2015

STRIDE | Southeastern Transportation Research, Innovation, Development and Education Center

Final Report

Comparative Analysis of Dynamic Pricing Strategies for Managed Lanes

Project #2012-089S

Jorge A. Laval, Ph.D., Georgia Institute of Technology
Yafeng Yin, Ph.D., University of Florida;
Yingyan Lou, Ph.D., Arizona State University
Hyun W. Cho, Georgia Institute of Technology

June 2015

Disclaimer

Acknowledgment of Sponsorship

This work was sponsored by a grant from the Southeastern Transportation Research, Innovation, Development and Education(STRIDE) Center, a U.S. DOT Region 4 grant-funded University Transportation Center.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

AUTHORS

Jorge A. Laval, Ph.D., Georgia Institute of Technology; jorge.laval@ce.gatech.edu

Yafeng Yin, Ph.D., University of Florida; yafeng@ce.ufl.edu

Yingyan Lou, Ph.D., Arizona State University; Yingyan.Lou@asu.edu

Hyun W. Cho, Georgia Institute of Technology; hwcho@gatech.edu

ABSTRACT

The objective of this research is to investigate and compare the performances of different dynamic pricing strategies for managed lanes facilities. These pricing strategies include real-time traffic responsive methods, as well as refund options and tradable credit schemes. 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. Performances of pricing strategies are compared by simulation experiments.

EXECUTIVE SUMMARY

The objective of this research is to investigate the performances of different dynamic pricing strategies for managed lanes facilities. This research provide answers to questions such as: what pricing strategies produce the least total system delay? What pricing strategies would produce the maximum revenue while minimizing total system delay? What are the impacts of including a refund option when advertised travel times underestimate trip times experienced by drivers? Can a tradable credit scheme to use managed facility compensate for productivity losses due to congestion?

Linear pricing strategies, as defined in this study, 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. How to determine this parameter depends on the operator's objective. From simulation experiments, linear pricing strategies were proved to be superior to fixed toll pricing strategy to the perspective of the total delay. However, more revenue is collected from the fixed toll pricing strategy than linear pricing strategies.

Approaches to determining optimal operational parameters for a proposed Managed Lane pricing scheme with refund option were investigated. Deterministic utility functions are adopted for each individual traveller with an underlying VOT distribution across the population. A modified point queue model for traffic propagation is developed to account for the intrinsic randomness in traffic flow. An optimization model with a chance constraint is established to determine the desired inflow to the HOT lane during each tolling interval. The relationship among the optimal operational parameters (including the toll rate, the refund amount, the premium for the refund option, the travel time saving guaranteed by the operator) is discussed for two operation paradigms.

This research also proposed and analyzed a tradable credit scheme to alleviate the negative impact of staggered work schedules on firms. The results of a numerical example show that the proposed scheme can act as a relief for the productivity loss resulted from not having all employees at the desired work start time.

CHAPTER 1 BACKGROUND

In the U.S., a prevalent form of congestion pricing is managed lanes, such as express toll lanes, which can be viewed as a first step toward more widespread pricing of congested roads. In a typical setting, lanes on a given freeway are designated either as general purpose or managed toll lanes. The former have no toll while the latter can only be accessed by paying a toll. If high-occupancy vehicles do not need to pay, the lane is widely known as high-occupancy/toll (HOT) lane. Since the first managed toll lane was implemented in 1995 on State Route 91 in Orange County, California, the concept is becoming quite popular and widely accepted by many transportation authorities. Currently, there are 20 managed toll lanes in operation, with more being constructed or planned in the country. To achieve their corresponding operational objectives, managed-lane operators often implement time-of-day or dynamic pricing. In the former, toll rate varies by time of day as per a pre-determined schedule. In the latter, toll rate is adaptive to the real-time traffic condition.

In the research community, 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), only a few studies are devoted to pricing strategies of managed lanes. 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). Yin et al. (2012) compared the pricing algorithm implemented on the 95 Express in south Florida with static and time-of-day tolls. The study suggested that when the demand pattern is predictable, time-of-day or even static tolling could perform as well as dynamic tolling, provided that the toll profiles are optimized for the demand pattern. Nonetheless, dynamic tolling performs in a more robust and stable manner due to its adaptive nature to demand fluctuations. Recognizing that dynamic tolling is beneficial but

more costly to implement, the study further conducted a cost-benefit analysis to examine whether the benefits from dynamic tolling can justify its additional implementation cost or not.

It can be observed that quite a few pricing strategies have been implemented in practice or developed in the literature, but little has been done to compare these strategies and provide guidance on when a particular one should be implemented. Additionally, most of the existing methods are numerical instead of analytical, and therefore little insight can be gained. To fill these voids, the project aims to compare existing and novel pricing strategies to understand the pros and cons of each one.

OBJECTIVE

The objective of this research is to investigate the performances of different dynamic pricing strategies for managed lanes facilities. These pricing strategies include real-time traffic responsive methods, as well as refund options and tradable credit schemes. The focus will be on the traffic dynamics resulting from each pricing strategy and the benefits and costs thereof. The problem will be analyzed from three different perspectives: the users, the tolling authority (i.e., DOT) and the society, which leads to three different performance measures.

CHAPTER 2 RESEARCH APPROACH

Here we briefly describe the main methodological aspects of the following three chapters, each describing different approaches for congestion pricing of managed lane facilities.

PRICING STRATEGIES FOR TOLL FACILITIES

A simplified system configuration is studied analytically while keeping traffic dynamics realistic. This simplified network consists of two parallel links with finite capacity and common origin and destination. While analytical results exist today for both User Optimum (UO) and System Optimum (SO) in the case of constant toll (Muñoz and Laval(2006)), this project generalized this methodology to account for time-dependent tolls as in the below strategies.

Time-of-day pricing

Under this scheme toll rate varies by time of day as per a pre-determined schedule. Typically, the hourly flows over a rolling horizon are examined to identify time periods when the facility is oversaturated, in which case the toll rate is set marginally higher. Conversely, the tolls for time periods where the flow is lower than a given threshold are marginally decreased.

Real-time pricing based on traffic conditions on the managed lanes and/or general purpose lanes

In this case, the toll rate is adaptive to real-time traffic conditions on the managed lanes and/or general purpose lanes. Our preliminary results suggest that depending on the optimization objective, this strategy may not be optimal in terms of total system benefits and can lead to unstable equilibrium patterns and excessive delays. In fact, an operator willing to guarantee free-flow travel time in the managed lanes may be forced to charge unreasonably high amounts to deter excess demand, which will worsen the conditions on the general purpose lanes and probably underutilize the managed lanes.

Strategies analyzed in the project were implemented numerically in order to obtain solutions for larger networks. This allowed us to conjecture the analytical insights against larger networks possibly containing multiple bottlenecks. Although each strategy required different numerical techniques for its resolution, traffic dynamics were given by the same model in all cases. We used *GTsim*, a simulation package that has been developed by Georgia Tech. *GTsim* includes the latest advancements in lane changing models that are capable of explaining congestion dynamics, which also useful to the users' lane choice behavior and effect.

All the strategies were implemented in each case and simulated under a typical rush hour demand pattern. For each strategy the two different objectives considered here (from perspectives of the society and the tolling authority) were simulated independently. Also, the parameters defining each strategy were optimized separately for each objective; e.g., in the

simplest example of a toll inversely proportional to the speed in the ML, the coefficient of proportionality has to be optimized in each case. For comparing large number of simulation results, we defined a suitable performance measure for each objective.

DYNAMIC PRICING FOR HIGH-OCCUPANCY/TOLL LANES WITH REFUND OPTION

While priced managed lanes provide an alternative travel choice for road users, travelers in general may have a negative attitude towards pricing. One plausible reason is that travelers may not receive the benefits they expected when choosing to pay to use managed lanes due to traffic uncertainties. This strategy is specifically proposed to address this issue. The idea is to offer a "price guarantee option" to a traveler when he or she is choosing to pay for managed lanes. Part of the toll paid by the traveler will be refunded if the travel time saving does not reach the minimal amount guaranteed. The goal of this pricing scheme is to achieve the operational objectives of managed lanes such as desired level of service and sufficient revenue return that can cover option claims, and at the same time enhance travelers' experiences with managed lanes and boost public acceptance of managed lanes pricing.

A TRADABLE CREDIT SCHEME FOR STAGGERED WORK TIME

A new tradable credit scheme is proposed to facilitate the implementation of staggered work schedules in firms. In the scheme, a government agency issues a certain number of mobility credits and charges one credit from any traveler who wishes to enter the central business district where the firms are located during the morning peak period. The mobility credits are directly allocated to the firms, who can either distribute them to their employees or sell them to other firms. Employees without credits will be shifted to a secondary work start time. The proposed scheme is analyzed in a simplified morning commute setting and travelers' equilibrium travel costs are derived using Vickrey's bottleneck model. To analyze the credit market equilibrium, the behavior of firms is characterized by the sensitivity of their productivity to their employees' work start time. Moreover, a problem of finding the optimum number of issued credits is formulated to maximize social benefit.

CHAPTER 3 REAL-TIME STRATEGIES FOR TOLL FACILITIES

MOTIVATION

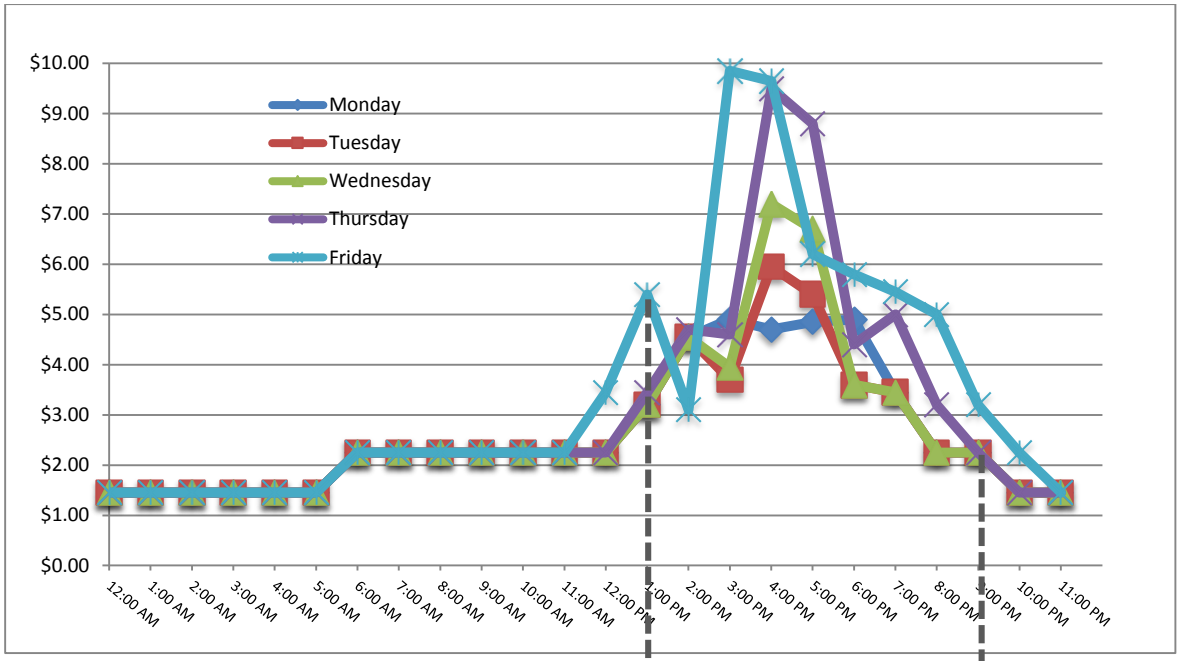
SR-91 Express Lanes were opened on 1995 as a first toll road to apply dynamic congestion pricing in U.S. SR-91 Express Lanes adopt Time-of-day pricing, which toll rate varies by predetermined time schedule. The purpose of designing the toll is to maintain SR-91 Express Lanes traffic flow at free-flow speeds. To accomplish this goal, the toll authority monitors hourly traffic volumes, and adjust the toll every six month if traffic volumes consistently exceed the threshold, which results in managing demand. Figure 3-1a depicts the weekday toll rate for eastbound traffic of SR-91 Express Lanes on July, 2014.

Our preliminary study suggested that SR-91 Express Lanes' Time-of-day pricing strategy appears to be consistent with the theory in Muñoz and Laval(2006), where the marginal costs (expressed in units of time) of an alternative at a given time is equal to the remaining duration of congestion, which decreases linearly. In the theory, the System Optimum toll is imposed by user, which is a difference between each alternative's externalities (marginal cost minus the delay experienced by the user).

From the California Department of Transportation's Performance Measurement System (PeMS) result, we found that SR-91Express Lanes' toll rate is rather similar to the shape of delay that is experienced by users. PeMS Manual defines the delay as "the amount of extra time spent by all the vehicles beyond the time it takes to traverse a freeway segment at a threshold speed." The average delay of the weekday on July, 2014 for the tolled section of SR-91(27~37 Postmile range) is shown in Figure 3-1b.

Although PeMS does not specify the type of lanes for aggregated time series data, we infer that the delay is experienced by general purpose lane users considering the purpose of the toll, which is to maintain SR-91 Express Lanes traffic flow at free-flow speeds. Comparing Figure 3-1a and b, it is clear that time range and peak amplitude of both graphs are similar, we concludes that the toll is also the consequence of the traffic conditions on the ML and/or GP lanes, which will be explained in the next section.

(a) SR-91E Toll Rate



(b) Delay ($V_{threshold}$ = 60mph)

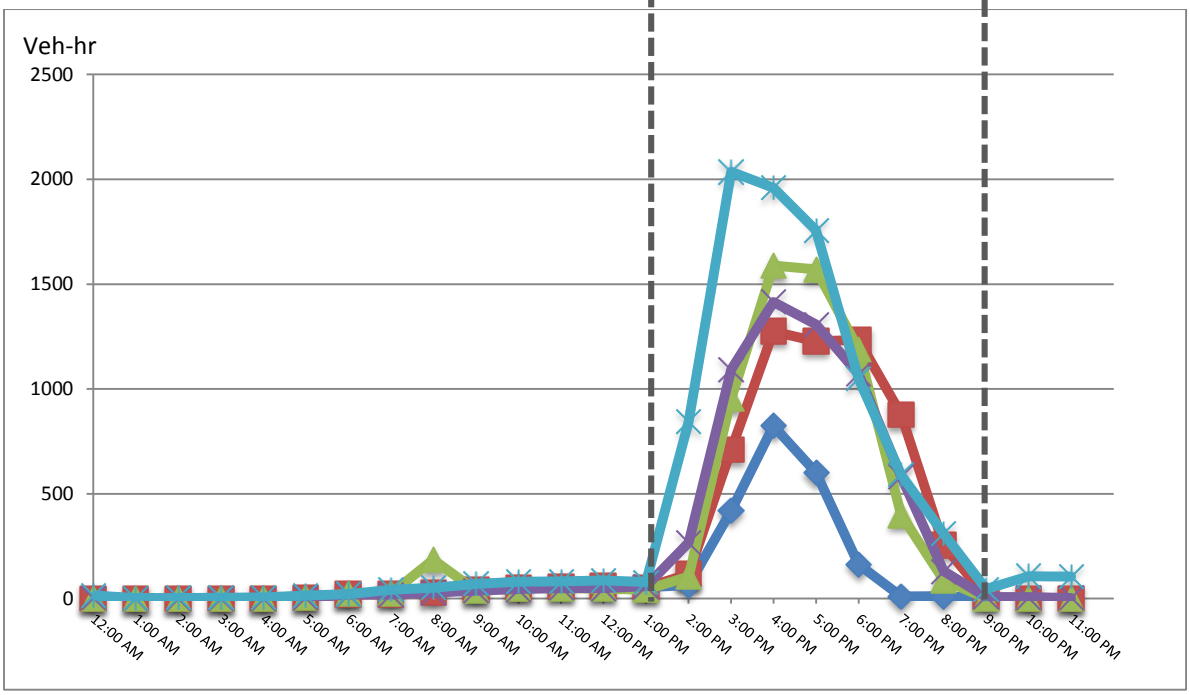


Figure 3-1. (a) SR-91 Eastbound weekday toll rate(July,2014). (b) Average Delay of SR-91 Eastbound weekday (July,2014).

REAL-TIME PRICING BASED ON TRAFFIC CONDITIONS ON THE ML AND/OR GP LANES

Analytical Models

Problem Formulation

Let $A(t)$ be the cumulative number of vehicles at time t that have entered a freeway segment containing a Managed Lane (ML) entrance. All vehicles are bound for a single destination past a General Purpose Lane (GPL) bottleneck of capacity μ_0, which may be bypassed by paying a toll $\pi(t)$ to use a ML that has a bottleneck of capacity μ_1; see Figure 3-2.

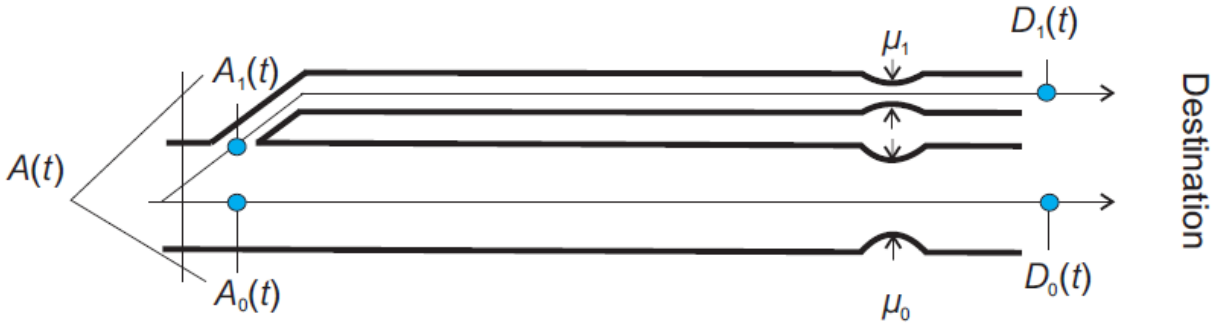


Figure 3-2 . Schematic representation of the network.

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 for maximum generality. It is convenient,

however, to fix the sign of Δ now to simplify the exposition. Let us 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.

System Optimum

The SO solution is presented in Figure 3-4, which shows the system input-output diagram using total arrivals $A(t) = A_0(t) + A_1(t)$ and total virtual departures $D^*(t)$. 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 (2006), 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 Figure 3-3. 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 (2006)):

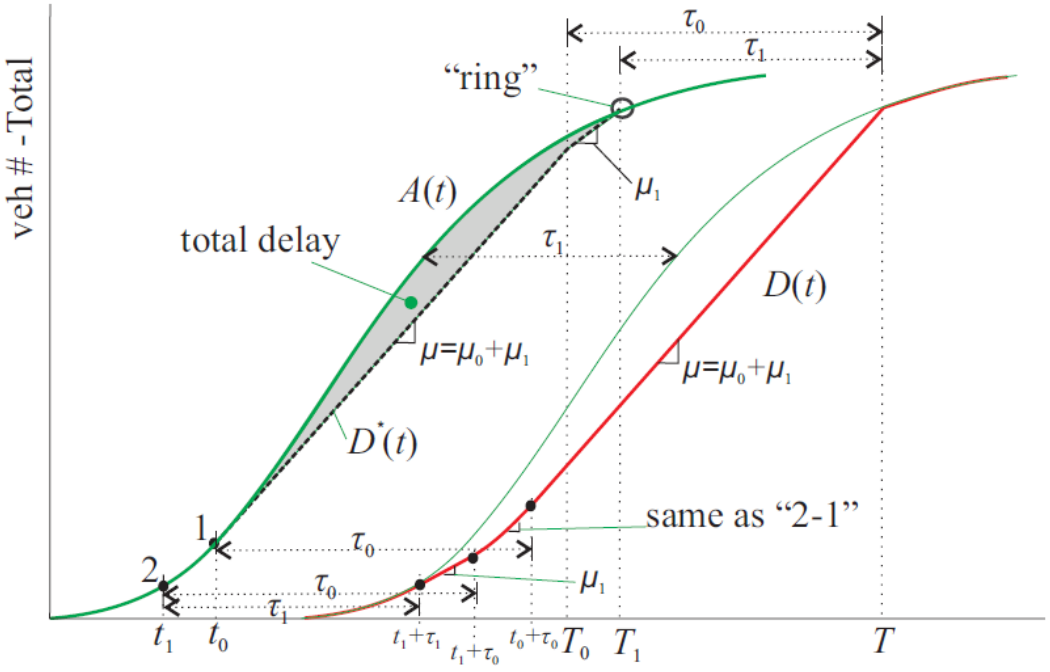


Figure 3-3. System Optimum input-output diagram.

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

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

Notice that these SO conditions say nothing about the alternative-specific arrivals $A_r(t)$; r= 0; 1 in $t_0 \le t \le T_0$, which means that they are not unique in this time interval. Therefore, hereafter we focus on $t_0 \le t \le T_0$ 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 Figure 3-4a, and streamlines the derivation of alternative-specific input-output diagrams in Figures 3-4b,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 Figures 3-5b,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 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 Figure 3-4a: (i) the total delay defined previously (area 0-4-3-0), (ii) the fixed travel time τ_1 incurred by all users (stripped 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 stripped and slightly shaded areas in Figure 3-5a correspond to the sum of the respective areas in parts b and c of the figure.

Figures 3-4b,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 $t_0 \le t \le T_0$ 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 Figure 3-4c 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.

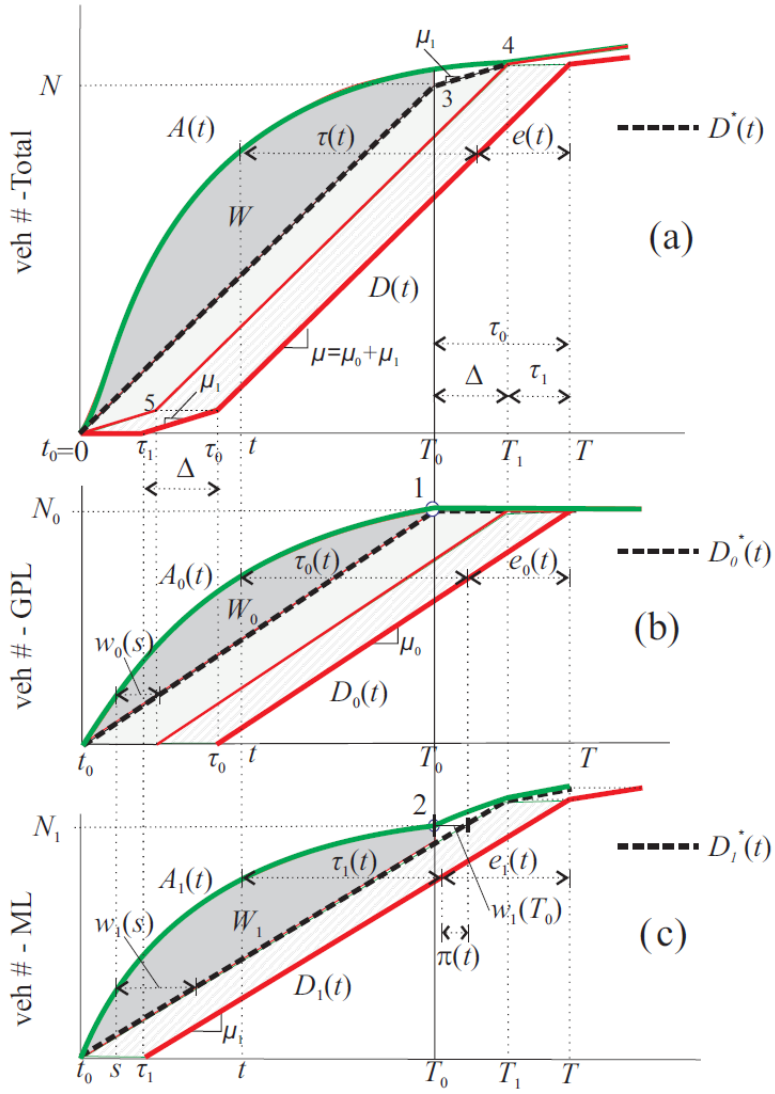


Figure 3-4. System Optimum input-output diagram for users arriving at $t \geq t_0$.

