

## **Road Pricing and Compensation for Delay**

By David Levinson  
Assistant Professor  
Department of Civil Engineering  
University of Minnesota  
500 Pillsbury Drive SE  
Minneapolis, MN 55455 USA

levin031@tc.umn.edu  
Work: 612-625-6354  
Fax: 510-626-7750

DRAFT November 6, 2001

Submitted for 2002 Transportation Research Board Conference

### **ABSTRACT**

The equity issues facing congestion pricing are an impediment to its adoption. A criticism that gets very little attention is that not only does a toll road enable some to buy their way out of congestion, under certain circumstances such as a queue jumper, they do so at the expense of others - that is, they may make others wait longer so that they can avoid delay, in both cases of take-away capacity and additional capacity. They, along with the toll road authority, are in a sense stealing time from those who don't pay. What to do with the revenue from congestion pricing is a critical question that needs to be answered before toll roads will become widely adopted. This paper investigates the issue of compensation and several possible alternatives. The equity and efficiency problem of conventional (uncompensated) congestion pricing is outlined. Then several of the previous alternatives are discussed and developed. A new compensation mechanism is suggested, called the "delayer pays" principle. This principle ensures that those who are undelayed but delay others pay a toll to compensate those who are delayed. Issues of imperfect information and gaming the system are addressed. Such a system can potentially eliminate some of the disadvantages of congestion pricing while ensuring that the money stays within the transportation sector, and is returned to those delayed.

**KEY WORDS:** Value Pricing, Road Pricing, Compensation, Transportation Equity,

## INTRODUCTION

The equity issues facing congestion pricing are an impediment to its adoption. In part there is resistance due to people's dated perceptions of how toll roads operate, people still envision stopping at toll booths and paying the toll, a situation where the toll road causes more delay than it relieves. Electronic toll collection will obviate these concerns. There is additional resistance to the idea of paying twice for the same thing. If gas taxes already paid for the road, why should tolls now be put in place? A third criticism is the idea of so-called "Lexus Lanes", the idea that toll roads (in parallel with free roads) are only for the wealthy, so that they can bypass congestion while the poor and middle class sit stuck in traffic. Research on the operations of SR91 in Southern California suggests that income effects are not very strong (Sullivan 2000). While logic argues that the rich do have a higher value of time than the poor, and so would in general be more willing to pay a toll, working class individuals may have a greater penalty for being late to work or pick up a child from day care. A related criticism, and one that gets very little attention, is that not only does a toll road enable some to buy their way out of congestion, they may do so at the expense of others if the toll lanes function as queue jumpers - that is, some toll road users may make others wait longer so that they can avoid delay. They, along with the toll road authority, are in a sense stealing time from those who don't pay.

What to do with the revenue is a critical question that needs to be answered before toll roads will become more widely adopted. This paper investigates the issue of compensation and several possible alternatives. First, the equity and efficiency problem of conventional (uncompensated) congestion pricing is outlined. Then several of the previous alternatives are discussed and developed. These include HOT Lanes, Fair Lanes, and Combined Toll/Rationing schemes. Finally, a new compensation mechanism is suggested, called the "delayer pays" principle. These alternatives are in contrast with the efficiency arguments put forward about marginal cost pricing presented in most research on the subject.

## STEALING TIME

At least as early as 1975, a number of environmentalists have called for imposing *The Polluter Pays Principle*. The Polluter Pays Principle argues that the parties who impose environmental costs should either pay to avoid it or compensate those who suffer because of it.

Any social cost takes at least two parties, for instance the polluter and the polluted upon. In the absence of either one, no economic externality would take place. The party responsible for mitigating the externality depends on the circumstances. Two examples illustrate the point:

- If a new (previously unplanned) airport is built in an existing community, can the airport make as much noise as it wants to?
- If an airport has long been located in the middle-of-nowhere, and then a new subdivision moves in, should the new neighbors be able to require the airport to become quieter?

The "common sense" answer to these two questions is "no" as we have an existing status quo that is disrupted by a change. It is the disrupter who creates the externality. In

contrast to the Polluter Pays Principle, we could establish a *Disrupter Pays Principle* to deal with externalities.

