INSTITUTE OF TRANSPORTATION STUDIES
UNIVERSITY OF CALIFORNIA, BERKELEY

Morning Commute with Competing Modes and Distributed Demand: User Equilibrium, System Optimum, and Pricing

Eric J. Gonzales and Carlos F. Daganzo

WORKING PAPER
UCB-ITS-VWP-2011-2

UC Berkeley Center for Future Urban Transport
A VOLVO Center of Excellence

June 2011

Morning Commute with Competing Modes and Distributed Demand: User Equilibrium, System Optimum, and Pricing

Eric J. Gonzales and Carlos F. Daganzo

June 10, 2011

Abstract

The morning commute problem for a single bottleneck, introduced in Vickrey (1969), is extended to model mode choice in an urban area with time-dependent demand. This extension recognizes that street space is shared by cars and public transit. It is assumed that transit is operated independently of traffic conditions, and that when it is operated it consumes a fixed amount of space.

As a first step, a single fixed-capacity bottleneck that can serve both cars and transit is studied. Commuters choose which mode to use and when to travel in order to minimize the generalized cost of their own trip. The transit agency chooses the headway and when to operate. Transit operations reduce the bottleneck's capacity for cars by a fixed amount. The following results are shown for this type of bottleneck:

1. If the transit agency charges a fixed fare and operates at a given headway, and only when there is demand, then there is a unique user equilibrium.
2. If the transit agency chooses its headway and time of operation for the common good, then there is a unique system optimum.
3. Time-dependent prices exist to achieve system optimum.

Finally, it is also shown that results 2 and 3 apply to urban networks.

1 Introduction

Cities around the world face growing demand for limited street space to meet the transportation needs of their residents. The reality is that these streets can be used by multiple transport modes, and people can choose both when and how they travel. This paper considers the morning commute when cars and transit compete to serve a population of travelers in an urban area who are identical except for their wished time to finish their trips. This problem is important because it lends insights for how to allocate street space and price multiple modes efficiently in urban networks.

The morning commute problem for a single mode was introduced in Vickrey (1969) which considers a population of car commuters who must use a single route with a fixed-capacity first-in, first-out (FIFO) bottleneck to get to work at a desired time. If the demand ever exceeds the capacity of the bottleneck, then commuters will adjust when they travel in response to the resulting delays. Specifically, each commuter chooses their own arrival time at the bottleneck in order to minimize the sum of their own cost of travel, their delay, and the penalty associated with their schedule deviation; i.e., the difference between their wished and actual departure times from the bottleneck. This problem was later expanded to consider a population with a distribution of wished bottleneck departure times which can be represented by a cumulative wish curve (Hendrickson and Kocur, 1981). Smith (1984) showed that an equilibrium exists in which no commuter can unilaterally reduce their own travel cost when the wish curve is S-shaped and the penalty function is smooth and convex. Daganzo (1985) proved that this equilibrium is unique.

The bottleneck model of the morning commute has been studied extensively for the case where all commuters are identical and wish to depart the bottleneck at a common time. For example, Arnott et al. (1990b) proposed an optimal time-dependent pricing scheme (or fine toll) which eliminates the delay. This toll is the difference between the user equilibrium and system optimum costs. These authors also investigated a system with route choice in which identical commuters can choose between multiple parallel congestible routes (Arnott et al., 1990a). Work on the morning commute problem with cars and transit has also been extensive (Tabuchi, 1993; Braid, 1996; Huang, 2000; Danielis and Marcucci, 2002), but it is limited in two main ways. First, commuters have been assumed to share an identical desired bottleneck departure time, and second, only unrealistically simple families of transit mode cost functions have been considered. Existing models, for example, do not recognize that transit operations reduce the remaining capacity for cars, and that the frequency of real transit service is adapted to the number of transit riders. Furthermore, unlike the case of the single bottleneck with distributed demand, the literature does not provide a system optimum solution with two modes, and whether it can be achieved with pricing.

An important extension of the bottleneck model is the morning commute problem on urban networks where origins and destinations are distributed across space. As explained in Daganzo (2007) and experimentally verified in Geroliminis and Daganzo (2008), a network can often be macroscopically modeled as a single bottleneck with state-dependent capacity. The network capacity is a function of the number of vehicles in the network and decreases as queues grow on the streets. The congestion resulting from this reduced capacity has been called hypercongestion (Small and Chu, 2003). Geroliminis and Levinson (2009) employs a macroscopic method to examine pricing strategies for the morning commute problem in a city with only cars. No reference has explored the effect of dedicating space to transit operations on the remaining road capacity for cars and the effect this has on prices. Thus, a natural next step is to look at urban networks where streets can be shared by cars and transit.

As a preliminary step toward this goal, this paper first presents an analysis of the morning commute for a general S-shaped wish curve and a choice between passing a fixed-capacity bottleneck by car or using a general uncongestible alternative transit mode. This capacity depends on whether or not transit service is being provided. Section 2 shows that there is a unique user equilibrium if the transit agency charges a fixed fare and operates at a given headway, and only when there is demand. Section 3 shows that there is a unique system optimum if the transit agency chooses its headway and time of operation for the common good. Section 4 presents and discusses a dynamic pricing strategy which moves the user equilibrium to system optimum. It shows that when modes share the bottleneck, the optimum toll is not always the difference between the user equilibrium and system optimum user costs. Finally, Section 5 shows that even though the user equilibrium for the network problem with state-dependent capacity is somewhat complex, the system optimum version of the problem reduces to the fixed-capacity bottleneck model. More specifically, with suitably modified cost functions, the system optimal travel pattern, pricing strategies, and insights identified in Sections 3 and 4 apply to multimodal urban networks.

2 User Equilibrium

We first review the bottleneck model for a single mode from Hendrickson and Kocur (1981) and then add a transit mode. Consider the morning commute problem with a population of commuters who are identical (e.g., values of travel time and queuing delay) except for when they wish to get to their destination. If commuters drive, they must pass a bottleneck with capacity μ. The total number of commuters that wish to depart from the bottleneck by time t is described by a wish curve, $W(t)$, which is S-shaped. The slope of this curve is the time-derivative of $W(t)$ (denoted by a dot), $\dot{W}(t)$, and it satisfies:

$$\begin{aligned} \dot{W}(t) > \mu \quad & \text{for } t \in (t_1, t_2) \\ \dot{W}(t) \leq \mu \quad & \text{otherwise} \end{aligned} \tag{1}$$

as shown in Figure 1. As a result of the first inequality, there will be a rush period starting at $t_e \leq t_1$ and ending at $t_L \geq t_2$ during which N commuters will experience queuing delay. Suppose that each commuter experiences a penalty for schedule deviation from their wished departure time which is described by a piecewise linear penalty function. Each minute of earliness is associated with a penalty of e equivalent minutes of travel time such that $0 < e < 1$, and each minute of lateness is equivalent to L minutes of travel time such that $L > 0$.

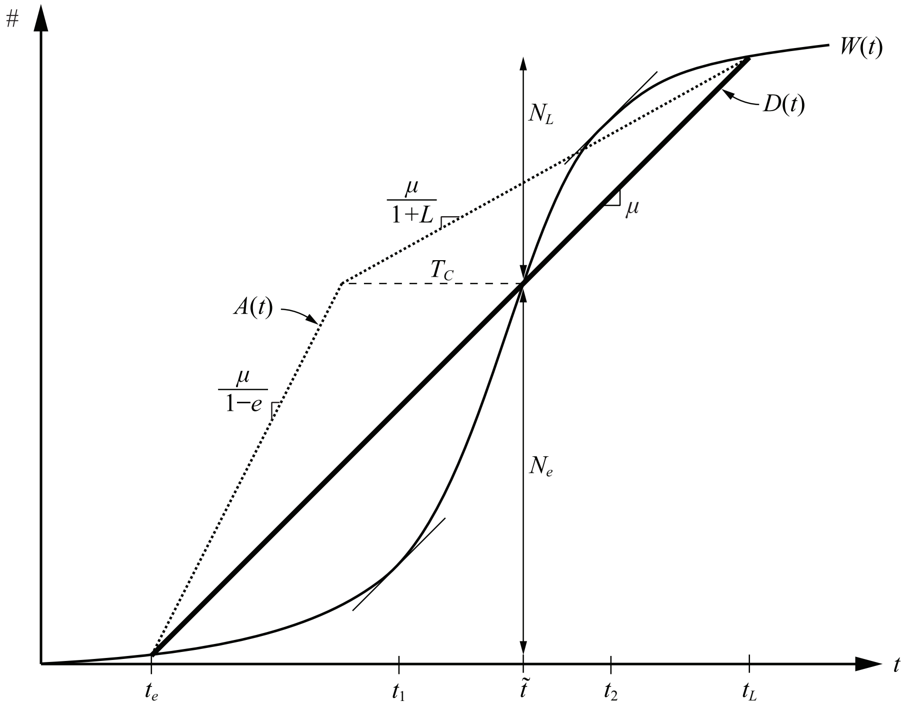


Figure 1: User Equilibrium for a fixed capacity bottleneck using a single mode.