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.

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 $\dot{\tau}_r(t) = \frac{\lambda_r(t)}{\mu_r} - 1$, r = 0,1. This gives in our case:

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

where we have defined the demand-capacity ratios $\rho_r(t) = \frac{\lambda_r(t)}{\mu_r}, r = 0,1$. 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 μ=μ_0+ μ_1 and $\bar{\mu}_r = \mu_r/\mu$ and $\rho(t) = \lambda(t)/\mu$ is demand-capacity ratio. It can be seen that for constant 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:

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

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

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 on 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 Figure 3-5a, 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 $t_0 \le t \le T_0$, where SO flows are not unique, and nor are τ_r (t) and $e_r(t)$. It follows that in the interval $t_0 \le t \le T_0$ the toll $\pi(t)$ is also not unique and can be chosen freely but within the following constraints:

(i) Boundary conditions constraints:

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

(ii) Active bottleneck constraints:

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

$$\dot{\pi}(t) \le \frac{\lambda(t)-\mu}{\mu_0}, \text{ if ML at capacity with no queue} \quad (10c)$$

$$\frac{\lambda(t)}{\mu_1} \le \dot{\pi}(t) \le \frac{\lambda(t)}{\mu_0}, \text{ if } w_r(t) > 0, r = 1,0 \quad (10d)$$

The boundary condition constraints (10a)–depicted as points "1" and "2" in Figure 3-5b– are a consequence of the SO conditions in the time intervals $t \le t_0$ and $t \ge T_0$, which force pricing to be either fixed or arbitrary. Before t_1 there is no congestion and therefore as long as $\pi(t) \le \Delta$ all drivers will choose the ML, as required by the SO condition. This is shown in Figure 3-5b by the shaded rectangles, which indicates that the toll could be anywhere inside this area. During the time interval $t_1 \le t \le t_0$ 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 $t_0 \le t \le T_0$ and under all situations. In particular, if there is no queue on alternative r one should impose $\lambda_r(t) \ge \mu_r$ in (8a) or (8b), which gives (10b) or (10c). If there is a queue on both alternatives, the less restrictive condition $\lambda_r(t) \ge 0$ should be imposed, which gives (10d).

Delays

The Total delay for users arriving in $t_0 \le t \le T_0$, $W = \int_{t_0}^{T_0}(A(t) - \mu t)dt$, is a constant in our problem and is given by the dark shaded area in Figure 3-4a. The delay in each alternative, $W_r(\pi) = \int_{t_0}^{T_0}(A(t) - \mu_r t)dt$, are functions of the pricing strategy. Using (9) gives:

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

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

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

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

which can also be verified in Figure 3-5a. In this figure, the shaded areas correspond to $\int_{t_0}^{T_0} w_r(t)dt = \int_{t_0}^{T_0} \frac{\mu_r w_r(t)dt}{\mu_r} = \frac{W_r}{\mu_r}, r = 0,1$, respectively.

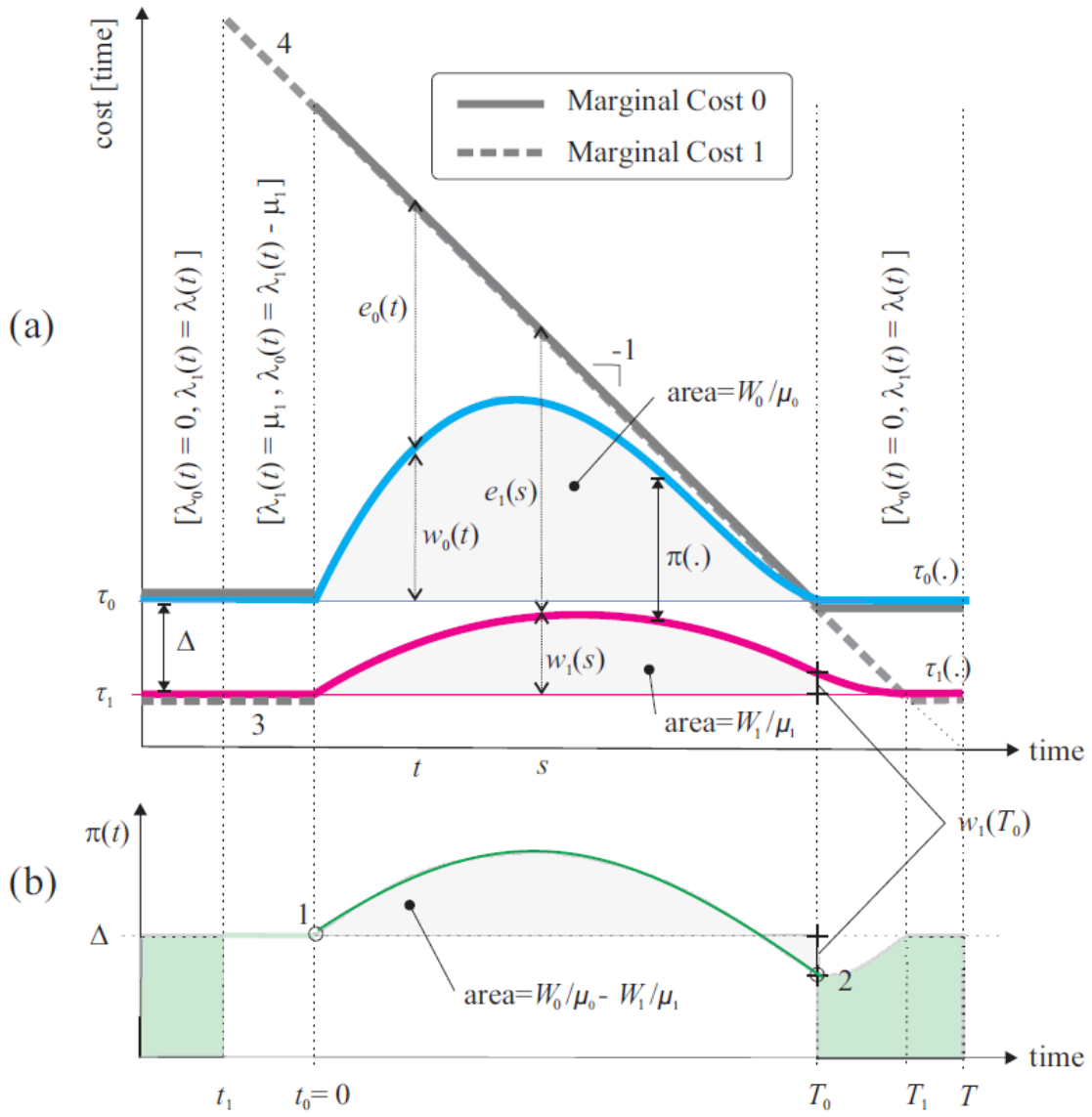


Figure 3-5. Evolution of the system (a) marginal cost, externality, travel time and (b) toll. The shaded rectangles in (b) indicate that the toll could be anywhere inside this area.

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:

$$R(\pi) = \bar{\mu}_1 \int_{t_0}^{T_0} \lambda_1(t)\pi(t)dt - \bar{\mu}_0\mu_1 \int_{t_0}^{T_0} \dot{\pi}(t)\pi(t)dt, \quad (13a)$$

$$= \bar{\mu}_1 \int_{t_0}^{T_0} \lambda_1(t)\pi(t)dt - C \quad (13b)$$

where $C = \bar{\mu}_0\mu_1(\Delta^2 - \big(\Delta - w_1(T_0)\big)^2)/2$ is a constant that follows from $\int_{t_0}^{T_0} \dot{\pi}(t)\pi(t)dt = 1/2\pi(t)^2|_{t_0}^{T_0}$ and (11a). Therefore, maximizing revenue can be expressed as the following mathematical program:

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

and we have the following result:

Result. (*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 Figure 3-6a).

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) λ(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 $\big(T_0, \Delta - w_1(T_0)\big)$, 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 Figure 3-6b, 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 $t_0 \le t \le t^*$ 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 $\frac{W_0(\pi)}{\mu_0} - \frac{W_1(\pi)}{\mu_1}$, which is obtained by imposing the highest and the lowest possible travel time to the GPL and ML, respectively.

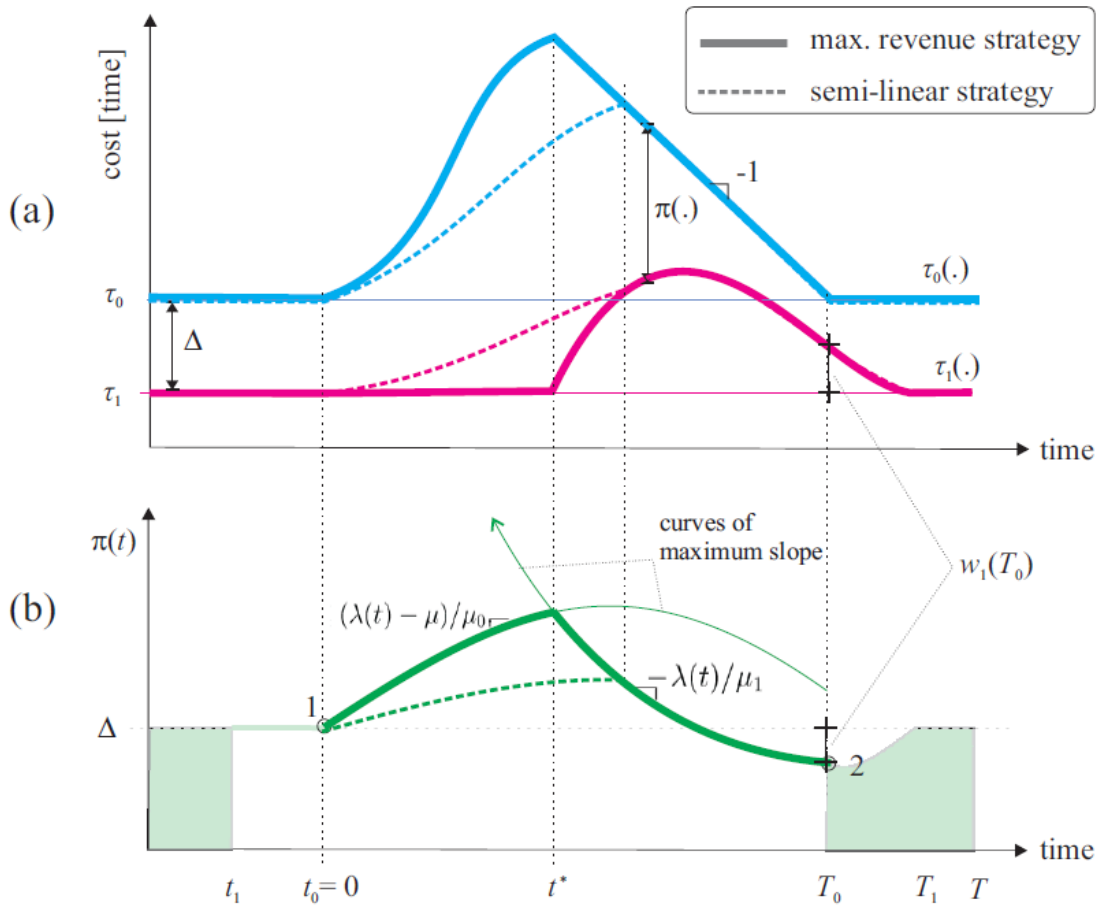


Figure 3-6. Toll of maximum revenue.

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 make such 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 Figure 3-4 that in this case $w_1(T_0) = 0$, and therefore the boundary condition (10a) changes to:

$$\pi(t_0) = 0,\ \pi(T_0) = 0 \qquad (15)$$

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.

Tolls Linear In the Arrivals

We say that tolls are linear (in the total arrival curve) if there is a constant, a, called the **pricing coefficient**, such that:

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

or equivalently (letting t_0=0),

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

Which means that the toll is proportional to the system queue $A(t)$-μt, or to the delay $w(t)=(A(t)-\mu t)/\mu$; see Figure 3-4a. Notice that a is dimensionless. This strategy is "real-time" because from (10) it is clear that to determine the toll at time t, all that needed is the demand-capacity ratio at the same time, which can be measured in real-time.

Result: *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:

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

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

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

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

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

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

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

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

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

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

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:

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

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.

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 Results of Assignment, delays and revenues under linear tolls that any objective function $f(\cdot)$ that is a linear combination of delays and revenue, e.g.:

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

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 $f'(a) = \bar{\mu}_1 W\big(1 + \bar{\mu}_0(c_0 - c_1)\big)$. Of course, nonlinear objectives are also possible but the optimal reduces to finding the extreme 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:

$$a = \frac{\bar{\mu}_1 r - \bar{\mu}_0}{\bar{\mu}_0 \bar{\mu}_1 (1+r)} \tag{23}$$

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 $R(a^*) = (r - \frac{\mu_0}{\mu_1})/(1+r)R_{max}$. This implies that under this policy, revenue decreases by a factor of $= (1+r)/(r - \frac{\mu_0}{\mu_1})$ compared to the maximum revenue policy.

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 3-1 summarizes the results.

Table 3-1. Pricing Strategies Summary

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

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 $q(t) = \frac{\kappa - k(t)}{w}$, 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.

Numerical Example

To illustrate our method, consider the HOT problem with the parameter values shown in Figure 3-7. Tolls are given by (17) and the traffic assignment by (18a), (18b). Figure 3-6 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.

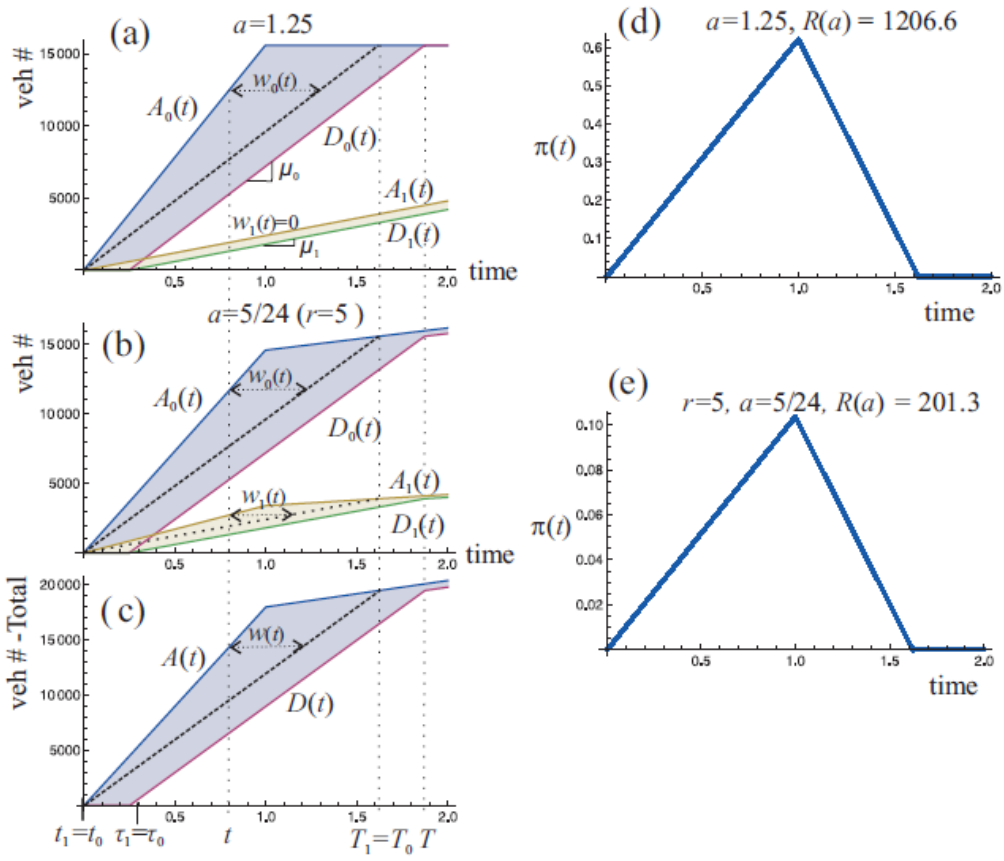


Figure 3-7. Numerical example. Parameter values: μ_0=9,600 vph, μ_1=2,400 vph, $\tau_0 = \tau_1 = 0.25\ hr$ $(\Delta = 0)$; the arrival rate $\mu(t)$ is 18,000 vph in 0 < t < 1 hr, and 2,400 vph in t > 1 hr, where $t_1 = t_0$ = 0hr. (a) input-output diagram for scenario of maximum revenue under no constraints, where a_{max} = 1.25; (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 $a = a^* = 5/24$; (c) system input-output diagram; (d) and (e) give the toll corresponding to (a) and (b).

VARIABLE BOTTLENECK CAPACITY LINEAR TOLL PRICING

In the real-time linear toll pricing strategy, we assumed that the bottleneck capacity (μ_0, μ_1) of general purpose lanes and managed lanes are constant. However, in reality the bottleneck capacity varies with the dynamics of traffic congestion. In this section, we relay this assumption and develop traffic assignment model. In this variable bottleneck capacity model, the predictive travel time is calculated based on the bottleneck capacity at current time *t*,$\mu(t)$, as in Figure 3-8.

$$\tau(t) = \frac{A(t)-D(t)}{\mu(t)} \tag{24}$$

The predicted travel time is the simplest estimation for vehicles arriving at time *t* ,but has an obvious prediction error; see Figure 3-8.

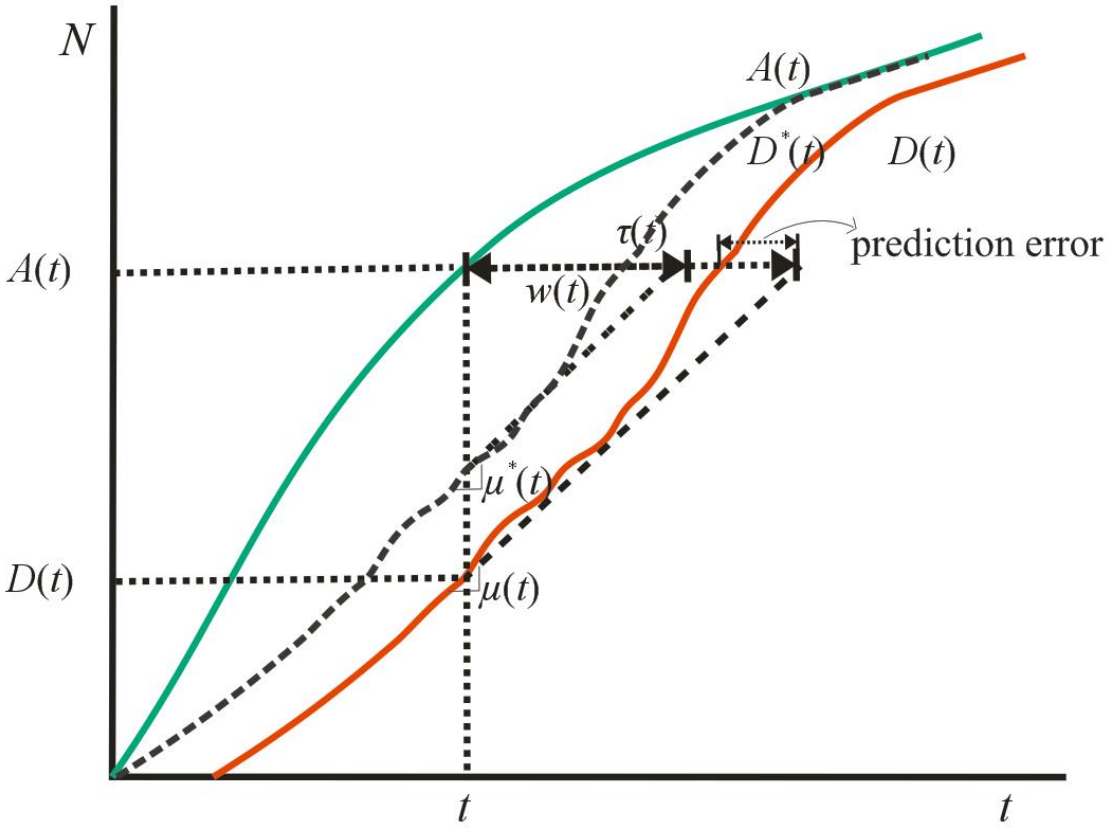


Figure 3-8. Input-Output diagram of variable bottleneck capacity linear toll pricing strategy.

In the variable bottleneck capacity model, the real-time linear toll is now expressed as:

$$\pi(t) = aw(t) = a\frac{A(t)-D^*(t)}{\mu^*(t)}, \tag{25}$$

where $D^*(t)$ and $\mu^*(t)$ are virtual departure and virtual departure rate, and assuming that $\mu(t) \approx \mu^*(t)$, the toll is proportional to the predicted delay at time *t*. Combining (24) and (25) into UE condition (6) in differential form gives us the following traffic assignment equations:

$$\rho_0(t) = \rho(t) + \bar{\mu}_1(t)\dot{\pi}(t) + \frac{\tau_0(t)}{\mu_0(t)}\bar{\mu}_1(t)\dot{\mu}_0(t) - \frac{\tau_1(t)}{\mu(t)}\dot{\mu}_1(t) \quad (26a)$$

$$\rho_1(t) = \rho(t) - \bar{\mu}_0(t)\dot{\pi}(t) + \frac{\tau_1(t)}{\mu_1(t)}\bar{\mu}_0(t)\dot{\mu}_1(t) - \frac{\tau_0(t)}{\mu(t)}\dot{\mu}_0(t) \quad (26b)$$

where $\dot{\pi}(t) = \frac{a\left(\left(\lambda(t)-\mu^*(t)\right)\mu^*(t)-\left(A(t)-D^*(t)\right)\dot{\mu^*}(t)\right)}{\mu^*(t)^2} = a(\rho^*(t)-1) - \pi(t)\frac{\dot{\mu^*}(t)}{\mu^*(t)}$ (26c)

Unfortunately, it is not easy to derive analytical solutions of delays and revenue. Therefore, we will introduce a simulation method to analyze delays and revenues of the variable bottleneck capacity real-time linear toll model.

COMPARISON TO FIXED TOLL PRICING

Fixed toll is the simplest pricing method, but the toll needs to be set reasonable to fully utilize the managed lane. Under the fixed toll pricing, ML is used only when the toll is beneficial. Under UE, drivers will use ML only if the toll is equal or less than the difference between the travel time of GPL and the ML, i.e.

$$\tau_0(t) - \tau_1(t) \geq \pi \quad (27)$$

If the condition is met, the traffic is allocated based on the UE assignment (28) as in (8) (s.t. $\dot{\pi}(t) = 0$); i.e.:

$$\rho_0(t) = \rho_1(t) = \rho(t), \quad (28)$$

The fixed toll pricing model can be interpreted as Laval (2009)'s User Optimum equilibrium, where the fixed toll is the same as Δ_r^* in the paper, which is "a constant travel time-independent of flow incurred when taking off-ramp *r*." In that paper, vehicles in freeway do not divert to off-ramp until delay of freeway is equal to the Δ_r^*. When the delay is as large as Δ_r^*, excess freeway demand diverts to the off-ramp, and when the off-ramp is also congested, traffic is assigned by rule (28). Therefore, under the fixed toll pricing, the capacity of managed lane is "wasted" until the GPL delay equals to the toll. Figure 3-9 depicts examples of the fixed toll pricing scheme at three levels (π_a, π_b, π_c) and their input-output diagram. t_{xr} is the time that vehicles start using the managed lane with toll π_x, and T_{xr} is when congestion ends with toll π_x(*r*=0 GPL, *r*=1 ML). The revenue of the fixed toll pricing can be expressed as $\pi_r \times \int_{t_r}^{t_r^*} \lambda_1(t)dt$, and using (28),

$$R(\pi_r) = \pi_r \bar{\mu}_1 \int_{t_r}^{t_r^*} \lambda(t)dt \quad (29)$$

Delays are also expressed similarly, $W(\pi_r) = \int_{t_0}^{T_r}(A(t) - D(t))dt$, where t_0 is the time when the GPL begins to be congested.

