REPORT | March 2020

HOW TO PRICE CONGESTION: THE BENEFITS OF DYNAMIC VARIABLE TOLLING

Alex Armlovich
Fellow

MI

About the Author

Alex Armlovich is a fellow at the Manhattan Institute, where he writes primarily on urban issues, such as housing, transportation, and infrastructure. He is the author of MI's *Poverty and Progress in New York* series, which benchmarks quality-of-life indicators across NYC. Armlovich has also co-authored several reports with MI's Howard Husock on nonprofit philanthropy and housing policy.

Armlovich's articles have appeared in, among others, the *New York Daily News*, *New York Post*, *U.S. News & World Report*, *Washington Examiner*, and *City Journal*. Previously, he was a Taubman Senior Fellow at the Massachusetts Bay Transportation Authority and interned with the offices of Senator Charles Schumer (D-NY) and with the deputy leader of the U.K. Labour Party. He holds a B.A. in economics and political science from the University of Rochester and an M.P.P. in social and urban policy from Harvard Kennedy School.

Contents

Executive Summary

Congestion in Manhattan's Central Business District (CBD) has reached its worst levels on record, with average daytime traffic speeds falling from 9.1 mph in 2010 to 7 mph in 2019. This is no surprise. Overuse of Manhattan's roads is an entirely predictable outcome: a "tragedy of the commons"—of treating the roads as an unpriced common-pool resource.[1]

The situation should improve with the implementation of tolls for Manhattan's CBD that were authorized in New York State's 2019 budget. The tolls are intended not only to reduce traffic but also to raise enough revenue to support $15 billion in bonds for the Metropolitan Transit Authority's 2020–24 capital budget, with "any additional revenues … available for any successor programs."[2]

Though the tolls have been authorized by the state legislature, the precise details and structure of the policy have not yet been formulated. The city and state, therefore, have a unique opportunity to devise an efficient congestion-pricing system using dynamic tolling. The system would raise revenue by narrowly tailoring variable tolls to actual congestion levels—more "invisible hand" of the price system, less money grab.

The Regional Plan Association (RPA), a prominent NYC-area urban-policy think tank, recently issued a report with recommendations for congestion pricing in New York City. In this paper, I will build on its analysis and propose an alternative policy that more narrowly tailors tolls in order to achieve target traffic speeds in the CBD. While tolling of any sort is an improvement from the unpriced status quo, the more responsive the tolling method is to traffic, the more likely prices are to match supply and demand for road space at any given time. Otherwise, prices will be unnecessarily high when traffic is light, or too low to prevent gridlock when it is heavy.

The goal of NYC's congestion-pricing program should be to reduce traffic congestion with narrowly tailored tolls that nonetheless raise enough revenue to satisfy the authorizing legislation's requirement. To that end, this report proposes that the city set a maximum peak cordon toll cap equal to the highest current regional headline toll of $20 in each direction, with a target speed of 10 mph in the CBD. Overnight congestion tolls should be at or near zero, and, drawing on the experience with high-occupancy toll (HOT) lanes elsewhere in the U.S., the city should implement real-time toll discounts such that drivers pay only the toll necessary to achieve the target policy speed, up to the maximum of $20.

The tolls on each crossing into Manhattan should float independently. For example, if traffic is moving fast on the Queensboro but slow on the Queens-Midtown Tunnel, the tolls on each crossing should temporarily readjust to maintain optimal traffic volumes on each crossing. Because the city already has the ability to track the location of for-hire vehicles (FHVs) such as taxis and Ubers, it should adopt a more efficient per-mile or per-minute congestion fee for FHVs, in lieu of subjecting them to the cordon toll.

Dynamic tolling—adjustment of prices in three- or six-minute intervals to achieve a particular target speed—will charge the minimum necessary toll to achieve the target speed. Rather than using estimates based on historical averages, dynamic pricing will automatically begin relieving tolls the moment that traffic volumes begin to abate for any reason. Transit improvements, recessions, holidays, gas price shocks—anything that causes congestion to decrease for any reason, for a given hour on a given day or over a long period of time, will automatically be reflected in lower tolls.

