), there is a simple relation between tolls and costs in an optimal system. If facilities are built to optimum levels so as to minimize the combined costs of facilities and traffic delays, revenues from tolls set at levels to reflect marginal congestion costs will just cover the cost of the facility. If the facility is too large, revenues will fall short of covering total costs; if the facility is inadequate, revenues will exceed costs. This can be seen as follows:

As before, let \( z(q_t)/w \) be the average delay experienced at time \( t \) with traffic \( q_t \) over a facility of capacity (width) \( w \); \( x_t \) is the density of traffic relative to capacity. Let the cost of the facility itself be \( cw \), \( c \) being the cost per unit of capacity or width. Total cost to be minimized is then

\[
C = cw + \sum_t a q_t z(x_t)
\]

(5)

where \( a \) is the money value of a vehicle minute of delay. Minimizing \( C \) with respect to \( w \), keeping \( q_t \) constant, we get

\[
\frac{\partial C}{\partial w} = c + \sum_t a q_t \frac{dz}{dx_t} \frac{dx_t}{dw} = c + \sum_t a q_t \frac{dz}{dx_t} \left( -\frac{q_t}{w^2} \right) = 0.
\]

(6)

To determine the toll at time \( t \), we get the marginal cost with respect to an increment of traffic at \( t \):

\[
M_t = \frac{\partial C}{\partial q_t} = a \left[ z(x_t) + q_t \frac{dz}{dx_t} \left( -\frac{1}{w} \right) \right]
\]

(7)

of which the first term \( a z(x_t) \) represents the cost increment borne by the marginal vehicle itself, leaving for the appropriate toll

\[
r_t = a q_t \frac{dz}{dx_t} \frac{1}{w}
\]

(8)

which yields total toll revenues of

\[
R = \sum_t r_t q_t = \sum a q_t^2 \frac{1}{w} \frac{dz}{dx_t}
\]

(9)

From this and (6) we see that \( R = cw \).

In general, in order to know how the cost of a given facility should be distributed over traffic at various times one would have to know not only the shape of the function \( z(x) \), but also where, along this function, the pattern of traffic was to be located in the light of variations in \( c \) at different locations and the consequent degree of congestion to be tolerated in the light of the higher or lower cost of expansion of the facility. One of the conveniences afforded by the "constant elasticity" function given in equation (1), i.e.,

\[
z = a x^k
\]

(11)

is that after differentiation the variable \( w \) can be factored out, so that the relative
share of the costs to be borne by different traffic components remains the same for any given pattern of traffic, regardless of whether the corridor is in a low-cost location leading to the justification of the provision of relatively ample facilities and low levels of congestion, or in a high-cost one where the optimum level of congestion would be somewhat higher.

If the form of congestion function given in (11) is accepted, the rule for the variation of the toll becomes fairly simple: the toll rate should vary as the $k$th power of the traffic flow at a given moment. Applying this formula to given diurnal patterns of traffic flow can provide us with a somewhat loosely defined but nevertheless fairly useful parameter: this is the peak-to-average cost ratio, a factor by which average cost figures can be multiplied to get an approximation of what fairly typical peak cost figures should be. Peaking factors for existing patterns of traffic, which of course are without the benefit of the smoothing effect of congestion tolls, run from 3 to 5 on fairly typical metropolitan radial expressways; it is likely, however, that if the indicated tolls were actually applied the peaking factors might be somewhat less. In any event, the variation in the appropriate toll can be fairly severe: for $k = 4.5$, the appropriate toll for a traffic flow of 70 per cent of the peak would be only 20 per cent of the peak hour toll. The peak hours remain, even under this more sophisticated treatment, the dominant factor in the justification of additional facility construction.

**Peak Pricing of Transit Service**

The pricing of peak and off-peak mass transit services is perhaps in even worse shape, in both practice and theory, than pricing of highway services. As to practice, there is a strong element of perverseness in the typical commuter railroad fare structure, since the lowest fares are paid by users of monthly commutation tickets, riding predominantly during the rush hours; the next lowest fares are paid by riders on special off-peak "shoppers" round-trip tickets, who are predominantly riding in the direction of the dominant traffic; the highest one-way fares are paid in significantly higher proportions by those riding against the dominant flow, who are in the fullest sense "by-product" users.

