Zettel, R., and R. Carll. 1964. The Basic Theory of Efficiency Tolls: The Tolled, the Tolled Off, and the Un-Tolled. In *Highway Research Record* 47, HRB, National Research Council, Washington, D.C.

The Basic Theory of Efficiency Tolls

The Tolled, the Tolled-Off, and the Un-Tolled

RICHARD M. ZETTEL and RICHARD R. CARLL
Institute of Transportation and Traffic Engineering, University of California, Berkeley

•TRAFFIC CONGESTION is attracting the increasing attention of economists of theoretical and mathematical bent, especially those of the welfare school. Some among them have the temerity to suggest that the application of economic analysis to the congestion problem, followed by the use of sophisticated pricing techniques, could contribute to the solution of the congestion problem, increase highway efficiency, and promote the public welfare. Requiring, as it does, a large departure from both "traditional" reasoning and current practice, the idea has not been well received nor even understood in much of the noneconomic world. It is popularly regarded as an effort to discourage highway travel by pricing it out of the market by high tolls or prices.

This paper endeavors within small compass to explain the basic ideas dispassionately and to show that the underlying concepts are neither frivolous nor diabolical. In fact, an understanding of the reasoning behind them may contribute to everyone's understanding of traffic phenomena. At the same time, the paper raises certain questions which seem not to have been fully thought through, and which suggest that one might well have reservations about efficiency pricing, not only on broad social or political grounds, but also within the confines of rigorous economic analysis.

THE TOLLSTERS

Traffic congestion can, of course, be found on any highway facility and at any hour of the day. In the usual circumstance and in its more extreme form, however, congestion is a temporal affair occasioned by uneven traffic flows which reflect the rhythms in patterns of living. One form of peaking takes place on weekends and over holidays as people endeavor to satisfy their desires for recreation and social activity. Because of its very nature, perhaps because this sort of travel is regarded as part of leisure activity itself, this kind of congestion seems rather more acceptable than the second kind.

Congestion in the daily movement to and from work in urban areas is attracting much attention. It is directly related to the grim business of making a living. It takes place regularly—on the order of twice a day, 250 days per year. It is generally regarded as irksome. With increasing urbanization, it becomes ever more difficult and costly to do something about it by engineering works.

Traffic Flow

The man in the street is well aware that traffic congestion reduces the quality of highway travel, most obviously through reductions in speed. But it will be useful to state this formally. At low volumes, the average speed of traffic is independent of the flow rate and is determined by geometric design and speed laws. With increases in traffic flow, a point is reached where the average speed begins to drop slightly. As flow builds up, speed drops more rapidly. Also, the possibility of a breakdown (the result of stall or accident) in the smooth flow increases, causing both speed and traffic volume to fall. Without a breakdown, however, traffic builds up to a maximum level

Paper sponsored by Committee on Highway Taxation and Finance and Committee on Equitable Allocation of Highway Costs.

as speed decreases. The maximum is reached when the effect of increasing vehicle density on the road is no longer sufficient to offset the effect of a speed reduction. It will be seen that maximum traffic flow can only be brought about through reductions in speed—that is to say, increases in travel time. As a matter of convenient exposition, the emphasis in this paper is on time costs and savings; but it should be recognized that congestion may increase other user costs (operations and accidents, for example) and that reduction of such costs through highway improvements are properly regarded as highway benefits.

The Engineer's Approach to Congestion

The underlying rationale of toll rationing is not as foreign to the thinking of highway and traffic engineers as is often supposed. They are eternally concerned with capacity and efficiency problems. For example, highway designers deliberately fail to provide for free-flowing traffic under all anticipated traffic conditions. In the typical rural situation, they endeavor to design for the anticipated 30th-highest hourly volume of traffic during the design year, "because to design for the highest single-hour volume of traffic to be expected during an entire year would be wasteful" (1). In an urban situation, "the average of the 52 weekly peak hours is a suitable traffic volume for design. In a broad sense, this volume is close to the 26th-highest volume" (1). What this means is that maximum hourly volumes are expected to be greater than design hourly volumes, but this traffic can be carried at reduced levels of comfort and convenience to drivers, and, of course, with losses in time. One example shows a median speed of 32 mph on urban freeways at "possible" capacity, and 50 percent greater speed (48 mph) under "freely flowing" traffic conditions.

The design speeds developed by highway engineers intentionally provide delays to some highway users. They are "predicated upon acceptability to nearly all drivers; this is in recognition of the fact that it would be unwise economically to make provision for the foolhardy actions of irresponsibly fast drivers" (1).

Still another seemingly intuitive economic judgment is exercised. Design speeds are established at higher levels for rural freeways than for urban, and at higher levels for urban areas of moderate development than for such areas of concentrated development.

In the design and spacing of freeway interchanges efforts may be made to keep some traffic off freeways so that traffic on the freeways will have less congestion (smoother flow). In this connection, priority seems to be given to the longer-distance traffic, presumably because larger chunks of usable time will be saved to individual users.

Traffic engineers constantly work to increase the efficiency of operations on existing highway facilities. Traffic control systems are designed to improve traffic flows, often by increasing waiting times of some users in order to reduce waiting times of a larger number of users. A more drastic measure, practiced in a few instances and under serious discussion in many more, is the closing of selected freeway ramps at certain times in order to increase the efficiency of flows on the freeway system. There is also talk of rationing highway space by decree; for example, the prohibition of trucks on congested highway facilities during peak periods.

The foregoing engineering techniques are usually regarded either as "practical" economics or temporary expedients. The ultimate goal of engineers is to expand highway capacity to provide free-flowing traffic at "desired" speeds of most of the drivers during most of the year. Much of the economic justification for the expansion, however, is based on the finding that travel times will be reduced. These reductions in travel times, converted into money equivalents, are regarded as highway benefits to be set off against highway costs. This is but another way of saying that current congestion conditions occasion costs to highway users.

The Economist's Approach to Congestion

The "traditional" approach of the economist is not unlike that of the engineer. Although he might insist upon a more rigorous finding of benefit and a more exact accounting of costs than is found in many highway analyses, he would sanction highway expan-

sion as long as highway benefits exceed highway costs. If benefits are not sufficient to justify added costs, he might conclude that letting congestion develop is a better choice. But where congestion will exist either as a short-run, a long-run, or an inescapable condition, a third policy alternative may deserve consideration: to take positive measures to control the volume of road users in order to mitigate congestion.

The economic question concerning the wisdom of this course can be phrased in the same way as for highway expansion: Would the benefits of traffic restriction be greater than the costs created? The benefits at issue are similar to those occurring from highway expansion: by reducing traffic flow, "savings" in travel time, accidents, operating costs, etc., are provided for those who continue to use the highway. The traffic flow curve mentioned previously indicated that after a certain point, as flow increases, speeds are reduced and travel times increase. Each vehicle added to a road during a given time period increases the total travel time for all users, not merely by the amount of time of the added user, but by a fractional addition to the travel times of all other users. Thus, in terms familiar to the economist, congestion creates costs. As traffic flow rises, average travel times increase; the marginal time increment is greater than the actual time experienced by the marginal user. Take away the marginal user, and the average time for each of the remaining users is reduced. The saving of congestion costs is, of course, exactly the same as the "benefit" of reduced travel time due to highway expansion, with this exception: if traffic is restricted, instead of road capacity added, fewer users enjoy the benefits.

However, the costs to be compared with the benefits are altogether different. Instead of prices of land and other resources needed to provide highways, the cost arising from traffic restriction is the loss to users who must be prevented or induced not to use a congested road. The amount of the loss depends on what alternatives are available to those who are diverted.

Some motorists might be simply diverted to other highway routes. Others could choose to use public transportation, to share rides with other motorists, to select different trip destinations, or simply to travel less. Some might be persuaded to change their times of travel to uncongested hours of the day. Downtown shoppers could be induced to avoid the afternoon rush hour, if traffic congestion has not already discouraged them. The more difficult goal, but at the same time the more meaningful, would be to stimulate a staggering of working hours. On a more drastic scale there might be substantial changes in ways of life: homeowners might change their residences, employed persons their jobs, employers the places of employment.