What happens on a highway? Congestion, like air pollution, noise, and other externalities results from a lack of well-defined property rights. In the absence of property rights, we have a first-come, first-serve priority system. First-come, first-serve (FCFS) is an arrangement brought about by the technology and the social norms applied to it. Vehicles line up in narrow lanes. Vehicles arriving at the back of the queue rarely drive to the front while other cars are still ahead of them. One occasionally sees cheaters (people driving on shoulders) who violate this norm. Roads with clearly striped lanes thus differ from the mob behavior seen in other bottleneck environments (e.g. a crowded elevator). Transit passengers have different customs in different locations, for instance, everyone is in a well defined queue boarding San Francisco's BART but not on Washington DC's Metro.<sup>i</sup>

On a roadway with a queue, the vehicle in front delays the vehicle in the back. By the "polluter pays principle", the front vehicle should compensate the back vehicle for their delay. On the other hand, the vehicle in front was there first (that is why they are in the front), and the vehicle in the back disrupted the status quo. So by the common sense "disrupter pays principle", it is the person in the back who causes the delay on themselves by arriving later - and of course they already bear the costs in terms of congestion and time lost.

Most congestion pricing proposals argue that because vehicle A delays vehicle B, a government authority should be able to impose tolls on vehicle A (or on both vehicles A and B). It is as if person A robs person B and the police captures person A and keep the loot themselves. This robbery example is socially unacceptable because we have a well-defined system of property rights and clearly the stolen property originally belonged to B. Who does stolen time belong to? Is vehicle B complicit in its delay, or is it solely the responsibility of A? In the case of the crime, is it possible that person B was "asking for it", by walking around and flashing money in a well-known crime-infested area? If the government authority gets the money, what does it do with it? These are issues that should be addressed in an equitable congestion pricing system.

The Coase Theorem famously argues two points, assuming rational behavior, no transaction costs, and bargaining (Coase 1992). First, the efficiency hypothesis posits that, regardless of how rights are initially assigned, the resulting allocation of resources will be efficient. Second, the invariance hypothesis suggests that the final allocation of resources will be invariant to how rights are assigned (Medema and Zerbe 1998). Coase shows how it takes two to have positive or negative externalities, and depending on one's view of the property rights, the prices, taxes, costs, or negotiations will differ. Traffic manifests high transaction costs, no property rights, and little bargaining, perhaps explaining the lack of efficient outcomes.

If property rights are to be assigned, and a low transaction cost exchange mechanism to be established (for instance electronic toll collection), perhaps a more efficient and equitable outcome could be achieved. An efficient outcome suggests maximizing net social benefit, which will consider the weighted sum of delay, schedule delay, and out-of-pocket costs for users, the costs of providing the infrastructure, and the social costs of externalities. Any analysis must assess the appropriate weights -- different individuals have different values of time and different types of delay are perceived

differently. An equitable outcome is less clear, perhaps equalizing the weighted sum of delay, schedule delay, and out-of-pocket costs for all members of some group (say, people who want to use the facility at a given time).

In the absence of private roads, we can consider at least two extreme alternatives regarding the initial distribution of rights:

- Everyone has the right to free (unpriced) travel.
- Everyone has the right to freeflow (undelayed) travel.

If everyone has the right to free (no monetary cost) travel, then the mechanism for more efficient travel requires the delayed to pay the delayers not to delay (a congestion prevention mechanism), or the delayed will continue to suffer congestion. Alternatively, if everyone has the right to freeflow (undelayed) travel, then the burden is on the delayers to compensate the delayed (a congestion damages mechanism). These comport with the *disrupter pays* and *polluter pays* principles respectively. Whether drivers impose costs on those behind them depends on one's point of view vis-a-vis property rights.

A major difficulty is that traffic and congestion externalities are time sensitive. By the time the delayed vehicle arrives, it is too late to pay the delaying vehicle not to be there. Furthermore, the delayer delays multiple vehicles, and so if the delayed tried to pay the delayers not to be there, he may pay significantly more than his own benefit would warrant. These dynamics suggest that conventional economic arguments concerning externalities cannot be simply applied. If the delayer pays scheme were in effect, then those behind would be imposing a cost (the price or the tax or the fine or whatever you want to call it) on those in front, in contrast with the traditional first-come, first-serve approach we have now.