The costing of transit service is not a simple matter, if a fully sophisticated allocation of equipment costs is to be made. Charges for right-of-way costs, in the case of bus service, can perhaps be treated in more or less the same manner as for other users, allowing of course for the greater contribution to congestion of the larger and clumsier vehicle. In a somewhat similar manner, the cost of rights of way and appurtenances for rail transit can be allocated at the margin to peak-hour use, the allocation being distributed in such a way as to even out the traffic appropriately; in most cases, however, excess capacity may exist more or less inevitably because of indivisibilities, and economies of scale may be such that only a part of the total cost of these facilities is marginally related to use, leaving a substantial intra-marginal residue. But I would like to defer discussion of the issue of economies of scale to a later point.

**Allocation of Equipment Costs**

The cost of equipment, however, is an element in which the economies of scale are relatively slight, but the allocation problem can be quite intricate. In the simple
case of an operation with a single route and no short-routing possibilities, the stock of equipment required can be determined by taking the round trip time and finding that period of this length within which traffic is at its maximum at the peak load point on the line. One might then suppose that the annual capital charges on the equipment might be assessed against traffic within this period, traffic outside this period being charged only the running costs. If this is done literally, however, and traffic is at all elastic, this may result in creating new peaks for the periods adjacent to that for which the charge is made, so that the equipment required would be greater than that indicated by the traffic in the period charged for. If this happens, the theoretical remedy, at least, is to divide the equipment carrying charges between the traffic on the trip just before the peak and that on the trip just after the peak (later by just one round-trip time) so that the traffic generated at the two times is equal. However, this is a refinement not likely to find much practical application.

A more typical case occurs where a number of routes are operated from a central terminal, either by turning back some runs short of the outermost terminus or by operating different routes in different directions. Sometimes each route is operated independently as a separate entity, but ordinarily there will be considerable advantage from integrating the operation in various ways, particularly if the peak times on the various routes do not coincide. One way of assigning equipment-carrying charges in such a case is to consider that all equipment storage takes place “virtually” at the central terminal, even though actual storage takes place at other points, whether for lack of space or to save deadhead mileage. An inbound morning run is then considered to “use” equipment from the time it would have had to leave the central terminal to be ready for the run until it reaches the central terminal again. Similarly an evening rush-hour trip uses equipment from the time of actual departure from the central terminal until the time the equipment could have returned. The schedule thus derived of virtual arrivals and departures from the central terminal will define a critical period during which equipment virtually held at the central terminal is limited to the minimum turnaround for equipment arriving and departing; before this period virtual departures exceeded arrivals, and after this period arrivals exceed virtual departures. Any trip requiring equipment to be away from the central terminal during the entire critical period would be charged for the carrying costs of the equipment; trips arriving before or leaving after the critical period would not be charged, and trips arriving or leaving during the critical period would be charged a *pro rata* amount. Thus peak charges would in effect cover a longer part of the day for long trips than for short; on the other hand, charges for the carrying cost of equipment would be a smaller amount per mile of travel for these longer trips.

**Depreciation according to Use and the Optimal Husbandry of Equipment**

This procedure for allocating costs assumes that capital charges are properly assessable entirely on a time basis rather than on a usage basis, and indeed in most accounting procedures equipment is depreciated over a lifetime specified in years. In fact, however, the time at which it becomes economical to retire and replace equipment does depend on the intensity of its use. While depreciation according to lapse of time may be adequate for fiscal purposes, it can be quite misleading both for purposes of planning efficient husbandry of equipment and for the computation of marginal cost for pricing purposes.
Utility companies have long operated their plants on the principle of using the latest, most efficient, generators for base load and older, less efficient, units for peaking purposes; the economy of so doing is fairly obvious. Where equipment varies fairly obviously and substantially in operating cost, it is likely that most transit operators would follow a similar procedure. However, where differences of this sort are minor or non-existent, the conventional wisdom seems rather to run in the direction of using equipment in a random manner so as to keep the entire fleet, or at least that part of it of a given vintage, more or less at the same level of use.

Actually, even where equipment is of identical operating characteristics, provided only that its useful life is limited in part by total mileage, it pays in terms of reduced capital charges to use some of the equipment more heavily than the rest, working towards a situation in which equipment is replaced at a more or less steady rate, the newest equipment being assigned to the heaviest runs and then demoted as it ages to less and less heavy runs. This remains true even for the "one hoss shay" case where operating characteristics do not deteriorate with age. The total number of units of equipment on hand is the same, but the stock of mileage tied up in a unit of equipment is exhausted more rapidly at first and more slowly later on, so that on average the stock of unused mileage is smaller than it would be if the equipment were rotated in random fashion.