In the absence of an alternative mode, and assuming that commuters arrive and pass the bottleneck in order of wished departure (first-wished, first-in, first-out or FWFIFO), we look for the beginning and end of the rush and for the equilibrium departure curve from the bottleneck which has slope $\dot{D}(t) = \mu$ for $t \in (t_e, t_L)$. This determines the time $\tilde{t}$ when a delayed commuter departs on time, as well as the number of commuters delayed by the bottleneck, N, of which N_e depart early and N_L depart late; see Figure 1. We also look for the user equilibrium arrival curve at the bottleneck, $A(t)$, which allows no commuter to reduce their travel cost by unilaterally changing their own arrival time. The slope of the arrival curve in equilibrium, shown in the figure, must satisfy:

$$\dot{A}(t) = \begin{cases} \frac{\mu}{1-e} & \text{for commuters who depart early} \\ \frac{\mu}{1+L} & \text{for commuters who depart late.} \end{cases} \tag{2}$$

Otherwise, early and late commuters could reduce their travel cost by arriving earlier if the slope was greater or arriving later if the slope was less than those specified in (2). The result is that a critical commuter with wished time $\tilde{t}$ departs the bottleneck on time but experiences the maximum travel cost as queuing delay:

$$T_C = \frac{NeL}{\mu(e+L)} \tag{3}$$

All travelers wishing to pass the bottleneck before $\tilde{t}$ are early in equilibrium, and all travelers wishing to pass after $\tilde{t}$ are late in equilibrium. Their excess costs (queuing and schedule penalty) are less than T_C.

If an alternative public transit mode becomes available, then commuters are able to choose when to travel and which mode to use. It is assumed in this section that the transit agency charges a fixed fare and operates a fixed headway. Suppose that when transit is operating, it is fully segregated on its own lane so that transit services are not subject to traffic congestion. The transit system requires a fixed amount dedicated space, so the bottleneck's remaining capacity to serve cars when both modes are operating is $\tilde{\mu} \leq \mu$. Transit users can always choose to pass the bottleneck at their wished time because use of the mode is not limited by congestion.[1] Therefore, given our assumptions, each transit rider has an identical generalized cost, z_T. This quantity and all costs appearing in this paper are expressed in units of equivalent queuing time (hours). A car trip without delay has a generalized cost of z_C (hours) which is independent of the number of car drivers. Thus, the total cost of driving through the bottleneck will be the sum of this free-flow cost and the excess costs of queuing delay and schedule penalty.

Following Wardrop (1952), it is assumed that at equilibrium each commuter chooses the mode and travel time which minimizes their own generalized cost. Transit will be competitive with the car for at least part of the rush hour if z_T is less than the generalized cost that the critical commuter would experience if transit is not provided: $z_C + T_C$. At equilibrium, the generalized cost of car and transit must be the same when both modes are used, and the generalized cost of a car trip cannot exceed that of a transit trip when only cars are used. Therefore, z_T is an upper bound for the cost of a trip by either mode. When competitive transit is provided, the maximum delay by car, T, satisfies:

$$T = z_T - z_C < T_C. \tag{4}$$

In order to distinguish between the travel patterns of cars and transit, we will consider the arrival and departure curves for each mode. Again, we assume FWFIFO in both cases. $D_C(t)$ is the cumulative number of car departures at the bottleneck, and $A_C(t)$ is the cumulative number of car arrivals. $D_T(t)$ is the cumulative number of transit departures, and the arrival curve of transit is the same curve, $A_T(t) = D_T(t)$, because all transit trips can be completed on time.

[1] This assumption is reasonable for a service using sufficiently large vehicles operated at regular headways but without a fixed schedule. The traveler cannot avoid the waiting time at a transit stop, but they can always board the next vehicle.

An equilibrium is easy to find in two cases: if $z_T < z_C$, then a transit trip is less costly than even a free-flow car trip, and all trips will be made by transit; if $z_T > z_C + T_C$, then there is always a lower cost for traveling by car and the equilibrium will be the same as the single mode problem. The following proposition addresses the remaining cases.

Proposition 1 (User Equilibrium, 2 Modes)**.** *If $W(t)$ is S-shaped, and each commuter can choose between traveling by car (with free-flow cost z_C per trip) through the bottleneck and an alternative transit mode with given cost $z_T \in (z_C, z_C + T)$, there is a unique FWFIFO user equilibrium with the following properties (see Figure 2):*

1. *N_e, the number of early car commuters, is given by $N_e = \mu T/e$. They travel at the beginning of the rush, $t \in (t_e, \tilde{t}_e)$.*
2. *N_L, the number of late car commuters, is given by $N_L = \mu T/L$. They travel at the end of the rush, $t \in (\tilde{t}_L, t_L)$.*
3. *N_o, the number of on-time car commuters in the rush, is a strictly decreasing function of T, $N_o = N_o(T)$. They travel in the middle of the rush, $t \in (\tilde{t}_e, \tilde{t}_L)$.*
4. *N_T, the number of transit riders, is a strictly decreasing function of T, $N_T = N_T(T)$. They also travel in the middle of the rush, $t \in (\tilde{t}_e, \tilde{t}_L)$.*

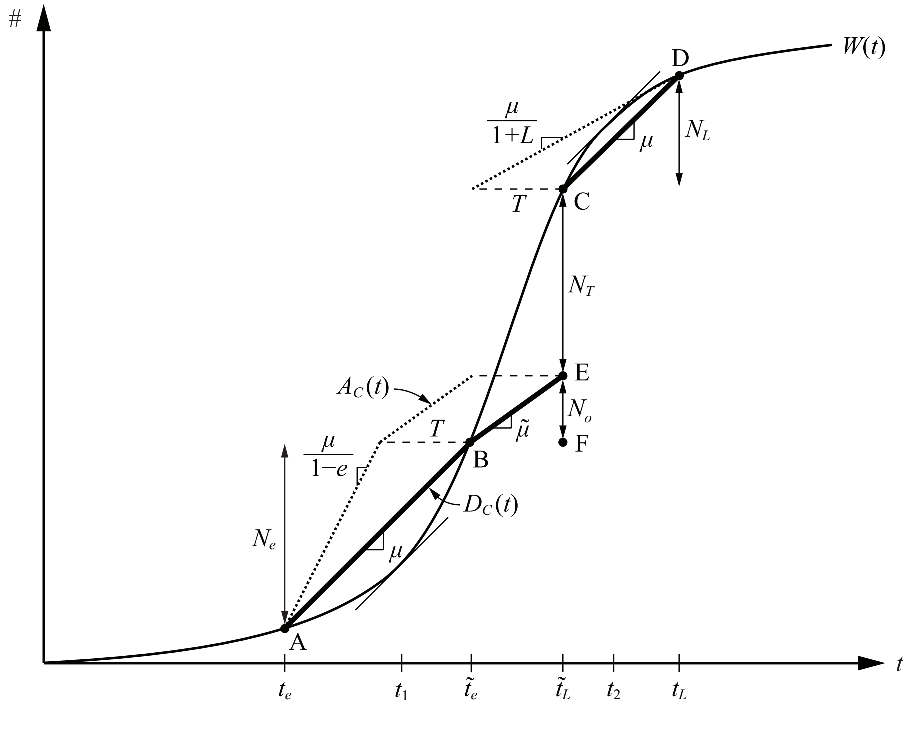


Figure 2: User Equilibrium for a bottleneck with a transit alternative.

Proof. Consider point A ($t = t_e$) where the first early commuter departs. Since the excess cost of driving (queuing delay and schedule penalty) is less than T shortly after this time, only cars are used and therefore $\dot{D}_C(t) = \mu$. For an equilibrium, the slope of the arrival curve should be as in Figure 2: $\dot{A}_C(t) = \mu/(1 - e)$, so that queuing time increases at rate e/μ with each additional commuter. Clearly, the queuing time is T for the $N_e = \mu T/e$ early commuters in agreement with property 1. We choose the unique location of point A such that commuter N_e departs on time at $\tilde{t}_e$ (point B) as shown in the figure. This ensures that the excess cost of driving increases monotonically from 0 to T for commuters departing in $(t_e, \tilde{t}_e)$ in FWFIFO order. Therefore, no transit is used during the early interval. This establishes property 1.

A similar FWFIFO construction is used for the late part of the rush to identify the unique segment $\overline{\text{CD}}$ and the time interval $(\tilde{t}_L, t_L)$ where the excess cost of driving declines monotonically from T to 0. In this interval, queuing time declines at the rate L/μ with each departing commuter, so the number of late commuters is $N_L = \mu T/L$. This establishes property 2.

In the middle of the rush $(\tilde{t}_e, \tilde{t}_L)$, the number of commuters served is the length of segment $\overline{\text{FC}}$. Both car and transit are used. Thus, cars depart in FWFIFO order at rate $\dot{D}(t) = \tilde{\mu}$. They experience a queuing delay T and no schedule penalty. Therefore, their number is $N_o = (\tilde{t}_L - \tilde{t}_e)\tilde{\mu}$ as shown by segment $\overline{\text{EF}}$. The number of transit users, N_T, is given by the length of segment $\overline{\text{CE}}$. They also pass the bottleneck in FWFIFO order, and experience no delay. Note, N_T is always greater than 0 because $\dot{W}(t) > \mu \geq \tilde{\mu}$ in the middle of the rush.

