Real-Time Congestion Pricing Strategies for Toll Facilities

Jorge A. Laval[a,*], Hyun W. Cho[a], Juan C. Muñoz[b], Yafeng Yin[c]

[a] *School of Civil and Environmental Engineering, Georgia Institute of Technology*
[b] *Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile*
[c] *Department of Civil and Coastal Engineering, University of Florida*

Abstract

This paper analyzes the dynamic traffic assignment problem on a two-alternative network with one alternative subject to a dynamic pricing that responds to real-time arrivals in a system optimal way. Analytical expressions for the assignment, revenue and total delay in each alternative are derived as a function of the pricing strategy. It is found that minimum total system delay can be achieved with many different pricing strategies. This gives flexibility to operators to allocate congestion to either alternative according to their specific objective while maintaining the same minimum total system delay. Given a specific objective, the optimal pricing strategy can be determined by finding a single parameter value in the case of HOT lanes. Maximum revenue is achieved by keeping the toll facility at capacity with no queues for as long as possible. Guidelines for implementation are discussed.

Keywords: system optimum, user equilibrium, congestion pricing

1. Introduction

There are currently more than a dozen cities around the world that implement zone- or cordon-based congestion pricing, and around 20 toll facilities in the United States subject to congestion pricing. The pricing strategies in these facilities are inspired by the "first-best toll" concept borrowed from the economics literature, which can be stated as "System Optimum (SO) will be equivalent to User Equilibrium (UE) with tolls derived from the SO solution" (see e.g. Carey and Watling, 2012). This concept has been adapted to the

*Corresponding author. Tel. : +1 (404) 894-2360; Fax : +1 (404) 894-2278
Email address: jorge.laval@ce.gatech.edu (Jorge A. Laval)

case of traffic flow rather directly, and it is our view that some important traffic dynamics properties may have been overlooked in doing so.

Although there are a number of studies examining the performance of High Occupancy Toll (HOT) lanes (see, e.g. Supernak et al., 2003, 2002a,b; Burris and Stockton, 2004; Zhang et al., 2009)) and travelers' willingness to pay (Li, 2001; Burris and Appiah, 2004; Podgorski and Kockelman, 2006; Zmud et al., 2007; Finkleman et al., 2011), SO tolling policies have received little attention. Existing studies focused on ad-hoc objectives that the tolling agencies may seek to achieve, such as ensuring free-flow conditions on HOT lane. For example, Li and Govind (2002) developed a toll evaluation model to assess the optimal pricing strategies of the HOT lane that can accomplish different objectives such as ensuring a minimum speed on the HOT lane, or in the general-purpose lanes (GPL), or maximizing toll revenue. Zhang et al. (2008) proposed the logit model to estimate dynamic toll rates of the HOT lane after calculating the optimal flow ratios by using feedback-based algorithm on the basis of keeping the HOT lane speed higher than 45mph. Yin and Lou (2009) explored two approaches including feedback and self-learning methods to determine dynamic pricing strategies for the HOT lane, and the comparative results showed that the self-learning controller is superior to the feedback controller in view of maintaining a free-flow traffic condition for managed lanes. Lou et al. (2011) further developed the self-learning approach in Yin and Lou (2009) to incorporate the effects of lane changing using the hybrid traffic flow model in Laval and Daganzo (2006). Burris et al. (2009) examined the potential impacts of different tolling strategies on carpools, which includes removing or reducing the preferential treatment for them in the HOV lane.

In our formulation the social cost to be minimized is total system delay, and does not include the effects that tolls may have on trip generation or departure-time choice. The proposed pricing strategies are real-time, in the sense that they respond to real-time traffic arrivals in a way that minimizes total system delay for that particular rush hour. Therefore, the underlying assumption is that demand is inelastic within the day, but it could very well be elastic from day to day. In this context, this paper proposes a real-time pricing mechanism that is consistent with known properties of marginal costs under inelastic demands, i.e.: the cost of adding an additional user to a specific alternative is given by the time until congestion clears, it is not well defined when capacity is reached, and the SO assignment is not unique (Muñoz and Laval, 2005; Kuwahara, 2007). Towards this end, this paper is

organized as follows. Section 2 presents the problem formulation along with the SO and UE solutions. Section 3 summarizes the general properties of SO tolls, including expressions for delays and revenue. Section 4 examines the special case of HOT lanes, and finally section 5 presents a discussion.

2. Problem Formulation

Consider the equilibrium between two alternatives with finite capacity, one of which is priced. To fix ideas, we take the example of a Managed Lane (ML) competing with the general-purpose lanes (GPL), but the formulation to be developed also applies to other cases such as toll roads or zone-based pricing. Our focus is on real-time pricing strategies and therefore we do not assume that traffic demand is known in advance, but only as it realizes.

Let $A(t)$ be the cumulative number of vehicles at time t that have entered a freeway segment containing a ML entrance, and let the corresponding flow be $\lambda(t) = \dot{A}(t)$. All vehicles are bound for a single destination past a GPL bottleneck of capacity μ_0, which may be bypassed by paying a toll $\pi(t)$ upon entering the ML at time t, which has a bottleneck of capacity μ_1; see Fig. 1.

The cumulative count curve of vehicles using route r (r=0 for the GPL and $r = 1$ for the ML) is denoted $A_r(t)$ and the flow, $\lambda_r(t) = \dot{A}_r(t)$. Clearly,

$$\lambda(t) = \lambda_0(t) + \lambda_1(t), \tag{1}$$

and is assumed unimodal. Let $\tau_r(t)$ be the trip time in route r experienced by a user arriving at time t:

$$\tau_r(t) = \tau_r + w_r(t), \tag{2}$$

where τ_r is the free-flow travel time, and $w_r(t)$ is the queuing delay, which can be expressed as:

$$w_r(t) = \frac{A_r(t) - A_r(t_r)}{\mu_r} - (t - t_r), \qquad t_r < t < T_r, \tag{3}$$

where t_r and T_r represent the times when route r begins and ends being congested, respectively. Let:

$$\Delta = \tau_0 - \tau_1 \tag{4}$$

be the extra free-flow travel time for using the free alternative. Although in many cases one would expect $\tau_0 \approx \tau_1$, this will not be assumed here for

maximum generality. To simplify the exposition, we assume that $\Delta > 0$ hereafter; the other two cases will be discussed in the last section of this paper. Under this assumption, we will see that $t_1 < t_0$ in the SO solution, i.e. the ML is used at capacity before the GPL, as shown next.

2.1. System Optimum

The SO solution to our problem (without pricing) is presented in Fig. 2, which shows the system input-output diagram using total arrivals $A(t) = A_0(t) + A_1(t)$ and total virtual departures $D^*(t)$. [1] The area between these curves is the total system delay, i.e. the total time spent queuing in the system. The method to obtain the curve $D^*(t)$ was introduced in Muñoz and Laval (2005), and is best visualized by imagining a ring connected to the rightmost end of $D^*(t)$ that is slid along $A(t)$ from right to left until $D^*(t)$ "touches" $A(t)$ again (at point "1" in the figure). This point corresponds to the time when both alternatives start being used at capacity (t_0 in our case since $\Delta > 0$, and $\lambda(t_0) = \mu_0 + \mu_1$), and from here one can identify the arrival time of the last vehicles to experience delay in each alternative, $T_r, r = 0, 1$, and the time when the shorter alternative starts being used at capacity, (t_1 in our case, and $\lambda(t_1) = \mu_1$); see Fig. 2. This figure also shows how to obtain the total system departure curve $D(t)$, which gives the count of vehicles reaching the destination at time t. Notice that total arrivals and departures in the system are not first-in-first-out. The resulting flow pattern is summarized below (Muñoz and Laval, 2005):

System Optimum Conditions: The SO assignment when $\Delta > 0$ for users arriving at t satisfy:

1. $0 \leq t \leq t_1$: everybody uses the ML
2. $t_1 \leq t \leq t_0$: the ML is used at capacity, excess inflow uses the GPL
3. $t_0 \leq t \leq T_0$: both alternatives are used at capacity
4. $t \geq T_0$: everybody uses the ML □

Notice that in $t_0 \leq t \leq T_0$ the solution in terms of the alternative-specific arrivals $A_r(t), r = 0, 1$ is not unique. This is because the SO conditions simply state that within this time interval both alternatives be used at capacity.

[1] Virtual departures are defined as the arrival curve shifted to the right by the free-flow travel time.

Where to store the queues is up to the operator. Therefore, hereafter we focus on $t_0 \leq t \leq T_0$ because it is the only time interval where we have flexibility to define $A_r(t)$. Without loss of generality and for simplicity we also set $t_0 = 0, A(t_0) = 0$. This implies that the delay to users arriving in $t < t_0$ will not be considered. But this is irrelevant because such a delay is a constant of our problem, i.e. independent of the pricing strategy.

