Simple vs. Optimal Congestion Pricing

Devansh Jalota
Columbia
dj2757@columbia.edu

Xuan Di
Columbia
sharon.di@columbia.edu

Adam N. Elmachtoub
Columbia
adam@ieor.columbia.edu

March 8, 2026

Abstract

Congestion pricing has emerged as an effective tool for mitigating traffic congestion, yet implementing welfare or revenue-optimal dynamic tolls is often impractical. Most real-world congestion pricing deployments, including New York City's recent program, rely on significantly simpler, often static, tolls. This discrepancy motivates the question of how much revenue and welfare loss there is when real-world traffic systems use static rather than optimal dynamic pricing.

We address this question by analyzing the performance gap between static (simple) and dynamic (optimal) congestion pricing schemes in two canonical frameworks: Vickrey's bottleneck model with a public transit outside option and its city-scale extension based on the Macroscopic Fundamental Diagram (MFD). In both models, we first characterize the revenue-optimal static and dynamic tolling policies, which have received limited attention in prior work. In the worst-case, revenue-optimal static tolls achieve at least half of the dynamic optimal revenue and at most twice the minimum achievable system cost across a wide range of practically relevant parameter regimes, with stronger and more general guarantees in the bottleneck model than in the MFD model. We further corroborate our theoretical guarantees with numerical results based on real-world datasets from the San Francisco Bay Area and New York City, which demonstrate that static tolls achieve roughly 80-90% of the dynamic optimal revenue while incurring at most a 8-20% higher total system cost than the minimum achievable system cost.

1 Introduction

Traffic congestion has surged in major cities worldwide, straining infrastructure, degrading air quality, and imposing substantial economic costs (Fleming 2019, Fernandez 2025). Against this backdrop, congestion pricing has emerged as an effective tool for mitigating the inefficiencies of traffic congestion and empirical evidence from recent large-scale deployments underscores its promise. For instance, New York City's newly implemented congestion pricing program in January 2025 has substantially reduced vehicle volumes, with the potential to generate billions of dollars in revenue for transit investments (Cook et al. 2025, Ostrovsky and Yang 2024). Similar successes have been documented in Stockholm and Singapore (Hu et al. 2023).

While many large-scale congestion pricing deployments, including those in London (2003), Stockholm (2007), and New York City (2025), have emerged in the past two decades, their intellectual roots date back to Pigou (1920). Under Pigou's framework, the socially optimal toll charges each traveler the marginal cost they impose on others in the network. However, implementing Pigouvian tolls in practice is challenging and often impractical, as it requires time-varying, state-dependent tolls that respond to evolving traffic conditions, potentially at the granularity of individual network links. Such fine-grained dynamic pricing demands sophisticated sensing, communication, and computation infrastructure, while placing substantial informational and cognitive burdens on travelers. Consequently, despite its theoretical appeal, fully dynamic marginal-cost or Pigouvian tolling remains largely infeasible in real-world congestion pricing deployments.

Instead, most operational congestion pricing systems rely on much simpler, often static, tolling structures. For instance, fixed (static) charges are levied on specific bottlenecks, bridges, or crossings, such as the San Francisco (SF)–Oakland Bay Bridge in the SF Bay Area (Gonzales and Christofa 2015). Likewise, large

urban cordon or area-based systems, such as those in London and New York City, typically impose a fixed (static) entry or usage fee for designated congestion zones during peak periods. Such static tolling schemes are appealing as they are transparent, easy to communicate to users, and require less real-time information and computational infrastructure than theoretically optimal dynamic pricing.

Yet, the simplicity of static tolling means that it may deviate substantially from idealized marginal-cost pricing, raising a central question: *How much performance is sacrificed when real-world transportation systems use simple static pricing instead of optimal dynamic congestion pricing?* Understanding this gap is crucial for policymakers as cities seek congestion pricing schemes that are operationally feasible and deliver substantial welfare and revenue gains. This question is also increasingly relevant for transportation authorities exploring more dynamic approaches, such as managed express lanes that adjust tolls based on real-time conditions (Pandey and Boyles 2018, Jang et al. 2014), or Singapore's ERP 2.0 system, which aims to enable real-time, link-level pricing across its road network (of Transport 2025).

This work addresses the above question by analyzing the performance gap in terms of revenue and welfare (or system cost) between simple static and optimal dynamic congestion pricing schemes in two canonical modeling frameworks that capture the dominant real-world use cases of congestion pricing. In particular, we study (i) Vickrey's bottleneck model (Vickrey 1969), a foundational and analytically tractable representation of peak-period congestion that captures queue formation and dissipation at a bottleneck, and (ii) its city-scale urban system extension based on the Macroscopic Fundamental Diagram (MFD) (Daganzo 2007), which incorporates spatially distributed origins and destinations and models congestion through a state-dependent capacity, rather than a fixed bottleneck capacity. We focus on these models as they provide clean, tractable frameworks in which to isolate the core mechanisms governing static versus dynamic tolling performance, while also reflecting the structure of contemporary deployments, from bottleneck-priced corridors such as the SF–Oakland Bay Bridge to city-scale congestion pricing systems in New York City and London.

In studying these models, we focus on revenue-maximizing tolls, which have received limited attention in the literature that has largely emphasized welfare-optimal or system-cost-optimal tolls. Beyond filling this research gap, our focus on revenue maximization is motivated by several practical considerations. First, revenue generation is an explicit policy objective of many congestion pricing programs (e.g., New York City), where toll revenues are earmarked for transit investments (MTA 2025). Next, revenue objectives are central to public–private toll road partnerships, including tolled road networks operated by firms such as Transurban (O'Neill 2022). Finally, our numerical experiments in Section 6 demonstrate that static revenue-optimal tolls often coincide, or perform nearly identically to, system-cost-optimal static tolls across empirically relevant parameter regimes, reinforcing our focus on revenue-optimal tolling. Overall, while a key focus of this work is to study revenue-optimal tolling policies, we also evaluate their welfare (system cost) implications, providing a comprehensive assessment of performance across the metrics most relevant to policymakers.

Our Contributions: This work analyzes the performance of static versus dynamic congestion pricing in two canonical settings: Vickrey's bottleneck model with a public transit outside option (e.g., a subway) and its MFD-based urban system extension. In doing so, a central contribution is the characterization of revenue-optimal static and dynamic tolls in both models. For the bottleneck model with an outside option, we leverage the equilibrium properties of Vickrey's framework to show that computing revenue-optimal static and dynamic tolls reduces to solving *single-variable quadratic programs* that admit closed-form solutions. For urban systems, such closed-form results are typically intractable for arbitrary MFD relations; thus, we focus on the widely studied and empirically supported triangular fundamental diagram (see Section 5 for more details). We characterize the revenue-optimal dynamic toll, showing that it maintains the system at the throughput-maximizing capacity, and derive a closed-form expression for the revenue generated by any static toll.

Leveraging these characterizations, we compare static revenue-optimal tolling with its optimal dynamic counterparts on both revenue and total system cost metrics (see Section 3.3 for details), which are of direct relevance to policymakers. Our results establish lower bounds on the revenue and upper bounds on the total system costs achievable under static-revenue optimal tolls relative to the dynamic benchmarks. In Vickrey's bottleneck model with an outside option, we show that static revenue-optimal tolls obtain *at least half* of

the revenue achieved by dynamic revenue-optimal tolls across all parameter regimes and *at least two-thirds* in many practically relevant regimes. Static revenue-optimal tolls do not always guarantee constant-factor approximations for system cost. However, when the public transit outside option is not substantially less attractive than driving a car, a condition that holds in many real-world settings, static revenue-optimal tolls incur *at most twice* the optimal system cost. In the MFD framework, the guarantees we obtain are weaker (as the capacity is state-dependent) but still robust: across a range of practically relevant parameter regimes, static revenue-optimal tolls still achieve *at least two-thirds* of the dynamic optimal revenue and incur *at most twice* the minimal system cost.

Finally, we complement our theory with numerical experiments based on two real-world case studies: the SF–Oakland Bay Bridge for the bottleneck model and New York City's congestion reduction zone for the MFD framework. Our results show that static revenue-optimal tolls achieve roughly 80-90% of the dynamic optimal revenue while incurring at most a 8-20% higher total system cost than the minimum achievable for practically relevant parameter regimes, though this gap can widen when the public transit option is significantly less attractive (i.e., more costly), consistent with theory. Moreover, when evaluated through the lens of our model, we find that the static tolls currently implemented in practice in both case studies correspond to parameter regimes in which static revenue-optimal tolls capture nearly all the benefits of dynamic tolling, achieving roughly 98% of the dynamic optimal revenue and a 3% higher total system cost than the minimum achievable. Beyond the attractiveness of public transit relative to driving, a key determinant of the performance of static revenue-optimal tolling is the ratio of the maximum system throughput to the desired user arrival rate. When this ratio is close to one, as in the Bay Bridge study, static revenue-optimal tolls capture a larger fraction of the dynamic optimal revenue but incur a higher multiplicative gap relative to the minimum achievable system cost. This is in comparison to the New York City study, where this ratio is substantially lower due to its much higher public transit mode share. Thus, in addition to highlighting the efficacy of static revenue-optimal tolls, our numerical results illustrate a fundamental trade-off between revenue and system cost objectives under static tolling.

Overall, our results demonstrate that simple static tolls can deliver robust performance on both revenue and system cost metrics, underscoring its practical efficacy and shedding light on the strong empirical performance of the many static congestion pricing systems already in operation. Moreover, since our comparisons abstract from the substantial informational and computational burdens that dynamic tolling places on both transportation planners and users, our theoretical bounds should be interpreted as worst-case guarantees. Incorporating realistic behavioral or operational constraints would only further strengthen the case for static tolling in real-world deployment.

Organization: The remainder of this paper is organized as follows. Section 2 reviews related literature. Section 3 introduces our model and reviews equilibrium outcomes in Vickrey's bottleneck model with and without an outside option, which form the foundation of our analysis. Then, Section 4 derives the revenue-optimal static and dynamic tolling policies in Vickrey's bottleneck model with an outside option and compares the performance of revenue-optimal static tolls to its dynamic benchmarks. Section 5 extends these results to urban systems based on the MFD. Section 6 presents numerical experiments. Section 7 concludes and provides directions for future work.

2 Related Literature

The study of peak-period congestion originates from Vickrey's bottleneck model (Vickrey 1969), which characterizes the equilibrium departure time choices of users traversing a bottleneck with a fixed capacity. This framework has served as the foundation for an extensive literature, with numerous extensions including models of parallel-route choice (Arnott et al. 1990a), modal split (Tabuchi 1993), heterogeneous preferences (Newell 1987, Lindsey 2004), carpooling (Xiao et al. 2016, Ostrovsky and Schwarz 2025), stochastic demand (Arnott et al. 1999, de Palma et al. 1983), and bounded rationality (Mahmassani and Chang 1987), among others. While the bottleneck model and its extensions capture queuing delays as a point queue, they do not account for hypercongestion, wherein system throughput declines once road density exceeds a critical threshold. This limitation has motivated city-scale extensions, including bathtub models of downtown

traffic (Arnott 2013) and queuing formulations (Fosgerau and Small 2013) in which the bottleneck capacity is state-dependent and degrades under heavy congestion. Relatedly, the MFD (Daganzo 2007, Geroliminis and Daganzo 2008) provides a parsimonious representation of the relationship between system-wide vehicle density and throughput. For a comprehensive survey of the bottleneck model and its extensions, see Li et al. (2020).

Building on these foundations, a substantial literature has examined congestion pricing as a mechanism for mitigating traffic, with a focus on system-cost-optimal tolls that internalize congestion externalities (Pigou 1920). In the bottleneck model, the system-cost-optimal toll is time-varying, requiring continuously adjustable charges (Vickrey 1969, Arnott et al. 1990b, Newell 1987). While such dynamic pricing is theoretically efficient, it is operationally complex, motivating the design of more implementable schemes, including uniform, stepwise, and coarse time-of-day tolls (Chu 1999, Arnott et al. 1993a, Braid 1989). For example, Laih (1994) showed that an optimal n-step toll can eliminate at most $\frac{n}{n+1}$ of the queuing delay compared to the time-varying optimal toll. Related studies extend these ideas beyond a single bottleneck to parallel networks (Arnott et al. 1990a, Braid 1996). Like these works, we compare simple static tolls with (optimal) dynamic congestion pricing. However, we consider a model with an outside option (e.g., public transit), which substantially alters equilibrium behavior, and evaluate tolling policies under a revenue-maximization objective.

In this regard, our work connects to prior studies that incorporate an outside option or elastic demand into Vickrey's bottleneck model (Arnott et al. 1993b, Gonzales and Daganzo 2012), and to the literature on revenue-maximizing tolls for privately operated facilities (De Palma and Lindsey 2000, de Palma and Lindsey 2002, Fu et al. 2018). Existing revenue-maximization studies in the bottleneck framework primarily examine equilibrium outcomes arising from competition among firms, and work incorporating outside options predominantly focus on devising system-cost-optimal tolling policies. In contrast, we study the design of revenue-maximizing tolling policies and provide explicit guarantees that quantify the performance gap between static and dynamic tolling.

Beyond the tolling literature rooted in Vickrey's bottleneck framework, our work also connects to the broader work on congestion pricing in settings where first-best Pigouvian pricing is infeasible due to policy, behavioral, or infrastructural constraints. This includes research on second-best congestion pricing, which examines toll design when only a subset of network links can be priced (Verhoef 2002, Labbé et al. 1998, Larsson and Patriksson 1998, Patriksson and Rockafellar 2002, Di et al. 2016), and optimal cordon pricing, which uses zone-based tolls to manage congestion in urban systems (il Mun et al. 2003, Zhang and Yang 2004). Like these works, we study simple, implementable tolling policies; however, unlike the large-scale bi-level or mixed-integer optimization models used in these works, we derive closed-form optimal static and dynamic tolls in both the bottleneck model and its MFD extension, yielding transparent structural insights and theoretical guarantees.

Our work also connects to the literature on simplicity versus optimality in algorithm and mechanism design (Hartline and Roughgarden 2009, Hart and Nisan 2017), including studies of static pricing in revenue management (Elmachtoub and Shi 2025, Besbes et al. 2022) and online platforms (Banerjee et al. 2015), which show that simple, time-invariant prices can achieve strong performance relative to their optimal dynamic pricing counterparts. In this spirit, we also show that simple static tolls can achieve performance guarantees relative to optimal dynamic tolling.

Finally, since we compare tolling policies in terms of revenue and welfare (or system cost), our work relates to the broader economics literature on the tension between revenue and welfare-maximizing pricing rules. This trade-off is well known in auction theory, most notably in the contrast between welfare-maximizing Vickrey auctions (Vickrey 1961) and Myerson's revenue-maximizing mechanisms (Myerson 1981), and has also been studied in settings including security games (Jalota et al. 2024), congestion games (Zhang et al. 2024), and revenue management (Chen and Gallego 2019, Zhang and Dong 2025). Contributing to this literature, we analyze revenue-welfare tradeoffs in a congestion pricing context and quantify how much revenue can be retained and how much welfare is lost when moving from optimal dynamic pricing to simple static tolls.

3 Model and Background: Bottleneck with Outside Option

This section presents Vickrey's bottleneck model with an outside option (Section 3.1), reviews its untolled equilibrium outcome established in Gonzales and Daganzo (2012) (Section 3.2), and defines the revenue and system cost metrics to evaluate the tolling policies we study (Section 3.3).

3.1 Setup

We study Vickrey's bottleneck model (Vickrey 1969), a canonical representation of peak-period congestion in capacity-constrained facilities such as bridges and tunnels, augmented with a public transit outside option. In this framework, a mass of Λ users makes morning commute trips using either (i) a car, which requires traversing a bottleneck of capacity (service rate) μ, or (ii) a public transit alternative (e.g., a subway). As in Vickrey (1969), users' desired bottleneck crossing times are uniformly distributed over the time interval $[t_1, t_2]$ (e.g., morning rush). In the absence of the bottleneck, users would arrive at their destination at their desired times, with a corresponding desired arrival rate $\lambda = \frac{\Lambda}{t_2 - t_1}$. Unlike stochastic queuing models, e.g., based on Poisson processes, we clarify that arrivals and service are deterministic. In addition to bottleneck congestion that may induce deviations from users' desired bottleneck crossing times and influence mode choices, users' travel decisions can also be influenced via a (possibly time-varying) toll $\tilde{\tau}(t)$ on cars crossing the bottleneck at time t.

The key distinction between car and public transit is how they influence users' travel costs. For car users, the travel cost depends on their desired bottleneck crossing time t^*, the prevailing level of congestion, and tolls. Normalizing the queuing delay at the bottleneck to zero under free-flow conditions (i.e., in the absence of congestion), the cost of a car trip comprises four components: (i) a congestion independent free-flow generalized cost $\tilde{z}_C$, (ii) queuing (or waiting) delay $w(t)$ to traverse the bottleneck due to congestion, (iii) schedule delay $|t^* - t|$, denoting the deviation between a user's desired and actual bottleneck crossing time, and (iv) the toll $\tilde{\tau}(t)$. Let c_W denote the penalty for incurring a unit of waiting time delay, and c_e and c_L be the schedule delay penalties for arriving early or late, respectively, which is assumed to be identical across users (Vickrey 1969, Ostrovsky and Schwarz 2025), with $0 < c_e < c_W$ and $c_L > 0$, as is standard in the literature (Vickrey 1969). Then, the total travel cost for a car user with a desired bottleneck crossing time t^* who crosses the bottleneck at time t is:

$$\tilde{c}(t, t^*) = \tilde{z}_C + c_W w(t) + c_e(t^* - t)_+ + c_L(t - t^*)_+ + \tilde{\tau}(t). \tag{1}$$

Unlike car users, we assume that transit users incur a fixed generalized cost $\tilde{z}_T$, as is standard in the literature (Gonzales and Daganzo 2012). This specification is reflective of settings in which the public transit system charges a fixed fare, operates at fixed headways, and is segregated from the road network, features typical of metro or subway systems (e.g., New York City's subway), and thus does not experience congestion delays. While richer models could also allow the transit cost to depend on ridership (e.g., to capture crowding or discomfort effects), a fixed transit cost enables analytical tractability, allowing us to isolate the core mechanisms through which static and dynamic pricing influence user behavior and system outcomes. Nevertheless, relaxing this assumption to incorporate crowding effects or other dependencies of the transit cost on ridership is a valuable direction for future work.

In the remainder of this work, for analytical clarity, we normalize user costs by the waiting time penalty c_W. Then, the normalized transit cost is $z_T = \frac{\tilde{z}_T}{c_W}$, and letting $z_C = \frac{\tilde{z}_C}{c_W}$, $\tau(t) = \frac{\tilde{\tau}(t)}{c_W}$, $\frac{c_e}{c_W} = e$, and $\frac{c_L}{c_W} = L$, the normalized cost for a car user from Equation (1) is given by:

$$c(t, t^*) = z_C + w(t) + e(t^* - t)_+ + L(t - t^*)_+ + \tau(t). \tag{2}$$

3.2 Equilibria in the Untolled Bottleneck Model

Under the setup and user costs defined in the previous section, an *equilibrium* arises when all users minimize travel costs by choosing a mode (car or transit) and, if traveling by car, a bottleneck crossing time, resulting in a pattern of departure times, mode choices, and induced congestion effects under which no user has an incentive to deviate (Vickrey 1969, Hendrickson and Kocur 1981). Given cost-minimizing user behavior, we now review well-established equilibrium characterization results of the bottleneck model in the untolled

setting ($\tau(t) = 0$ for all t), both without and with a public transit outside option, which forms the foundation for our analysis of static and dynamic tolling schemes in Sections 4 and 5.

Bottleneck without Outside Option: To build intuition, we first consider the setting without an outside option, in which all users travel by car and choose only their bottleneck crossing time t. Specifically, a user with a desired crossing time t^* selects a time t to pass the bottleneck to minimize their travel cost $c(t, t^*)$ in Equation (2). If the bottleneck service rate is at least the desired arrival rate, i.e., $\mu \geq \lambda$, no queue forms at equilibrium and all users cross the bottleneck at their desired times, incurring a cost z_C. When $\mu < \lambda$, as is typical during morning and evening rush periods, the bottleneck cannot serve all Λ users during the interval $[t_1, t_2]$ and congestion delays arise.

In this congested regime, users incur waiting and schedule delay costs, and the Λ users are served at the bottleneck capacity μ over an interval $[t_A^{(No)}, t_D^{(No)}]$, where $t_D^{(No)} - t_A^{(No)} = \frac{\Lambda}{\mu} > t_2 - t_1$. Here, the superscript *No* denotes the setting with *no* outside option. Over this interval, the equilibrium waiting-time function $w(t)$ is determined by the first-order optimality condition for cost-minimizing users, $c'(t, t^*) = 0$, under which no user can reduce their travel cost through a marginal adjustment in their bottleneck crossing time. This condition implies that the slope of the waiting-time function satisfies $w'(t) = e$ for early arrivals and $w'(t) = -L$ for late arrivals, which is independent of users' desired crossing times t^*, thus characterizing the equilibrium condition for all users.

To characterize the waiting time profile from this differential equation, note that the user with the maximum waiting time at equilibrium must cross the bottleneck at their desired time; otherwise, a marginal shift in their departure time would reduce schedule delay without increasing waiting time, thus lowering their travel cost. Let $t^* = \tilde{t}$ denote this user's desired crossing time. Then, the waiting time profile is triangular with $w'(t) = e$ for $t < \tilde{t}$ and $w'(t) = -L$ for $t > \tilde{t}$, as shown in Figure 1 (left). The critical user with a desired crossing time $\tilde{t}$ incurs the maximum travel cost, and experiences the maximum waiting time $T_C = \frac{\Lambda e L}{\mu(e+L)}$. For more details on bottleneck equilibria and the resulting equilibrium bottleneck arrival and departure (crossing) time profiles of users, see Vickrey (1969), Hendrickson and Kocur (1981) and Figure 7 in Appendix A.1.

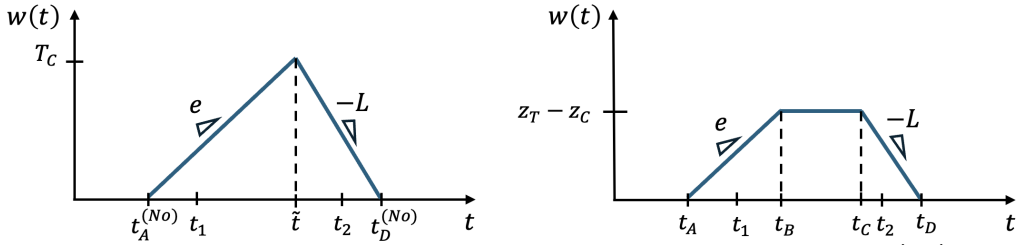


Figure 1: Equilibrium waiting time profiles in the bottleneck model without an outside option (left) and with a public transit outside option in a mixed-mode equilibrium where both car and transit are used (right) in the setting when $\mu < \lambda$. The slopes e and L denote the normalized schedule-delay penalties for early and late arrivals, respectively, and $[t_1, t_2]$ represents users' desired bottleneck crossing times.