For example, suppose an operation requiring 12 buses, which can be used 6 on a light schedule of 50,000 miles/year and 6 on a heavy schedule of 100,000 miles/year. If we start with a new fleet with a life per bus of 200,000 miles, and the buses are rotated to run 75,000 miles per year each, they will all be due for replacement after 12 years. If, however, half the buses are run more intensively 90,000 miles a year for 10 years, they will then be due for replacement, the remaining buses still having 90,000 miles of service left. If thereafter the new buses bought as replacement are run on the 100,000 mile schedule for the next 6 years, and the old buses put on the 50,000 mile schedule, no further replacements will be needed for 6 years, after which replacements will occur each 6 years for half the fleet. In effect the average date of replacement has been deferred one year and the average amount of unamortized capital has been reduced by about one-sixth.

More generally, in the absence of technological change, for equipment having a lifetime specified in terms of mileage the optimum husbandry for a steady state dictates that the schedules be arranged to produce the maximum possible concentration of usage (provided this does not involve other costs such as extra dead-head mileage) and the assignment of equipment to schedules in order of age of the equipment, the newest equipment to the heaviest schedules. The equipment rental to be charged for mileage operated at a given time of day will be the same for all equipment (there being by assumption no difference in operating cost). In linear programming fashion, each unit of equipment will have a shadow reservation price for its mileage at a given date, and on that date will be used for those hours for which the rental is above this reservation price. This shadow reservation price for any given piece of equipment will itself increase over time at the current rate of interest: in this way it becomes impossible to increase the discounted value of the rentals received for the using up of the stock of mileage embodied in any one vehicle by shifting the utilization of that mileage from one time to another. This implies that the ratio between the equipment

131
rental charges made for any two particular times of the day will be equal to the discount factor for an interval of time equal to the difference between the ages of the oldest equipment in use at the respective times.

On this basis, then, differentials between peak and off-peak costs will tend to be somewhat less extreme for modes that use equipment of relatively short life or for which the equipment costs are large in proportion to right-of-way costs. Moreover, consideration of this aspect tends to make the use of mass transit for peak travel compare better than otherwise with private automobile use, since off-peak use of a private car has much less influence in mitigating its high peak-use cost.

It is much more difficult to be definite about the peak-off-peak characteristics of labour costs of mass transit operations. I can only say at this stage that I feel fairly strongly that existing wage formulas and working rules are very far from being conducive to an optimum adaptation between the need of the travelling public for efficient and convenient service and the legitimate interest of employees in convenient working schedules and attractive working conditions. Many working rules, indeed, seem to be particularly onerous in their effect on the cost of peak-hour service. But it would take me too far afield to go into this in any detail.

Increasing Returns

There remains the problem of dealing adequately with increasing returns and economies of scale, especially since these vary significantly among the various modes. In the case of highways, the provision of facilities that cost more than the amounts paid for them by users is often justified on the ground that there are important economies of scale. It is important to bear in mind that this justification is valid primarily for lightly travelled roads where congestion does not occur, and for roads which provide greatly improved access to new areas, or significantly shorter routings for a substantial portion of the traffic they carry. One can claim economies of scale up to the point where we have four-lane highways, but beyond this there seems to be little, if any, reduction in the cost per vehicle mile of providing for traffic along a given alignment. Increasingly, highway construction in metropolitan areas consists of widening existing routes or providing additional parallel routes to provide more capacity for traffic already fairly well provided for at lower rates of flow. Mass transit, on the other hand, still remains very largely in the area of substantial economies of scale, not only in terms of lowered cost per passenger as intensity of utilization increases, but also—and this is perhaps more important—in terms of improvement in the frequency and variety of the service offered. Optimum utilization of mass transit therefore requires a subsidy sufficient to permit fares to be set at levels reflecting incremental cost, net, where appropriate, of credits for any improvement in the quality of the service that accompanies increases in volume. This is quite distinct from the case that might be made for such a subsidy in terms of income redistribution or as a counterpoise to subsidies afforded to competing modes, notably the private automobile.

Restricted-form Subsidies

In providing such a subsidy, however, care must be taken not to cause distortions by the form in which the subsidy is made. The case of New York provides a horrible
example, as the subsidy, being limited to providing capital free of charge, has induced the Transit Authority to embark on large construction projects of limited value, while paring operating and maintenance expenses to the bone, to the detriment of current services. Nor should any facile analogy of equal provision of free rights of way for motorist and for transit rider, resulting in a ten-cent-per-trip subsidy for the one and a fifty-cent-per-trip or more subsidy for the other, be allowed to delude one into thinking that this constitutes either even-handed justice or a promotion of an efficient modal split.

