How to Price Congestion: The Benefits of Dynamic Variable Tolling

Alex Armlovich

Contents

Executive Summary

In 2019, congestion in Manhattan's Central Business District (CBD) reached its worst levels on record, with average daytime traffic speeds falling to 7 mph from 9.1 mph in 2010. 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.[i]

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."[ii]

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. This system would raise revenue by narrowly tailoring variable tolls to actual congestion levels, using the "invisible hand" of variable price signals to achieve consistent travel speeds in the CBD."

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. While policymakers will ultimately have to decide on the minimum target speeds and the requisite congestion charges necessary to achieve them, this report focuses on the impact of two different tolling scenarios: one in which the maximum peak cordon toll is capped at our best estimate of the "economically optimal dynamic toll"[iii] and another in which it is capped at a lower but perhaps more politically feasible price.

With a projected target speed of 10 mph in the CBD, an optimal dynamic toll would briefly go up to a peak of $26 in each direction during rush hours. Mid-day tolls between the morning and evening rushes would average $1.75 each way or $3.50 round-trip, while overnight congestion tolls should be at or near zero. A peak toll of $26 in each direction would mean a maximum round-trip toll of $52 but most drivers would pay far less: the weighted average round-trip toll would be about $18. 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 projected target policy speed, up to the maximum cap of $26.

This report also considers a peak toll that is considerably lower than the economic optimum and consistent with a slower target speed: a peak toll of $15 in each direction (which yields a weighted average round-trip toll of about $4) and a rush-hour speed target of 8mph—illustrating the model's best estimate of the policy tradeoffs of a lower but perhaps more politically feasible peak toll. There is a prudential case to be made for careful, incremental change with respect to tolling. The tolling framework presented here relies on flexible, automatically responsive pricing to accommodate unanticipated reductions in

congestion—including prolonged reductions like the one that the city is now experiencing due to the Covid-19 pandemic.

It is also true that the costs of current road policy are extremely high. These costs include travel time losses (analyzed in this paper) as well as costs that are outside this paper's scope, such as lengthy emergency vehicle response times or unpriced vehicle exhaust emissions in the nation's most densely populated area. While a policy response—and tolling—commensurate with the scale of New York's congestion problem may be advisable and achievable in the long run, it would be unwise to overreach on such a solution before a broad public consensus is achieved.[iv]

Regardless of the toll levels chosen, dynamic 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 tracks 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. Such a fee would also obviate less-efficient restrictive regulations on FHVs, such as the onerous cap on the number of FHV drivers.

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 traffic volumes & 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. Under the economically optimal dynamic tolling scenario with a target speed of 10mph and a weighted average round-trip toll of about $18, revenues exceed $5 billion annually in a commonly used transportation model even though overnight tolls are zero.[v] This revenue figure would be a bit lower if, as this paper recommends, there is broad crediting for upstream tolls (such as those paid by drivers crossing the George Washington bridge prior to entering the CBD). The RPA estimates that upstream toll credits cost about $120 million annually. In the lower-toll scenario with a rush-hour speed target of 8MPH and a weighted average toll of about $4.37, estimated revenues are roughly $2 billion annually before accounting for broader toll credits.

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, a preset maximum toll cap will provide some pricing certainty during rush hour and help to avoid the legislative blowback that would be likely to result from uncapped surge pricing.

In short, the plans 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 that, if used appropriately (see Appendix II), would allow the MTA to make critical improvements to transit services for New Yorkers.

This paper's model was written for and calibrated to pre-COVID pandemic traffic volumes and speeds. As such, the annual revenue projections apply only to a medium-term return to normal. But the lower revenues consistent with any given target speed during this period of reduced economic activity are a feature, not a bug of this paper's "congestion reduction over revenue" approach: Daytime average tolls calibrated to traffic speeds automatically fall in a deep recession and only recover when traffic volumes recover and begin to drive travel speeds below the target once again[vi].

Introduction: How to Set Prices in Time and Space

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.

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.

Tolling Across 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:[vii] 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 4). 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.[viii] 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.

Tolling Across 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.[ix] In Manhattan, this would mean a toll for entering or exiting the grid 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.[x]

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" periods.

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;[xi]
- 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.

Capped dynamic tolls 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 high during an emergency. And a capped maximum price gives people an anxiety-relieving rule of thumb.

Recommendations for System Design and Implementation

Improve driving alternatives before implementing congestion pricing.