Bottleneck with Outside Option: When a public transit outside option is available, users choose not only when to travel but also whether to travel by car or transit. While car users may incur congestion delays, transit users (e.g., subway riders) face a fixed cost z_T and can arrive at their destination at their desired time; hence, they do not face a departure-time decision. Then, equilibrium mode choice and, for car users, departure-time decisions are determined by the relative magnitudes of the costs z_T, z_C, and the maximum waiting time T_C in the car-only bottleneck. This yields three parameter regimes, corresponding to different equilibrium mode-use patterns, summarized below.

Case 1 ($z_T < z_C$): All users take transit, as it is strictly preferred to a car trip at free-flow.

Case 2 ($z_T \geq z_C + T_C$): Transit is never cost-effective relative to car travel, even for the user facing the highest travel cost with a wait time of T_C in the car-only bottleneck. Thus, all users travel by car, and the equilibrium coincides with the bottleneck model without an outside option.

Case 3 ($z_T \in [z_C, z_C + T_C]$): In this intermediate regime, a mixed-mode equilibrium arises in which both car and transit are used, as characterized by the following proposition from Gonzales and Daganzo (2012).

Proposition 1 (Two-Mode Equilibrium (Gonzales and Daganzo 2012))**.** *Suppose users can choose between two modes, traveling by car through a bottleneck with a free-flow cost z_C, and using public transit with a fixed cost z_T, where $z_T \geq z_C$. Letting $z_T - z_C < T_C$ and assuming users pass the bottleneck in order of their desired bottleneck departure times, there exists an equilibrium such that (see right of Figure 1):*

1. *The number of early car users $N_e = \frac{\mu(z_T - z_C)}{e}$, who travel at the start of the rush between $[t_A, t_B]$. The number of late car users $N_L = \frac{\mu(z_T - z_C)}{L}$, who travel at the end of the rush between $[t_C, t_D]$.*
2. *The number of on time car users and transit riders $N_o(\cdot)$, $N_T(\cdot)$ are strictly decreasing functions of the cost difference $z_T - z_C$ and they travel in the middle of the rush between $[t_B, t_C]$.*

Proposition 1 characterizes the equilibrium structure in the regime $z_T \in [z_C, z_C + T_C]$, which is arguably the most practically relevant case: while public transit is typically less attractive than driving under free-flow conditions, it is often not sufficiently worse for all users to travel by car, so that both modes are used at equilibrium. In this regime, Proposition 1 shows that the waiting time profile for car users is trapezoidal (see right of Figure 1), which arises as the queuing delays at the car bottleneck can never exceed $z_T - z_C$, since any such user would prefer to switch to transit. We also note that the interval $[t_A, t_D]$ in Proposition 1 need not coincide with, and is generally a subset of, the corresponding interval $[t_A^{(\mathrm{No})}, t_D^{(\mathrm{No})}]$ in the car-only bottleneck. For more details on equilibria in the bottleneck model with an outside option and the resulting equilibrium bottleneck arrival and departure (crossing) time profiles of car users, see Gonzales and Daganzo (2012) and Figure 8 in Appendix A.1.

3.3 Performance Metrics to Evaluate Equilibria Induced by Tolling Policies

Having presented the untolled equilibrium outcomes in the bottleneck model with and without an outside option, we now present the metrics to assess the equilibria induced by a (possibly time-varying) tolling policy $\tau(\cdot)$. We focus on revenue and total system cost metrics, both of which we normalize by the waiting time penalty c_W. The revenue of a tolling policy, denoted $R(\tau(\cdot))$, is defined as the total toll payments collected from car users at the resulting equilibrium. The total system cost, denoted $SC(\tau(\cdot))$, measures the aggregate travel burden borne by users at equilibrium and is defined as the sum of waiting time costs, schedule delay costs, and generalized costs (i.e., z_T or z_C) associated with car or transit use. Equivalently, the total system cost is the sum of users' total travel costs at equilibrium *minus* their toll payments. Consistent with standard practice in congestion pricing (Vickrey 1969, Gonzales and Daganzo 2012) and economics (Vickrey 1961) that excludes tolls (or prices) when defining system cost or welfare, toll payments are excluded from total system cost, as they represent transfers between users and the planner. For formal mathematical definitions of both metrics, see Appendix B.

4 Static vs. Dynamic Tolling in the Bottleneck Model

This section begins our performance comparison between static revenue-optimal tolling and its revenue and system-cost-optimal dynamic tolling counterparts in the bottleneck model with an outside option. In this setting, we first derive closed-form expressions for revenue-optimal static and dynamic tolls in Sections 4.1 and 4.2, respectively. Leveraging these characterizations, we then quantify the performance gap between static revenue-optimal tolls and its dynamic optimal benchmarks by establishing lower bounds on the fraction of optimal revenue attained and multiplicative upper bounds on total system cost in Sections 4.3 and 4.4.

We focus on the non-trivial and empirically relevant regimes when $\mu < \lambda$, i.e., the congested regime when the bottleneck service rate is strictly below the desired bottleneck departure rate, and $z_T \geq z_C$, i.e., car travel under free-flow is preferred to transit. Note that when $\mu \geq \lambda$, congestion does not arise and the static and dynamic optimal tolls on both revenue and system cost metrics coincide, corresponding to a uniform toll of $\max\{z_T - z_C, 0\}$. Moreover, when $z_T < z_C$, all users strictly prefer transit under any tolling policy, yielding no revenue and a total system cost of $z_T \Lambda$.

4.1 Static Revenue-Optimal Tolls

This section characterizes the revenue-optimal static toll τ in the bottleneck model with an outside option, where the toll is constant over time (i.e., $\tau(t) = \tau$ for all t). Under a static toll τ, let $N_T(\tau)$ denote the number of users that choose transit at equilibrium. The toll revenue is given by $R(\tau) = \tau(\Lambda - N_T(\tau))$. Proposition 2 derives an expression for the revenue-optimal static toll τ_s^*, showing that it depends on the magnitude of the cost difference between transit and car travel at free-flow, given by $z_T - z_C$. When this difference is below a threshold, revenue is maximized by setting the toll at its highest feasible level, $z_T - z_C$, beyond which all users would switch to transit. In contrast, when this difference is larger than that threshold, the revenue-optimal static toll is strictly lower and depends on the remaining model parameters, as characterized in Proposition 2.

Proposition 2 (Revenue-Optimal Static Tolls)**.** *Suppose $\mu < \lambda$ and users choose between two modes, traveling by car through a bottleneck with a free flow cost of z_C, and using a public transit alternative with a fixed cost of z_T, where $z_T \geq z_C$. Then, the revenue-optimal static toll is given by:*

$$\tau^* = \begin{cases} \max\left\{\frac{z_T - z_C}{2} + \frac{\Lambda e L}{2(\lambda-\mu)(e+L)}, z_T - z_C - \frac{\Lambda e L}{\mu(e+L)}\right\}, & \text{if } z_T - z_C \geq \frac{\Lambda e L}{(\lambda-\mu)(e+L)} \\ z_T - z_C, & \text{if } 0 \leq z_T - z_C < \frac{\Lambda e L}{(\lambda-\mu)(e+L)}. \end{cases} \tag{3}$$

Proof. First note that a static toll $\tau \geq 0$ must be such that $z_T - z_C - \tau \leq T_C$. Note that if $z_T - z_C - \tau > T_C$, then τ can be increased to $z_T - z_C - T_C$ without changing the number of car users, resulting in a higher revenue. Next, for any $\tau \geq \max\{0, z_T - z_C - T_C\}$, note that the equilibrium is as specified in Proposition 1 other than that the maximum waiting time is $\bar{w} = z_T - z_C - \tau$.

Next, we derive an expression for the number of users traveling by transit under the toll τ. By the linearity of the waiting time profile, all car users that pass the bottleneck either early or late (i.e., between $[t_A, t_B]$ and $[t_C, t_D]$, respectively, on the right of Figure 1) correspond to a $\frac{\bar{w}}{T_C}$ fraction of all users. The remaining $1 - \frac{\bar{w}}{T_C}$ fraction arrive exactly on time (either using transit or car), where the total on-time car users is $N_o(\tau) = (1 - \frac{\bar{w}}{T_C})(t_2 - t_1)\mu = (1 - \frac{\bar{w}}{T_C})\frac{\Lambda}{\lambda}\mu$ and transit users is $N_T(\tau) = (1 - \frac{\bar{w}}{T_C})\Lambda\left(1 - \frac{\mu}{\lambda}\right)$. We then obtain the following expression for the revenue as a function of τ: $R(\tau) = \tau\left[\Lambda - N_T(\tau)\right] = \tau\left[\Lambda - \left(1 - \frac{z_T - z_C - \tau}{T_C}\right)\Lambda\left(1 - \frac{\mu}{\lambda}\right)\right] \overset{(a)}{=} \mu\tau\left[\frac{\Lambda}{\lambda} + \frac{(z_T - z_C - \tau)(e+L)}{eL}\left(1 - \frac{\mu}{\lambda}\right)\right]$, where (a) follows as $T_C = \frac{\Lambda e L}{\mu(e+L)}$. Consequently, the static revenue optimization problem is:

$$\max_{\tau \in \mathbb{R}} R(\tau) := \mu\tau\left[\frac{\Lambda}{\lambda} + \frac{(z_T - z_C - \tau)(e+L)}{eL}\left(1 - \frac{\mu}{\lambda}\right)\right] \quad \text{s.t.} \quad \max\{0, z_T - z_C - T_C\} \leq \tau \leq z_T - z_C. \tag{4}$$

Taking the first-order condition of this single dimensional quadratic optimization problem, we can show that its optimal solution corresponds to the static tolls in the statement of the proposition. □

Proposition 2 characterizes the revenue-optimal static toll by casting the static revenue optimization problem as a single-variable quadratic program, where the optimal toll depends on the magnitude of $z_T - z_C$ relative to the threshold $\frac{\Lambda e L}{(\lambda-\mu)(e+L)}$. A key step in proving this result requires extending the equilibrium analysis in Proposition 1 (Gonzales and Daganzo 2012), which characterizes monotonicity properties of the number of users choosing transit at equilibrium, to obtain an explicit expression for this quantity. This explicit characterization enables closed-form expressions for both revenue and total system cost under the optimal static toll across parameter regimes, which underpins our performance comparison between static and dynamic tolling in Sections 4.3 and 4.4.

4.2 Dynamic Revenue-Optimal Tolls

We now characterize the revenue-optimal dynamic tolling policy $\tau_d^*(\cdot)$. Theorem 1 establishes that the revenue-optimal dynamic toll has the trapezoidal structure shown in Figure 2, where the tolling policy is constant and equal to $z_T - z_C$ over the middle of the rush, i.e., over the interval $[t_B^*, t_C^*]$, and decreases linearly on either side of this interval. The corresponding equilibrium arrival and departure (crossing) time profiles of users at the bottleneck under this tolling policy are as depicted in Figure 9 in Appendix A.2.

Theorem 1 (Revenue-Optimal Dynamic Tolls)**.** *Suppose $\mu < \lambda$ and users choose between traveling by car through a bottleneck with a free flow cost z_C and using public transit with a fixed cost z_T, where $z_T \geq z_C$. Then, the revenue-optimal dynamic tolling policy $\tau_d^*(\cdot)$ is such that (see Figure 2):*

1. *The toll is constant and fixed at $\tau_d^*(t) = z_T - z_C$ for all $t \in [t_B^*, t_C^*]$ during the middle of the rush. Moreover, the time interval $[t_B^*, t_C^*]$ over which the toll remains constant corresponds to a fraction $f^* = \max\{1 - \frac{(z_T - z_C)\mu(e+L)}{\Lambda e L}\left(1 - \frac{\mu}{\lambda}\right), 0\}$ of the interval $[t_1, t_2]$, i.e., $\frac{t_C^* - t_B^*}{t_2 - t_1} = f^*$.*
2. *For all $t \in [t_A^*, t_B^*]$, $\tau^*(t) = z_T - z_C - e(t_B^* - t)$, and for all $t \in [t_C^*, t_D^*]$, $\tau^*(t) = z_T - z_C - L(t - t_C^*)$, where $t_B^* - t_A^* = \frac{L}{e+L}\frac{(1-f^*)\Lambda}{\mu}$, and $t_D^* - t_C^* = \frac{e}{e+L}\frac{(1-f^*)\Lambda}{\mu}$.*

Moreover, under this tolling policy, the optimal revenue is given by:

$$R^* = \begin{cases} (z_T - z_C)\Lambda\frac{\mu}{\lambda} + \frac{(z_T - z_C)^2\mu(e+L)}{2eL}\left(1 - \frac{\mu}{\lambda}\right)^2, & \text{if } z_T - z_C \leq \frac{\Lambda e L}{(\lambda-\mu)(e+L)}\frac{\lambda}{\mu}, \\ (z_T - z_C)\Lambda - \frac{\Lambda^2}{2\mu}\frac{eL}{e+L}, & \text{if } z_T - z_C > \frac{\Lambda e L}{(\lambda-\mu)(e+L)}\frac{\lambda}{\mu}. \end{cases} \tag{5}$$

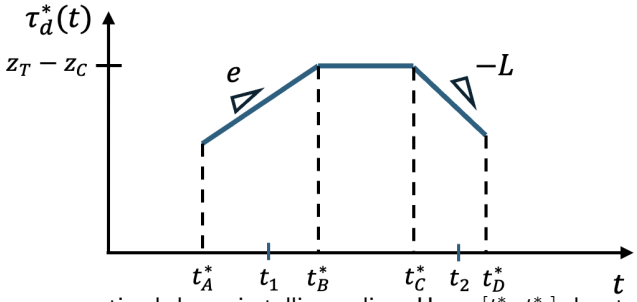


Figure 2: Depiction of the revenue optimal dynamic tolling policy. Here, $[t_A^*, t_D^*]$ denotes the equilibrium interval over which users pass the bottleneck under the dynamic revenue-optimal tolling policy $\tau_d^*(\cdot)$, analogous to the no-toll equilibrium interval $[t_A, t_D]$ shown in the right of Figure 1. The sub-interval $[t_B^*, t_C^*]$ corresponds to the middle of the rush, during which the toll is constant and equal to $z_T - z_C$.

Proof. Under the cost function in Equation (2), a necessary condition at equilibrium is that all users minimize travel costs when using a car, i.e., the following condition must hold: $c'(t, t^*) = 0$, which implies that $w'(t) + \tau'(t) = e$ for early arrivals and $w'(t) + \tau'(t) = -L$ for late arrivals. Moreover, at equilibrium, $w(t) + \tau(t) \leq z_T - z_C$ for all t, as otherwise users would switch to transit. Thus, the sum of the equilibrium waiting and toll costs take the form depicted in Figure 3, where the exact times t'_A, t'_B, t'_C, t'_D shown in the figure are determined endogenously by the tolling policy.

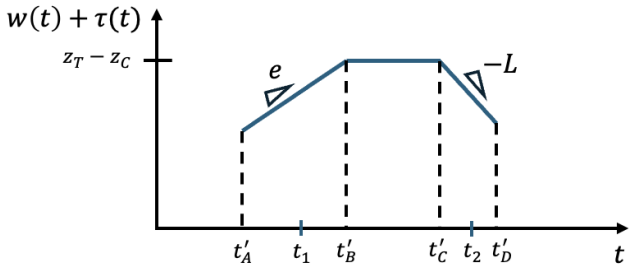


Figure 3: Sum of the waiting plus toll costs at equilibrium. The time points t'_A, t'_B, t'_C, t'_D are determined endogenously by the tolling policy and may, in general, differ from the corresponding time points shown in Figure 1 (right) and Figure 2, which correspond to the no-toll and revenue-optimal dynamic toll settings, respectively.

Leveraging the structure of the sum of the waiting and toll costs at equilibrium in Figure 3, we proceed by showing that in analyzing dynamic revenue-optimal policies, it suffices to restrict attention to tolling

schemes under which the waiting time satisfies $w(t) = 0$ for all periods t. To see this, for any policy $\tau(\cdot)$ with associated periods t'_A, t'_B, t'_C, t'_D as depicted in Figure 3, we construct an alternate policy $\tilde{\tau}(\cdot)$, where $\tilde{\tau}(t) = w(t) + \tau(t)$ for all t. Then, we have $R(\tau(\cdot)) = \int_{t'_A}^{t'_D} \mu\tau(t)dt \leq \int_{t'_A}^{t'_D} \mu\tilde{\tau}(t)dt = R(\tilde{\tau}(\cdot))$, where the inequality follows as $\tilde{\tau}(t) \geq \tau(t)$ for all t.

Thus, focusing on tolling policies $\tilde{\tau}(\cdot)$ where the waiting time $w(t) = 0$ for all t, we have:

$$\begin{aligned} R(\tilde{\tau}(\cdot)) &= \int_{t'_A}^{t'_D} \mu\tilde{\tau}(t)dt = \int_{t'_A}^{t'_B} \mu\tilde{\tau}(t)dt + \int_{t'_B}^{t'_C} \mu\tilde{\tau}(t)dt + \int_{t'_C}^{t'_D} \mu\tilde{\tau}(t)dt, \\ &= \mu\left[\int_{t'_A}^{t'_B} z_T - z_C - L(t'_B - t)dt + \int_{t'_B}^{t'_C} (z_T - z_C)dt + \int_{t'_C}^{t'_D} z_T - z_C - L(t - t'_C)dt\right], \\ &= \mu\left[(z_T - z_C)(t'_D - t'_A) - \frac{e(t'_B - t'_A)^2}{2} - \frac{L(t'_D - t'_C)^2}{2}\right]. \end{aligned}$$

Let f be the fraction of time in the horizontal portion of the waiting time curve, i.e., $\frac{t'_C - t'_B}{t_2 - t_1} = f$. Then, $t'_C - t'_B = f(t_2 - t_1) = f\frac{\Lambda}{\lambda}$. Since the remaining $(1 - f)$ fraction of users all use a car, it follows that $\mu(t'_B - t'_A + t'_D - t'_C) = (1 - f)\Lambda$. Thus, the above revenue expression reduces to:

$$R(\tilde{\tau}(\cdot)) = (z_T - z_C)\left(f\Lambda\frac{\mu}{\lambda} + (1 - f)\Lambda\right) - \frac{e\mu(t'_B - t'_A)^2}{2} - \frac{L\mu(t'_D - t'_C)^2}{2}.$$

Next, we know that $t'_B - t_1 + t_2 - t'_C = (1 - f)\frac{\Lambda}{\lambda}$, where $\lambda(t'_B - t_1) = \mu(t'_B - t'_A)$ and $\lambda(t_2 - t'_C) = \mu(t'_D - t'_C)$. Then, it is straightforward to see that to maximize $\frac{e\mu(t'_B - t'_A)^2}{2} + \frac{L\mu(t'_D - t'_C)^2}{2}$, it must be that $t'_B - t'_A = \frac{L}{e+L}\frac{(1-f)\Lambda}{\mu}$ and $t'_D - t'_C = \frac{e}{e+L}\frac{(1-f)\Lambda}{\mu}$. We thus obtain the following expression for the revenue, which, with a slight abuse of notation, we re-express as a function of the fraction f:

$$R(f) = (z_T - z_C)\left(f\Lambda\frac{\mu}{\lambda} + (1 - f)\Lambda\right) - \frac{\Lambda^2}{2\mu}\frac{eL}{e + L}(1 - f)^2. \tag{6}$$

Given this relation for the revenue, we have the following revenue optimization problem:

$$\max_{f\in[0,1]} R(f) = (z_T - z_C)\left(f\Lambda\frac{\mu}{\lambda} + (1 - f)\Lambda\right) - \frac{\Lambda^2}{2\mu}\frac{eL}{e + L}(1 - f)^2 \text{ s.t. } 1 - \min\left\{\frac{z_T - z_C}{T_C}, 1\right\} \leq f \leq 1. \tag{7}$$

We have thus reduced the dynamic revenue optimization problem from one of optimizing over a set of tolling functions to optimizing over a single dimensional variable f. Note here that when $z_T - z_C \leq T_C = \frac{\Lambda eL}{\mu(e+L)}$, the lower bound constraint reduces to $f \geq 1 - \frac{z_T - z_C}{T_C}$ to ensure that the tolls are non-negative at all periods. Then, taking the above problem's first order condition, we obtain $f^* = \max\left\{1 - \frac{(z_T - z_C)\mu(e+L)}{\Lambda eL}\left(1 - \frac{\mu}{\lambda}\right), 0\right\}$. Finally, substituting f^* in the revenue expression in Equation (6), we obtain the expression for the optimal revenue in the theorem statement. □

Theorem 1 characterizes the revenue-optimal dynamic toll in the bottleneck model with an outside option. To achieve this result, we reduced the revenue maximization problem, which requires optimizing over tolling functions, to a single-variable quadratic program. To our knowledge, this is the first analysis of revenue maximization in the bottleneck model with an outside option. The resulting revenue-optimal policy closely resembles the system-cost-optimal dynamic toll in Gonzales and Daganzo (2012), which also has a trapezoidal form and is, in particular, identical to the equilibrium waiting time profile shown on the right of Figure 1 and eliminates congestion delays at equilibrium. Despite this similarity, the two policies differ in key ways. First, by comparing the functions in Figure 1 (right) and Figure 2, the revenue-optimal policy can be viewed as a vertical shift of the system-cost-optimal tolling profile. Additionally, the duration of the flat segment, where the toll equals $z_T - z_C$, is chosen to maximize revenue and thus generally differs from that under the system-cost-optimal policy.

4.3 Revenue Comparison for Static vs. Dynamic Revenue-Optimal Tolls

Leveraging the characterizations of the static and dynamic revenue-optimal tolls, we compare their performance by deriving lower bounds on the fraction of the optimal dynamic revenue attained under the static

revenue-optimal policy. Theorem 2 shows that the revenue gap between the two policies depends on the magnitude of $z_T - z_C$, the difference between the cost of transit and car travel at free-flow. When $z_T - z_C$ lies below a lower threshold or above an upper threshold, the static revenue-optimal toll achieves at least two-thirds of the optimal dynamic revenue. In the intermediate regime, this revenue fraction is bounded below by one-half, implying that static revenue-optimal tolls achieve at least half of the dynamic optimal revenue across all parameter regimes. Sharper, regime-specific bounds are provided in the theorem statement.