In any case, the purpose of restricting traffic would be to produce a net gain—an excess of benefits over losses. On a congested road, the object would be to provide improved highway service for some of the users, at the cost of denying the road's service altogether for the remainder. The merits of this policy might be weighed against the net gain, if any, from enlarging the road system, or against the virtues of doing nothing. The alternative which produced the greatest return (or the least loss) would be favored.

Of course, the comparison of costs with the benefits of traffic restriction is far more difficult than in the case of analyzing road improvement projects. In either instance, the benefits of service improvement can be cautiously estimated. But on the cost side, traffic restriction offers no solid measure such as the actual costs of right-of-way acquisition and highway construction. The losses incident to diverting traffic can take the many forms previously mentioned. If vehicles were only diverted to other highways, then savings in time, accidents, and operating costs on a decongested road might be set off directly against increases in these same factors for motorists who were induced to use less satisfactory routes. But how would the cost of car pooling, use of public transit, travel to different destinations, or not traveling at all, be balanced against the gains from decongestion?

The Toll Proposal

Economists, long familiar with the idea of equating marginal cost and price, have suggested pricing as the appropriate way of restricting traffic. A toll charge upon a congested road, which tested the value of the service to users by their willingness to

pay, would sort out the motorists which had the least to lose by being removed. The toll would be set equal to what was believed to be the value of the marginal benefit produced by removing a user. Any motorist who thought his loss from being diverted would be greater than the size of the toll would pay the price. Thus, the loss to any motorist who refused to pay could be considered less than the size of the gain due to his absence. To complete its virtues, the toll method would do the job of restriction by voluntary action on the part of motorists.

The novel contribution of this proposal is that the prices, which are sometimes called "efficiency tolls," would be unrelated to the actual cost of highways. The revenue earned by the tolls would be no more than an incidental by-product of the pricing program. This itself would be a substantial departure from familiar highway tax policy, but some economists have argued that modern circumstances compel a revision of pricing objectives. Instead of earning revenue to build more roads, the primary purpose of pricing should be, it is argued, to achieve optimum utilization of the existing highway system, by controlling and directing the volume of use. For example, Vickrey (2) sees the need to avoid "wasteful expenditure on excessively extensive facilities." Buchanan (3) contends that "the reduction of congestion by road expansion can never represent an adequate solution in major urban centers." It appears, he says, that "with relatively little change or modification the system of highway user taxation now employed could be made to approach one which would achieve efficient operation of the existing highway structure."

Basic Theory of Efficiency Tolls

On first impression, it seems quite unreasonable to collect tolls that are unrelated to actual highway costs. However, the initial capital costs of any existing highway, for which use is to be manipulated by tolls, are regarded as sunk; in this sense, the highway is costless. The use of any toll revenue that may be forthcoming is irrelevant to the pure theory, except that it cannot be given back directly to those who pay it or no rationing would take place. Perhaps what bothers many people most about the efficiency toll proposal, in its pure form, is this apparently pointless production of revenues unneeded to meet expenses.

However, economists invoke a theory of "social cost" to justify rationing charges. Instead of normal highway expenses, they look to the costs of congestion which users inflict upon each other as a cost basis for the prices. They argue that when optimum utilization of highways is the pricing objective, congestion costs should be compensated with user charges; just as actual highway costs should be defrayed by user taxes when expanding the road system is the objective. A motorist, it is said, should be made aware of all costs involved in his decision to drive. The individual motorist is aware only of his own travel time on a congested highway, not the increased time that his presence causes others. If a money value is given to the time of road users, it may be said that a motorist inflicts "social costs" upon other motorists by reason of his presence.

No motorist should decide to use a road, the theory continues, unless the value of his trip suffices to cover both the "private" and "social" costs associated with his use. Therefore, a toll charge which reflects social cost to him provides a fair test of how badly he really wants to travel.

The current vogue of this theory is apparent to any transportation student. Analysts have claimed that automobile users are not paying the full cost of their highway use; and that this fact especially distorts the urban transport scene in favor of the automobile. It is argued that costs per user go up rather than down with growing utilization of highways. It is even suggested that the failure to price the social costs of congestion amounts to an outright subsidy to motorists.

It has been observed, in connection with the toll proposal, that the price would be set according to some estimate of marginal benefit. Is this pricing basis the same as the one which asserts that highway users are obliged to pay for social costs? On the surface, the concepts seem to be the same. In terms of social cost, marginal benefit refers to the congestion cost eliminated by removing one vehicle; marginal social cost refers to the congestion cost occasioned by adding a vehicle. At any given volume of

traffic, marginal benefit and marginal social cost are equal, because one is simply the reverse of the other. Thus, setting a price to cover marginal social cost would appear to be the same as making it equal to marginal benefit.

But it should be noted that a price imposed to produce a benefit by means of restriction is not identical to a price charged to cover a cost as an obligation. The difference is subtle but important. Marginal social cost indicates what the benefit would be to users of a congested highway if one of their number were removed. But that benefit will not necessarily result from setting a toll charge equal to marginal social cost. Suppose that all users were willing to pay such a toll. No benefit would result. Or suppose that on a highly congested highway a very large toll brought but a very small response. In this case, motorists would be asked to pay a large price in return for a small benefit.

Those who accept the social cost theory must begin with an assumption that only a congestion-free highway should be toll-free, and that any degree of traffic congestion on highways gives sufficient justification for charging restrictive prices. Beesley and Roth (4), for example, find that there is reason to restrain the use of highways if two conditions are found: (1) that "use of space by one unit of traffic affects the terms on which other units can use it (which is their way of defining "congestion"), and (2) "when each road user makes his own separate decision when and to what extent to use roads," (which is the typical behavior of an automobile driver). If, then, a toll charge set to equal marginal benefit is very much larger than the average benefit produced to each user who pays it—if indeed any benefit is produced—should there be any complaints? It might be concluded that the toll should have been charged anyhow, in order to pay for social costs. But what would be the attitude of the tolled? Does one know that these users, as individuals, are actually better off than they were before the toll? Would they rather have suffered the congestion (and time losses), and saved the toll, which represents hard purchasing power extracted from their pockets? The fact that they are willing to pay the toll gives no answer. Although it is obvious that they prefer the tolled facility to any alternative, it can scarcely be asserted that they willingly pay because of the benefits of time saving they are supposed to enjoy.

The "Costless" Transfer.—Fortunately for the economists at this point, they have a coup de grace to administer. Let none of the tolled harbor the thought that he might be better off with a little more congestion and no toll to pay. Leaving aside the toll collection process itself (for purposes of discussion it is assumed throughout that the technical problems of toll collection can be solved at favorable cost), no resources have been used and society has lost nothing. The tolls simply transfer income from users to the government. It will be redistributed in one manner or another: perhaps general taxes will be cut; perhaps general governmental services will be improved. All that will have occurred is a "costless" transfer of income.

Such is the happy outcome that "social costs" have been converted into community income with no exhaustion of real resources. If there has been no response to charging a toll, no individual has gained, but society has not lost. If there is any response, the toll payers are unequivocably better off. The toll charge, as one advocate blithely remarks (5) is "a transfer payment involving no net loss to the community (i.e., the rest of the community gains what the vehicle loses) and therefore can be disregarded." Because the toll charge is seen merely as an instrument to bring about a reduction in traffic which is thought to be beneficial—and an instrument which tends to remove the right users—writers have been wont to hurry nervously over this subject, as though it were a minor irritation, leaving to politicians the task of explaining to toll-payers the "costlessness" of their loss of income. (Politicians might also have to explain why users should pay "costless" money for social costs they inflict upon others, but should receive no compensation for the social costs that others inflict upon them.)

The Marginal Cost.—Actually, a more thoughtful view of the matter shows that the social cost theory need not regard toll charges as being "costless," but rather as legitimate prices based on marginal cost which ought to be charged. Welfare economics suggests that the individual motorist should not be permitted the option of choosing more congestion and no toll charge, except in the far-fetched case that no user put any value on faster, safer, and more confortable travel. The existence of congestion

creates the obligation to pay. Obviously, this theory involves much more than the idea that pricing is an apt way to restrict traffic and control highway use. The economic principle on which the theory rests is fundamental: price should be based on marginal cost, to achieve the best organization of the economy. With marginal cost prices, nobody will consume a commodity that he values less than it costs to provide, and the value of all goods and services rendered by the economy will exceed by the largest margin possible the cost of supplying them. This principle applies as much to highways as to any other activity where price policy is in question. If social costs are clearly a consequence of added road use, the marginal cost criterion indicates that the benefit to each user should be as much as the cost imposed on fellow users (and any others).