Furthermore, as long as the toll is permitted to float high enough during rush hour, this approach will raise more revenue than the law's approximately $1 billion annual minimum target. With an initial $20 maximum two-directional toll, net revenues exceed $5 billion annually in a commonly used transportation model[3] —even though overnight tolls are zero. If rush-hour traffic proves to be more responsive to pricing than currently expected, the target policy speed can be increased to ensure that the overall scheme yields enough revenue. If rush-hour traffic is less responsive to pricing, the preset maximum guarantees that prices will not become so high as to produce national headlines and legislative blowback.

In short, the plan presented in this paper will allow the city to maintain target traffic speeds in the CBD, cap the maximum toll charged for entry into the CBD, automatically charge drivers lower or zero tolls during periods of reduced congestion, and raise additional revenue with which the MTA can make critical improvements to transit services for New Yorkers.

HOW TO PRICE CONGESTION: THE BENEFITS OF DYNAMIC VARIABLE TOLLING

Introduction: How to Set Prices in Time and Space

Transportation is more like electricity than it is like other typical goods or services: it is hard or impossible to store, must be consumed in real time, and relies on large fixed infrastructure built long in advance of daily consumption. Its optimal pricing therefore resembles wholesale electricity pricing rather than, say, that of televisions. Optimal price signals need to vary in line with the social scarcity of road space, so prices vary across time and across space.

Time

The following possible schemes are used to express road prices across time:

Flat tolls: Tolls are flat all day, negating one of the main "traffic-shifting" goals of congestion tolling. Under flat tolls, prices will be far too high at night and far too low during rush hours. Nevertheless, it is theoretically possible to set flat tolls equal to the daylong average cost of congestion.

Scheduled variable tolls: Tolls go up or down during the day, based on historical average travel speeds at a given time of day. Prices will be correct on average but will still be too high or too low when congestion is higher or lower than expected because of holidays, events, or other idiosyncratic factors. This approach prioritizes price predictability instead of travel time predictability.

Dynamic variable tolls:[4] Tolls go up or down by intervals as short as three minutes, based on real-time traffic conditions, to maintain a target policy speed. The dynamic approach prioritizes travel time predictability over price predictability.

These pricing approaches are not always mutually exclusive: Singapore's current generation of Electronic Road Pricing uses a mixed approach (see **sidebar** on page 6). Its daily toll schedule is divided into 30-minute increments, which are reset on a quarterly basis, aiming to keep speeds at 45–65 kmph on expressways and 20–30 kmph on arterial roads.[5] The more frequently a toll schedule is adjusted, the closer it approximates a real-time variable toll.

Another mixed approach is used on dynamically variable tolls for HOT lanes in the U.S.: the price varies dynamically in three- to six-minute intervals but is still subject to a maximum toll, regardless of traffic conditions. This delivers certainty about the maximum toll but at the cost of losing some travel time predictability when the toll is limited by the cap.

Space

After determining when tolls will be applied, the next step is to figure out where they will be applied. Again, there are three basic designs:

Cordon tolls: Tolls are charged when crossing a "cordon point." Typically, the cordon is around a central business district (CBD), as in London.[6] In Manhattan, this would mean a toll for entering or exiting the grid

Singapore: Road-Pricing Pioneer

Singapore's Area Licensing Scheme, launched in 1975 with paper tickets, is widely considered the world's first successful urban congestion-pricing scheme.

Today, Singapore's Land Transport Authority (LTA) uses Electronic Road Pricing, a network of electronic gantries that charge scheduled variable tolls along various arterial and highway road segments throughout the city.

LTA employs scheduled tolls but adjusts them in a manner that approximates some of the benefits of dynamic tolling. LTA resets the 30-minute scheduled increments of its daily toll schedule on a quarterly basis, aiming to keep speeds at 45–65 kmph on expressways and 20–30 kmph on arterial roads. The focus on achieving a policy speed, rather than a revenue target, is a particularly useful distinction.