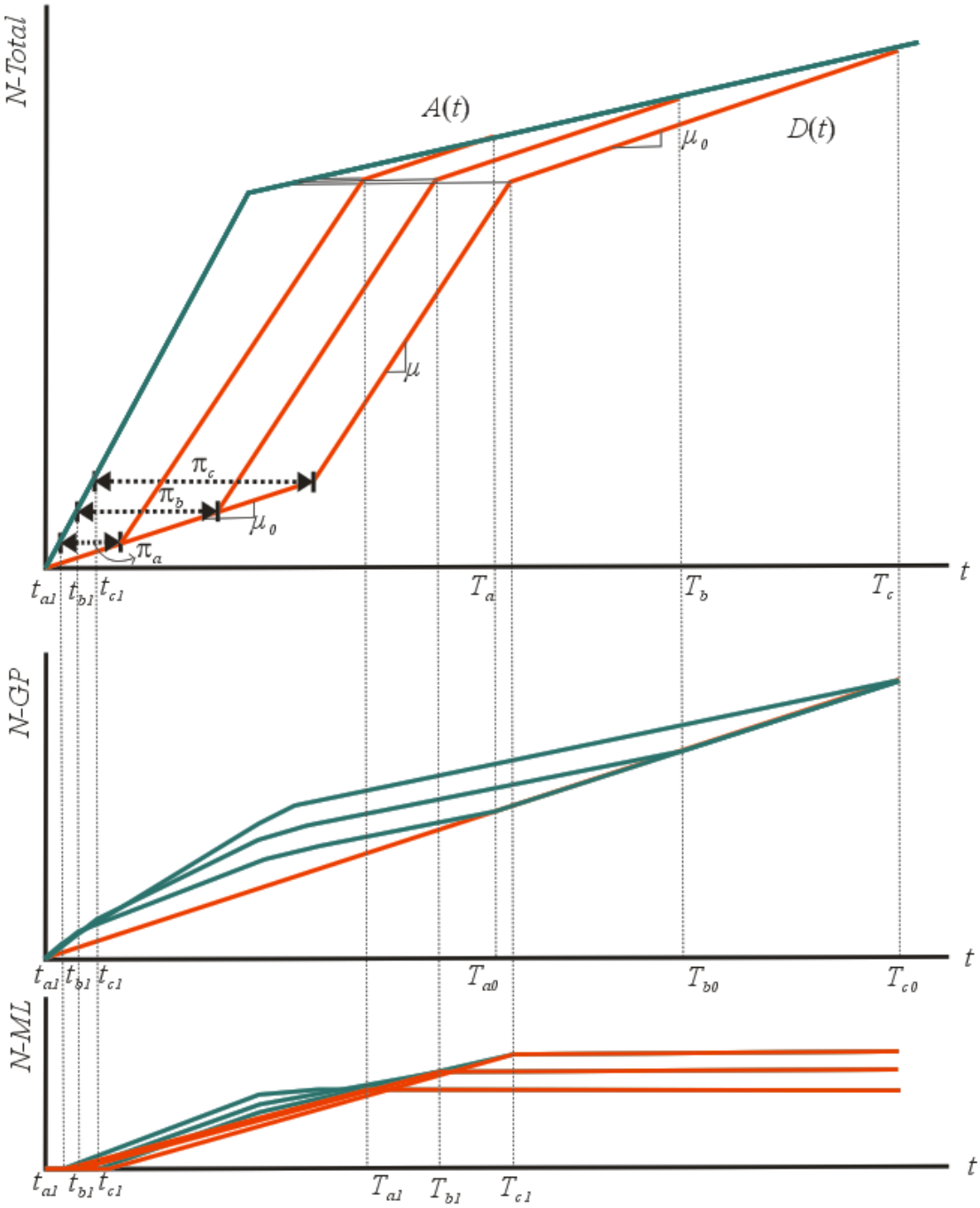


Figure 3-9. Input-Output diagram of Fixed Toll pricing strategy at 3 different levels $\pi_a < \pi_b < \pi_c$.

COMPARATIVE ANALYSIS USNIG SIMULATION

In this section, we compare the two previously developed real-time strategies with a fixed toll pricing strategy using the simulation model *GTsim*. Here after we use the following acronyms to refer to these three strategies:

- FT: fixed toll pricing strategy
- CLT: constant bottleneck capacity linear toll pricing strategy , and
- VLT: variable bottleneck capacity linear toll pricing strategy

It is well known that the existing off-the-shelf traffic simulation software are deficient in simulating congested traffic dynamics on freeways. This is mainly due to the lane-changing models that produce very unrealistic outcomes and which result in overall traffic dynamics that do not work well to empirical observations. Our group at Georgia Tech has developed a micro-simulation application, called hereafter *GTsim*, which is based on the kinematic wave model. This application includes the latest car-following and lane-changing models that replicate bounded vehicle accelerations, realistic lane-change maneuvers, congestion dynamics such as capacity drop and lateral propagation of congestion. The latest mandatory lane-change models used in *GTsim* also replicated spatially realistic lane changes and vehicle accumulations across lanes . *GTsim* was built in JAVA to perform faster than real-time simulation.

As in the analytical framework, the network consists of two parallel roads, general purpose lanes and a High Occupancy Toll lane as in Figure 3-10. In the figure, traffic demand heads eastbound, the top lane is the HOT lane (in blue), and other lanes are general purpose lanes (in red). HOT lane vehicles are inserted to the HOT lane directly at the input section. An exit ramp is set 1*km* downstream of the end of the HOT lane. In this way, mandatory lane-changing maneuvers to exit the freeway from HOT lane users will create a bottleneck of variable capacity. Figure 3-11 depicts the congestion formation on each lane.

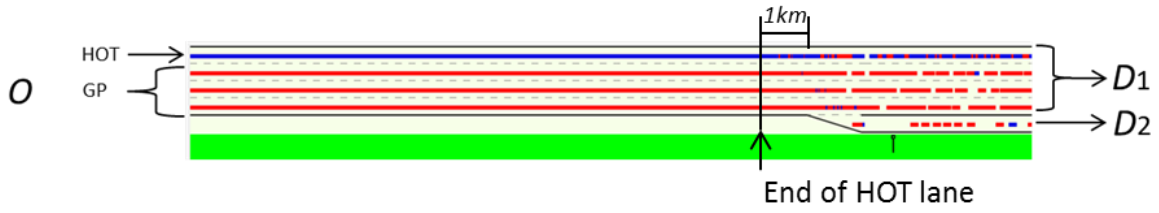


Figure3-10. Diagram of simulation model.

Figure 3-11. Congestion formation of the traffic in the simulation model. Green means free flow speed, and red, congested.

Traffic demand is set as follows: $\lambda(t)$ is 7,500vph in t < 4,800s, and 5,000vph in $t \geq$ 4,800s. Note that ideal capacity (without Lane-changes) is 2,500$vph/lane$. Origin-Destination distribution and portions of through vehicles and exit vehicles are summarized in Table 3-2.

Table 3-2. O/D Distribution of simulation.

Time(s)	From\to	D_1(vph)	D_2(vph)	Exit Veh. Portion
t < 4800	*O*	6,000	1,500	1/5
t > 4800		4,000	1,000	1/5

We simulated two-hour experiments that include the formation and dissipation of queues in all lanes. Tolls and HOT lane assignment are updated every two minutes according to equations (17), (25) and (18a,b), (26a,b) and (28)--, respectively. This means that the toll is constant for two minute period for all cases.

In the fixed toll pricing strategy (FT), we set the toll in units of time from 0.01hr to 0.1hr in 0.01hr interval (10 cases), and traffic is assigned using (28) under condition (27). Note that when the HOT lane bottleneck is inactive, $\mu_1(t)$ is equal to the ideal capacity of 2500 veh/h/lane.

In the constant bottleneck capacity linear toll pricing strategy (CLT), the toll is set using (17), by changing the pricing coefficient "a" in the range 0.1~ 1.0 with a 0.1 interval , and traffic assignments were followed by (18a, b). Although the bottleneck capacity μ_r can change in time every time-step, in this strategy the bottleneck capacity is assumed constant in equations (17, 18a, b).

In the variable bottleneck capacity linear toll pricing strategy (VLT), the toll is decided by (25), also by changing the pricing coefficient "a" as in CLT, and traffics are allocated using (26). This method requires the time derivative of the bottleneck capacity $\dot{\mu}_r(t)$, which we approximate using Euler's method; i.e as the rate of change of the bottleneck capacity within two consecutive time steps.

In the following, we will investigate the performances of each strategy in terms of social cost (delays) and benefits to the operator (revenue). Also, we will verify our results with analytical equations whenever possible.

The total delay (in units of veh/h) — which is composed of GP lane delay (W_0) and HOT lane delay (W_1) — of all pricing strategies are summarized in Figure 3-12. It is found that for all ranges of the parameter *a* in the linear toll pricing strategies (CLT, VLT), the total delay is smaller than that of the fixed toll pricing strategy for all fixed tolls that we experimented. Note that in FT strategy, the total delay tends to increase as the toll increases. This can be explained by Figure 3-9 in the previous section. The total delay of the linear toll pricing strategies seems to random variables drawn from a distribution with constant mean. This is consistent with our theoretical results that indicate that the total delay, W, is a constant independent of the pricing coefficient.

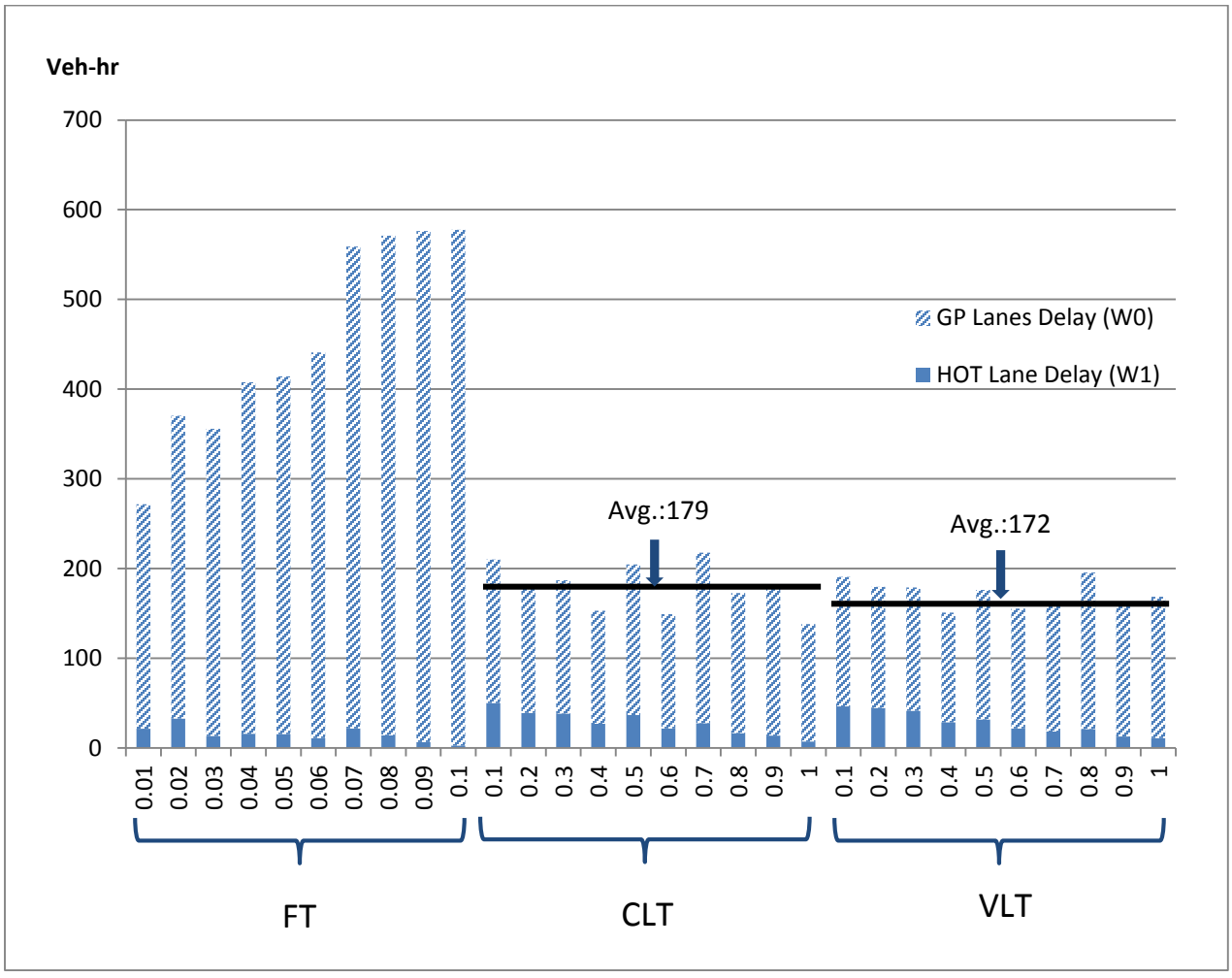


Figure 3-12. Total Delays of Pricing Strategies.

In Figure 3-12, it is interesting to note that HOT lane delays are decreasing as the pricing coefficient a increases in the CLT and VLT strategies. This tendency reminds us of equation (18d) of the linear toll strategy. Converting (18d) into a linear equation in a :

$$\frac{W_1(a)}{W} = (1 - a\bar{\mu}_0)\bar{\mu}_1 = ac_1 + c_0 \qquad (30)$$

To verify (30) with our results, we extracted from the simulation the average values of $\bar{\mu}_0, \bar{\mu}_1$ of each strategy for all cases as in Figure 3-13. Note that these values are measured only when the bottleneck is active.

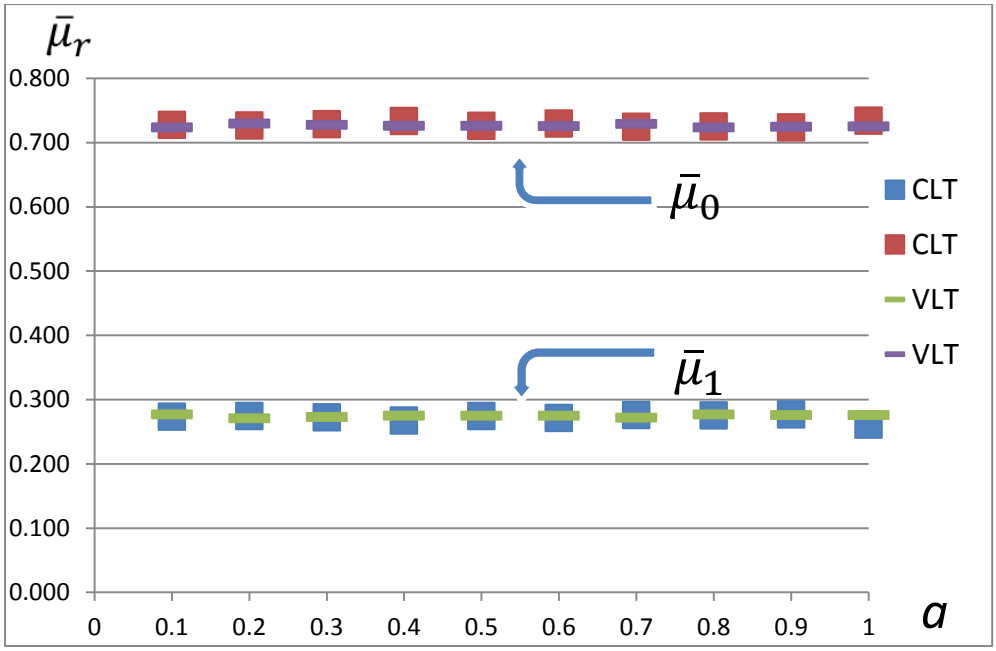


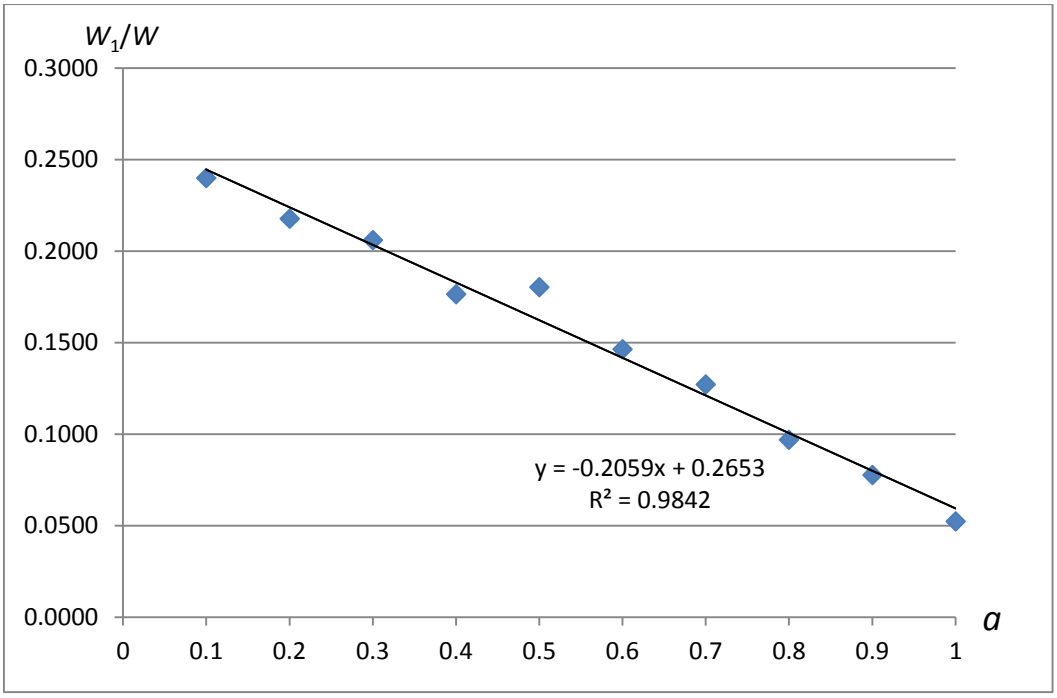
Figure 3-13: Average of $\bar{\mu}_0, \bar{\mu}_1$ in CLT and VLT when bottleneck is active.

The average and standard deviations for $\bar{\mu}_0$ and $\bar{\mu}_1$ across all cases in CLT and VLT were used to compute the 95% confidence intervals for $c_0$and c_1. These confidence intervals were compared to the ones obtained using linear regression with the simulation data shown in Figure 3-14a and b. The results of this analysis are summarized in Table 3-3, where can be seen that the confidence intervals for a given parameter and strategy overlap. This indicates that the theoretical approximations are equivalent to the simulation results.

Table 3-3. Comparing (30) and Simulation Results.

		95% C.I. c_1	95% C.I. c_0
CLT	eq.(30)	(-0.2035, -0.1919)	(0.2603, 0.2833)
	Simulation	(-0.2272, -0.1846)	(0.2521, 0.2785)
VLT	eq.(30)	(-0.2010, -0.1969)	(0.2697, 0.2789)
	Simulation	(-0.2510, -1932)	(0.2653, 0.3011)

(a) CLT



(b) VLT

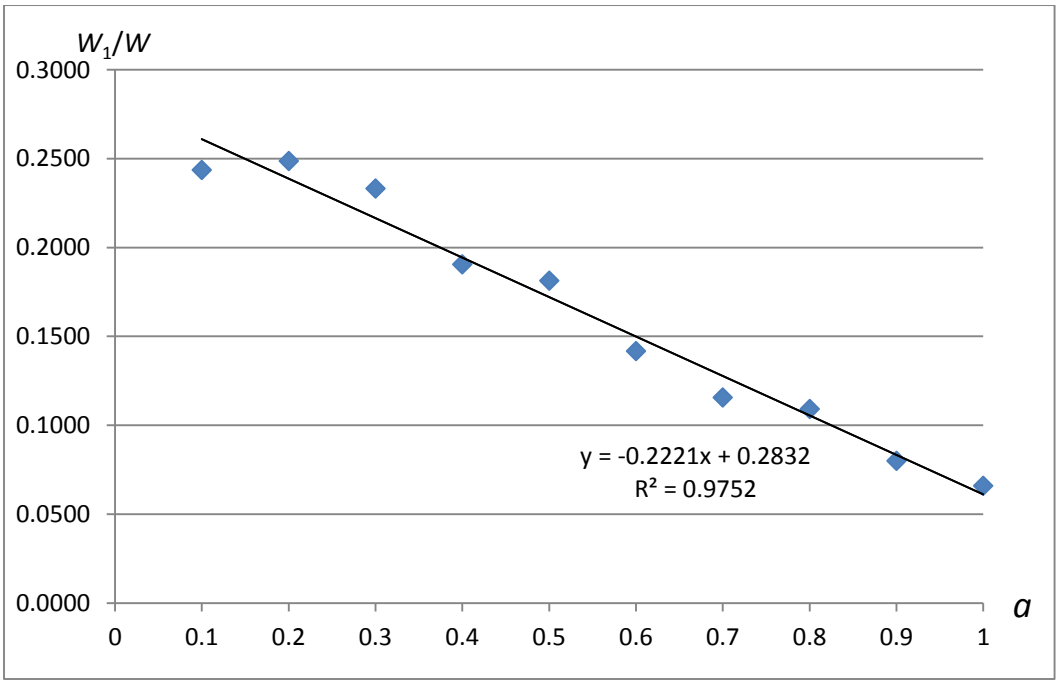


Figure 3-14. Relations of W_1/W and a in (a) CLT and (b) VLT.

The revenue for each pricing strategy is summarized in Figure 3-15. Note that as the toll in our experiment has units of time, the units of revenue are *veh-hr*.