To the greatest extent possible given the tight timeline for the implementation of congestion pricing (originally January 2021), infrastructure improvements that provide better alternatives to driving, such as mass transit and bike lanes, should be implemented before the 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.

Advance complementary markets in parking to relieve the burden on tolling to achieve congestion reduction.

Travel lanes aren't the only unpriced commons on the road: On-street parking is also a scarce, overused open commons[xii]. Cruising for over-occupied free parking adds to congestion directly. But there's more to it than that: Another key part of the reason the optimal Pigouvian toll for travel is typically so high in Manhattan is that substantial valuable land in the CBD is given over as an in-kind subsidy for free car storage, even as the market rate for private parking approaches $30 per day. Parking subsidies of all kinds matter: In a 2006 study following his report on congestion pricing for the Manhattan Institute, Bruce Schaller estimated that "free parking" placards given to municipal employees nearly double the share of city employees who drive to work in Manhattan.[xiii]

Tolls and parking costs interact in setting the total cost of a rush hour trip by car versus other modes or car trips at other times. To get weekday morning 6am to 9am Manhattan traffic speeds to 10MPH, we need to shrink hourly traffic volumes by 11% from the pre-Pandemic baseline. To get speeds up to 20MPH for the same period, we'd need to shrink traffic volume by a whopping 43%. The lower the average parking fees in Manhattan, the higher the average congestion tolls necessary to achieve the hourly volume reductions corresponding to any given speed target.

A system of residential permit auctions, and dynamic metered parking on commercial streets, would be efficiency-enhancing on its own terms and reduce the dynamic toll necessary to achieve any given speed target. But exposing street parking to market discipline would also improve the politics of congestion pricing by reducing typical toll levels and by protecting parking garage owners from the up to 50% decline in the price of off-street parking that Pigouvian tolls might otherwise cause.

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.[xiv] This report does not recommend complicating the "day one" launch with this scheme, but it is worth exploring in future system upgrades.[xv]

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 MTA would begin with cordon tolling that it referred consideration of area tolling alternatives to future research.[xvi]

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 distance-based tolling within the zone.

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.

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.[xvii]

Recommendations for Pricing

Vary the congestion fee dynamically in response to real-time traffic volume to achieve a specific target policy speed in the CBD, subject to a maximum one-way toll on light-duty vehicles.

Twenty-six dollars is 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—and we expect a peak toll at that level to achieve a 10-mph average speed target within the toll cordon.[xviii] In the congestion model employed by RPA, in order to achieve a 10-mph speed target, one-way weekday cordon tolls must approach $20 for the morning peak and $26 for the evening peak. While the maximum round-trip toll would be $52 under this scenario (the rare case in which a driver enters and leaves the CBD during rush hour), most drivers would pay far less—the weighted average round-trip toll would be about $18, yielding $5.2 billion before wider toll credits (to be detailed in the next recommendation).

A lower speed target would correspond with lower maximum and weighted average tolls. For example, a slower target speed of 8MPH (modeled in Appendix III) would bring a maximum round-trip toll of $30, with a very low weighted average round-trip toll of about $4, yielding $2 billion before wider toll credits.

For context on the notion of double-digit round-trip tolls for driving to and from Manhattan, consider the commuter rail fares that current and future LIRR riders are expected to pay: From Jamaica, Queens, the round-trip peak commuter fare is $21.50. From Far Rockaway, Hempstead, and Port Washington, it's $25. The next fare zones are $28, $33.50, $39.50, $47, and finally $61 round-trip from Montauk or Greenport.[xix] LIRR's weighted average round-trip fare in 2018, including discounted "non-commutation" trips, was slightly over $17[xx], while the weighted average round-trip fare on Metro-North was $18—both near the $18 weighted average toll modeled in our highest-toll scenario, and far above the weighted average toll in the lower-toll scenario.

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.[xxi] For example, imagine a driver who crosses the GW Bridge right at the 9 AM peak and leaves the CBD at 8PM. On the way in, this driver would pay the $13.75 Port Authority toll plus another $10.75 for a net of $26 to enter the CBD from New Jersey. On the way out, the driver would likely pay no toll since, by 8PM, the congestion toll necessary to achieve the 10pmh speed target in the Manhattan CBD will typically be low or zero and the Port Authority only charges tolls on the GW Bridge in one direction. The net round-trip toll for such a trip is $26 compared to $13.75 today—but without the toll credit, that same trip would impose an unduly strong price signal to NJ commuters that is not necessary to correct the congestion externality. 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.

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.