Theorem 2 (Revenue Ratio: Static vs. Dynamic Revenue-Optimal Tolls)**.** *Suppose $z_T \geq z_C$ and $\mu < \lambda$. Moreover, let τ_s^* be the static revenue-optimal toll, $\tau_d^*(\cdot)$ be the dynamic revenue-optimal policy, and let $s = \frac{\Lambda e L}{(z_T - z_C)(\lambda - \mu)(e+L)} \geq 0$. Then, the ratio of the revenues of the two policies satisfies:*

$$\frac{R(\tau_s^*)}{R(\tau_d^*(\cdot))} \geq \begin{cases} \frac{2}{3-\frac{\mu}{\lambda}}, & \text{if } 0 \leq z_T - z_C < \frac{\Lambda e L}{(\lambda-\mu)(e+L)} \\ \min\left\{\frac{2+s}{4}, \frac{1}{2(1-\frac{\mu}{\lambda})}\right\}, & \text{if } z_T - z_C \in \left[\frac{\Lambda e L}{(\lambda-\mu)(e+L)}, \frac{\Lambda e L}{(e+L)}\left(\frac{1}{\lambda-\mu} + \frac{2}{\mu}\right)\right] \\ \frac{2}{3}, & \text{if } z_T - z_C > \frac{\Lambda e L}{(e+L)}\left(\frac{1}{\lambda-\mu} + \frac{2}{\mu}\right). \end{cases}$$

We establish this result by bounding the ratio of the revenue of the static revenue-optimal toll to the dynamic optimal revenue in Equation (5) in the three regimes for $z_T - z_C$ in the theorem statement, each corresponding to a distinct static revenue-optimal toll characterized in Proposition 2. For a proof of Theorem 2, see Appendix D.1. Theorem 2 establishes that revenue-optimal static tolls achieve a constant fraction of the optimal dynamic revenue, at least one-half in some parameter regimes and at least two-thirds in others, highlighting the efficacy of simple static tolls.

4.4 System Cost Bounds of Revenue-Optimal Tolling Policies

We now compare the total system cost under static and dynamic revenue-optimal tolls to that under the dynamic system-cost-optimal policy. Unlike the revenue comparisons in the previous section, constant-factor system cost guarantees do not hold across all parameter regimes (see Proposition 3 at the end of this section). Nevertheless, in the regime where both car and transit are used at the untolled equilibrium, i.e., when $z_T - z_C \leq T_C = \frac{\Lambda e L}{\mu(e+L)}$, a condition characteristic of cities with viable public transit and holds for real systems in our numerical experiments in Section 6, we establish a constant-factor guarantee. Specifically, in this regime, both static and dynamic revenue-optimal tolls incur at most twice the minimum achievable system cost.

Theorem 3 (System Cost Comparison)**.** *Suppose $\mu < \lambda$, and let τ_s^* be the static revenue-optimal toll, $\tau_d^*(\cdot)$ be the dynamic revenue-optimal toll, and SC^* be the minimum achievable system cost. If $z_T - z_C \leq \frac{\Lambda e L}{\mu(e+L)}$, then $SC(\tau_d^*(\cdot)) \leq 2SC^*$ and $SC(\tau_s^*) \leq 2SC^*$.*

Proof (Sketch). To prove this claim, we first leverage the characterizations of the revenue-optimal static and dynamic tolling policies in Proposition 2 and Theorem 1, respectively, to derive expressions for their corresponding total system costs at equilibrium. In particular, for both policies, we compute the total system cost by, summing across all users, the following four terms: (i) the cost of using transit, (ii) the free-flow cost of using a car, (iii) schedule delay costs, and (iv) waiting time delays, where, recall from the proof of Theorem 1 that there are no waiting time delays under the dynamic revenue-optimal policy. Summing these relations, we then prove that the total system cost under both policies remains upper bounded by $z_C \Lambda \frac{\mu}{\lambda} + z_T \left(1 - \frac{\mu}{\lambda}\right) \Lambda$. Then, comparing this upper bound to the minimum achievable total system cost derived in Gonzales and Daganzo (2012), combined with the fact that $z_T - z_C \leq \frac{\Lambda e L}{\mu(e+L)}$, we obtain our desired bounds. □

For a complete proof of Theorem 3, see Appendix D.2. Theorem 3 shows that when transit provides a viable outside option, i.e., when $z_T - z_C \leq \frac{\Lambda e L}{(\lambda-\mu)(e+L)}$, revenue-optimal tolls achieve total system costs within a modest factor of the system-cost-optimal benchmark. Together, Theorems 2 and 3 show that static revenue-optimal tolling simultaneously achieves strong revenue and system cost guarantees by capturing at least a constant fraction of the optimal dynamic revenue while incurring at most twice the minimum achievable

system cost (when $z_T - z_C \leq \frac{\Lambda e L}{\mu(e+L)}$). Since our comparisons abstract from the substantial informational and computational burdens that dynamic tolling places on traffic planners and users, these bounds should be interpreted as worst-case relative guarantees. Incorporating realistic behavioral or operational constraints would likely strengthen these bounds, further reinforcing the case for static tolling in practical applications.

The constant-factor guarantees in Theorem 3 rely on the cost difference between transit and car travel at free flow being bounded by $\frac{\Lambda e L}{\mu(e+L)}$, ensuring both modes are used at equilibrium (see Section 3.2). When this condition fails, constant-factor system cost ratios generally do not hold.

Proposition 3 (Unbounded System Cost Ratios)**.** *Let τ_s^* be the static revenue-optimal toll and SC^* be the minimum achievable total system cost. Further, suppose that $z_T - z_C > \frac{\Lambda e L}{\mu(e+L)}$, with $z_C = 0$ and $z_T > \frac{\Lambda e L}{(e+L)(\lambda-\mu)} \left(\frac{2\lambda-\mu}{\mu}\right)$. Then, $SC(\tau_s^*) = \left(1 + \frac{1}{1-\frac{\mu}{\lambda}}\right) SC^*$, which is unbounded when $\frac{\mu}{\lambda} \to 1$.*

For a proof of Proposition 3, see Appendix D.4. This result also extends when comparing the system cost of dynamic revenue-optimal tolls to the minimum achievable system cost. The main driver behind these unbounded system cost ratios is that when $z_T - z_C > \frac{\Lambda e L}{\mu(e+L)}$, transit is sufficiently unattractive compared to car travel that revenue-optimal tolls scale with $z_T - z_C$ to extract maximal revenue. In contrast, system-cost-optimal tolls are upper bounded by $\frac{\Lambda e L}{\mu(e+L)}$ and are calibrated to eliminate congestion rather than exploit users' willingness to pay. Thus, when $z_T - z_C > \frac{\Lambda e L}{\mu(e+L)}$, the revenue and system-cost-optimal policies operate on fundamentally different scales, precluding constant factor system cost guarantees. Despite this unboundedness result, Theorem 3 demonstrates the robustness of revenue-optimal tolling even on the system cost metric in regimes where both car and public transit are used at the untolled equilibrium, an empirically relevant regime that characterizes many transportation systems with viable public transit alternatives.

5 Revenue-Optimal Tolling in Urban Systems

The previous section analyzed static and dynamic revenue-optimal tolling in the bottleneck model with an outside option, where the bottleneck represents a single capacity-constrained facility (e.g., bridge or tunnel) and has a fixed capacity independent of the number of vehicles in the system. We now extend this analysis to city-scale urban traffic systems, where congestion emerges from interactions across an entire urban area rather than at a single facility. In such settings, the outflow capacity (or service rate) depends on the number of vehicles in the urban system and decreases once vehicle accumulation exceeds a critical threshold, e.g., due to queue spill back across intersections. We model these system-wide congestion dynamics using the Macroscopic Fundamental Diagram (MFD), which provides a parsimonious yet empirically grounded characterization of aggregate traffic behavior in urban systems (Daganzo 2007). While the MFD is commonly expressed as a relationship between average system-wide vehicle density and aggregate flow, when the average trip distance D is constant across users, it can be interpreted as describing a state-dependent service rate as a function of total vehicle accumulation (Gonzales and Daganzo 2012).

We focus on a well-established class of MFDs satisfying two standard properties:

- **Property 1:** There exists a critical vehicle accumulation level n_c beyond which additional vehicles reduce total outflow due to spill-back and gridlock effects.
- **Property 2:** For vehicle accumulations up to n_c, the throughput is linear, corresponding to free-flow conditions, achieving its maximum throughput μ_f at $n = n_c$.

Figure 4 illustrates the canonical triangular MFD satisfying these properties. For vehicle accumulations below the threshold n_c, the system operates at free-flow with no queuing delays. Once accumulation exceeds n_c, congestion sets in, reducing the effective system capacity, which drops to zero at the jam accumulation level n_j. Vickrey's bottleneck model can be interpreted as a special case of such MFD relations, with a capacity plateau rather than a drop in the congested regime (see Appendix C).

For urban systems governed by MFDs satisfying the above properties, in Section 5.1, we first characterize the dynamic revenue-optimal policy and show that it maintains the system at the throughput-maximizing

capacity μ_f achieved at n_c. Then, in Section 5.2, we derive a closed-form expression for the revenue generated by any static toll for triangular MFDs (see Figure 4). Finally, while static revenue-optimal tolls generally achieve weaker guarantees than in the bottleneck setting (as the system service rate is state-dependent), in Section 5.3, we show that, across a broad and practically relevant range of parameters, static revenue-optimal tolls attain at least two-thirds of the dynamic optimal revenue while incurring at most twice the minimum attainable system cost.

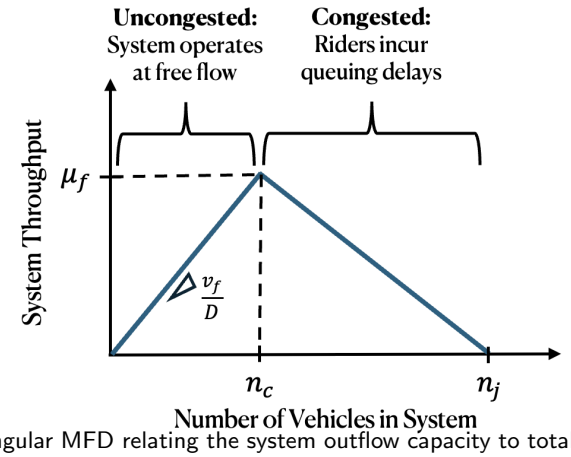


Figure 4: Depiction of the triangular MFD relating the system outflow capacity to total vehicle accumulation. The slope of the segment connecting $(0, 0)$ to any point on the curve determines the average system speed, equal to average trip distance D times the slope. The MFD consists of an uncongested regime in which the system operates at free-flow with a speed v_f up to the critical threshold n_c. Beyond n_c, the system enters a congested regime in which speeds decline, eventually dropping to zero at the jam accumulation level n_j.

5.1 Revenue-Optimal Dynamic Tolling

This section characterizes the revenue-optimal dynamic tolling policy for urban systems governed by MFDs satisfying Properties 1 and 2, such as the triangular MFD in Figure 4, and shows that it maintains the system at the critical vehicle accumulation level that maximizes system throughput. In proving this result, akin to Section 4, we focus on the setting when the maximum achievable system throughput μ_f is strictly below the desired arrival rate λ, i.e., $\mu_f < \lambda$. For a depiction of the arrival and departure time profiles of users in the urban system under the revenue-maximizing tolling policy, see Figure 10 in Appendix A.2.

Theorem 4 (Revenue-Optimal Dynamic Tolls under MFD)**.** *Suppose $\mu_f < \lambda$ and users choose between two modes, traveling by car through an urban system characterized by an MFD satisfying properties 1 and 2, and using public transit with a fixed cost z_T, where $z_T \geq z_C$. Then, under the revenue-optimal dynamic tolling policy, the system always operates at the critical vehicle accumulation level n_c corresponding to a throughput-maximizing capacity μ_f.*

Proof (Sketch). First, as with the bottleneck model, a necessary equilibrium condition is that $w'(t)+\tau'(t) = e$ for early arrivals and $w'(t)+\tau'(t) = -L$ for late arrivals, and that $w(t)+\tau(t) \leq z_T - z_C$ for all t, i.e., the sum of the waiting and toll costs are as depicted in Figure 3. Here, the times t'_A, t'_B, t'_C, t'_D shown in the figure are determined endogenously by the tolling policy and the waiting time depends on the number of users $n(t)$ in the system at time t. Under this equilibrium condition, we compare two policies: (i) a candidate revenue-optimal policy $\tau^*(\cdot)$ with associated times $t^*_A, t^*_B, t^*_C, t^*_D$ in Figure 3, and (ii) a policy $\tilde{\tau}(\cdot)$, with associated times $\tilde{t}_A, \tilde{t}_B, \tilde{t}_C, \tilde{t}_D$, under which the system operates at the capacity μ_f with no waiting delays, where $\tilde{t}_B = t^*_B$ and $\tilde{t}_C = t^*_C$, such that $\tilde{\tau}(t) = z_T - z_C$ for the period $[t^*_B, t^*_C]$. Finally, using Properties 1 and 2, we show $R(\tilde{\tau}(\cdot)) \geq R(\tau^*(\cdot))$. □

For a complete proof of Theorem 4, see Appendix D.3. The challenge in establishing this result and, in particular, proving $R(\tilde{\tau}(\cdot)) \geq R(\tau^*(\cdot))$ lies in handling a key trade-off under a revenue-maximization

objective. Specifically, we compare the revenue of a candidate optimal policy $\tau^*(\cdot)$ that operates over a longer interval $[t_A^*, t_D^*]$ over which both throughput and tolls may be lower, with that of an alternative policy $\tilde{\tau}(\cdot)$ that enforces operation at the throughput-maximizing level μ_f by charging (weakly) higher tolls over a shorter interval $[\tilde{t}_A, \tilde{t}_D]$. Apriori, it is not obvious which effect dominates: higher tolls over a shorter interval or lower tolls over a longer interval. Our analysis shows that despite this apparent tradeoff between toll intensity and duration, revenue is maximized by maintaining the system at the throughput-maximizing operating point. We note that this trade-off does not arise in the bottleneck setting, where throughput is fixed at the bottleneck capacity; in contrast, under the MFD, throughput is endogenous and depends on vehicle accumulation in the system, making the analysis fundamentally more involved.

Theorem 4 establishes that the dynamic revenue-optimal toll eliminates queuing delays and maintains the system at the throughput-maximizing capacity μ_f, a desirable property in real-world traffic systems. Consequently, as immediate corollaries of Theorems 1 and 4, the dynamic revenue-optimal policy is akin to the tolling policy depicted in Figure 2 (as it sets waiting times point-wise to zero) and the revenue and system cost expressions under this policy coincide with those in the bottleneck setting in Section 4, with the bottleneck capacity μ replaced by μ_f.

Since the dynamic revenue-optimal toll eliminates queuing delays and sustains an equilibrium at the throughput-maximizing point of the MFD, it closely resembles the system-cost-optimal policy in Geroliminis and Levinson (2009), which also maintains the system at the critical accumulation n_c. Nevertheless, the two policies differ in important respects, for the same reasons discussed in Section 4.2.

More broadly, Theorem 4 can be interpreted through the lens of standard revenue-maximization trade-offs. In many settings, tolls or prices directly affect the number of users served, and the optimal revenue is achieved by balancing higher payments per user against reduced demand. In the setting studied here, Theorem 4 shows that this balance is achieved at the boundary of the feasible operating region, with the revenue-optimal policy maintaining the system at the critical accumulation n_c, rather than at higher accumulation levels on the congested branch of the MFD.

5.2 Static Tolling under Triangular MFD

This section characterizes the revenue under static tolls for urban systems governed by a triangular MFD (see Figure 4). Unlike the dynamic tolling analysis in the previous section, which applies to general MFDs satisfying Properties 1 and 2, static tolling can induce operation on the congested branch of the MFD and thus requires additional structure for analytical tractability. Thus, focusing on triangular MFDs, we derive an expression for the revenue as a function of any static toll τ.

Theorem 5 (Revenue under Static Tolls in MFD Framework)**.** *Suppose $\mu_f < \lambda$ and users choose between two modes, traveling by car through an urban system characterized by a triangular MFD (see Figure 4), and using public transit with a fixed cost z_T, where $z_T \geq z_C$ and all trips are of a fixed distance D. Moreover, let $\underline{\tau} \geq 0$ be the minimum toll at which there is some user that is indifferent between using car and transit. Then, the revenue under any static toll $\tau \in [\underline{\tau}, z_T - z_C]$ is given by:*

$$R(\tau) = \tau \left[\frac{\Lambda}{\lambda} \frac{n_j}{\frac{n_j}{\mu_f} + z_T - z_C - \tau} + n_j \frac{e+L}{eL} \ln\left(1 + \frac{(z_T - z_C - \tau)\mu_f}{n_j}\right) \left(1 - \frac{\frac{n_j}{\frac{n_j}{\mu_f} + z_T - z_C - \tau}}{\lambda}\right) \right]. \tag{8}$$

Proof. We prove this claim in two steps. First, under a triangular MFD, we characterize the equilibrium system throughput $\mu(n(t))$, which is time-varying and depends on the vehicle accumulation in the system, unlike in the bottleneck setting, under a static toll τ. Using this relation for the system throughput, we then derive the revenue expression in the statement of the theorem.

To establish a relation for $\mu(n(t))$, first note that the static toll $\tau \geq 0$ must be such that there is some user with a cost of z_T at equilibrium. If not, all users would take the car at equilibrium and thus the toll can be increased without reducing car ridership or revenues. Thus, let $\underline{\tau} \geq 0$ be the minimum toll at which there is some user who is indifferent between using car and transit.

Next, recall that under a static toll $\tau \geq \underline{\tau}$, the equilibrium waiting time is akin to the right of Figure 1, with the time points t_A, t_B, t_C, t_D determined endogenously by the tolling policy and where the peak waiting time is $z_T - z_C - \tau$ during the interval $[t_B, t_C]$. Moreover, note that in the interval $[t_A, t_B]$, the waiting time is given by $z_T - z_C - \tau - e\Delta$ where $\Delta \in [0, t_B - t_A]$. Furthermore, in the interval $[t_C, t_D]$, the waiting time is given by $z_T - z_C - \tau - L\Delta$ where $\Delta \in [0, t_D - t_C]$.

Next, let v_f be the free-flow speed in the uncongested part of the MFD, so free-flow travel time is $\frac{D}{v_f}$. Moreover, letting $v(n(t))$ be the system speed at time t, the corresponding travel time with vehicle accumulation $n(t)$ is $\frac{D}{v(n(t))}$. Here, $v(n(t)) = D\frac{\mu(n(t))}{n(t)}$, since the system speed equals to the trip distance D times the slope from the origin to any point on the triangular MFD in Figure 4. Consequently, with slight abuse of notation, we have the following relation for the waiting time: $w(n(t)) = \frac{D}{v(n(t))} - \frac{D}{v_f} = \frac{n(t)}{\mu(n(t))} - \frac{n_c}{\mu_f}$. Moreover, by our triangular MFD assumption, $\frac{\mu_f}{\mu(n(t))} = \frac{n_j - n_c}{n_j - n(t)}$, which implies $n(t) = n_j - \frac{\mu(n(t))}{\mu_f}(n_j - n_c)$. Substituting this relation in the above waiting time equation and rearranging, we obtain that: $\mu(n(t)) = \frac{n_j}{\frac{n_j}{\mu_f} + w(n(t))}$.

Let μ_τ be the fixed system throughput during the period $[t_B, t_C]$ when the waiting time function is flat and at its peak under the static toll τ. Then, using the above relation for the system throughput as a function of t, we have the following expression for the revenue given a fixed toll τ:

$$
\begin{aligned}
R(\tau) &= \tau \int_{t_A}^{t_D} \mu(n(t))dt = \tau \left[\int_{t_A}^{t_B} \mu(n(t))dt + \int_{t_B}^{t_C} \mu(n(t))dt + \int_{t_C}^{t_D} \mu(n(t))dt \right], \\
&\overset{(a)}{=} \tau \left[\int_0^{t_B - t_A} \frac{n_j}{\frac{n_j}{\mu_f} + (z_T - z_C - \tau - \Delta e)} d\Delta + \int_0^{t_D - t_C} \frac{n_j}{\frac{n_j}{\mu_f} + (z_T - z_C - \tau - \Delta L)} d\Delta + \mu_\tau (t_C - t_B) \right], \\
&\overset{(b)}{=} \tau \left[\frac{n_j}{e} \ln \left(1 + \frac{(z_T - z_C - \tau)\mu_f}{n_j} \right) + \frac{n_j}{L} \ln \left(1 + \frac{(z_T - z_C - \tau)\mu_f}{n_j} \right) + \frac{n_j}{\frac{n_j}{\mu_f} + z_T - z_C - \tau} (t_C - t_B) \right], \\
&= \tau \left[n_j \frac{e + L}{eL} \ln \left(1 + \frac{(z_T - z_C - \tau)\mu_f}{n_j} \right) + \frac{n_j}{\frac{n_j}{\mu_f} + z_T - z_C - \tau} (t_C - t_B) \right],
\end{aligned}
$$

where (a) follows by the variable transformation $\Delta = t_B - t$ for the integral between $[t_A, t_B]$ and the variable transformation $\Delta = t - t_C$ for the integral between $[t_C, t_D]$, and (b) follows by evaluating the integral and noting that the waiting time at t_A and t_D is zero.

Next, noting that $\int_{t_A}^{t_B} \mu(n(t))dt = \lambda(t_B - t_1)$ and $\int_{t_C}^{t_D} \mu(n(t))dt = \lambda(t_2 - t_C)$, we have: $t_C - t_B = \frac{\Lambda}{\lambda} - \frac{1}{\lambda}\left[\int_{t_A}^{t_B} \mu(n(t))dt + \int_{t_C}^{t_D} \mu(n(t))dt \right] = \frac{\Lambda}{\lambda} - \frac{1}{\lambda} n_j \frac{e+L}{eL} \ln \left(1 + \frac{(z_T - z_C - \tau)\mu_f}{n_j} \right)$. Substituting this into the above relation for the revenue and simplifying, we obtain our desired result. □

While Theorem 5 yields a closed-form expression for the revenue under a static toll τ, its complexity makes it challenging to characterize the revenue-optimal static toll in closed form, in contrast to the bottleneck setting in Section 4.1. This difficulty arises since, unlike dynamic revenue-optimal tolls, static revenue-optimal tolling can, in general, induce operation on the congested branch of the MFD, resulting in congestion delays. That said, since the revenue function in Equation (8) is Lipschitz continuous over $\tau \in [0, z_T - z_C]$, an approximately optimal toll can be computed by discretizing this interval and selecting the toll with the highest revenue from this set.

