U.S. Department of Transportation
Federal Highway Administration

Economics: Pricing, Demand, and Economic Efficiency

A PRIMER

Quality Assurance Statement

The Federal Highway Administration (FHWA) provides high quality information to serve Government, industry, and the public in a manner that promotes public understanding. Standards and policies are used to ensure and maximize the quality, objectivity, utility, and integrity of its information. FHWA periodically reviews quality issues and adjusts its programs and processes to ensure continuous quality improvement.

Contents

The Primer Series and the Purpose of This Volume

About This Primer Series

The Congestion Pricing Primer Series is part of FHWA's outreach efforts to introduce the various aspects of congestion pricing to decision-makers and transportation professionals in the United States. The primers are intended to lay out the underlying rationale for congestion pricing and some of the technical issues associated with its implementation in a manner that is accessible to non-specialists in the field. Titles in this series include:

- Congestion Pricing Overview.
- Non-Toll Pricing.
- Technologies That Enable Congestion Pricing.
- Technologies That Complement Congestion Pricing.
- Transit and Congestion Pricing.
- Economics: Pricing, Demand, and Economic Efficiency.
- Income-Based Equity Impacts of Congestion Pricing.

States and local jurisdictions are increasingly discussing congestion pricing as a strategy for improving transportation system performance. In fact, many transportation experts believe that congestion pricing offers promising opportunities to cost-effectively reduce traffic congestion, improve the reliability of highway system performance, and improve the quality of life for residents, many of whom are experiencing intolerable traffic congestion in regions across the country.

Because congestion pricing is still a relatively new concept in the United States, the Federal Highway Administration (FHWA) is embarking on an outreach effort to introduce the various aspects of congestion pricing to decision-makers and transportation professionals. One element of FHWA's congestion-pricing outreach program is this Congestion Pricing Primer Series. The aim of the primer series is not to promote congestion pricing or provide an exhaustive discussion of the various technical and institutional issues one might encounter when implementing a particular project; rather, the intent is to provide an overview of the key elements of congestion pricing, to illustrate the multidisciplinary aspects and skill sets required to analyze and implement congestion pricing, and to

provide an entry point for practitioners and others interested in engaging in the congestion-pricing dialogue.

The concept of tolling and congestion pricing is based on charging for access and use of our roadway network. It places responsibility for travel choices squarely in the hands of the individual traveler, where it can best be decided and managed. The car is often the most convenient means of transportation; however, with a little encouragement, people may find it attractive to change their travel habits, whether through the consolidation of trips, car-sharing, by using public transportation, or by simply traveling at less-congested times. The use of proven and practical demand-management pricing that we freely use and apply to every other utility is needed for transportation.

The application of tolling and road pricing provides the opportunity to solve transportation problems without Federal or state funding. It could mean that further gas tax, sales tax, or motor-vehicle registration fee increases are not necessary now or in the future. Congestion pricing is not a complete plan of action. It has to be coordinated with other policy measures to maximize success.

This volume describes the underlying economic rationale for congestion pricing and how it can be used to promote economic efficiency. It lays out the basic theory of travel demand and traffic flow and shows how inefficient pricing of the road network helps create an economic loss to society, as well as the means by which this can be alleviated through pricing. The impact of congestion pricing on highway infrastructure investment and the revenue implications of congestion pricing will be discussed in a separate volume in this primer series.

Introduction

Traffic congestion represents a significant and growing threat to mobility in the United States. According to recent data, the average driver in U.S. urbanized areas traveling in peak periods experienced 38 hours of total delay in 2005 (up from 14 hours in 1982), at a total cost of $78 billion in excess travel time and wasted fuel consumption.

To some extent, these trends reflect the vibrancy of our Nation's cities, which have seen significant economic growth over the past 3 decades. They also reflect growing imbalances between travel growth and the development of new highway capacity, as the former has exceeded the latter for many years. However, in an important way, traffic congestion also results from the inefficient pricing of highway use, leading drivers to travel more than what is economically desirable, particularly during the height of morning and evening rush hours.

The fundamental economic problem in highway congestion is that highway users consider the costs that they bear themselves (such as fuel costs and travel time), but not the additional delay that they impose on others when using a congested facility. Economists refer to these latter costs as "external costs" or "externalities," because they are not borne by the person whose use of the highway creates them. The goal of congestion pricing is to make drivers "internalize" these costs by directly charging them for use of the highway in proportion to the external costs that they cause for society. In doing so, the economic inefficiencies associated with underpriced and overutilized highways can be significantly reduced.

Congestion pricing is becoming increasingly viewed as an important strategy for improving the operational performance of the highway system. By moving from a system based on fixed or flat charges to highway users (e.g., vehicle fees and fuel taxes) to a more refined system of charges that can vary by the time and location of travel, price signals can be used to more efficiently and effectively allocate scarce capacity on our Nation's highways.

Basic Economic Concepts: Highway Supply and Demand

This section introduces some of the basic economic principles that provide a foundation for understanding the economic rationale for congestion pricing, discussed in the following sections. Key concepts include:

- The explicit and implicit costs associated with highway travel;
- The demand for using highway facilities;
- Basic relationships of traffic volumes and traffic flow; and
- The interaction of supply and demand to determine traffic volumes and highway user costs.

THE COST OF TRAVEL

We live in a world of scarcity, meaning that resources are not unlimited. As a result, when we consume a good or service, we forego the opportunity to consume something else. Economists refer to this as an "opportunity cost," and it is perhaps the most fundamental concept in all of economics. In an important sense, the "price" that people pay for consumption is this opportunity cost.