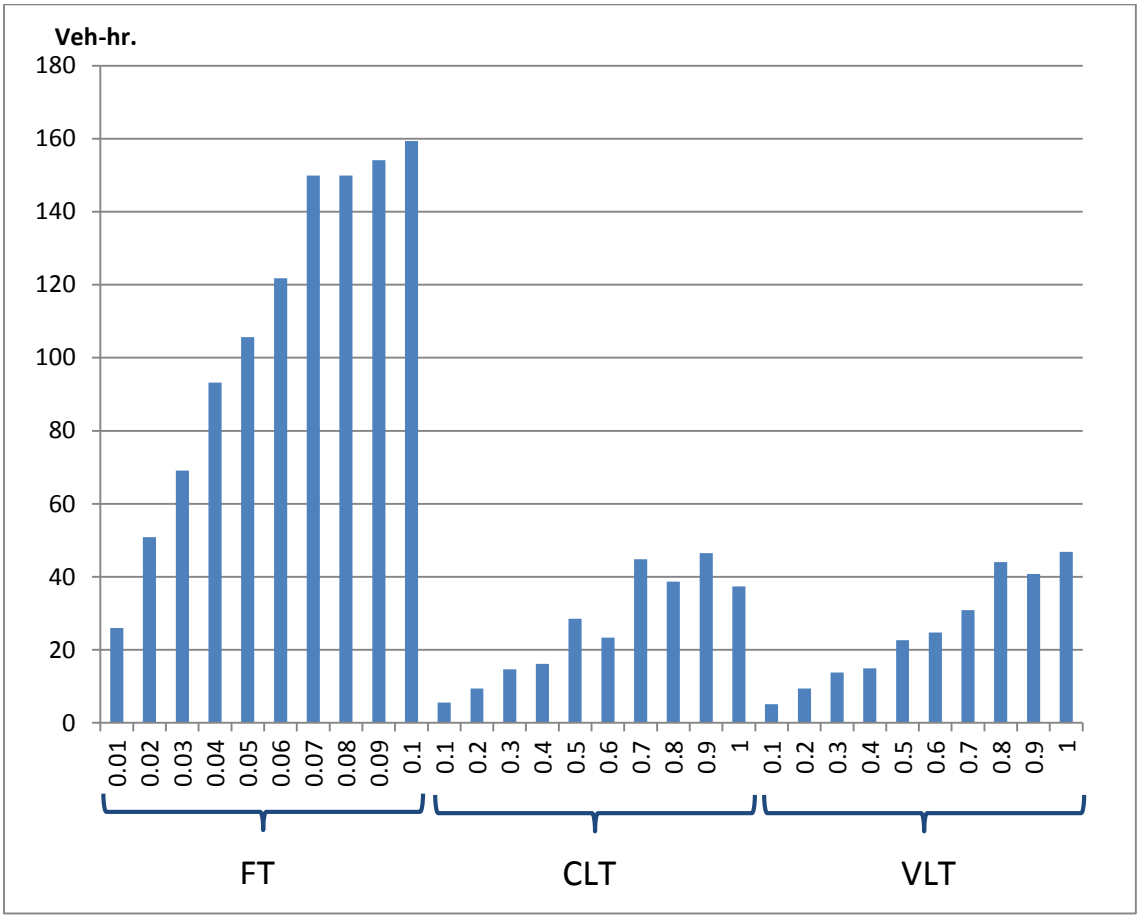


Figure 3-15. Revenues of Pricing Strategies.

It can be seen that the FT revenue is always larger than the revenue of other linear pricing strategies except when the fixed toll is 0.01 *hr*. Specifically, the maximum revenue of the fixed toll pricing strategy is obtained at the highest fixed toll analyzed here , i.e. 0.1 *hr*, and is more than three times of the maximum revenue of the linear pricing strategies. This is not surprising since the FT strategy is not constrained by minimizing total system delay, as is the case with the linear pricing strategies. In fact, it can be seen from Figure 3-12 that the FT strategy imposes very high delays to GPL users compared to the linear pricing strategies.

It is interesting to examine the benefit-cost ratio in our experiments by interpreting the total delay (W) as a social cost and the revenue (R) as a private benefit to the operator. In the case of the linear toll pricing this ratio is given by equation (18e):

$$\frac{R(a)}{W} = a\bar{\mu}_1 \qquad (18e)$$

A relation between *R/W* ratios of the fixed toll pricing and the fixed toll π is depicted in Figure 3-16 below. As in the figure, the below quadratic function is estimated from the data:

$$R/W = -39.114\pi^2 + 6.1286\,\pi + 0.0396$$

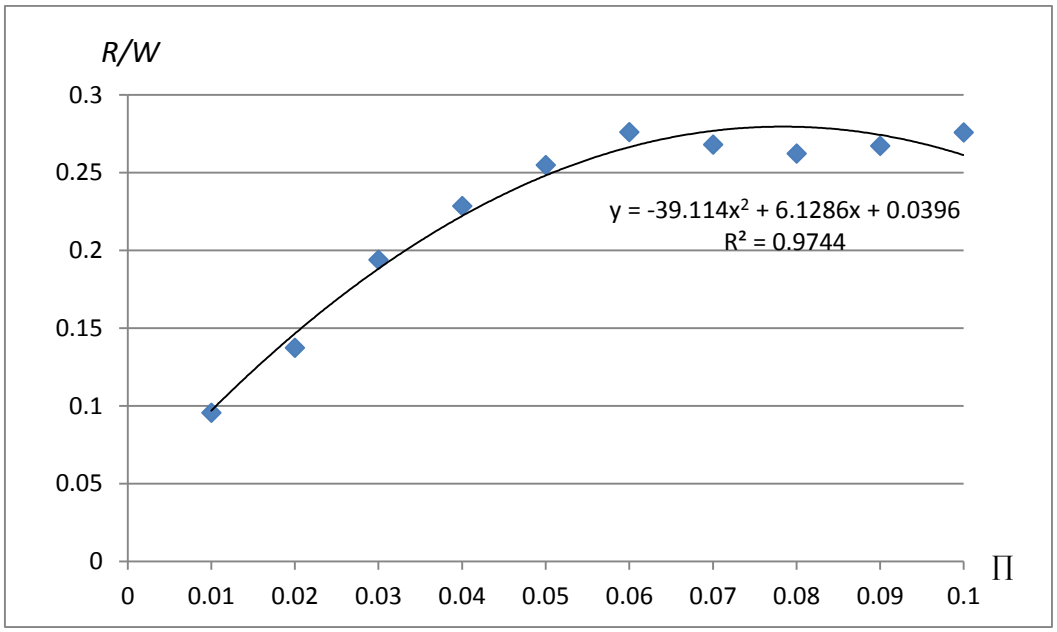


Figure 3-16. Relations of *R/W* and π of the Fixed Toll Pricing Strategy.

For the linear pricing strategies, we compare our results with the analytical equations as in the previous $\frac{W_1(a)}{W}$ analysis. Converting (18e) into linear equation in terms of *a* gives us :

$$\frac{R(a)}{W} = a\bar{\mu}_1 = ac_1 + c_0. \qquad (31)$$

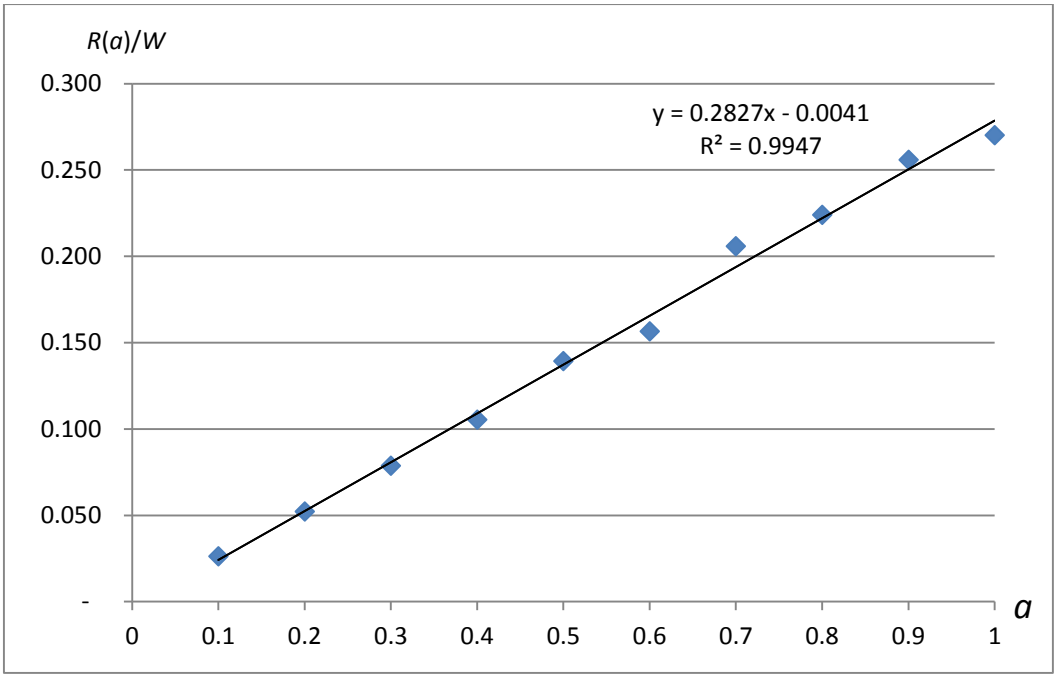
Repeatedly, average $\bar{\mu}_1$ of all cases in CLT and VLT are 0.2718 and 0.2743 respectively. After substituting these numbers into (31), and obtaining coefficients of linear regression equations of simulation data as Figure 3-17a and b, we compared 95% confidence intervals of coefficients; see Table 3-4.

Table 3-4. Comparing (31) and Simulation Results.

		c_1	c_0
CLT	eq.(31)	(0.2603, 0.2789)	-
	Simulation	(0.2658, 0.2996)	(-0.0146, 0.0064)
VLT	eq.(31)	(0.2697, 0.2789)	-
	Simulation	(0.2723, 0.2972)	(-0.0153, 0.0001)

Again, as in the $\frac{W_1(a)}{W}$ analysis, we conclude that our analytical approximations are statistically equivalent to the simulation results.

(a) CLT



(b) VLT

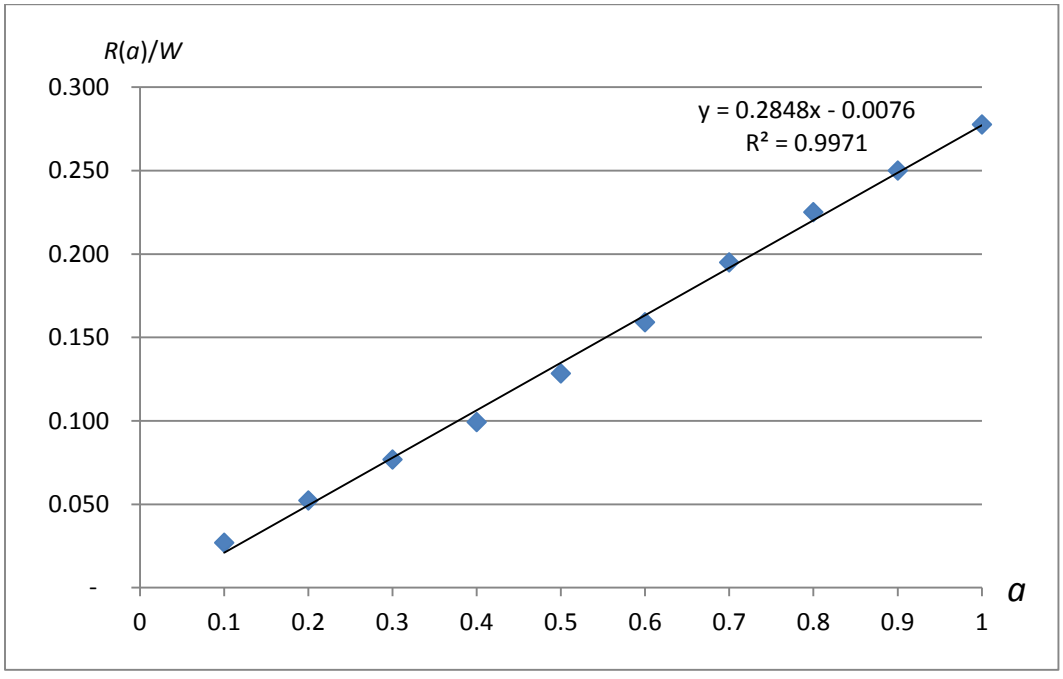


Figure 3-17. Relations of R(*a*)/W and *a* in (a) CLT and (b) VLT.

We compare the relative magnitude delay and revenue produced by the three pricing strategies analyzed here; see Table 3-5. It can be seen that VLT minimizes total delay, that FT maximizes revenue at the expense of a high GP delay and minimizes HOT delay. Finally, it should be noted that CLT and VLT perform very similarly. This indicates that from a practical perspective CLT is more appealing since one does not need to update the estimates of bottleneck capacity.

Table 3-5. Comparison of Pricing Strategies' MoE.

MoE		**FT**	**CLT**	**VLT**
Delay	W	●	◓	◔
	W_0	●	◓	◔
	W_1	○	◓	◓
Revenue	R	●	◔	◔

(Magnitutes of MoE: ○<◔<◓<●)

Sample input-output diagrams, oblique input-output diagrams, departure rates, and $\bar{\mu}_0, \bar{\mu}_1$ by time for each pricing strategy, i.e. π = 0.06 *hr* for FT, and a=0.8 for CLT and VLT are presented in the following Figures 3-18 to 3-21.

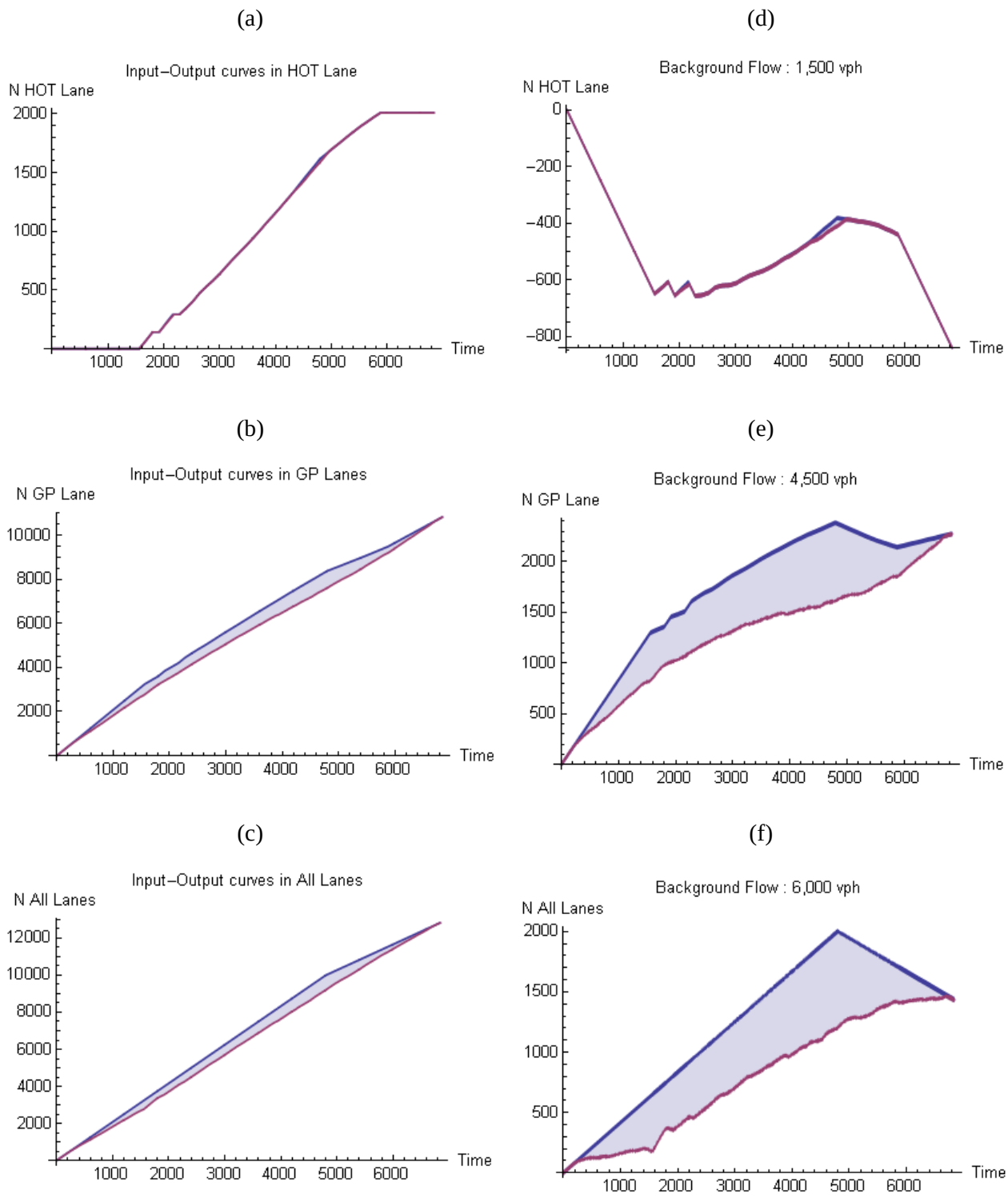


Figure 3-18. Fixed Toll Pricing Strategy's (π=0.06 *hr*) (a) input- output diagram for HOT Lane; (b) GP Lanes; (c) All lanes; (d) Oblique input-output diagram for HOT Lane; (e) GP lanes; (f) All lanes.

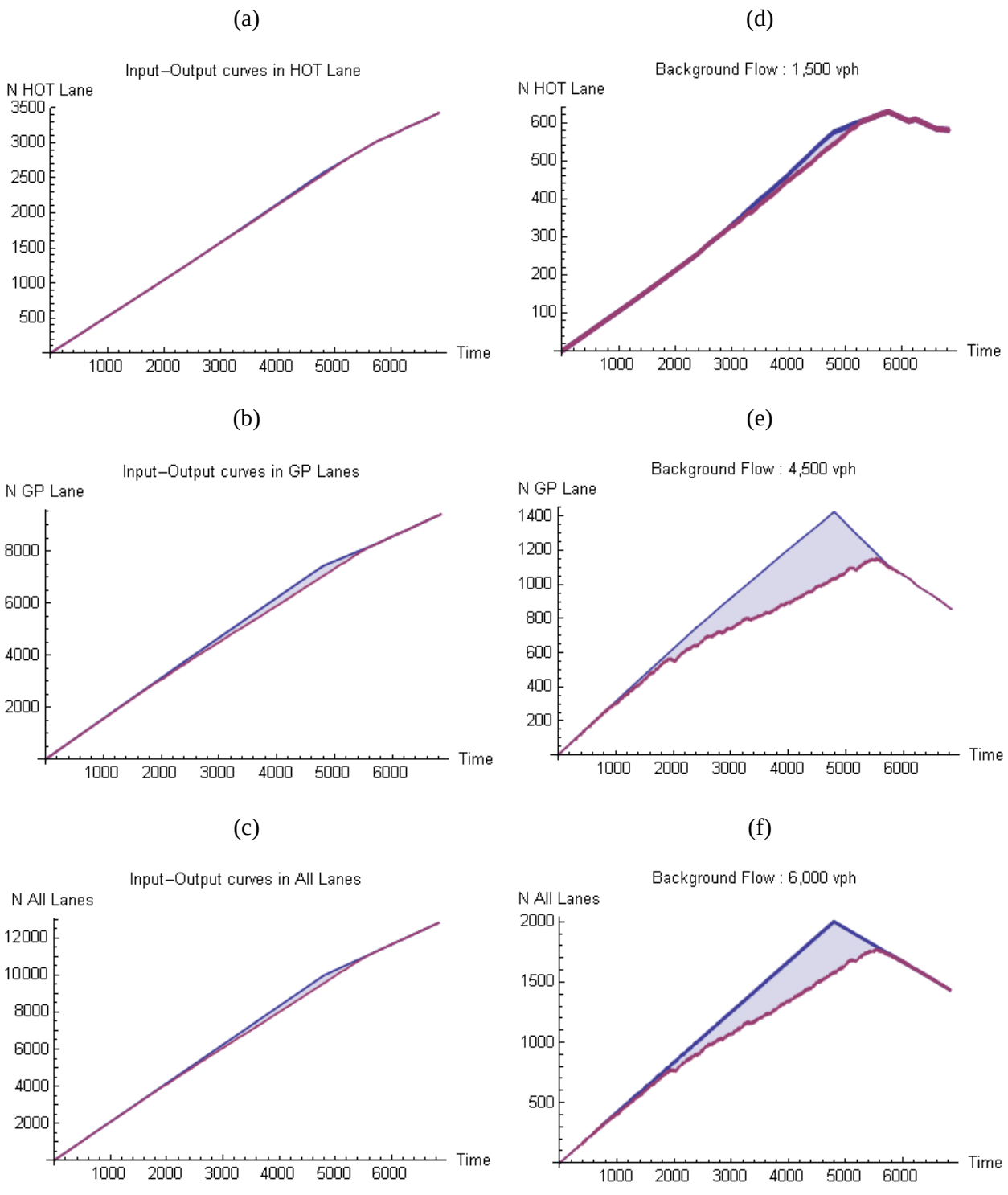


Figure 3-19. Constant Bottleneck Capacity Linear Toll Pricing Strategy's (a=0.8) (a) input-output diagram for HOT Lane; (b) GP Lanes; (c) All lanes; (d) Oblique input-output diagram for HOT Lane; (e) GP lanes; (f) All lanes.

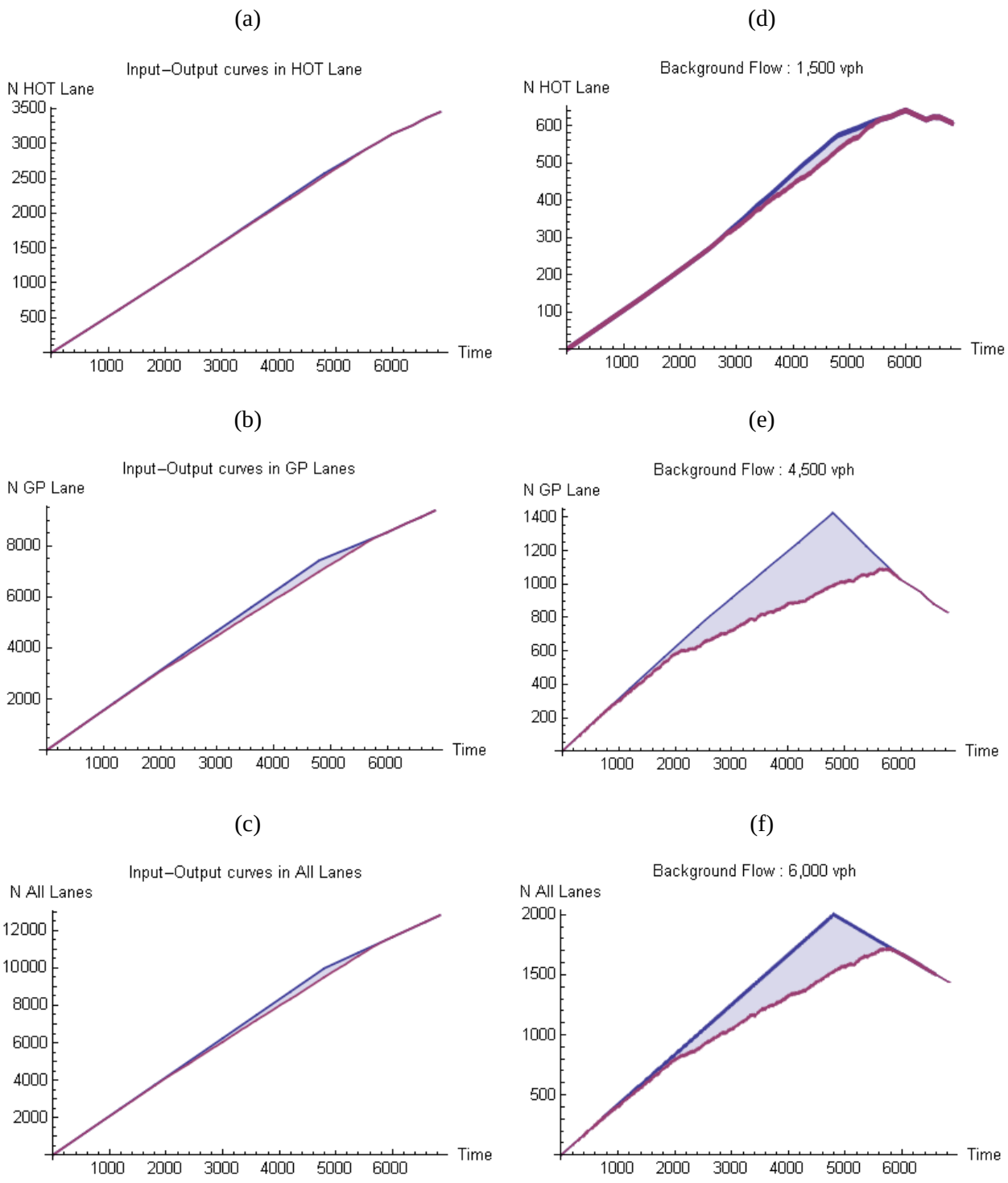


Figure 3-20. Variable Bottleneck Capacity Linear Toll Pricing Strategy's (a=0.8) (a) input-output diagram for HOT Lane; (b) GP Lanes; (c) All lanes; (d) Oblique input-output diagram for HOT Lane; (e) GP lanes; (f) All lanes.