5.3 Static vs. Dynamic Tolling under Triangular MFD

Due to the state-dependent capacities under the MFD framework, static tolls generally admit weaker performance guarantees than in the bottleneck model. For instance, while static revenue-optimal tolls achieve at least half of the dynamic optimal revenue in the bottleneck model (Theorem 2), our numerical experiments in Section 6 show that this revenue ratio may drop below half for urban systems governed by a triangular MFD. Nevertheless, for such systems, we show that there exists a toll that, over a broad and practically relevant range of parameters, achieves at least two-thirds of the dynamic optimal revenue and incurs at most twice the minimum attainable system cost.

We consider the regime in which the cost difference between transit and the free-flow cost of using a car satisfies $z_T - z_C \leq \frac{\Lambda e L}{(\lambda - \mu_f)(e+L)}$, an empirically relevant condition for urban systems where public transit provides a viable alternative to car travel, as supported by our numerical experiments in Section 6. In this regime, the toll $\tau = z_T - z_C$, the revenue-optimal static toll in the bottleneck setting (see Proposition 2), maintains the system at the throughput-maximizing capacity μ_f with no congestion delays, as any delays would induce car users to switch to transit. As a result, both the revenue expression in Equation (8) and the corresponding system cost reduce to their bottleneck counterparts, with the capacity μ replaced by μ_f. Then, following the proofs of Theorem 2 and 3, if $z_T - z_C \leq \frac{\Lambda e L}{(\lambda - \mu_f)(e+L)}$, the static toll $\tau = z_T - z_C$ (i) achieves at least a $\frac{2}{3 - \frac{\mu_f}{\lambda}}$ fraction of the optimal dynamic revenue and (ii) incurs at most twice the minimum achievable system cost.

While the above guarantees do not extend to the regime in which $z_T - z_C > \frac{\Lambda e L}{(\lambda - \mu_f)(e+L)}$, they nonetheless highlight that static tolling can achieve robust performance across both revenue and system cost metrics over a broad and practically relevant range of parameters, even for urban systems governed by a triangular MFD. That said, owing to the complexity of the revenue expression under static tolls in Equation (8), a general characterization of the performance gap between static revenue-optimal tolling and its dynamic benchmarks is considerably more challenging than in the bottleneck setting. Therefore, we provide a more comprehensive assessment of this performance gap through numerical experiments in the next section.

6 Numerical Experiments

This section presents numerical experiments comparing the efficacy of static revenue-optimal tolling against its dynamic revenue and system-cost-optimal counterparts based on two real-world application cases of congestion pricing: (i) the SF–Oakland Bay Bridge for the bottleneck model and (ii) New York City's congestion reduction zone (CRZ) for the MFD framework. Our results, which are grounded in practical datasets, validate our theoretical guarantees and demonstrate that simple static tolling policies can deliver robust performance on both revenue and system cost metrics, underscoring their practical effectiveness and shedding light on the strong empirical performance of the many static congestion pricing systems already in operation. In the following, we provide an overview of our experimental setup (Section 6.1) and present results comparing the relative efficacy of static and dynamic tolling for both application cases (Section 6.2). For complete details on our data sources and model calibration procedure for both case studies, see Appendix E. The code to generate the results are available at the following Link.

6.1 Overview of Experimental Setup

For both the SF-Oakland Bay Bridge and New York City (NYC) case studies, we focus on the weekday morning commuting period between 5-10 AM, when congestion is the most pronounced and a majority of cars are subject to a *flat, time-invariant* toll of \$8.50 for westbound trips on the Bay Bridge and \$9 for entry into NYC's Congestion Relief Zone (CRZ). The use of simple static tolls in both settings makes them natural empirical settings for evaluating the performance of static tolling relative to its dynamic benchmarks. In each case, we consider a setting where users choose between traveling by car or using a local public transit alternative, Bay Area Rapid Transit (BART) in the Bay Bridge corridor or the NYC subway system, which we model as the outside option.

To calibrate the parameters of the bottleneck model with an outside option for the Bay Bridge study and the MFD framework for the NYC study, we draw on a broad set of data sources. These include BART and NYC subway ridership data (Bay Area Rapid Transit (BART) 2026, State of New York 2025a), vehicular demand and flow data (such as hourly traffic counts for the Bay Bridge and CRZ entry counts for NYC) (California Department of Transportation 2024, State of New York 2025b), taxi and for-hire vehicle trip records in NYC (New York City Taxi and Limousine Commission 2025), and OpenStreetMap data (OpenStreetMap contributors 2025) to compute total roadway length within the CRZ for calibrating the parameters of the triangular MFD. We combine these data sources with standard estimates of the value of time and schedule delay penalties from the literature.

A key quantity we vary in our experiments is the cost difference between transit and car travel at free-flow, $z_T - z_C$, which plays a central role in our theoretical results in Sections 4 and 5. To do so, while we fix z_C to reflect the sum of the costs of parking and the average free-flow travel time for morning commutes in each setting, we vary z_T through a single parameter: *a discomfort multiplier* η. Specifically, in our experiments, we model z_T as the sum of the transit fare and a time-based cost composed of walking to and from stations, waiting time on platforms, and in-vehicle BART or subway travel time, with these time components scaled by η to reflect the empirically observed fact that time spent using transit is perceived as more onerous than time spent driving (Wardman 2012).

Because there is no single, well-established value for the discomfort multiplier and it can vary substantially across settings, we conduct a sensitivity analysis by varying $\eta \in [1.5, 18]$, with values in the range $[1.5, 5]$ consistent with empirical estimates (Wardman 2012). Exploring a broader range allows us to capture heterogeneity in user preferences and assess how technological or infrastructural changes, such as improvements in transit quality or the increased attractiveness of car travel due to autonomous vehicles (Ostrovsky and Schwarz 2025), affect the cost differential $z_T - z_C$. In addition, varying η over this wider range enables us to capture all the regimes for $z_T - z_C$ characterized in our theoretical results.

For further details on our data sources and model calibration procedure, see Appendix E.1. Moreover, we summarize the parameter values used in our experiments for the Bay Bridge and NYC case studies in Tables 2 (Appendix E.3) and 3 (Appendix E.6), respectively.

6.2 Results

This section compares the performance of static and dynamic tolls on both revenue and system cost metrics for the Bay Bridge and NYC case studies. To this end, Figures 5 and 6 plot the revenue and system cost ratios for a set of static and dynamic tolling policies, normalized by the corresponding dynamic revenue-optimal and system-cost-optimal benchmarks, as the discomfort multiplier η is varied for the Bay Bridge and NYC settings, respectively. We denote the static revenue-optimal and system-cost-optimal policies by *Static-RO* and *Static-SO*, and their dynamic counterparts by *Dynamic-RO* and *Dynamic-SO*. For the experiments, the static and dynamic revenue-optimal policies are computed using the derivations in Sections 4 and 5. The dynamic system-cost-optimal policy is computed using established results from the literature (Gonzales and Daganzo 2012, Geroliminis and Levinson 2009), while the static system-cost-optimal policy under the bottleneck and MFD models are computed based on the derivations in Appendices F.1 and F.2.

Our results show that for practically relevant values of the discomfort multiplier (i.e., $\eta \in [1.5, 5]$), static revenue-optimal tolls incur at most a 10% revenue loss in the Bay Bridge study and 20% in the NYC study compared to the dynamic revenue-optimal policy. As public transit becomes substantially less attractive ($\eta >> 5$), this gap widens to around 15% in the Bay Bridge study and can exceed 50% in NYC, consistent with our theory. While static revenue-optimal tolling achieves better revenue ratios in the Bay Bridge study, it performs worse on the system cost metric. Specifically, in the NYC study, static revenue-optimal tolls incur at most an 8% increase relative to the minimum achievable system cost, whereas, in the Bay Bridge study, the system cost increase reaches 25% for $\eta \in [1.5, 5]$. Furthermore, when $\eta \gg 5$, the system cost ratio in the Bay Bridge study can even exceed two, incurring more than twice the optimal cost. Thus, while static revenue-optimal tolls in NYC capture a smaller fraction of dynamic optimal revenue, they achieve a system cost much closer to the dynamic optimum compared to the Bay Bridge study, highlighting a core trade-off between the revenue and system cost objectives under static tolls.

Additionally, our results for the Bay Bridge study in Figure 5 highlight a non-monotonic relationship between the discomfort multiplier and the performance gap between static and dynamic policies. Static revenue-optimal tolls perform well when transit is only slightly less desirable than driving (i.e., low values of η), their performance deteriorates as transit becomes moderately worse corresponding to intermediate values of η, and then improves somewhat again when transit becomes highly unattractive (i.e., very large values of η). This non-monotonic behavior is consistent with our theoretical bounds on the revenue ratio between static and dynamic revenue-optimal policies in Theorem 2. Moreover, this non-monotonic behavior contrasts the strictly monotonic pattern observed in the NYC study under the MFD framework in Figure 6,

where the performance gap of static revenue optimal tolls relative to its dynamic benchmarks worsens as η increases.

Figure 5 further shows that the two static tolling policies (Static-RO and Static-SO) exhibit near identical performance across most values of the discomfort multiplier η. Notably, in the practically relevant range $\eta \in [1.5, 5]$, the two policies coincide, with a slight divergence between these policies at intermediate values of η. A similar pattern holds for the NYC study under the MFD framework, where the static system-cost-optimal and revenue-optimal tolls coincide for all discomfort multipliers shown in Figure 6; hence, we plot only the latter for visual clarity. Overall, these results suggest that if a central planner is required to set static tolls, as is common in practice, both revenue and system cost objectives can be achieved simultaneously particularly over practically relevant parameter ranges of the discomfort multiplier.

Beyond the discomfort multiplier, a key driver of performance is the ratio of the maximum system throughput to the desired user arrival rate ($\frac{\mu}{\lambda}$ for the bottleneck model and $\frac{\mu_f}{\lambda}$ for the MFD framework). When this ratio is much lower than one, as in NYC, where it is around 0.2 due to its high public transit mode share, the dynamic revenue and system-cost-optimal tolls nearly coincide, and static tolls achieve lower revenues and higher system costs compared to both dynamic benchmarks. Despite this Pareto dominance of dynamic tolling across both metrics in the transit-intensive NYC study, the performance losses of static tolling remain small for most practically relevant values of $\eta \in [1.5, 5]$. In contrast, when the ratio of the maximum system throughput to the desired user arrival rate is closer to one, as in the Bay Bridge morning peak where it is roughly 0.7, the two dynamic tolling policies differ considerably. In this case, while dynamic revenue-optimal tolls Pareto dominate both static policies by achieving higher revenues and lower system costs for all η, such a Pareto dominance does not hold for dynamic system-cost-optimal tolls, which can even generate up to 20% less revenue than both static policies for practically relevant values of η.

Finally, in both case studies, the static and dynamic revenue-optimal tolls align closely with the currently implemented tolls on the Bay Bridge and in NYC's CRZ for most practical ranges of the discomfort multiplier η. Notably, the static revenue-optimal toll equals the \$8.50 Bay Bridge toll at approximately $\eta = 2.1$ and the \$9 CRZ toll at $\eta = 1.7$. This lower value of the discomfort multiplier in NYC is reflective of its robust public transit system, which contrasts the dominance of car travel in the Bay Bridge corridor. At these values of η, we find that static tolls capture nearly all of the benefits of dynamic tolling, achieving roughly 98% of dynamic optimal revenue and increasing total system cost by only 3% relative to the minimum achievable system cost in both settings.

Overall, our results highlight the merits of static revenue optimal tolling policies relative to their dynamic optimal tolling counterparts on both revenue and total system cost metrics.

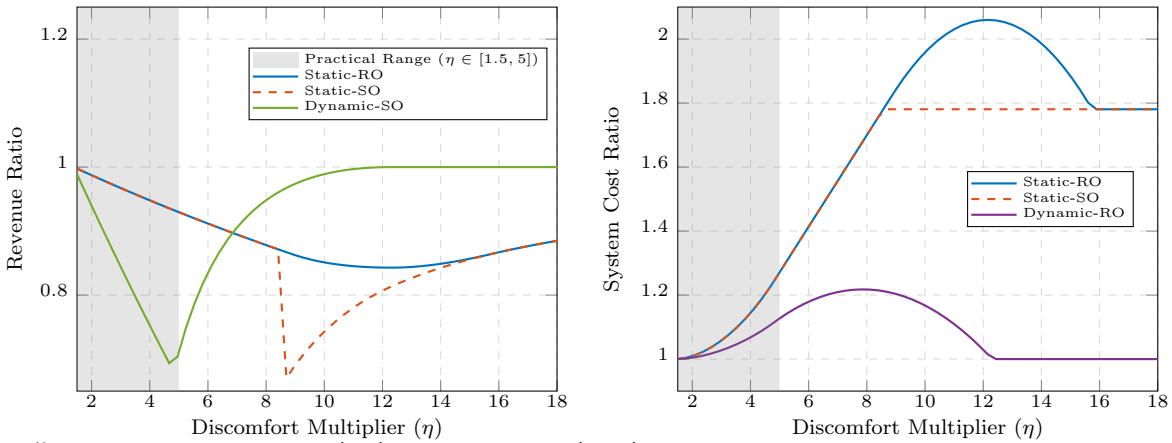


Figure 5: Depiction of the revenue (left) and system cost (right) ratios of a set of static and dynamic tolling policies, normalized by the corresponding dynamic revenue-optimal and system-cost-optimal benchmarks, as the discomfort multiplier η is varied for the Bay Bridge case study.

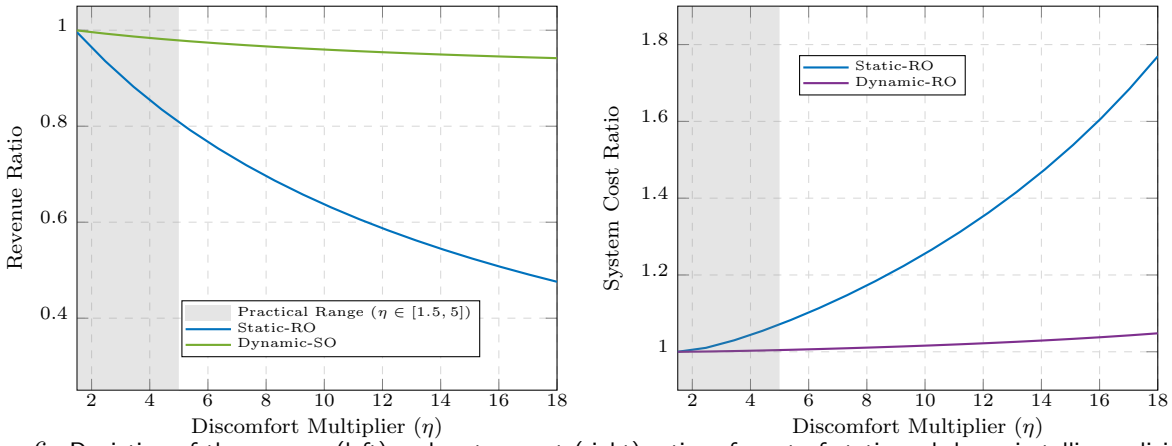


Figure 6: Depiction of the revenue (left) and system cost (right) ratios of a set of static and dynamic tolling policies, normalized by the corresponding dynamic revenue-optimal and system-cost-optimal benchmarks, as the discomfort multiplier η is varied for the New York City case study.

7 Conclusion and Future Work

This work analyzed the performance gap between simple static and optimal dynamic congestion pricing schemes in two canonical models capturing the dominant real-world use cases of congestion pricing. In both models, we derived, in closed form, the revenue-optimal static and dynamic tolling policies and showed that static revenue-optimal tolls deliver robust performance, achieving constant-factor approximations, on both revenue and system cost metrics for a wide range of practically relevant parameter regimes. We further validated our theory with experiments based on two real-world congestion pricing case studies. Overall, our results demonstrate the practical efficacy of simple static tolls and help shed light on the strong empirical performance of the many static congestion pricing systems already in operation. Moreover, since our work abstracts from the substantial informational and computational burdens dynamic tolling places on transportation planners and users, our results can be interpreted as worst-case guarantees. Incorporating realistic behavioral or operational constraints would only strengthen the case for static tolling.

There are several future research directions. First, it would be valuable to study the performance gap between static and dynamic tolling under standard extensions of the bottleneck model, including heterogeneous values of time, stochastic arrivals and utilities, non-uniform desired bottleneck departure distributions, and congestion-dependent outside options whose cost varies with utilization. It would also be worthwhile to examine tolling policies for MFDs, such as Greenshield's relation (Greenshields 1947), that do not satisfy Properties 1 and 2 or are non-triangular in the case of static tolling. Finally, there is scope to study objectives beyond revenue and system cost, and differentiated tolling schemes (e.g., carpool or taxi discounts) that better reflect real-world congestion pricing structures.

Acknowledgments

This work was supported by the Data Science Institute Postdoctoral Fellowship at Columbia University. We also thank Michael Ostrovsky for insightful discussions and Luyang Han for his support with the code to process the OpenStreetMap and NYC subway origin-destination data, required for our experiments in Section 6.

References

Sean Fleming. Traffic congestion cost the us economy nearly $87 billion in 2018. World Economic Forum, 2019. URL https://www.weforum.org/agenda/2019/03/traffic-congestion-cost-the-us-economy-nearly-87-billion-in-2018/. Accessed February 19, 2021.

Celia Fernandez. U.s. drivers lost 43 hours—a full work week—to traffic in 2024: 'if you're driving in new york, it's brutal,' expert says, 2025. URL https://www.cnbc.com/2025/01/07/inrix-most-congested-us-cities.html. Accessed November 17, 2025.

Cody Cook, Aboudy Kreidieh, Shoshana Vasserman, Hunt Allcott, Neha Arora, Freek van Sambeek, Andrew Tomkins, and Eray Turkel. The short-run effects of congestion pricing in new york city. Technical report, National Bureau of Economic Research, 2025.

Michael Ostrovsky and Frank Yang. Effective and equitable congestion pricing: New york city and beyond. Technical report, 2024.

Winnie Hu, Ana Ley, Stephen Castle, and Christina Anderson. Congestion pricing's impact on new york? these 3 cities offer a glimpse., 2023. URL https://www.nytimes.com/2023/12/02/nyregion/new-york-congestion-pricing-london-stockholm-singapore.html. Accessed November 17, 2025.

A. C. Pigou. *The economics of welfare*. Macmillan London, 1920.

Eric J. Gonzales and Eleni Christofa. Empirical assessment of bottleneck congestion with a constant and peak toll: San francisco–oakland bay bridge. *EURO Journal on Transportation and Logistics*, 3(3):267–288, 2015. ISSN 2192-4376. doi: 10.1007/s13676-013-0036-y. URL https://www.sciencedirect.com/science/article/pii/S2192437620301424.

Venktesh Pandey and Stephen D. Boyles. Dynamic pricing for managed lanes with multiple entrances and exits. *Transportation Research Part C: Emerging Technologies*, 96:304–320, 2018. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2018.09.017. URL https://www.sciencedirect.com/science/article/pii/S0968090X18302766.

Kitae Jang, Koohong Chung, and Hwasoo Yeo. A dynamic pricing strategy for high occupancy toll lanes. *Transportation Research Part A: Policy and Practice*, 67:69–80, 2014. ISSN 0965-8564. doi: 10.1016/j.tra.2014.05.009. URL https://www.sciencedirect.com/science/article/pii/S0965856414001232.

Ministry of Transport. Electronic road pricing (erp), 2025. URL https://www.mot.gov.sg/what-we-do/motoring-road-network-and-infrastructure/Electronic-Road-Pricing. Accessed November 17, 2025.

William S. Vickrey. Congestion theory and transport investment. *The American Economic Review*, 59(2):251–260, 1969. ISSN 00028282. URL http://www.jstor.org/stable/1823678.

Carlos F. Daganzo. Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B: Methodological*, 41(1):49–62, 2007. ISSN 0191-2615. doi: 10.1016/j.trb.2006.03.001. URL https://www.sciencedirect.com/science/article/pii/S0191261506000282.

MTA. Congestion pricing makes for better transit, 2025. URL https://www.mta.info/fares-tolls/tolls/congestion-relief-zone/better-transit. Accessed November 17, 2025.

Phillip O'Neill. Privatising and financialising roads: The peculiar case of transurban. *Geographical Research*, 60(2):328–341, 2022. doi: https://doi.org/10.1111/1745-5871.12528. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1745-5871.12528.

R. Arnott, A. de Palma, and R. Lindsey. Departure time and route choice for the morning commute. *Transportation Research Part B: Methodological*, 24(3):209–228, 1990a. ISSN 0191-2615. doi: https://doi.org/10.1016/0191-2615(90)90018-T. URL https://www.sciencedirect.com/science/article/pii/019126159090018T.

Takatoshi Tabuchi. Bottleneck congestion and modal split. *Journal of Urban Economics*, 34(3):414–431, 1993. ISSN 0094-1190. doi: https://doi.org/10.1006/juec.1993.1044. URL https://www.sciencedirect.com/science/article/pii/S0094119083710442.

Gordon F. Newell. The morning commute for nonidentical travelers. *Transportation Science*, 21(2):74–88, 1987. doi: 10.1287/trsc.21.2.74. URL https://doi.org/10.1287/trsc.21.2.74.

Robin Lindsey. Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes. *Transportation Science*, 38(3):293–314, 2004. doi: 10.1287/trsc.1030.0045. URL https://doi.org/10.1287/trsc.1030.0045.

Ling-Ling Xiao, Tian-Liang Liu, and Hai-Jun Huang. On the morning commute problem with carpooling behavior under parking space constraint. *Transportation Research Part B: Methodological*, 91:383–407, 2016. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2016.05.014. URL https://www.sciencedirect.com/science/article/pii/S0191261515301922.

Michael Ostrovsky and Michael Schwarz. Congestion pricing, carpooling, and commuter welfare. Technical report, National Bureau of Economic Research, 2025.

Richard Arnott, André de Palma, and Robin Lindsey. Information and time-of-usage decisions in the bottleneck model with stochastic capacity and demand. *European Economic Review*, 43(3):525–548, 1999. ISSN 0014-2921. doi: https://doi.org/10.1016/S0014-2921(98)00013-0. URL https://www.sciencedirect.com/science/article/pii/S0014292198000130.

Andre de Palma, Moshe Ben-Akiva, Claude Lefevre, and Nicolaos Litinas. Stochastic equilibrium model of peak period traffic congestion. *Transportation Science*, 17(4):430–453, 1983. doi: 10.1287/trsc.17.4.430. URL https://doi.org/10.1287/trsc.17.4.430.