For most goods and services, the price that consumers pay is simply the out-of-pocket monetary cost to the user; the foregone opportunity would be to spend that money on something else. However, economists have also recognized that consumption requires time as well as money, which is itself in limited supply. This insight has been useful in analyzing the demand for time-saving devices (e.g., microwave ovens) and recreational services, among many others.

For highway travel, the time spent in transit is one of the most critical components of the price that users face. Time spent in travel is time that could be devoted to other pursuits, such as earning income or engaging in leisure activities. The value of these other pursuits represents the opportunity cost of travel time.

Safety is another implicit cost associated with highway travel. When people choose to travel, they also take on the risk of property damage, injuries, or even death due to crashes. These risks can be valued based on the willingness of people to pay for reducing the risk of such adverse outcomes.

Other types of costs borne by highway users do represent actual, direct costs. Vehicle operating costs include such items as fuel, oil, tires, maintenance, and depreciation. Travel-related taxes and tolls represent another direct cost to users. In the case of cash or coins paid at toll booths, these can literally be "out-of-pocket" costs.

The preceding discussion of user costs is based on the predominant form of highway use, namely individuals driving or riding in privately owned passenger vehicles. Similar considerations hold for commercial vehicles, with two key exceptions. First, the value of time for drivers is a direct cost rather than an implicit one, because the drivers are being paid for their services. Second, the user cost for truck transportation also includes the time value of the cargo while in transit.

THE DEMAND FOR HIGHWAY TRAVEL

The demand for highway transportation represents the value that consumers place on traveling in a particular time, manner, and place, as measured by their willingness to "pay" for a trip. Some trips will be valued very highly, whereas others will be valued much less so. This relationship between the cost of travel and the level of demand for travel is commonly depicted as the travel demand curve (see Exhibit 1).

The travel demand curve slopes downward, reflecting a basic economic truth: As the price of a good or service falls, the quantity that will be demanded increases, holding other factors constant. The demand for travel is no different: When the price of travel is high (in the generalized user-cost sense described above), fewer people will be willing to make fewer trips; when that price falls, there will be more people willing to make more trips.

The demand curve is characterized by two important qualities: its level and its shape. The level of demand (i.e., the position of the demand curve) is affected by a number of factors. For example, each trip has an origin and a destination. The more people there are at a particular origin and the more activities (e.g., shopping or employment) there are at a particular destination, the more will routes between the origin and destination be in demand for travel. As income levels rise, the willingness to pay for travel also increases, shifting the demand curve outward. Demand levels can also vary significantly (and importantly for the discussion here) by time of day, due to the simple fact that people prefer to sleep at night and be active during the day, leading to higher levels of demand for travel in the morning and early evening and lower levels of demand during mid-day and overnight hours. Finally, subjective qualities such as comfort and convenience can affect the level of demand.

The responsiveness of the quantity of travel demanded to changes in the price of travel is measured by *travel demand elasticity*. Mathematically, it is simply the percentage change in quantity demanded divided by the percentage change in price.

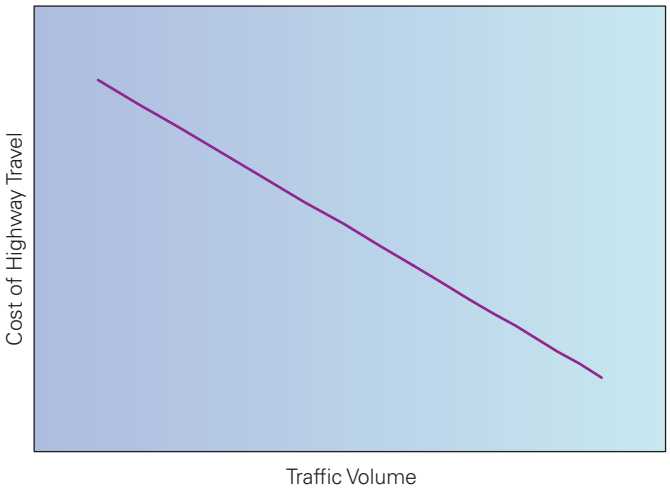

Exhibit 1. The travel demand curve.

Intuitively, elasticity represents the shape of the demand curve. If the quantity demanded changes significantly in response to small changes in price, demand is said to be relatively elastic; thus, the demand curve is fairly flat. Conversely, if demand changes only slightly in response to large changes in price, demand is said to be relatively inelastic; thus, the demand curve will be relatively steep. At the extremes, demand can be said to be perfectly elastic (i.e., any change in price results in an infinite change in quantity demanded) or perfectly inelastic (i.e., any change in price results in no change in quantity demanded).

The elasticity of demand also depends on a number of factors. Perhaps most important is the timeframe being considered: Demand is typically less elastic in the short run than in the long run. When the price of travel changes significantly (as it has recently with large runups in fuel prices), travelers initially have relatively few opportunities for adjusting their behavior. They may decide not to make some trips or to change their mode of travel to work, but their housing and employment locations, key determinants of the level of travel, are likely to remain fixed initially. In the long run, however, everything is variable. People may choose to move closer to their work or take jobs closer to home. Commercial real estate development patterns may

also respond to reduce the distance between consumers and activity centers. As a result, the long-term impact of an increase in travel costs on the volume of highway may be much higher than the short-term impact.