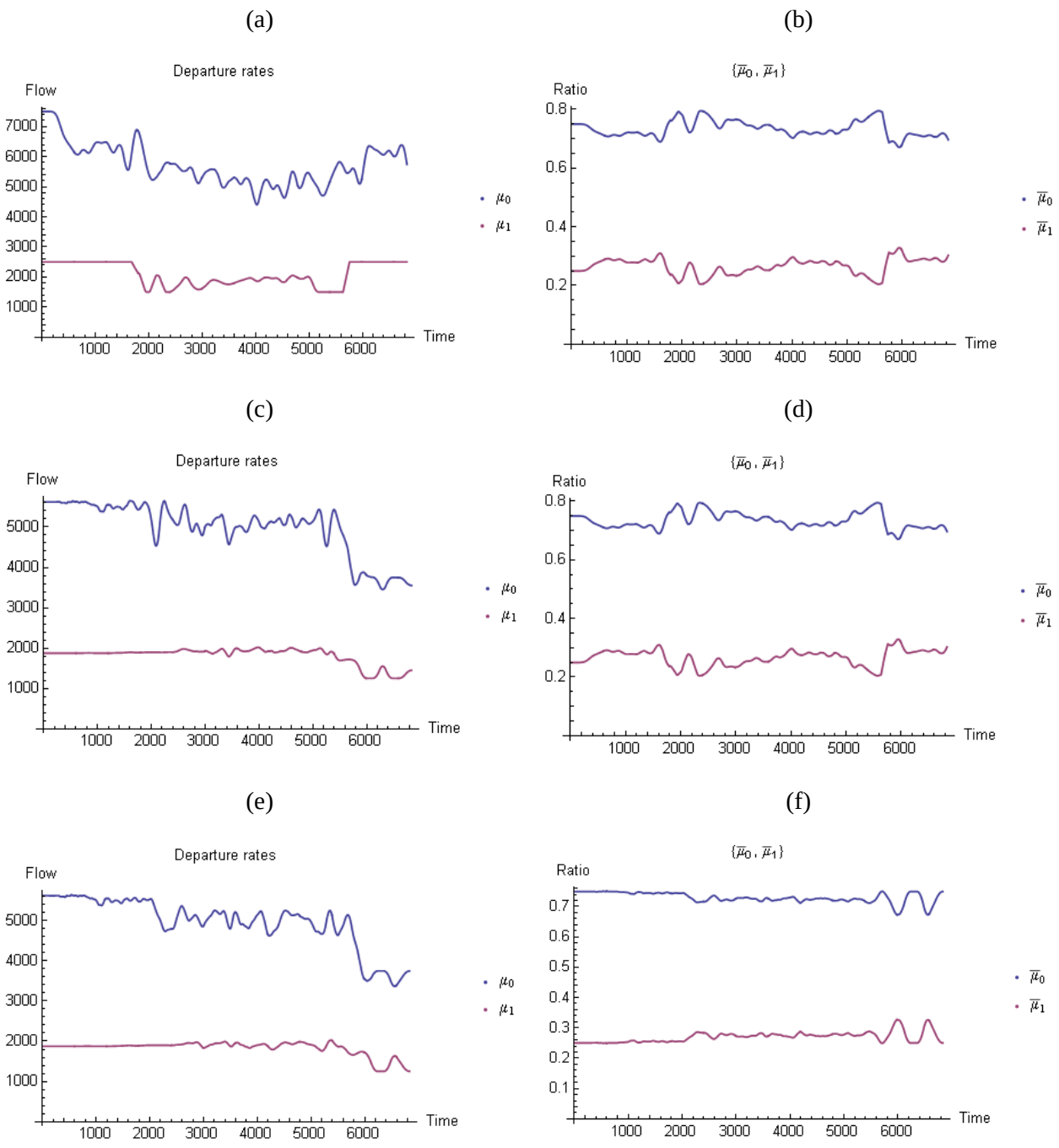


Figure 3-21. (a) Departure rates and (b)$\bar{\mu}_0, \bar{\mu}_1$ of the Fixed Toll Pricing Strategy's (π=0.06 *hr*); (c) Departure rates and (d)$\bar{\mu}_0, \bar{\mu}_1$ of the Constant Bottleneck Capacity Linear Toll Pricing Strategy's (a=0.8); (e) Departure rates and (f)$\bar{\mu}_0, \bar{\mu}_1$ of the Variable Bottleneck Capacity Linear Toll Pricing Strategy's (a=0.8)

CHAPTER 4 DYNAMIC PRICING FOR HIGH-OCCUPANCY/TOLL LANES WITH REFUND OPTION

INTRODUCTION

Operation strategies of managed lanes (ML) often employ a combination of vehicle eligibility and road pricing to manage the demand and to improve the traffic conditions of the facility. While MLs aim to provide an alternative travel choice for road users, travellers in general may have a negative attitude towards pricing of MLs (Ungemah et al. (2005)). One plausible reason is that paid ML users may not receive the benefits they expected due to uncertainties in traffic. This has created a growing challenge of developing innovative pricing strategies to support ML goals that range from operational efficiency to social benefits as well as public acceptance.

In order to improve travellers' experiences with MLs and boost the public acceptance (and thus the feasibility) of ML pricing, this chapter explores an innovative solution by introducing a refund option. When choosing to pay to gain access to MLs, a traveller is offered the chance to purchase an additional refund option. Part (or all) of the toll paid by the traveller will be refunded if the travel time saving does not reach some minimal amount guaranteed by the operator. The goal of this pricing scheme is to achieve the operational objectives of MLs such as desired facility level of service and revenue return that can cover refund claims, and at the same time make MLs appeal more to the travellers.

With the premise that appropriate advanced technologies, such as Connected Vehicle (Research and Innovative Technology Administration, 2014) applications, are in place for ML operators to obtain actual travel time of each individual vehicle, this study investigates approaches to determining optimal operational parameters for the proposed ML pricing scheme with refund option. The operational parameters include the toll rate π, the refund amount r, the premium for the refund option f, the travel time saving guaranteed by the operator $\bar{\tau}$.

METHODOLOGIES

Similar to the author's previous research (Lou (2013), Lou et al. (2011), Yin and Lou (2009)), this study considers dynamic pricing for High/Occupancy Toll (HOT) lanes. HOT lanes are a prevalent form of priced MLs in the US; and to achieve the operational objectives of HOT lanes, ideally the toll rate should be adjusted dynamically in response to real-time traffic condition as well as travellers' willingness to pay (WTP). The success of such operation depends on accurate prediction of travellers' lane choices and estimation of traffic conditions along the facility, as well as carefully designed toll rates. Lou (2013) has proposed a proactive self-learning framework that consists of two critical steps: system inference and toll optimization. Through mining real-time traffic data (such as speed, flow or occupancy) collected

at a regular time interval from loop detectors (often at limited locations), the first step learns travellers' WTP and predicts their lane choices when facing the tolls based on certain lane choice models, delivers a full picture of current traffic condition of the entire facility using certain traffic flow models, and forecasts short-term traffic demand by employing statistical modelling. The attained knowledge up to the current time point will then be used in the second step to determine the optimal toll rate for the next tolling interval in order to achieve the operational objectives of HOT lanes.

The focus of this chapter is the determination of the operational parameters for the proposed dynamic pricing scheme with refund option. The system inference component in the framework (Lou (2013)), therefore, is not considered. Instead, simple models are employed for travellers' lane choices and traffic propagation to allow more in-depth analysis of the proposed innovative pricing scheme.

LANE CHOICE MODEL

It is assumed that each traveller follows a set of deterministic utility functions with her own WTP parameters; and the WTP parameters follows a certain distribution across the population. The following deterministic utility functions are assumed for each individual vehicle not qualified for free access to the HOT lane. U_1, represents the utility of choosing to pay for HOT lane access with the purchase of the refund option; U_1 the utility of paying for HOT lane access without purchasing the refund option; and U_0 the utility of choosing to continue on the general purpose lane.

$$U_1, = -\hat{\tau}_1 - v(\pi + f) + vr \cdot \Pr(\tau_0 - \tau_1 < \bar{\tau})$$

$$U_1 = -\hat{\tau}_1 - v\pi$$

$$U_0 = -\hat{\tau}_0$$

In the above, *v* represents a traveller's WTP, which is essentially the inverse of the traveller's value of time (VOT); τ_0 and τ_1 are the travel times (random due to intrinsic uncertainties in traffic) for the general purpose and the HOT lanes respectively; and $\hat{\tau}_0$ and $\hat{\tau}_1$ are the expected travel times. Note it is assumed that$\hat{\tau}_0$, $\hat{\tau}_1$, as well as $\Pr(\tau_0 - \tau_1 < \bar{\tau})$ are provided to the traveller by the operator when the traveller approaches the HOT facility from the general purpose lane. They are not necessarily related to the actual travel experience of this traveller.

From the utility functions, the following can be derived.

- Choosing *0*(General purpose lane), implies $(U_1 > U_1,)$ and $(U_0 > U_1)$

$$\Rightarrow v > \frac{\hat{\tau}_0 - \hat{\tau}_1}{\pi + f - r \cdot \Pr(\tau_0 - \tau_1 < \bar{\tau})} \quad \text{and} \quad v > \frac{\hat{\tau}_0 - \hat{\tau}_1}{\pi}$$

- Choosing *1*(HOT lane without refund), implies $(U_1 > U_0)$ and $(U_1 > U_1,)$

$$\Rightarrow f - r \cdot \Pr(\tau_0 - \tau_1 < \bar{\tau}) > 0 \quad \text{and} \quad v < \frac{\hat{\tau}_0 - \hat{\tau}_1}{\pi}$$

- Choosing *1'*(HOT lane with refund), implies $(U_1, > U_0)$ and $(U_1, > U_1)$

$$\Rightarrow v < \frac{\hat{\tau}_0 - \hat{\tau}_1}{\pi + f - r \cdot \Pr(\tau_0 - \tau_1 < \bar{\tau})} \quad \text{and} \quad f - r \cdot \Pr(\tau_0 - \tau_1 < \bar{\tau}) < 0$$

To simplify the notation, the following is introduced.

$$A := \frac{\hat{\tau}_0 - \hat{\tau}_1}{\pi}$$

$$B := \frac{\hat{\tau}_0 - \hat{\tau}_1}{\pi + f - r \cdot \Pr(\tau_0 - \tau_1 < \bar{\tau})}$$

It is worth mentioning that the expected net cost of the refund option, $f - r \cdot \Pr(\tau_0 - \tau_1 < \bar{\tau})$, is a critical value in a user's lane choice. In fact, if the expected net cost is positive, all the travellers with $v > A$ will choose *0*; and all the travellers with $v < A$ will choose *1*. On the other hand, if the net cost is negative, all the travelers with $v > B$ will choose *0*; and all the travelers with $v < B$ will choose *1'*.

TRAFFIC MODEL

To account for the intrinsic uncertainty in traffic flow, this study adopts a modified point queue model. Suppose a vehicle enters the facility at a time point where x vehicles are in the downstream vertical queue and y vehicles are on the link but have not joined the queue yet. Then the travel time of this vehicle, denoted as $c(x, y)$, is a function of the free-flow travel time c_0, both x and y, and the discharge process at the downstream bottleneck. The discharge process is stochastic, where the discharge headways are assumed to follow independent and identical normal distributions with a mean of the saturation discharge headway $\bar{h}$ and a variance of σ^2, if the downstream bottleneck is oversaturated. Otherwise, the vehicle's travel time is simply the free-flow travel time if no queue is present when it reaches the downstream end of the facility.

Based on the above discussion, $c(x, y)$ is further approximated by a normal distribution with mean $\hat{c}(x, y)$ and variance $\tilde{c}^2(x, y)$, where

$$\begin{aligned} \hat{c}(x, y) &= c_0 + \max\left\{x + y - \frac{c_0}{\bar{h}}, \quad 0\right\} \cdot \bar{h} \\ \tilde{c}(x, y) &= \max\left\{x + y - \frac{c_0}{\bar{h}}, \quad 0\right\} \cdot \sigma \end{aligned} \tag{1}$$

DYNAMIC PRICING WITH REFUND OPTION

The primary operational objective of HOT lanes is to make full use of its available capacity while maintaining free-flow traffic condition. To this end, a chance constraint is employed to determine the desired inflow to the HOT facility during tolling interval k. Suppose the number of high occupancy vehicles (those qualified for free access) arriving the upstream end of the facility during interval k is θ_1^k, and the number of lower occupancy vehicles is θ_0^k. Inflows to the facility up to interval k are denoted as λ_1^l and λ_0^l for all $l < k$. The vertical queues at the downstream bottleneck of the facility at the beginning of tolling interval k are denoted as q_1^{k-1} and q_0^{k-1} respectively. Further assume t_0 is m (integer) times of a tolling interval Δt. The desired inflows $\tilde{\lambda}_1^k$ and $\tilde{\lambda}_0^k$ can be determined by the following optimization problem.

$$\begin{aligned} & \max \tilde{\lambda}_1^k \\ \text{s.t.} \quad & \Pr\left(c_1\left(q_1^{k-1}, \sum_{i=1}^{m} \lambda_1^{k-i} + \tilde{\lambda}_1^k\right) \le (m+1)\Delta t\right) \ge p \\ & \tilde{\lambda}_1^k + \tilde{\lambda}_0^k = \theta_1^k + \theta_0^k \\ & \tilde{\lambda}_1^k \text{ and } \tilde{\lambda}_0^k \text{ are integers} \end{aligned} \tag{2}$$

Essentially, the optimization model seeks to prompt as many travellers as possible to use the HOT lane, as long as the last user entering the HOT lane during interval k has a minimum chance of p to experience free-flow travel. The solution to the above model can be analytically derived as

$$\tilde{\lambda}_1^k = \left\lfloor \frac{\Delta t}{\sigma \Phi_N^{-1}(p) + \bar{h}_1} - \frac{1}{\bar{h}_1} \cdot \hat{c}\left(q_1^{k-1}, \sum_{i=1}^{m} \lambda_1^{k-i}\right) \right\rfloor \tag{3}$$

where Φ_N^{-1} is the inverse cumulative distribution function of the standard normal random variable, and $\bar{h}_1$ the average saturation discharge headway for the HOT lane.

If $\theta_1^k < \tilde{\lambda}_1^k < \theta_1^k + \theta_0^k$, pricing should be implemented to achieve $\tilde{\lambda}_1^k$. To this end, the operator needs to determine the toll rate π^k, the refund amount r^k, the premium for the refund option f^k, and the guaranteed travel time saving $\bar{\tau}^k$ for tolling interval k. In addition, the operator also needs to provide $\hat{\tau}_0^k$, $\hat{\tau}_1^k$, as well as $\Pr(\tau_0^k - \tau_1^k < \bar{\tau}^k)$, to all the travellers approaching the facility during tolling interval k. Since the focus of this study is on operational parameters, the operator-provided traffic information is set as

$$\begin{aligned} \hat{\tau}_1^k &= \hat{c}_1\left(q_1^{k-1}, \sum_{i=1}^{m} \lambda_1^{k-i} + \tilde{\lambda}_1^k\right) \\ \hat{\tau}_0^k &= \hat{c}_0\left(q_0^{k-1}, \sum_{i=1}^{m} \lambda_0^{k-i} + \tilde{\lambda}_0^k\right) \end{aligned} \tag{4}$$

Note that $\hat{\tau}_1^k$ and $\hat{\tau}_0^k$ represent the predicted expected travel times of the last traveller entering each lane during tolling interval k, if the inflows to the HOT and the general purpose lanes are exactly $\tilde{\lambda}_1^k$ and $\tilde{\lambda}_0^k$.

Without loss of generality, the superscription k will be dropped from now on for simplicity of the notation. Based on the discussion in Lane Choice Model Section, this study will investigate the cases where $f - r \cdot \Pr(\tau_0 - \tau_1 < \bar{\tau}) > 0$ and $f - r \cdot \Pr(\tau_0 - \tau_1 < \bar{\tau}) < 0$ separately.

Paradigm 1: Positive net expected cost of the refund option

In this case, the operator will set the values of f, r, and $\Pr(\tau_0 - \tau_1 < \bar{\tau})$ such that $f - r \cdot \Pr(\tau_0 - \tau_1 < \bar{\tau}) > 0$. The problem is now reduced to determining the value of τ such that $\Phi_v^{-1}(A) = \tilde{\lambda}_1$, where $\Phi_v^{-1}(\cdot)$ is the inverse cumulative distribution of v. Note that since $v = 1/\text{VOT}$, $\Phi_v^{-1}(x) = 1/\Phi_{\text{VOT}}^{-1}(1-x)$, where $\Phi_{\text{VOT}}^{-1}(\cdot)$ is the inverse cumulative distribution of VOT. The analytical solution to this problem is

$$\tau = \frac{\hat{t}_0 - \hat{t}_1}{\Phi_v^{-1}\left(\frac{\tilde{\lambda}_1 - \theta_1}{\theta_0}\right)} \tag{5}$$

Paradigm 2: Negative net expected cost of the refund option

In this case, the operator will set the values of π, f, r, and $\Pr(\tau_0 - \tau_1 < \bar{\tau})$ such that

$$f - r \cdot \Pr(\tau_0 - \tau_1 < \bar{\tau}) < 0 \quad \text{and} \quad \Phi_v^{-1}(B) = \tilde{\lambda}_1.$$

There are four decision parameters and only two equations. Therefore, this system is underdetermined, and multiple solutions exist.

Note that in both paradigms, the determination of f, r, and $\Pr(\tau_0 - \tau_1 < \bar{\tau})$ should have the financial feasibility of the operation as one of the considerations.

SIMULATION AND PRELIMINARY RESULTS

A simulation framework similar to the author's previous research (Lou et al. (2011)) is adopted to investigate the approach under Paradigm 1. The only difference here is that a Monte Carlo approach is implemented to simulate the randomness in lane choice (due to VOT distribution) and traffic propagation (due to random discharge headway).

The framework consists of three major components:

- A controller that implements equations (3) – (5) to calculate the desired inflows to the HOT and the general purpose lanes during tolling interval k ($\tilde{\lambda}_1^k$ and $\tilde{\lambda}_0^k$), the predicted expected travel times of the last traveller entering each lane ($\hat{\tau}_1^k$ and $\hat{\tau}_0^k$), and the optimal toll rate π;
- A lane-choice simulator that generates random numbers according to the assumed VOT distribution to simulate each traveler's lane choice;
- A traffic simulator that generates random travel time for each traveler according to equation (1).

Note that pricing should only be implemented when $\theta_1^k < \tilde{\lambda}_1^k < \theta_1^k + \theta_0^k$. When $\tilde{\lambda}_1^k \leq \theta_1^k$, the demand of HOVs is high enough to warrant an exclusive HOV lane, and no lower occupancy vehicles will be allowed to enter the HOV/HOT lane. When $\tilde{\lambda}_1^k \geq \theta_1^k + \theta_0^k$, the demand is low enough to open the HOT lane for general use. In this case, the toll rate will be 0; and the inflows are assumed proportional to the remaining capacity:

$$\lambda_1^k = \theta_1^k + \theta_0^k \cdot \frac{\mu_1 - \theta_0^k}{\mu_1 + \mu_0 - \theta_0^k} \quad (6)$$

where μ_1 and μ_0 are the saturation flow rates at the downstream bottleneck.

Note that in Paradigm 1, as long as the operator sets the values of f, r, and $\Pr(\tau_0 - \tau_1 < \bar{\tau})$ such that $f - r \cdot \Pr(\tau_0 - \tau_1 < \bar{\tau}) > 0$, the critical VOT value that governs the choice between general purpose and HOT lanes (and thus the traffic condition) only depends on π. Therefore, the experiments are focused on determining the toll rate π only. The financial feasibility of the operation (determination of f, r, and $\Pr(\tau_0 - \tau_1 < \bar{\tau})$) is not considered in the experiments.

Simulation Settings

Figure 4-1 illustrates the simulated HOT facility. It has one HOV/HOT lane and one general purpose lane. In order to create congested traffic condition, it is assumed that a downstream bottleneck is in effect. The free flow travel time for the simulated freeway segment is set as four times the tolling interval ($m = 4$), i.e., 8 minutes. If the free flow speed is 60 mph, the facility is 8 miles long. The average saturation flow rates at the downstream bottleneck (μ_1 and μ_0) are set to 1800 vph and 2400 vph for the HOT and the general purpose lanes respectively. This is equivalent to setting $\bar{h}_1 = 2$ seconds and $\bar{h}_0 = 1.5$ seconds. The standard deviation of the saturation headway is set as 10% of the mean, for both HOT and general purpose lanes. The upstream saturation flows are set to 1800 *vph* and 3600 *vph* respectively.

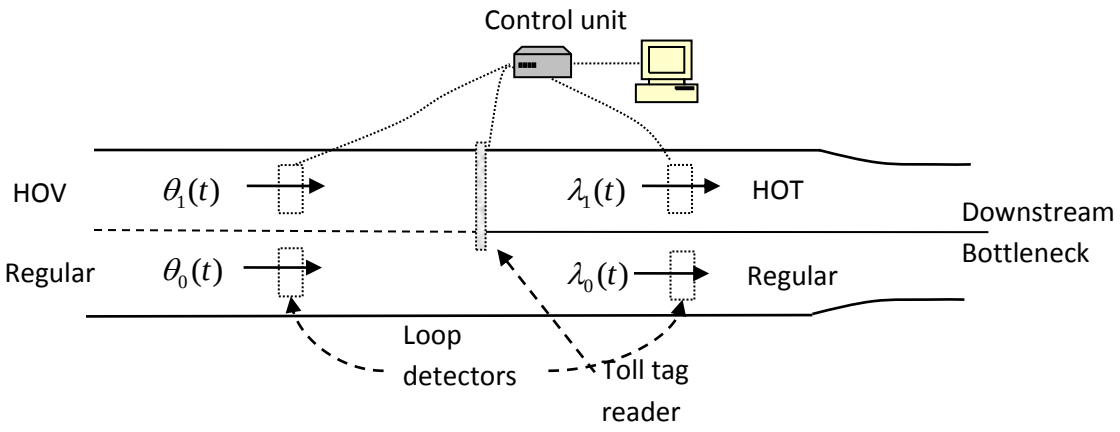


Figure 4-1. Simulated HOT Facility.

The dynamic tolling interval Δt is set to 2 minutes. A total of 44 time intervals (88 minutes) is simulated. The upstream inflow rate for the HOT lane θ_1^k is set to exactly 10 vehicles per time interval (equivalent to 300 *vph*) throughout the simulation duration. For the general purpose lane, the upstream inflow rate θ_0^k is set to exactly 120 vehicles per time interval (equivalent to 3600 *vph*) for the first 24 time intervals (48 minutes), and 60 vehicles per interval (equivalent to1800 *vph*) for the last 20 time intervals (40 minutes). The first 4 time intervals (8 minutes) of the entire simulation duration are warm-up periods, where the HOT lane is not activated. Therefore, the inflows λ_1^k and λ_0^k are 10 vehicles and 120 vehicles for every interval during the initial 4 intervals.