Swap the fixed "congestion fee" on taxis and for-hire vehicles for a dynamic per-mile toll equal to the average per-mile equivalent 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 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 per-mile equivalent of the dynamic cordon tolls entering the city, with the same illustrative maximum of $26.

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[xxii] estimate of about $13 per mile at the average daily CBD travel speed of 7 mph.

Subjecting FHVs to this dynamic fee tailored precisely to the congestion externality would also obviate the existing clumsy & heavy-handed regulations that indirectly target FHV congestion, especially the cap on FHV licenses. The use of price to control traffic volumes hour-by-hour strictly dominates the annual quantity cap approach. Like the archaic yellow medallion system it replicates, the FHV cap does little to deter rush hour traffic volumes in Manhattan but does unnecessarily reduce FHV availability overnight and outside the congested CBD.

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 planners cannot, and do not even need to, foresee—but it never levies a toll just for revenue's sake.[xxiii] 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 considered here will raise more revenue than the approximate $1 billion annual minimum target of the law. Under the scenario with a $26 maximum one-way toll, net revenues (inclusive of approximately $120 million in broad toll credits) should approach $5 billion annually versus the pre-Pandemic traffic baseline—even though overnight tolls are zero.[xxiv] Under the scenario with a $15 maximum one-way toll, net revenues are nearly $2 billion after broader toll credits. Until the region's economy and traffic volumes recover, rush hour tolls are unlikely to hit the cap and revenues are unlikely to significantly exceed the $1 billion statutory minimum. That economic sensitivity is a feature, not a bug, of a focus on congestion outcomes instead of revenue. On the other hand, Beijing's experience with post-COVID reopening anecdotally suggest car traffic recovers before transit ridership[xxv] as partial social distancing guidelines remain in effect. But if necessary due to a prolonged recovery of traffic volume, the MTA could explore an interim speed target higher than modeled here in order to ensure revenue collections at least hit the statutory minimum of $1 billion.

Again, this paper offers a modular set of separable recommendations. It is intended to illustrate the advantages of dynamic tolling, and the benefits of setting the toll cap near our best estimate of the socially optimal toll on travel inside the Manhattan CBD for both private and For-Hire Vehicles, while avoiding carveouts and loopholes for non-transit vehicles. The focus on congestion reduction—targeting a policy speed instead of revenue—has the further benefit of robustness to unforeseen shocks to car travel demand, from the trivial to the pandemic. This paper's illustrations nonetheless represent one proposal within a matrix of possible tradeoffs—a higher target speed will bring higher tolls, more revenue, and of course faster traffic. A lower speed target will bring the converse. A continuum of tradeoffs also exists between travel time certainty and price certainty: Policymakers could also choose the "hybrid approach" described in the Singapore sidebar, where scheduled tolls are adjusted every three months to achieve the average speed target—this would still be an improvement from a static schedule lacking a defined speed target as policy goal.

In sum, this paper offers four novel suggestions for congestion pricing in New York City:

1. A speed target should be at the center of policy, rather than revenue, and the fee should vary dynamically in response to traffic volumes to achieve the target speed subject to a maximum peak toll. (Subject to political constraints, the closer the peak toll can get to $26, the better the economically estimated balance of speed and toll price).
2. Upstream tolls should be credited broadly to increase regional equity and ensure the incentives to choose any given route to Manhattan depend only on traffic management, not on revenue considerations.
3. Dynamic tolls on each entry point to Manhattan should float independently to incentivize only useful "toll shopping". Current toll differences on priced and unpriced crossings are arbitrary and divert traffic to the busiest free crossings, while independently floating tolls would equilibrate to balance traffic volumes across all crossings.
4. The cap on licenses for For-Hire Vehicles should be removed and the fixed pickup tax should be replaced by a dynamic per-mile toll tailored directly to the per-mile congestion externality. While distance-based charging inside the cordon would work better for all vehicles in theory, in practice the privacy and political challenges of vehicle tracking have only yet been solved for taxis and FHVs.