Singapore is preparing Electronic Road Pricing 2.0, an upgrade that will use satellite location tracking of new on-board units in vehicles to enable real-time pricing by distance traveled, not merely by passage of a fixed gantry cordon point.

below 60th Street, per the authorizing legislation, but not for trips that begin and end without crossing the boundary. A trip from 86th Street to 14th Street, therefore, would incur a toll, while a trip from 59th to 14th would not.

Area tolls: Tolls are charged for travel anywhere within an area, not just when crossing a cordon point. In Manhattan, this would mean a toll for driving anywhere below 60th Street, even if the entire trip is within that section of the city. They can be charged as a single toll for any amount of travel, or scaled per mile or minute of travel within the tolled area.

Both area and cordon tolls can be flat, scheduled variable, or dynamically variable.

High-occupancy toll (HOT) lanes: A form of congestion pricing currently used on U.S. highways. Tolls are levied on single-occupant vehicles in a designated highway lane but are free or discounted for transit or carpool vehicles. As implemented in the U.S., tolls are generally dynamically variable, in three- to six-minute increments, to achieve a particular target speed—for example, HOT lanes permitted by the federal Value Pricing Pilot program generally require a minimum average speed above 45 mph 90% of the time.[7]

Dynamic Cordon Tolls Could Work Better than Scheduled Variable Tolls

In its congestion-pricing proposal, the Regional Plan Association (RPA) suggests a set of scheduled variable toll scenarios tailored to raise no more than the minimum $1 billion revenue target of the authorizing legislation. It begins with a baseline scenario of a flat daylong toll with an overnight discount. As it explains, variability in price is necessary because of the changing "social marginal cost" of driving throughout the day and to prevent inefficient "toll shopping" caused by current arbitrary pricing. As RPA details, for any given $1 billion in annual revenue raised, the more dynamic the pricing—higher peak prices during the morning and evening rush hours and steeper discounts midday and overnight—and the greater the reduction in congestion and total social benefit. The plan represents a huge advance over current policy, and RPA provides an excellent illustration of the social dividends of road pricing.

Nonetheless, extending the logic of the argument for variable tolling suggests that dynamic tolls could work even better.

Optimal Tolls Vary More than a Fixed Schedule Can Communicate

Scheduled variable tolls use time of day as a proxy for expected congestion in order to match prices to traffic conditions. Why not cut out the proxy and charge based on real-time, actual congestion, as is done on existing HOT lanes around the country?

On any given day, subway maintenance could cause rush-hour bridge volumes to persist into the evening. An accident on one crossing could make a higher toll necessary to redirect traffic to other crossings. A holiday could cause lighter rush-hour and heavier off-peak travel than usual.

Over months and years, optimal toll levels vary as well. Recessions, gas price spikes, parking price increases, and major transit improvements all reduce the toll levels necessary to achieve, say, a 10-mph average speed into Manhattan (all else equal). Shocks in the other direction increase the toll level necessary to maintain a given speed target. Even the very introduction of scheduled pricing can require subsequent toll adjustment for "peak shifting," as the first round of tolling shifts "peak" traffic into "shoulder" immediately before or after the scheduled peak.

In practice, the most sophisticated jurisdictions, such as Singapore, adjust their toll schedules every three months to keep average speeds on each road close to target. This hybrid approach takes care of the long-term adjustments discussed above but still wouldn't send price signals about unexpected L train maintenance jamming the Williamsburg Bridge overnight—or about any of the one-off idiosyncrasies or random events that can make traffic better or worse.

If Dynamic Tolls Are So Good, Why Haven't They Been Done Already?