Similar to Gardner et al. (2013), a Burr function is adopted for VOT distribution with two parameters ξ and γ.

$$\Phi_{\mathrm{VOT}}(x) = \Pr(\mathrm{VOT} \le x) = 1 - \frac{1}{1 + \left(\frac{x}{\xi}\right)^{\gamma}}$$

Therefore,

$$\Phi_v^{-1}(x) = \frac{1}{\Phi_{\mathrm{VOT}}^{-1}(1-x)} = \frac{1}{\xi} \cdot \left(\frac{1-x}{x}\right)^{\gamma}$$

ξ represents the median VOT, and γ is a shape parameter. In this simulation, ξ is set to 15 ($/hour), and γ is set to 2.

Results

Two experiments are performed for Paradigm 1 with different target p values (the probability for the last user entering the HOT lane during a tolling interval to experience free-flow travel).

For each experiment, ten simulation replications are performed. The inflows, queue lengths, and toll rates are recorded for both lanes at each time interval. The results are presented below.

Experiment 1: p =0.85.

The performance of the facility from the first simulation replication of this experiment (Exp. 1, Run 1) is shown in Figure 4-2. The corresponding toll rate is shown in Figure 4-3. During the warm-up period (first 8 minutes), the HOV/HOT lane operates as an HOV-only lane. The toll rates are zeros during these time period. The inflows to HOT and general purpose lanes (λ_1^k and λ_0^k) are constant and equal to the upstream inflows (θ_1^k and θ_0^k). Since it takes 4 time intervals (8 minutes) in free flow for a vehicle to arrive at the downstream bottleneck, the queue starts to build up starting at the 8th minute for the general purpose lane. At the same time, the HOV/HOT lane starts to open to lower occupancy vehicles with a toll. The toll rates (Figure 4-3) vary between \$1 and \$1.5 from time interval 5 to 24 (minute 8 to 48) when the demand is higher. After the upstream arrival rate drops at minute 48, the facility keeps operating as an HOT lane for another two time intervals with significantly lower rates (Figure 4-3). When the toll rate completely drops to zero after minute 52 and both lanes are open to all traffic, the inflows to both lanes become stable, proportional to available capacities (see equation (6)).

Note that the downstream discharge headway is considered a random variable in order to model the uncertainty in travel time. Although the downstream saturation flow rate is 60 vehicles per time interval, the desired HOT inflow will be lower than 60 to satisfy the chance constraint (2). For this simulation replication, the desired inflow to HOT lane is calculated as 54 vehicles per interval from equation (3); and it can be seen from Figure 4-2 that the actual inflow to HOT lane varies around the desired value due to random VOT of the approaching vehicles.

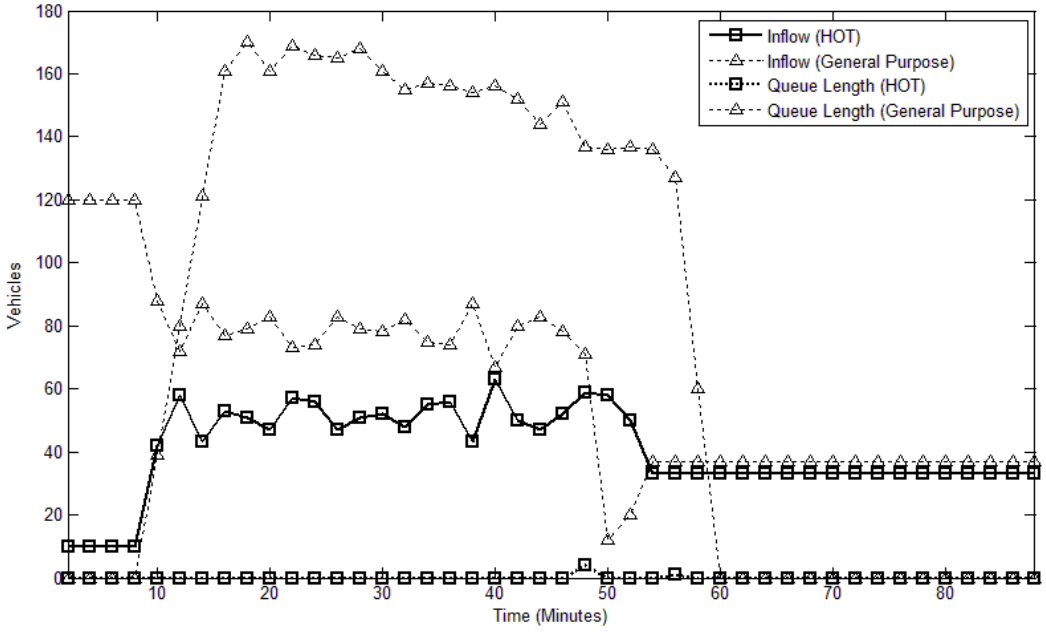


Figure 4-2. Facility Performance (Exp. 1, Run 1).

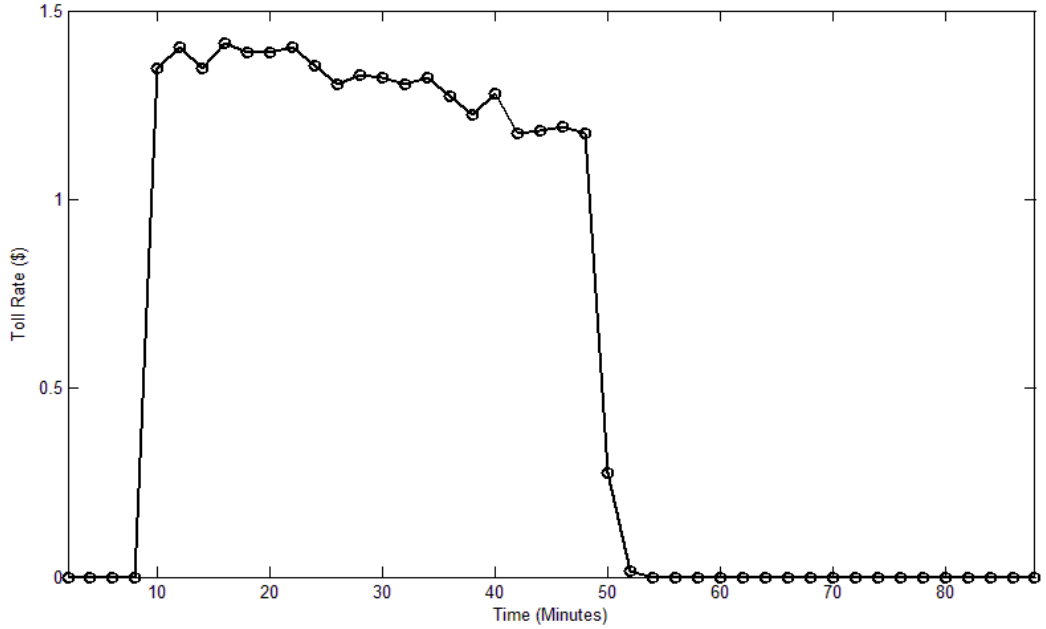


Figure 4-3. Toll Rates (Exp. 1, Run 1).

The inflow to HOT lane during interval 5 to interval 26 (minute 8 to minute 52) when the facility is active ranges from 43 to 63 vehicles per time interval, with an average value of 51.73. This value is lower than the calculated desired value of 54 vehicles per interval. Therefore, it is no surprise that there is no queue on the HOT lane for the most part of the simulation duration (see Figure 4-2). In fact, there are only two time intervals where queue is present on the HOT lane. At the end of interval 24 (minute 48), the HOT lane has a 4-vehicle queue. This is corresponds well with the highest inflow observed for the HOT lane during interval 20 (minute 38 to 40, see Figure 4-2). At the end of interval 28 (minute 56), the HOT lane has a 1-vehicle queue. This is because that the HOT lane is still active with a relatively high inflow during intervals 25 and 26 (minute 48 to 52), and it is still possible to form queues due to the random discharge headway at the downstream bottleneck.

On the other hand, although the inflows to general purpose lane are significantly reduced when the HOT lane is active, the general purpose lane is still severely congested because the total demand is much higher than the total capacity at the downstream bottleneck. When the total demand drops at the beginning of interval 25 (minute 49), the HOT lane is still in effect but with significantly lower toll rates, prompting a large percentage of travellers to use HOT lane. This leads to a substantial drop of inflow to the general purpose lane during interval 25 and 26 (minute 48 to 52). It can be seen from Figure 4-2 that the inflow to the general purpose lane during these 4 minutes is significantly lower than that to the HOT lane. This has prevented the existing queue on the general purpose lane from growing and has helped it to dissipate. No new queue is formed after the existing queue is discharged, since the new demand and inflow is much less than the downstream capacity.

Two performance measures are of interest: the maximum queue length and the percentage time when queue is present for the HOT lane. The latter is related to the chance constraint (2), and represents the probability for the last user entering the HOT lane during a tolling interval to encounter a queue. Suppose the HOT lane is active starting at interval k_1 and ending at interval k_2. Because the free-flow travel time is m time intervals, this metric should be calculated as the number of intervals when queue is present between intervals $k_1 + m$ and $k_2 + m$ divided by $(k_2 - k_1 + 1)$. Any HOT queue present after interval $k_2 + m$ is not related to the p value in chance constraint (2). It does indicate, however, a failure in HOT operation, as too many vehicles have opted in the facility and have created additional congestion. For Exp. 1, Run 1, the HOT lane is active between intervals 5 and 26 (minute 8 to 52 for a total of 44 minutes), and no queue is present after interval 30. The percentage time when queue is present for the HOT lane is $2/22 = 9.09\%$, which is lower than the target value of $1 - p = 1 - 0.85 = 15\%$. The maximum HOT queue length is 4.

Results from another simulation replication (Exp. 1, Run 2) are presented in Figure 4-4 and Figure 4-5. It can be observed that the general trends of the inflows, the queue lengths, and the toll rates are similar to the results of Exp. 1, Run 1. The toll rates (Figure 4-5) are a little

lower comparing to the previous replication, varying between \$0.7 and \$1.35 when the demand is higher. The facility keeps operating as an HOT lane for only one additional time interval with a toll rate of \$0.08 after the upstream arrival rate drops at minute 48. The inflow to HOT lane ranges from 41 to 66 vehicles per time interval with an average value of 54.14, when HOT facility is active. The average is very close to the target inflow of 54 vehicles per time interval. However, due to a series of relatively high HOT inflows during intervals 21 to 25 (minute 40 to 50), queue is present on the HOT lane during intervals 25 to 27 (minute 48 to 54). A fourth interval when queue is present on the HOT lane is interval 17. No queue is present after interval $25 + 4 = 29$ (minute 58). The percent time when queue is present for or the HOT lane is $4/21 = 19.05\%$, higher than the target value of 15%. The maximum HOT queue length is 9.

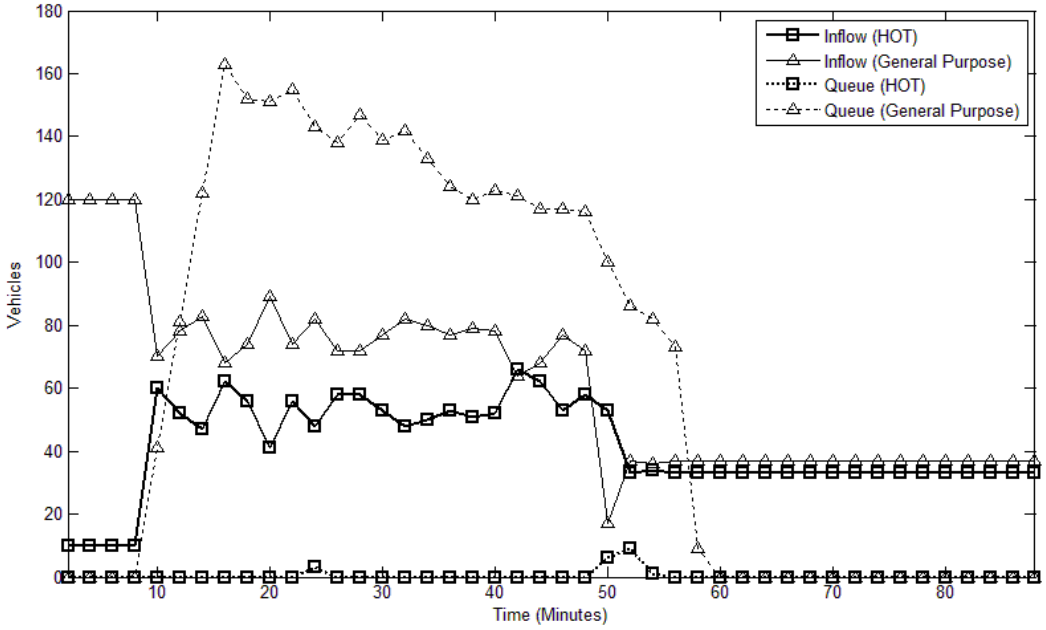


Figure 4-4. Facility Performance (Exp. 1, Run 2).

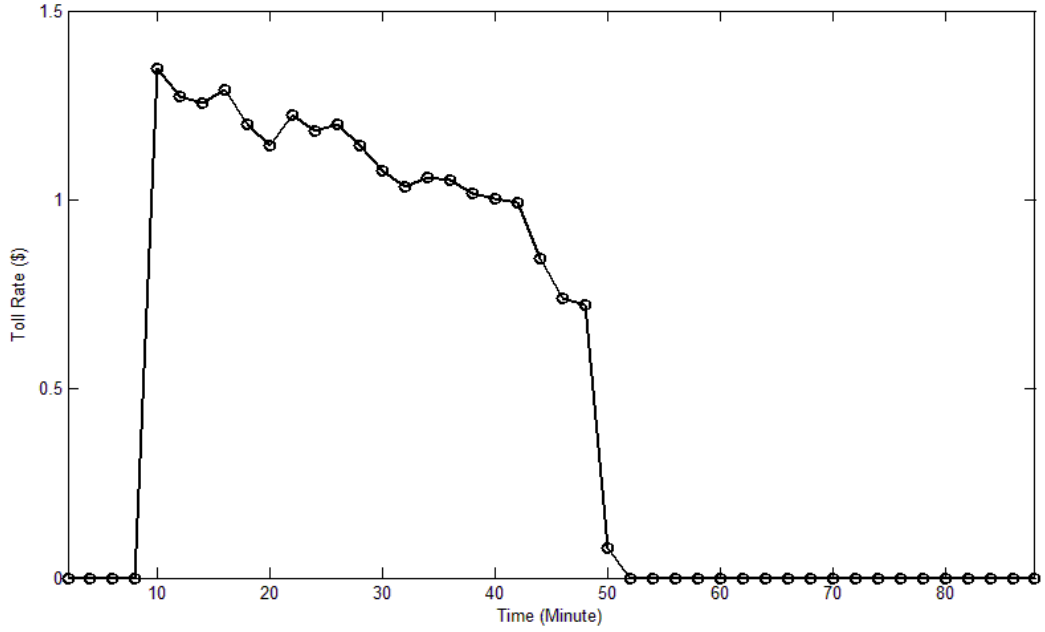


Figure 4-5. Toll Rates (Exp. 1, Run 2).

Across the 10 simulation replications, the maximum HOT queue length varies from 1 to 9 vehicles, with an average value of 4.30. No queue is present beyond 4 time intervals after the last tolling interval. The average percentage time when queue is present for the HOT lane is 16.43%, slightly higher than the operational goal of 15%. The toll rate ranges from $0.01 to $1.42.

Experiment 2: p =0.95.

In this experiment, the p value in the chance constraint (2) is increased to 0.95—a higher safety margin in order to deliver the superior travel condition for the HOT lane.

Figure 4-6 and Figure 4-7 present the results from one of the ten simulation replications (Exp. 2, Run 3). It can be observed that the general trends of the inflows, the queue lengths, and the toll rates are similar to the results of Experiment 1. The toll rates (Figure 4-6) range from $0.97 to $1.50 when the demand is higher. The facility keeps operating as an HOT lane for only one additional time interval with a toll rate of $0.24 after the upstream arrival rate drops at minute 48. The inflow to HOT lane ranges from 45 to 62 vehicles per time interval with an average value of 52.57, when HOT facility is active. The average is slightly higher than the target inflow of 51 vehicles per time interval. Queue is present on the HOT lane during three time intervals: 13, 16, and 17 (minute 24 to 26 and 30 to 34). No queue is present after interval $25 + 4 = 29$ (minute 58). The percent time when queue is present for or the HOT lane is $3/21 = 14.29\%$, much higher than the target value of 5%. The maximum HOT queue length is 5.

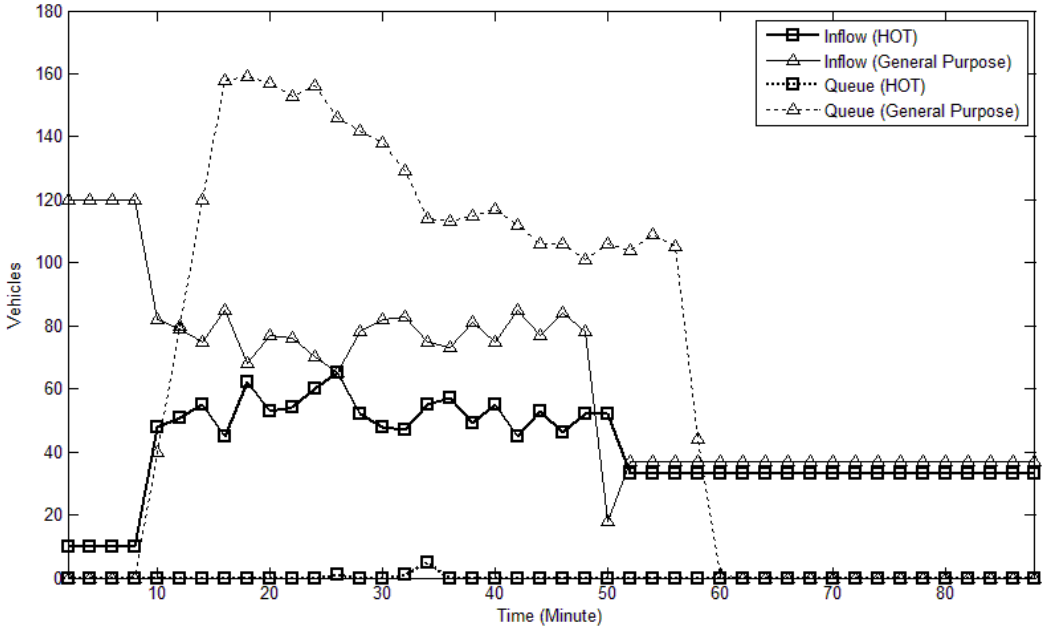


Figure 4-6. Facility Performance (Exp. 2, Run 3).

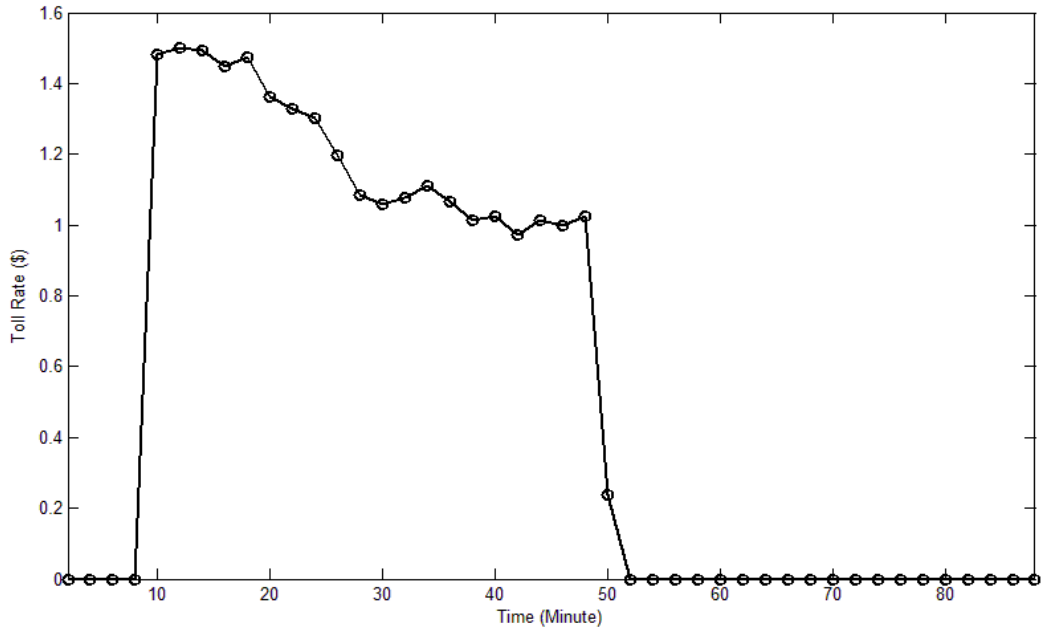


Figure 4-7. Facility Performance (Exp. 2, Run 3).

Across the ten replications, the maximum queue length for the HOT lane varies from 0 to 5 vehicles, with an average value of 2.60. No queue is present beyond 4 time intervals after the last tolling interval. The average percentage time when queue is present for the HOT lane is 6.43%, slightly higher than the operational goal of 5%. The toll rate ranges from $0.06 to $1.71.

CHAPTER 5 A TRADABLE CREDIT SCHEME FOR STAGGERED WORK TIME

INTRODUCTION

Traffic congestion is a direct result of spatial and temporal concentration of travel demand. Historically, capacity expansion has been one of the primary solutions to the congestion problem. This approach turns out to be not viable, as it requires significant financial resource and right-of-way. More importantly, the added capacity would soon be consumed by induced travel demand (Duranton and Turner (2012)).

In contrast, travel demand management (TDM) strategies directly target travel demand and aim to spread it across space and time. Many TDM strategies have been proposed and some of them have been implemented in practice, e.g., congestion pricing (De Palma and Lindsey (2011), Tsekeris and Voss (2009)) and traffic rationing (Wang et al. (2010), Han et al. (2010)).

Alternative work schedules (Arnott et al. (2005), Mun and Yunekawa (2006)) are another TDM strategy that has been implemented to reduce traffic congestion. There are three types of alternative work schedules, i.e., staggered work hours where the firms assign different groups of travelers different work start times; flextime in which employees can adjust their arrival times, but need to be at workplace for some core times and fulfill the required working hours, and compressed work weeks in which employees work more hours per working days in order to compensate working for fewer days (Transportation Research Board (1980)).

Arnott et al. (2005) reported several real-world implementations of alternative work schedules in Manhattan in 1970, Toronto in 1970, Washington D.C. in 1970, Tel-Aviv in 1980, and Kuala Lumpor in 1998. He also remarked the voluntary staggered work plans of BMW and Siemens in Germany as instances of involvement of firms in traffic congestion mitigation. In a four-week pilot program in Honolulu, Hawaii, the work hours of state, city, and county employees were changed from 7:45 AM–4:30 PM to 8:15 AM–5:15 PM. The subjected employees were 20% of 60,000 employees in downtown Honolulu. Although the program yielded reduction in average travel time, many employees did not like the mandatory shifting (Giuliano and Golob (1989)). Indeed, despite its considerable potential in mitigating traffic congestion, staggered work time has not been successful in practice due to the opposition of both employees and firms (Yoshimura and Okumura (2001)).