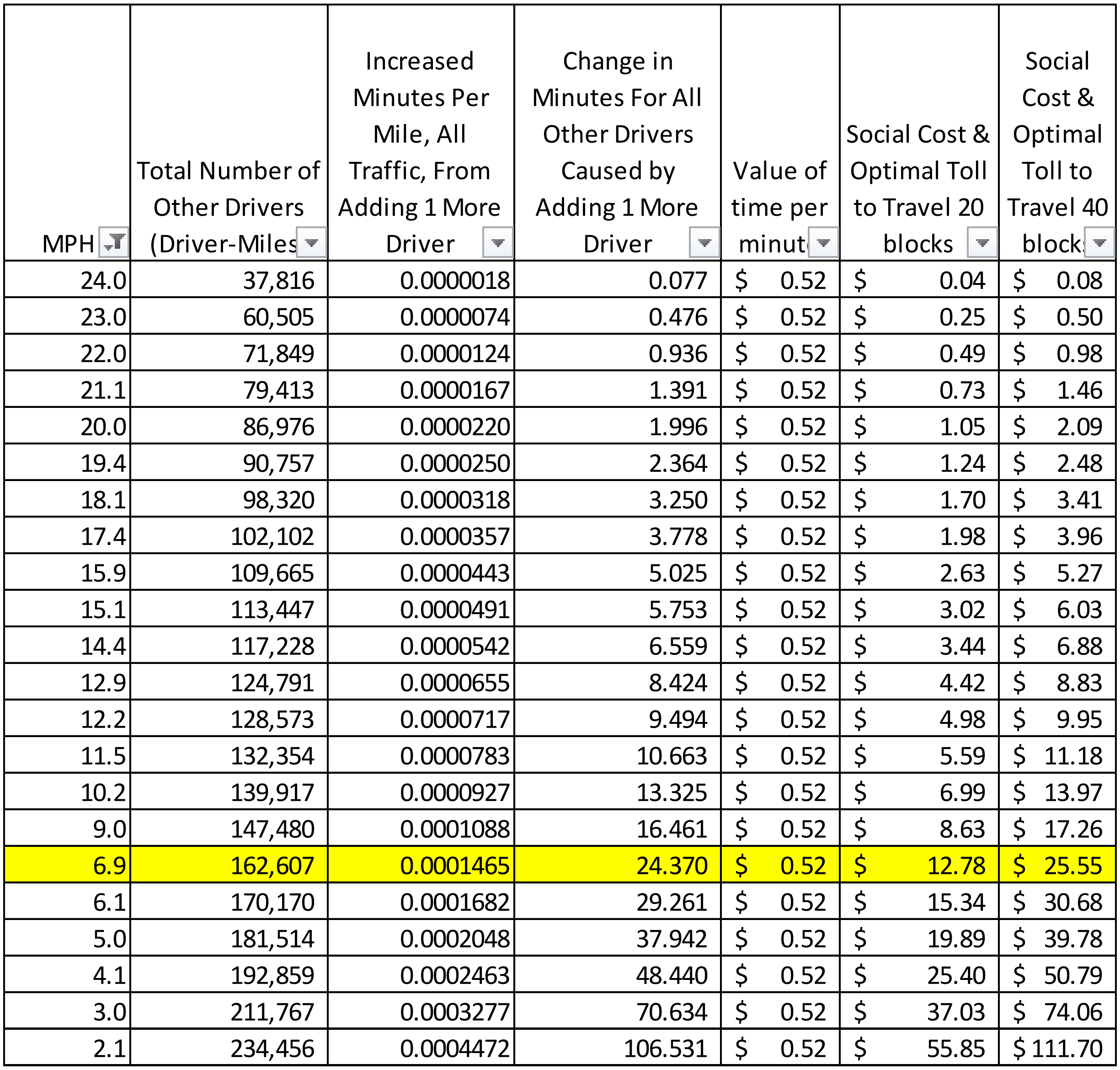
Appendix I. What Is Congestion Pricing? Different Methods with Similar Goals

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.[xxvi]

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,"[xxvii] 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 this 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:

Figure 1. The Social Cost of, and Optimal Toll for, an Incremental Vehicle-Mile Traveled at Various Levels of Manhattan Congestion[xxviii]



MPH	Total Number of Other Drivers (Driver-Miles	Increased Minutes Per Mile, All Traffic, From Adding 1 More Driver	Change in Minutes For All Other Drivers Caused by Adding 1 More Driver	Value of time per minut	Social Cost & Optimal Toll to Travel 20 blocks	Social Cost & Optimal Toll to Travel 40 block
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
20.0	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

Source: Charles Komanoff, Balanced Transportation Analyzer

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.

Using the September 2019 BTA's speed-volume curve for Manhattan, we can also illustrate how much the hourly traffic volumes need to change in order to achieve the volume consistent with an average traffic speed. Manhattan's grid can deliver average speeds of about 20MPH as long as average hourly VMT stays below 91,000. At 10 MPH, it can handle 143,000 VMT. Prices vary to achieve the traffic volumes corresponding to policy speed target.