All of this accords with marginal theory; but the marginal cost criterion is not blindly accepted as a welfare principle in the economic world, and some of the questions concerning it are important here. Marginal cost is considered mainly as an explanation of "short-run" pricing decisions. Prices are set at the marginal cost level to achieve the best use of facilities. If prices do not influence use, the pricing basis is indeterminate, except as a guide to "long-run" economic behavior. Marginal cost prices that produce revenues greater than total costs will stimulate plant expansion and attract entry into the enterprise. Marginal cost prices which result in deficits will induce contraction and withdrawal. In short, time is expected to correct over-use or under-use; at least there will be a constant tendency toward best use.

The case for marginal cost prices that produce revenue surpluses or deficits over the long run is much less certain, and is the subject of intense economic debate. Enterprises with marginal costs below average costs would have to be subsidized to remain active. Activities with high marginal costs (greater than average costs) would earn excess returns. Marshall's suggestion of a "tax-and-bounty" system of pricing, whereby excess returns are preempted by the community for the purpose of subsidizing long-run deficits, is relevant to the pricing of the lightly and heavily used links in a highway system.

In a word, marginal cost is not an infallible theoretical rule for pricing. It must be demonstrated that a gain in social welfare will be obtained from charging prices which produce revenue surpluses or deficits. The highway case calls for particularly close analysis; this is not a case dealing with the simple substitution of one good or service for another. Nor is one dealing in this context with "costs" that have been incurred by any but the users themselves. There must be consideration of whether a positive gain might actually be achieved by pricing these costs.

FOR WHOM THE TOLLS TOLL

How might the results of a rationing toll policy be evaluated in practice? Let it be supposed that on a congested highway a toll charge has been levied and free-flowing traffic movement restored. The highway authority desires to learn whether a net gain has been achieved. How would it reckon up the gains and losses that had occurred?

At the outset, it is assumed specifically that there will be losses which deserve consideration. The attitude of those who are panicked by the growing flood of automobile traffic into thinking that any rationing of traffic is good per se is not adopted. Rather, it is insisted that the losses of those forced by tolls to consider inferior alternates be balanced against the benefits to those who remain and pay tolls.

The Tolled

Who are the gainers from restrictive tolls? In general it would be desired that those tolled would stand to benefit most from the improvement of road service—that is, they would place the higher values on time savings, comfort and convenience, safety, and other service factors. More exactly, these users would include the following groups:

1. Motorists who had a pressing need to move rapidly but were unable to foresee the need to take a trip. In this category would be emergency calls for ambulances and fire trucks and other such irregular traffic desires.
2. Motor vehicles carrying larger numbers of passengers. Express buses would have to be credited with a high value on rapid movement because the value would

represent the total demand of all riders in the bus. Car pools would also be in this class.

3. Motorists willing to pay for "luxury" service. Many students have felt increasingly that the ownership and operation of automobiles in metropolitan areas is a costly privilege, that this privilege might be reserved for those who can afford to pay for exclusiveness.

4. Motorists who have no convenient choice but to use a congested road. Because they are "captive" users to a large degree, they cannot easily disengage themselves if road service is not to their liking; thus, they may have a special interest in good service.

This list is not complete, but it suggests the nature of the gains. It points up that there may be considerable variation in the way that individual users would value the gains, which in turn depend on different motives for traveling. In the usual analysis of highway benefits, average values are given to highway service factors, which permits the evaluation to proceed but is not altogether satisfactory. This becomes evident when an attempt is made to value the gains of traffic restriction. The problem might be solved if the toll charge itself measured the gain, but it does not.

Most analyses of toll rationing concentrate attention on the comparison of travel time "before and after" decongestion. As a typical example, one may look to Tanner's verbal illustration (5, p. 1). On a given facility, 1,100 users during a given time period are reduced to 800 by a sixpence toll. The 1,100 traveled only 10 mph; the 800 travel 15 mph. Each user's individual time cost for the given trip was assumed to be 10 pence when travel was at 10 mph. For the 800 who remain to pay the toll, the individual time cost is reduced to 6.67 pence. Each user saves 3.33 pence, for which he pays a sixpence toll.

For the moment the fact that this appears to be a bad bargain for the toll payer is ignored. Instead a look is taken at the 300 who would not pay the toll. Each of these is unwilling to invest the 6.67 pence in individual time cost, which he would incur by using the decongested road and the sixpence toll. Thus, their value on using the road, although high enough to have made them endure 10 pence of travel time cost, is less than 12.67 pence.

Several points concerning the valuation of time, as used here, deserve comment. First, it may be noted that the time dealt with is generally regarded as leisure, that is to say nonworking time. This seems reasonable because such time is probably far more significant than working time in the usual congestion situation; it is also far more difficult to deal with in economic terms.

The second point is that in the typical analysis time values are established independently. One method is to equate leisure time with take-home wage rates. As something of a digression, the authors raise an eyebrow at this.

With working hours set institutionally for the most part, and with fairly short workdays and workweeks, there would seem to be reasonable doubt that individuals value nonworking time at wage rates. Witness the vast expansion of "do-it-yourself" work which involves an investment of individual time very often at returns much less than the do-it-yourselfer's wage rate. It may be said that time used in such activity is not lost but is part of leisure itself. But, by the same token, it could be argued that transporting one's self in a motor vehicle is a "do-it-yourself" activity, and the time used is not necessarily pure waste.

Time spent in highway travel might be put to constructive use. It does afford change. It can furnish relaxation. It can be part of leisure. It has even been suggested that the driving task itself may be "psychologically rewarding." At least, allowance should be made for the possibility that travel time is less onerous than working time. What this amounts to is that "social costs" of congestion in monetary terms may be much less than the magnitudes often estimated. But use of lower time values would not dispose of other fundamental questions.

Certainly, returning to Tanner's (5) example, none of the users who are diverted from the road can value the time saving—made possible by their decisions—by more than the amount of the toll. Otherwise, they would quickly return. Of those who pay the toll, it is known that their value upon using the road is 12.67 pence or more. This much can be said. But it does not follow from this fact that the value of the time sav-

ing for them is larger than the toll and thus explains why they remain to pay. Indeed, by assuming that all users have the same value upon time saving, a peculiar dilemma is posed: If the toll were equal to or less than the value of the time saving for every user, why would any one of them shift to an alternative? But, of course, if none of them shifted there would be no time saving. And if the toll were greater than the value of any possible time saving to any individual user, why would any one of them remain and pay toll? For it has been necessary to worsen the lot of all users, by 2.67 pence each, in order to shift any of them.

Some users, of course, would continue to pay because the utility of the trip to them is greater than the utility of any other alternative, even after cost has risen a net 2.67 pence. In other terms, it might be said that the net result of the toll (the algebraic sum of the time saving and the money payment), even though negative, is still not sufficient to offset the comparative disutility of the next alternative for anyone willing to pay the toll.

Average time saving values have been used mainly for convenience in exposition. It is recognized in reality that different users have different time values, and hence will respond differently to the imposition of tolls. Thus, it might be imagined that those who had the higher value upon the benefits would be more willing to pay the tolls than those with lower time values—as indeed they would if time savings were the only factor in evaluating the decision to take a trip. One who would pay $1.00 to save 10 minutes could be said to have a valuation on time equal to, or more than, $0.10 per minute. One who would not pay the toll would place a valuation on his time of less than $0.10 per minute. If an average value of $0.10 per minute sufficed to establish the justification for restricting traffic, it would not matter by how much the tollpayer's time saving values exceeded the average—as long as it was thought that they were equal to, or exceeded, the average. (This is simple economic reasoning. No sellers of a product can tell, or need to know, just how much a product is worth to buyers. What matters is that the product be worth enough to equal the price charged. In economic jargon, whatever excess there may be in the value over the price is the consumer's "surplus.")