The elasticity of demand is also affected by the quality and availability of close substitutes. For example, if two companies make very similar products, then consumers are likely to readily switch from one product to the other in significant numbers if the price of one of the products changes, resulting in high-demand elasticity for each product. Conversely, if there are no good substitutes for a good or service, then consumers might simply be faced with a choice between paying a higher price or going without, in which case demand is likely to be inelastic.

For highway travel, public transit use can often serve as a reasonable alternative to highway use. In general, the higher the quality of that service (such as frequency and convenience), the better a substitute it will be to driving and the more elastic highway travel demand will be. Telecommuting can also serve as a substitute for highway commuting during peak periods; the more readily it is offered by employers, and the better it is at substituting for time spent in an office (e.g., through high-quality telecommunications), the more highway travel will respond to changes in the cost of travel.

THE SUPPLY SIDE

In determining market outcomes, supply is the counterpart to demand. For most goods and services, supply is simply related to production: how many toys will be fabricated, how many tennis lessons will be taught, etc. For highway travel, however, the relationship is more complicated. The highway infrastructure may be built and maintained by a public agency or private entity, but consumers supply the vehicles (and their persons) to travel on that infrastructure. In order to understand the supply side of highway travel, it is first helpful to review some basic principles of traffic flow and how highways become congested.

Traffic engineers typically characterize traffic flow as a relationship between travel speeds, traffic volumes, and traffic density (e.g., number of vehicles occupying a given space on the road). Exhibit 2 shows the general shape of these relationships.

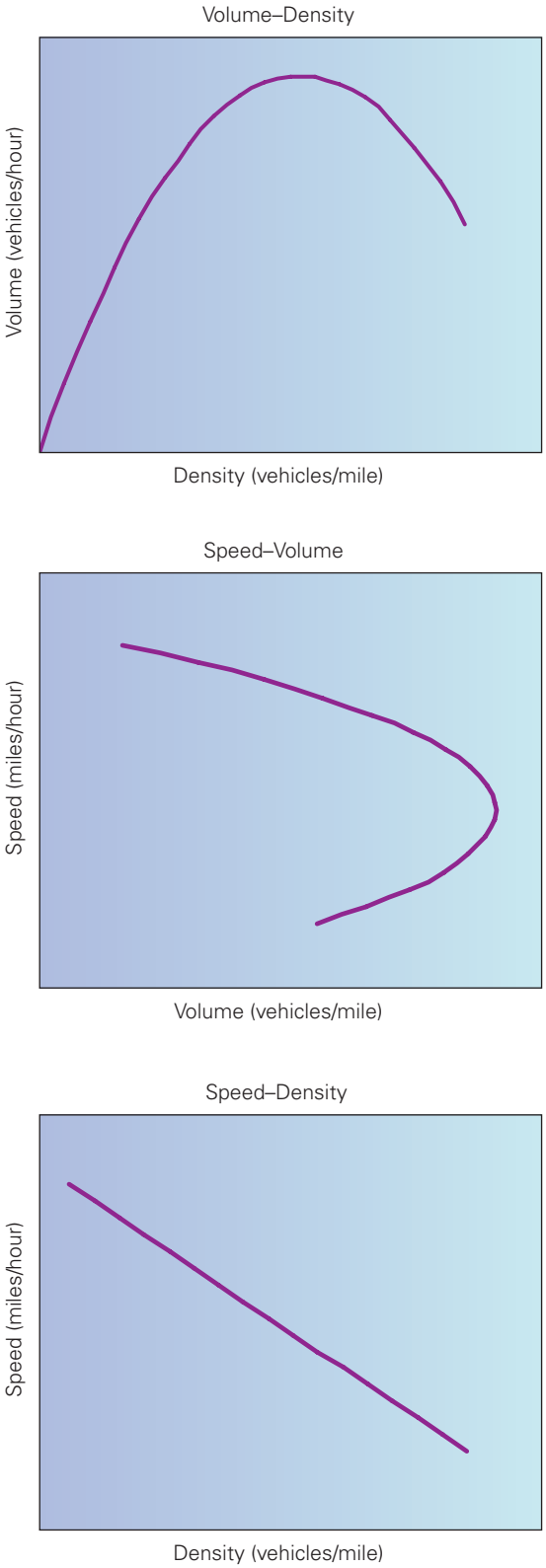

Exhibit 2. Fundamental traffic flow relationships.

When traffic volumes are very low, vehicles have minimal impact on one another, and their travel speeds are limited only by traffic-control devices and the geometry of the road. As traffic volumes increase, however, traffic density increases, and the freedom for vehicles to maneuver is more constricted. As a result, travel speeds begin to decline, relatively slightly at first, but falling significantly as traffic volumes approach the maximum capacity (service flow rate) on the facility. As traffic density continues to increase beyond this saturation point, the speed–volume relationship actually bends backward, as traffic flow breaks down and fewer vehicles are able to get through.

The decline in travel speeds as traffic volumes approach roadway capacity, of course, is what we all know as *congestion delay*. The important implication of this is that there will be a relationship between highway-user costs and traffic volumes on a particular road. At lower volumes, user costs will be relatively constant with respect to volume. As traffic volumes increase, however, user costs will eventually begin to rise at an increasing rate; the point at which this occurs depends on the capacity of the road (see Exhibit 3). This relationship is sometimes referred to as a generalized user cost curve.