	11pm-5am	5am-6am	6am-9am	9am-10am	10am-2pm	2pm-8pm	8pm-11pm
Hourly Pre-Pandemic VMT	**72,672**	**84,940**	**160,753**	**160,844**	**145,895**	**172,834**	**150,693**
Hourly Average Manhattan VMT @ 20mph	91,000	91,000	91,000	91,000	91,000	91,000	91,000
Volume decrease from baseline to hit 20MPH	N/A	N/A	43%	43%	38%	47%	40%
Hourly Average Manhattan VMT @ 10mph	143,000	143,000	143,000	143,000	143,000	143,000	143,000
Volume decrease from baseline to hit 10MPH	N/A	N/A	11%	11%	2%	17%	5%

Appendix II. What should the MTA do with up to $5+ billion per year?

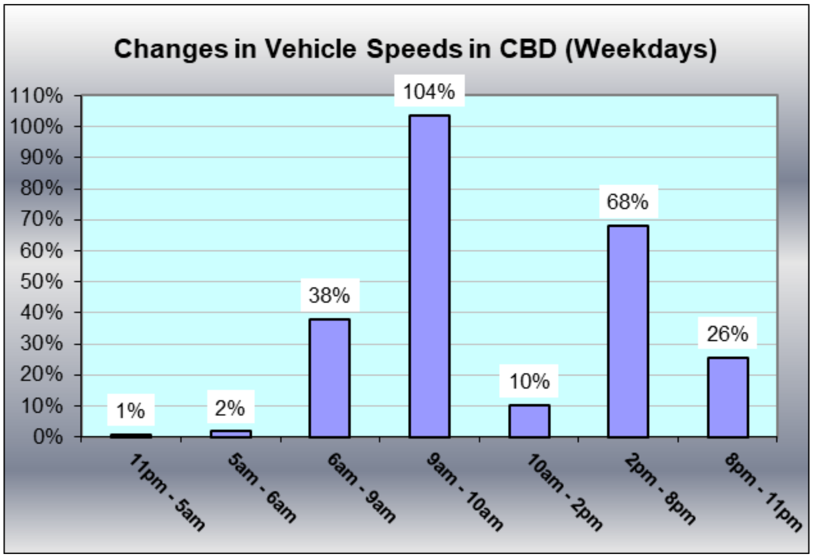
In the long run, the MTA faces three broad management challenges: Capital commitment, capital efficiency, and operating efficiency. The capital commitment problem is the MTA generally struggles to get capital budget money "out the door" on time within the window of each capital plan. The capital efficiency problem is that the MTA spends up to 10X the rate of its global peers on comparable capital projects when it does manage to actually commit its capital budget. Last, the operating efficiency problem is that the MTA runs its day-to-day business much less efficiently than it did in the mid-20th Century[xxix], and less efficiently than peer agencies today.[xxx] So while congestion reduction is valuable as an end in itself, there is ample reason to wonder about the MTA's ability to productively spend new money. For these reasons we propose that congestion pricing proceeds be used to self-fund the MTA's capital plan in lieu of currently planned discretionary state and local appropriations from general funds. All other excess proceeds can be used to restructure existing MTA debt, the annual service on which is projected to exceed $3 billion annually.

Assumption of debt service will indirectly relieve the operating budget in order to help bridge the immediate pandemic-related operating budget crisis. Once the regional economy has recovered, the operating funds freed up by retiring debt will support service-positive bus network redesigns and other service increases to accommodate riders attracted from congestion-priced roads. Depending on the mix of operating budget increases versus accelerated debt retirement in any restructuring, an accelerated drawdown of the MTA's current debt burden will provide the requisite borrowing space for future capital plans following reforms of MTA's capital procurement process. The looming operating deficit should also be closed in this process.

In short: The MTA should use congestion pricing proceeds to retire debt and indirectly relieve the operating budget of debt service immediately, and eventually obviate discretionary general fund appropriations to future capital plans. This should be done both to restore its borrowing capacity in anticipation of future capital procurement reform, and to employ fungible operating funds freed by restructuring debt service to close the looming operating deficit—all while still expanding bus service frequencies in pending bus network redesigns.