When the toll proposal has been expounded in general terms, it has usually been implied that the tollpayers will be those more eager for time saving. An oft-made comparison, illustrating the idea, is that of the hard-pressed business executive, whose time is valued in dollars per minute, and the carefree joy-rider, who jogs along mindless of the delays he creates for others and undisturbed by the slow pace. The businessman would gladly pay a price to expel this nuisance blocking his way. The other driver, because he counts little the cost of being slowed by congestion, would be equally unlikely to pay much for the privilege of going faster and would balk at a price. The toll, in this instance, would retain on the road the motorist who had the persuasive reason for wanting decongestion, and would remove the motorist who had no urgent reason to take his trip. The comparison links a high trip utility with a high value of time, and a low trip utility with a low value of time.

But the executive might find himself in competition with another class of motorist, one whose desire and urgency for using the highway compared with his own, but who would have little desire to travel faster if he had to pay for it. The very existence of intense traffic congestion suggests that this type of individual would be present in far greater numbers than joy-riders. The tenacity of the typical urban commuter, driving in traffic congestion that would be discouraging to almost any other travel purpose, attests to the utility of his trip. And this is small wonder, for the journey to work enables him to earn his living. Now the commuter might have a value upon time saving that was above or below the average; his evident willingness to pay a toll would by itself give no clue to this. What can be said for certain is that, if there is no convenient alternative to the highway for going to and from work, it would take an imposingly high toll to make much of a dent in rush-hour traffic. The business executive might give up in disgust, rearrange his hours of work, or even move his office to the suburbs before paying tolls that provided little or no time savings.

Only one case has been found where, considering the congested road itself and the toll charge, the price gives a reliable indication of the value of the gains to the tollpayers. Any driver who has not previously used the highway, and appears when a toll

is charged and traffic volume is reduced, must value the benefits of decongestion at least as much as the toll. A shopper, for example, may have been induced by congestion to use suburban rather than downtown stores. If he is persuaded to travel downtown to shop because congestion is reduced, the gain to him equals or exceeds the price he pays.

The essential reason why the rationing toll fails to reveal much about the value of the gains from traffic restriction is that it does not ordinarily offer the tollpayer a true "before-and-after" evaluation. If a 10-min time saving is achieved by charging a $1.00 toll, the typical motorist does not have the easy option of paying to save the time, or not paying and receiving no benefit. Only if there is an alternate travel route which required the same travel time as the congested road before the toll was charged, and afterwards takes 10 min longer, does the toll give an "either-or" choice. If the alternate to paying roughly simulates the "before" situation, the motorist can exercise his opinion about the road service improvement, for if he elects not to pay the toll, he will at least be no worse off than before it was charged.

Otherwise, the benefit values represent an independent judgment of the analyst. And there is danger that his discretion will establish a self-fulfilling condition. The higher the price it may take to remove a vehicle from a congested road, the greater may appear to be the social cost of the congestion and the greater the saving in this cost from reducing traffic; hence, the more justifiable the toll.

The Tolled-Off

There is, in fact, little said in most analyses of toll rationing about those who choose not to pay the toll. It has been noted that their valuation on the trip via the tolled highway must be less than those who pay the toll. It also has been indicated that they have found an alternative. Now, it must be noted that the tolled-off user prefers the facility with congestion to any alternative he selects. The toll has motivated him to use a less desirable alternative and to incur a loss. Although it might be concluded that the loss cannot be larger than the amount of the toll payment (otherwise the user would pay and not be diverted), the main question is whether the loss is less than, or exceeds, the benefits of decongestion.

The following considers a number of alternative choices for those "tolled off" the highway:

1. The Two-Road Network.—This provides the simplest and most defensible illustion of the rationing toll, and it is used frequently to demonstrate the basic theory.

All users travel between points A and B, which are served by two roads, one superior and one inferior. The superior route becomes clogged with congestion. By diverting some users to the inferior highway, a saving in travel time for those remaining can be achieved. If this saving is greater than the increase in time for those diverted, a net reduction of total travel time for all users results. A toll charge will offer a direct choice to users: paying to go faster, or not paying and going slower. Therefore, the toll guarantees that the tollpayers value time saving as much or more than those who are diverted, and any reduction in total travel time for all users can be considered an unequivocal gain. Allowing for different time valuations among users, a toll that caused an increase in aggregate travel time might also be shown to produce a net gain, but the problem then becomes rather more complicated.

The example, however, can be further refined to remove the last shadow of doubt about whether a net benefit is possible. Let it be assumed traffic on the superior highway, before the toll restraint is introduced, increases in volume until travel time is the same as time on the inferior route. Then, anyone who is tolled off the superior highway has not suffered (here it is assumed that there is no congestion problem on the other road), and all those who pay the toll gain. Moreover, the toll charge, because it need be no higher than necessary to keep the diverted users from coming back to the superior highway once a reduction in congestion has been effected, will not be any larger than the value of the gain to the tollpayer. And if the tolled, value time more highly than the tolled off, they will come out ahead, even after the toll is subtracted from their gain.

This is the classic case of a clear gain in economic welfare: nobody is made worse off, some may be made better off, and the "community" has an income which was not previously available. The case is very stringently drawn, but it can serve as a benchmark from which to view other possibilities.

When the diverted users must incur losses, a reduction in aggregate travel time still indicates that a gain has resulted, but the toll charge cannot be so easily dismissed. Suppose that the congested superior route requires 20 min travel time between A and B, the inferior alternate 30 min. and that a $0.30 toll diverts enough drivers to cut travel time on the superior highway to 15 min. It can be assumed that those who pay the toll value time at $0.02 per minute or more; otherwise, they wouldn't pay $0.30 cents to save 15 min. But they would have to value time at least at $0.06 per minute in order to think the toll payment worthwhile, for they are asked to pay $0.30 to save 5 min over the congested situation. The worse that the alternate choice is in relation to the pretoll situation, the more significant this consideration. For example, suppose the alternate route to require 40 instead of 30 min, and the toll charge to be $0.50 instead of $0.30. The toll informs us that, as before, motorists who pay it must value time at $0.02 per minute or more, because paying the $0.50 toll saves 25 min over using the inferior route. But to be satisfied with having the toll at all, the tollpayers would have to value time at $0.10 per minute, at least five times the amount which their willingness to pay, as measured by assumed time values, seemed to indicate.

Of course, the toll itself has been discounted as a cost because the redistribution of income effected by the tolls is said to benefit all in the community, including those who have chosen alternatives to the tolled facility. The worse the alternatives, the more that this assumption—sure to be questioned as a practical matter—must be relied upon. (Some cities will argue that the whole procedure is inequitable in that it violates ability-to-pay concepts. They will argue that the basic ideas require acceptance of the propriety of existing distributions of income. Most assuredly, the efficiency toll proposal has nonequalitarian overtones. But for purposes of discussion at this point the proposition is here being treated as an impersonal pricing problem.) But there is a further word of comfort for those tolled off the facility. In principle, they might be compensated for their losses out of the income from the tolls. In fact, the tolls required to divert a given number of motorists would not have to be so high if some of them could be "bribed off" with compensation. It can be proved, as a theoretical point, that if tolls in the two-road-network case are used in this manner, then the toll can be made equal to, or less than, the value of benefit to each user who pays, and every motorist's welfare is advanced: the tollpayers receive value greater than price, the losers are compensated more than the amount of their loss. Only the "community" is left out.

The mechanics of compensation are less easily imagined, but the principle potentially has broad application. For example, high tolls could be charged on heavily used highways and low tolls on lightly used roads, which might stimulate some evening out of traffic flow while bringing in the same amount of revenue as with an average user charge assessed equally upon all road use. Or the "bribery" could be indirect. Toll revenue could be used to subsidize rapid transit, which would reduce the losses of those who chose that alternative to the tolled facilities, while at the same time reducing the money payments of the motorists.

2. The Multi-Road Network.—There is no shortage of intellectual apparatus to demonstrate the toll argument as a problem of network flow. Wardrop's (6) fundamental paper some years ago recognized that there are two different distributions of traffic which correspond to (a) the action of all motorists seeking the minimum travel time route between any two points, and (b) the minimum average journey time per motorist, which is the same as the minimum aggregate journey time consumed by all motorists collectively. The first of these criteria explains the network flow that is likely to result from the free interaction of motorists. The second might be brought about by an appropriate toll policy. If travel time is accepted as the measure of efficiency in road use, the second condition could be said to represent maximum efficiency in the use of a highway system.