These ideas are not new. The FixNYC panel commissioned by Governor Cuomo in 2016 urged the adoption of "dynamic" pricing without giving specific recommendations for how to do so. As far back as the 1960s, Columbia's William Vickrey, a Nobel Prize winner and the intellectual father of congestion pricing, proposed real-time transponder tolling. Those ideas eventually transformed into today's E-ZPass system as well as Singapore's first congestion-pricing policy, the Area Licensing Scheme, in the 1970s. Why, then, haven't real-time dynamic tolls been implemented outside a dozen or so U.S. highway projects?

The most important objection is a political argument: uncapped real-time dynamic tolls could create driver anxiety about price uncertainty and the availability of alternatives once a trip has started. As noted, some HOT lanes in the U.S. do use dynamic tolls, but HOT lanes run parallel to free lanes; when the digital toll for the next HOT lane segment rises to double digits, drivers can simply switch lanes. But what could I do if real-time tolling were enacted on every Manhattan crossing and the toll for the Midtown Tunnel spikes while I'm already on the Long Island Expressway?

The solution is threefold:

- A realistic, stakeholder-calibrated cap on the maximum dynamic toll;[8]
- Aggressive advertisement of price and congestion information on navigation apps, signs, and radio traffic updates; and
- Formation of driver expectations: if you expect traffic, expect to face the maximum toll.

Capping the maximum dynamic toll around $20, the current highest headline toll in the New York region,[9] trades off travel time certainty in exchange for high-end price certainty. Two things must be balanced: the need to send a price signal about road scarcity; and drivers' perception that they are not being egregiously punished beyond their ability to adapt to changing circumstances.

If an emergency on the Queensboro shuts down one of its decks, a price increase is much more effective in redirecting traffic to other crossings than the impotent pleading of a DOT road sign to "choose alternate routes." It's not about punishment; it's about giving everyone skin in the game to motivate collectively coordinated adaptation to an unexpected real-time constraint.

But this social need to coordinate around a strong price signal has a political limit. Uncapped dynamic tolls, though reflecting the true marginal social cost of occupying the road in any moment, could cause national headlines if they rise too far above any other headline toll in the country during an emergency. And a capped maximum price gives people an anxiety-relieving rule of thumb.

Although real-time adjustment is theoretically ideal for matching prices to congestion levels, in practice drivers may only accept a system that offers more time to respond to price signals and make alternative plans. A 15- or 30-minute interval, for example, might give drivers the ability to choose a different crossing, or delay their trip, if they hear reports of abnormally high traffic. Such a system would not be quite as efficient as real-time dynamic tolling, but it would still represent a major improvement over either the status quo or over simple flat tolls.

Recommendations for System Design and Implementation

1. ***Improve driving alternatives before implementing congestion pricing.*** Infrastructure improvements that provide better alternatives to driving, such as mass transit and bike lanes, should be implemented before any congestion-pricing policy goes into effect. In the dynamically variable tolling system proposed here, such improvements will reduce the toll amount necessary to achieve a given target policy speed at any time.

2. ***Begin research on localized social costs of driving beyond congestion for consideration in future upgrades to congestion pricing.*** London has implemented an Ultra-Low Emissions Zone alongside its congestion-pricing scheme, in order to improve air quality. A number of other EU jurisdictions have implemented a mix of differential pricing and outright bans of older vehicles in the densely populated areas—again, with improvements in air quality in mind.[10] This report does not recommend complicating the "day one" launch with this scheme, but it is worth exploring in future system upgrades.[11]

3. ***Design the system to incorporate new technologies that can transition to more area tolling and distance-based pricing.*** The New York State legislation authorizing the tolling of Manhattan's CBD requires cordon tolling, and it permits, but does not require, area tolling. The political and technological burdens of tracking all vehicle movements on the Manhattan grid, by location and time of day, however, make area tolls unlikely on day one. Indeed, RPA was confident enough that the city would begin with cordon tolling that it did not even discuss area tolling alternatives.[12]

   The city should begin with cordon tolling; but to the extent possible, the tolling system design should be procured to facilitate future upgrades to area tolling and even location-based tolling.