Setting $t_0 = 0, A(t_0) = 0$ simplifies the construction of total arrivals and departures, as shown in Fig. 3a, and streamlines the derivation of alternative-specific input-output diagrams in Figs. 3b,c, which are first-in-first-out. Recall that arrivals $A_r(t)$ are not unique; the only requirement is that they start at the origin, remain above the virtual departures, and pass through points "1" and "2" in Figs. 3b,c, respectively. The departure curves at each alternative measured at the destination, $D_r(t)$, are obtained by shifting the virtual departures by the free-flow travel time τ_r; total system departures are then $D(t) = D_0(t) + D_1(t)$.

The total system cost is the area between total arrivals and departures, which can be partitioned into the three components shown in Fig. 3a: (i) the total delay defined previously (area 0-4-3-0), (ii) the fixed travel time τ_1 incurred by all users (striped area), and (iii) the extra travel time $\Delta\mu_0 T_0$ incurred by GPL users (lightly shaded area). The reader can verify that the striped and lightly shaded areas in Fig. 3a correspond to the sum of the respective areas in parts b and c of the figure.

Figs. 3b,c also show the delay, travel time and externality in each alternative, $w_r(t), \tau_r(t)$ and $e_r(t)$, respectively. It can be seen that:

$$e_r(t) = T - t - \tau_r(t). \tag{5}$$

The marginal cost $\tau_r(t)+e_r(t)$ in each alternative gives the extra cost incurred by the system if an additional unit of flow uses such alternative. In $t_0 \leq t \leq T_0$ the marginal cost is given by the time remaining until the end of congestion in the system, and it is identical on both alternatives, as expected. Outside this time interval only the alternative with the least marginal cost (ML in this case) is used.

It is worth noting that point "2" in Fig. 3c implies that at T_0 there has to be a queue in the ML, and therefore completely eliminating queues from the ML facility is not system optimal (when $\Delta > 0$). The reason is that starting at this time the GPL must not be used since its marginal cost is greater than the ML marginal cost.

2.2. User Equilibrium with pricing

The UE condition for our problem under any pricing strategy $\pi(t)$–not necessarily SO tolls–can be expressed as

$$\tau_0(t) = \tau_1(t) + \pi(t) \tag{6}$$

when both alternatives are used; otherwise, only the less expensive alternative is used. Notice that in this formulation the toll has units of time, and implies that all users have the same value of time. The case of heterogeneous users is discussed in section 5.

Following Laval (2009) it is more convenient to express the UE condition (6) in differential form, which equalizes the rate of change in travel cost among alternatives, i.e. $\dot{\tau}_0(t) = \dot{\tau}_1(t) + \dot{\pi}(t)$, with $\dot{\tau}_r(t) = \dot{w}_r(t) = \lambda_r(t)/\mu_r - 1, r = 0, 1$ by (3). This gives in our case:

$$\rho_0(t) = \rho_1(t) + \dot{\pi}(t), \tag{7}$$

where we have defined the demand-capacity ratios $\rho_r(t) = \lambda_r(t)/\mu_r, r = 0, 1$. Notice that the differential UE condition is applicable only when the initial condition is in UE equilibrium. Substituting (1) into (7) gives the UE assignment when both alternatives are used:

$$\rho_0(t) = \rho(t) + \bar{\mu}_1\dot{\pi}(t), \tag{8a}$$

$$\rho_1(t) = \rho(t) - \bar{\mu}_0\dot{\pi}(t), \tag{8b}$$

where $\mu = \mu_0 + \mu_1$, $\bar{\mu}_r = \mu_r/\mu$ and $\rho(t) = \lambda(t)/\mu$ is the system demand-capacity ratio. It can be seen that for constant or no tolls ($\dot{\pi}(t) = 0$) the UE condition implies that each alternative and the system have the same demand-capacity ratio. Arrival curves are obtained by integrating (8) from the time when both alternatives start being used, say t_{ini}, and thus:

$$A_r(t) = (-1)^r\bar{\mu}_0\mu_1(\pi(t) - \pi(t_{ini})) + \bar{\mu}_r A(t), \qquad r = 0, 1. \tag{9}$$

where we have used $A(t_{ini}) = 0$ without loss of generality.

3. Properties of System Optimum tolls

In this section we identify and examine the properties of the SO toll, $\pi(t)$, that produces a SO assignment under UE. The goal of SO tolls is for every

user to perceive the marginal cost it imposes to the system. This could be accomplished in our case by charging the externality in each alternative given by (5). Equivalently, since we want to maintain the GPL toll-free we will charge the difference in the externalities to the ML only. This is illustrated in Fig. 4a, which shows the marginal cost in equilibrium along with travel times, delays, and externalities on each alternative, as a function of time. The figure also shows the SO flow pattern in each relevant time interval, with the exception of $t_0 \leq t \leq T_0$, where SO flows are not unique, and nor are $\tau_r(t)$ and $e_r(t)$. It follows that in $t_0 \leq t \leq T_0$ the toll $\pi(t)$ is also not unique and can be chosen freely but within the following constraints:
(i) boundary conditions constraints:

$$\pi(t_0) = \Delta, \qquad \pi(T_0) = \Delta - w_1(T_0), \qquad \text{and} \tag{10a}$$

(ii) active bottleneck constraints:

$$\dot{\pi}(t) \geq (\mu - \lambda(t))/\mu_1, \qquad \text{if GPL at capacity with no queue} \tag{10b}$$

$$\dot{\pi}(t) \leq (\lambda(t) - \mu)/\mu_0, \qquad \text{if ML at capacity with no queue} \tag{10c}$$

$$-\lambda(t)/\mu_1 \leq \dot{\pi}(t) \leq \lambda(t)/\mu_0, \quad \text{if GPL and ML have queues} \tag{10d}$$

The boundary condition constraints (10a)–depicted as points "1" and "2" in Fig. 4b–are a consequence of the SO conditions in the time intervals $t \leq t_0$ and $t \geq T_0$, which force pricing to be either fixed or arbitrary. Before t_1 there is no congestion and therefore as long as $\pi(t) \leq \Delta$ all drivers will choose the ML, as required by the SO condition. This is shown in Fig. 4b by the shaded rectangles, which indicates that the toll could be anywhere inside this area. During the time interval $t_1 \leq t \leq t_0$ the ML has to operate at capacity with no queues while the excess demand should be diverted to the GPL, which is achieved using $\pi(t) = \Delta$. After T_0 only the ML should be used, which can be achieved, again, by pricing within the shaded area in the figure.

The active bottleneck constraints (10b), (10c) and (10d) ensure that the bottlenecks will be used at capacity in $t_0 \leq t \leq T_0$ and under all situations. In particular, if there is no queue on alternative r one should impose $\lambda_r(t) \geq \mu_r$ in (8a) or (8b), which gives (10b) or (10c). If there is a queue on both alternatives, the less restrictive condition $\lambda_r(t) \geq 0$ should be imposed, which gives (10d).

3.1. Delays

The total delay for users arriving in $t_0 \leq t \leq T_0$, $W = \int_{t_0}^{T_0} (A(t) - \mu t) dt$, is a constant in our problem and is given by the dark shaded area in Fig. 3a.

The delay in each alternative, $W_r(\pi) = \int_{t_0}^{T_0}(A_r(t) - \mu_r t)dt$, are functions of the pricing strategy. Using (9) gives:

$$W_r(\pi) = \int_{t_0}^{T_0} (-1)^r \bar{\mu}_0\mu_1(\pi(t) - \Delta) + (\bar{\mu}_r A(t) - \mu_r t)dt, \tag{11a}$$

$$= (-1)^r \bar{\mu}_0\mu_1 \int_{t_0}^{T_0} (\pi(t) - \Delta)dt + \bar{\mu}_r W \tag{11b}$$

where one can see that $W = W_0(\pi) + W_1(\pi)$, as expected. It is interesting to note that manipulation of (11b) gives

$$W_0(\pi)/\mu_0 - W_1(\pi)/\mu_1 = \int_{t_0}^{T_0} (\pi(t) - \Delta)dt, \tag{12}$$

which can also be verified in Fig. 4a: the shaded areas correspond to $\int_{t_0}^{T_0} w_r(t)dt = \int_{t_0}^{T_0} \mu_r w_r(t)dt/\mu_r = W_r/\mu_r, r = 0, 1$, respectively.