Hani S. Mahmassani and Gang-Len Chang. On boundedly rational user equilibrium in transportation systems. *Transportation Science*, 21(2):89–99, 1987. doi: 10.1287/trsc.21.2.89. URL https://doi.org/10.1287/trsc.21.2.89.

Richard Arnott. A bathtub model of downtown traffic congestion. *Journal of Urban Economics*, 76:110–121, 2013. ISSN 0094-1190. doi: https://doi.org/10.1016/j.jue.2013.01.001. URL https://www.sciencedirect.com/science/article/pii/S0094119013000107.

Mogens Fosgerau and Kenneth A. Small. Hypercongestion in downtown metropolis. *Journal of Urban Economics*, 76:122–134, 2013. ISSN 0094-1190. doi: https://doi.org/10.1016/j.jue.2012.12.004. URL https://www.sciencedirect.com/science/article/pii/S0094119012000836.

Nikolas Geroliminis and Carlos F. Daganzo. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42(9):759–770, 2008. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2008.02.002. URL https://www.sciencedirect.com/science/article/pii/S0191261508000180.

Zhi-Chun Li, Hai-Jun Huang, and Hai Yang. Fifty years of the bottleneck model: A bibliometric review and future research directions. *Transportation Research Part B: Methodological*, 139:311–342, 2020. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2020.06.009. URL https://www.sciencedirect.com/science/article/pii/S0191261520303490.

Richard Arnott, André de Palma, and Robin Lindsey. Economics of a bottleneck. *Journal of Urban Economics*, 27(1):111–130, 1990b. ISSN 0094-1190. doi: https://doi.org/10.1016/0094-1190(90)90028-L. URL https://www.sciencedirect.com/science/article/pii/009411909090028L.

Xuehao Chu. Alternative congestion pricing schedules. *Regional Science and Urban Economics*, 29(6):697–722, 1999. ISSN 0166-0462. doi: https://doi.org/10.1016/S0166-0462(99)00021-6. URL https://www.sciencedirect.com/science/article/pii/S0166046299000216.

Richard Arnott, André de Palma, and Robin Lindsey. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *The American Economic Review*, 83(1):161–179, 1993a. ISSN 00028282. URL http://www.jstor.org/stable/2117502.

Ralph M. Braid. Uniform versus peak-load pricing of a bottleneck with elastic demand. *Journal of Urban Economics*, 26(3):320–327, 1989. ISSN 0094-1190. doi: https://doi.org/10.1016/0094-1190(89)90005-3. URL https://www.sciencedirect.com/science/article/pii/0094119089900053.

Chen-Hsiu Laih. Queueing at a bottleneck with single- and multi-step tolls. *Transportation Research Part A: Policy and Practice*, 28(3):197–208, 1994. ISSN 0965-8564. doi: https://doi.org/10.1016/0965-8564(94)90017-5. URL https://www.sciencedirect.com/science/article/pii/0965856494900175.

Ralph M. Braid. Peak-load pricing of a transportation route with an unpriced substitute. *Journal of Urban Economics*, 40(2):179–197, 1996. ISSN 0094-1190. doi: https://doi.org/10.1006/juec.1996.0028. URL https://www.sciencedirect.com/science/article/pii/S0094119096900285.

Richard Arnott, André de Palma, and Robin Lindsey. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *The American Economic Review*, 83(1):161–179, 1993b. ISSN 00028282. URL http://www.jstor.org/stable/2117502.

Eric J. Gonzales and Carlos F. Daganzo. Morning commute with competing modes and distributed demand: User equilibrium, system optimum, and pricing. *Transportation Research Part B: Methodological*, 46(10):1519–1534, 2012. ISSN 0191-2615. doi: 10.1016/j.trb.2012.07.009. URL https://www.sciencedirect.com/science/article/pii/S0191261512001026.

Andre De Palma and Robin Lindsey. Private toll roads: Competition under various ownership regimes. *The Annals of Regional Science*, 34(1):13–35, 2000.

André de Palma and Robin Lindsey. Private roads, competition, and incentives to adopt time-based congestion tolling. *Journal of Urban Economics*, 52(2):217–241, 2002. ISSN 0094-1190. doi: https://doi.org/10.1016/S0094-1190(02)00002-5. URL https://www.sciencedirect.com/science/article/pii/S0094119002000025.

Xinying Fu, Vincent A.C. van den Berg, and Erik T. Verhoef. Private road supply in networks with heterogeneous users. *Transportation Research Part A: Policy and Practice*, 118:430–443, 2018. ISSN 0965-8564. doi: https://doi.org/10.1016/j.tra.2018.09.025. URL https://www.sciencedirect.com/science/article/pii/S0965856417314490.

Erik T. Verhoef. Second-best congestion pricing in general static transportation networks with elastic demands. *Regional Science and Urban Economics*, 32(3):281–310, 2002. ISSN 0166-0462. doi: https://doi.org/10.1016/S0166-0462(00)00064-8. URL https://www.sciencedirect.com/science/article/pii/S0166046200000648.

Martine Labbé, Patrice Marcotte, and Gilles Savard. A bilevel model of taxation and its application to optimal highway pricing. *Management Science*, 44(12-part-1):1608–1622, 1998. doi: 10.1287/mnsc.44.12.1608. URL https://doi.org/10.1287/mnsc.44.12.1608.

Torbjörn Larsson and Michael Patriksson. *Side Constrained Traffic Equilibrium Models—Traffic Management Through Link Tolls*, pages 125–151. Springer US, Boston, MA, 1998. ISBN 978-1-4615-5757-9. doi: 10.1007/978-1-4615-5757-9_7. URL https://doi.org/10.1007/978-1-4615-5757-9_7.

Michael Patriksson and R. Tyrrell Rockafellar. A mathematical model and descent algorithm for bilevel traffic management. *Transportation Science*, 36(3):271–291, 2002. doi: 10.1287/trsc.36.3.271.7826. URL https://doi.org/10.1287/trsc.36.3.271.7826.

Xuan Di, Henry X. Liu, and Xuegang (Jeff) Ban. Second best toll pricing within the framework of bounded rationality. *Transportation Research Part B: Methodological*, 83:74–90, 2016. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2015.11.002. URL https://www.sciencedirect.com/science/article/pii/S0191261515002337.

Se il Mun, Ko ji Konishi, and Kazuhiro Yoshikawa. Optimal cordon pricing. *Journal of Urban Economics*, 54(1):21–38, 2003. ISSN 0094-1190. doi: https://doi.org/10.1016/S0094-1190(03)00022-6. URL https://www.sciencedirect.com/science/article/pii/S0094119003000226.

Xiaoning Zhang and Hai Yang. The optimal cordon-based network congestion pricing problem. *Transportation Research Part B: Methodological*, 38(6):517–537, 2004. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2003.08.001. URL https://www.sciencedirect.com/science/article/pii/S0191261503000900.

Jason D. Hartline and Tim Roughgarden. Simple versus optimal mechanisms. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, EC '09, page 225–234, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584584. doi: 10.1145/1566374.1566407. URL https://doi.org/10.1145/1566374.1566407.

Sergiu Hart and Noam Nisan. Approximate revenue maximization with multiple items. *Journal of Economic Theory*, 172:313–347, 2017. ISSN 0022-0531. doi: https://doi.org/10.1016/j.jet.2017.09.001. URL https://www.sciencedirect.com/science/article/pii/S0022053117300935.

Adam N. Elmachtoub and Jiaqi Shi. *The Power of Static Pricing for Reusable Resources*, page 357. Association for Computing Machinery, New York, NY, USA, 2025. ISBN 9798400719431. URL https://doi.org/10.1145/3736252.3742552.

Omar Besbes, Adam N. Elmachtoub, and Yunjie Sun. Technical note—static pricing: Universal guarantees for reusable resources. *Operations Research*, 70(2):1143–1152, 2022. doi: 10.1287/opre.2020.2054. URL https://doi.org/10.1287/opre.2020.2054.

Siddhartha Banerjee, Ramesh Johari, and Carlos Riquelme. Pricing in ride-sharing platforms: A queueing-theoretic approach. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, EC '15, page 639, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334105. doi: 10.1145/2764468.2764527. URL https://doi.org/10.1145/2764468.2764527.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37, 1961. ISSN 00221082, 15406261. URL http://www.jstor.org/stable/2977633.

Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981. doi: 10.1287/moor.6.1.58. URL https://doi.org/10.1287/moor.6.1.58.

Devansh Jalota, Michael Ostrovsky, and Marco Pavone. When simple is near optimal in security games. *CoRR*, abs/2402.11209, 2024.

Zhanhao Zhang, Ruifan Yang, and Manxi Wu. Designing high-occupancy toll lanes: A game-theoretic analysis. *CoRR*, abs/2408.01413, 2024. URL https://arxiv.org/abs/2408.01413.

Ningyuan Chen and Guillermo Gallego. Welfare analysis of dynamic pricing. *Management Science*, 65(1):139–151, 2019. doi: 10.1287/mnsc.2017.2943. URL https://doi.org/10.1287/mnsc.2017.2943.

Yukun Zhang and Qi Dong. Revenue vs. welfare: a comprehensive analysis of strategic trade-offs in online food delivery systems. *IET Conference Proceedings*, 2025:68–83, 2025. doi: 10.1049/icp.2025.2957. URL https://digital-library.theiet.org/doi/abs/10.1049/icp.2025.2957.

Chris Hendrickson and George Kocur. Schedule delay and departure time decisions in a deterministic model. *Transportation Science*, 15(1):62–77, 1981. doi: 10.1287/trsc.15.1.62. URL https://doi.org/10.1287/trsc.15.1.62.

Nikolas Geroliminis and David M. Levinson. *Cordon Pricing Consistent with the Physics of Overcrowding*, pages 219–240. Springer US, Boston, MA, 2009. ISBN 978-1-4419-0820-9. doi: 10.1007/978-1-4419-0820-9_11. URL https://doi.org/10.1007/978-1-4419-0820-9_11.

Bay Area Rapid Transit (BART). Ridership reports. https://www.bart.gov/about/reports/ridership, 2026. Accessed: 2026-01-09.

State of New York. Mta subway origin-destination ridership estimate: Beginning 2025. https://catalog.data.gov/dataset/mta-subway-origin-destination-ridership-estimate-beginning-2025, 2025a. Accessed: 2026-01-10.

California Department of Transportation. Performance measurement system (pems). https://pems.dot.ca.gov/, 2024. Accessed: Month Year.

State of New York. Mta congestion relief zone vehicle entries: Beginning 2025. https://data.ny.gov/Transportation/MTA-Congestion-Relief-Zone-Vehicle-Entries-Beginni/t6yz-b64h, 2025b. Accessed: 2026-01-10.

New York City Taxi and Limousine Commission. Tlc trip record data. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page, 2025. Accessed: 2026-01-10.

OpenStreetMap contributors. Openstreetmap. https://www.openstreetmap.org, 2025. Accessed: 2026-01-10.

Mark Wardman. Review and meta-analysis of u.k. time elasticities of travel demand. *Transportation*, 39(3):465–490, 2012. doi: 10.1007/s11116-011-9333-7.

Bruce D Greenshields. The potential use of aerial photographs in traffic analysis. In *Proc. 27th Annu. Meeting Highway Res. Board Held*, pages 291–297, 1947.

Metropolitan Transportation Commission. Travel time reliability and value of time in the san francisco bay area. https://mtc.ca.gov/sites/default/files/Current_Conditions.pdf, 2015. Accessed: 2026-01-09.

Kenneth A. Small. The scheduling of consumer activities: Work trips. *American Economic Review*, 72(3):467–479, 1982.

Bay Area Rapid Transit (BART). Bart gtfs schedule data. https://www.bart.gov/schedules/developers/gtfs, 2025a. Accessed: 2026-01-09.

Bay Area Rapid Transit (BART). Station access guidelines. https://www.bart.gov/sites/default/files/docs/access_guidelines.pdf, 2025b. Accessed: 2026-01-09.

Bay Area Rapid Transit. Fy26 adopted budget memo, 2025. URL https://www.bart.gov/sites/default/files/2025-09/FY26%20Adopted%20Budget%20Memo_FINAL_1.pdf. Accessed January 9, 2026.

Victor L. Knoop. *Traffic Flow Theory: An Introduction with Exercises*. TU Delft OPEN Publishing, 2021. ISBN 978-94-6366-378-6. URL https://www.victorknoop.eu/research/book/Knoop_Intro_traffic_flow_theory_edition3.pdf. Accessed: 2026-01-10.

U.S. Bureau of Labor Statistics. Occupational employment and wages in new york-newark-jersey city — may 2024. https://www.bls.gov/regions/northeast/news-release/occupationalemploymentandwages_newyork.htm, 2025. News release published April 29, 2025; Accessed: 2026-01-10.

New York City Comptroller. Subway frequencies: Left in the dark. https://comptroller.nyc.gov/wp-content/uploads/documents/Subway-Frequencies_Left-in-the-Dark.pdf, 2023. Accessed: 2026-01-10.

StreetEasy. Nyc subway access: Neighborhoods with the best and worst access. https://streeteasy.com/blog/nyc-subway-neighborhoods-best-worst-access/, 2018. Accessed: 2026-01-10.

A Arrival and Departure Time Profiles

A.1 Equilibrium Arrival and Departure Time Profiles

In the following, we depict the equilibrium arrival and departure time profiles in the bottleneck model without an outside option (Figure 7) and with a public transit outside option in a mixed-mode equilibrium where both car and transit are used (Figure 8) in the setting when $\mu < \lambda$. For more details on these equilibrium arrival and departure time profiles, see Gonzales and Daganzo (2012). Note that the waiting time profiles depicted in Figure 1 for car users is given by $w(t) = D(t) - A(t)$ for the bottleneck model without an outside option and is given by $w(t) = D_C(t) - A_C(t)$ for the bottleneck model with a public transit outside option. We note that in the bottleneck model, the arrival time represents the time at which users arrive at the point queue depicted in Figure 11 and the departure or cross time is the time at which they exit the point queue.

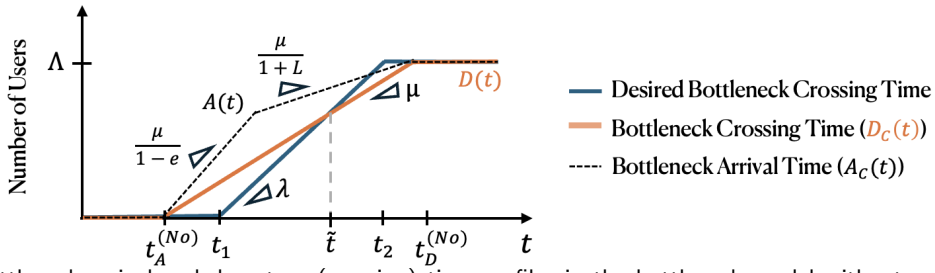


Figure 7: Bottleneck arrival and departure (crossing) time profiles in the bottleneck model without an outside option in the setting when $\mu < \lambda$. Here, e and L denote the normalized schedule delay penalties for early and late arrivals, respectively, and $[t_1, t_2]$ represents users' desired bottleneck crossing times. Moreover, $A(t)$ represents the arrival time distribution of the users at the bottleneck and $D(t)$ represents the number of car users that cross the bottleneck at their desired time. The curve in blue represents the desired bottleneck cross time of the users.

A.2 Arrival and Departure Time Profiles under Dynamic Revenue Optimal Toll

Figure 9 depicts the equilibrium bottleneck arrival and departure (crossing) time profiles in the bottleneck model with a public transit outside option under the dynamic revenue-optimal tolling scheme depicted in Figure 2 in the setting when $\mu < \lambda$ and $z_T \geq z_C$. Since there are no waiting delays under the dynamic revenue optimal tolling policy, the departure and arrival time profiles coincide, i.e., $w(t) = D_C(t) - A_C(t) = 0$ for all t.

Figure 10 depicts the equilibrium arrival and departure time profiles in an urban system governed by an MFD satisfying Properties 1 and 2 (see Section 5) under the dynamic revenue optimal tolling scheme in the setting when $\mu_f < \lambda$ and $z_T \geq z_C$. Since there are no waiting delays under the dynamic revenue optimal tolling policy, the departure and arrival time profiles are offset by a constant free-flow travel time of $t_f = \frac{D}{v_f}$, which corresponds to the time taken to travel through the urban system if all users traverse a fixed distance D.

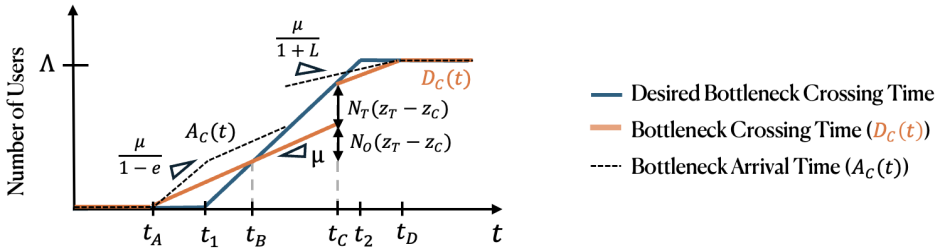


Figure 8: Bottleneck arrival and departure (crossing) time profiles for car users in the bottleneck model with a public transit outside option in a mixed-mode equilibrium where both car and transit are used in the setting when $\mu < \lambda$ and $z_T \geq z_C$. Here, e and L denote the normalized schedule delay penalties for early and late arrivals, respectively, and $[t_1, t_2]$ represents users' desired bottleneck crossing times. $A_C(t)$ represents the arrival time distribution of car users at the bottleneck and $D_C(t)$ represents the departure or crossing time of car users from the bottleneck. Moreover, $N_T(z_T - z_C)$ represents the number of transit users and $N_O(z_T - z_C)$ represents the number of car users that cross the bottleneck at their desired time. The curve in blue represents the desired bottleneck cross time of the users.

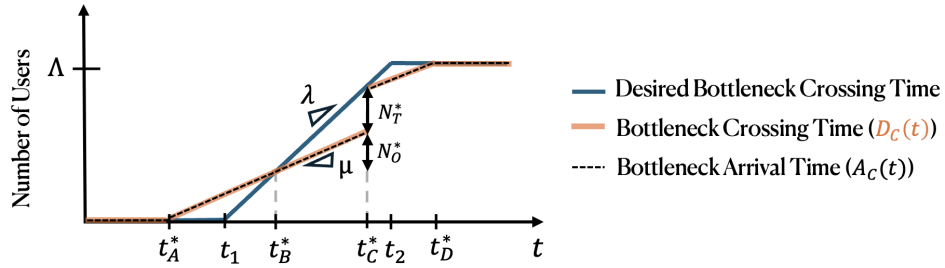


Figure 9: Bottleneck arrival and departure (crossing) time profiles for car users under the dynamic revenue-optimal toll in the bottleneck model with a public transit outside option in the setting when $\mu < \lambda$ and $z_T \geq z_C$. Here, $[t_1, t_2]$ represents users' desired bottleneck crossing times and the curve in blue represents the desired bottleneck cross time of the users. Since waiting or queuing delays are eliminated under the dynamic revenue-optimal toll, both arrival and departure time profiles for car users coincide, i.e., $D_C(t) = A_C(t)$. Moreover, N_T^* represents the number of transit users and N_O^* represents the number of car users that cross the bottleneck at their desired time.

Note that the bottleneck model does not inherently include a notion of distance—hence the arrival and departure curves coincide in that setting. However, distance can be incorporated by modeling the bottleneck as comprising an uncongested physical segment of length D followed by a point queue (see Figure 11). In this case, the arrival and departure time profiles in the bottleneck framework will look akin to that depicted in Figure 10.

B Formal Definitions of Revenue and Total System Cost Metrics

We evaluate tolling policies along two performance metrics of direct relevance to a central planner: (i) cumulative toll revenue and (ii) total system cost, defined below.

Revenue: The cumulative toll revenue is defined as the total toll payments collected from all users at the equilibrium induced by a tolling policy $\tau(\cdot)$. Under such a policy, let $t^{\tau(\cdot)}(t^*)$ denote the time at which a user with a desired bottleneck departure time t^* passes the bottleneck under $\tau(\cdot)$ at equilibrium, and let $x_C^{\tau(\cdot)}(t^*) \in [0,1]$ denote the equilibrium fraction of users of type t^* who travel by car (with the remaining fraction traveling by transit). Then, the cumulative toll revenue under $\tau(\cdot)$, normalized by c_W, is given by:

$$R(\tau(\cdot)) := \int_{t_1}^{t_2} \lambda \cdot x_C^{\tau(\cdot)}(t^*) \cdot \tau(t^{\tau(\cdot)}(t^*)) dt^*.$$

While the above expression for the revenue is in terms of users' desired bottleneck departure times t^*, it is

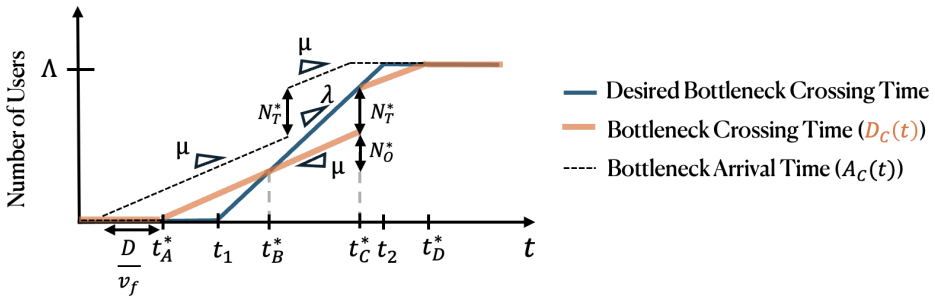


Figure 10: Arrival and departure time profiles for car users in an urban system under the dynamic revenue-optimal toll in the MFD model in the setting when $\mu < \lambda$ and $z_T \geq z_C$. Here, the curve in blue represents users' desired departure time profiles from the urban system (or arrival times at the destination). Since waiting or queuing delays are eliminated under the dynamic revenue-optimal toll, both arrival and departure time profiles for car users are offset by a constant free-flow travel time of $t_f = \frac{D}{v_f}$, which corresponds to the time taken to travel through the urban system if all users traverse a fixed distance D. Moreover, N_T^* represents the number of transit users and N_O^* represents the number of car users that cross the urban system at their desired time.

often more convenient to express the above integral in the time at which users actually pass the bottleneck. Accordingly, let $[t_A^{\tau(\cdot)}, t_D^{\tau(\cdot)}]$ denote the equilibrium interval over which users pass the bottleneck under the tolling policy $\tau(\cdot)$, analogous to the no-toll equilibrium interval $[t_A, t_D]$ shown in the right of Figure 1. Then, in the case $z_T \geq z_C$, the above revenue expression simplifies to:

$$R(\tau(\cdot)) = \int_{t_A^{\tau(\cdot)}}^{t_D^{\tau(\cdot)}} \min\{\mu, \lambda\}\tau(t)dt$$

as the rate at which users can be processed at the bottleneck is the minimum of the user arrival rate and the bottleneck service rate. Note that when $z_T < z_C$, all users strictly prefer transit, so $x_C^{\tau(\cdot)}(t^*) = 0$ for all t^* and no toll revenue is generated. In the remainder of this work, for brevity of notation, we drop the superscript $\tau(\cdot)$ in the notation for the equilibrium interval $[t_A^{\tau(\cdot)}, t_D^{\tau(\cdot)}]$ when it is clear from context. Furthermore, in studying equilibrium outcomes, when users are indifferent between car and transit, without loss of generality, we break ties in favor of higher revenue outcomes.