4. ***Install congestion-pricing devices to allow for a simple method of identifying vehicles bypassing the zone.*** The day-one launch of congestion pricing is set to employ cordon tolling, not area tolling, for trips that begin and end inside the cordon. Furthermore, through-traffic on the FDR and the West Side Highway that does not enter the Manhattan grid is statutorily exempt from tolling. Rather than deploying toll gantries at every highway exit, exempt through-traffic can be identified by a smaller number of gantries on river crossings and the northern borders of the FDR and the West Side Highway.

5. ***Introduce two-way tolling in the congestion zone.*** One-way tolls provide a price signal only in one direction. For example, a driver might enter Manhattan during the morning rush but leave at midday or overnight. Only two-way tolling can provide the discount for the second leg of the trip taken off-peak, when the toll should be low or zero. Two-way tolling also reduces congestion created by "toll shopping," such as the New Jersey–bound truck traffic on Canal Street in Manhattan.[13]

Recommendations for Pricing

1. ***Vary the congestion fee dynamically in response to real-time traffic volume to achieve at least a 10-mph target policy speed in the CBD, subject to a $20 maximum one-way toll on light-duty vehicles.*** Twenty dollars is below this paper's best estimate of the marginal congestion cost of a personal vehicle-mile of travel in Manhattan during the worst moments of the evening rush hour—but still high enough to reach a 10-mph average speed target within the toll cordon. In the congestion model employed by RPA, one-way weekday cordon tolls to achieve a 10-mph speed target approach $20 for the morning peak and $23 for the evening peak. When capping the toll at $20, average projected CBD traffic speeds dip to 9.9 mph during the weekday evening peak, slightly below the minimum target speed proposed.

2. ***Credit upstream regional tolls broadly.*** Tolls are already charged currently on crossings that enter the cordoned tolling area, like the Lincoln and Holland Tunnels. There are also "upstream" tolls on crossings that do not directly enter the toll cordon, such as the George Washington or the Triboro. Crediting upstream tolls reduces net revenue from the new pricing scheme but reduces unintended toll shopping and enhances perceived regional equity.[14] In a model where maximizing net revenue, not narrowly targeting congestion, is the goal, denying toll credits to bridges that do not directly enter the Manhattan CBD, or the island as a whole, would be one available means of increasing revenue. In this paper's proposal, revenue maximization is not the goal.

3. ***Allow the dynamic toll on each entry point to float independently.*** Toll shopping for free bridges is inefficient when the prices are arbitrary, as they are today. But when prices are set dynamically to match demand to each crossing's capacity, toll shopping can actually be helpful in redirecting traffic to available road space. If a crash or other idiosyncratic factor clogs one crossing more than the others, the dynamic pricing algorithm will passively begin raising the price on that crossing to direct incoming traffic to other crossings.

4. ***Swap the fixed "congestion fee" on taxis and for-hire vehicles for a dynamic per-mile toll equal to the average of the cordon tolls on each entry point.*** The best model of congestion pricing, with dynamic pricing based on distance traveled in real-time traffic conditions, is possible only with detailed location tracking. While infeasible for private vehicles on the day-one launch, taxis and for-hire vehicles (FHVs) are already ubiquitously tracked

and priced by distance and time. This paper proposes that the new system take advantage of this existing technology to implement a dynamic per-mile fee for taxi and FHV travel below 60th Street equal to the average of dynamic cordon tolls entering the city, with the same maximum of $20.

The current fixed fee of $2.75 for solo FHV pickups is low and unconnected from the congestion externality. As shown in the Appendix, even a basic speed-volume model of Manhattan congestion yields a congestion externality[15] estimate of about $13 per mile at the average daily CBD travel speed of 7 mph.

Conclusion:

Capped Dynamic Cordon Tolls Are a Better Way to Reduce Congestion and Raise Revenue, but They Also Minimize Excessive Tolling

Congestion pricing ought to be primarily about reducing congestion, not just raising revenue. Real-time dynamic tolling automatically raises or lowers tolls to align supply and demand for road space for reasons that we as planners cannot, and do not even need to, foresee—but it never levies a toll just for revenue's sake.[16] Tolls should not price drivers off the road when the roads aren't full. The whole point of high-fixed-cost infrastructure is to maximize use without ruining quality.