Appendix III. Alternatives Analysis:

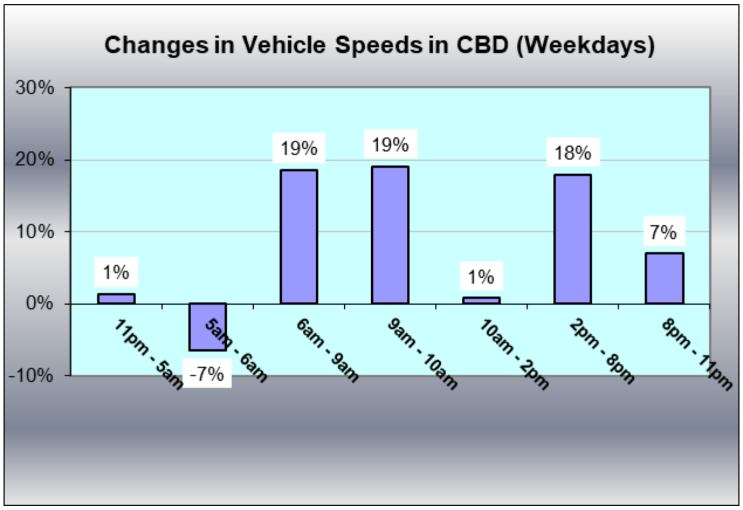
Higher Dynamic Tolls $26 Pigouvian Toll Cap, 10MPH speed target at all other times:



Revenue: $5 billion after wider toll credits
Weighted Average Round-Trip Toll: $18
Max Round-Trip Toll: $52.18

This scenario allows the cordon toll to float as high as the known maximum rush-hour congestion externality per mile of travel within the Manhattan grid. Based on the model used for estimates in this paper, this price cap will correspond to a 10MPH speed target. The FHV cap and fixed trip fee are replaced with a dynamic per-mile charge for FHV and taxi travel below 60th Street, in lieu of the cordon toll. Tolls on existing MTA and Port Authority tolled crossings are credited for trips that enter Manhattan below 60th Street.

Scheduled Variable Tolls: RPA Scenario D w/status quo pickup fee on FHVs:



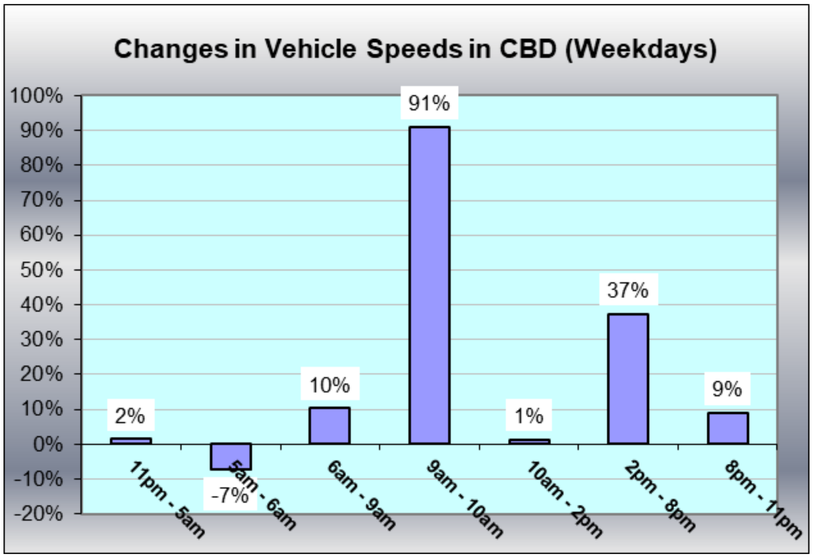
Revenue: $1 billion after wider toll credits
Weighted Average Round-Trip Toll: $10.97
Max Round-Trip Toll: $18.36

This is our replication of the RPA's Scenario D of scheduled tolls without any particular speed target. The FHV cap and flat pickup fee remain in place, and there is no cordon toll or per-mile charge for taxis or FHVs. The RPA does not commit to any particular toll credit scheme but expects broad toll credits to transfer roughly $120 million annually, which we included in the net revenue estimate here.

The primary observable difference in this scheduled toll plan is the lack of the dynamic per-mile FHV and taxi charge. The other substantive differences are harder to model, because the largest differences arise precisely in the cases when the real world differs from our model expectation. If the world could be perfectly captured in a spreadsheet model and always behaved the same way, we could design the perfect toll schedule down to the minute in advance. Indeed, a narrowly tailored dynamic price that "passively" floats to accommodate daily, seasonal, or longer-term recessionary traffic shocks, subject to some politically tolerable toll cap, is most different from a scheduled variable toll precisely in the states of the world that differ from what we predict in advance.