Total System Cost: We define the total system cost as the sum (across all users) of waiting costs, schedule delay costs, and generalized travel costs associated with car or transit use at the equilibrium induced by a tolling policy $\tau(\cdot)$. Equivalently, the total system cost is the sum of users' total travel costs at equilibrium *minus* their toll payments, and is given by:

$$SC(\tau(\cdot)) = \int_{t_1}^{t_2} \lambda \left[x_C^{\tau(\cdot)}(t^*) \Big(c\big(t^{\tau(\cdot)}(t^*), t^*\big) - \tau\big(t^{\tau(\cdot)}(t^*)\big) \Big) + \big(1 - x_C^{\tau(\cdot)}(t^*)\big) z_T \right] dt^*,$$

where $c(t, t^\star)$ denotes the travel cost (in time units) for a car user from Equation (2). The above expression for the total system cost subtracts toll payments from the user costs, since tolls represent transfers between users and the central planner, consistent with standard practice in congestion pricing (Vickrey 1969, Gonzales and Daganzo 2012) and economics (Vickrey 1961) that excludes tolls (or prices) when defining system cost or welfare metrics.

C Conceptual Relationship Between Bottleneck and MFD Models

This section describes a high-level conceptual connection between the bottleneck and MFD models, illustrating how Vickrey's bottleneck model can be interpreted as a special case of the MFD model. Figure 11 depicts a bottleneck system as an uncongestible physical section followed by a downstream point queue,

and shows the resulting MFD-like flow–accumulation relationship. Akin to the triangular MFD in Figure 4, throughput in the bottleneck setting increases linearly with vehicle accumulation up to the critical threshold n_c. Beyond n_c, the throughput remains at the bottleneck capacity μ rather than declining in the congested regime as in the case of the triangular MFD.

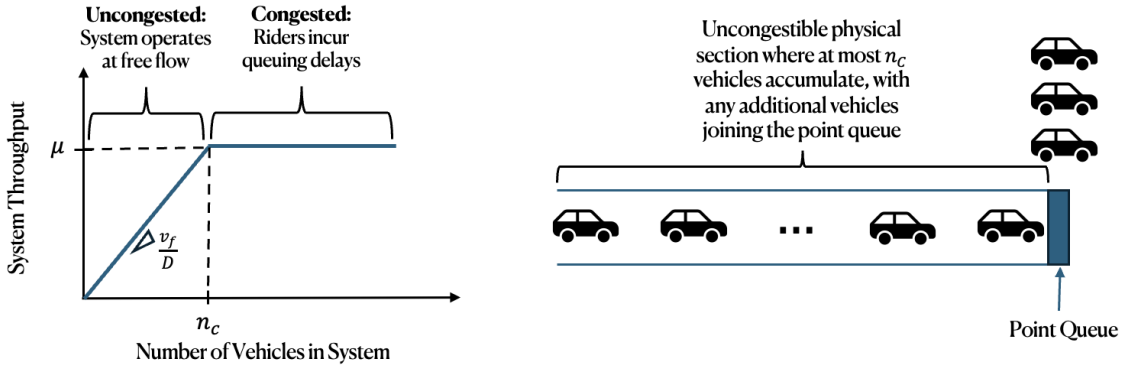


Figure 11: Bottleneck system with a point queue and its implied MFD-like relation. The bottleneck system can be represented via an uncongestible physical section (e.g., a bridge of length D) followed by a downstream point queue that occupies no physical space and captures congestion and queuing delays in the bottleneck. The physical section can hold at most n_c vehicles and any additional vehicles join the point queue. In this system, when the total vehicle accumulation (i.e., the sum of the vehicles in the physical section and the point queue) satisfies $n \leq n_c$, the system operates in free flow with a speed v_f. Beyond this accumulation threshold n_c, the system throughput remains at the bottleneck capacity μ and vehicles incur queuing delays. Note that the slope of the segment connecting $(0,0)$ to any point on the curve determines the average system speed, equal to trip distance D times the slope, which declines when the vehicle accumulation exceeds n_c.

D Proofs

D.1 Proof of Theorem 2

We establish this result by analyzing three regimes for $z_T - z_C$, each corresponding to a distinct revenue expression under the static revenue-optimal toll characterized in Proposition 2:

$$R(\tau_s^*) = \begin{cases} (z_T - z_C)\Lambda\frac{\mu}{\lambda}, & \text{if } 0 \leq z_T - z_C < \frac{\Lambda eL}{(\lambda-\mu)(e+L)} \\ \frac{\mu\, eL}{4\,(e+L)\left(1-\frac{\mu}{\lambda}\right)}\left[\frac{\Lambda}{\lambda} + \frac{e+L}{eL}\left(1-\frac{\mu}{\lambda}\right)(z_T - z_C)\right]^2, & \text{if } z_T - z_C \in \left[\frac{\Lambda eL}{(\lambda-\mu)(e+L)}, \frac{\Lambda eL}{(e+L)}\left(\frac{1}{\lambda-\mu}+\frac{2}{\mu}\right)\right] \\ (z_T - z_C)\Lambda - \frac{\Lambda^2 eL}{\mu(e+L)}, & \text{if } z_T - z_C > \frac{\Lambda eL}{(e+L)}\left(\frac{1}{\lambda-\mu}+\frac{2}{\mu}\right). \end{cases}$$

Case i $\left(z_T - z_C < \frac{\Lambda eL}{(\lambda-\mu)(e+L)}\right)$**:** In this case, note that $z_T - z_C < \frac{\Lambda eL}{(\lambda-\mu)(e+L)}\frac{\lambda}{\mu}$ as $\lambda > \mu$; hence, from Theorem 1, the revenue of the revenue-optimal dynamic toll is $(z_T - z_C)\Lambda\frac{\mu}{\lambda} + \frac{(z_T - z_C)^2\mu(e+L)}{2eL}\left(1-\frac{\mu}{\lambda}\right)^2$. Moreover, let $s \geq 1$ be a constant such that $s(z_T - z_C) = \frac{\Lambda eL}{(\lambda-\mu)(e+L)}$. Then, we have:

$$\begin{aligned} R(\tau_d^*(\cdot)) &= (z_T - z_C)\Lambda\frac{\mu}{\lambda} + \frac{(z_T - z_C)^2\mu(e+L)}{2eL}\left(1-\frac{\mu}{\lambda}\right)^2 \overset{(a)}{=} R(\tau_s^*) + \frac{(z_T - z_C)^2\mu(e+L)}{2eL}\left(1-\frac{\mu}{\lambda}\right)^2, \\ &\overset{(b)}{=} R(\tau_s^*) + (z_T - z_C)\frac{\Lambda eL}{s(\lambda-\mu)(e+L)}\frac{\mu(e+L)}{2eL}\left(\frac{\lambda-\mu}{\lambda}\right)^2 = R(\tau_s^*) + \frac{1}{2s}(z_T - z_C)\Lambda\frac{\mu}{\lambda}\left(1-\frac{\mu}{\lambda}\right), \\ &= R(\tau_s^*) + R(\tau_s^*)\frac{1}{2s}\left(1-\frac{\mu}{\lambda}\right) = R(\tau_s^*)\left(1 + \frac{1}{2s}\left(1-\frac{\mu}{\lambda}\right)\right) \end{aligned}$$

where (a) follows as $R(\tau_s^*) = (z_T - z_C)\Lambda\frac{\mu}{\lambda}$ and (b) follows as $s(z_T - z_C) = \frac{\Lambda eL}{(\lambda-\mu)(e+L)}$. Thus, the optimal static toll achieves at least $\frac{2s}{2s+1-\frac{\mu}{\lambda}} \geq \frac{2}{3-\frac{\mu}{\lambda}}$ fraction of the optimal dynamic revenue.

Case ii $\left(z_T - z_C \in \left[\frac{\Lambda eL}{(\lambda-\mu)(e+L)}, \frac{\Lambda eL}{(e+L)}\left(\frac{1}{\lambda-\mu} + \frac{2}{\mu}\right)\right]\right)$**:** In this case, letting $s \leq 1$ be a constant such that $s(z_T - z_C) = \frac{\Lambda eL}{(\lambda-\mu)(e+L)}$, we have

$$\begin{aligned}
R(\tau_s^*) &= \frac{\mu eL}{4(e+L)\left(1-\frac{\mu}{\lambda}\right)}\left[\frac{\Lambda}{\lambda} + \frac{e+L}{eL}\left(1-\frac{\mu}{\lambda}\right)(z_T - z_C)\right]^2, \\
&\overset{(a)}{=} \frac{\Lambda s}{4}\frac{\mu}{\lambda}(z_T - z_C) + \frac{\Lambda}{2}\frac{\mu}{\lambda}(z_T - z_C) + \frac{(z_T - z_C)^2\mu(e+L)}{4eL}\left(1-\frac{\mu}{\lambda}\right), \\
&\geq \min\left\{\frac{1}{2} + \frac{s}{4}, \frac{1}{2(1-\frac{\mu}{\lambda})}\right\}\left[\Lambda\frac{\mu}{\lambda}(z_T - z_C) + \frac{(z_T - z_C)^2\mu(e+L)}{2eL}\left(1-\frac{\mu}{\lambda}\right)^2\right] \\
&\overset{(b)}{\geq} \min\left\{\frac{2+s}{4}, \frac{1}{2(1-\frac{\mu}{\lambda})}\right\} R(\tau_d^*(\cdot)),
\end{aligned}$$

where (a) follows as $s(z_T - z_C) = \frac{\Lambda eL}{(\lambda-\mu)(e+L)}$ and (b) follows as the term in the square brackets is the unconstrained optimal objective of the dynamic revenue optimization Problem (7). Thus, optimal static tolls achieve at least $\min\left\{\frac{2+s}{4}, \frac{1}{2(1-\frac{\mu}{\lambda})}\right\}$ fraction of the optimal dynamic revenue.

Case iii $\left(z_T - z_C > \frac{\Lambda eL}{(e+L)}\left(\frac{1}{\lambda-\mu} + \frac{2}{\mu}\right)\right)$**:** In this case, $\tau_s^* = z_T - z_C - \frac{\Lambda eL}{\mu(e+L)}$, which implies:

$$\begin{aligned}
&z_T - z_C - \frac{\Lambda eL}{\mu(e+L)} > \frac{z_T - z_C}{2} + \frac{\Lambda eL}{2(\lambda-\mu)(e+L)}, \\
&\implies z_T - z_C > \frac{\Lambda eL}{(\lambda-\mu)(e+L)} + 2\frac{\Lambda eL}{\mu(e+L)} = \frac{\Lambda eL}{e+L}\left(\frac{1}{\lambda-\mu} + \frac{2}{\mu}\right) > \frac{\Lambda eL}{e+L}\left(\frac{1}{\lambda-\mu} + \frac{1}{\mu}\right) = \frac{\Lambda eL}{(\lambda-\mu)(e+L)}\frac{\lambda}{\mu}
\end{aligned}$$

When $z_T - z_C > \frac{\Lambda eL}{(\lambda-\mu)(e+L)}\frac{\lambda}{\mu}$, the dynamic optimal revenue is given by $(z_T - z_C)\Lambda - \frac{\Lambda^2}{2\mu}\frac{eL}{e+L}$. Then:

$$\begin{aligned}
\frac{R(\tau_s^*)}{R_d^*} &= \frac{(z_T - z_C)\Lambda - \frac{\Lambda^2}{\mu}\frac{eL}{e+L}}{(z_T - z_C)\Lambda - \frac{\Lambda^2}{2\mu}\frac{eL}{e+L}} = 1 - \frac{\frac{\Lambda^2}{2\mu}\frac{eL}{e+L}}{(z_T - z_C)\Lambda - \frac{\Lambda^2}{2\mu}\frac{eL}{e+L}} = 1 - \frac{1}{\frac{(z_T - z_C)\Lambda}{\frac{\Lambda^2}{2\mu}\frac{eL}{e+L}} - 1} = 1 - \frac{1}{\frac{(z_T - z_C)(e+L)}{\Lambda eL}(2\mu) - 1} \\
&\overset{(a)}{\geq} 1 - \frac{1}{\left(\frac{1}{\lambda-\mu} + \frac{2}{\mu}\right)(2\mu) - 1} = 1 - \frac{1}{\frac{2(2\lambda-\mu)}{\lambda-\mu} - 1} = 1 - \frac{\lambda-\mu}{3\lambda-\mu} \geq 1 - \frac{\lambda-\mu}{3(\lambda-\mu)} = \frac{2}{3}
\end{aligned}$$

where (a) follows as $z_T - z_C > \frac{\Lambda eL}{(\lambda-\mu)(e+L)}\frac{\lambda}{\mu}$. This establishes our claim.

D.2 Proof of Theorem 3

To prove this claim, we first derive expressions for the total system cost under the static and dynamic revenue-optimal tolling policies as well as the minimum achievable system cost SC^*.

Total System Cost of Dynamic Revenue-Optimal Tolls: Under the dynamic revenue-optimal policy, recall from the proof of Theorem 1 that there are no queuing delays. Letting, N_e be the number of users that pass the bottleneck early (i.e., between $[t_A^*, t_B^*]$ in Figure 2) and N_L be the number of users that pass the bottleneck late (i.e., between $[t_C^*, t_D^*]$ in Figure 2), the total system cost under the dynamic revenue-optimal policy for the fraction f^* defined in Theorem 1 is given by:

$$SC(\tau_d^*(\cdot)) = z_T f^*\Lambda\left(1-\frac{\mu}{\lambda}\right) + z_C\left(f^*\Lambda\frac{\mu}{\lambda} + (1-f^*)\Lambda\right) + eN_e\frac{t_1 - t_A^*}{2} + LN_L\frac{t_D^* - t_2}{2},$$

which corresponds to the sum of (i) the cost of using transit, where $f^*\Lambda\left(1-\frac{\mu}{\lambda}\right)$ users avail transit, (ii) the free-flow cost of using a car, where $f^*\Lambda\frac{\mu}{\lambda}+(1-f^*)\Lambda$ users avail a car, and (iii) schedule delay costs. Then, noting that $N_e=\frac{L}{e+L}(1-f^*)\Lambda$, $N_L=\frac{e}{e+L}(1-f^*)\Lambda$, and substituting the corresponding relations for $t_1-t_A^*$ and $t_D^*-t_2$ based on the analysis in the proof of Theorem 1, we obtain the following relation for the total system cost of the revenue-optimal dynamic tolling policy:

$$SC(\tau_d^*(\cdot))=z_Tf^*\Lambda\left(1-\frac{\mu}{\lambda}\right)+z_C\left(f^*\Lambda\frac{\mu}{\lambda}+(1-f^*)\Lambda\right)+\frac{\Lambda^2eL}{2\mu(e+L)}(1-f^*)^2\left(1-\frac{\mu}{\lambda}\right).$$

In the regime $z_T-z_C\leq\frac{\Lambda eL}{\mu(e+L)}$, we have $z_T-z_C\leq\frac{\Lambda eL}{\mu(e+L)}\frac{\lambda}{\lambda-\mu}$, and therefore Theorem 1 implies that $f^*=1-\frac{(z_T-z_C)\mu(e+L)}{\Lambda eL}\left(1-\frac{\mu}{\lambda}\right)$, yielding the following expression for the total system cost:

$$SC(\tau_d^*(\cdot))=z_C\Lambda\frac{\mu}{\lambda}+z_T\left(1-\frac{\mu}{\lambda}\right)\Lambda-\frac{(z_T-z_C)^2\mu(e+L)}{2eL}\left(1-\frac{\mu}{\lambda}\right)^3. \tag{9}$$

Total System Cost of Static Revenue-Optimal Tolls: For a static toll τ, we show that the total system cost as a function of the toll τ is given by (see Appendix F.1 for a derivation):

$$\begin{aligned}SC(\tau)=&z_T\left[1-\frac{z_T-z_C-\tau}{\frac{\Lambda eL}{\mu(e+L)}}\right]\Lambda\left(1-\frac{\mu}{\lambda}\right)+z_C\left[\mu(z_T-z_C-\tau)\left(\frac{1}{e}+\frac{1}{L}\right)+\left(1-\frac{z_T-z_C-\tau}{\frac{\Lambda eL}{\mu(e+L)}}\right)\frac{\Lambda\mu}{\lambda}\right]\\&+\frac{\mu(z_T-z_C-\tau)^2}{2}\left(1-\frac{\mu}{\lambda}\right)\left(\frac{1}{e}+\frac{1}{L}\right)+(z_T-z_C-\tau)\left(\left(1-\frac{z_T-z_C-\tau}{\frac{\Lambda eL}{\mu(e+L)}}\right)\Lambda\frac{\mu}{\lambda}+\frac{\mu(z_T-z_C-\tau)}{2}\left(\frac{1}{e}+\frac{1}{L}\right)\right),\end{aligned} \tag{10}$$

which corresponds to the sum of the (i) cost of using transit, (ii) the free-flow cost of using a car, (iii) schedule delay costs, and (iv) waiting time delays. From this expression, in the regime $z_T-z_C\leq\frac{\Lambda eL}{\mu(e+L)}$, we derive the total system cost under the two feasible static revenue-optimal tolls in Equation (3): (i) $\tau_s^*=z_T-z_C$ and (ii) $\tau_s^*=\frac{z_T-z_C}{2}+\frac{\Lambda eL}{2(\lambda-\mu)(e+L)}$. Specifically, we obtain:

$$SC(\tau_s^*)=\begin{cases}z_C\Lambda\frac{\mu}{\lambda}+z_T\left(1-\frac{\mu}{\lambda}\right)\Lambda, & \text{if } \tau_s^*=z_T-z_C\\ z_C\Lambda\frac{\mu}{\lambda}+z_T\left(1-\frac{\mu}{\lambda}\right)\Lambda-\frac{\mu(2\lambda-\mu)}{8\lambda eL(e+L)(\lambda-\mu)^2}\left((z_T-z_C)(e+L)(\lambda-\mu)-\Lambda eL\right)^2, & \text{if } \tau_s^*=\frac{z_T-z_C}{2}+\frac{\Lambda eL}{2(\lambda-\mu)(e+L)}\end{cases}$$

We note that the third candidate revenue-optimal toll, $\tau_s^*=(z_T-z_C)-\frac{\Lambda eL}{\mu(e+L)}$ (see Proposition 2), is non-positive in the parameter regime of interest and thus is not relevant for our analysis here.

Optimal System Cost: From Gonzales and Daganzo (2012), the minimum achievable system cost is $SC^*=z_C\Lambda+\left(1-\frac{\mu}{\lambda}\right)\Lambda(z_T-z_C)-\left(1-\frac{\mu}{\lambda}\right)\frac{\mu(e+L)}{2eL}(z_T-z_C)^2$ in the regime when $0<z_T-z_C\leq\frac{\Lambda eL}{\mu(e+L)}$.

System Cost Comparisons: From the above derived relations for the total system cost under static and dynamic revenue-optimal tolls, it follows that $SC(\tau_s^*),SC(\tau_d^*(\cdot))\leq z_C\Lambda\frac{\mu}{\lambda}+z_T\left(1-\frac{\mu}{\lambda}\right)\Lambda$. Then, we obtain the following relation for the ratio of the system costs under these two policies (which we present here for the static revenue-optimal policy):

$$\begin{aligned}\frac{SC(\tau_s^*)}{SC^*}&\leq\frac{z_C\Lambda\frac{\mu}{\lambda}+z_T\left(1-\frac{\mu}{\lambda}\right)\Lambda}{z_C\Lambda+\left(1-\frac{\mu}{\lambda}\right)\Lambda(z_T-z_C)-\left(1-\frac{\mu}{\lambda}\right)\frac{\mu(e+L)}{2eL}(z_T-z_C)^2},\\&=1+\frac{\left(1-\frac{\mu}{\lambda}\right)\frac{\mu(e+L)}{2eL}(z_T-z_C)^2}{z_T\Lambda-(z_T-z_C)\frac{\mu}{\lambda}\Lambda-(z_T-z_C)\frac{\Lambda}{2}\left(1-\frac{\mu}{\lambda}\right)},\\&\overset{(b)}{\leq}1+\frac{(z_T-z_C)\frac{\Lambda}{2}\left(1-\frac{\mu}{\lambda}\right)}{z_T\Lambda-(z_T-z_C)\frac{\mu}{\lambda}\Lambda-(z_T-z_C)\frac{\Lambda}{2}\left(1-\frac{\mu}{\lambda}\right)},\\&\overset{(c)}{\leq}1+\frac{\frac{(z_T-z_C)}{2}\left(1-\frac{\mu}{\lambda}\right)}{\frac{(z_T-z_C)}{2}\left(1-\frac{\mu}{\lambda}\right)},\\&=2,\end{aligned}$$

where (b) follows as $z_T - z_C < \frac{\Lambda e L}{\mu(e+L)}$, and (c) follows by subtracting $z_C\Lambda$ in the denominator and simplifying. These inequalities establish our desired system cost ratio for static revenue-optimal tolling. Note that the same set of inequalities applies for the dynamic revenue-optimal tolling policy, as its total system cost also satisfies $SC(\tau_s^*), SC(\tau_d^*(\cdot)) \leq z_C\Lambda\frac{\mu}{\lambda} + z_T\left(1-\frac{\mu}{\lambda}\right)\Lambda$. This establishes our claim.

D.3 Proof of Theorem 4

First, as with the bottleneck model, a necessary equilibrium condition is that $w'(t) + \tau'(t) = e$ for early arrivals and $w'(t) + \tau'(t) = -L$ for late arrivals, and that $w(t) + \tau(t) \leq z_T - z_C$, i.e., the sum of the waiting and toll costs are as depicted in Figure 3. Note here that the waiting time $w(t)$ can be expressed as a function of the number of users $n(t)$ in the system at time t, i.e., $w(t) = w(n(t))$, and the exact times t'_A, t'_B, t'_C, t'_D shown in Figure 3 are determined endogenously by the tolling policy. We now show that even for urban systems governed by an MFD, the dynamic revenue-optimal tolls correspond to operating on the throughput-maximizing point of the MFD where the waiting times are set point-wise to zero.