3.2. Revenue

Let $R(\pi)$ be the revenue under strategy $\pi(t)$. It can be expressed as $\int_{t_0}^{T_0} \lambda_1(t)\pi(t)dt$, which by (8b) is also:

$$R(\pi) = \bar{\mu}_1 \int_{t_0}^{T_0} \lambda(t)\pi(t)dt - \bar{\mu}_0\mu_1 \int_{t_0}^{T_0} \dot{\pi}(t)\pi(t)dt, \tag{13a}$$

$$= \bar{\mu}_1 \int_{t_0}^{T_0} \lambda(t)\pi(t)dt - C \tag{13b}$$

where $C = \bar{\mu}_0\mu_1(\Delta^2 - (\Delta - w_1(T_0))^2)/2$ is a constant that follows from $\int_{t_0}^{T_0} \dot{\pi}(t)\pi(t)dt = \frac{1}{2}\pi^2(t)|_{t_0}^{T_0}$ and (10a). Therefore, maximizing revenue can be expressed as the following mathematical program:

$$\max_{\pi(t)} \int_{t_0}^{T_0} \lambda(t)\pi(t)dt, \quad \text{subject to} \quad (10), \tag{14}$$

and we have the following result:

Result 3.1. (Maximum Revenue) *Revenue is maximized for the highest possible $\pi(t)$ that does not violate the SO condition; i.e., the ML is maintained at capacity with no queues for as long as possible (see Fig. 5a).*

Proof Maximizing $\int_{t_0}^{T_0} \lambda(t)\pi(t)dt$ is equivalent to maximizing $\int_{t_0}^{T_0} \pi(t)dt$ because (i) $\lambda(t)$ is exogenous and nonnegative, and (ii) the active bottleneck constraints are in terms of $\dot{\pi}(t)$, which means that the highest possible $\pi(t)$ value at a given time t is obtained only if it is preceded by the highest possible $\pi(t')$ value at an earlier time t'. Therefore, the optimal solution can be obtained in a (t, π) diagram starting from each boundary point (t_0, Δ) and $(T_0, \Delta - w_1(T_0))$, and drawing curves of maximum slope from each one in the direction of increasing and decreasing time, respectively, until they cross, say at time t^*. This is shown in Fig. 5b, where these points have been labeled "1" and "2", respectively. It can be seen that maximum slopes are constrained by (10c) and (10d), respectively, because at $t = t_0$ there is no ML queue, and right before $t = T_0$ there is a queue on both alternatives. It follows that in $t_0 \leq t \leq t^*$ the ML is maintained at capacity with no queues, and in $t > t^*$ a queue on both alternatives is allowed. □

Intuitively, from (12) one can see that maximizing $\int_{t_0}^{T_0} \pi(t)dt$ also maximizes the difference $W_0(\pi)/\mu_0 - W_1(\pi)/\mu_1$, which is obtained by imposing the highest and the lowest possible travel time to the GPL and ML, respectively.

3.3. Social benefits of S0 tolls

The benefits of SO tolling compared to no tolling can also be quantified. Fig. 6 presents the input-output diagram in Fig. 3a, which assumes $\Delta > 0$, along with the total departure curve assuming no tolls, namely $D_{\pi=0}(t)$. The details for obtaining this curve are omitted here in the reader is referred to Laval (2009). It can be seen that the social (i.e., delay) benefits of SO pricing is bounded by $(T - \tau_0)\Delta\bar{\mu}_1$. To see this, we note that the extra travel time with no tolls is bounded by $\Delta\bar{\mu}_1$, which is a tight bound when the slope between the origin and point "1" in the figure is μ.

4. HOT lanes under linear tolls

System optimum tolls on HOT lanes can be characterized within the proposed framework using $\Delta = 0$; typically $\mu_1 \ll \mu_0$ but we do not need this assumption. For simplicity and without loss of generality we neglect high occupancy vehicles (who do not pay the toll to use the HOT lane) in this analysis. The reader can verify using Fig. 3 that in this case $w_1(T_0) = 0$, and therefore the boundary condition (10a) changes to:

$$\pi(t_0) = 0, \qquad \pi(T_0) = 0. \tag{15}$$

We now show that when the pricing strategy is linear, as defined momentarily, we can obtain closed-form expressions for revenue, delay and flows. It turns out that these quantities are all linear functions of a single parameter, which makes the optimization of this system very simple, to the point where the appropriate pricing strategy to accomplish a given objective is reduced to choosing a single parameter value.

Definition: We say that tolls are linear in the arrivals if there is a constant, a, called the pricing coefficient, such that:

$$\dot{\pi}(t) = (\rho(t) - 1)\, a, \qquad t_0 \le t \le T_0, \tag{16}$$

or equivalently (letting $t_0 = 0$),

$$\pi(t) = (A(t) - \mu t)\, a/\mu, \qquad t_0 \le t \le T_0, \tag{17}$$

which means that the toll is proportional to the system queue at t, $A(t) - \mu t$, or delay $w(t) = (A(t) - \mu t)/\mu$; see Fig. 3a. Notice that a is dimensionless. This strategy is "real-time" because from (16) it is clear that to determine the toll at time t all that is needed is the demand-capacity ratio at the same time, which can be measured in real-time.

Result 4.1. (Assignment, delays and revenues under linear tolls) *Under linear pricing the flow assigned to each alternative, delays and revenue are linear functions of the pricing coefficient; i.e., in dimensionless form:*

$$\rho_0(a,t) = (1 + a\bar{\mu}_1)\rho(t) - a\bar{\mu}_1, \qquad t_0 \le t \le T_0, \tag{18a}$$

$$\rho_1(a,t) = (1 - a\bar{\mu}_0)\rho(t) + a\bar{\mu}_0, \qquad t_0 \le t \le T_0, \tag{18b}$$

$$W_0(a)/W = (1 + a\bar{\mu}_1)\bar{\mu}_0, \tag{18c}$$

$$W_1(a)/W = (1 - a\bar{\mu}_0)\bar{\mu}_1, \tag{18d}$$

$$R(a)/W = a\bar{\mu}_1. \tag{18e}$$

Proof For the flow assigned to each alternative, combining (8) and (16) gives the desired result. For the delays, we notice that on alternative r it is given by (11b) using $\int_{t_0}^{T_0} \pi(t)dt = aW/\mu$, which follows from (17), and simplifies to (18c) and (18d) as sought. In the case of the revenue, from Result 3.1 the

revenue is proportional to $\int_{t_0}^{T_0} \lambda(t)\pi(t)dt$, which integrated by parts gives:

$$\int_{t_0}^{T_0} \lambda(t)\pi(t)dt = A(t)\pi(t)|_{t_0}^{T_0} - \int_{t_0}^{T_0} A(t)\dot{\pi}(t)dt \tag{19a}$$

$$= a\int_{t_0}^{T_0} A(t)(1-\lambda(t)/\mu)dt \tag{19b}$$

$$= a\left(\int_{t_0}^{T_0} A(t)dt - \frac{1}{2}A(T_0)T_0\right) \tag{19c}$$

$$= aW \tag{19d}$$

The first term in (19a) is zero because of (15), while (19c) results from $\int A(t)\lambda(t)dt = A(t)^2/2$ and noting that $A(T_0) = \mu T_0$. The revenue is obtained by substituting (19d) into (13), which gives (18e). □

It is interesting to note that all relevant measures of performance in our problem are not only a linear function of a single parameter, a, but also linear functions of all the constants that define the problem: $\bar{\mu}_0, \bar{\mu}_1$ and W.

Imposing nonnegative delays gives the bounds for the pricing coefficient:

$$a_{\max} = 1/\bar{\mu}_0, \qquad a_{\min} = -1/\bar{\mu}_1, \tag{20}$$

which also can be derived by imposing $\rho_0(t) \geq 1$ for $a_{\min}$ and $\rho_1(t) \geq 1$ for $a_{\max}$. Since the revenue is a linearly increasing function of a, it follows that the maximum revenue is $R(a_{\max})$, namely:

$$R_{\max} = \frac{\mu_1}{\mu_0}W. \tag{21}$$

Replacing $a = a_{\max}$ in (18) shows that maximum revenue implies the HOT lane is used at capacity with no queues.

4.1. Optimizing operator objectives

Since all performance measures become analytical under linear pricing, it is a simple matter to optimize any particular objective set by the operator. For example, it follows from Result 4.1 that any objective function $f(\cdot)$ that is a linear combination of delays and revenue, e.g.:

$$f(a) = c_0 W_0(a) + c_1 W_1(a) + R(a), \qquad \text{with } c_0, c_1 = \text{constants}, \tag{22}$$

is also a linear function of the pricing coefficient. Therefore, the optimal solution will be either $a_{\min}, a_{\max}$ or an arbitrary value within these bounds, depending on the sign of $f'(a) = \bar{\mu}_1 W(1 + \bar{\mu}_0(c_0 - c_1))$. Of course, nonlinear objectives are also possible but the optimum reduces to finding the extremum of a scalar function.