Drivers and passengers on high-value or inflexible trips—plumbers, contractors, delivery people losing appointments to long travel times, job seekers headed to an interview, travelers at risk of missing flights—should find value in high-dollar tolls at busy times, knowing that the toll sent a signal to others to use available alternatives and reduce congestion. Bus riders on routes without dedicated lanes, currently stuck in slow traffic, obviously benefit from faster bus travel as well—indeed, many who currently drive might be tempted to switch to express bus service. Drivers on flexible trips who shift their travel to midday or overnight likewise deserve their low- or no-toll trip.

Rather than estimating optimal tolls based on historical traffic patterns, dynamic pricing will automatically begin relieving tolls the moment that traffic volumes begin to abate for any reason in real time. Any time traffic improves—for whatever idiosyncratic, unforeseen reason—tolls will fall. And if traffic conditions worsen on a particular route, price signals are the necessary prerequisite for efficient collective action in response to new conditions.

Although revenue should not be the main goal of a congestion-pricing regime, dynamically variable pricing of the kind proposed here will raise more revenue than the approximate $1 billion annual minimum target of the law. With a $20 maximum two-directional toll, net revenues exceed $5 billion annually—even though overnight tolls are zero.[17]

Appendix

What Is Congestion Pricing? Different Methods with Similar Goals

The key intuition of congestion pricing is that each additional driver on a given segment of road at a given time slows down all the other drivers by some small amount. However, each driver chooses a trip based only on his own travel time, not on the tiny incremental delay to all other drivers on the relevant road segment. In order to have drivers choose their trips optimally, a public or private road operator must set a toll for access to the road segment that is equal to the slowdown cost imposed on all other drivers at any given time. The slowdown cost imposed by one driver on all others is called the "congestion externality." This congestion framework is called the "speed-volume" model of congestion, in which increasing volume brings decreasing speed.[18]

In a theoretically ideal congestion-pricing scheme, every road segment would be charged in real time, based on the real-time congestion externality. This price for each vehicle on each road segment would be calculated from the known volume capacity of the road segment, the size of the vehicle in "passenger-car equivalents,"[19] and real-time traffic conditions. Singapore, in its Electronic Road Pricing 2.0 procurement, is developing the satellite tracking and on-board units for vehicles that would be necessary to accomplish universal road pricing. But the logistical and political barriers of such an effort have thus far led other regions to pursue second-best versions of congestion pricing.

To illustrate the intuition behind the flow model of congestion, consider the speed-volume model from Charles Komanoff's Balanced Transportation

Analyzer, fit to the number of vehicles traveling one mile simultaneously in Manhattan below 60th Street, depicted in **Figure 1**.

When traffic is flowing smoothly at 24 mph, and there are only about 40,000 cars traveling a mile in Manhattan, adding one more driver doesn't slow traffic down much and only affects those 40,000 vehicle-miles. Adding one more driver slows down the 40,000 other cars by 0.0000018 minutes per mile, for a total delay of 0.077 minutes. Not much of a social cost there.

When, however, there are ~234,000 drivers and traffic is gridlocked at 2 mph, adding one more driver-mile slows traffic by a larger, but still seemingly small, 0.004472 minutes per mile. But now that tiny slowdown also applies to about 234,000 other drivers, for an aggregate delay of 106.5 minutes. If the travel time value of Manhattan drivers is $30/hour, as in this example, those 106.5 minutes of delay from an additional driver cost $55.85. In the best of all worlds, drivers should drive only an additional mile when Manhattan traffic is at 2 mph if they value that mile trip in excess of the $55.85 delay imposed on the street network. In this way, we can express a complete schedule of optimal per-mile tolls for any given set of traffic conditions.