To establish this claim, we consider two tolling policies corresponding to possibly different values of the endogenously determined time points t'_A, t'_B, t'_C, t'_D in Figure 3. First, consider the dynamic revenue-optimal policy $\tau^*(\cdot)$ with associated times $t^*_A, t^*_B, t^*_C, t^*_D$ in Figure 3. Moreover, let f be the fraction of time corresponding to the horizontal portion of the curve in Figure 3 under the optimal dynamic tolling policy. Additionally, let f_1 be the fraction of users that seek to arrive in the interval $[t^*_A, t_1]$ and f_2 be the fraction of users that seek to arrive in the interval $[t^*_C, t_2]$, where $f_1 + f_2 = 1 - f$.

We define the second tolling policy as $\tilde{\tau}(\cdot)$, with no waiting time delays (i.e., $w(t) = 0$ for all t and the system operates at maximum throughput μ_f) and $\tilde{\tau}(t) = z_T - z_C$ for the period $[t^*_B, t^*_C]$. In this case, we define the associated times in Figure 3 as $\tilde{t}_A, \tilde{t}_B, \tilde{t}_C, \tilde{t}_D$, where $\tilde{t}_B = t^*_B$ and $\tilde{t}_C = t^*_C$.

Then, letting the system throughput $\mu(t) = \mu(n(t))$ be time-varying and depend on the vehicle accumulation in the system, the optimal revenue is given by:

$$R^* = \int_{t^*_A}^{t^*_D} \mu(n(t))\tau^*(t)dt = \int_{t^*_A}^{t^*_B} \mu(n(t))\tau^*(t)dt + \int_{t^*_B}^{t^*_C} \mu(n(t))\tau^*(t)dt + \int_{t^*_C}^{t^*_D} \mu(n(t))\tau^*(t)dt,$$

and the revenue corresponding to the second policy is given by:

$$\tilde{R} = \int_{\tilde{t}_A}^{\tilde{t}_D} \mu_f\tilde{\tau}(t)dt = \int_{\tilde{t}_A}^{t^*_B} \mu_f\tilde{\tau}(t)dt + \int_{t^*_B}^{t^*_C} \mu_f\tilde{\tau}(t)dt + \int_{t^*_C}^{\tilde{t}_D} \mu_f\tilde{\tau}(t)dt.$$

To establish our claim, we show $\tilde{R} \geq R^*$ by proving three inequalities: (i) $\int_{t^*_B}^{t^*_C} \mu(n(t))\tau^*(t)dt \leq \int_{t^*_B}^{t^*_C} \mu_f\tilde{\tau}(t)dt$, (ii) $\int_{t^*_A}^{t^*_B} \mu(n(t))\tau^*(t)dt \leq \int_{\tilde{t}_A}^{t^*_B} \mu_f\tilde{\tau}(t)dt$, and (iii) $\int_{t^*_C}^{t^*_D} \mu(n(t))\tau^*(t)dt \leq \int_{t^*_C}^{\tilde{t}_D} \mu_f\tilde{\tau}(t)dt$.

Proof of (i): Note that $\mu(n(t))\tau^*(t) \leq \mu_f(z_T - z_C)$ for all t, as $\mu(n(t)) \leq \mu_f$ and $w(n(t)) + \tau^*(t) \leq z_T - z_C$ for all t. Then, we obtain:

$$\int_{t^*_B}^{t^*_C} \mu(n(t))\tau^*(t)dt \leq \int_{t^*_B}^{t^*_C} \mu_f(z_T - z_C)dt = \mu_f(z_T - z_C)(t^*_C - t^*_B) = \mu_f(z_T - z_C)f\frac{\Lambda}{\lambda} = \mu_f\int_{t^*_B}^{t^*_C} \tilde{\tau}(t)dt.$$

Proof of (ii): First note that the revenue under the policy $\tilde{\tau}(\cdot)$ in the range $[\tilde{t}_A, t^*_B]$ is given by:

$$\begin{aligned}\int_{\tilde{t}_A}^{t^*_B} \mu_f\tilde{\tau}(t)dt &\overset{(a)}{=} \int_0^{t^*_B - \tilde{t}_A} \mu_f(\tilde{\tau}(t^*_B) - e\Delta)d\Delta = \mu_f\left[\tilde{\tau}(t^*_B)(t^*_B - \tilde{t}_A) - \frac{e(t^*_B - \tilde{t}_A)^2}{2}\right], \\ &= \mu_f\left[\tilde{\tau}(t^*_B) - \frac{e(t^*_B - \tilde{t}_A)}{2}\right](t^*_B - \tilde{t}_A) \overset{(b)}{=} \left[z_T - z_C - \frac{e(t^*_B - \tilde{t}_A)}{2}\right] f_1\Lambda, \end{aligned} \tag{11}$$

where (a) follows from the structure of the tolling policy in Figure 3 and (b) follows as $f_1\Lambda = \lambda(t_B^* - t_1) = \mu_f(t_B^* - \tilde{t}_A)$. Then, to establish (ii), we use a variable transformation $\Delta = t_B^* - t$ to get:

$$\begin{aligned}\int_{t_A^*}^{t_B^*} \mu(t)\tau^*(t)dt &= \int_0^{t_B^*-t_A^*} \mu(\Delta)\tau^*(\Delta)d\Delta = \int_0^{t_B^*-t_A^*} \mu(\Delta)(z_T - z_C - e\Delta - w(\Delta))d\Delta, \\ &\overset{(a)}{\leq} \int_0^{t_B^*-t_A^*} \mu(\Delta)(z_T - z_C - e\Delta)d\Delta, \\ &= \int_0^{t_B^*-\tilde{t}_A} \mu(\Delta)(z_T - z_C - e\Delta)d\Delta + \int_{t_B^*-\tilde{t}_A}^{t_B^*-t_A^*} \mu(\Delta)(z_T - z_C - e\Delta)d\Delta, \end{aligned} \tag{12}$$

where (a) follows as $w(\Delta) \geq 0$ for all $\Delta \in [t_A^*, t_B^*]$. Furthermore, for some $g \in [0,1]$, we define:

$$\int_0^{t_B^*-\tilde{t}_A} \mu(\Delta)d\Delta = gf_1\Lambda, \quad \int_{t_B^*-\tilde{t}_A}^{t_B^*-t_A^*} \mu(\Delta)d\Delta = (1-g)f_1\Lambda.$$

We use these expressions to upper bound the two terms in Equation (12). To this end, note that:

$$\int_{t_B^*-\tilde{t}_A}^{t_B^*-t_A^*} \mu(\Delta)(z_T - z_C - e\Delta)d\Delta \leq (1-g)f_1\Lambda(z_T - z_C - e(t_B^* - \tilde{t}_A)), \tag{13}$$

which follows as $\Delta \geq t_B^* - \tilde{t}_A$ for all $\Delta \in [t_B^* - \tilde{t}_A, t_B^* - t_A^*]$. Next, for the other term in Equation (12):

$$\begin{aligned}\int_0^{t_B^*-\tilde{t}_A} \mu(\Delta)(z_T - z_C - e\Delta)d\Delta &= \int_0^{t_B^*-\tilde{t}_A} \mu(\Delta)(z_T - z_C)d\Delta - e\int_0^{t_B^*-\tilde{t}_A} \mu(\Delta)\Delta d\Delta, \\ &\overset{(a)}{=} gf_1\Lambda(z_T - z_C) - e\int_0^{t_B^*-\tilde{t}_A} \mu(\Delta)\Delta d\Delta, \end{aligned} \tag{14}$$

where (a) follows as $\int_0^{t_B^*-\tilde{t}_A} \mu(\Delta)d\Delta = gf_1\Lambda$ and $z_T - z_C$ is a constant.

Next, from Equations (13) and (14), we obtain the following upper bound on the revenue:

$$\begin{aligned}\int_{t_A^*}^{t_B^*} \mu(t)\tau^*(t)dt &\leq gf_1\Lambda(z_T - z_C) - e\int_0^{t_B^*-\tilde{t}_A} \mu(\Delta)\Delta d\Delta + (1-g)f_1\Lambda(z_T - z_C - e(t_B^* - \tilde{t}_A)), \\ &= f_1\Lambda(z_T - z_C) - (1-g)f_1\Lambda e(t_B^* - \tilde{t}_A) - e\int_0^{t_B^*-\tilde{t}_A} \mu(\Delta)\Delta d\Delta. \end{aligned} \tag{15}$$

Now, to bound the term $\int_0^{t_B^*-\tilde{t}_A} \mu(\Delta)\Delta d\Delta$, we define $F(\Delta) = \int_0^{\Delta} \mu(x)dx$. Note that F is absolutely continuous with $F'(\Delta) = \mu(\Delta)$. Consequently, using integration by parts, it follows that:

$$\begin{aligned}\int_0^{t_B^*-\tilde{t}_A} \mu(\Delta)\Delta d\Delta &= \int_0^{t_B^*-\tilde{t}_A} F'(\Delta)\Delta d\Delta = [\Delta F(\Delta)]_0^{t_B^*-\tilde{t}_A} - \int_0^{t_B^*-\tilde{t}_A} F(\Delta)d\Delta, \\ &= (t_B^* - \tilde{t}_A)\int_0^{t_B^*-\tilde{t}_A} \mu(\Delta)d\Delta - \int_0^{t_B^*-\tilde{t}_A} \left(\int_0^{\Delta} \mu(x)dx\right) d\Delta, \\ &= (t_B^* - \tilde{t}_A)gf_1\Lambda - \int_0^{t_B^*-\tilde{t}_A} \left(\int_0^{\Delta} \mu(x)dx\right) d\Delta. \end{aligned} \tag{16}$$

Next, substituting Equation (16) in Equation (15), we get:

$$\begin{aligned}
\int_{t_A^*}^{t_B^*} \mu(t)\tau^*(t)dt &\leq f_1\Lambda(z_T - z_C) - (1-g)f_1\Lambda e(t_B^* - \tilde{t}_A) - e\left[(t_B^* - \tilde{t}_A)gf_1\Lambda - \int_0^{t_B^* - \tilde{t}_A}\left(\int_0^{\Delta}\mu(x)dx\right)d\Delta\right], \\
&= f_1\Lambda(z_T - z_C) - f_1\Lambda e(t_B^* - \tilde{t}_A) + e\int_0^{t_B^* - \tilde{t}_A}\left(\int_0^{\Delta}\mu(x)dx\right)d\Delta, \\
&\overset{(a)}{\leq} f_1\Lambda(z_T - z_C) - f_1\Lambda e(t_B^* - \tilde{t}_A) + e\int_0^{t_B^* - \tilde{t}_A}\left(\int_0^{\Delta}\mu_f dx\right)d\Delta, \\
&= f_1\Lambda(z_T - z_C) - f_1\Lambda e(t_B^* - \tilde{t}_A) + e\int_0^{t_B^* - \tilde{t}_A}\mu_f \Delta d\Delta, \\
&= f_1\Lambda(z_T - z_C) - f_1\Lambda e(t_B^* - \tilde{t}_A) + e\mu_f\frac{(t_B^* - \tilde{t}_A)^2}{2}, \\
&\overset{(b)}{=} f_1\Lambda(z_T - z_C) - f_1\Lambda e(t_B^* - \tilde{t}_A) + ef_1\Lambda\frac{t_B^* - \tilde{t}_A}{2}, \\
&= f_1\Lambda(z_T - z_C) - f_1\Lambda e\frac{t_B^* - \tilde{t}_A}{2} \overset{(c)}{=} \int_{\tilde{t}_A}^{t_B^*}\mu_f\tilde{\tau}(t)dt
\end{aligned} \tag{17}$$

where (a) follows as $\mu(x) \leq \mu_f$ for all $x \in [0, t_B^* - \tilde{t}_A]$, (b) follows as $\mu_f(t_B^* - \tilde{t}_A) = f_1\Lambda$, and (c) follows by our derived relation for $\int_{\tilde{t}_A}^{t_B^*}\mu_f\tilde{\tau}(t)dt$ in Equation (11). This establishes inequality (ii).

Proof of (iii): Using an entirely analogous line of reasoning to that in the proof of inequality (ii), we can show that $\int_{t_C^*}^{\tilde{t}_D}\mu_f\tilde{\tau}(t)dt \geq \int_{t_C^*}^{t_D^*}\mu(n(t))\tau^*(t)dt$. We omit the details here for brevity.

Concluding the Proof: Combining inequalities (i), (ii), and (iii), it is immediate that $R^* \leq \tilde{R}$, i.e., for any tolling policy $\tau^*(\cdot)$ there exists another policy $\tilde{\tau}(\cdot)$ that operates the system at the throughput-maximizing capacity at all periods and achieves a weakly higher revenue. We note that if the throughput induced by the policy $\tau^*(\cdot)$ is strictly below μ_f over any time sub-interval, then at least one of inequalities (i)–(iii) is strict, implying that $R^* < \tilde{R}$, a contradiction to the optimality of $\tau^*(\cdot)$, thus establishing our claim.

D.4 Proof of Proposition 3

In the regime when $z_T - z_C > \frac{\Lambda eL}{\mu(e+L)}$, the minimum achievable total system cost reduces to that in the classical bottleneck model without an outside option and is given by $SC^* = z_C\Lambda + \frac{eL}{2(e+L)}\Lambda^2\left(\frac{1}{\mu} - \frac{1}{\lambda}\right)$. To illustrate the implications of this regime, consider an instance in which $z_C = 0$ (the argument extends directly to settings in which z_C is a small positive constant) and $z_T > \frac{\Lambda eL}{(e+L)(\lambda-\mu)}\left(\frac{2\lambda-\mu}{\mu}\right)$. In this setting, the optimal static toll is $\tau_s^* = (z_T - z_C) - \frac{\Lambda eL}{\mu(e+L)}$ and its total system cost is given by $z_C\Lambda + \frac{\Lambda^2 eL}{2\mu(e+L)}\left(1 - \frac{\mu}{\lambda}\right) + \frac{\Lambda^2 eL}{2\mu(e+L)}$. Then, we have:

$$\frac{SC(\tau_s^*)}{SC^*} = \frac{z_C\Lambda + \frac{\Lambda^2 eL}{2\mu(e+L)}\left(1 - \frac{\mu}{\lambda}\right) + \frac{\Lambda^2 eL}{2\mu(e+L)}}{z_C\Lambda + \frac{\Lambda^2 eL}{2\mu(e+L)}\left(1 - \frac{\mu}{\lambda}\right)} \overset{(a)}{=} 1 + \frac{1}{1 - \frac{\mu}{\lambda}},$$

where the final equality follows from letting $z_C = 0$. Consequently, if $\frac{\mu}{\lambda} \to 1$, the system cost of the static revenue-optimal tolling policy can be unbounded in the regime when $\frac{\mu}{\lambda} \to 1$.

E Additional Details on Numerical Experiments

E.1 Model Calibration Details

This section describes our assumptions and methodology to calibrate our model parameters for the bottleneck and MFD frameworks based on the SF-Oakland Bay Bridge and New York City's CRZ case studies, respectively.

Calibration of Bottleneck Model for Bay Area Case Study: We consider westbound commuting trips from the East Bay into San Francisco, which can occur either by car via the SF–Oakland Bay Bridge or by the Bay Area Rapid Transit (BART) system, a local subway network that runs parallel to the bridge and serves east–west travel across the Bay. All westbound vehicles using the SF–Oakland Bay Bridge are subject to a fixed (static) toll of \$8.50 for most commuter vehicles that does not vary by time of day, and empirical evidence indicates that the SF–Oakland Bay Bridge operates as a true bottleneck (Gonzales and Christofa 2015). Together, these features make the Bay Bridge–BART corridor a well-suited empirical setting for studying the performance gap between static and dynamic tolling policies in a bottleneck model with an outside option.

For our study, we focus on the weekday morning commuting period between 5:00 AM and 10:00 AM, during which westbound travelers choose between driving across the SF–Oakland Bay Bridge, modeled as a bottleneck in our framework, or using BART, which serves as the outside option. To calibrate the total number of users Λ seeking to travel during this time window, we aggregate average weekday westbound trips by car and by BART for August 2025. Using BART ridership data (Bay Area Rapid Transit (BART) 2026), we obtain that the average weekday westbound BART ridership between 5:00 AM and 10:00 AM in August 2025 is 27,132 trips. Vehicular demand is estimated using PeMS data from a sensor (VDS 426389) located immediately upstream of the Bay Bridge, which yields an average flow of 41,369 vehicles over the same period. Summing these two components gives a total cumulative demand of $\Lambda = 68,501$, which we round to $70,000$ users for our experiments. This corresponds to an average desired arrival rate of $\lambda = \frac{\Lambda}{5} = 14,000$ users per hour over the five-hour commuting window.

Next, to calibrate the bottleneck service rate, we use the observed maximum flow at detector VDS 426389 on the SF–Oakland Bay Bridge. The peak flow at this location is approximately $\mu = 9,600$ vehicles per hour, which is consistent with standard capacity ranges for a five-lane highway such as the SF–Oakland Bay Bridge. Under these calibrated parameters, observe that the bottleneck service rate during the morning commute is strictly lower than the desired arrival rate (i.e., $\mu < \lambda$), implying congestion delays on the Bay Bridge. These delays will, in turn, affect commuters' departure-time decisions and may induce mode shifts toward the outside option.

Next, we calibrate the parameters governing user cost function in Equation (2) for westbound car trips from East Bay to San Francisco. We set the value of waiting time to $c_W = \$22$, based on an inflation-adjusted estimate of the average value of travel time in the Bay Area (Metropolitan Transportation Commission 2015). For schedule delay costs, we follow the estimates from the seminal work of Small (Small 1982), setting the earliness parameter to $e = 0.61$ and the lateness parameter to $L = 2.4$. Moreover, we model the free-flow cost of car travel, z_C, in time-equivalent units as the sum of (i) the average free-flow travel time for westbound trips from the East Bay into San Francisco and (ii) the daily parking cost in San Francisco, converted into time-equivalent units through a normalization with the value of waiting time c_W. Daily parking costs in San Francisco are assumed to be \$30, consistent with typical weekday parking fees in the city. Free-flow travel times are calibrated using origin–destination shares and distances for westbound AM trips into San Francisco, yielding a weighted average free-flow travel time of approximately 21 minutes. Full details of the free-flow travel time calibration are reported in Appendix E.2. Then, the resulting free-flow cost of car travel is $z_C = 1.714$ hrs.

Finally, we calibrate the average cost z_T of westbound trips from the East Bay into San Francisco using BART. Analogous to the free-flow cost of car travel, we model z_T as the sum of a monetary fare component and a generalized time cost. The generalized time cost is composed of three elements: (i) walking time to

and from the BART station, (ii) travel time in the BART, and (iii) waiting time at the station. In line with the empirical evidence that time spent using public transit is perceived as more onerous than time spent driving (Wardman 2012), we weight these time components by a multiplicative discomfort factor η when computing z_T.

To estimate these components, we use BART GTFS data to compute ridership volume-weighted averages for east–west Bay trips during the morning rush, yielding an average fare of \$6.14 and an average in-vehicle travel time of approximately 32 minutes (Bay Area Rapid Transit (BART) 2025a). Walking time is assumed to be 10 minutes each way, consistent with BART access guidelines, resulting in 20 minutes of total access and egress time (Bay Area Rapid Transit (BART) 2025b). Moreover, the waiting time is taken to be half of the average train headway of 20 minutes, implying an average waiting time of 10 minutes for randomly arriving commuters (Bay Area Rapid Transit 2025). We combine these components to define z_T as the sum of the fare, converted into time-equivalent units by normalizing with the value of waiting time c_W, and the generalized travel time (walking, waiting, and travel time on BART) weighted by the discomfort factor.

Because the discomfort multiplier associated with transit use may vary substantially across settings, we conduct a sensitivity analysis in our numerical experiments by varying the multiplier over the range $\eta \in [1, 30]$. Multiplier values in the range $\eta \in [1.5, 5]$ are consistent with empirical estimates in practice (Wardman 2012). Exploring a broader range allows us to capture heterogeneity in traveler preferences and to examine how technological or infrastructural changes, such as improvements in transit quality or the increased attractiveness of car travel due to the proliferation of autonomous vehicles (Ostrovsky and Schwarz 2025), affect the cost differential $z_T - z_C$. In addition, varying η over this wider range enables us to study the full set of regimes for $z_T - z_C$ characterized by our theoretical results in Section 4.

We summarize the above parameter values for the SF-Oakland Bay Bridge case study in Table 2 in Appendix E.3.

Calibration of MFD Framework for New York City (NYC) Case Study: We now describe the calibration methodology of our model parameters for the MFD framework in Section 5 based on New York City's recently implemented congestion pricing program. Under this program, which began on January 5, 2025, vehicles are tolled when entering the Congestion Relief Zone (CRZ), defined as the area of Manhattan south of 60th Street. While toll levels vary by vehicle class and other factors, most commuter vehicles are subject to a flat (static) toll of \$9 to enter the CRZ during weekday hours between 5:00 AM and 9:00 PM. These features make New York City's congestion pricing program an especially compelling empirical setting for evaluating the performance gap between static and dynamic tolling policies in city-scale systems, which can be effectively represented through an MFD.

For the purposes of our empirical calibration, we consider commuter trips that either originate or terminate within the CRZ, so that at least a portion of each trip takes place inside the tolled region. Users may choose between traveling by car, incurring the congestion toll when traveling through the CRZ, or using the local subway system, which is separated from the road network. To isolate the key mechanisms of interest in this work, we focus on subway as the only outside option and thus do not explicitly model bus services or other transit modes in New York City for our empirical study.

As with the Bay Area case study, we focus on the weekday morning commuting period between 5:00 AM and 10:00 AM. During this period, we estimate the cumulative travel demand Λ as the sum of subway trips and vehicle trips that occur at least partially within the CRZ. Averaged across weekdays in August 2025, we estimate 700,000 subway trips and 200,000 vehicle trips during the morning peak. Details of the demand calibration are provided in Appendix E.4. Combining these components yields a total calibrated demand of $\Lambda = 900,000$ users, corresponding to a desired arrival rate of $\lambda = \frac{\Lambda}{5} = 180,000$ users per hour.

Next, we calibrate the parameters of the triangular MFD for the New York City case study. To estimate the average trip distance D, we use New York City taxi and for-hire vehicle trip data (New York City Taxi and Limousine Commission 2025) and find that the average trip length for vehicle trips within the CRZ is approximately 3.7 miles (5.95 km). We round this value to $D = 6$ km and use it as the representative trip distance for all vehicle trips in the zone. Moreover, we set the free-flow speed in Manhattan to $25mph$, consistent with posted speed limits on local Manhattan streets.