Another type of objective could be maximizing revenue while ensuring that the GPL delay does not exceed the HOT lane delay by a factor of, say, r; i.e.: $\max_a R(a)$ subject to $W_0(a) \leq rW_1(a)$. Since $R(a)$ is a linearly increasing function of a, the optimum a, namely a^*, is the highest possible value of a, which in this case is given by the condition $W_0(a^*) = rW_1(a^*)$, or:

$$a^* = \frac{\mu_1 r - \mu_0}{\mu_0 \bar{\mu}_1 (1 + r)}, \tag{23}$$

provided that it is not larger than $a_{\max} = 1/\bar{\mu}_0$. The corresponding revenue $R(a^*)$ is given by (18e), which can be written as $R(a^*) = (r - \mu_0/\mu_1)/(1 + r)R_{\max}$. This implies that under this policy, revenue decreases by a factor of $(1 + r)/(r - \mu_0/\mu_1)$ compared to the maximum revenue policy.

4.2. Other real-time pricing strategies

It turns out that a wide family of real-time pricing strategies that may arise in practice are linear in the arrivals and therefore share the properties outlined in the preceding section. In these strategies, tolls are calculated as linear functions of the traffic conditions on (i) the HOT lane, (ii) the GPL, and/or (iii) all lanes. The appendix shows this when the traffic condition is the delay or the number of vehicles in queue, and Table 1 summarizes the results.

This result extends to any traffic condition that is a linear function of the delay in each alternative $w_r(t)$. They include the number of vehicles in the queue $\mu_r w_r(t)$, travel time $\tau_r + w_r(t)$, pace $(\tau_r + w_r(t))/L$, density $k(t) = k_c + \mu_r w_r(t)/L$; if we assume a linear congestion branch in the flow-density relationship one may also include the flow in congestion $q(t) = (\kappa - k(t))w$, where κ is the jam density and $-w$ is the wave speed. The only difference is the way each one would be implemented in practice. Each strategy would keep track of different traffic variables, such as queue length, delay, density, etc. An operator should choose to track the traffic variables that can be measured more accurately with the available technology. In most cases, it is more reliable to estimate speeds so that a delay-based strategy may be advisable.

4.3. Numerical example

To illustrate our method, consider the HOT problem with the parameter values shown in Fig.7. Tolls are given by (17) and the traffic assignment by (18a), (18b). Fig.7 illustrate the cases $a = a_{\max}$ and a given by (23), which correspond to the scenario of maximum revenue under no constraints, and constrained such that the GPL delay does not exceed the HOT lane delay by a factor of $r = 5$, respectively. It can be seen that in the unconstrained scenario, the GPL users experience all the delay while HOT users enjoy no queues, as expected. In contrast, in the constrained scenario both alternatives are congested with $W_0/W_1 = 5$, and the revenue decreases by a factor of 6, as expected. Notice in part e of the figure that the system input-output diagram for both scenarios is identical, which illustrates that two different pricing strategies yield the same SO solution.

5. Discussion

As stated in the introduction, our definition of social costs does not include the effects that tolls may have on trip generation or departure-time choice. This does not mean that demand is considered inelastic here, at least in the traditional sense: demand is assumed inelastic for each particular rush hour (because users do not know the toll in advance), but next day demand may change due to pricing. In particular, we provided an answer to the question: given the current state of the system, what is the toll that minimizes total system delay until the end of the rush hour? This approach allowed the closed-form solutions and insights derived here, and it may not be far-fetched: empirical results concerning departure time variations due to dynamic pricing on managed lanes that have and untolled alternative are mixed (Ozbay et al., 2006; Lam and Small, 2001).

Not surprisingly, our findings differ sharply from the classical result in the congestion pricing literature that the SO solution is unique with the tolled facility always queue-free. This result applies for both the (i) departure-time choice literature (see e.g., Vickrey, 1969; Arnott et al., 1990; Braid, 1996) and the (ii) static assignment literature that explicitly includes cost elasticity (see e.g., Verhoef et al., 1996; Liu and McDonald, 1998; Small and Yan, 2001; Verhoef and Small, 2004). If the framework proposed here is supplemented with a departure-time and/or cost elasticity component for day-to-day demand variations it remains to be seen if the solution presented

here becomes unique and queue-free in the long run. This is currently being investigated by the authors.

Similarly as in section 3.3, it can be shown that the social (delay) benefits of SO tolling compared to no tolling in the case $\Delta < 0$ are bounded by $(T-\tau_0)|\Delta|\bar{\mu}_0$. It follows that the only case where SO pricing does not provide social benefit is when $\Delta = 0$. In any case, SO tolls have the advantage of providing a revenue source and giving flexibility to the operator where to store the queues, while keeping system costs at a minimum. User costs, however, increase with SO pricing since tolls have the effect of increasing user costs in both alternatives. To see this, note that for ML users the variable part of the total user cost is $W_1(\pi) + R(\pi)$, and for GP users it is $W_0(\pi)$; see Fig. 3b,c. In the case $\Delta = 0$ we can use (18) to show that user costs are linear in a, i.e. $W_0 = (1 + a\bar{\mu}_1)\bar{\mu}_0 W$ and $W_1 + R = (1 + a\bar{\mu}_1)\bar{\mu}_1 W$, which shows that user costs in both alternatives increase with revenue $R = a\bar{\mu}_1$.

The concept of marginal costs for an alternative needs some clarification as it becomes ambiguous when capacity is reached, i.e. the left and right derivatives of the cost function are different. This happens in our problem starting at time t_1 when the ML is first used at capacity and until time t_0. It can be seen in Fig. 4a that during this time interval an extra vehicle to the ML would induce a jump in the marginal cost curve from dashed segment "3" to "4" in the figure (right derivative), whereas one fewer vehicle to this alternative would keep the marginal cost constant at τ_1 (left derivative). Disambiguity can pose some difficulty when implementing the congestion pricing principle. But this paper has shown that the left derivative is the only one consistent with the desired SO flow pattern.

Linear pricing strategies, as defined in this paper, are intuitive to apply in practice and exhibit appealing properties. They allowed us to derive analytical expressions for all variables of interest for HOT lanes, including revenues and total delay in each alternative, which are linear functions of a single parameter, the pricing coefficient a. How to determine this parameter depends on the operator's objective, as outlined in section 4.1. Note that Pareto efficient and/or equitable pricing strategies could be devised in a similar fashion, and this is currently being investigated by the authors. To implement this strategy all the operator has to do is start charging when total demand exceeds total capacity, and apply (16) until the end of the rush. This mechanism only requires measuring total system demand in real-time, which will vary stochastically and does not need to be known in advance. Linear pricing may be applied when $\Delta \neq 0$ but only up to the time when the

active bottleneck restriction is hit; see e.g. the semi-linear strategy in Fig. 5.

The results from section 3 also apply to the case $\Delta < 0$ if one interchanges t_1 and t_0, and also T_1 and T_0; see Fig. 8. It can be seen from the figure that negative tolls may be necessary to ensure SO, which can be implemented as a credit to ML users. In this case too, maximizing revenue is equivalent to maximizing the difference between GPL delay and ML delay. Therefore, maximum revenue is achieved by eliminating queues from the ML facility while remaining at capacity, up to time T_1 when the facility is no longer used. This differs fromthe case $\Delta > 0$, where one has to allow congestion in both alternatives at T_0. In any case, except for these nuances at the end of the rush, this maximum revenue strategy coincides with common practice, where the toll is set to guarantee a free-flow ML facility. It has to be noted, however, that there are other non-SO strategies that would generate even more revenue, but at the expense of a greater total delay.

Finally, it is worth noting that a natural extension of this framework would be to incorporate driver heterogeneity. The linearity property unveiled in section 4 could be exploited for analyzing driver differences in the willingness to pay. In our formulation, if drivers' value-of-time is a random variable, so is the pricing coefficient a. The conjecture is that Result 4.1 is still valid replacing a by its expected value, for the expected values of flows, delays and revenue. In the same direction, one might want to incorporate a discrete choice model to capture the effects of income and other socioeconomic variables (e.g., Wu et al., 2012). Subsequently, the SO will maximize the expected utility rather than total delay as in this paper. It remains open whether the conclusions obtained in this paper still stand, another topic for our future investigation.

Acknowledgments.

This research was supported by STRIDE/GDOT research project 2012-089S, and by CEDEUS, CONICYT/FONDAP 15110020. The authors would like to thank two anonymous referees and the Associate Editor for their valuable comments and suggestions, which greatly improved this paper.