There is also the issue of behavioral response of the paid driver. If I am compensated not to do something, I won't do it. But what if I weren't going to do it initially? For instance, as a non-smoker, I will gladly take any compensation you want to give me for not smoking. Under a compensation regime, I may threaten to smoke just to extort money from you. Similarly, as a driver, I may make the threat to drive on a congested route just to be paid not to. Table 1 categorizes alternative payment and compensation schemes.

These difficulties with internalizing the delay externality are, in part, associated with treating the road as a commons, and trying to give rights to drivers, rather than having the road owner have the right to charge for use. However private ownership does not guarantee an absence of delay. This paper does not consider private roads.

## **BUYING TIME: HOT LANES**

In 1998 the Congestion Pricing Policy Project at the Humphrey Institute released a short video entitled *Buying Time*. It argued that individuals with a high value of time, because of a business meeting, doctor's appointment, departing late for the airport, or picking up a child at day care should be able to buy into a toll lane that moves faster than the freeway it parallels. It is well established that HOV lanes are often underutilized (Dahlgren 1998). While Dahlgren argues that most HOV lanes should be reverted back to general purpose lanes, an alternative has emerged in recent years. High Occupancy/Toll Lanes (HOT) are an innovative solution, suggested by Fielding and Klein (1993) to implement what is now called "value pricing" by selling the available

High Occupancy Vehicle (HOV) lane capacity to those willing to pay extra. Those who pay to use the HOT lanes save time. Other HOV travelers don't noticeably lose time because the additional flow is managed to keep it sufficiently below capacity. What happens to traffic in the general purpose lanes (serving low occupancy vehicles or LOV), however, depends on the geometric configuration of the roads, as well as weather, travel demand, etc.

Figure 1 illustrates two cases of special (diamond) lanes which are used for HOV traffic and might be used as HOT Lanes. In the first case, the bottleneck jumpers, the diamond lane traffic does not interfere with the regular LOV traffic, and avoids the queue entirely. The presence of the additional lane provides a net benefit to regular traffic, by taking cars out of the stream and thus reducing total delay, ignoring any induced demand effects.

In the second case, queue jumpers, the diamond lane traffic simply moves to the head of the queue, displacing the regular LOV traffic (making regular cars wait longer). The total delay in the second case is the same as the baseline, and regular traffic views it as a net loss unless they are compensated. These two outcomes have very different equity implications.

Assume the diamond lanes allow toll users to buy-in. Like a corrupt maître d'hôtel at an expensive restaurant, the toll authority receives payment for allowing the bribers to pass the honest.<sup>ii</sup>

Compensation is required to make the situation fair. Assume the toll-payers have a higher value of time than the no-toll traffic, otherwise they wouldn't pay the toll. The maximum payment that should or could be made to the no-toll traffic is the price of the toll. If the payment were too high however (congested no-toll travelers were paid more than their extra delay would warrant), travelers would be induced by the compensation payment to travel more. But we again run the risk that people with very low values of time would drive to generate income. To avoid this kind of scheming, a two tier pricing system must be established. Part 1 would be a fixed cost assessed to all travelers to pay for maintenance and operation of the roads, as well as other non-delay externalities. Part 2 would be a premium for avoided congestion. The part 2 revenue collected from toll-payers could offset the congested travelers part 1 charge, but should not exceed it.

## **BORROWED TIME: FAIR LANES**

Patrick DeCorla Souza has put forward an idea he has called *Fair Lanes*. Noting that congested facilities often have lower throughput than uncongested facilities, he would separate currently free, but congested, freeway lanes into two sections: toll lanes (our diamond lanes) and "Credit" lanes, but not add any lanes. Electronically tolled express lanes would bear tolls dynamically set to maximize throughput. Electronic credits, funded from tolls, would be given to travelers in the Credit Lanes where congestion continues. The credits could be spent on the toll lanes or for other priced transportation goods (e.g. transit fares or parking), or could be taken as cash. DeCorla Souza claims credit lane travelers would benefit two ways. By better traffic management, the toll lanes now have a higher vehicle throughput than they did previously. Since more vehicles per hour (and fewer vehicles per mile) are on the toll road, fewer vehicles per hour are attempting to use the other lanes. Second, credit lane

travelers receive credits to compensate them for their frustration and for seeing free lanes converted to tolls. While this might again induce travelers with low values of time to drive just to receive credits, perhaps some control could be placed on that. Second, the claim of higher throughput needs to be established empirically.