Finally, note from the geometrical construction that if T increases, then $\tilde{t}_e$ increases and $\tilde{t}_L$ decreases; i.e., point B moves to the right along $W(t)$, and point C to the left. Clearly then, both N_o and N_T strictly decrease with T. This establishes properties 3 and 4. □

Note from Figure 2 that the departure curve for cars is piecewise linear in the rush with the slope always equal to the capacity for cars. From Proposition 1, it follows that the number of commuters who depart early and late must satisfy:

$$\frac{N_e}{N_L} = \frac{L}{e}. \tag{5}$$

In this equilibrium, all commuters with wished times in $(\tilde{t}_e, \tilde{t}_L)$ travel on time and experience the same travel cost, $z_T = z_C + T$. The transit service is only used during this period. All early and late commuters travel only by car and experience a lower cost. The total number of travelers in the rush is given by the sum:

$$N = N_e + N_L + N_o + N_T. \tag{6}$$

Each of these values, including N, is uniquely determined for any given $\{W(t),\, e,\, L,\, \mu,\, \tilde{\mu}\}$.

Note by comparing Figures 1 and 2 that the maximum cost of a trip in the two-mode equilibrium is less than that of a single-mode equilibrium. Since $N_T > 0$ implies $T < T_C$, it follows from properties 1 and 2 of Proposition 1 that there are fewer early and late commuters. These are represented by shorter segments $\overline{\text{AB}}$ and $\overline{\text{CD}}$ in Figure 2, which implies that the rush starts later and ends earlier with two modes than with all commuters traveling by car. Therefore, the rush period with multiple modes is shorter and involves fewer commuters. Provision of a competitive public transit alternative to congested driving is a Pareto improvement because every delayed commuter experiences a reduced travel cost, even those who travel by car at the beginning and end of the rush when no transit service is used.

3 System Optimum

The system optimal travel pattern will minimize the total system cost (or maximize welfare) associated with the bottleneck. Since queuing delay is an avoidable waste of time, $A_C(t)$ must equal $D_C(t)$ at system optimum. Thus, to find the system optimum, it suffices to identify the departure curves for car and transit that minimize the monetary mode costs (e.g., vehicles, fuel, infrastructure, etc.), the free-flow travel time, and the schedule penalty. We do this in general and then for a Z-shaped wish curve.

3.1 General Wish Curve

In order to minimize the total system cost, we must consider the total generalized cost function of each mode. It is assumed that the transit spatial coverage is given, but its headway is chosen to minimize the sum of the agency and the user costs (including the out-of-vehicle wait) for the given number of transit riders, N_T. Thus, the system optimum transit cost is a function of the number of transit users, $Z_T(N_T)$.[2] The system optimum problem is approached in two steps. First, we determine how car users and transit riders should behave if we are given that there are a total of N_T transit riders by the end of the peak period, t_{max}. The resulting costs are also determined. Then, the optimal number of transit riders, N_T^*, is identified by minimizing the system costs. All values associated with the system optimum are denoted with $*$.

To start, let us define the curve $W_L(t) \doteq W(t) - N_T$. This is a lower bound to $W_C(t)$, the number of car users that wish to depart the bottleneck by time t when there are N_T transit users. Logically, $W(t)$ is an upper bound for $W_C(t)$.

Proposition 2. *For a given wish curve, $W(t)$, and a given number of transit riders, N_T, there is a unique system optimal departure curve for cars, $D_C(t)$, and transit, $D_T(t)$. The $D_C(t)$ curve is piecewise linear with 3 segments going from $W(t)$ to $W_L(t)$ (see Figure 3):*

Phase 1. $\dot{D}(t) = \mu$ while above $W(t)$, serving N_e^* trips (segment $\overline{AB}$); $W_C(t) = W(t)$; no transit is used.

Phase 2. $\dot{D}(t) = \tilde{\mu}$ from $W(t)$ to $W_L(t)$, serving N_o^* trips (segment $\overline{BC}$); $W_C(t) = D_C(t)$; and $D_T(t) = W(t) - W_C(t)$.

Phase 3. $\dot{D}(t) = \mu$ below $W_L(t)$, serving N_L^* trips (segment $\overline{CD}$); $W_C(t) = W_L(t)$; no transit is used.

Proof. Since N_T is given all monetary costs and free-flow travel times are fixed. Thus, the optimal departure curves must minimize only the remaining schedule delay for cars. In order to identify the optimal $D_C(t)$ and $D_T(t)$ we must also identify $W_C(t)$. This curve is bounded above by $W(t)$ and below by $W_L(t) = W(t) - N_T$ and must satisfy the following criteria (illustrated in Figure 3): $W_C(t)$ must start on $W(t)$ and end on $W_L(t)$; and for all t, $0 \leq \dot{W}_C(t) \leq \dot{W}(t)$.

We now show that there is a unique system optimal solution with the stated properties, as depicted in Figure 3. Consider the point B where the $W_C(t)$ diverges from $W(t)$. To the left of B, the schedule delay is minimized because $D_C(t)$ is as low as possible and $W_C(t)$ is as high as possible. Note, there is no transit use because $W(t) - W_C(t) = 0$. Thus, for the given B, phase 1 is optimum. Likewise, to the right of point C where $W_C(t)$ joins $W_L(t)$, the schedule delay is minimized when $D_C(t)$ is as high as possible and $W_C(t)$ is as low as possible. There is also no transit because $\dot{W}_C(t) = \dot{W}(t)$. Thus for a given point C, phase 3 is optimum.

The schedule penalty can be made equal to 0 in phase 2 by choosing $W_C(t) = D_C(t)$. Therefore, for a given B, $D_C(t)$ should be chosen to minimize the schedule delay in phase 3. This is achieved by choosing the highest possible slope for $D_C(t)$. Note, $\dot{W}(t) > \mu$

[2] $Z_T(N_T)$ is a concave function that increases with $\sqrt{N_T}$ when the headway is determined endogenously to minimize the total generalized cost of the transit system (Gonzales, 2011).

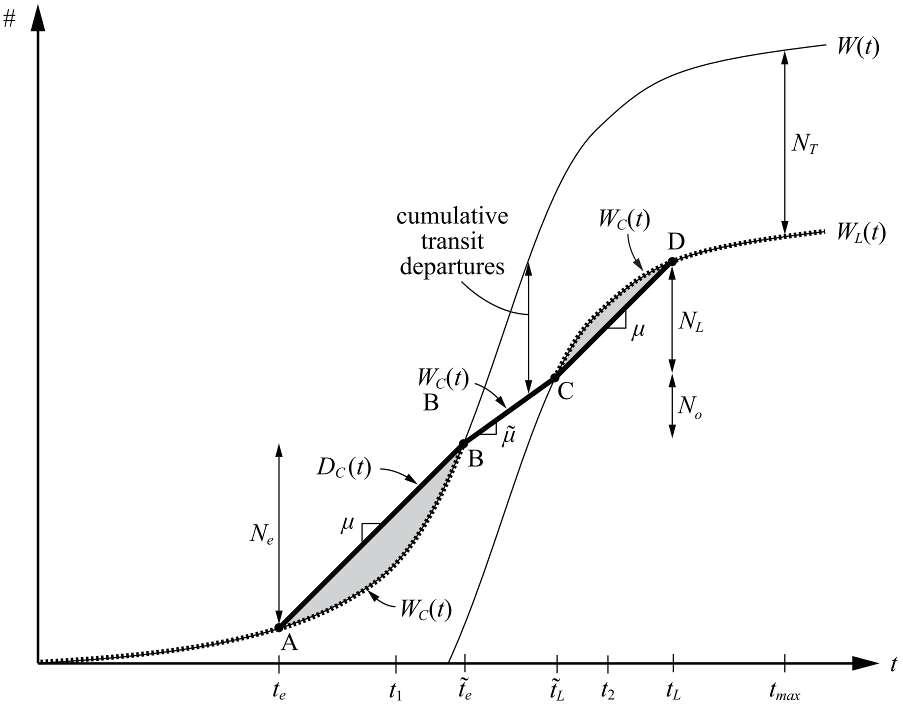


Figure 3: The departure curve for car giving minimum schedule delay for a given N_T.

in phase 2, so transit is used. Thus, $\dot{D}(t) = \tilde{\mu}$ in phase 2, as shown in Figure 3, and $D_T(t) = W_T(t) = W(t) - W_C(t)$ because all transit trips are served on time. Clearly, for a given B, there is a unique segment $\overline{\text{BC}}$ in phase 2, representing $W_C(t)$ and $D_C(t)$, that minimizes the schedule penalty.

Since point C is uniquely determined by point B, it only remains to pick the point B which corresponds to the minimum total schedule cost of earliness and lateness. An upward shift of B along $W(t)$ corresponds to a shift of the departure curve for early car commuters by dn_e and for late car commuters by dn_L; see Figure 4. Consideration of Figures 3 and 4 shows that this corresponds to an increase in earliness $(eN_e/\mu)dn_e$, and a decrease in lateness $(LN_L/\mu)dn_L$. Therefore, the schedule cost is minimized at the unique point when these two quantities (i.e., the shaded areas in Figure 4) are equal. This unique B defines the optimal solution. □

Proposition 2 allows us to uniquely define the number of car users for a given N_T, $N_C(N_T)$. Therefore, we can define three functions of N_T: the total transit system cost, $Z_T(N_T)$; the total car cost, $Z_C(N_T)$; and the total schedule cost, $S(N_T)$. Thus, the minimum total cost for a given N_T can be defined as:

$$Z(N_T) = Z_T(N_T) + Z_C(N_T) + S(N_T). \tag{7}$$

The system optimum cost is the global minimum of this function and the optimum number of transit riders is $N_T^* = \arg\min\{Z(N_T)\}$.