The multi-road network case differs from the two-road network in that not all users on a given highway are moving between the same origin and destination, so that the same

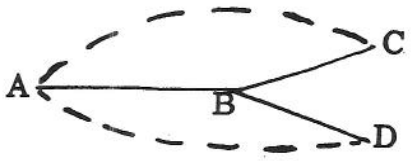

Figure 1.

alternate routes are not equally available. On a congested road, AB (Fig. 1), some motorists are traveling between A and C and have alternate route AC. Others move between A and D and have the alternate AD. Thus, the utility to users of taking a trip on AB will reflect the quality of the alternates. If it is assumed that AC is a very poor alternate and AD is relatively more satisfactory, and if all users are credited with the same value on time saving, those going between A and C would be relatively more disposed to pay a toll charge for AB than those going between A and D. A reduction in total travel time could be achieved if diverting some users moving on ABD to alternate AD would cause less increase in travel time than the saving on AB.

In this circumstance, however, the toll does not guarantee that those who refuse to pay it have the lower value on time saving. The fact that one user's alternate to the best route takes 1 hr longer, and another user's alternate route takes 15 min longer, reveals nothing about how each would value a saving of 10 min on the main route.

Thus, it is possible to suggest that a toll charge could divert users who had the higher value on time saving but the better alternative choice, while retaining users who had a lower value on time saving but a poor alternate. If this is true, the vehicle flow which minimizes aggregate travel time in a road network does not necessarily maximize value in excess of cost.

The foregoing is but a theoretical possibility. However, it illustrates how the evaluation of losses may be complicated by demand factors not directly related to the value of gaining or losing travel time.

3. Smoothing Out the Peak.—Motorists diverted to uncongested time periods, although continuing to use the same route, are asked to accept an inferior time to travel, rather than an inferior route.

The great advantage of network flow evaluation is that gains and losses are measured in identical units. Thus, a simple comparison of physical pluses and minuses may suffice, without having recourse to economic valuation. If time saving, for example, exceeds time loss, it is enough to know that the unit value of time to the savers equals or exceeds the unit value of time to the losers.

For alternatives other than inferior routes, one cannot escape placing some kind of valuation, no matter how general, on both the benefits and the losses so they may be compared. If the pricing objective is to divert users to the off-peak, there is no ready way to estimate the losses occasioned except in terms of the size of the toll charge. Hence, the benefits must also be valued in terms of money.

The success of a rationing toll in achieving this objective depends on the existence of drivers traveling during the peak who have a relatively low value on time delays. A low congestion cost explains why such users tolerate traffic conditions during rush periods and would be responsive to a toll.

4. Car Pooling.—Perhaps the best general case for the rationing toll is the inducement it would provide for highway users to pool their rides. Motorists might be more willing to share each other's vehicles if they could be certain that they would benefit from doing so. The toll charge would provide the necessary instrument to promote the result, and all could partake of the benefits to some degree. However, the losses in this case are almost all in intangible service factors. These include the irritations of conforming to group schedules, the lack of privacy, tensions, crowding, etc., none of which have a quantitative dimension.

5. Diversion to Public Transportation.—The amount of the loss depends on how good a substitute transit is for the automobile, in the mind of the motorist. The choice between modes depends on the proximity of transit routes to travel origins and destinations, the convenience of door-to-door service afforded by the automobile, the reliability of scheduled transport, the value of personal privacy, etc.

It would be possible, of course, to ignore all such service intangibles and consider only the time and cost of public transport. An attempt would be made to determine whether aggregate time and cost could be reduced by a shift of people from automobiles

to trains and buses. Some analysts have urged that making the most efficient use of the transportation system, rather than the road system alone, is the proper objective for pricing, claiming that the "movement of people" may constitute a "higher-level criterion" for defining an optimal solution than does the movement of vehicles.

6. Diversion to Different Destinations.—To grasp this situation, one might rule out all other transport choices, such as the diversion to other highway routes, to different time periods, to riding pools, or to other travel modes. It is assumed that the motorist must either pay the toll or take a trip to or from a different place than he had planned.

Within these limits, the willingness of the user to pay reflects his interest in reaching a particular destination, or departing from a particular origin, compared with others. All manner of theoretical possibilities can be imagined. Giving each user the same value upon time saving, the toll will divert those whose interest is least urgent in getting to a particular place, but the toll tells nothing about how the loss compares with the benefit until an exact value is specified for time saving. It is possible that none of the toll-paying group cares anything about the reductions in traffic congestion. On the other hand, the urgency to reach a particular destination may be directly related to the need for speed, as in the case of a fire engine. Again, it is not impossible that those users with a higher value upon reaching their destinations would be diverted. This could happen if those with the lower trip utility nevertheless thought themselves so benefitted by an improvement in road service that they would pay the toll charge which the others would not.

In short, the losses due to a revision of origins and destinations are fairly indeterminate, apart from the toll. One extremely restrictive assumption is worth mentioning. If it is said that the purpose of a trip is satisfied equally well at any of several destinations, except for travel considerations, then the loss due to choosing an inferior destination is the increase in travel time and cost of going to the alternative.

7. Reduced Trip-Making.—There is no way known of estimating losses in utility when persons are discouraged from traveling altogether.

It would be the objective of a toll rationing policy to "toll off" users until the loss of the last one diverted would equal the gain achieved. But a toll on any particular facility would be likely to produce a wide variety of responses, ranging from a simple change in routing to not taking a trip at all. Although there is some basis for estimating the losses to the tolled off in the simpler cases, estimating losses becomes progressively more difficult and indeterminate as the responses become more drastic. Furthermore, it cannot be ascertained precisely whether a toll will end up in tolling off the very users for whose benefit the toll is charged, while not diverting the user who cares little about saving a few minutes but regards his trip as vitally important.

How High the Toll?

According to some writings, it would seem that all this bother about comparing costs and benefits is beside the point. Perhaps one has only to avail himself of the principle that price should cover marginal social cost in order to determine a "correct price." It is possible that an estimate might be made of social costs at any given traffic flow, and price set accordingly. (Nothing has been found in the literature to suggest how an economically "correct" toll could be established in practice, other than the general rule that it should be set at the intersection of the demand and the marginal cost curve.)

To see how this theory works out, it is helpful to examine the typical economic diagram for a congested highway as it appears in toll rationing discussions. In Figure 2, all curves are expressed in units of travel time corresponding to a given vehicle flow on a section of highway. The AC curve is average time per user. It rises as traffic flow increases and congestion develops. The amount of increase in the total travel time of all vehicles brought about by the addition of one vehicle is depicted by the MC curve. Any increase (a point on the MC curve) is composed of (a) the marginal vehicle's travel time, as measured on the AC curve, and (b) the delay that the marginal vehicle causes other vehicles, as measured by the difference between the AC and MC curves. This difference is the "social cost" inflicted by one user upon others. Likewise, it is the "benefit" created by the removal of one vehicle.

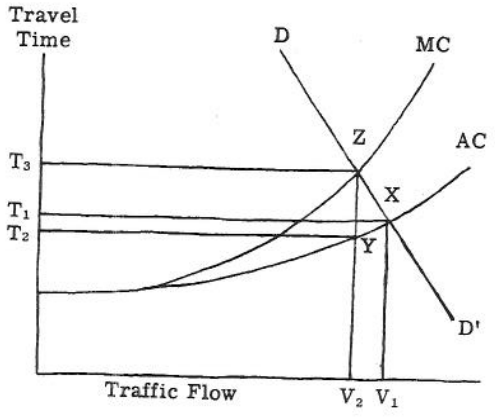

Figure 2.

The demand curve DD' follows from the fact that highway users are responsive to traffic conditions. With increasing congestion and higher travel times, fewer motorists will use the road.