### **SHARING TIME**

A Pareto-efficient outcome is one where some people are better off while no one is made worse off. Unless revenues are returned to drivers, conventional congestion pricing or marginal cost pricing is not Pareto-efficient. Hau (1991) speaks of the tolled or tolled-on and the tolled-off. The wealthy minority with a very high value of time clearly benefit from congestion pricing, but others lose. Losers are those who either pay a toll but would prefer the congestion to the toll, or those who are tolled-off and don't pay the toll. Further, some people will switch routes to avoid the toll, making the individuals onto whose route they switch worse off. To overcome such difficulties, Daganzo and Garcia (1999) suggest drivers should take turns. By combining rationing (some fraction of users get a free pass every day) with tolling (the remaining fraction of users pay a daily toll that depends on the length of the queue), a Pareto-efficient outcome results, even if revenues are not returned to the original drivers. Their analysis considers commuters driving through a single bottleneck during the morning commute, who each have a desired arrival time, and early and late penalties if they miss that time. Each commuter selects an arrival time at the bottleneck to minimize the weighted sum of tolls, queuing time and deviation from the desired passage time. This system is Pareto-efficient where others aren't because everyone alternates paying the toll and receiving the benefits of others paying the toll. Unless the benefits of traveling faster are shared among the entire population, congestion pricing benefits some (those with a high value of time) at the cost of others, who either pay the toll and save time, but not enough to make it worth while, or who defer the trip altogether.

### **REIMBURSING TIME: DELAYER PAYS**

The system we will introduce and explore in this paper is a variation on the polluter pays scheme applied to congestion. Imagine a cumulative arrival and departure pattern as in Figure 2. This is represented numerically in Table 2, where the numbers 1 - 9 indicate the 1st through 9th vehicle. Each row is a time increment (or turn) for instance a two second headway, reflecting the capacity of the roadway of 1800 vehicles per hour.<sup>iii</sup> Vehicle 1 delays nobody. However after that first vehicle, the arrival rate exceeds the departure rate (say 3600 vehicles per hour for several seconds). As a consequence, Vehicle 2 delays Vehicle 3 by one turn. Vehicle 3 delays vehicles 4 and 5 by 1 turn. Vehicle 4 delays vehicles 5, 6, and 7 by 1 turn and so on. We can tabulate the direct payments and income from such a system, shown the right hand columns of Table 2.

We define this *short-run* marginal cost as the change in the *short-run* total cost, because we only know information about the present (the number of vehicles in the queue at the time a vehicle leaves), not the full consequences of delay on vehicles yet to join the queue. The short-run marginal cost scheme above would then charge 1 unit of toll to

vehicles 2, 3 and 6. It would charge 2 units of charge to vehicles 4, and 5. Vehicles 7, 8, and 9 would get refunds of 1, 2, and 4 units of toll respectively. If everyone has the same value of time, which can be monetized in units of tolls, this seems fair.

However, the short-run marginal costs imposed by a vehicle are not its only costs. Rather a vehicle's presence has a reverberation much longer in time. For instance, in the absence of vehicle 2, the queue looks like the cumulative arrival and departures given in Figure 3, shown numerically in Table 3. Note that the total difference in costs with and without vehicle 2 is now  $16 - 9 = 7$ , implying a true long-run marginal cost of vehicle 2 of 7 units, rather than the 1 unit shown above.

In the absence of vehicle 3, the total costs are again only 9 units. In the absence of vehicle 4, the total costs are 10 units. But those savings are not additive, that is, initially there were 16 units of cost, the savings from vehicle 2 is 7 units, from vehicle 3 is also 7 units and vehicle 4 is 6 units. Yet, we cannot add  $7 + 7 + 6 = 20$ , which exceeds the total delay. Rather, the total cost is 4 units and only  $16 - 4 = 12$  units are saved. So even eliminating vehicles 2, 3, and 4 does not completely eliminate congestion. Thus we can identify two complications, the long-run marginal cost of a vehicle depends on how many other vehicles there are and when each vehicle arrives.