It is shown in Appendix A that at system optimum

$$\frac{N_e^*}{N_L^*} = \frac{L(\lambda_e^* - \tilde{\mu})(\lambda_L^* - \mu)}{e(\lambda_e^* - \mu)(\lambda_L^* - \tilde{\mu})} \tag{8}$$

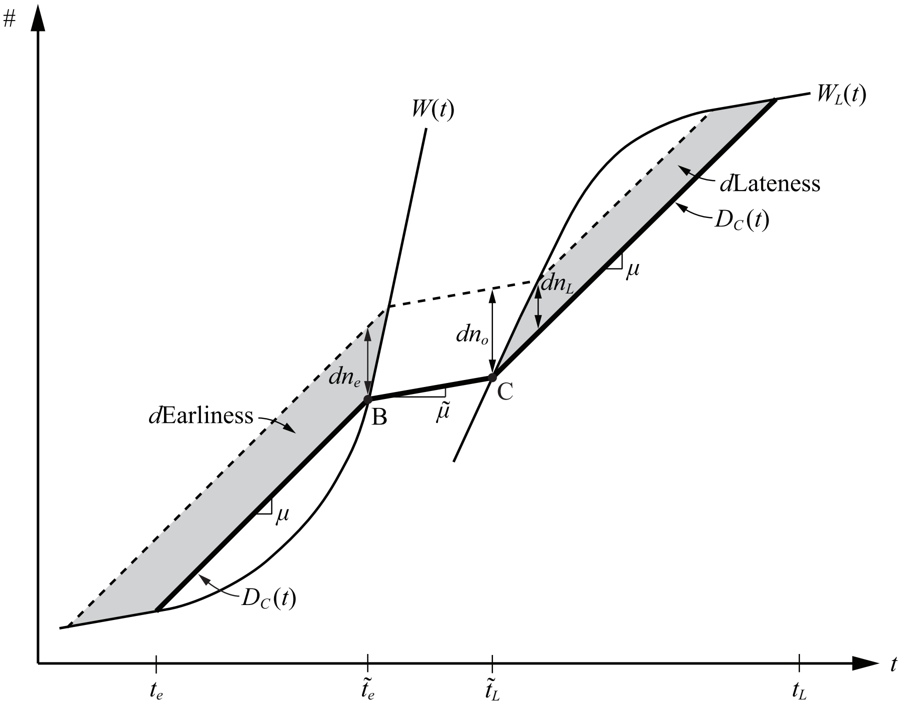


Figure 4: Decrease in earliness and increase in lateness resulting from a shift of B.

where we define $\lambda_e^* \doteq \dot{W}(\tilde{t}_e^*)$ and $\lambda_L^* \doteq \dot{W}(\tilde{t}_L^*)$, which are the slopes of $W(t)$ at the system optimum points B and C, respectively. Note that $N_e^*/N_L^* \neq L/e$ in general. Thus, the ratio of early to late commuters can be different in the user equilibrium and system optimum. However, if transit is operated on a separate right of way, so that $\mu = \tilde{\mu}$, or if $W(t)$ is Z-shaped so that $\lambda_e^* = \lambda_L^*$, then the ratio is the same: $N_e^*/N_L^* = L/e$.

3.2 Z-shaped Wish Curve

Now we examine the Z-shaped case as shown in Figure 5 in more detail. In this case, explicit forms for N_T^* and $Z^*(N_T^*)$ are derived below. As a first step, we use the geometry of the Z-shaped wish curve to derive $S(N_T)$.

Lemma 1. *If $W(t)$ is Z-shaped with slope λ during a peak of length t_p and 0 otherwise, then the optimum schedule delay for a given number of transit riders is given by:*

$$S(N_T) = \begin{cases} \left(t_p - \frac{N_T}{\lambda-\tilde{\mu}}\right)^2 \frac{\lambda e L(\lambda-\mu)}{2\mu(e+L)} & \textit{for } N_T < t_p(\lambda - \tilde{\mu}) \\ 0 & \textit{otherwise.} \end{cases} \quad (9)$$

Proof. Consider Figure 5 illustrating the Z-shaped $W(t)$ and the optimal $D_C(t)$ for N_T as given by Proposition 2. In the middle of the rush, transit demand is $\lambda - \tilde{\mu}$, so to serve N_T commuters, transit is operated for a duration of time, $N_T/(\lambda - \tilde{\mu})$. All of the demand in the remaining time is served only by cars. This demand, $N_e + N_L$, is the difference between the total demand λt_p and the total demand in the middle of the rush $\lambda N_T/(\lambda - \tilde{\mu})$, i.e.:

$$N_e + N_L = \lambda \left(t_p - \frac{N_T}{\lambda - \tilde{\mu}}\right). \quad (10)$$

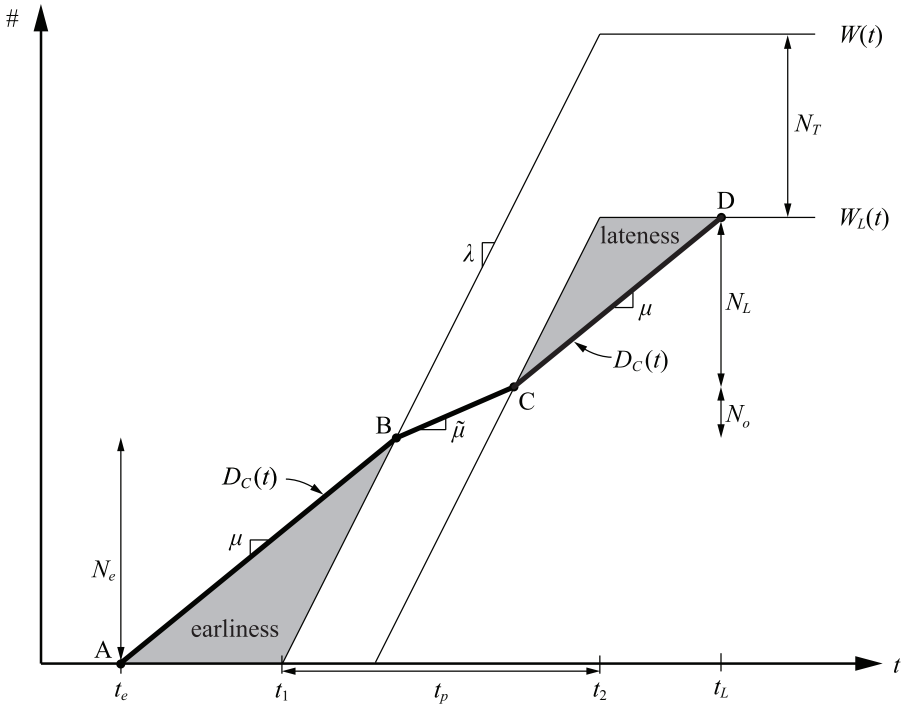


Figure 5: System optimal schedule delay for Z-shaped $W(t)$.

Since the demand rate is always λ during the peak, then $\lambda_e^* = \lambda_L^* = \lambda$. Then following from (8), $N_e/N_L = L/e$ at system optimum. Substituting this ratio into (10), N_e and N_L are each defined by N_T as:

$$N_e = \frac{L}{e+L}\lambda\left(t_p - \frac{N_T}{\lambda - \tilde{\mu}}\right) \tag{11}$$

$$N_L = \frac{e}{e+L}\lambda\left(t_p - \frac{N_T}{\lambda - \tilde{\mu}}\right) \tag{12}$$

when $N_T < t_p(\lambda - \tilde{\mu})$. Otherwise, transit operates for the full duration of the rush and there are no early or late commuters.

The total earliness is the area between $D_C(t)$ and $W_C(t)$ when commuters depart early (the triangle below segment $\overline{\text{AB}}$): $N_e/2\,(N_e/\mu - N_e/\lambda)$. The cost of the earliness is the product of this area and e. Similarly, the total lateness is the area between $D_C(t)$ and $W_L(t)$ when commuters depart late (the triangle above $\overline{\text{CD}}$): $N_L/2\,(N_L/\mu - N_L/\lambda)$. The cost of the lateness is the product of this area and L. The sum of these two costs is the total schedule cost, S, and by simplifying we find:

$$S = \left(eN_e^2 + LN_L^2\right)\frac{\lambda - \mu}{2\lambda\mu}. \tag{13}$$

Now S is expressed in terms of N_e and N_L which are both functions of N_T. Substituting (11) and (12) into (13), we obtain $S(N_T)$, which reduces to (9). □

This result is now used to look for the optimum N_T. To this end, let us denote by N the total number of trips, $N = \lambda t_p$. Then, the optimal transit ridership, N_T^*, is the global minimum of:

$$Z(N_T) = Z_T(N_T) + (N - N_T)z_C + S(N_T). \tag{14}$$

This global minimum is identified below.

Proposition 3. *If $W(t)$ is Z-shaped with slope λ during a peak of length t_p and 0 otherwise, then $Z(N_T)$ has as most one unconstrained local minimum in [0,N] at some point $N_T^u \in (0, N)$. The point, N_T^u is the solution of:*

$$Z_T'(N_T^*) - z_C - \frac{eL}{\mu(e+L)}\left(t_p - \frac{N_T^*}{\lambda - \tilde{\mu}}\right) = 0 \tag{15}$$

and must satisfy $N_T^u \in (0, t_p(\lambda - \tilde{\mu}))$ and

$$Z_T''(N_T^*) + \frac{eL}{\mu(e+L)(\lambda - \tilde{\mu})} > 0. \tag{16}$$

If such a solution does not exist, then there is no unconstrained minimum.