We calibrate the jam density using standard traffic flow benchmarks. Assuming an average vehicle length of 5 meters and accounting for inter-vehicle spacing under congested conditions implies approximately 7 meters of road space per vehicle, corresponding to a jam density of about 142 vehicles per kilometer per lane. We round this number to set the jam density to 140 veh/km/lane, consistent with estimates in the literature (Knoop 2021). To extend this to the entire CRZ, we approximate the total road supply within the CRZ using OpenStreetMap data (OpenStreetMap contributors 2025), yielding an estimate of approximately 1,000 lane-kilometers. This implies a maximum system-level jam accumulation level of roughly 140,000 vehicles. Given that the true jam accumulation level may be lower than this value, we conduct sensitivity analyses over jam accumulation levels ranging between 14,000 to 140,000 vehicles, capturing plausible variation in the jam density values in practice. Finally, to calibrate the throughput-maximizing capacity μ_f in the CRZ, we follow an analogous procedure to that used to calibrate the arrival rate λ. Using this approach, we estimate the peak achievable flow through the CRZ to be $\mu_f = 45,000$ vehicles per hour (see Appendix E.5 for details). We find that our results are not sensitive to moderate variations in μ_f, indicating that this value captures the key effects relevant for our analysis.

Next, we calibrate the user cost function for car travel through the CRZ. To this end, following the estimate in Cook et al. (2025), we set the value of waiting time to $c_W = \$40$, consistent with mean hourly wage levels in New York City (U.S. Bureau of Labor Statistics 2025). As in the Bay Area case study, we set the earliness parameter to $e = 0.61$ and the lateness parameter to $L = 2.4$. Moreover, we calibrate free-flow cost of car travel, z_C, as the generalized cost of an uncongested trip within the CRZ expressed in time-equivalent units, combining parking fees and the free-flow travel time. The daily parking fee is assumed to be \$30, consistent with an average parking rate of approximately \$3.50 per hour for an eight-hour workday in New York City. We estimate the free-flow travel time as the ratio of the average trip distance within the CRZ to the free-flow speed, given by $\frac{D}{v_f} = \frac{6km}{40km/h} = 0.15$ hours.

We calibrate the subway cost z_T for trips passing through the CRZ using the same formulation as the Bay Area case study. The trip fare is set to \$3 (as of January 2026), reflecting the flat fare structure of the New York City Subway, which applies uniformly across origins and destinations. On weekdays, most subway lines operate with headways of less than five minutes, implying an average waiting time of approximately 2.5 minutes (New York City Comptroller 2023). Walking time to and from the subway is assumed to be 10 minutes per access and egress leg, resulting in a total walking time of 20 minutes (StreetEasy 2018). We further assume that the average time spent traveling on the subway within the CRZ is approximately 12 minutes, corresponding to roughly half the travel time between Columbus Circle (59th Street) and South Ferry along the subway's Red Line, under the assumption of spatially uniformly distributed trips within the CRZ. Finally, as in the Bay Area case, we vary the discomfort multiplier over the range $\eta \in [1, 30]$ in our numerical experiments to examine the implications of differing relative attractiveness of transit versus car travel.

We summarize the above parameter values in the New York City case study in Table 3 in Appendix E.6.

E.2 Free-flow Travel Time Calibration for Bay Area Case Study

In this section, we present the calibration methodology for the average free-flow travel time for westbound trips from East Bay into San Francisco. We use origin–destination (OD) data from Metropolitan Transportation Commission (2015), which reports distances and observed OD shares for major East Bay sub-regions traveling to San Francisco, which we summarize in Table 1. The remaining OD share required to sum to 100 percent corresponds to trips originating in other East Bay locations for which such data are not available. For the origin locations listed in Table 1, the free-flow travel times are computed assuming a constant speed of 50mph, consistent with the posted speed limit on the SF–Oakland Bay Bridge. These travel time estimates are then aggregated using the reported OD shares for westbound AM trips to obtain a weighted average free-flow travel time of **0.35 hours (21 minutes)**. This weighted average provides a parsimonious and empirically grounded estimate of the free-flow travel time for westbound trips from East Bay into San Francisco, required to calibrate the parameter z_C.

Table 1: Calibration of free-Flow Travel Times for Westbound AM Trips into San Francisco. Free-flow times are computed assuming a travel speed of 50 mph. OD shares and distances are based on westbound AM trips reported in Metropolitan Transportation Commission (2015).

Origin	**Distance (miles)**	**Free-Flow Time (hours)**	**OD Share (%)**
Oakland	12.6	0.252	29.8
Richmond	17.9	0.359	12.4
Berkeley	13.8	0.276	12.4
Hayward	27.2	0.544	8.6
Walnut Creek	25.3	0.506	5.0
Concord	31.1	0.622	4.8

E.3 Summary of Parameter Values for Bay Area Case Study

Table 2: Calibrated parameter values of the bottleneck model with an outside option for the SF-Oakland Bay Bridge case study.

Parameter	**Value**
Demand, Arrival Rate, and Bottleneck Service Rate	
Λ	70,000
λ	14,000 users/hr
μ	9,600 vehicles/hr
Components of User Cost of Driving	
c_W	\$22/hr
e	0.61
L	2.4
Parking fee	\$30
Free-flow travel time	21 min
z_C	1.714 hrs
Transit (BART) Parameters to Calibrate z_T	
Fare	\$6.14
Walking time	20 min
Waiting time	10 min
On-board BART travel time	32 min
Discomfort Multiplier (η)	$[1.5, 18]$

E.4 Calibration of Total Travel Demand for New York Case Study

In this section, we describe the calibration methodology for total travel demand within the CRZ for the New York City case study. We estimate total demand as the sum of subway trips and vehicle trips, where some portion of the trip occurs within the CRZ during the weekday morning peak between 5-10 AM, averaged over weekdays in August 2025.

For subway demand, we classify stations according to whether they are located inside or outside the CRZ and aggregate trips across three origin–destination (OD) categories: (i) trips originating within the CRZ and ending outside the CRZ, (ii) trips both originating and ending within the CRZ, and (iii) trips originating outside the CRZ and ending within the CRZ. Together, these OD categories account for the vast majority of subway trips that pass through the CRZ. Using this approach, we estimate an average of 684,306 subway trips traversing the CRZ during the 5–10 AM period on weekdays in August 2025 using the subway OD ridership dataset (State of New York 2025a). For our numerical experiments, we round this figure to 700,000

to conservatively account for any remaining trips not explicitly captured by the above procedure.

For vehicle trips, we begin with CRZ entry data (State of New York 2025b), which indicate an average of around 128,000 vehicle entries into the CRZ between 5-10 AM on weekdays in August 2025, of which around 35,000 are taxi or for-hire vehicle entries. Because these data record only vehicle entries into the zone, they do not capture all vehicle trips occurring within the CRZ. To account for internal trips, we additionally use New York City taxicab and for-hire vehicle (FHV) trip data (New York City Taxi and Limousine Commission 2025), including yellow taxis, green taxis, and FHVs, to estimate the total number of taxi trips traversing the CRZ. Aggregating across the same three origin–destination categories used for the subway calibration, we find approximately 50,000 taxi and FHV trips passing through the CRZ during the morning peak, compared with roughly 35,000 taxi and FHV entries recorded in the CRZ entry data. We therefore scale total vehicle entries by the corresponding factor, yielding an adjusted estimate of approximately $128,000 \times \frac{50,000}{35,000} \approx 183,000$ vehicle trips. For our numerical experiments, we conservatively round this figure to 200,000 vehicle trips to account for any remaining trips not explicitly captured by this procedure.

E.5 Calibration of Throughput Maximizing Capacity for New York Case Study

In this section, we describe the calibration methodology for throughput-maximizing capacity μ_f within the CRZ for the New York City case study. To estimate μ_f, we follow a similar procedure to estimate the average vehicle trips in the CRZ in Appendix E.4 and instead of looking at averages, we look at the maximum travel demand.

To this end, we begin with CRZ entry data (State of New York 2025b), which indicate a maximum of approximately 137,000 vehicle entries into the CRZ between 5-10 AM on weekdays in August 2025, of which about 36,400 correspond to taxi and for-hire vehicle (FHV) entries. Because this data captures only vehicle entries into the zone, we supplement it with New York City taxicab and FHV trip data (New York City Taxi and Limousine Commission 2025), including yellow taxis, green taxis, and FHVs, to account for trips occurring within the CRZ. Aggregating across the same three origin–destination categories described in Appendix E.4, we estimate approximately 54,600 taxi and FHV trips traversing the CRZ during the morning peak, compared with the 36,400 taxi and FHV entries recorded in the CRZ entry data. We therefore scale total vehicle entries by the corresponding ratio, yielding an adjusted estimate of approximately $137,000 \times \frac{54,600}{36,400} \approx 205,000$ vehicle trips. Applying the same conservative rounding adjustment used in Appendix E.4, this implies a total of approximately $205,000 \times \frac{200,000}{183,000} \approx 225,000$ vehicle trips over the five-hour morning period, corresponding to a throughput of $\mu_f = 45,000$ vehicles per hour.

E.6 Summary of Parameter Values for New York Case Study

F System Cost Optimal Static Tolling

F.1 Derivation of System Cost Optimal Static Toll for Bottleneck Model

In the following, we derive the system-cost-optimal static toll in the bottleneck model with an outside option. To this end, we first present the expression for the total system cost as a function of the static toll τ, which is the sum of the cost of using transit and the generalized cost of using the car plus scheduling and queuing delays. In presenting this expression, we focus on the case when $\tau \in [\max\{z_T - z_C - T_C, 0\}, z_T - z_C]$ and $z_T - z_C \geq 0$. Note that in this regime, we have a mixed-mode equilibrium as characterized in Proposition 1. Specifically, letting $\bar{w} = z_T - z_C - \tau$, we have:

- $N_e = \frac{\mu\bar{w}}{e}$ early arrivals
- $N_L = \frac{\mu\bar{w}}{L}$ late arrivals
- $N_o = (1 - \frac{\bar{w}}{T_C})\frac{\Lambda}{\lambda}\mu$ on-time arrivals via car

Table 3: Calibrated parameter values for the MFD framework in the New York City Case Study

Parameter	**Value**
Demand and Arrival Rates	
Λ	900,000 users
λ	180,000 users/hr
MFD Parameters	
Jam density	140 veh/km/lane
Total road length in CRZ	1,000 lane-km
Jam accumulation level (n_j)	$\{14,000, 42,000, 70,000, 140{,}000\}$ vehicles
Throughput-maximizing capacity μ_f	45,000 veh/hr
Average trip distance D	6 km
Car Travel Cost Parameters	
c_W	\$40/hr
e	0.61
L	2.4
Parking fee	\$30
Free-flow speed v_f	40 km/hr
z_C	0.9
Transit (Subway) Parameters to Calibrate z_T	
Fare	\$3
Walking time	20 min
Waiting time	2.5 min
On-board subway time	12 min
Discomfort multiplier η	$[1.5, 18]$

- $N_T = (1 - \frac{\bar{w}}{T_C})\Lambda\left(1 - \frac{\mu}{\lambda}\right)$ arrivals via transit.

Then, we obtain the following expression for the total system cost as a function of the static toll τ:

$$\begin{aligned} SC(\tau) = {} & z_T\left[\Lambda - \frac{\mu(e+L)(z_T - z_C - \tau)}{eL}\right]\left(1 - \frac{\mu}{\lambda}\right) \\ & + z_C\mu(z_T - z_C - \tau)\frac{(e+L)}{eL} + z_C\left[\Lambda - \frac{\mu(e+L)(z_T - z_C - \tau)}{eL}\right]\frac{\mu}{\lambda} \\ & + \frac{\mu(z_T - z_C - \tau)^2}{2}\frac{(e+L)}{eL}\left(1 - \frac{\mu}{\lambda}\right) \\ & + (z_T - z_C - \tau)\left(\Lambda - \frac{\mu(e+L)(z_T - z_C - \tau)}{eL}\right)\frac{\mu}{\lambda} + \frac{\mu(z_T - z_C - \tau)^2}{2}\frac{(e+L)}{eL}. \end{aligned}$$

To compute the optimal uniform toll, we first compute the derivative $SC'(\tau)$ of the total system cost, which is given by:

$$SC'(\tau) = \tau\frac{\mu(e+L)}{eL}\left(2 - 3\frac{\mu}{\lambda}\right) - \frac{\Lambda\mu}{\lambda} + (z_T - z_C)\frac{\mu(e+L)}{eL}\left(\frac{2\mu}{\lambda} - 1\right).$$

We now compute the system-cos-optimal static toll in three regimes: (i) $\frac{\mu}{\lambda} = \frac{2}{3}$, (ii) $\frac{\mu}{\lambda} > \frac{2}{3}$, and (iii) $\frac{\mu}{\lambda} < \frac{2}{3}$.

Case (i): In the regime when $\frac{\mu}{\lambda} = \frac{2}{3}$, the total system cost is linear with $SC'(\tau) = -\frac{2\Lambda}{3} + \frac{1}{3}(z_T - z_C)\frac{\mu(e+L)}{eL}$. Thus, we have:

$$\tau^* = \begin{cases} \max\{0, z_T - z_C - T_C\}, & \text{if } z_T - z_C > \frac{2\Lambda eL}{\mu(e+L)} \\ z_T - z_C, & \text{if } z_T - z_C < \frac{2\Lambda eL}{\mu(e+L)} \\ [\max\{0, z_T - z_C - T_C\}, z_T - z_C], & \text{if } z_T - z_C = \frac{2\Lambda eL}{\mu(e+L)}. \end{cases} \tag{18}$$

Note that in the regime when $z_T - z_C > \frac{2\Lambda eL}{\mu(e+L)} = 2T_C$, it follows that $\max\{0, z_T - z_C - T_C\} = z_T - z_C - T_C$.

Case (ii): In the regime when $\frac{\mu}{\lambda} > \frac{2}{3}$, the total system cost is concave and hence the minimum must occur at one of the end-points 0 or $z_T - z_C$. Leveraging the fundamental theorem of calculus and the linearity of the derivative $SC'(\tau)$, it follows that:

$$\begin{aligned} SC(z_T - z_C) - SC(0) &= \int_0^{z_T - z_C} SC'(\tau)d\tau = \frac{z_T - z_C}{2}(SC'(0) + SC'(z_T - z_C)) \\ &= \frac{\mu}{\lambda}\left((z_T - z_C)\mu\frac{e+L}{eL} - 2\Lambda\right). \end{aligned}$$

Thus, we have:

$$\tau^* = \begin{cases} z_T - z_C - T_C, & \text{if } z_T - z_C > \frac{2\Lambda eL}{\mu(e+L)} \\ z_T - z_C, & \text{if } z_T - z_C < \frac{2\Lambda eL}{\mu(e+L)} \\ [z_T - z_C - T_C, z_T - z_C], & \text{if } z_T - z_C = \frac{2\Lambda eL}{\mu(e+L)}. \end{cases} \tag{19}$$

Case (iii): In the regime when $\frac{\mu}{\lambda} < \frac{2}{3}$, the total system cost is convex and hence there is a possibility that the minimum occurs between the range $(0, z_T - z_C)$. In this case, note that the unique unconstrained minimizer satisfies $SC'(\underline{\tau}) = 0$, i.e.,

$$\underline{\tau} = \frac{\frac{\Lambda\mu el}{\lambda\mu(e+L)} + (z_T - z_C)\left(1 - \frac{2\mu}{\lambda}\right))}{2 - 3\frac{\mu}{\lambda}}$$

We have that if $\frac{\Lambda\mu el}{\lambda\mu(e+L)} \leq (z_T - z_C)(\frac{2\mu}{\lambda} - 1)$, then $\underline{\tau} \leq 0$, i.e., the minima occurs at 0, if $\frac{\Lambda\mu el}{\lambda\mu(e+L)} > (z_T - z_C)(1 - \frac{\mu}{\lambda})$, then $\underline{\tau} \geq z_T - z_C$.

Then, we have:

$$\tau^* = \begin{cases} \max\{0, z_T - z_C - T_C\}, & \text{if } \frac{\Lambda\mu el}{\lambda\mu(e+L)} \leq (z_T - z_C)(\frac{2\mu}{\lambda} - 1) \\ z_T - z_C, & \text{if } \frac{\Lambda\mu el}{\lambda\mu(e+L)} \geq (z_T - z_C)(1 - \frac{\mu}{\lambda}) \\ \max\left\{\frac{\frac{\Lambda\mu el}{\lambda\mu(e+L)} + (z_T - z_C)\left(1 - \frac{2\mu}{\lambda}\right))}{2 - 3\frac{\mu}{\lambda}}, z_T - z_C - T_C\right\}, & \text{otherwise.} \end{cases} \tag{20}$$

F.2 Derivation of Total System Cost Under Static Tolling for Triangular MFD

We now derive the total system cost corresponding to any static toll τ under a triangular MFD, which consists of the following four terms: (i) cost of using transit, (ii) cost of using a car at free-flow, (iii) waiting time or queuing delays when using the car, and (iv) schedule delays. We now derive expressions for each of these four terms in the regime when $\tau \geq \underline{\tau}$, where $\underline{\tau} \geq 0$ is the minimum toll at which there is some user who is indifferent between using car and transit.

Transit Costs: We have the following expression for the cost of using transit:

$$C_T(\tau) = z_T \left(\Lambda - \int_{t_A}^{t_D} \mu(n(t))dt \right),$$
$$= z_T \left[\Lambda - \frac{\Lambda}{\lambda} \frac{n_j}{\frac{n_j}{\mu_f} + z_T - z_C - \tau} - n_j \frac{e+L}{eL} \ln \left(1 + \frac{(z_T - z_C - \tau)\mu_f}{n_j} \right) \left(1 - \frac{\frac{n_j}{\frac{n_j}{\mu_f} + z_T - z_C - \tau}}{\lambda} \right) \right], \quad (21)$$

where the equality follows from the analysis in Theorem 5.

Free-flow Car Costs: We have the following expression for the cost of using car:

$$C_F(\tau) = z_C \int_{t_A}^{t_D} \mu(n(t))dt,$$
$$= z_C \left[\frac{\Lambda}{\lambda} \frac{n_j}{\frac{n_j}{\mu_f} + z_T - z_C - \tau} + n_j \frac{e+L}{eL} \ln \left(1 + \frac{(z_T - z_C - \tau)\mu_f}{n_j} \right) \left(1 - \frac{\frac{n_j}{\frac{n_j}{\mu_f} + z_T - z_C - \tau}}{\lambda} \right) \right]. \quad (22)$$

Waiting and Queuing Delays: The total waiting and queuing delays are given by the sum of the corresponding delays for the (a) on-time car users, (b) early car users, and (l) late car users. For the on-time car users, we have the following expression for the queuing delay:

$$C_Q^O(\tau) = (t_C - t_B)\mu_\tau(z_T - z_C - \tau),$$
$$= \left(\frac{\Lambda}{\lambda} - \frac{1}{\lambda} n_j \frac{e+L}{eL} \ln \left(1 + \frac{(z_T - z_C - \tau)\mu_f}{n_j} \right) \right) \frac{n_j}{\frac{n_j}{\mu_f} + z_T - z_C - \tau} (z_T - z_C - \tau) \quad (23)$$

For the early car users, we have the following expression for the queuing delay:

$$C_Q^E(\tau) = \int_{t_A}^{t_B} \mu(n(t))w(n(t))dt,$$
$$= \int_{t_A}^{t_B} \mu(n(t)) \left(\frac{n(t)}{\mu(n(t))} - \frac{n_c}{\mu_f} \right) dt,$$
$$= \int_{t_A}^{t_B} n_j \left(1 - \frac{\mu(n(t))}{\mu_f} \right) dt,$$
$$= \int_0^{t_B - t_A} n_j \left(1 - \frac{\frac{n_j}{\frac{n_j}{\mu_f} + (z_T - z_C - \tau - e\Delta)}}{\mu_f} \right) d\Delta,$$
$$= \int_0^{t_B - t_A} n_j \frac{z_T - z_C - \tau - e\Delta}{\frac{n_j}{\mu_f} + z_T - z_C - \tau - e\Delta} d\Delta. \quad (24)$$

Analogously, we have the following expression for the queuing delay of the late car users:

$$C_Q^L(\tau) = \int_0^{t_B - t_A} n_j \frac{z_T - z_C - \tau - L\Delta}{\frac{n_j}{\mu_f} + z_T - z_C - \tau - L\Delta} d\Delta. \quad (25)$$

Schedule Delay: We first derive the total earliness delay in the system. To this end, for a user that exits the system at t with a desired exit time $t^*(t)$, their earliness delay is given by:

$$t^*(t) - t = t_1 + \frac{t - t_A}{t_B - t_A}(t_B - t_1) - t = \frac{(t_1 - t_A)(t_B - t)}{t_B - t_A}.$$

We now use this expression to derive the total earliness delay, which is given by:

$$\begin{aligned}
C_S^E(\tau) &= e\int_{t_A}^{t_B} \mu(n(t))\frac{(t_1-t_A)(t_B-t)}{t_B-t_A}dt, \\
&= e\int_{t_A}^{t_B} \frac{n_j}{\frac{n_j}{\mu_f}+(z_T-z_C-\tau-e(t-t_A))}\frac{(t_1-t_A)(t_B-t)}{t_B-t_A}dt, \\
&= \frac{n_j}{e}\left(z_T-z_C-\tau-\frac{n_j}{\lambda}\ln\left(1+\frac{(z_T-z_C-\tau)\mu_f}{n_j}\right)\right)\left[1-\frac{n_j}{\mu_f(z_T-z_C-\tau)}\ln\left(1+\frac{(z_T-z_C-\tau)\mu_f}{n_j}\right)\right],
\end{aligned}$$

where the final equality is obtained by solving the integral.

Analogously, the total lateness delay is given by:

$$C_S^L(\tau) = \frac{n_j}{L}\left(z_T-z_C-\tau-\frac{n_j}{\lambda}\ln\left(1+\frac{(z_T-z_C-\tau)\mu_f}{n_j}\right)\right)\left[1-\frac{n_j}{\mu_f(z_T-z_C-\tau)}\ln\left(1+\frac{(z_T-z_C-\tau)\mu_f}{n_j}\right)\right].$$

Then, we have that the total social cost is given by:

$$SC(\tau) = C_T(\tau)+C_F(\tau)+C_Q^O(\tau)+C_Q^E(\tau)+C_Q^L(\tau)+C_S^E(\tau)+C_S^L(\tau).$$