Charging the long-run marginal cost (rather than the short-run marginal cost) and paying people the amount of their delay, would produce the result shown in Figure 4. The figure shows that more money is paid in than paid out. This discrepancy is because eliminating a vehicle will sharply reduce delay, but to the delayed vehicle, it matters not which vehicle ahead is eliminated, any one of them will reduce delay significantly. So using long-run marginal cost accounting will generate surpluses. This can be described mathematically with the equations and description given in Table 4.

If people vary in their values of time, people with a high value of time may not be fully compensated, while those with a low value of time would get more dollars back than the value of the time they wasted. This may induce more travel by clever people with low values of time trying to scam the system; however clever people rarely have low values of time for long.

Moreover, the system would send price signals back to drivers, who would then change their departure times in some fashion, probably smoothing out their behavior. A new, less peaked, arrival pattern would come about. So after equilibration between price and demand, the system would have a lower price and lower net turnover than suggested by Table 2.

One can imagine problems with this scheme, getting on queue becomes a gamble that there is not a large platoon of vehicles behind you. Can the technical "gamble" problem be solved? I believe we can come very close with the technology available, but it will require implementing a detailed traffic monitoring system, as illustrated in Figure 5.

Strictly speaking the correct charge (either short-run or long-run marginal cost) is unknown until some time after the driver exits (the front) of the queue, but some approximations could be made. The charge depends not only on how many vehicles were behind the driver at the time the driver exits, but how many vehicles are behind those vehicles -- that is on how much delay that vehicle actually caused. Figure 5 represents a freeway with an on and off ramp just before a bottleneck. If we know the mainline traffic flow, on-ramp flow and off-ramp flow, we can post the expected price at

the Variable Message Sign (VMS) just before the bottleneck. This will not be strictly accurate, as the mainline flow may suddenly spike upward, or the off-ramp may suddenly get more traffic. But with experience, the forecasting system would get more and more accurate.

This leads to a modified strategy that distributes the revenue back to the delayed, but would only charge drivers based on what they were promised at the VMS. In this case the Toll Authority would assume the risk of under/over forecasting, and someone would monitor it to ensure it behaved well.

The delayer pays scheme, using short-run marginal cost enables a straightforward solution to "what to do with congestion pricing revenue" -- return it directly to those who were delayed almost instantly. The system can be perfectly revenue neutral, stay within the roadway sector, and be economically efficient. Overall, the amount of revenue collected equals the amount distributed. But those who delay others the most pay the most, while those who are delayed more than they imposed delay on others are compensated for their delay. Again to avoid scheming, a two-tier pricing system could be established.

## CONCLUSIONS

Equity and efficiency form the two pillars on which transportation decisions should be made. However, determining what is efficient, much less what is equitable, is far from simple.

When considering whether and how to compensate for congestion pricing, we have a number of alternatives:

- continue with First Come, First Serve, using delay as the cost of travel - the "no-toll" option.
- Marginal cost pricing in peak times, without compensation.
- implement a delayer pays scheme to charge based on the actual congestion caused.
- split the difference between delayer and delayed.
- convert HOV lanes to HOT lanes,
- convert general purpose lanes to "Fair" lanes, or
- construct a toll and rationing system.

Who owns the right to travel on the roadway? Currently the system is first-come first-serve. Unfortunately the conventional marginal cost pricing approach often ignores traffic dynamics and tends to treat time in large discrete blocks rather than continuously. How significant a problem this is depends on the conditions of the case. The delayer pays scheme outlined in this paper implies everyone has a right to free-flow, and the individuals who deny that right to others are the ones who should pay. So is delayer pays a good idea? This depends on answers to two questions:

- Empirical question - What will the magnitude of cheating/gaming the system be?
- Technical question - What is the cost of the added data collection and toll redistribution?

There are also several key philosophical questions that need to be addressed. These very much parallel the fundamental question of whether people should be guaranteed equality

of opportunity or equality of outcome. Congestion externalities required two actors, the delayer and the delayed. If both parties have equal opportunity to arrive, than one should not compensate the other. But if we want to guarantee an equal outcome in terms of a combination of time and money, those who save time should pay more money and those who spend more time should be paid by those causing their delay.