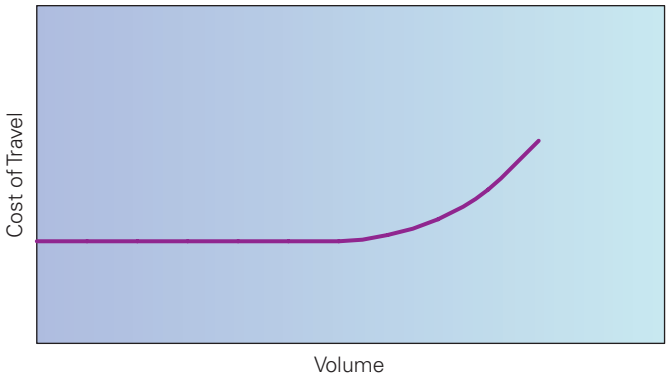

Exhibit 3. Generalized user costs and traffic volumes.

MARKET EQUILIBRIUM

When supply and demand are in balance, a market is said to be in *equilibrium*. This is often represented as the intersection of a supply curve and a demand curve, which determines the market-clearing price and quantity (see Exhibit 4). At this point, everyone who purchases the good is willing to (collectively) buy that amount at that price, and producers are willing to supply that quantity at that price. If either the supply or demand curves shift, the market price and quantity will also change.

For highway travel, demand is determined as described above. The "supply" curve, however, is essentially represented by the generalized cost curve. The intersection of these two curves determines how high traffic volumes will be and what the associated average highway-user costs will be at that volume level. When the level of demand is low relative to the capacity of the road, it will be uncongested, and prices will be relatively constant even as volumes increase (the "flat" part of the user cost curve in Exhibit 4). However, when demand levels are high and the road is congested, both user costs and traffic volumes will be higher, potentially rising sharply as demand continues to increase.

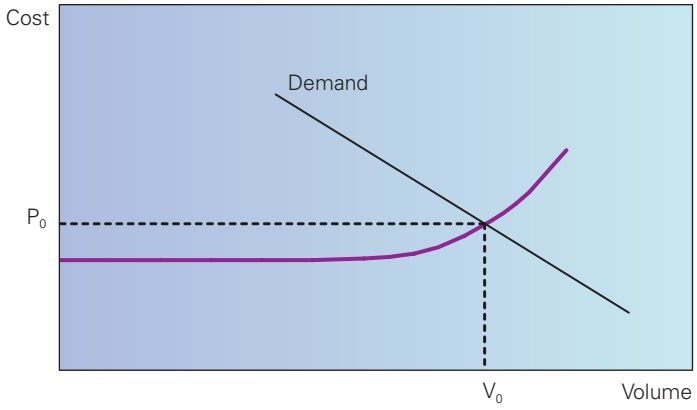

Exhibit 4. Equilibrium user costs and traffic volumes. P = price. V = volume.

Pricing and Economic Efficiency

CONGESTION EXTERNALITIES AND ECONOMIC EFFICIENCY

Economics is focused on the allocation of scarce resources in the economy. When resources are allocated to their highest valued use in the economy, the outcome is said to be *economically efficient*. Under most circumstances, the setting of prices in markets through the interaction of consumers and producers can help achieve an efficient outcome. In many other cases, however, market forces can lead to excessive consumption in a particular sector; highways are a good example of this.

The key to achieving economic efficiency in a market is ensuring that prices reflect the opportunity cost to society of producing and consuming a particular good or service. For most goods and services, this opportunity cost will be equivalent to the incremental cost of the resources that are required to produce the good. For some goods, however, one person's consumption can have a negative impact on other users (or non-users) of the good, a situation that economists refer to as an *external cost* or *negative externality*. If the market price does not reflect these external effects, then the opportunity cost to society will be greater than the price that consumers face, leading to excess consumption from the point of view of society and thus creating economic inefficiency.

From an economic point of view, externalities are a key problem of congested highways. To see why, consider the generalized cost curve shown in Exhibit 3. This curve is an average cost curve and reflects the rise in travel costs that individual users face as traffic volumes increase. What it does not reflect, however, are the incremental costs that each successive vehicle that enters the traffic stream imposes on all the others by causing the speed of the entire flow of vehicles to decrease as congestion builds. The true opportunity cost to society of using highways thus includes both the costs that individual users face and the congestion externality. This is sometimes referred to as the *marginal social cost* of congestion (see Exhibit 5).

From the point of view of society, the efficient traffic level would occur at the point where the marginal social cost curve meets the demand curve, shown in Exhibit 5 as V*. At this traffic level, all the users of the highway value their trips at least as much as the incremental cost to society of adding more users. The equilibrium traffic level, however, is somewhat higher, at V_0. At this point, there are a large number of drivers who are using the facility, because the value they place on travel is greater than the cost that they face (the demand curve is above the average user-cost curve), but the cost to society is greater than the value they receive. Thus, there is an economic efficiency loss due to excessive traffic volumes.