Optimum Allocation of Land to Transportation

Still another factor tending to bias decisions in the area of transportation planning is the way in which land taken for transportation purposes is handled. Much of the land used for transportation is treated as a free good, having been "dedicated" to this use at some time in the remote past; even where new land is taken for transportation use, it is seldom that its full value in the alternative use is taken into account. Sometimes this is because the land was already in the public domain, as parks or in other public use; even where the land was purchased from a private holder, the price will usually have been depressed by the discounting of future property taxes which a private owner would have had to pay. The corresponding property taxes are not ordinarily included as a cost when the land is used for highways.

A more subtle bias in the taking of land for transportation arises from the nature of urban commerce. In the perfectly competitive spatial model, every commodity and service has a shadow price at every geographical location, and prices for a given commodity vary among locations by amounts which never exceed its cost of transportation and are exactly equal to that cost over every interval over which the commodity is actually transported. Competitively determined land rentals correctly reflect the increased cost of transportation which would result from displacing an activity from its equilibrium location to a next best location, possibly with the bumping of other activities until an activity is removed to the peripheral land. Under these circumstances, in comparing alternative modes of transportation it would be correct to use the going rental value as a basis for comparison between modes using different quantities of land to accomplish the same transportation task.

Actual land rents, however, are determined in no such perfect market, and the deviations seem on the whole to be in a direction which would make the going rental value fall considerably below the figure for the cost of land that would, when used as the basis for choice among transportation modes, lead to optimal choices. There is a strong tendency for nominal prices of goods and services to be more or less uniform over a metropolitan area, costs of transportation being borne in some cases by the seller and in some cases by the buyer. Under these circumstances, when a firm is deciding upon the location at which to conduct an activity involving receipt or delivery of goods and services, it is likely to find that it will have to bear only a part of the difference in transportation costs incurred as a result of its choice, so that the premium it can afford to pay for the location providing the lower transportation cost will be considerably less than the full difference in the transportation cost affected by its choice. As a result the tendency will be for market-determined land rents to
be lower than the value which it would be appropriate to use as a guide for evaluating the desirability of taking land for transportation purposes. This undervaluation would be by as much as 50 per cent in the limiting case where all prices were blanketed over the metropolitan area and the transportation costs were divided at random between sender and receiver. This undervaluation is in addition to the depression of land values by the anticipation or property taxes which would have to be taken into account if the transportation agency failed to continue the payment of these taxes.

**Consequences of Peak Use Charges**

Optimum patterns of urban transportation facilities and their use are thus not easy to achieve. Indispensable for such optimization, however, is a system of pricing for the use of the various facilities which will come a lot closer than anything we now have to revealing the costs of alternative choices to the decision makers in such a way that their decisions will be made with proper consideration for these costs. Once the concept of such pricing is accepted, however, the way is open for far-reaching changes in patterns of transportation planning, leading to substantial economies in construction and greater convenience and facility in use.

One important consequence is likely to be a greater emphasis on circumferential routes and a reduction in the required capacity of expensive and amenity-destroying central arteries, with their tendency to block transverse movement. In the absence of pricing, it is often found that, if both circumferential and central artery routes are provided on a scale to keep congestion to low bounds, much through traffic is attracted to the central artery, which then must be enlarged to take care of this traffic if congestion is to be avoided. The alternatives are to enlarge the central artery at very great expense, to accept a degree of congestion on it so that traffic legitimately needing to travel to and from the centre is denied the possibility of a low-congestion route, and possibly to deliberately degrade the design of the central artery so as to make it unattractive to through traffic, in many cases thereby making it also less convenient for the local traffic with no alternative. Pricing makes it possible to divert traffic to an economical extent to the circumferential route, preserving a low-congestion route for those not served by the circumferential, while at the same time keeping the cost of the entire system to a minimum.

Another important consequence of the pricing of highway use is to permit fares on mass transit also to be varied in a rational manner in accordance with the costs of peak and off-peak service, so as to encourage the shifting of riding away from the peak. To some extent, to be sure, such spreading of the peak can be accomplished by the staggering of working hours, achieved by some form of public encouragement or inducement. The success of staggering is likely to be greatly enhanced, however, if, even after the congested conditions have been abated by the staggering, the temptation for each firm separately to return towards the peak schedule (since its employees would no longer be faced with crowded conditions) is counterbalanced by the consideration that higher fares would have to be paid. In the absence of a special charge for peak use of highway facilities, peak pricing on mass transit facilities might have the undesirable effect of diverting traffic to highways at just the time when this is least desirable from the standpoint of overall cost.