Demand, as represented by the DD' curve, reflects the alternative choices open to users of the highway. To illustrate, suppose that the diagram depicts one road in a two-road network. Travel time on the other road between the points served is T_3. Under this hypothesis, the demand curve would be horizontal at the level of T_3, because no motorist would have to endure a greater travel time than the alternate route provides. Then assume that the demand curve follows the T_3ZXD'. The downward slope beyond point Z has this explanation: if traffic flow on the congested route were V_1 and average travel time T_1, not all users could be diverted to the alternate route by a rise in travel time above T_1. Some would discontinue their trips, changing to public transit, to another destination, etc.

It is evident that the demand curve, as portrayed here, reflects the utility of trip-making on the highway to users, because the behavior of any user willing to endure travel time for the sake of the trip is an index of the value he places on traveling over that road. The loss in value to any user diverted from the road is his total utility, measured by the demand line, minus the actual time cost of travel, which is the point on the AC line corresponding to the volume of vehicles using the road.

Whenever this loss in value is less than the marginal benefit (the difference between the AC and MC curves) created by the removal of one user, the theory argues that economic efficiency can be improved by restricting traffic. Thus, for any traffic flow greater than V_2, the gains from restriction will be more than the losses. At V_2, the marginal loss and benefit exactly balance. This volume of traffic can be achieved with a toll charge which has an equivalent time value of T_3-T_2. Any user whose trip utility does not exceed T_3 would be diverted by this toll. The toll also dissuades any user from coming on the road whose gain in value is less than the added cost, ZY, which he inflicts upon others.

To establish an actual dollars-and-cents toll, of course, marginal social cost, expressed in time units, must be given a money value. Analysts of the efficiency toll proposal customarily transform the diagram in Figure 2 into money terms by multiplying the time units by a given value (such as $0.03 per minute), thus leaving all curves in the same relative positions.

Setting the Toll.—Here one might pause to reflect how the "correct" toll would be set in the real world. Suppose one begins with traffic flow V_1. The only datum about demand is a single point on the curve. Nevertheless, to proceed according to the social cost rule, the toll could be set to equal marginal social cost at V_1 (the vertical distance between AC and MC). But this toll would be "correct" only if there were no response to it and no benefit produced. Otherwise it would be too high; it would divert more users from the road than could be justified by the value of the gains. Then a lower toll, equal to marginal social cost at a lower traffic volume, might be set and succeeding adjustments made until an equilibrium was ultimately reached. (The path toward equality between marginal gain and loss could be described by the "cobweb" diagram of the firm.)

Rather than juggle prices, a highway authority would probably try to establish a "reasonable" toll in the first instance. This would require a preliminary estimate of the slope of the demand curve under the particular circumstances, which perhaps could best be gained by evaluating alternatives available to users and the losses incident to them. The authority would then determine whether the actual response to tolls fulfilled its expectations.

Evaluating the Result.—What if the toll charge failed to produce the result expected of it? Would the toll authority then decide instead that losses incident to vehicle diversion had been undervalued and conclude that the tolls failed to promote efficiency and welfare?

The easiest case to deal with is the two-road network situation. The difference in travel times between the roads would be given a certain money value, based on an estimate of the "average" users' valuation of a minute saved or lost. A calculation would be made of the amount of traffic diversion to the inferior road required to minimize aggregate travel time. At this point no further reduction in total time could be achieved by diverting one more user. The toll actually charged would be based on the assumed valuation of marginal benefit at the desired traffic flow level. If this toll failed to divert the required number of users, one would conclude that the time unit had been undervalued, not that the loss through diversion to the alternate route was greater than had been supposed. The theory would call for an increase in tolls. In short, in the two-road network example, marginal social cost can be given any magnitude required to justify the toll needed to cut traffic to the desired volume.

In the more general situation where the nature of the demand curve is unknown and no desideratum for the removal of traffic is given, some unlikely conclusions are reached by pricing according to marginal social cost estimates.

A toll authority, seeking to set a price, would pick a value for time saving which was thought to be reasonable. Then by a process of calculation and experiment it would arrive at a toll which would be equal to marginal social cost (at traffic volume V_2 in Figure 2). At this volume the toll would equal the marginal benefit believed to occur from the removal of one vehicle.

If demand were highly elastic, the response to a toll would be a relatively large drop in traffic and in social cost. If demand were fairly inelastic, traffic would not drop off as much in response to the toll and marginal social cost would remain relatively high. These different responses are depicted in Figure 3. In each case, the toll charge is T_3-T_2, the benefit to tollpayers is T_1-T_2, and the reduction in traffic is from V_1 to V_2.

When demand is inelastic (Fig. 3b), it takes a rather large toll to produce a small benefit. In the other instance (Fig. 3a), a smaller toll (smaller because marginal social cost at V_2 is less than in Fig. 3b) produces a relatively large benefit. An interesting corollary is that a toll that completely restores free-flowing traffic conditions would have no justification on the basis of social cost. On the other hand, the highest justifiable toll would be charged when demand was totally inelastic (the curve would be a vertical line) but no benefit at all would be gained from its imposition.

An elastic demand suggests that alternatives to trips on the road are readily available and their comparative disutilities are not great. A toll charge then gets results. An inelastic demand suggests that the disutility of alternatives is great and that little is to be gained by charging a toll. Yet the social cost theory claims that in this latter situation an even higher toll ought to be charged.

Here one begins to appreciate the theory's need for the crutch that the tolls are "costless." The social cost theory begins by assuming that tolls equal to marginal

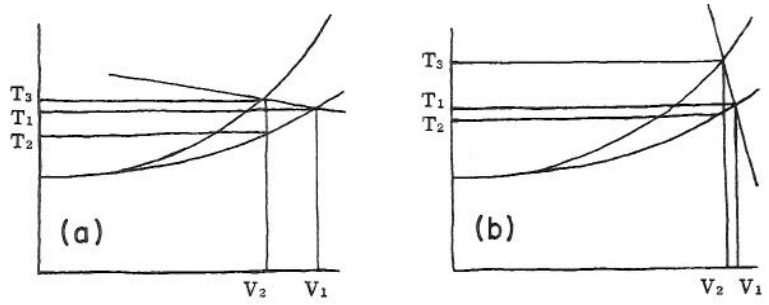

Figure 3.

social cost at any traffic level are justified. Thus, any response to a toll which equals marginal social cost can be counted as a net gain, for the toll revenue remains in the community and represents no loss of value to society. But surely this is a slender thread upon which to hang an entire theory of pricing and income redistribution.

Allowing for Different Benefit Values.—In Figure 2, it is seen that the toll charge, T_3-T_2, is larger than the reduced travel time to each user, T_1-T_2. The best that can happen to a tolled motorist is that he will be no worse off after the toll is paid; but the demand curve would have had to be horizontal at the level of T_1 for this solution. Otherwise, the toll would necessarily leave all motorists in a worse position, in order that any might be diverted. Only "the community" can gain from a restrictive toll policy, given the assumption of a uniform value on the time unit for all users.

This assumption is made, as previously noted, primarily as a convenience in presenting the toll theory. Of course, one might assume that those who pay the toll have the higher value on the benefits it provides, and would think themselves better off even after the toll payment. But we have discovered no proof of this result through a rigorous use of social cost theory. Indeed, the theory does not rule out the opposite result: that those with the lower value on benefits remain to pay because they have the more intense trip demands. Even then a net gain might be shown, as long as the value of benefits exceeded the loss in utility to those diverted. But in this event, it is essential that the toll charge deprive those who pay it of any benefit they gain—and by a considerable margin—in order to keep those diverted from coming back on the road.

The Untolled

At this juncture, consideration might be given to the impact on still another group—those who are already using the alternatives to which some of the former users of the toll facility shift. One can easily visualize a number of situations. For example, alternative facilities might have ample capacity so that no new social costs would be created as a result of the transfers. (This was assumed in the simple two-road network case.) On the contrary, greater utilization of an alternative (a transit route, for example) might actually reduce the average costs of providing the service and stimulate better service, which would reduce "social costs" of all who use it by reducing waiting times. On the other hand, if the shifts were made to certain other facilities (previously uncongested highway routes, for example) it is possible that congestion would begin to build up there and a new epidemic of social costs would be encountered.