EFFICIENT PRICING

Improving economic efficiency under these circumstances requires the elimination of trips that are valued less than their social cost. One way to do this might be to somehow identify all of the lower valued trips and to ensure that they do not occur during this demand period on that particular facility, perhaps through prohibition. A much simpler means, however, is to adjust the price signals that potential users of the highway facility receive. This can be done by imposing a toll on all

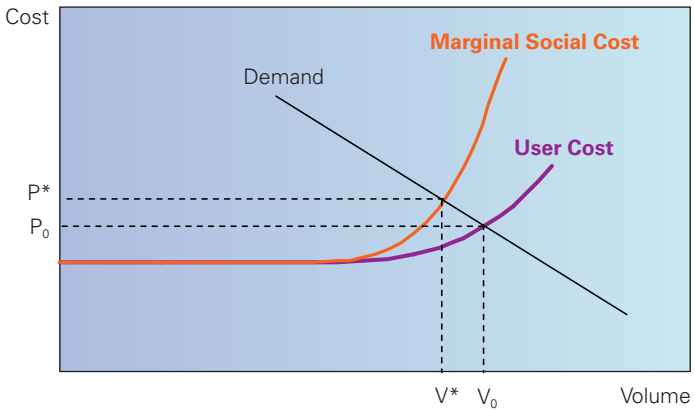

Exhibit 5. Marginal social cost and user cost.
P = price. V = volume.

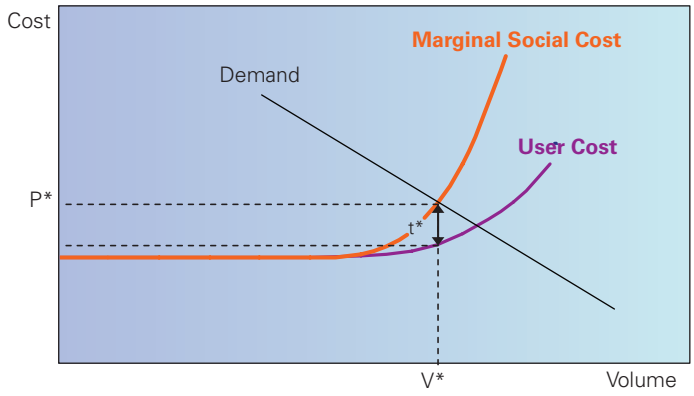

Exhibit 6. Optimal congestion pricing.
P = price. V = volume. t = toll.

users of the facility during congested periods corresponding to the magnitude of the congestion externality; thus, the price that users face is equivalent to the marginal cost to society. At the optimum toll level (t* in Exhibit 6), lower valued trips will be shifted to other routes or time periods (or not made at all), such that the new equilibrium traffic volume is also the socially optimum level. This is congestion pricing.

There are several important points to note with regard to efficient pricing. The first is to emphasize that the optimal congestion toll is intended to reduce congestion on the facility but not eliminate it. To see why, consider Exhibit 7. Eliminating congestion in this example would require much more significant reductions in traffic volumes to V_1, just before congestion begins to build. Doing so with pricing would require a toll of t_1, much greater than the "optimal" toll of t*. This would yield an inefficiently *low* level of traffic on this facility, because many users would be priced based on those who value their travel more than it costs society, even when accounting for the external costs that they impose on others.

The second point is to note that revenue generation is not the primary purpose of congestion pricing, which is intended to better align supply and demand to overcome the economic inefficiencies imposed by congestion. In reality, of course, any such tolling scheme could potentially generate significant revenues, whose use becomes a significant policy and operational issue. The relationship of congestion pricing to user revenues and infrastructure investment will be discussed in a separate volume in this primer series.

For a given facility, the level of the optimal congestion toll and the resulting effect on congestion levels is largely a function of the elasticity of demand for travel. If demand is relatively inelastic, the optimal congestion toll will be high, and the effect on reducing congestion will be relatively low. Conversely, if demand is more elastic, then a greater reduction in traffic volumes and congestion can be achieved with a smaller toll. As was discussed above, a key factor in determining the elasticity of demand is the quality and availability of alternatives to highway use in peak periods, such as transit or teleworking. The implication of this is that measures to improve such alternatives can greatly improve the effectiveness of efficient pricing in reducing congestion.

VARIABILITY IN DEMAND AND PRICING

The level of the congestion toll is also a function of the level of demand. Although this is intuitively obvious, it has important implications for the design of congestion-pricing schemes and helps demonstrate its superiority to other forms of user charges. As was noted above, one of the key features of the demand for highways is that it varies significantly by time of day as well as by location. This implies

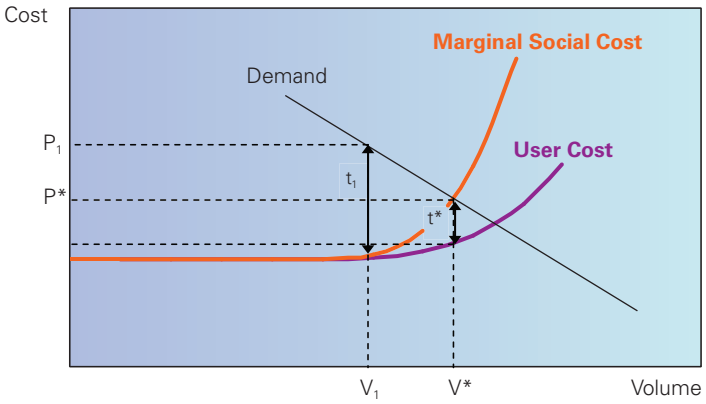

Exhibit 7. Inefficiently high congestion pricing.
P = price. V = volume. t = toll.

that the optimal congestion tolls will vary over the course of the day on the basis of fluctuations in demand. Suppose, however, that a "flat" toll (e.g., fixed tolls and fuel taxes that do not discriminate based on when travel occurs) must be charged that does not vary in this manner. In theory, the flat toll could still be set to match the optimal toll during congested, high-demand periods. However, if the same toll rate applies during uncongested, low-demand times (when there are no congestion externalities and thus no optimal toll), traffic volumes will be inefficiently low in those periods. In fact, because the tolls are the same throughout the day, some of that traffic might even move back into the congested period. Because of the potential for this to occur, the optimal toll during the peak period would be lower (and less effective at reducing congestion) if the tolls cannot vary by time of day.