References

Arnott, R., de Palma, A., Lindsey, R., 1990. Departure time and route choice for the morning commute. Transportation Research Part B 24 (3), 209–228.

Braid, R. M., 1996. Peak-load pricing of a transportation route with and unpriced substitute. Journal of Urban Economics 40 (2), 179–197.

Burris, M. W., Appiah, J., 2004. Examination of houstons quickride participants by frequency of quickride usage. Transportation Research Record 1864, 22–30.

Burris, M. W., Stockton, B. R., 2004. Hot lanes in houston-six years of experience. Journal of Public Transportation 7 (3), 1–21.

Burris, M. W., Ungemah, D. H., Mahlawat, M., Pannu, M. S., 2009. Investigating the impact of tolls on high-occupancy-vehicle lanes using managed lanes. Transportation Research Record 2065, 113–122.

Carey, M., Watling, D., 2012. Dynamic traffic assignment approximating the kinematic wave model: System optimum, marginal costs, externalities and tolls. Transportation Research Part B 46 (5), 634–648.

Finkleman, J., Casello, J., Fu, L., 2011. Empirical evidence from the greater toronto area on the acceptability and impacts of hot lanes. Transport Policy 18 (6), 814–824.

Kuwahara, M., 2007. A theory and implications on dynamic marginal cost. Transportation Research Part A 41 (7), 627 – 643.

Lam, T. C., Small, K. A., 2001. The value of time and reliability: measurement from a value pricing experiment. Transportation Research Part E 37 (2-3), 231 – 251, advances in the Valuation of Travel Time Savings. URL http://www.sciencedirect.com/science/article/pii/S1366554500000168

Laval, J., Daganzo, C., 2006. Lane-changing in traffic streams. Transportation Research Part B 40 (3), 251–264.

Laval, J. A., 2009. Graphical solution and continuum approximation for the single destination dynamic user equilibrium problem. Transportation Research Part B 43 (1), 108–118.

Li, J., 2001. Explaining high-occupancy-toll lane use. Transportation Research Part D 6 (1), 61–74.

Li, J., Govind, S., 2002. An optimization model for assessing pricing strategies of managed lanes. No. 03-2082. Proc., 82nd Annual Meeting of the Transportation Research Board.

Liu, L. N., McDonald, J. F., 1998. Efficient congestion tolls in the presence of unpriced congestion: A peak and off-peak simulation model. Journal of Urban Economics 44 (3), 352–366.

Lou, Y., Yin, Y., Laval, J. A., 2011. Optimal dynamic pricing strategies for high-occupancy/toll lanes. Transportation Research Part C 19 (1), 64–74.

Muñoz, J. C., Laval, J. A., 2005. System optimum dynamic traffic assignment graphical solution method for a congested freeway and one destination. Transportation Research Part B 40 (1), 1–15.

Ozbay, K., Yanmaz-Tuzel, O., Holgun-Veras, J., 2006. Evaluation of combined traffic impacts of time-of-day pricing program and e-zpass usage on new jersey turnpike. Transportation Research Record 1960, 40–47.

Podgorski, K. V., Kockelman, K. M., 2006. Public perception of toll roads: A survey of the texas perspective. Transportation Research Part A 40 (10), 888–902.

Small, K. A., Yan, J., 2001. The value of "value pricing" of roads: Second-best pricing and product differentiation. Journal of Urban Economics 49 (2), 310–336.

Supernak, J., Golob, J., Golob, T. F., Kaschade, C., Kazimi, C., Schreffler, E., Steffey, D., 2002a. San Diegos interstate 15 congestion pricing project: Attitudinal, behavioral, and institutional issues. Transportation Research Record 1812, 78–86.

Supernak, J., Golob, J., Golob, T. F., Kaschade, C., Kazimi, C., Schreffler, E., Steffey, D., 2002b. San Diegos interstate 15 congestion pricing project: Traffic-related issues. Transportation Research Record 1812, 43–52.

Supernak, J., Steffey, D., Kaschade, C., 2003. Dynamic value pricing as instrument for better utilization of high-occupancy toll lanes. Transportation Research Record 1839, 55–64.

Verhoef, E., Nijkamp, P., Rietveld, P., 1996. Second-best congestion pricing: The case of an untolled alternative. Journal of Urban Economics 40 (3), 279–302.

Verhoef, E. T., Small, K. A., 2004. Product differentiation on roads: Constrained congestion pricing with heterogeneous users. Journal of Transport Economics and Policy 38 (1), 127–156.

Vickrey, W. S., May 1969. Congestion theory and transport investment. American Economic Review, American Economic Association 59 (2), 251–60.

Wu, D., Yin, Y., Lawphongpanich, S., H, Y., 2012. Design of more equitable congestion pricing and tradable credit schemes for multimodal transportation networks. Transportation Research Part B 46 (9), 1273–1287.

Yin, Y., Lou, Y., 2009. Dynamic tolling strategies for managed lanes. Journal of Transportation Engineering 135 (2), 45–52.

Zhang, G., Wang, Y., Wei, H., Yi, P., 2008. A feedback-based dynamic tolling algorithm for high-occupancy toll lane operations. Transportation Research Record 2065, 54–63.

Zhang, G., Yan, S., Wang, Y., 2009. Simulation-based investigation on high-occupancy toll lane operations for washington state route 167. Journal of Transportation Engineering 135 (10), 677–686.

Zmud, J., Bradley, M., Douma, F., Simek, C., 2007. Attitudes and willingness to pay for tolled facilities: a panel survey evaluation. Transportation Research Record 1996, 58–65.

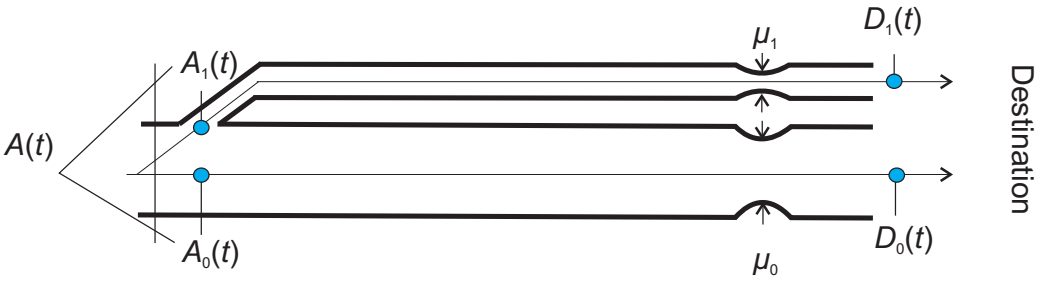

Figure 1: Schematic representation of the network.

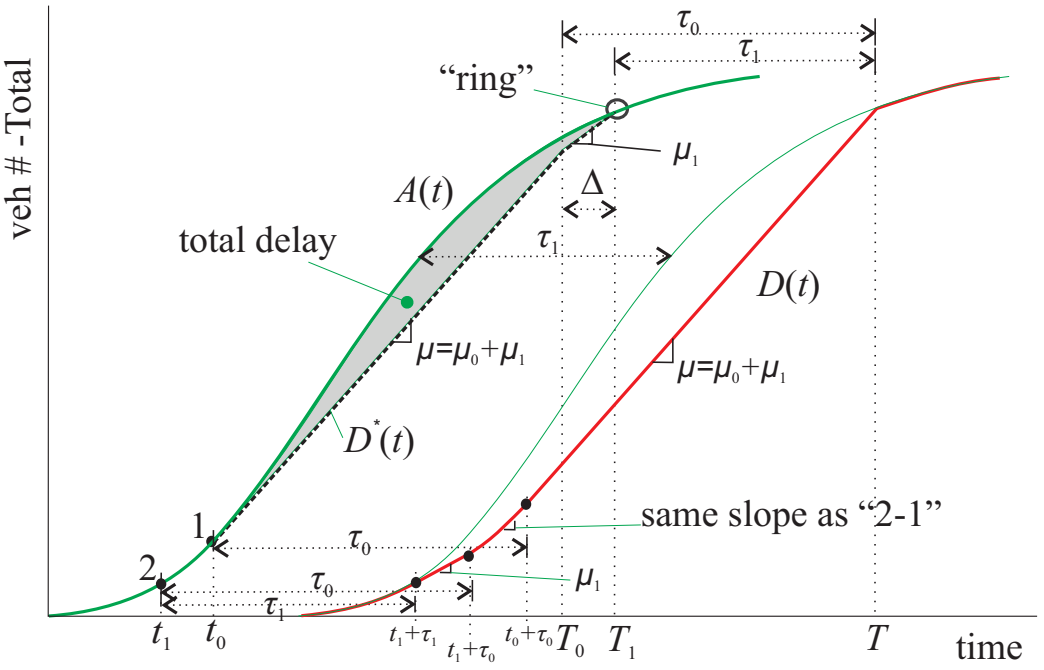

Figure 2: System Optimum input-output diagram.