The observation that firms do not tend to participate in staggered work plans can be explained by the theory of economic agglomeration, which suggests that the productivity of employees increases as the number of employees who work simultaneously increases. Staggering employees reduces the overlap of working hours and thus leads to some productivity loss. This explains why firms often do not voluntarily enroll in a staggered work schedule. However, firms'

production technology may vary substantially. At some firms, employees complement each other and thus its productivity will be largely compromised if employees do not work at the same time. At others, employees work more independently and their productivity is not affected too much if a portion of employees start to work after the morning peak period. Such heterogeneity can be utilized to design policies to encourage firms to stagger their employees to reduce traffic congestion. This chapter is one of such attempts.

More specifically, this chapter proposes a market-based mechanism to mitigate the negative consequence that firms may experience in staggering the work start time of their employees. In our proposed scheme, mobility credits are first allocated by a government agency to all firms in a central business district (CBD), and firms are responsible for redistributing the credits to their employees. In addition, the credits can be traded freely among firms. During the morning peak, each traveler who enters the CBD will be charged one credit. Because the number of allocated credits to each firm is not enough to fulfill all the travel needs of their employees, each firm may eventually have two groups of employees: employees with credits who can arrive during the charging interval, i.e. the morning peak, and employees without credits who will shift to another work start time.

The proposed scheme belongs to a category of tradable mobility credit schemes that recently have received considerable attention. The potential of tradable permits in regulating traffic congestion externality was first noted by Verhoff et al. (1997) and Viegas (2001). Recently, Yang and Wang (2011) proposed a mathematical framework for analyzing tradable mobility credits. Their work has been extended to capture the heterogeneity of travelers (Wang et al. (2012), Zhu et al. (2014)), transaction cost (Nie (2012)) and income effect (Wu et al. (2012)). Shirmohammadi et al. (2013) showed formally that there is one-to-one correspondence between tradable credits and congestion pricing in idealized situations with perfect certainty. They further investigated a safety valve policy to balance regulation success and the volatility of credit price under demand and/or supply uncertainty. More recently, He et al. (2013) studied a tradable scheme when a finite number of Cournot-Nash (CN) players and an infinite number of Wardrop-equilibrium (WE) players compete in the network simultaneously. They analyzed how transaction costs would affect the trading and route-choice behaviors of both CN and WE players. For a more comprehensive review, see, e.g., Fan and Jiang (2013).

Our proposed scheme differs from those discussed above in that credits are allocated to eligible firms rather than individual travelers and trading in credits market are only allowed between firms instead of travelers. As such, the government agency needs only deal with a limited number of players and subsequently the credit market is relatively smaller and easier to establish and monitor.

DESCRIPTION OF THE PROPOSED SCHEME

We describe the proposed scheme in a simplified morning commute setting where all travelers who are traveling to the CBD every morning are employees of the firms in the CBD. The current common work start time is $t^*$and employees should be at workplace before t^*, i.e. no late arrival is allowed. To reduce congestion, a government agency now decides the number of travelers who can enter the CBD during the morning peak, denoted by K, and issue K mobility credits. The mobility credits are allocated to firms, and firms are responsible for assigning them to their employees. During a charging period that starts from t^+ and ends on t^*, every traveler who wishes to enter the CBD will be charged one mobility credit. Eventually, firm i with N_i employees has two groups of employees: $\widehat{N}_i$ employees with mobility credits, and $\widetilde{N}_i$ employees without credits. Clearly, only employees with credits can arrive during the charging interval. For employees without credits, firms can either change their work start time or require them to be at work place before t^+. The latter option is less practical or desirable because it creates much dissatisfaction among employees and causes a loss in productivity too. So, staggering work start time becomes a plausible option available to the firms. They can shift the work start time of employees without credits to $\bar{t}^*$, where $\bar{t}^* > t^*$. For simplicity, we further assume that $\bar{t}^*$ is the same for all firms and the duration between t^* and $\bar{t}^*$ is long enough that all employees without credits would be able to enter the CBD before $\bar{t}^*$. More specifically, we assume $\bar{t}^* - t^* = \frac{\sum_l \widetilde{N}_l}{s}$, where s is the capacity of the highway leading to the CBD.

In the proposed scheme, mobility credits are not distributed directly among travelers and travelers are not allowed to trade their credits. Instead, mobility credits are initially endowed to the firms. Firms can distribute them among their employees or trade them with other firms. Conceptually, firms with more complementary technology value more of having more employees at t^* and will be likely a buyer in the credit market. On the other hand, firms with "independent" technology are less affected by staggering and would find themselves as a seller in the credit market.

Allocation of credits to firms rather travelers reduces the number of players in the market substantially. Hence, the market may be more tractable, and transaction costs associated with searching and negotiating with trading partners, monitoring the market and enforcement can be reduced as well.

MODELING FRAMEWORK

This section introduces the modeling framework for analyzing the proposed scheme. Given the total number of credits issued by the government agency, K, we attempt to model firms' decisions in the credit market and derive the equilibrium departure and arrival patterns of employees. Subsection "Impacts on travelers" mathematically describes how the decisions of departure time of employees shape traffic demand pattern and presents the equilibrium cost of each group of employees based on the bottleneck model initialized by Vickrey (1969). Subsection "Impacts on firms" is devoted to the modeling of firms' behaviors in the credit market. Subsection "Optimal design" formulates an optimal credit design problem based on the combination the equilibrium outcomes of firms and employees.

Impacts on Travelers

Morning commute problem

The morning commute problem was first introduced by Vickrey (1969) to describe the temporal distribution of morning commutes. Vickrey (1969) argued that, in addition to travel time, the deviation of actual arrival time from desired arrival time is an important factor in traveler's departure time decision. In fact, travelers who want to avoid traffic delay depart relatively early or late. On the other hand, travelers who arrive closer to the desired arrival time incur more travel delay. Smith (1984) and Daganzo (1985) proved, respectively, the existence and uniqueness of the equilibrium arrival pattern at a single bottleneck. For recent comprehensive reviews on the morning commute problem, see, e.g., Arnott et al. (1998) and de Palma and Fosgerau (2011).

Bottleneck model has been applied to investigate different tradable mobility credit schemes. Xiao et al. (2013) studied the efficiency of a tradable credits scheme with time-varying charging rate, and showed a system optimum charging scheme can be designed for heterogeneous travelers if the distribution of value of travel time is known. Nie (2012) proposed a tradable credit scheme that charges uniformly the travelers passing the bottleneck inside a peak time window, and rewards mobility credits to travelers who travel outside of the peak time window. Nie and Yin (2013) further considered rewarding travelers who divert to alternative route or mode, in addition to those who travel during off-peak times.

In this chapter, we utilize the bottleneck model to derive the equilibrium travel cost of employees. To do so, the bottleneck model is briefly reviewed and then twisted to capture the specifications of our proposed scheme. Assume N homogenous individuals are commuting every morning from their origin, i.e. home, to their workplaces located in the CBD. The capacity of the highway connecting the origin to the destination is s. Obviously, when $N \geq s$, not all travelers can arrive at workplace on time. Normalizing the free-flow travel time to zero and assuming no

late arrival, the travel cost of a traveler who departs from home at time t can be expressed as follows:

$$c(t) = \alpha T(t) + \beta \max\{0, t^* - t - T(t)\} \quad (1)$$

where $T(t)$ is the queuing delay; α is the value of travel time and β is the unit cost of arriving early to the destination, and $\alpha > \beta$. Therefore, each individual incurs a cost associated with being at the queue, referred as the delay cost, and a cost associated with not arriving on time, referred as the schedule cost. At equilibrium, the travel cost of all commuters would be the same and no commuter can reduce his or her travel cost by changing his or her departure time unilaterally. $T(t)$ can be estimated by dividing the length of queue at the bottleneck at time t, i.e., $q(t)$, by the capacity of the bottleneck, i.e., s. At equilibrium, the last traveler must arrive at t^*; otherwise he or she can save by arriving at t^*. The first departing traveler at t_s will experience no queuing delay; otherwise he or she can save by departing earlier. With a similar logic, it can be inferred that the bottleneck should be fully utilized during the departing period $[t_s, t^*]$ and the departure rate from the bottleneck would be s. Thus $t_s = t^* - \frac{N}{s}$. Consequently, $q(t)$, which is the difference between the cumulative departure from home and the cumulative departure from the bottleneck can be written as:

$$q(t) = \int_{t_s}^{t} r(t)dt - s \cdot (t - t_s), t \in [t_s, t^*] \quad (2)$$

where $r(t)$ is the departure rate at time t. From the equilibrium definition of $\frac{\partial c}{\partial t} = 0$, (1) and (2) yield:

$$r(t) = \frac{\alpha}{\alpha - \beta} s, \qquad t_s \le t \le t^*$$

It is straightforward to obtain the equilibrium travel cost as $\beta \frac{N}{s}$ and $q(t^*) = \frac{\beta}{\alpha} N$. The departure pattern before the implementation of the proposed scheme is depicted in Figure 5-1. In this situation, total travel cost and total travel delay are $\beta \frac{N^2}{s}$ and $0.5\beta \frac{N^2}{s}$ respectively.

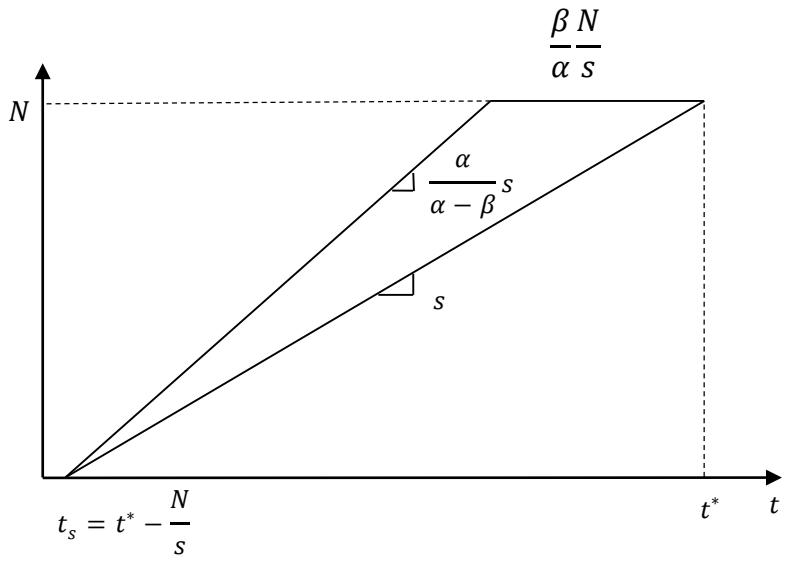


Figure 5-1. Departure pattern of employees before the implementation.

Temporal Distribution Under The Proposed Scheme

Recall that under the proposed scheme, firms will shift the work start time of employees without credits to $\bar{t}^*$, where $\bar{t}^* > t^*$. As no late arrival is allowed, t^* and $\bar{t}^*$ are set in a way that all employees without credits can arrive before $\bar{t}^*$. More specifically, it is assumed that $\bar{t}^* - t^* = \frac{\sum_l \bar{N}_l}{s}$. This assumption completely separates the departures of the two groups of employees so that the temporal distribution of each group can be analyzed separately.

For the group with credits, the first employee departs from home at $\hat{t}_s = t^* - \frac{\sum_i \hat{N}_i}{s}$, and the last employee departs at $\hat{t}_e = t^* - \frac{\beta}{\alpha}\frac{\sum_i \hat{N}_i}{s}$. The equilibrium travel cost is $\beta\frac{\sum_i \hat{N}_i}{s}$ and the total travel cost and travel delays are $\beta\frac{(\sum_i \hat{N}_i)^2}{s}$ and $0.5\beta\frac{(\sum_i \hat{N}_i)^2}{s}$, respectively. Similarly, for the group without credits, the first employee departs from home at $\bar{t}_s = \bar{t}^* - \frac{\sum_i \bar{N}_i}{s}$, and the last employee departs at $\bar{t}_e = t^* - \frac{\beta}{\alpha}\frac{\sum_i \bar{N}_i}{s}$. The equilibrium travel cost is $\beta\frac{\sum_i \bar{N}_i}{s}$ and the total travel cost and travel delays are $\beta\frac{(\sum_i \bar{N}_i)^2}{s}$ and $0.5\beta\frac{(\sum_i \bar{N}_i)^2}{s}$, respectively. The equilibrium departure pattern associated with the proposed scheme is depicted in Figure 5-2.

Given the total number of credits issued, we can compare total travel cost, total travel delay, and total schedule cost before and after the implementation of the proposed scheme. The comparision suggests that all these measures are reduced by $2\left(\frac{K}{N}\right)\left(1-\frac{K}{N}\right)$. The maximum possible reduction is 50%, which can be achieved by issuing credits of the half of the total number of employees.

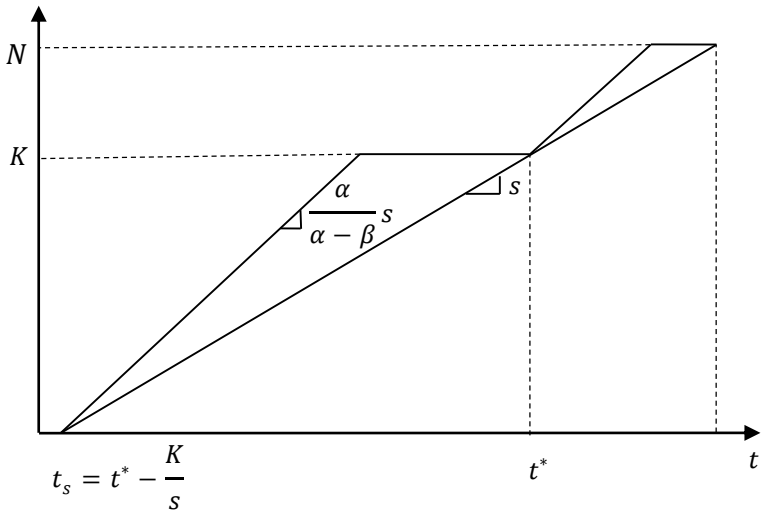


Figure 5-2. Departure pattern of employees under the proposed scheme.

Impacts On Firms

Productivity Effect Of Work Start Time

It appears that no empirical study has investigated the relationship between firm productivity and the work start time of its employees. Some researchers have suggested an indirect way to study the relationship (Yushimito et al. (2013)). Since a profit-maximizing firm determines its employees' wages to reflect their marginal productivity, investigating the variation of wages versus work start time may lead us to understand how productivity varies with the work start time. It should be emphasized that firm productivity can be affected by many factors, and establishing a function that addresses all these factors is challenging, if not impossible.

To our best knowledge, there are only two empirical studies that explored the relationship between wage and work start time. Wilson (1988) observed a strongly U-inverse relationship between them with the average wage of peak worker being twice of off-peak ones. However, Arnott et al.(2005) argued that such a considerable difference in wage between peak and off-peak workers in Wilson's observation cannot "be explained by intraday productivity effect alone". He believed that such large differences stem from employees' abilities that are noticed by firms but "not observable to the empirical researcher".

In contrast to Wilson (1988), Gutiérrez-i-Puigarnau and Ommeren (2012) used panel data to have more control on time-invariant characteristic of firms and workers, and found a slight inverse U-shaped relation between wage and work start time. Yushimito et al. (2013) explained the difference between Wilson (1988) and Gutiérrez-i-Puigarnau and Ommeren (2012) by referring to the time of these studies. With advancements in telecommunication technology, firms and employees can be in closer contact outside of workplace than they could at the time of

the former study. Hence, the firms' productivity loss would be less, and wage is less sensitive to the work start time. Consistent with these earlier studies, Yushimito et al. (2013) defined the productivity obtained from a worker who arrive at time k as $\exp(\beta_0 + \beta_1 k + \beta_2 k^2)$, where, β_0, β_1, and β_2 are the parameters.

In this chapter, we assume that the productivity of firm *i* contributed by $N_{i,j}$ employees with work start time t_j^* follows the following form:

$$\rho_{i,j}(N_{i,j}) = a_i e^{-\theta_i (t_j^* - t^*)^2} N_{i,j} \quad (3)$$

where a_i is the productivity of employees whose work start time is t^*, and θ_i is a parameter reflecting the sensitivity of firm's productivity to employees' work start time. A higher value of θ_i implies that firms are more affected by the deviation of work start time from t^*, while a lower value of θ_i is for less-sensitive firms. In other words, higher θ_i means firm's technology is complementary while lower θ_i suggests that employees are less dependent on each other. The term of $e^{-\theta_i (t_j^* - t^*)^2}$ describes the relative productivity of an employee in firm *i* with work start time t_j^* as compared to his or her co-worker who is assigned to initial work start time t^*.

It should be emphasized that Yushimito's productivity function is based on the arrival time to the workplace. In contrast, in Eq. (3) the productivity only depends on the work start time assigned to employees, not the actual arrival time to the workplace. Total productivity of firm *i*, denoted as ρ_i, is the sum of productivity resulted from all groups of employees, i.e.,

$$\rho_i = \sum_j a_i e^{-\theta_i (t_j^* - t^*)^2} N_{i,j} \quad (4)$$

According to (4), the productivity of firm i is maximized if all of its employees start to work on t^*, which correspond to the situation without implementing the proposed scheme.

Firms' Behavior In The Credit Market

Under the proposed scheme, each firm should determine the number of credits needed. Firms are assumed to be profit maximizers and, therefore, each of them chooses the number of credits to maximize its own profit. Here, the profit of firm i, π_i, is the total productivity minus the expense of purchasing credits from the market, i.e.,

$$\pi_i = a_i \widehat{N}_i + a_i e^{-\theta_i \left(\frac{\sum_l \tilde{N}_l}{s}\right)^2} \tilde{N}_i - p\left(\widehat{N}_i - k_i^0\right) \quad (5)$$

Where p is the market price of credits and k_i^0 is the number of credits initially allocated to firm i. The first term in (5) is the total productivity resulted from employees with credits, who start to work at the primary work start time t^*, and therefore have no loss in their productivity. The second term in (5) is the total productivity resulted from employees without credits, who are shifted to the secondary work start time, $\bar{t}^*$, and the productivity resulted from each of them is

$a_i . e^{-\theta_i(\bar{t}^* - t^*)^2}$ which is equal to $a_i . e^{-\theta_i\left(\frac{\sum_l \tilde{N}_l}{s}\right)^2}$. Finally, the last term in (5) is the expense of purchasing extra mobility credits from the market. Note that when $\hat{N}_i < k_i^0$ the firm is a seller of credits and receive additional profit from selling extra credits; if $\hat{N}_i > k_i^0$, the firm is a buyer of credits and will pay to purchase the extra mobility credits needed.

The decision that firm i faces can be expressed as a mathematical model as follows,

$$\max_{\hat{N}_i, \tilde{N}_i} \pi_i \equiv a_i \hat{N}_i + a_i e^{-\theta_i\left(\frac{\sum_l \tilde{N}_l}{s}\right)^2} \tilde{N}_i - p\left(\hat{N}_i - k_i^0\right)$$

s.t.

$$\hat{N}_i \geq 0 \tag{6}$$

$$\tilde{N}_i \geq 0 \tag{7}$$

$$\hat{N}_i + \tilde{N}_i = N_i \tag{8}$$

The decision variables of each firm, i.e. $\hat{N}_i$and $\tilde{N}_i$, not only have effect on the firm's profit, but also affect the profit of other firms. In addition, the market clearing condition can be written as,

$$0 \leq p \perp \sum_i k_i^0 - \sum_i \hat{N}_i \geq 0 \tag{9}$$

Equation (9) states that if the market is not cleared, i.e. there are more credits than needed, the price of credits would be zero.

The first-order optimality conditions for the above problem can be written as follows,

$$-a_i + p - \hat{\mu}_i + \xi_i = 0 \tag{10}$$

$$-a_i e^{-\theta_i\left(\frac{\sum_l \bar{N}_l}{s}\right)^2} \left[1 - \bar{N}_i . \left(\frac{2\theta_i}{s}\right) . \left(\frac{\sum_l \tilde{N}_l}{s}\right)\right] - \bar{\mu}_i + \xi_i = 0 \tag{11}$$

$$0 \leq \hat{\mu}_i \perp \hat{N}_i \geq 0 \tag{12}$$

$$0 \leq \bar{\mu}_i \perp \tilde{N}_i \geq 0 \tag{13}$$

$$\hat{N}_i + \tilde{N}_i = N_i \tag{14}$$

where $\hat{\mu}_i$, $\tilde{\mu}_i$, and ξ_i are the lagrangian multipliers associated with constraints (6), (7), and (8), respectively. At equilibrium, conditions (10)-(14) should be satisfied for each firm.

Assume that the credit market is cleared, and all firms staggered a positive number of employees, i.e., both $\hat{N}_i$and $\tilde{N}_i$ are strictly positive. From (12) and (13) we have $\hat{\mu}_i = \tilde{\mu}_i = 0$. In addition, by assuming a strictly positive of credit price, we have $\sum_i \hat{N}_i = \sum_i k_i^0 \equiv K$, and

$$\sum_i \widetilde{N}_i = \sum_i N_i - K \tag{15}$$

From (10) $\xi_i = a_i - p$, and plugging this into (14) yields:

$$-a_i\, e^{-\theta_i\left(\frac{\sum_i N_i - K}{s}\right)^2}\left[1 - \widetilde{N}_i\left(\frac{2\theta_i}{s}\right)\left(\frac{\sum_l \bar{N}_l^a}{s}\right)\right] + a_i - p = 0 \tag{16}$$

After some algebraic manipulation, we reach the following:

$$p = \frac{\sum_i \frac{1}{\theta_i} e^{\theta_i A^2} - \sum_i \frac{1}{\theta_i} + 2A^2}{\sum_i \frac{1}{\theta_i a_i} e^{\theta_i A^2}} \tag{17}$$

$$\widetilde{N}_i = \frac{s}{2\theta_i A}\left[1 + \frac{p - a_i}{a_i} e^{\theta_i A^2}\right] \tag{18}$$

where $A = \frac{\sum_i N_i - K}{s}$. From (17), it can be found that the credit price is independent of initial allocation of credits, which is consistent with [14]. In addition, $\widetilde{N}_i$ depends on the total number of employees, total issued credits, and productivity parameters of firm i, i.e. θ_i and a_i. That suggests that for a given total number of employees, changing the number of employees of a firm would not change its number of staggered employees. Such a property largely stems from the specification of the productivity function in which the productivity of each employee solely depends on his or her work start time, and would not be affected by the portion of employees who are shifted. In fact, given the total numbers of employees and issued credits, the total number of shifted employees will be determined and firms based on their initial productivity and their sensitivity find their shares of shifted employees. Note that in deriving Eqs. (17) and (18), it is assumed that corner solutions do not exist, i.e. $\widetilde{N}_i$ and $\widehat{N}_i$ are strictly positives.

Optimal Design

As we discussed above, firms suffer by staggering their employees due to the projected loss in their productivity. The proposed scheme is able to compensate the firms to some extent. On the other hand, travelers are better off, because their travel cost can be reduced by implementing the staggered work time. Therefore, one portion of players, i.e., firms, is worse off while the other portion, i.e., travelers, is better off. The goal of this section is to find an optimum number of credits to be issued by the government agency to maximize the social benefit, which is the total productivity of firms minus the total travel cost of employees, i.e.,

$$SB = \sum_i a_i\left(N_i - \widetilde{N}_i\right) + \sum_i a_i e^{\theta_i\left(\frac{\sum_i N_i - K}{s}\right)^2}\widetilde{N}_i - \beta\frac{K^2}{s} - \beta\frac{(\sum_i N_i - K)^2}{s} \tag{19}$$

where $\widetilde{N}_i$ can be obtained from the previous Section. It is not difficult to verify from (19) that the social benefit only depends on K. Thus the maximization of social benefit, i.e., $\max_{0 \le K \le \sum_i N_i} SB(K)$, is a single-variable maximization problem. Unfortunately, a closed-form solution to the problem is not available. We use a numerical scheme to solve it instead.