This example can be extended to multiple demand periods. Suppose the congestion toll is structured to be charged at a fixed rate when it is collected but may be zero (i.e., not collected) during off-peak periods. To the extent that demand varies during the congestion period (such as between the height of rush hour and "shoulder hours" adjacent to the peak), the same considerations would apply. Congestion tolls for the shoulder period would either be inefficiently high or inefficiently low. The result is that the more variability there is in the congestion-pricing scheme, the more effective it can be at reducing congestion and the closer it can come to achieving an economically efficient outcome.

Similar considerations apply when looking at pricing across facilities in different locations. Although traffic congestion can be quite severe in large urban areas, there is much less congestion in smaller cities or rural areas. As a result, a user charge (such as fuel taxes) that is applied uniformly to both congested and uncongested areas will not be nearly as effective as congestion-based pricing in efficiently allocating resources within the economy.

Another consideration in limited congestion pricing concerns alternate routes. Suppose there are two parallel roads between the same origin and destination, both subject to congestion, but only one of them can be priced. If congestion tolls are only charged on the one facility, some of the trips that are priced off that facility will shift to the unpriced facility, exacerbating congestion there. Given the possibility of this to occur, the optimal toll under such a situation will be lower, as will the impact on reducing congestion levels. An important implication of this is that congestion pricing will be much more effective if it is jointly applied to all major highway facilities in a congested corridor, rather than to only a subset of the alternative routes.

In reality, of course, highway networks are comprised of a whole series of interconnected segments, supporting both alternative and connecting routes, as well as access to connections with other modes (such as train stations). The Federal Highway Administration's (FHWA's) Office of Operations has been conducting research to examine the potential impact of congestion pricing when applied to such real-world transportation networks.

CONGESTION PRICING AND OTHER EXTERNALITIES

Highway travel creates other external costs for society, in addition to congestion. Perhaps the most significant of these other impacts is the damage

caused to the environment and human health by highway vehicle emissions. In a broad sense, the marginal social cost of highway use would include these costs. Market and regulatory strategies for dealing with environmental externalities also exist, such as taxes on fuel or regulations on vehicle-emissions control systems and fuel content, and such strategies are the most effective way of dealing with these particular externalities. However, to the extent that current mechanisms do not fully capture these environmental externalties, congestion pricing can help reduce their magnitude by reducing excessive highway use (at least during congested travel periods).

TOLL-COLLECTION COSTS AND THE ROLE OF TECHNOLOGY

The preceding discussion of the economic principles underlying congestion pricing does not address the cost of actually collecting such tolls from users. Indeed, although William Vickrey articulated the basic theory of congestion pricing in his Nobel Prize-winning work in the 1950s, its potential for widespread application was limited by the available methods of collecting tolls. With manned tollbooths and the traffic-flow restrictions that they impose on roadways, it was a case of the "cure being worse than the disease."

Advancements in tolling technology over the past two decades, however, have made variable tolling in urban cores and on limited-access highways a viable option, and future developments in global-positioning satellite (GPS) and dedicated short-range communications technologies may make it feasible to adopt congestion pricing on an even wider basis. Two other volumes in this primer series, *Technologies That Enable Congestion Pricing* and *Technologies That Complement Congestion Pricing*, describe these technologies in greater detail.

Forms of Congestion Pricing

The discussion above describes the classic economic case for congestion pricing, based on mitigating congestion externalities. However, as used today, the term congestion pricing has come to be applied more broadly, to any projects involving tolls or other charges for road use that vary by the level of demand. The degree of variation in such applications can range from simple policies that apply different prices in peak and off-peak periods (with no toll at all being sometimes applied in the off-peak) to rates that change within peak periods, often targeted at maintaining a specific level of service. Variable charges may be applied on the basis of a pre-set schedule (updated periodically) or dynamically (changing as often as every 3–6 minutes) in response to real-time changes in demand.

Variable Pricing in Other Sectors

Although a relatively new concept to highway transportation, charging different prices in peak and off-peak demand periods is a well-established practice in many other sectors and industries. This is especially common in industries in which supply (e.g., the number of available hotel rooms or the capacity of the water delivery system) does not vary significantly in different demand periods. By charging higher rates during high demand periods, proprietors are able to better allocate demand to optimize the utilization of the available capacity.

Examples of such common practices include higher rates for lodging and other amenities in tourist areas during the "high season," discounts for afternoon showings at movie theaters, and evening and weekend discounts for telephone use. More recently, other industries have moved in this direction, including professional sports teams (which have begun charging more for tickets to more desirable games, reflecting long-standing practices in the aftermarket for tickets) and electric utilities (in which advanced electronic meters now allow usage at different times of day to be recorded).