Congestion pricing generates revenue that can substitute for conventional transportation financing (such as the gas tax). Few argue against substitution, as it makes sense as a demand management measure. However, what to do with excess congestion pricing revenue has been a hurdle for its adoption. In the absence of private roads, this is a political problem. Suggestions range from the government keeping the money, to building more roads, to providing transit, to compensating the poor (redistributing the money by income class). There is a clear alternative however that is fair, returning the excess congestion pricing revenue to those who are congested, in the form of cash or credits, in such a way to avoid encouraging gaming the system or driving for dollars.

## REFERENCES

- Daganzo, C.F., R. Garcia. A Pareto improving strategy for the time-dependent morning commute problem, *Transportation Science* (in press). Presented at the 8th World Conference on Transport Research, Antwerp, Belgium, July, 1998
- Dahlgren, Joy. 1998. High occupancy vehicle lanes : not always more effective than general purpose Transportation research. Part A, Policy and practice. Vol. 32A, no. 2 (Feb. 1998) p. 99-114
- DeCorla-Souza, Patrick. 2000. Making Value Pricing on Currently Free Lanes Acceptable to the Public with "Credit" Lanes . unpublished manuscript <http://pdecorla.tripod.com/lane3.htm>
- Fielding, Gordon J., and Daniel B. Klein. 1993. How to Franchise Highways. *Journal of Transport Economics and Policy*. May 1993. pp. 113-130. (University of California Transportation Center No. 134 reprint).
- Hau, Timothy D. 1991. *Economic Fundamentals of Road Pricing: A Diagrammatic Analysis*. The World Bank Washington DC
- Sullivan, Edward. 2000. Continuation Study to Evaluate the Impacts of the SR91 Value-Priced Express Lanes Final Report. Submitted to State of California Department of Transportation Traffic Operations Program HOV Systems Branch, Sacramento CA 94273
- Transportation Research Board. 1994. *Curbing Gridlock: Peak Period Fees to Relieve Traffic Congestion: Special Report 242*. Washington DC

*Table 1: Alternative monetary payment and compensation schemes*

Delayer	Delayed	Road	Label
0	0	0	First-Come First Serve (unpriced)
Paid	Pays	0	Disrupter Pays
Pays	Paid	0	Polluter Pays
Pays	0	Paid	\
0	Pays	Paid	- "Marginal Cost Pricing"
Pays	Pays	Paid	/

*Table 2: Short-run marginal cost payment scheme with all vehicles.*

Time	Queue	Veh	Payment	Income	Net Income
0:00	1	1	0	0	0
0:02	23	2	1	0	-1
0:04	345	3	2	1	-1
0:06	4567	4	3	1	-2
0:08	56789	5	4	2	-2
0:10	6789	6	3	2	-1
0:12	789	7	2	3	1
0:14	89	8	1	3	2
0:16	9	9	0	4	4
Total			16	16	0

*Note:* Vehicle 1 arrives and departs before vehicle 2 arrives.

*Table 3 Payment scheme in the absence of vehicle 2.*

Time	Queue	Veh	Payment	Income	Net Income
0:00	1	1	0	0	0
0:02	3	3	0	0	0
0:04	45	4	1	0	-1
0:06	567	5	2	1	-1
0:08	6789	6	3	1	-2
0:10	789	7	2	2	0
0:12	89	8	1	2	1
0:14	9	9	0	3	3
			9	9	0

Table 4 Mathematical model of delayer pays compensation schemes

Cost and Income Variables	Expression
$S_v =$ Own cost	$S_v = A_v - D_v$
$T_{[l]}$ = Total cost [for arrival pattern containing vehicles in bracket]	$T_{[l]} = S_v$
$J_v =$ Short-run marginal cost	$J_v = Q(D_v) - l$
$M_v =$ Long-run marginal cost	$M_v = T_{[l-v]} - T_{[l-v-l, v+l-v]} - S_v$
$R_v =$ Reimbursement income	$R_v = S_v / \mu$
$N_v =$ Net income	Short-run marginal cost $N_v = J_v - R_v$ Long-run marginal cost $N_v = M_v - R_v$

Notes: Subscript  $v$  denotes vehicle  $v$ .  $A_v =$  Arrival time (at back of queue).  $D_v =$  Departure time (from front of queue).  $Q(t) =$  Number of vehicles in queue at time 't'.  $\mu =$  Service time (headway between vehicles departing queue).