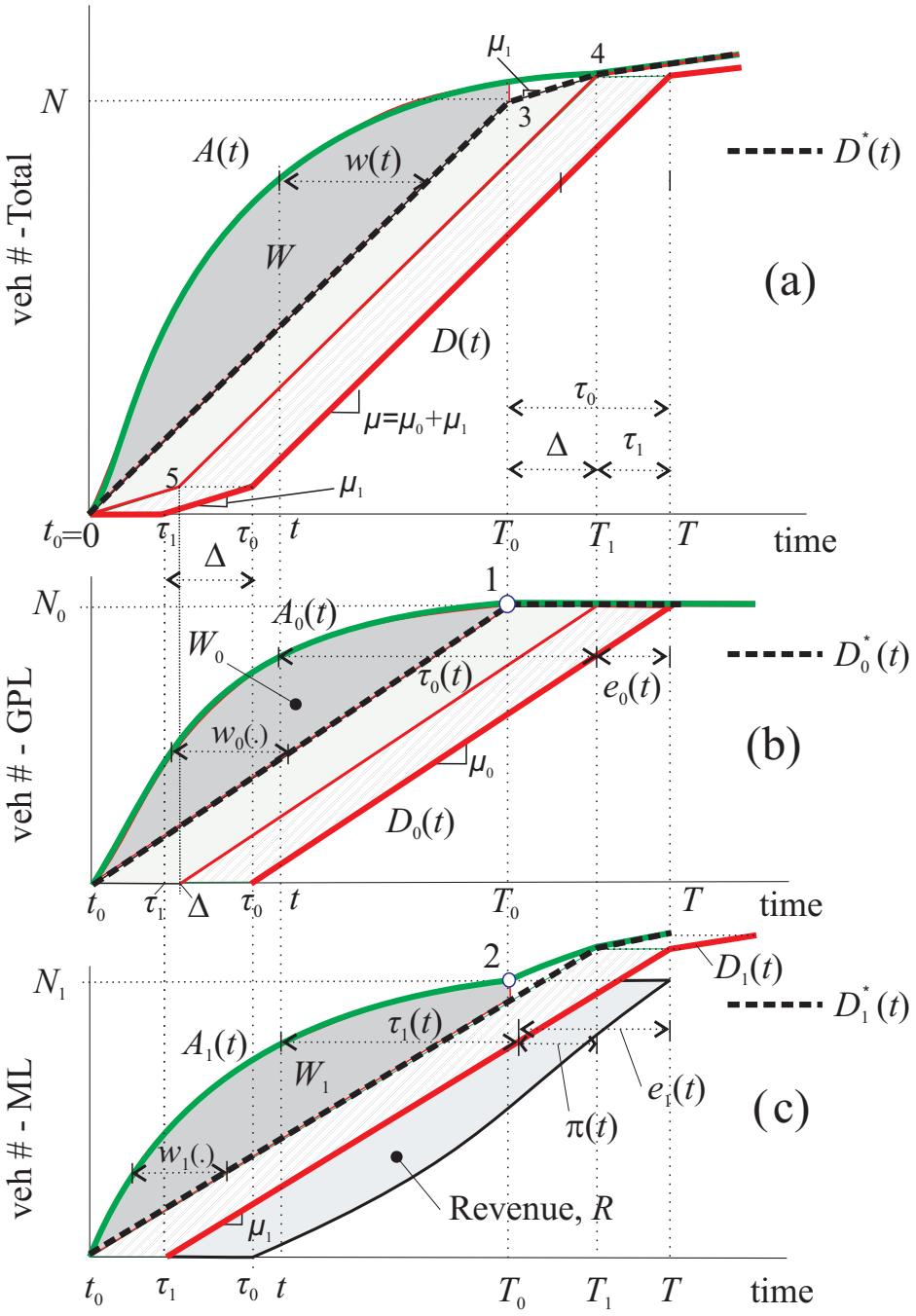

Figure 3: System Optimum input-output diagram for users arriving at $t \geq t_0$.

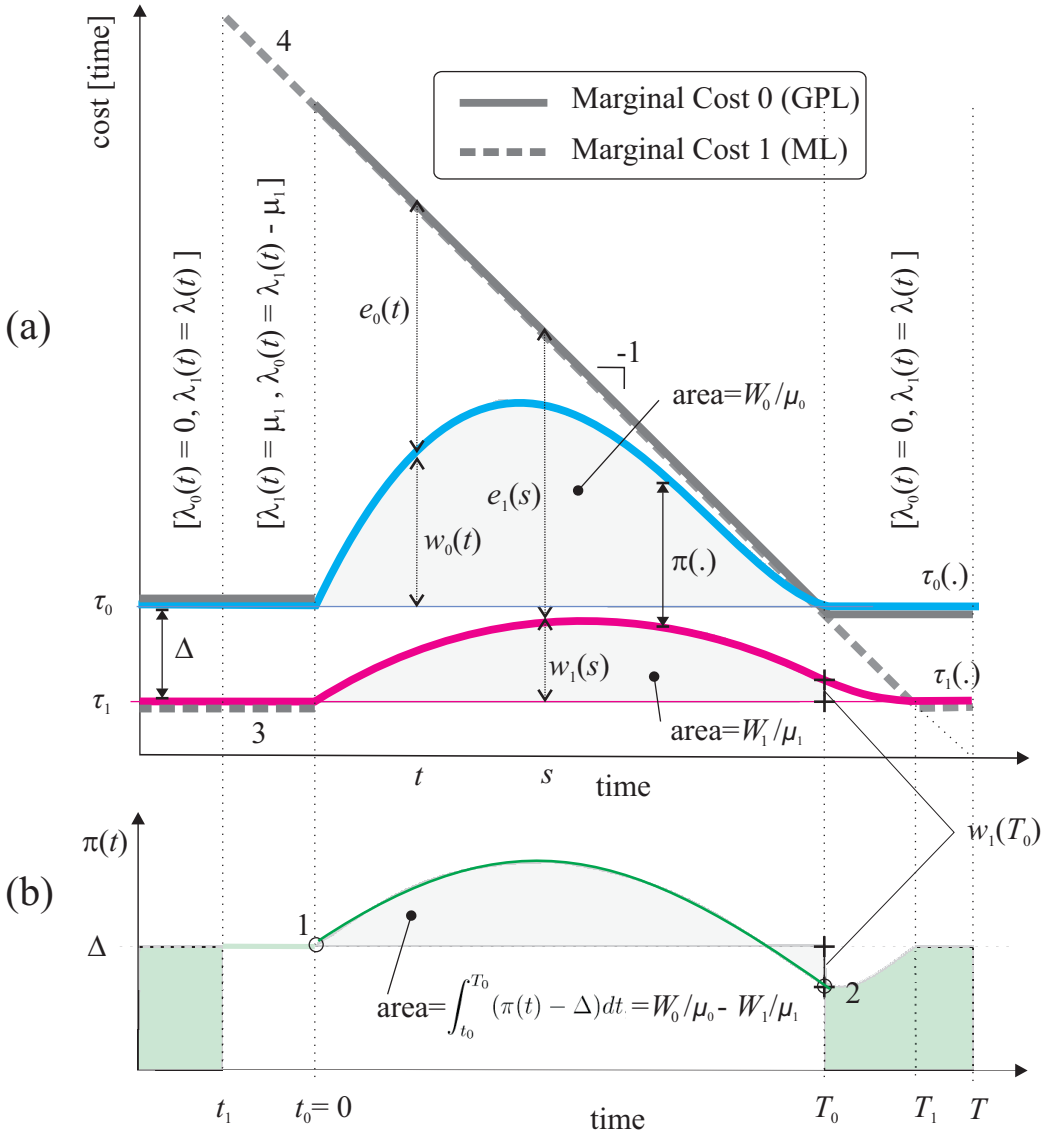

Figure 4: Evolution of the system (a) marginal cost, externality, travel time and (b) toll. The shaded rectangles in (b) indicate that the toll could be anywhere inside this area.