The system optimum N_T^* is either N_T^u or at an extreme of the interval [0, N], whichever value produces the least cost. Thus, three cases are possible:

1. *Trips served by a mix of cars and transit , $Z_M = Z(N_T^u)$, if N_T^u exists.*
2. *All trips served by car, $Z_C = z_C N$.*
3. *All trips served by transit, $Z_T = Z_T(N)$.*

Proof. The total cost is composed of three terms: $Z(N_T) = Z_T(N_T) + Z_C(N_T) + S(N_T)$. Recall that $Z_T(N_T) = A + BN_T + C\sqrt{N_T}$, $Z_C(N_T) = (N - N_T)z_C$, and that $S(N_T)$ is given by (9).

We first examine the existence of unconstrained local minima of $Z(N_T)$. Note that $Z(N_T)$ is a twice differentiable function in the range $[0, N]$, although the second derivative is discontinuous at $N_T^0 = t_p(\lambda - \tilde{\mu})$. Note also that $Z(N_T)$ is concave for $N_T \in [N_T^0, N]$. Thus, it can only have a local unconstrained minimum in $[0, N_T^0)$. The necessary and sufficient conditions for such a minimum are (15) and (16).

We now show that this unconstrained minimum is unique. Consideration shows that $Z''(N_T)$ is monotonic increasing in $[0, N_T^0]$ and that it can be 0 at a unique inflection point N_T^1 in the interval. In other words, $Z(N_T)$ is concave in $(0, N_T^1]$ and convex in $[N_T^1, N_T^0]$. Therefore, any unconstrained local minimum N_T^U must be unique.

Finally, since $Z(N_T)$ has at most one unconstrained local minimum, it follows that the global minimum must be either the local minimum (if it exists) or an extreme point of the optimization interval $[0, N]$. □

Note from Proposition 2 that to identify the system optimum solution it is necessary to know $W(t)$, which is not directly observable. However, if $W(t)$ is Z-shaped, we see from Proposition 3 that we only need the observable values: N, μ, $\tilde{\mu}$, and $\lambda - \tilde{\mu}$ (demand rate on transit). The values of e and L can be estimated from revealed preferences in equilibrium by measuring the rate at which delays increase and decrease over the rush.

3.3 Captive Transit Riders

With heterogeneous populations that include people who are transit captives, public transportation service should never be completely turned off. The car-only scenario (case 2 in Proposition 3) should be modified to include transit service operated in traffic with a minimum acceptable frequency. If desired, transit can be given priority in this modified car-only scenario, provided that car traffic is allowed into the bus lanes when buses are not present. The analysis with a binary population including transit captives and modal choice-makers is straight-forward. The total generalized cost would combine the total cost of the transit captives, which is easy to express as a function of the duration of the middle period when intensive and segregated transit service is provided since captives do not have a choice, with the cost of the remaining population which can be analyzed as in the paper since these choice-makers do not interact at all with the captive population. The results, including the pricing mechanism, should be qualitatively similar to those provided in the paper. More refined classifications of the population (e.g., by income) can also be considered. In this case, we expect the policies proposed in this paper to increase welfare, but a rigorous analysis of optimality is difficult because the FWFIFO rule may be sub-optimal in this case. Simulations may be the best option to analyze the distribution of benefits.

4 System Optimal Pricing of Cars and Transit

Now that the user equilibrium and system optimum have been identified for a bottleneck serving cars and transit, we will turn our attention to a pricing strategy that will achieve system optimal behavior in equilibrium. Commuters are assumed to choose when to travel and which mode to use based on the generalized cost of their own trip, which includes as components: the travel time, vehicle costs, schedule delay, and any pricing fees. As before, this generalized cost is expressed in units of equivalent queuing time.

Suppose that in the absence of pricing, the users of each mode must cover its costs, so drivers pay z_C as a base rate and transit riders pay z_T. The pricing strategy will define the additional car toll $\$_C(t)$ and transit fare $\$_T(t)$ that users passing through the bottleneck at time t must pay. Therefore, the user cost of a free-flow car trip at time t is $z_C + \$_C(t)$ (hours) and the user cost of a transit trip is $z_T + \$_T(t)$ (hours). A negative price represents a subsidy. We look for a set of prices that produces an equilibrium when added to the system optimum costs of Section 3.

Proposition 4 (Optimal Prices)**.** *For any time-dependent car price satisfying*

$$\dot{\$}_C(t) = e \quad \textit{for } t \in (t_e^*, \tilde{t}_e^*) \tag{17a}$$

$$\dot{\$}_C(t) = -L \quad \textit{for } t \in (\tilde{t}_L^*, t_L^*) \tag{17b}$$

$$-L < \dot{\$}_C(t) < e \quad \textit{otherwise}, \tag{17c}$$

the following time-dependent price for transit,

$$\$_T(t) = z_C - z_T + \$_C(t) \quad \textit{for } t \in (\tilde{t}_e^*, \tilde{t}_L^*), \tag{18}$$

produces an equilibrium at system optimum.

Proof. Equations (17) are considered first. To this end, note from Figures 3 and 5 that if an early driver departs dt later in the system optimum solution, then his or her schedule penalty is reduced by edt for each additional dt in the departure time. Therefore, to cancel this benefit and ensure equilibrium we must increase the toll by an additional edt as a toll. Thus, the optimal toll must increase at rate e when commuters depart early in agreement with (17a). If this happens, early commuters do not have an incentive to choose any other departure time (early or late) Likewise, the system optimum toll must decrease at rate $-L$ for commuters who depart late as expressed in (17b). Finally, any commuter who departs on-time by car or transit will not have an incentive to change their departure time to any other time if the the rate at which costs change with time is in $(-L, e)$. This establishes (17c).

It now remains to show that commuters do not have an incentive to change modes. This is achieved by setting the transit prices. Note that the transit service is only used during the middle of the peak, $t \in (\tilde{t}_e^*, \tilde{t}_L^*)$, so its price must only be set for this interval. Since car and transit are used simultaneously during this period, the user cost of travel by both modes must be equal in order to maintain the Wardrop equilibrium; i.e., $\$_T(t) + z_T = \$_C(t) + z_C$, which reduces to (18). $\square$

Note that for the case without transit (i.e., $\tilde{t}_e^* = \tilde{t}_L^*$) only (17a) and (17b) apply. Thus, the prices defined in Proposition 4 are the Vickrey (1969) prices. Figure 6 illustrates optimal prices for a special case in which the car price is fixed at $\$_C(t) = \$_C^{\text{off-peak}}$ outside the rush. From the system optimum described in Section 3, N_e^* car commuters depart the bottleneck early at rate μ between t_e^* and $\tilde{t}_e^*$ (points A and B). Since the toll must increase at rate e during this interval, the car toll increases by $\Delta\$_e^* = eN_e^*/\mu$ from the first to last early commuter. Likewise, the car toll decreases by $\Delta\$_L^* = LN_L^*/\mu$ for late commuters from $\tilde{t}_L^*$ to t_L^*. In the middle of the rush, $(\tilde{t}_e^*, \tilde{t}_L^*)$, all commuters are on time, so the optimal price can follow any curve from point B to C satisfying the third condition of (17); e.g., the solid curve shown. Feasible prices are bounded by the dashed diamond. The system optimal price of transit is the same shape as $\$_C(t)$ translated down by $z_T - z_C$ as defined by (18).

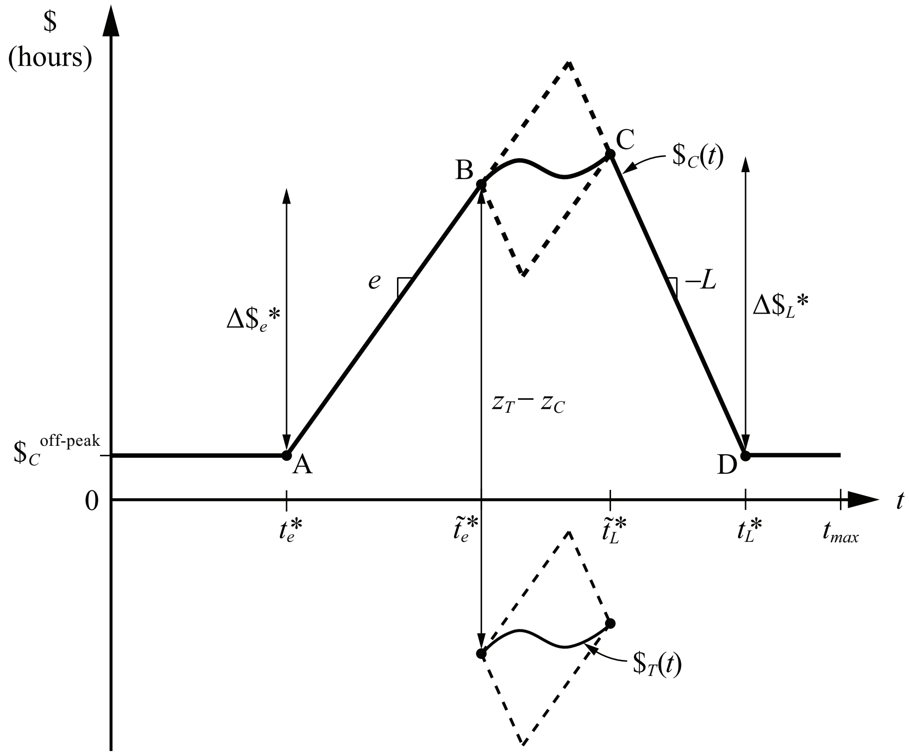


Figure 6: The system optimal time-dependent price for car and transit for the special case when the car toll is fixed at $\$_C(t) = \$_C^{\text{off-peak}}$ in the off-peak.