At this point, one suffers mental indigestion trying to picture the tolled, the tolled-off, and the untolled, the users and the nonusers, bouncing around among the alternatives, all the while a blinking giant of a computer is fixing and refixing tolls, shadowing users, and redistributing income to promote the general welfare through an optimal arrangement, not only of travel, but also of nontravel.

Yet one more thing must be said about social costs and gains. It is assumed that the private gains to those paying tolls are greater (or no less) than the private losses to those using alternatives. But the original case for efficiency tolls was based on a finding of social costs for which prices should be laid. Hopefully, tolls will have reduced social costs somewhat or caused them to be converted into community income. But what of other consequences resulting from the rearrangements of travel and living patterns which have been wrought by the tolls. In short, one must ask whether the removal of users from congested facilities may result in hitherto unmentioned social costs or, for that matter, social gains.

Economists do not often carry the argument into this field. Yet one suspects that many who flirt with the efficiency toll proposal hold the thought that anything that may discourage highway travel will be socially good. Thus, tolls might be useful regardless of the underlying economic rationale; the economic concepts of "social costs" and "costless transfers" are conveniently available to bolster the case (or to obfuscate the issues).

Space does not permit of joining the attack on this new battleground. One can simply pose the question: possibly a discouragement of highway travel will result in significant losses in preexisting external economies, bring about less desirable living conditions, or even create a less rather than a more viable and amenable urban structure. The

case for toll rationing cannot be closed without acknowledging the possibility of adverse as well as beneficial social consequences. No economic calculus, presently available, is known which could balance all of the gains and losses, the private as well as the social.

HIGHWAY EXPANSION VS TOLL RATIONING

The question that will have occurred to highway engineers, and perhaps everyone else, throughout this discussion is this: Why, if congestion is so bad, is not highway capacity simply expanded in order to eliminate "social costs" by providing for free-flowing traffic. The answer, stated simply, is that it is vital to the pure logic of the toll rationing proposal that capacity will not be added, because it is (a) impossible, (b) uneconomic, or (c) undesirable to do so. Physical or technological impossiblity of expansion will be extremely rare. The other cases go to the heart of the matter.

Rental Tolls

An instance of uneconomic expansion of the highway plant would be one in which the cost of an added facility through a traffic corridor would be so great that informed users would not willingly bear the actual costs of a particular investment (even notwithstanding its contribution to "the system"). With increasing land and construction costs in urban areas, it is quite conceivable that a new facility could not now be economically justified, even though an existing facility which has now become congested was warranted when built.

Whether or not it would be "undesirable" to expand highway capacity involves a more difficult judgment, part of which is beyond the ken of the economist. It might be found by planning decree, for example, that highway expansion would seriously diminish some specific community value, perhaps an esthetic view. On a broader scale, intersubjective community agreement might hold that continued catering to the effective demands of highway users would adversely affect city form. Obviously, important trade-offs are made in such planning decisions.

The costs and benefits of highway expansion should be included among the data to be weighed in the planning process. But the relevant costs will be the actual costs of highway expansion rather than the social costs inflicted by congestion. In fact, costs of congestion will be entered on the other side of the equation; their elimination will be shown as benefits to be offset against actual highway costs. Only after a planning decision has been reached against highway expansion would it be pertinent to evaluate the conversion of social costs to community income through toll rationing.

In this event, the pricing of social cost becomes a long-run proposition, and tolling brings a permanent return to the community in excess of actual expenses. How could these profits on a public service be satisfactorily explained? The question might be turned around to ask, as Kuhn (7) does, why users should be charged any less than they could be charged—why should there be any discrepancy between the potential and actual revenues earned by a public enterprise? Obviously, highway users, by and large, could be made to pay much more than they do.

Although Kuhn observes that there are a wide variety of motives involved in pricing decisions for public enterprises, it may be set down as a general presumption that government will not attempt to exploit a monopoly position without good reason. Thus, the rationale of a long-run excess return from highways must be that the excess is properly earned as a "rental" for resources which, for natural or artificial causes, are in fixed supply. As Lewis (8) puts it:

> If marginal cost exceeds average cost, there will be a profit. This profit is a rent for the use of some scarce resources owned by the undertaking. For example, suppose that the road system can be extended only at rising marginal cost; then if the government based motor taxation on this marginal cost it would probably receive each year a sum well in excess of its expenditure on roads. The difference is a rental for the use of existing roads, and is a very proper and necessary charge if road transport is not to develop beyond the economic point.

The concept of "rent" is usually associated with the pricing of land, but its application in other economic sectors—and specifically to the pricing of highways—has long been recognized in basic economic literature. Knight (9) offered the two-road network case to illustrate the rental charge due to the owner of a superior opportunity for investment. The "toll or rent" on the superior road, he said, would be adjusted to just the amount a user would pay rather than use the inferior road. This adjustment, he demonstrated, would be the one that would maximize the total product of both roads.

In general: "The point is that any opportunity, whether or not it represents a previous investment of any sort, is a productive factor if there is sufficient demand for its use to carry into the state of diminishing returns the application to it of transferable investment. The charge made by private owners for the use of such an opportunity serves the socially useful purpose of limiting the application of investment to the point where marginal product per unit is equal to the product of investment in free (rentless) opportunities; . . ."

Nevertheless, the analogy between rent for land and rent for roads deserves close inspection. The economic rationale of land rent is this: since the same piece of land cannot be used for more than one purpose, the rental charge assigns the land to its most valuable use—that which will pay most for locating on the land. It is implied that the rental charge forces a choice between one use and another. But land is capable of combining two or more uses if they are not thoroughly incompatible. Even though the uses interfere with each other to some degree, like an apartment over a laundromat, each user might think himself better off by sharing the site than by paying for exclusive occupancy. The fact that a user would be willing to pay the full rental for exclusive occupancy, rather than not occupy the site at all, tells nothing of his attitude about joint use. The landowner might decide that he could get his greatest return by limiting the site to one user, and excluding the other, without this arrangement being in accord with the private wishes of either user.

In essence, if a highway is a scarce resource, it is also one that may be jointly used. The toll proposal is designed to reduce the amount of joint tenancy. A toll may make the gain from more exclusive use of the road larger than the loss to those who are denied use. On the other hand, it might leave all users dissatisfied because they would prefer to endure each other in order to avoid tolls. The "rental" rationale for the toll, in short, does not escape the evaluation of welfare gains and losses which has already been explored at some length.

Short-Run Tolls

Some economic theorists who embrace sophisticated pricing techniques will take exception to the conditions just laid down. Some regard efficiency tolling only as a temporary expedient. They see inadequate highway capacity as a short-run phenomenon and would improve the "efficiency" of highway operations by toll rationing until "adequate" capacity can be provided. It would be an incidental but welcome result if the tolls were to provide revenue for additions to capacity.

No overriding virtue is found in this approach. The questions that have been raised and the reservations that are held about tolls over the long pull apply with equal force to short-run tolls. In fact, they are somewhat reinforced by the feeling that short-run tolls are likely to be quite capricious, penalizing some users for the benefit of others because of faulty planning for which none of the users was responsible. The "temporary" nature of such tolls would make it difficult for anyone to make a rational adjustment in travel or living patterns.

The only real difference introduced here is that highway expansion becomes part of the toll package. The basic question is whether the result would be any different than under a rational use of more conventional financing methods, whereby funds would be borrowed, the facility built, and tolls or prices collected to repay actual costs. If the result of the toll approach were no different, it would seem that theoreticians are straining themselves unduly to provide an esoteric rationale in terms of social costs for a fiscal policy that can be handled straightforwardly.

Efficiency Tolls as User Charges

There is a school of thought, however, that feels the need to rely on the social cost theory in order to justify any highway prices at all (other than for current maintenance and operation). Included in this school are some whose thinking is far from "anti-highway"; in fact, they sometimes chide engineers as being too modest in advancing the case for highway expansion. This school uses the economic reasoning whose description has been attempted to demonstrate that capacity should be added to the plant. They would use tolls to convert "social costs" into highway income which would be used to eliminate "social costs," thus to create benefits for users and for the community-at-large.