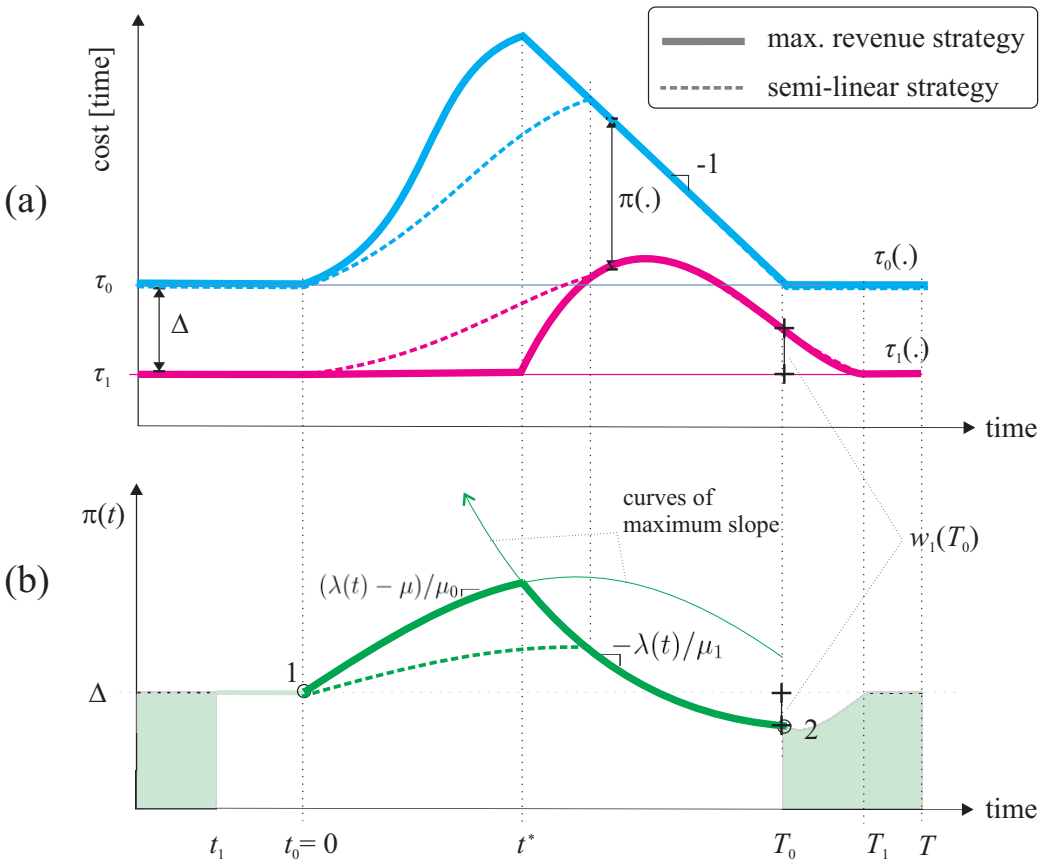

Figure 5: Toll of maximum revenue.

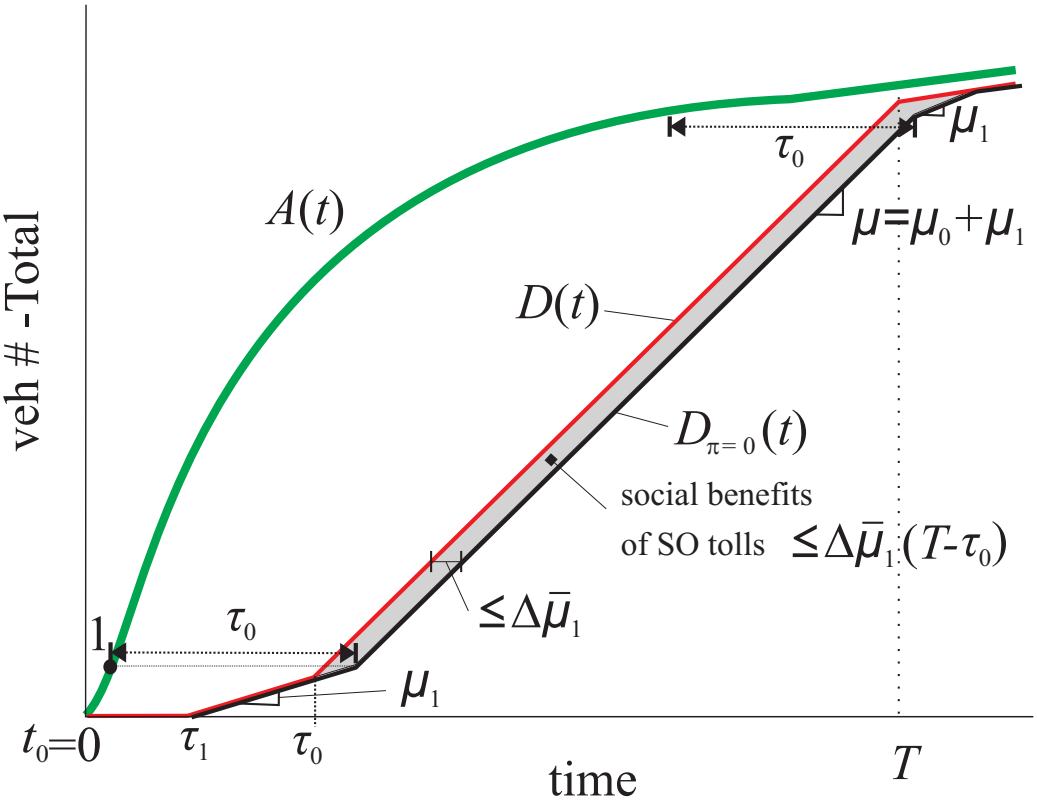

Figure 6: Input-output diagram showing the delay benefits of SO tolling compared to no tolling; case $\Delta > 0$.

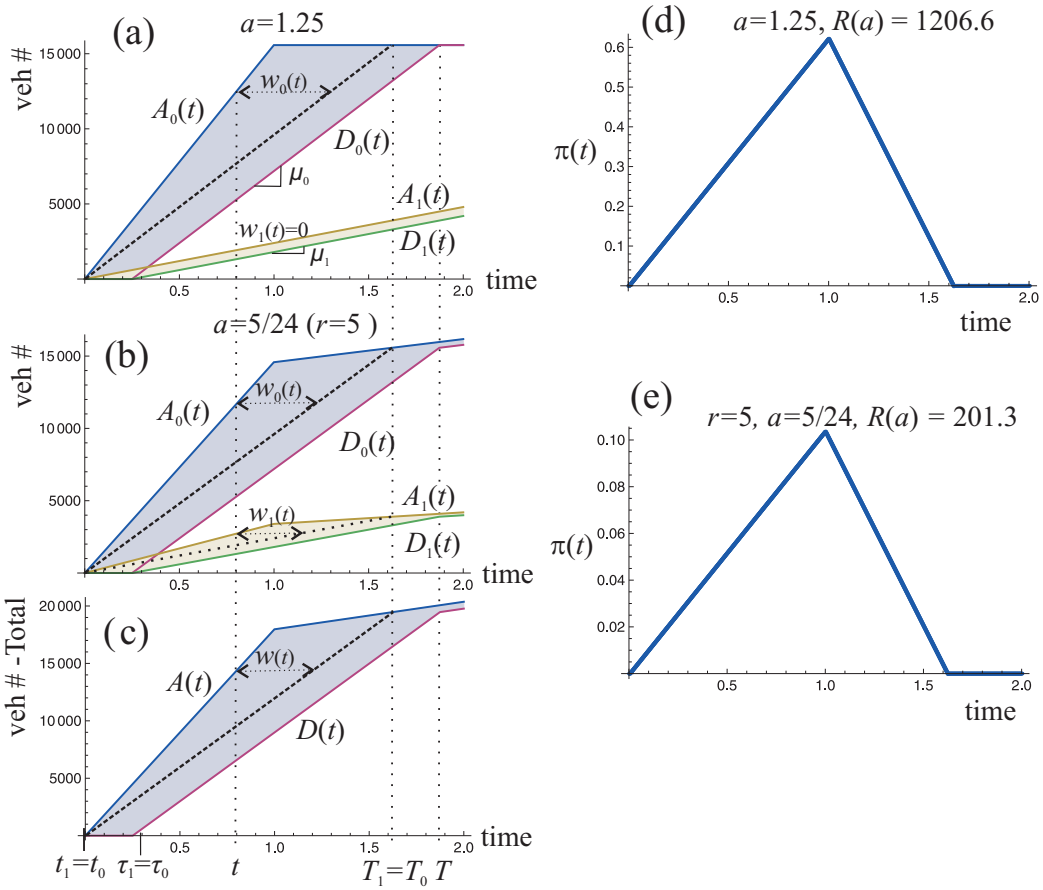

Figure 7: Numerical example. Parameter values: μ_0=9,600 vph, μ_1=2,400 vph, $\tau_0 = \tau_1$ =0.25 hr (Δ=0); the arrival rate $\lambda(t)$ is 18,000 vph in $0 < t < 1$ hr, and 2,400 vph in $t > 1$ hr, where $t_1 = t_0 = 0$hr. (a) input-output diagram for scenario of maximum revenue under no constraints, where $a_{\max} = 1.25$ from (20); (b) input-output diagram for scenario of maximum revenue constrained so that the GPL delay does not exceed the HOT lane delay by a factor of $r = 5$, respectively, where $a = a^* = 5/24$ given by (23); (c) system input-output diagram; (d) and (e) give the toll corresponding to (a) and (b).