In an ideal policy, cities would use traffic engineers' knowledge of the volume and speeds of traffic passable in any given road segment to set the toll for that road segment in that moment equal to its marginal social cost. Singapore's bid solicitation for the next generation of congestion-pricing equipment will use satellite-assisted location tracking, known as Electronic Road Pricing 2.0, and will come very close to accomplishing this.

FIGURE 1.

The Social Cost of, and Optimal Toll for, an Incremental Vehicle-Mile Traveled at Various Levels of Manhattan Congestion[20]

MPH	Total Number of Other Drivers (Drivers-Miles)	Increased Minutes per Mile, All Traffic, from Adding 1 More Driver	Changes in Minutes for All Other Drivers Caused by Adding 1 More Driver	Value of Time per Minute	Social Cost & Optimal Toll to Travel 20 Blocks	Social Cost & Optimal Toll to Travel 40 Blocks
24.0	37,816	0.0000018	0.077	$0.52	$0.04	$0.08
23.0	60,505	0.0000074	0.476	$0.52	$0.25	$0.50
22.0	71,849	0.0000124	0.936	$0.52	$0.49	$0.98
21.1	79,413	0.0000167	1.391	$0.52	$0.73	$1.46
20	86,976	0.0000220	1.996	$0.52	$1.05	$2.09
19.4	90,757	0.0000250	2.364	$0.52	$1.24	$2.48
18.1	98,320	0.0000318	3.250	$0.52	$1.70	$3.41
17.4	102,102	0.0000357	3.778	$0.52	$1.98	$3.96
15.9	109,665	0.0000443	5.025	$0.52	$2.63	$5.27
15.1	113,447	0.0000491	5.753	$0.52	$3.02	$6.03
14.4	117,228	0.0000542	6.559	$0.52	$3.44	$6.88
12.9	124,791	0.0000655	8.424	$0.52	$4.42	$8.83
12.2	128,573	0.0000717	9.494	$0.52	$4.98	$9.95
11.5	132,354	0.0000783	10.663	$0.52	$5.59	$11.18
10.2	139,917	0.0000927	13.325	$0.52	$6.99	$13.97
9.0	147,480	0.0001088	16.461	$0.52	$8.63	$17.26
6.9	162,607	0.0001465	24.370	$0.52	$12.78	$25.55
6.1	170,170	0.0001682	29.261	$0.52	$15.34	$30.68
5.0	181,514	0.0002048	37.942	$0.52	$19.89	$39.78
4.1	192,859	0.0002463	48.440	$0.52	$25.40	$50.79
3.0	211,767	0.0003277	70.634	$0.52	$37.03	$74.06
2.1	234,456	0.0004472	106.531	$0.52	$55.85	$111.70

Chart: Manhattan Institute

Note: Highlighted row represents current average speed in Manhattan CBD / Source: Author's calculations based on Charles Komanoff, Balanced Transportation Analyzer

Endnotes

[1] Road pricing is a special thing in public finance. It raises revenue without imposing a new net cost on society, as higher income or sales taxes would do. The total price of driving in Manhattan is already high, whether roads are priced with money or not—the Appendix slows how high the current travel time losses are in typical traffic conditions. Drivers can pay with time or with money, but wasted time is a pure loss to society, whereas tolls create transferable revenue that can be put to some other useful purpose, like transit investment. This is why business groups like the Partnership for NYC support road pricing. Residents and businesses are already losing billions of dollars' worth of time to congestion, so converting those time losses into toll revenue is a very attractive alternative to new income or sales taxes.

[2] The Laws of New York, Title 8, Article 44-C, Section 1704-A: Central Business District Tolling Program.

[3] Charles Komanoff, "Balanced Transportation Analyzer." This model is used by RPA in its analysis of scheduled variable tolls.

[4] U.S. Department of Transportation (DOT), Federal Highway Administration, "Congestion Pricing."

[5] Singapore Ministry of Transport, "How ERP Works as a Speed Booster."

[6] The cordon does not always wrap around an enclosed area; it can also be as simple as a toll on a single bottleneck point on a highway, as in part of Stockholm's system.