Lower Dynamic Tolls

8MPH Speed Target At All Times, FHV fee at half the Pigovian per-mile rate:



Revenue: Nearly $2 billion after wider toll credits
Weighted Average Toll: $4.37
Maximum Round Trip Toll: $30

This scenario modifies the "Higher Dynamic Toll" scenario to generate a projected minimum rush-hour speed of 8MPH. The FHV cap and fixed trip fee are replaced with a dynamic per-mile charge for FHV and taxi travel below 60th Street, equal to half the maximum level of the "Higher Dynamic Toll" scenario, in lieu of the cordon charge. Tolls on existing MTA and Port Authority tolled crossings are credited for trips that enter Manhattan below 60th Street.

The main caveat here is that the 8MPH speed target and $30 peak round-trip toll is somewhat arbitrarily chosen, as is the "half the per-mile rate cap" for FHVs.

The BTA model as a whole expects a less-than-linear relationship between the speed target and the weighted average toll: It projects lowering the speed target by 20%, from 10MPH to 8MPH, will reduce the weighted average toll necessary to achieve the target by 76%, from $18 to $4. But that less-than-linear relationship is projected from price elasticity estimates measured from past behavioral responses to toll increases much smaller than modeled here. By contrast, the core speed-volume model is more reliable because it is built on traffic data we already have and re-sample regularly: We know how fast traffic currently travels at night and during rush hour, and how many vehicles are on the road at each time.

The core speed-volume model tells us that each mile driven on the Manhattan grid during the worst congestion of the day slows down all other vehicles by about $26 worth of time loss when roughly 184,000 hourly VMT slow the grid to 5.8MPH from the hours of 2PM-8PM in the pre-Pandemic baseline. Since that $26 is the slowdown cost imposed on all other vehicles, one can make the case that it is a reasonable price for driving into the CBD during this period — it is the dollar amount by which one is slowing down traffic.

How many people will respond to such a price by using alternative modes of transportation or avoiding peak period CBD trips altogether? Resulting in what vehicle volume and speed in equilibrium? The BTA says traffic volumes will fall enough to speed traffic up to 10MPH when a peak $26 cordon toll is imposed, hence this paper's estimated "optimal" target speed.

In sum: I am confident in the core speed-volume model, including the point estimate of a $26 per-mile cost of pre-Pandemic travel at 5.8MPH based on currently observable traffic. I am less confident in the full BTA model's translation of that $26 slowdown cost into an equivalently "optimal" equilibrium speed using behavioral elasticities estimated from previous, much smaller toll changes.

[i] 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—see Appendix to see 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.

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

[iii] By "optimal Pigouvian toll", I mean the toll that is exactly equal to the marginal time cost imposed on other drivers by the average trip within the CBD, similar to the pro forma speed-volume schedule in the appendix. It is "optimal" in the "partial equilibrium" sense that drivers will only take trips when they value that trip in excess of the average delay cost imposed on the grid.

[iv] This paper relies on the Hayekian logic of dynamic pricing to build in some robustness to error in our expectation of the responsiveness of traffic to price—much more so than a fixed toll schedule that policymakers would scramble *ex post* to change in the event of an economic depression or other unexpected negative exogenous shock to traffic volumes. But intellectual humility is warranted insofar as the headline speed target is still subject to some model uncertainty.

Specifically: The speed-volume curve tells us exactly how much to charge the next car to travel a mile for any given number of hourly vehicle-miles of travel in Manhattan. For any given volume of travel, we know how much the next vehicle mile will slow down all the other cars already on the road, and can then reliably convert those minutes of aggregate travel time loss into a dollar value of delay per mile without knowing anything about human behavior in response to tolls. We know, for example, that when hourly traffic volumes hit the PM peak of a little over ~170,000 VMT in Manhattan, corresponding to 6MPH on average, then the next mile trip causes about ~$26 in slowdown to all the other cars combined. That's why we set the ideal maximum toll cap at $26. Such a toll is "optimal" in the sense that trips valued in excess of the delay cost per mile should keep happening, while trips valued less than that should not.