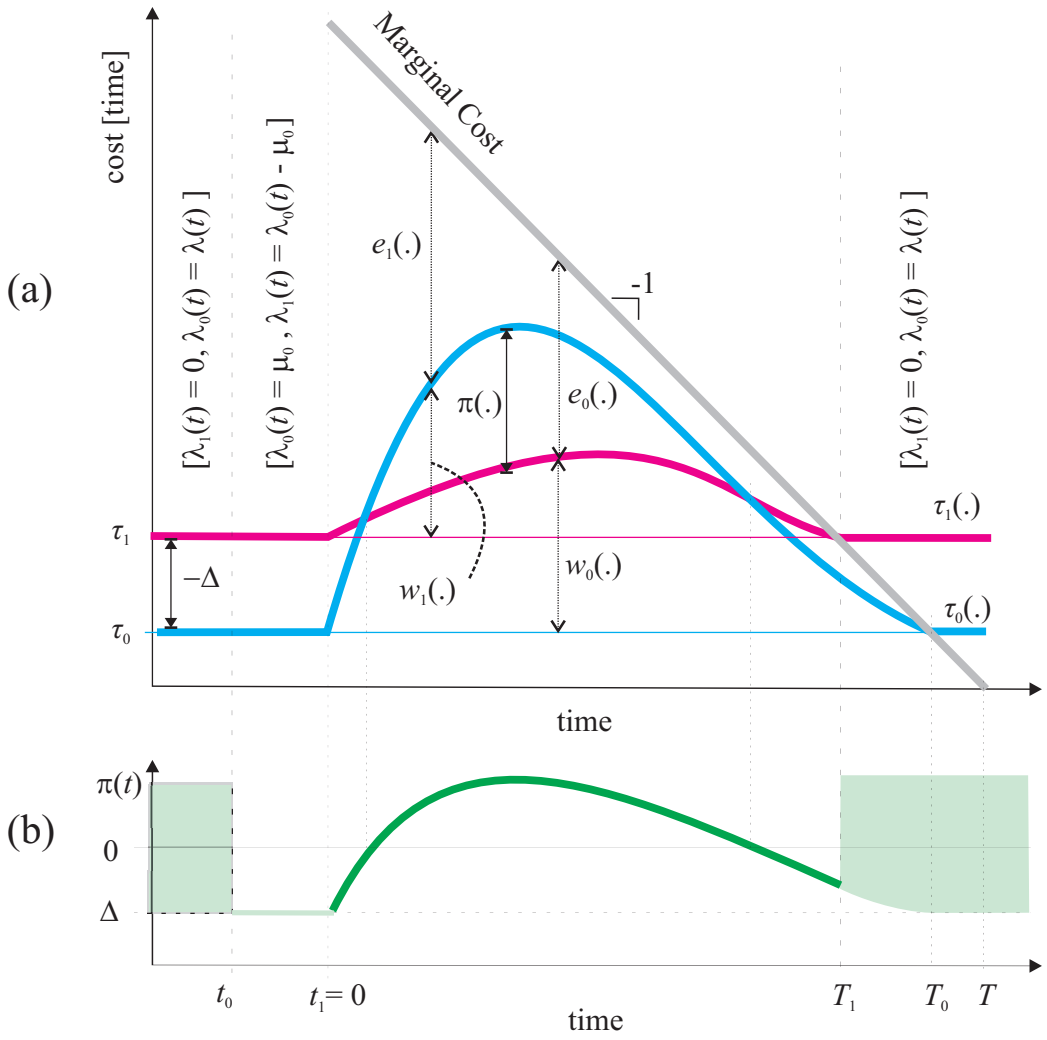

Figure 8: Evolution of the system when $\Delta < 0$: (a) marginal cost, externality, travel time and (b) toll. The shaded rectangles in (b) indicate that the toll could be anywhere inside this area.

Table 1: Pricing Strategies summary

Toll linear in	$\pi(t)$	$\lambda_0(t)/\mu_0$	$c_{\min}$	$c_{\max}$	a
ML queue	$c\left(A_1(t)-\mu_1 t\right)$	$\frac{\lambda(t)+c\mu_1(\lambda(t)-\mu_1)}{\mu+c\mu_0\mu_1}$	$-\frac{\mu}{\mu_1(\mu_0+\mu\mu_1)}$	∞	$\frac{c\mu_1\mu}{\mu+c\mu_0\mu_1}$
GPL queue	$c\left(A_0(t)-\mu_0 t\right)$	$\frac{\lambda(t)-c\mu_0\mu_1}{\mu-c\mu_0\mu_1}$	$-\frac{\mu}{(\mu-1)\mu_0\mu_1}$	$1/\mu_0$	$\frac{c\mu\mu_0}{\mu-c\mu_0\mu_1}$
Queue on All lanes	$c\left(A(t)-\mu t\right)$	$\frac{\lambda(t)+c(\lambda(t)-\mu)\mu_1}{\mu}$	$-1/\mu_1$	$1/\mu_0$	$c\mu$
ML delay	$c\left(\frac{A_1(t)}{\mu_1}-t\right)$	$\frac{\lambda(t)(1+c)-c\mu_1}{\mu+c\mu_0}$	$-\frac{\mu}{\mu_0+\mu\mu_1}$	∞	$\frac{c\mu}{\mu+c\mu_0}$
GPL delay	$c\left(\frac{A_0(t)}{\mu_0}-t\right)$	$\frac{\lambda(t)-c\mu_1}{\mu-c\mu_1}$	$-\frac{\mu}{(\mu-1)\mu_1}$	1	$\frac{c\mu}{\mu-c\mu_1}$
Delay on All lanes	$c\left(\frac{A(t)}{\mu}-t\right)$	$\frac{\mu_0(\lambda-c\mu_1)+\mu_1(\lambda+c(\lambda-\mu_1))}{\mu^2}$	$-\frac{\mu}{\mu_1}$	$\frac{\mu}{\mu_0}$	c

Appendix A. Comparison of Real-Time Pricing Strategies on HOT lanes

This appendix shows that strategies where tolls are calculated as linear functions of the (a) delay or (b) the number of vehicles in queue on (i) the HOT lane, (ii) the GPL, and/or (iii) all lanes, are linear in the arrivals and therefore are in fact mathematically equivalent. Since the methodology is identical in all cases, we illustrate the analysis for one strategy, case (b)(i) above, and summarize all results in Table 1.

Toll linear in the queue on the HOT lane

The queue on the HOT lane at time t is $A_1(t) - \mu_1 t$, and therefore under this strategy we have:

$$\pi(t) = (A_1(t) - \mu_1 t)\, c, \qquad t_0 \le t \le T_0, \tag{A.1}$$

where c is a constant that involves the value of time and has units of time in this case. To obtain the UE assignment under this strategy we substitute

$$\dot{\pi}(t) = (\lambda_1(t) - \mu_1)\, c \tag{A.2}$$

in (8a) and (8b), which leads to

$$\lambda_0(t) = \mu_0 \frac{\lambda(t) + c\mu_1(\lambda(t) - \mu_1)}{\mu + c\mu_0\mu_1} \tag{A.3a}$$

$$\lambda_1(t) = \lambda(t) - \lambda_0(t) = \mu_1 \frac{\lambda(t) + c\mu_0\mu_1}{\mu + c\mu_0\mu_1} \tag{A.3b}$$

Combining (A.2) and (A.3b) reveals that this strategy is linear in the arrivals with pricing coefficient $a = (c\mu_1\mu)/(\mu + c\mu_0\mu_1)$; i.e.:

$$\dot{\pi}(t) = (\frac{\lambda(t)}{\mu} - 1)\frac{c\mu_1\mu}{\mu + c\mu_0\mu_1} \tag{A.4}$$

To ensure that the SO conditions are met at all times $t_0 \le t \le T_0$, we can identify a feasible interval $c_{\min} \le c \le c_{\max}$ similarly as in the main text; i.e., by imposing $\rho_0(t) \ge 1$ and $\rho_1(t) \ge 1$. Finally, Table 1 summarizes the results for all the strategies considered in this appendix. Notice that using $c_{\max}$ and $c_{\min}$ to evaluate a in Table 1 gives the same values for all strategies, and they coincide with (20) as expected.