Note that any vertical translation of the transit and car curves satisfies (17) and (18) and therefore will result in the same system optimal travel pattern. Thus, by shifting these prices up or down, it is possible to achieve additional policy objectives such as any particular car toll during the off-peak or revenue neutrality. From the system optimal cumulative departures of cars $D_C(t)$ and of transit $D_T(t)$, the net revenue $\$_{net}$ is given by

$$\$_{net} = \int_0^{t_{max}} \$_C(t)\dot{D}_C(t) + \$_T(t)\dot{D}_T(t)dt \tag{19}$$

where t_{max} is the amount of time until the next rush period. Since the car price can take any value in the off-peak, there is always a system optimal pricing strategy which is also revenue neutral.

5 Competing Modes on Urban Networks

The previous sections have shown the user equilibrium and system optimum for two competing modes using a single bottleneck with fixed capacity. Here, we extend the results to urban networks. A bottleneck on a road will discharge vehicles at fixed capacity as long as there are vehicles in a queue feeding it.[3] However, the capacity of an urban network to discharge vehicles to their destinations depends on the number of vehicles circulating in the network. Unlike a bottleneck on a single road, queues of vehicles in a network tend to block other streets and impede network flow. Recent work suggests that there is a consistent macroscopic relationship between the average network vehicle density and average network flow called a Macroscopic Fundamental Diagram (MFD), and when the average trip length is not changing, the MFD defines a consistent function relating the number vehicles in the network to the discharge flow of exiting vehicles (Daganzo, 2007; Geroliminis and Daganzo, 2008). We call this second relationship the Network Exit Function (NEF). This relationship describes the state-dependent discharge rate (capacity) of a network as a function of the number of vehicles in the network.

Consider a network with a general concave MFD as illustrated in Figure 7(a). The average vehicle flow on the network, q (veh/lane-hr), is a function of the average vehicle density on the network, k (veh/lane-km). So, the MFD describes $q = Q(k)$ for all possible vehicle densities, and the shape depends on the properties of the network (e.g., saturation flow per lane, free-flow vehicle speed, block lengths, and signal timings). As presented in Daganzo (2007), the MFD can be used to derive the NEF which expresses the flow of vehicles exiting the network, f (veh/hr), as a function of the total number of vehicles circulating in the network, n (veh):

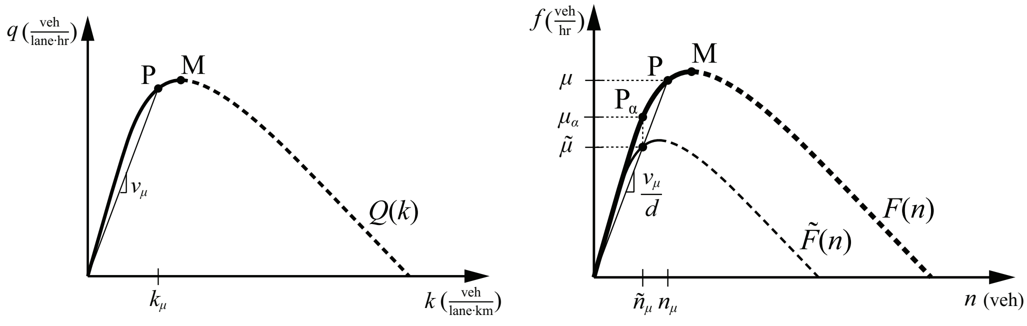
$$f = F(n) = \frac{l}{d} Q\left(\frac{n}{l}\right) \tag{20}$$

where l (lane-km) is the total length of the network, and d (km) is the length of a vehicle trip. Note that the exit function, $F(n)$, is simply a rescaling of the MFD, $Q(k)$, to account for the size of the network and length of trips; see the heavy curve in Figure 7(b). We will study this system assuming that the instantaneous exit flow depends only on the number of vehicles in the network at that time.[4] A vehicle exiting a network is analogous to a vehicle departing a bottleneck, so we can think of the network as a bottleneck with the state-dependent capacity given by the exit function.

The maximum feasible exiting flow is associated with point M in Figure 7. For a given traffic state on the MFD (such as point P), the slope from the origin represents the average vehicle speed across the network, v_μ, which includes time spent at signals and in queues. The total time required to complete a trip of length d is the reciprocal of the analogous

[3] This is approximately true, although evidence suggests that the queue discharge rate is reduced when queues grow very long (Koshi et al., 1992).

[4] This assumption holds when traffic is in a steady state. Transitions between steady states are not instantaneous but have durations comparable to a trip time (Daganzo, 2007). The effect of these transitions in the system optimum will be small if the rush period is long compared to the duration a trip.



(a) Macroscopic Fundamental Diagram (MFD)

(b) Network Exit Function (NEF)

Figure 7: The MFD and NEF for a network with and without transit. Dashed lines indicate congested traffic states, and solid lines are uncongested states where a target operating state P may be reasonably chosen.

slope on the NEF, v_μ/d. Traffic states to the right of M (dashed lines in Figure 7) are congested and should always be avoided because the same average flow and exit flow can achieved with greater traffic speeds and fewer vehicles on the road to the left of M (solid lines in Figure 7).

Geroliminis and Levinson (2009) provides a numerical method to construct the user equilibrium for a single mode on a network with a stable, single-valued NEF. The user equilibrium problem in networks is complicated by the reduced exit flow when the network becomes congested. Fortunately, the system optimal network problem is not affected by this complication because only uncongested traffic states (to the left of M) should occur. Geroliminis and Levinson (2009) also presents the system optimum and optimal pricing strategies for a network with a single mode taking advantage of this result. Now, by keeping the traffic states only on the uncongested side of the NEF, we look at the network system optimum problem for two modes.

Although point M corresponds to the maximum feasible exit flow, a city could choose to put a limit on exit flow by capping it at a target level μ associated with point P to the left of M. A lower target exit flow lengthens the rush but serves each vehicle with less travel time. Figure 7(b) shows that at point P, μ is associated with a critical accumulation of vehicles on the network, n_μ, such that $\mu = F(n_\mu)$. We will define delay as the excess travel time over d/v_μ for a trip of length d. So, in system optimum where delays are avoided, d/v_μ can be interpreted as the maximum travel time guarantee.

If the transit service uses a separate right of way (e.g., metro or permanent dedicated bus lanes), it has no impact on the street network. In this case, $F(n)$ does not change when transit is provided, and $\tilde{\mu} = \mu$. By applying system optimal pricing as described in Section 4, car commuters will choose to travel at rate μ, so the network will maintain a steady accumulation of n_μ vehicles during the rush. No delay will be experienced.

In reality, transit services often use the same street space as other vehicles, so deploying buses will reduce the remaining capacity available for cars. Suppose that the spatial structure of the transit system is given but the headway is endogenously determined by the transit demand.[5] We assume that a fixed number of lanes are dedicated to transit, so $\tilde{\mu}$ is constant when transit is operated. If dedicated space for transit is deployed uniformly

[5] Daganzo (2010) shows that the optimal spatial structure of transit service is insensitive to demand, whereas the optimal headway is not. In a very well-run city, street space could be dedicated to transit with intermittent priority (Eichler and Daganzo, 2006) and the spatial requirement of transit would be a function of the transit demand, but here the spatial requirement is considered fixed.

across the network, the effect should be the same as reducing the network length uniformly leaving a fraction $\alpha < 1$ of the original network length remaining available for cars. For dedicated transit lanes, α will be directly related to the lane distance that is dedicated only to transit. For buses and trams operating in mixed traffic lanes, α must account for the losses due to conflicts between the different types of vehicles. The result is that the capacity of each individual street to serve cars is reduced on average to α times its original. This is the same effect as reducing the network length for cars from l to αl.

Since the change in network size is uniform and none of the other determinants of network capacity have been altered, the MFD as described by $Q(k)$ should remain unchanged. Thus, we see from (20) that the NEF when transit is operated, $\tilde{F}(n)$, is:

$$\tilde{F}(n) = \frac{\alpha l}{d} Q\left(\frac{n}{\alpha l}\right) \tag{21}$$

which is shown in Figure 7(b). Note that the point P associated with the target exit flow moves along the ray with slope v_μ/d towards the origin so the travel time per trip does not change. This peak is associated with the same density k_μ as before, so the optimal car accumulation when both modes are operating, $\tilde{n}_\mu$, and the exit flow (capacity) for cars, $\tilde{\mu}$, are given by:

$$\tilde{n}_\mu = \alpha n_\mu \tag{22}$$

$$\tilde{\mu} = \alpha \mu. \tag{23}$$

Note that $k_{\tilde{\mu}} = k_\mu$, because the network is managed to operate at the same point P on the MFD (Figure 7(a)) with and without transit operations. Expressions (22) and (23) describe the traffic state for cars when transit and cars are operating together on the network in the middle of the rush. This is shown in Figure 8 by the slope of the departure curve for cars exiting the network in the middle of the rush (segment $\overline{\text{BC}}$).