Another possible consequence is that in many situations the control of congestion
through pricing may so improve traffic conditions that bus service becomes more satisfactory, so that the more expensive forms of mass transit involving separate rights of way are less urgently needed. In effect, it may no longer be necessary to establish separate rights of way in order to preserve the mass transit vehicles from the delays and congestion at present encountered on general usage arteries. In many instances, given the availability of an improved alternative, construction of expensive special rights of way may no longer be worth while.

Still another consequence may be increased use of night-time pick-up and delivery in congested areas. At present, while the cost to the trucking firm of night-time delivery is often only a small fraction of that for daytime delivery, even allowing for premium wage rates, resistance of senders and consignees to the additional expense of overtime for their own personnel has kept night-time activity to a low level. This is in part due to the difficulty of adjusting freight tariffs to give appropriate concessions in relation to conditions of pick-up and delivery, but another reason is that under existing arrangements even private carriers can obtain only a fraction of the benefit of the total reduction in congestion that they would generate. Appropriate charges for the use of such areas during congested hours would go far to encourage more economical scheduling of goods handling, though to get the full benefit of this some means should be found of adjusting common-carrier freight charges to provide an incentive for consignors and consignees to co-operate in the change.

Appropriate charges for use by trucks of downtown city streets and highways may also bring about a rethinking of the relative roles of rail and road in the realm of terminal operations. The recent trend seems to have been to terminate the rail hauls at increasing distances from both origin and destination, giving over to trucks an increasing proportion of the delivery and pick-up task. In extreme cases we have the spectacle of trailers taken off flat-cars at a rail terminal on one side of town, trucked through the town to the other side and loaded on rail flat-cars again for the continuation of the routing. To some extent this is attributable to the breakdown of the railroad car-service rules and the inappropriateness of interline per diem charges for the use of equipment, but an important contributing factor is the failure to charge against the trucks an appropriate cost element representing either the congestion costs imposed on others or a scarcity rental for the street space they use. A more accurate appraisal of the cost of trucking on urban streets in terms of its total effect on other traffic, and on the amenities of the area, may lead to a shortening of the distance for which rail shipment is competitive with the truck, and to the use of rail terminals closer to the origin and destination of freight, so as to reduce the truck portion of the trip. To be sure, rail terminal operations, and especially switching between terminals, has always been considered one of the less efficient parts of railroad operation, but there are possibilities for automation and for the adjustment of freight tariffs in this area that have not as yet been exploited.

Impacts of Derived Information on Evaluation and Planning

Some of the methods of assessing charges for the use of streets produce as a by-product detailed information on traffic origins and destinations which will be invaluable in planning new facilities. But, even more important, the process of adjusting the charges in response to changes in traffic patterns will yield information as to
the value of the existing facilities and of possible enlargements which can be obtained in no other way with like the same degree of accuracy and reliability. A system of user charges related to congestion costs is essential if facilities are to be expanded intelligently.

**Minimizing the Impact of Errors in Planning**

But in any case construction of new facilities takes time, and facilities once in place last for even longer. Prediction of future traffic trends is at best an uncertain business, and it is almost inevitable that mistakes will be made. One of the advantages of pricing is that by suitable adjustment of the prices the loss resulting from mistakes in planning can be held to a minimum.

**Impacts on Public Finance**

Methods such as those suggested here to improve the efficiency of transportation will doubtless be opposed by many who regard any attempt to obtain money from highway users in new ways as an arbitrary and vindictive imposition – an infringement of what is at times referred to with some emotion as the “freedom of the road”. It is probably no accident that the term “freeway” has been applied to many traffic arteries. Yet it is not at all necessary, for the achievement of the objectives sought, for the new charges to be a net addition to total motor vehicle charges. Indeed, it may in some circumstances be appropriate to provide that any revenues derived from congestion tolls be used to reduce fuel and licence taxes. This would have the effect of bringing the charges related to rural highway use more nearly in line with costs. On the other hand, given the serious financial plight of many urban governments, it would perhaps be desirable to use added charges on urban vehicular users to provide an appropriate source of additional funds. This would on the one hand be free of the baneful economic impact of most other revenue sources, such as taxes on property improvements or sales taxes, and on the other constitute a local resource more conducive to economical use of the proceeds than grants from larger jurisdictions, the spending of which is more often decided upon without adequate consideration of the tax consequences.

But, whether the charges are made an additional source of general revenue or whether they are merely substituted for other forms of vehicle revenue, they are essential to the efficient use and development of transportation facilities. It is difficult to see any other means by which massive inefficiency in this area can be avoided.

*Columbia University, New York*