Here the main departure from traditional reasoning is the insistence that only the "social cost" theory justifies the imposition of sophisticated user charges which will provide revenue for capacity expansion. This notion stems from the reasoning that investment in all segments of the existing highway plant is sunk. Thus, the only justification for any user charges in excess of current operation and maintenance expenses must rest on a finding that the social welfare will be advanced by such charges.

The genesis of these ideas is to be found in cases discussed by leading welfare economists a generation or two ago. In striking contrast to modern times, the typical examples involved over-capacity (or under-utilization). In a toll bridge example, for instance, it could be demonstrated that welfare would be promoted if the toll were removed, because additional traffic would impose no costs on anyone; there would be neither additional highway costs (the average would decrease), nor additional user costs (without congestion the average would stay the same).

As a change is made to the case of under-capacity (over-utilization), the reasoning is turned around. Additional traffic still does not increase highway costs (indeed, it reduces average costs), but it does increase the costs—the "social costs"—of all users. Tolls which would reduce traffic would reduce these costs or convert them to revenue which would be available for highway expansion.

This line of reasoning has promise, but it may be impossible to demonstrate that the revenue earned by such a toll structure would provide for just the "right" amount of highway expansion, without the use of some extreme assumptions. Where demand is highly elastic, returns will be meager, and where demand is inelastic returns will be handsome—even though the cost of highway expansion might be the same in either case. The failure of congestion to build up on roads might bring in considerably less revenue than necessary to finance warranted road improvement. In addition, amounts of revenue produced would depend somewhat on values placed on benefits, and thus on social cost. The higher the unit value of time saving, the more would be total revenue. Or if the unit value of time were derived from an estimate of total revenue requirements, what is accomplished that is not accomplished by straightforward user financing?

A stronger argument for a varied price structure might be made if the road system were regarded as an interrelated network, so that the effect of charging high prices on congested roads would be to stimulate the use of lightly-traveled roads. In a road network situation, however, tax surpluses on some roads would be used as subsidies for others, and marginal social cost might not then serve as a satisfactory guide for correct price levels. A certain theoretical logic may be dug out of all this, but the concept of "social cost" in a multi-road situation needs more elucidation than it has had to date before it can be considered as a reliable basis for pricing.

Improving the Conventional Approach

Conventional highway financing, with its concentration on actual highway costs, can be made to produce satisfactory results.

It is now assumed that there is going to be provided capacity which will eliminate social costs "within reason." But the "costs" are looked on as the "benefits" of highway improvement. Highway investments are made accordingly. Through prices the amounts required to amortize investment costs and pay for current maintenance and operations are recovered. No reason is found to effect a redistribution of income from highway users to others in the community, or from others to highway users.

Obviously the model is oversimplified. It may not be possible to balance capacity and demand precisely. In some cases, indivisibilities of construction will require provision of too much capacity. In other cases, miscalculations in planning will result in over- or under-capacity. These are part of the risks of highway investment planning that must be borne by someone. In this case, why not by the users themselves? It might be said that a risk factor will be included in the costs assessed against users. Where there is under-capacity, the risk cost will be manifested in increases in users' time costs.

It should not be thought that present practices of highway finance are condoned on grounds of economic theory with its nearly uniform pricing structure over time and space, and the geographical cross-subsidies inherent in current expenditure policies. On the contrary, there is a strong appeal for greater sophistication in pricing policy and for advanced techniques of revenue collection. In fact, the end result of this thinking might look almost as though the efficiency toll argument had been embraced.

For example, close study of actual highway costs would unquestionable reveal a case for differential charges (call them tolls) during periods of high traffic demand. Thus, if capacity were needed simply to take care of weekday commuter loads on an urban freeway, the users during the peak period should pay for that capacity. Similarly, if a mountain road had to be expanded only to accommodate week-end travel, appropriate charges to the responsible traffic would be in order. But in each case the differentially higher charges would be based on actual costs rather than elusive social costs.

What is suggested here is not at all unconventional. In fact, it follows directly from the incremental cost theory which appeals to engineers as the "scientific" or "engineering" solution of familiar highway cost assignment problems.

To oversimplify an analogy: If it takes 4 in. of pavement to accommodate passenger cars, but 6 in. to accommodate trucks, the increment in cost (that is, the cost of the 2 in.) is assigned entirely to the trucks. But they are also to bear a share of the costs of the initial 4 in. of pavement, as well as a share of costs which are nonincremental in nature (such as right-of-way costs). Now suppose that four lanes of pavement are adequate to handle all traffic through a corridor except during given peak periods; and additional two lanes are needed to provide "reasonable" flow during the peaks. Peak users would be called upon to pay for the additional two lanes plus an allocated share of the basic four-lane facility.

It is not proposed in this paper to go into the nature of the added highway costs, nor how they, as well as basic costs, might be assigned. It is pointed out, however, that all evidence thus far indicates that additions to cost are less than proportional to additions to capacity for individual freeways. In other words, a six-lane facility costs less than 50 percent more than a four-lane freeway; eight lanes cost less than one-third more than six lanes. On the other hand, if an entirely new freeway were to be required through a corridor to handle peaking its costs would likely be proportional to, or perhaps even higher than an original facility which could handle off-peak traffic.

The important thing is that there are peaking costs which ought to be identified, measured and brought out into the open for evaluation. If these costs were brought home to peak users through prices, one might expect some responses not unlike those expected from efficiency toll pricing, but the case would be based on the solid foundation of actual highway costs assessed against those who gave rise to them.

It is concluded that a case for more "sophisticated" pricing of highways may be made without resting it on an elusive "social cost" or "costless transfer" line of attack. In fact, the typical case for efficiency tolls, notwithstanding the usual impressive mathematical convolutions, appears to rest on a remarkably "unsophisticated" line of reasoning.

A case may be made for differentiation in highway pricing based on geographical and temporal differences in highway demands in relation to carefully measured, but actual highway costs. Such pricing would make for more accurate appraisal of the effective demands of users and provide the wherewithall with which to meet them.

To go beyond this, however, and to base a system of pricing on social costs as measured by congestion alone leaves many questions. The social cost theory would have prices charged for congestion costs regardless of benefits that result or losses that

ensue. Prices are not charged for service improvements. Actual prices may be unreasonably high in relation to benefits. High tolls could provide luxuries for some while denying necessities to others. On the basis of social cost theory, there is no conviction that the results of vehicle rationing through tolls would be beneficial on balance. Therefore, there would be an inclination to let users shoulder their own social costs, in the absence of a showing that efficiency tolls would produce a clear improvement in transportation.

It is not denied that worthy reasons may be found to support attempts at restriction or redirection of motor vehicle use in some urban areas. Pricing might be one of the better tools to accomplish this. But the rationale of a rationing policy should be drawn up in broad planning terms, involving community amenities and esthetics, rather than in the narrow context of social costs which users impose on each other. This requires a balancing of the total consequences of rationing, the adverse as well as the beneficial, not only as they affect users but also as they affect the community-at-large. This part of the toll story seems to remain untold.

REFERENCES

1. Jones, J. H., "The Geometric Design of Modern Highways," pp. 1-37 (1961).
2. Vickrey, W., "Reading an Economic Balance Between Mass Transit and Provision for Individual Automobile Transit." (July 1958): Printed in "Transportation Plan for the National Capital Region," p. 478, 86th Congress, 1st Session (1959).
3. Buchanan, J. M., "The Pricing of Highway Services." Natl. Tax Jour., pp. 97, 102 (June 1952).
4. Beesley, M.E., and Roth, G. J., "Restraint of Traffic in Congested Areas." Town Planning Rev., p. 184 (Oct. 1962).
5. Tanner, J. C., "Pricing the Use of Roads—A Mathematical and Numerical Study." From "2nd Internat. Symposium on the Theory of Road Traffic Flow," British Road Res. Lab. (1963).
6. Wardrop, J. G., "Some Theoretical Aspects of Road Traffic Research." Road Paper No. 36, Proc. Inst. Civil Eng. (London) (1952).
7. Kuhn, T. E., "Public Enterprise Economics and Transport Problems." P. 78, Berkeley, Calif. (1962).
8. Lewis, W. A., "Fixed Costs." Economica, 13:245 (1946).
9. Knight, F. H., "Some Fallacies in the Interpretation of Social Cost." Quart. Jour. of Economics, p. 852 (1924).