The procedure for identifying the system optimum is the same as described in Section 3. Conditional on the segment $\overline{\text{BC}}$, the total earliness and lateness are minimized by serving car trips at the maximum possible rate before $\tilde{t}_e$ and after $\tilde{t}_L$. Then, segment $\overline{\text{BC}}$ can be slid up or down until the sum of the schedule penalties for all early and late commuters is minimized.

For most of the rush, early and late commuters can be served at rate μ associated with point P, and vehicle accumulation n_μ. In the middle of the rush, when both transit and cars operate together on the street network without delay, commuters are served at $\tilde{\mu}$, and the total car accumulation is $\tilde{n}_\mu$. Therefore, just before transit service begins at $\tilde{t}_e$, the vehicle accumulation must be reduced to $\tilde{n}_\mu$ so the network exit rate μ_α will be:

$$\mu_\alpha = F(\tilde{n}_\mu). \tag{24}$$

This results in a shift of the traffic state to the left from P to P$_\alpha$ along $F(n)$ in Figure 7(b). If the duration of this transition is very short compared to the length of the rush, then the effect is small and the departure curve for cars in the network system optimum (see Figure 8) is approximately the same piece-wise linear pattern identified in Proposition 2.

In reality, there are two competing secondary effects of the transition from n_μ to $\tilde{n}_\mu$: increased total earliness, and travel time savings for faster trips. More details about the transition and these effects are presented in Appendix B. The transition effects do not occur when transit service ends because cars are able to freely enter the network and rapidly raise the vehicle accumulation to n_μ at $\tilde{t}_L$.

In order to eliminate queuing delay, the optimal prices presented in Proposition 4 are used so that users choose to travel at the system optimal departure time. To be consistent with the same bottleneck model, these prices should ideally be applied at the moment when

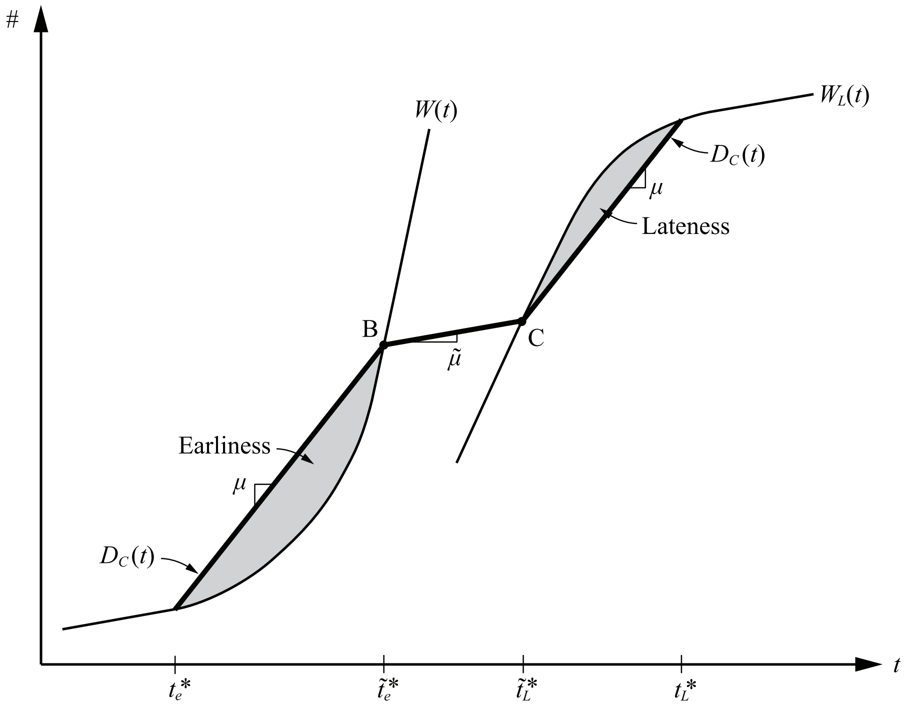


Figure 8: System optimal departure curve for cars exiting a network with transit operations.

cars leave the network. In equilibrium, the price for car drivers must increase at rate e for early drivers and decrease at rate L for late drivers; otherwise queuing delays will emerge as commuters seek to adjust their travel times in order to reduce their own experienced cost. In order to achieve the transition from P to P_α, some additional network control (e.g., adjusting signal timings) is required, but the optimal prices ensure that congestion does not develop. The time intervals $(t_e, \tilde{t}_e)$ and $(\tilde{t}_L, t_L)$ are determined by the target exit flow μ as shown in Figure 8. These determine the optimal car toll and transit fare as described in Section 4.

6 Conclusion

It has been shown that the provision of public transit is a Pareto improvement because everyone experiences a lower cost when transit is provided than if it is not. Public transit has also been shown to reduce the duration of the rush period in user equilibrium and the overall cost. When cars and transit share the same road capacity, the system optimum travel pattern can differ from the user equilibrium unless $\mu = \tilde{\mu}$ or $W(t)$ is Z-shaped. Optimal time-dependent prices always exist. For a Z-shaped $W(t)$, the optimal prices can be easily obtained. The optimal prices are unique up to an additive constant, so there is flexibility to pursue other policy objectives by choosing that constant.

The system optimum for a fixed capacity bottleneck has been shown to apply for networks which have state-dependent capacity. This can be done even accounting for the change in capacity to serve cars which results from dedicating some street space to transit operations. Therefore, the system optimum and optimal pricing strategy presented in Sections 3 and 4 also apply to multimodal urban networks.

The modeling of networks with multiple modes can be further improved by considering additional heterogeneity among users. This paper considered heterogeneity only in wished travel time, but commuters in real cities have varied values of time and lengths of trips which will also contribute to their choice of mode and departure time.

References

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

Arnott, R., De Palma, A., and Lindsey, R. (1990b). Economics of a bottleneck. *Journal of Urban Economics*, 27(1):111–130.

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

Daganzo, C. (2010). Structure of competitive transit networks. *Transportation Research Part B*, 44(4):434–446.

Daganzo, C. F. (1985). The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transportation Science*, 19(1):29–37.

Daganzo, C. F. (2007). Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B*, 41:49–62.

Danielis, R. and Marcucci, E. (2002). Bottleneck road congestion pricing with a competing railroad service. *Transportation Research Part E*, 38(5):379–388.

Eichler, M. and Daganzo, C. F. (2006). Bus lanes with intermittent priority: Strategy formulae and an evaluation. *Transportation Research Part B*, 40:731–744.

Geroliminis, N. and Daganzo, C. F. (2008). Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B*, 42(9):759–770.

Geroliminis, N. and Levinson, D. (2009). *Transportation and Traffic Theory*, chapter Cordon pricing consistent with the physics of overcrowding, pages 219–240. Springer Science.

Gonzales, E. J. (2011). *Allocation of space and the costs of multimodal transport in cities.* PhD thesis, University of California, Berkeley.

Hendrickson, C. and Kocur, G. (1981). Schedule delay and departure time decisions in a deterministic model. *Transportation Science*, 15(1):62–77.

Huang, H. (2000). Fares and tolls in a competitive system with transit and highway: The case with two groups of commuters. *Transportation Research Part E*, 36(4):267–284.

Koshi, M., Kuwahara, M., and Akahane, H. (1992). Capacity of sags and tunnels on japanese motorways. *ITE Journal*, 62(5):17–22.

Small, K. and Chu, X. (2003). Hypercongestion. *Journal of Transport Economics and Policy*, 37(1):319–352.

Smith, M. J. (1984). The existence of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transportation Science*, 18(4):385–394.

Tabuchi, T. (1993). Bottleneck congestion and modal split. *Journal of Urban Economics*, 34(3):414–431.

Vickrey, W. S. (1969). Congestion theory and transport investment. *The American Economic Review*, 59(2):251–260.

Wardrop, J. G. (1952). Some theoretical aspects of road traffic research. *ICE Proceedings: Engineering Divisions*, 1(3):325–378.

A System Optimum Necessary Condition

We use N_e^* and N_L^* to denote the values obtained with the construction of Figures 3 and 4 for the system optimum N_T^*. We will also define $\lambda_e^* \doteq \dot{W}(\tilde{t}_e^*)$ and $\lambda_L^* \doteq \dot{W}(\tilde{t}_L^*)$, which are the slopes of $W(t)$ at the system optimum points B and C, respectively.