NUMERICAL EXAMPLE

Suppose that there are two firms in the CBD, each with 1500 employees. The capacity of the bottleneck leading to the CBD is assumed to be 1000 veh/hr. Therefore, the peak period will be three hours. For other parameters, $a_1 = a_2 = \$500/day$, $\beta = \$4/hr$. For each combination of θ_1 and θ_2, the optimum number of K is obtained. Two scenarios are created based on the sensitivity parameter of firm 1. In the first case, $\theta_1 = 0.1$, which represents situations where firm 1 is a firm with moderately complementary employees. In the second case, $\theta_1 = 0.04$, representing a firm with less dependent employees. Figures 5-3 and 5-4 depict the number of shifted employees of each firm for the first and second scenario, respectively. It should be emphasized that the corner solutions are not activated for the ranges of θ_2 illustrated in Figures 5-3 and 5-4. It can be observed in the second case, which represents less sensitivity to staggering, fewer mobility credits are issued by the government agency as compared to the first case. As expected, by increasing the sensitivity of firm 2, its shares in staggered employees are reduced.

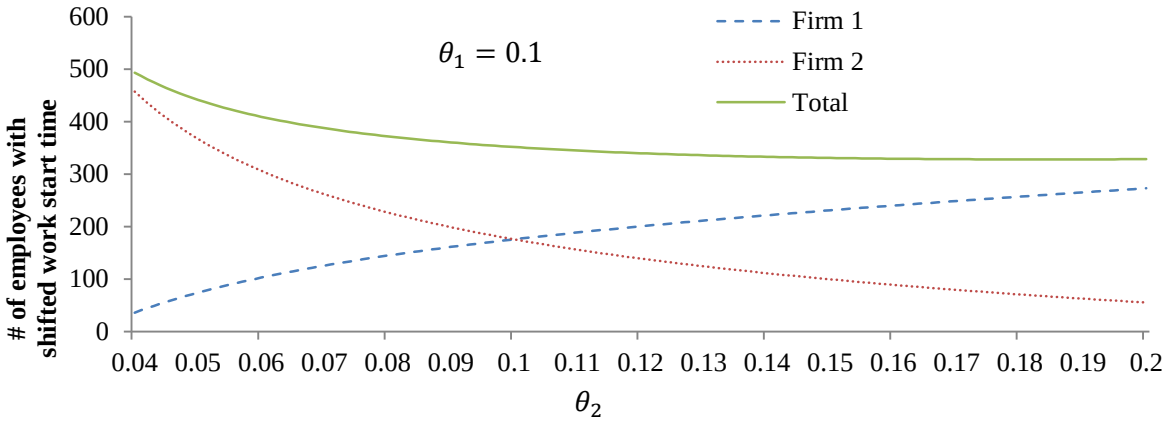


Figure 5-3. Number of shifted employees in scenario 1 (θ_1=0.1).

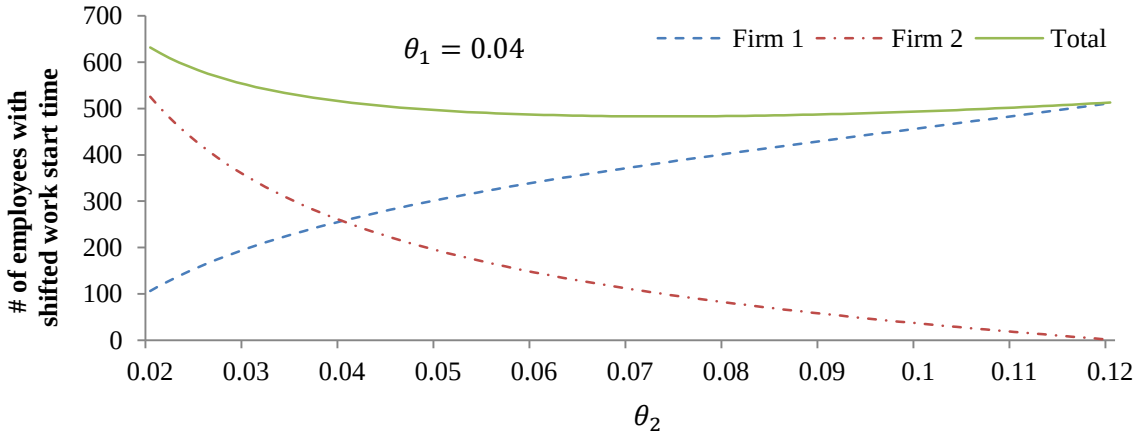


Figure 5-4. Number of shifted employees in scenario 2 (θ_1=0.04).

Figure 5-5 shows the relation of credit price and total number of employees. Among three combinations of low and moderate complementary technology, the lowest price is obtained when both firms have low complementary technology. Surprisingly, the price of credits is lower when both firms have moderately complementary technology, compared to the case where firms have different levels of sensitivity. It is mainly because when both firms have moderately complementary technologies, more credits will be issued by the government agency at social optimum.

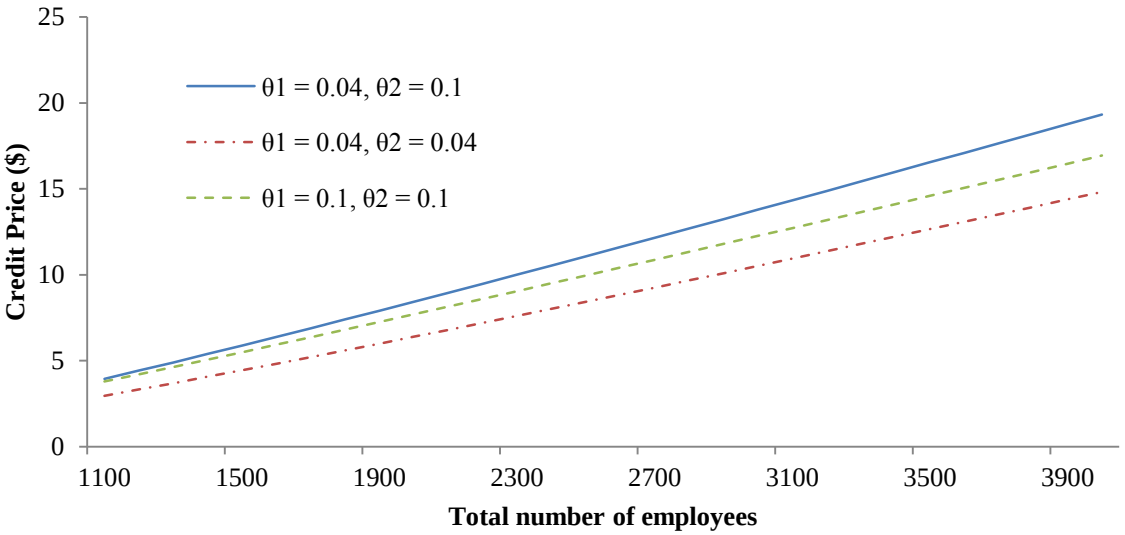


Figure 5-5. Variation of credit price.

The social benefit is improved by the implementation of the proposed tradable credit scheme for all scenarios. Figure 5-6 shows that the improvement in social benefit increases as the bottleneck becomes more congested. As expected, when firms are less sensitive and traffic congestion is more severe, the social benefit improvement is more significant. However, firms are not necessarily better off. For an equal allocation of credits between two firms, both firms will be worse off if firms have the same degree of sensitivity to staggering. However, when their degrees of sensitivity are different, the less sensitive firm would better off as shown in Figure 5-7, while the firm with more complementary technology would be made worse off.

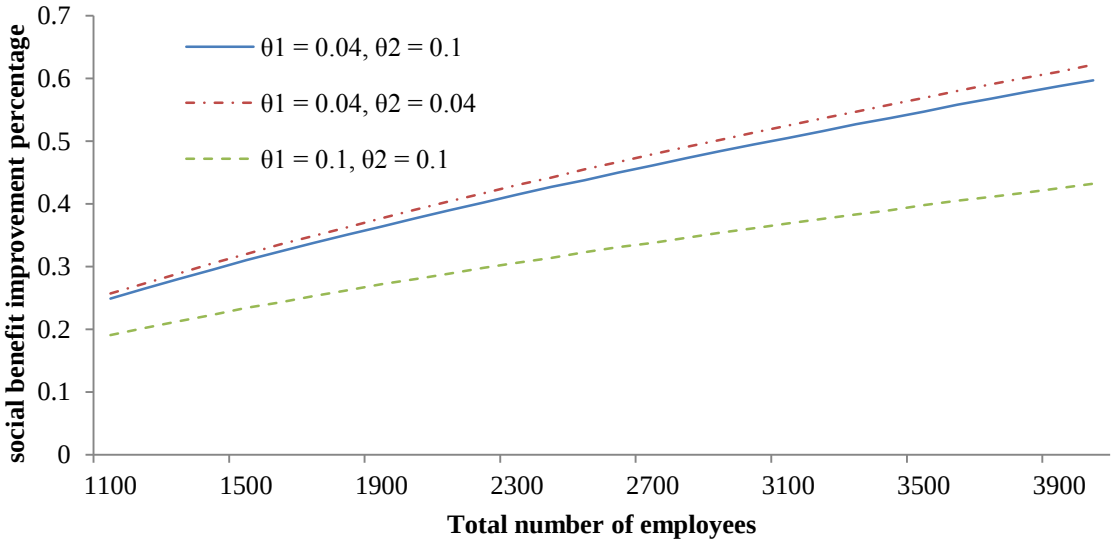


Figure 5-6. Social benefit percentage change compared to existing condition.

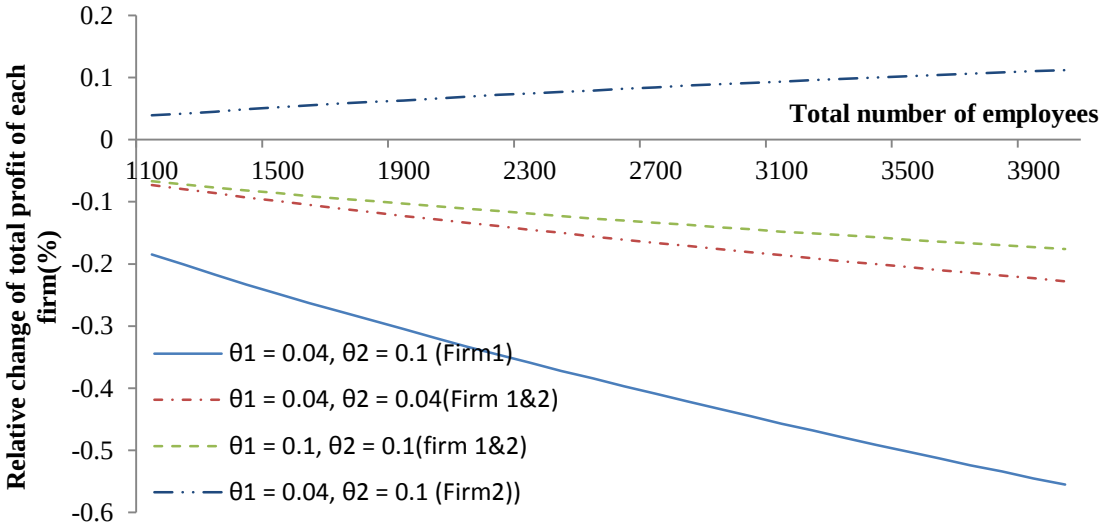


Figure 5-7. Relative change of total profit of firms.

CHAPTER 6 CONCLUSIONS

In Chapter 3, we provided answers to questions: given the current state of the system, what is the toll that minimizes total system delay? What pricing strategy would produce the least total system delay? Is the linear toll pricing applicable to reality? The concept of marginal costs for an alternative needs some clarification as it becomes ambiguous when capacity is reached, i.e. the cost function is not continuous, and 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. Dis-ambiguity can pose some difficulty when implementing the congestion pricing principle. However, we have shown that the left derivative is the only one consistent with the desired SO flow pattern. Linear pricing strategies, as defined in this chapter, 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. From simulation experiments using *GTsim*, VLT was selected as the most efficient pricing strategies to the perspective of the total delay, and FT produced the largest revenue.

In Chapter 4, we investigated approaches to determining optimal operational parameters for a proposed ML pricing scheme with refund option. Deterministic utility functions are adopted for each individual traveller with an underlying VOT distribution across the population. A modified point queue model for traffic propagation is developed to account for the intrinsic randomness in traffic flow. An optimization model with a chance constraint is established to determine the desired inflow to the HOT lane during each tolling interval. The relationship among the optimal operational parameters (including the toll rate, the refund amount, the premium for the refund option, the travel time saving guaranteed by the operator) is discussed for two operation paradigms. The optimal toll rate for Paradigm 1 is calculated based on the desired inflow. The preliminary results show that the models are able to capture important uncertainties in lane choices and traffic flow. The proposed approach in Paradigm 1 is able to deliver HOT performances that are very close to the target operational objectives.

In Chapter 5, we have proposed and analyzed a tradable credit scheme to alleviate the negative impact of staggered work schedules on firms. The results of a numerical example show that the proposed scheme can act as a relief for the productivity loss resulted from not having all employees at the desired work start time. Unfortunately, it does not necessarily provide incentive for firms to stagger their employees. Although it can improve the social benefit, it is not a Pareto-improving scheme. The productivity function utilized to model the behaviors of firms under the proposed scheme does not reflect the effect of congestion cost on their productivity. However, there exist empirical evidences on the negative effect of traffic congestion on firms' productivity. If this effect of congestion is somehow reflected in the productivity function, then firms may also be made better off by the proposed scheme due to the congestion reduction.

Eventually, their behaviors in the credit market would be to make a tradeoff between the negative impact of staggering employees and the positive impact from congestion mitigation. In this situation, we have more chance to find the proposed scheme Pareto improving. Consequently, the scheme may function as an incentive mechanism to foster staggered work time. Our future research will look into this direction.

REFERENCES

1. Supernak, J., Golob, J., Golob, T. F., Kaschade, C., Kazimi, C., Schreffler, E., and Steffey, D. (2002a). San diegos interstate 15 congestion pricing project: Attitudinal, behavioral, and institutional issues. *Transportation Research Record,* 1812, 78–86.
2. Supernak, J., Golob, J., Golob, T. F., Kaschade, C., Kazimi, C., Schreffler, E., and Steffey, D. (2002b). San diegos interstate 15 congestion pricing project: Traffic-related issues. *Transportation Research Record,* 1812, 43–52.
3. Supernak, J., Steffey, D., and Kaschade, C. (2003). Dynamic value pricing as instrument for better utilization of high-occupancy toll lanes. *Transportation Research Record,* 1839, 55–64.
4. Burris, M. W., and Stockton, B. R. (2004). Hot lanes in Houston-six years of experience. Journal of Public Transportation, 7 (3), 1–21.
5. Zhang, G., Yan, S., and 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.
6. Li, J. (2001). Explaining high-occupancy-toll lane use. *Transportation Research Part D,* 6 (1), 61–74.
7. Burris, M. W., and Appiah, J. (2004). Examination of Houstons quickride participants by frequency of quickride usage. *Transportation Research Record,* 1864, 22–30.
8. Podgorski, K. V., and Kockelman, K. M. (2006). Public perception of toll roads: A survey of the texas perspective. *Transportation Research Part A,* 40(10), 888–902.
9. Zmud, J., Bradley, M., Douma, F., and Simek, C. (2007). Attitudes and willingness to pay for tolled facilities: a panel survey evaluation. *Transportation Research Record,* 1996, 58–65.
10. Finkleman, J., Casello, J., and Fu, L. (2011). Empirical evidence from the greater Toronto area on the acceptability and impacts of hot lanes. *Transport Policy,* 18(6), 814–824.
11. Li, J., and 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*.
12. Zhang, G., Wang, Y., Wei, H., and Yi, P. (2008). A feedback-based dynamic tolling algorithm for high-occupancy toll lane operations. *Transportation Research Record,* 2065, 54–63.
13. Yin, Y., and Lou, Y. (2009). Dynamic tolling strategies for managed lanes. *Journal of Transportation Engineering,* 135(2), 45–52.
14. Lou, Y., Yin, Y., and Laval, J. A. (2011). Optimal dynamic pricing strategies for high-occupancy/toll lanes. *Transportation Research Part C,* 19(1), 64–74.
15. Laval, J., and Daganzo, C. (2006). Lane-changing in traffic streams. *Transportation Research Part B,* 40, 251–264.
16. Yin, Y., Washburn, S., Wu, D., Kulshrestha, A., Modi, V., Michalaka, D., and Lu, J. (2012). *Managed Lane Operations–Adjusted Time of Day Pricing vs. Near-Real Time Dynamic*

Pricing Volume I: Dynamic Pricing and Operations of Managed Lanes. Final Report to Florida Department of Transportation.

17. Muñoz, J. C., and Laval, J. A. (2006). System optimum dynamic traffic assignment graphical solution method for a congested freeway and one destination. *Transportation Research Part B,* 40 (1), 1–15.
18. 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.
19. Ungemah, D., Swisher, M., and Tighe, C. (2005). Discussing high-occupancy toll lanes with the denver, colorado, public. *Transportation Research Record: Journal of the Transportation Research Board*, (1932), 129-136.
20. Research and Innovative Technology Administration. (2014). Connected vehicle research in the United States. Retrieved January 25, 2014, from http://www.its.dot.gov/connected_vehicle/connected_vehicle_research.htm
21. Lou, Y. (2013). A unified framework of proactive self-learning dynamic pricing for high-occupancy/toll lanes. *Transportmetrica A-Transport Science,* 9(3), 205-222. doi:10.1080/18128602.2011.559904
22. Gardner, L. M., Bar-Gera, H., and Boyles, S. (2013). Development and comparison of choice models and tolling schemes for high-occupancy/toll (HOT) facilities. *Transportation Research Part B,* 55, 142-153.
23. Duranton, G., and Turner, M. A. (2012). Urban growth and transportation. *The Review of Economic Studies*, *79*(4), 1407-1440.
24. De Palma, A., and Lindsey, R. (2011). Traffic congestion pricing methodologies and technologies. *Transportation Research Part C,* 19(6): 1377–1399.
25. Tsekeris, T., and Voss, S. (2009). Design and evaluation of road pricing: state-of-the-art and methodological advances. *NETNOMICS: Economic Research and Electronic Networking*, 10, 5-52.
26. Wang, X., Yang, H., and Han, D. (2010). Traffic rationing and short-term and long-term equilibrium. *Transportation Research Record: Journal of the Transportation Research Board*, *2196*(1), 131-141.
27. Han, D., Yang, H., and Wang, X. (2010). Efficiency of the plate-number-based traffic rationing in general networks. Transportation *Research Part E, 46*(6), 1095-1110
28. Arnott, R., Rave, T., and Schöb, R. (2005). Alleviating urban traffic congestion. *MIT Press Books*, *1*.
29. Mun, S. I., and Yonekawa, M. (2006). Flextime, traffic congestion and urban productivity. *Journal of Transport Economics and Policy* (JTEP), 40(3), 329-358.
30. Transportation Research Board (1980). *Alternative work schedules: Impacts on transportation*. Tech. Rep. NCHRP Synthesis of Highway Practice 73, Transportation Research Board, Washington, DC.

31. Giuliano, G., and Golob, T. F. (1989). *Evaluation of the 1988 Staggered Work Hours Demonstration Project in Honolulu*: Final Report. *UCI-ITS-RR, 88-5, ISSN: 0193-5860;-UNTRACED*, (88-5).
32. Yoshimura, M., and Okumura, M. (2001). Optimal Commuting and Work Start Time Distribution under Flexible Work Hours System on Motor Commuting. *Proceedings of the Eastern Asia Society for Transportation Studies*, *10*(3), 455–69
33. Fan, W., and Jiang, X. (2013). Tradable mobility permits in roadway capacity allocation: Review and appraisal. *Transport Policy*, *30*, 132-142.
34. Verhoef, E., Nijkamp, P., and Rietveld, P. (1997). Tradeable permits: their potential in the regulation of road transport externalities. *Environment and Planning B*, *24*, 527-548.
35. Viegas, J. M. (2001). Making urban road pricing acceptable and effective: searching for quality and equity in urban mobility. *Transport Policy*, *8*(4), 289-294.
36. Yang, H., and Wang, X. (2011). Managing network mobility with tradable credits. *Transportation Research Part B*, *45*(3), 580-594.
37. Wang, X., Yang, H., Zhu, D., and Li, C. (2012). Tradable travel credits for congestion management with heterogeneous users. *Transportation Research Part E*, *48*(2), 426-437.
38. Zhu, D. L., Yang, H., Li, C. M., and Wang, X. L. (2014). Properties of the multiclass traffic network equilibria under a tradable credit scheme. *Transportation Science (in press)*.
39. Nie, Y. M. (2012). Transaction costs and tradable mobility credits. *Transportation Research Part B*, *46*(1), 189-203.
40. Wu, D., Yin, Y., Lawphongpanich, S., and Yang, H. (2012). Design of more equitable congestion pricing and tradable credit schemes for multimodal transportation networks. *Transportation Research Part B*, *46*(9), 1273-1287.
41. Shirmohammadi, N., Zangui, M., Yin, Y. and Nie, Y, (2013). Analysis and design of tradable credit schemes under uncertainty. *Transportation Research Record*, 2333, 27-36.
42. He, F., Yin, Y., Shirmohammadi, N., and Nie, Y. M. (2013). Tradable credit schemes on networks with mixed equilibrium behaviors. *Transportation Research Part B*, *57*, 47-65.
43. Vickrey, W. S. (1969). Congestion theory and transport investment. *The American Economic Review*, *59*(2), 251-260.
44. Smith, M. J. (1984). The existence of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transportation science*, *18*(4), 385-394.
45. Daganzo, C. F. (1985). The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transportation science*, *19*(1), 29-37.
46. Arnott R, de Palma, A. and Lindsey, R. (1998). Recent developments in the bottleneck model. In: Button, K.J., Verhoef, E.T. (Eds.), *Road Pricing. Traffic congestion and the Environment: Issues of Efficiency and Social Feasibility*. Aldershot, Edward Elgar, UK, 161-179.
47. de Palma, A. and Fosgerau, M. (2011). Dynamic traffic modeling. In: de Palma, A., Lindsey, R., Quinet, E., Vickeman, R. (Eds.), *Handbook in Transport Economics*. Cheltenham, Edward Elgar, UK, 29-37.

48. Xiao, F., Qian, Z. S., and Zhang, H. M. (2013). Managing bottleneck congestion with tradable credits. *Transportation Research Part B*, *56*, 1-14.
49. Nie, Y. M. (2012). A New tradable credit scheme for the morning commute problem. *Networks and Spatial Economics*, 1-23.
50. Nie, Y. M., and Yin, Y. (2013). Managing rush hour travel choices with tradable credit scheme. *Transportation Research Part B*, *50*, 1-19.
51. Yushimito, W. F., Ban, X., and Holguín-Veras, J. (2013). Correcting the market failure in work trips with work rescheduling: an analysis using bi-level models for the firm-workers Interplay. *Networks and Spatial Economics*, 1-33.
52. Wilson, P. W. (1988). Wage variation resulting from staggered work hours. *Journal of Urban Economics*, *24*(1), 9-26.
53. Gutiérrez-i-Puigarnau, E., and Van Ommeren, J. N. (2012). Start Time and Worker Compensation Implications for Staggered-Hours Programmes. *Journal of Transport Economics and Policy (JTEP)*, *46*(2), 205-220.