Congestion-pricing projects that involve tolls may be categorized according to their extent of application, as follows:

- **Priced Lanes:** Pricing is applied on a limited number of lanes of a roadway, leaving other travel lanes on the facility unpriced. The priced facility (commonly referred to as express lanes, managed lanes, or high-occupancy toll [HOT] lanes) is typically separated from the other lanes, with limited access points. Examples of priced lanes in the United States include SR-91 in Orange County, CA; I-15 in San Diego; I-394 in Minneapolis; SR-167 in Seattle; and I-10 and US-290 in Houston.
- **Priced Roadways:** Congestion pricing is applied on all lanes of a roadway facility. This is frequently done on existing toll facilities that convert their fixed toll regime to a system that varies by the time of travel. Examples include bridges in Lee County, Florida; the Dulles Greenway in Northern Virginia; and freeways in Singapore.
- **Zone-Based or System-Wide Pricing:** Pricing is applied on all roadways within a specified zone. Pricing can take the form of cordon pricing, in which vehicles are charged as they cross a cordon line into the priced zone, or broader based tolling that is applied to trips within the zone as well. Projects of this type that have been implemented to date have typically focused on congested central business districts, such as those in London, Stockholm, and Singapore; however, future applications may include larger districts, perhaps encompassing entire urban areas.

The *Congestion Pricing Overview* primer includes more information on the practical application of these different approaches.

Priced Lanes: The Economics of Multi-Class Service

The preceding discussion describes the fundamental economic motivation for efficient pricing in its "pure" form. However, as noted earlier, the term congestion pricing has come to be applied to a variety of different practices of applying tolls or other road-user charges that vary based on time and space. One such form is tolled express lanes, in which a freeway is partitioned into two different classes ("priced" and "general purpose"), and tolls are charged on the priced lanes to ensure that traffic remains free-flowing on them. Although similar in some ways to the classical case of congestion pricing, the economic justification for offering multiple classes of service is actually quite different.

HOT LANES

HOT lanes are a special case of tolled express lanes, in which high-occupancy vehicles (HOV; including carpools, vanpools, and transit vehicles) are allowed to use the special lanes for free, whereas low-occupancy vehicles are required to pay a toll to use the lanes. To date, most HOT lane projects in the United States have been conversions from existing HOV facilities.

The purpose of HOV lanes is to encourage more efficient use of highway capacity by increasing average vehicle occupancy, thereby accommodating high levels of passenger travel while reducing total traffic volumes. The lanes operate by establishing minimum occupancy thresholds (typically two or three passengers per vehicle) that are set high enough to keep traffic at low enough levels to ensure that the lanes remain free-flowing, even during peak demand periods. However, because these occupancy minimums must be set at whole numbers, the lanes often wind up with significant excess capacity available, even as the general-purpose lanes may be severely congested.

The purpose of converting such facilities to HOT lanes is to make use of the excess capacity in the HOV lanes, while still preserving the incentives for carpool and transit use (no toll and free flow). For example, several HOV-3 facilities in the United States, such as the reversible express lanes on I-395 in Northern Virginia, were initially constructed as exclusive busways; however, because the buses alone did not fully utilize the available capacity, carpools and vanpools were soon allowed to share those facilities. HOT lanes simply extend this further down the vehicle-occupancy scale, while introducing the element of pricing to manage demand more precisely.

Because most toll-paying users of the HOT lanes are likely to shift from the other lanes, congestion on these lanes will be reduced and travel times will be improved, whereas existing HOV users will see no reduction in the quality of the service they receive. The result is a pure gain to highway users.

Although many HOV lanes operate with excess capacity during peak hours, this is not always the case. In some places, growing traffic volumes have resulted in HOV lanes that no longer provide reliable, uncongested travel to HOVs, even if they did so when first established. One approach to returning such facilities to free-flow conditions is to increase the minimum vehicle occupancy standard for the special lanes, thereby reducing the number of vehicles that qualify to use them. The excess capacity that is created through such a move can then

make the lanes a candidate for HOT lanes—in some cases, additional capacity for HOV lanes (or for HOV-to-HOT conversions) can be created through restriping or other improvements to roadway geometry and operations. By taking these two actions simultaneously (increasing occupancy standards and conversion to HOT), a poorly performing HOV facility can be transformed into a highly functioning tolled facility, while still allowing users who would otherwise be forced off the HOV lanes to continue using them by paying a toll.

BENEFITS FROM PRODUCT DIFFERENTIATION

In addition to the capacity utilization benefits afforded by HOV-to-HOT lane conversions, express lanes can provide highway users with multiple classes of service quality from which to choose at different prices. This form of "product differentiation," as economists refer to it, can produce significant benefits for consumers. The key is that users of a facility may vary significantly in the value that they place on travel-time savings. For some users, reducing travel times for a particular trip may be quite valuable, whereas others may be less sensitive to such reductions.[1] By partitioning the highway into tolled express lanes and free general-purpose lanes, users are given the option of selecting the price–quality combination that best suits them, rather than being forced into a "one-size-fits-all" solution.

It is important to note that the value of time savings reflects to the total value of all passengers in a vehicle, not just the driver. Thus, some of the highest value trips are likely to be those in buses or other transit vehicles, even if the riders individually have lower values of time than single occupant vehicle drivers. In addition, the value of time savings can vary from day to day, even for the same user, depending on the time sensitivity and value of the activities being pursued at the user's destination (such as attending an important family function). This is reflected in the experiences on express lane projects such as SR-91, where a large number of facility users use the express lanes on an occasional basis, when their trips are most critical.