[7] DOT, Federal Highway Administration, "Federal-Aid Highway Program Guidance on High Occupancy Vehicle (HOV) Lanes," September 2016.

[8] There are policy grounds on which one could argue for cordon tolls over $100 during the evening peak in the Balanced Transport Analyzer model, when delays on the unpriced regional road network for trips en route to the CBD are attributed to the average CBD trip. In the best of all worlds in universal road pricing, each congested road segment gets a custom price. But in a second-best world, where all the dollarized congestion time losses from a Manhattan-bound Connecticut driver can be charged only at the border of Manhattan, quite a large bill is accrued. This paper takes it as politically impossible to implement a cordon charge at the Manhattan border for congestion costs imposed on New Jersey, Long Island, and Westchester roads by Manhattan CBD-bound trips.

Additionally, as noted in the Appendix, even in a simple model of Manhattan alone, without hypercongestion, the aggregate congestion delays caused by driving when traffic is jammed at 5 mph approaches $40 *per mile* traveled inside the CBD. A $20 cap *per cordon entry* is therefore on the low end of any plausible estimate of the typical daytime congestion externality.

[9] See, e.g., Tamar Lapin, "Verrazzano-Narrows Bridge Is Now Most Expensive Toll in U.S.," *New York Post*, Mar. 31, 2019.

[10] "EU: Low Emission Zones (LEZ)," DieselNet.

[11] See Appendix.

[12] RPA, "Congestion Pricing in NYC: Getting It Right," September 2019.

[13] Sydney Pereira, "Verrazzano Study Verifies: 2-Way Toll Would Slash Traffic," *The Villager*, July 19, 2018.

[14] See RPA, "Congestion Pricing in NYC," p. 16, for a detailed discussion and infographic of possible upstream toll credit schemes; the revenue trade-off is a concern only if a narrow $1 billion revenue goal is the primary policy objective, unlike this paper's 10-mph speed target. Crediting Port Authority tolls on the Lincoln and Holland Tunnels, for example, enhances regional equity by not double-tolling New Jersey drivers—and furthermore ensures that prices are narrowly targeted to congestion reduction, not to raising new revenue.

[15] An externality is generally any unpriced social cost not captured by participating parties to a market transaction. In this case, it is the external cost of congestion as derived in the Appendix.

[16] Indeed, the logic of this paper suggests that the Port Authority and existing MTA bridges should also adopt dynamic tolling, since overnight and midday drivers do not deserve the excess toll burdens that they currently face.

[17] As calculated in Komanoff, "Balanced Transportation Analyzer," the model used by RPA.

[18] See n. 20 below; this model doesn't feature "hypercongestion," the state of extreme gridlock in which, at the highest volumes and lowest speeds, the street grid's supply curve bends backward such that volumes begin to decline again as speed further declines. This is an evolving empirical research area, and, as a model feature, hypercongestion models produce more favorable outcomes from the introduction of pricing—by ending solid gridlock, it actually allows more vehicles to use the same road space. By building optimal price estimates without this favorable model feature, I ensure a fortiori that the speed target is achievable at the proposed prices while leaving open the possibility of an upside surprise in the effectiveness of this pricing scheme.

[19] Orazio Giuffrè et al., "Passenger Car Equivalents for Heavy Vehicles at Roundabouts: A Synthesis Review," *Frontiers in Built Environment* 5, no. 80 (June 18, 2019).

[20] This is a speed-flow model of Manhattan without the backward-bending supply curve at "LOS F," colloquially known as "hypercongestion." It is intended to provide intuition and illustrate the ballpark of the damage done by an incremental mile traveled in Manhattan under varying conditions. The actual pricing model for each crossing would use the more sophisticated "queued bottleneck" approach fit to each crossing's physical capacity. See, e.g., Michael L. Anderson and Lucas W. Davis, "An Empirical Test of Hypercongestion in Highway Bottlenecks," working paper, January 2020.

March 2020

MANHATTAN
INSTITUTE