But federal legislation allowing dynamic HOT lanes on federally funded Interstates requires the additional step of setting a concrete minimum policy speed target, not merely charging the abstract volume-conditional Pigouvian toll schedule. So we know that roughly $26 is the maximum correct price to charge per mile, according to the speed-volume curve and our knowledge of the average value of travel time, when traffic is at its average hourly worst in the pre-Pandemic baseline during the PM rush hour. But we must then translate the known $26 per-mile Pigouvian toll from the speed-volume curve into an estimated speed target achieved through a cordon toll. To do this the BTA model has to commit to an estimate of travel time and price elasticities and average travel distance per cordon entry by different vehicle and trip types—and the BTA happens to expect the human response to the $26 Pigouvian peak rush hour toll to coincide with roughly a 10MPH speed target. So in sum: We can be confident from the baseline speed-volume curve that $26 is the maximum necessary per-mile toll, but whether that optimum coincides with the 10MPH rush hour speed target in equilibrium relies upon the empirical accuracy of the BTA model. Less-responsive traffic would result in the $26 peak toll lasting longer throughout the day than currently projected. More responsive traffic could yield a higher optimal speed target than currently projected to result from the $26 optimal cap.

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

[vi] This paper's recommendations for the use of revenue in excess of the $1 billion annual target are similarly robust to surprise shocks: By using any excess revenue to restructure and retire existing debt and thereby indirectly relieve the operating budget, no particular capital project is at risk. Prepayment of outstanding debt, even if "lumpy" within the next few years of economic recovery, will provide operating budget relief in the near term and eventually stable borrowing capacity for next MTA capital plan.

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

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

[ix] 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.

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

[xi] 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, (and generally undesirable compared to better-tailored road pricing in other jurisdictions) 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 $26 cap *per cordon entry* is therefore on the low end of any plausible estimate of the typical daytime congestion externality.

[xii] Singapore, for example, achieves impressive speed targets with moderate tolls because it does not rely on Electronic Road Pricing tolls alone: It also prices urban parking and conducts auctions of “Certificates of Entitlement” for the right to own a car, where the price for a CoE can exceed the purchase price of the vehicle itself. While CoEs impose deadweight loss insofar as they ration cars unnecessarily, and priced parking & tolls are therefore more efficient, they are nonetheless illustrative of how alternative policies that raises the cost of car ownership can reduce the equilibrium toll.

[xiii] http://www.schallerconsult.com/pub/freeparking.htm [ed: the link is dead but the PDF can still be found on the Internet Archive’s 2008 snapshot. Not sure how to cite, so including the Streetsblog coverage]

See also : https://nyc.streetsblog.org/2006/06/16/the-46-million-parking-perk/

[xiv] “EU: Low Emission Zones (LEZ),” DieselNet.

[xv] See Appendix.

[xvi] RPA, “Congestion Pricing in NYC: Getting It Right,” September 2019.

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

[xviii] The optimal Pigouvian toll in the Balanced Transportation Analyzer’s speed-volume model of the Manhattan grid, considering only delays on the Manhattan grid and not delays on tristate regional roads en route to the CBD, is around $26 during rush hours. By ”optimal Pigouvian toll”, I mean the toll that is exactly equal to the marginal time cost imposed on other drivers by the average trip within the CBD. It is ”optimal“ in the “partial equilibrium” sense that drivers will only take trips when they value that trip in excess of the average delay cost imposed on the grid.

The minimum speed within the Manhattan grid when the toll cap is set to the Pigouvian rate of $26 is approximately 10MPH, leading to this paper’s illustrative recommendation that the speed target be set at 10MPH, as this is the model-expected speed that is coincident with the Pigouvian toll.

[xix] See LIRR fare chart: http://web.mta.info/lirr/about/TicketInfo/FareChart2019.pdf

[xx] LIRR collected roughly $790 million in fares on 91.1 million one-way trips in 2018, of which 38.9 million trips were discounted “non-commutation” trips. See: http://web.mta.info/mta/news/books/docs/LIRR-2018-Annual-Ridership-Report.pdf

For Metro-North’s $815 million in fares on 87.1 million one-way trips, see:

http://web.mta.info/mta/news/books/docs/MNR-2018-Annual-Ridership-Report.pdf

[xxi] 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.

[xxii] 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.

[xxiii] 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.

[xxiv] As calculated in Komanoff, “Balanced Transportation Analyzer,” the model used by RPA.

[xxv] https://streets.mn/2020/04/07/five-facets-of-recovery-from-covid-19/

[xxvi] 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.

[xxvii] 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).

[xxviii] 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.

[xxix] https://academiccommons.columbia.edu/download/fedora_content/download/ac:147390/content/Kirschling_Thesis__NYC_Rapid_Transit_1870-2010__05-13-12.pdf

[xxx] For one example among many, NYCT subway operations would earn an operating profit if work rules and systemwide productivity rose to that of the Chicago Transit Authority’s rail system.