An important component of the appeal that express lanes can have for highway users is their reliability. Variable pricing on express lanes (particularly dynamic pricing based on real-time traffic conditions) can be used to manage demand in such a way as to virtually guarantee a free-flowing trip at all times, not just "on average." Reducing the likelihood and severity of unexpected delays can be very valuable to users and may be an important component of their willingness to pay for the premium service afforded by express lanes. In this way, the introduction of express lanes can be viewed as a new product entirely (a guaranteed high-speed trip during peak hours), previously unavailable at any price (i.e., any price below that of private helicopter taxi services),

that enhances consumer welfare. Such services do exist, of course, for the most extremely high-valued trips, such as emergency trips to trauma centers or ferrying top business executives around major cities.).

IMPACTS ON CAPACITY AND DELAY

Although offering multi-class service can be appealing to road users, it may also have some drawbacks. When a road is partitioned into two facilities, it limits the ability of drivers to change lanes and fill gaps that develop in the traffic stream. Separating the roadways may also require barriers or pavement markings that take up space that could be otherwise used for shoulders or travel lanes. As a result, the combined capacity of the two separated facilities will be lower than the overall capacity that could be realized without the partitioning.

Providing free-flow conditions on the express lanes also requires tolls set sufficiently high to limit the number of vehicles that might want to switch from the free lanes. As a result, traffic and congestion levels may be quite high on the general-purpose lanes. Because delay functions are quite "non-linear" in the sense that each additional vehicle on a congested roadway causes more delay than the previous one, adding more vehicles to the general-purpose lanes causes overall user costs on the combined facility to be higher than they would be if traffic were more evenly allocated.

These capacity and delay-related drawbacks to providing multiple classes of service on freeways make it difficult to make general statements as to whether providing express-lane services will enhance or detract from overall economic welfare. Under some circumstances (particularly where there is significant variation in the value of time for different users), the gains from product differentiation may outweigh the increased congestion costs on the general-purpose lanes, whereas in other cases the latter effect may dominate.

Empirical Estimates of Congestion Tolls

Because congestion pricing is still a relatively new approach for managing congestion, especially in the United States, and has yet to be implemented comprehensively in its "pure" form, empirical estimates of the magnitude of optimal congestion tolls are limited. However, the results from two separate studies of potential congestion-pricing applications have yielded interesting insights into the potential of congestion pricing.

An analysis of future highway investment and performance that used FHWA's Highway Economic Requirements System (HERS) examined the potential impacts of implementing efficient pricing on all congested roads after 2020. The estimated congestion tolls averaged 34 cents per vehicle mile (in 2006 dollars) on roads where it was "implemented" in the analysis, which were predominantly located in major urbanized areas. Over 50 percent of the vehicle miles subject to congestion pricing in the scenario had estimated toll rates below 25 cents per mile, whereas approximately 7 percent had toll rates over $1.00 per mile (see Exhibit 8).

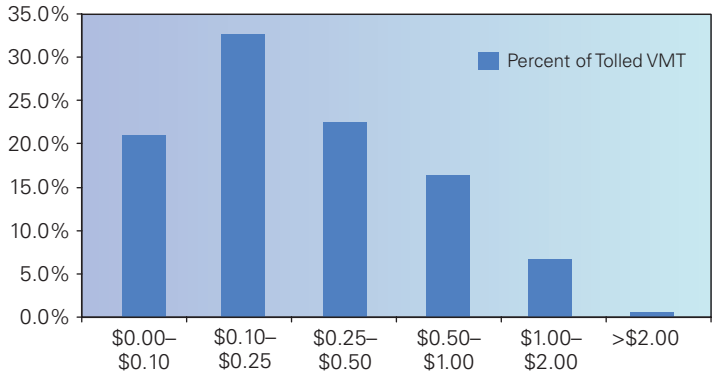

Exhibit 8. Distribution of toll rates per VMT (2020–2025). VMT = vehicle miles traveled.

Source: *Report of the National Surface Transportation Revenue and Policy Study Commission,* Volume III, Paper 6C-04, January 2008.

By using a grant from the Value Pricing Pilot Program, the Puget Sound Regional Council in Seattle recently completed a study[1] of the impact that congestion-based tolling could have on travel behavior. The project involved 275 households whose vehicles were equipped with GPS-based tolling meters. Participants were subjected to a per-mile toll schedule on major arterials in the region that varied significantly based on the time of day (with charges up to 50 cents per mile in the weekday afternoon peak) and the location of travel (with higher rates applied on freeways). The study found that households made several modifications to their travel patterns in response to the congestion tolls, including reductions in the number and length of trips, choosing alternate routes and travel times, and switching to transit. On the basis of the results of the pilot program, the study estimated that a full-scale implementation of congestion pricing on the area's road network could yield over $28 billion in net benefits.

Reference

1. Puget Sound Regional Council. (2008, April). *Traffic choices study: Summary report*. Retrieved on January 1, 2009, from http://psrc.org/projects/trafficchoices/summaryreport.pdf

For more information, contact:

Darren Timothy

Office of Innovative Program Delivery, HIN

Federal Highway Administration

U.S. Department of Transportation

1200 New Jersey Avenue, S.E.

Washington, DC 20590

Tel: 202-366- 4051

E-mail: darren.timothy@dot.gov

U.S. Department
of Transportation

Federal Highway Administration

Office of Transportation Management

Federal Highway Administration

U.S. Department of Transportation

1200 New Jersey Avenue, S.E.

Washington, DC 20590

Tel: 202-366-6726

www.ops.fhwa.dot.gov/siteindex.htm

November 2008

FHWA-HOP-08-041