Proposition 5. *At system optimum, the wished curves and departure curves for cars and transit are such that:*

$$\frac{N_e^*}{N_L^*} = \frac{L(\lambda_e^* - \tilde{\mu})(\lambda_L^* - \mu)}{e(\lambda_e^* - \mu)(\lambda_L^* - \tilde{\mu})}. \tag{25}$$

Proof. Figure 4 shows that the effect of an incremental shift of B up and to the right along $W(t)$ is associated with shifting segment $\overline{\mathrm{BC}}$ up by dn_o. This causes a upward shift of the departure curve for early car commuters by dn_e and for late car commuters by dn_L. Due to the geometry, these differentials are related by:

$$dn_o = dn_e \frac{\lambda_e - \tilde{\mu}}{\lambda_e - \mu} = dn_L \frac{\lambda_L - \tilde{\mu}}{\lambda_L - \mu}, \tag{26}$$

where $\lambda_e = \dot{W}(\tilde{t}_e)$ when the first on-time commuter departs the bottleneck and $\lambda_L = \dot{W}(\tilde{t}_L)$ when the last on-time commuter departs the bottleneck. At the system optimum, the schedule cost is minimized when the resulting change in total earliness balances the lateness:

$$\frac{eN_e}{\mu} dn_e = \frac{LN_L}{\mu} dn_L. \tag{27}$$

By manipulating (27) to express N_e/N_L in terms of dn_e and dn_L, then substituting expressions for these differentials from (26), it follows that the relative number of early and late commuters in system optimum is:

$$\frac{N_e^*}{N_L^*} = \frac{Ldn_L}{edn_e} = \frac{L(\lambda_e^* - \tilde{\mu})(\lambda_L^* - \mu)}{e(\lambda_e^* - \mu)(\lambda_L^* - \tilde{\mu})}, \tag{28}$$

which establishes (25). □

B Network System Optimum: Traffic State Transition

The Network Exit Function (NEF) describes the relationship between the number of cars in a network and the rate that vehicles exit the network as described in Section 5. In order to prevent delays for traffic when transit service begins, the vehicle accumulation in the network must be reduced to $\tilde{n}_\mu$ immediately before the start of transit service at $\tilde{t}_e$. This corresponds to a transition in the rate that cars discharge from the network from μ to μ_α as shown by points P and P_α in Figure 7(b). Since the NEF is concave, $\tilde{\mu} \leq \mu_\alpha$, and the slope from the origin to μ_α is no less than the slope to $\tilde{\mu}$. Therefore, before transit starts operating, vehicle trips are at least as fast as when both modes are operated together and

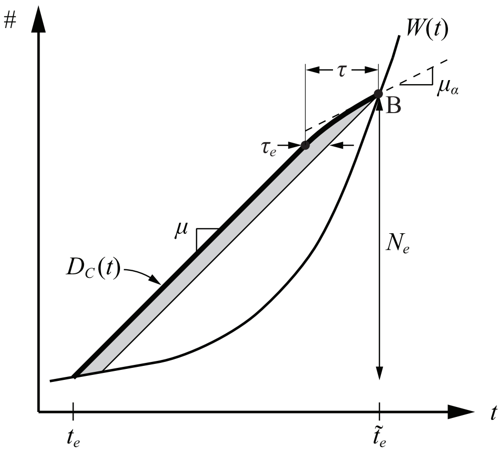


Figure 9: System optimal departure curve for early cars in a network transitioning from departure rate μ to μ_α.

no delays are incurred. The effect on vehicle departures is illustrated by $D_C(t)$ to the left of point B in Figure 9.

If the vehicle accumulation is expressed as a function of time, $n(t)$, then the state of the network follows the mass conservation equation (Daganzo, 2007):

$$\frac{dn}{dt} = I(t) - F(n(t)) \tag{29}$$

where $I(t)$ is the rate that vehicles enter the network. We define τ as the transition time for the vehicle accumulation to drop from n_μ to $\tilde{n}_\mu$. Then, τ is minimized if no vehicles enter the network ($I = 0$), and trips exit according to the NEF. The conservation equation (29) is an ordinary differential equation which can be solved as a boundary value problem to obtain τ:

$$\tau = -\int_{n_\mu}^{\tilde{n}_\mu} \frac{1}{F(n)} dn. \tag{30}$$

Recall from (22) that $\tilde{n}_\mu = \alpha n_\mu$.

The transition from n_μ to $\tilde{n}_\mu$ causes two competing effects. First, the total earliness cost is increased because the maximum departure rate for early commuters cannot be sustained at μ for the entire interval $(t_e, \tilde{t}_e)$. The reduced exit flow immediately preceding transit service adds τ_e additional earliness to nearly every early commuter (see Figure 9). This is the difference between the transition time, and the time it would have taken for the same $(1-\alpha)n_\mu$ trips to exit at rate μ:

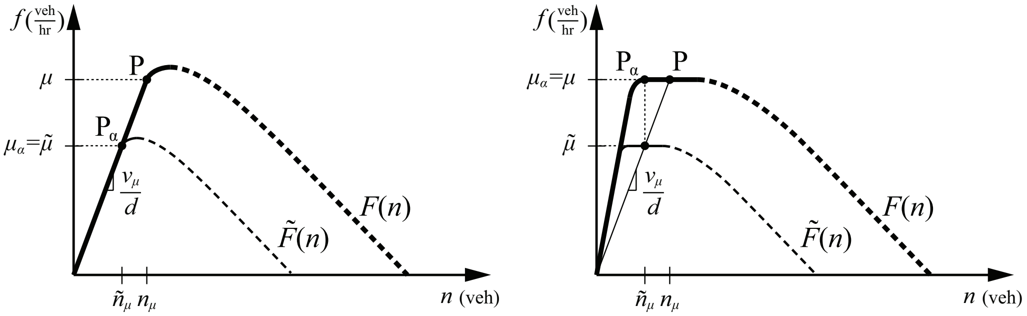
$$\tau_e = \tau - \frac{(1-\alpha)n_\mu}{\mu}. \tag{31}$$

Since nearly every early driver experiences τ_e additional earliness, the total system cost is increased by approximately $eN_e\tau_e$.

Second, some travel time savings are experienced by early commuters in the transition period of length τ which reduces the total system cost. This occurs because the transition from point P to P$_\alpha$ decreases the exit flow to μ_α. Since $\tilde{\mu} \le \mu_\alpha$ and both flows are associated with $\tilde{n}_\mu$, the travel time will be at least as short for P$_\alpha$ as P, if not shorter. The aggregated travel time savings, TT_s, is the difference between the total travel time when $(1-\alpha)n_\mu$ trips exit while the network accumulation is n_μ and the total travel time during the transition period when the same number of trips actually exit:

$$TT_s = \frac{(1-\alpha)n_\mu^2}{\mu} - \int_0^\tau n(t)dt \tag{32}$$

The first term of (32) is the product of the time it takes $(1-\alpha)n_\mu$ to exit the network at rate μ and the n_μ vehicles which are in the network at all times. Upper bounds for the magnitude of these effects can be determined by considering two NEFs: with $\mu_\alpha = \tilde{\mu}$, and with $\mu_\alpha = \mu$ (see Figure 10).



(a) Case 1: NEF when $\mu_\alpha = \tilde{\mu}$

(b) Case 2: NEF when $\mu_\alpha = \mu$

Figure 10: Example NEFs for cases with maximum earliness cost (Case 1) and maximum travel time savings (Case 2).

Case 1: Maximum Earliness The largest possible change in exit flow from P to P$_\alpha$ is a transition from μ to $\mu_\alpha = \tilde{\mu}$. In this case, NEF must be linear to the left of P as shown in Figure 10(a). This will result in the greatest possible transition time τ and additional earliness τ_e, because the exit rate can be no lower for each vehicle accumulation if $F(n)$ is concave. For this case, the exit flow is given by:

$$F_1(n) = \frac{\mu}{n_\mu} n \quad \text{for } n \in (0, n_\mu). \tag{33}$$

We can solve for τ by substituting (33) into (30), and solving the integral with $\tilde{n}_\mu = \alpha n_\mu$:

$$\tau = -\frac{n_\mu}{\mu} \ln \alpha. \tag{34}$$

Then, by substituting (34) into (31) and collecting terms, the added earliness for each early commuter is:

$$\tau_e = \frac{n_\mu}{\mu}\left(-\ln\alpha - 1 + \alpha\right). \tag{35}$$

This is an upper bound for the τ_e associated with any concave NEF. Note that n_μ/μ is the average travel time for a trip of length d, and τ_e will be small for many reasonable values of α (e.g., τ_e is less than 3% of the uncongested travel time for values of $\alpha > 0.8$).

The NEF in this case is linear to the left of P, so all traffic states are associated with the same slope to the origin for $n \leq n_\mu$ (see Figure 10(a)). The average travel time per trip in the network does not change over the course of the transition, and therefore there are no travel time savings experienced. This can be verified by solving (29) with (33) and substituting the result into (32).

Case 2: Maximum Travel Time Savings The largest possible reduction in travel time from P to P$_\alpha$ is when the exit flow transitions from μ to $\mu_\alpha = \mu$. In this case, the NEF has a constant value between P$_\alpha$ and P as shown in Figure 10(b). Although we would expect the point P always to be chosen as the left most point with exit flow μ, this case provides an upper bound for the total travel time savings as μ_α approaches μ.

Since the exit flow is always μ, the number of vehicles in the network at any time during the transition is given by:

$$n(t) = n_\mu - \mu t. \tag{36}$$

We also know that duration of the transition for $(1-\alpha)n_\mu$ vehicles to depart will be:

$$\tau = (1-\alpha)n_\mu/\mu. \tag{37}$$

Substituting (36) and (37) into (32), and solving the integral, the total travel time savings is:

$$TT_s = \frac{(1-\alpha)^2 n_\mu^2}{2\mu}. \tag{38}$$

This is an upper bound for the TT_s associated with any choice of P on the uncongested side of a concave NEF. Note that this value is independent of the number of early commuters as long as the length of the period when commuters travel early is longer than the transition period.

The NEF in this case does not contribute any additional earliness to the other early commuters ($\tau_e = 0$). This result can be easily verified by substituting (37) into (31) and occurs because the exit flow is always maintained at μ until transit service starts. Therefore, the departure curve for cars in this case is still represented by Figure 8, and the system optimal solution will be exactly the same as the travel pattern identifies in Proposition 3.