

Paying Customers for Patience: An approach for pricing slotted deliveries for Quick-Commerce

Milon Bhattacharya

`milon.bhattacharya25-08@iimv.ac.in`

Indian Institute of Management Bangalore <https://orcid.org/0000-0002-7382-5797>

Research Article

Keywords:

Posted Date: October 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11220649/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Paying Customers for Patience: An approach for pricing slotted deliveries for Quick-Commerce

Milon Bhattacharya 

Department of Operations, Indian Institute of Management Visakhapatnam
`milon.bhattacharya25-08@iimv.ac.in`

September 2026

Working paper version 3.4

Abstract

Quick-commerce platforms advertise sub-10-minute delivery, yet they increasingly offer slotted options where orders arrive with a delay within a time window, delivered via a shared route. Under current platform practices, customers lack financial incentives to choose slotted delivery despite the substantial operational savings it offers to the platform. This study models optimal pricing for such slotted deliveries and determines how platforms should incentivize patient customers during dark-store fulfillment. Our analytical framework incorporates exact solo delivery tour geometry, courier capacity constraints, stochastic batching, endogenous delay feedback, and two-part courier piece rates clearing reservation wages. Calibrated on open-source datasets from Amazon, Grubhub, Instacart, and Olist, the model yields four main insights. First, under uniform customer time values, the optimal slot discount is equal to the marginal cost savings multiplied by a pass-through rate $\rho = 1/(2 + \varepsilon)$, where ε is the elasticity of the delay of the slot to the batch size; the calibrated pass-through is 0.40 – 0.55 (versus 1.0 under competitive fee setting), passing \$0.50 – \$ 0.90 per order to customers. Second, cost savings are saturated in a single courier capacity load (C), as the marginal tour time per stop drops from 5.8 to less than 2 minutes by the twelfth stop; the optimal discount remains flat for slot lengths of 15 minutes to 3 hours, making the shortest capacity-filling window optimal. Third, distance-based fee schedules increase slotting surplus by 28% by grading discounts with distance, but empirical fees remain strictly flat within metros (distance explains only 1.4% of fee variance in Olist). Fourth, grocery demand follows a broad daytime plateau (10:00–16:00) rather than a bimodal distribution, making same-hour slots viable across most operational hours (08:00–21:00). A day-long simulation confirms analytical delay predictions and reveals that instant fulfillment requires fleet utilization below 0.50–0.65.

1 Literature Review

Our research bridges four major streams of literature: continuum approximations in vehicle routing, time-slot management in attended home delivery, dynamic batching in micro-fulfillment operations, and platform pass-through economics. Below, we review these streams and highlight our exact theoretical and empirical distinctions.

1.1 Continuum Approximations and Routing Geometry

The modeling of spatial routing costs relies on the foundational continuum approximation results of [5] and [11], who showed that optimal Traveling Salesperson Problem (TSP) tour lengths over

m points in area A scale as $k\sqrt{mA}$. This paradigm has been extensively deployed in pickup-and-delivery logistics [17] and collaborative urban transportation [10].

However, standard formulations charge a one-stop delivery an unphysical routing leg $k\sqrt{A}$, causing severe distortion at small batch sizes ($n = 1, 2$) and falsely indicating that small slots destroy value. We resolve this by introducing a modified tour-time function $H_1(n)$ featuring a $(m - 1)^+$ term under the square root, ensuring exact calibration for solo trips ($n = 1$) while preserving asymptotic scaling for large batches.

1.2 Time-Slotting and Incentive Schemes in Home Delivery

Incentive-based slotting and time-window management were established in traditional Attended Home Delivery (AHD) to smooth demand and increase density [8, 2, 12, 23]. [6] analyzed revenue management and environmental trade-offs in online grocery, while more recent work by [3], [14], and [24] introduced dynamic incentive discounts to shift customer demand across wider windows.

Whereas AHD literature focuses on multi-hour or multi-day windows served from regional warehouses, quick-commerce (Q-commerce) operates from decentralized dark stores under sub-hour time constraints. We demonstrate that in Q-commerce, marginal cost savings saturate rapidly at a single courier capacity load C , causing the optimal slot discount $d^*(\Delta)$ to flatten beyond 15–30 minutes while delay disutility continues to grow.

1.3 Micro-Fulfillment, Dynamic Batching, and On-Demand Operations

Modern hyperlocal fulfillment relies on micro-fulfillment centers (dark stores) and dynamic order pooling [18, 13]. High-frequency dispatching models have explored the tension between immediate order fulfillment and dynamic order bundling [21, 19, 9]. Furthermore, [4] demonstrated that driver batching significantly alters market thickness and unit fulfillment costs in online food platforms.

We build upon these operational foundations by incorporating endogenous delay feedback: as more customers adopt a slot, tour sizes grow, inflating in-tour delay $\theta(n)$ and deterring further adoption. We prove that this delay feedback fundamentally alters optimal pricing, establishing that local profit maxima sit directly on capacity kinks $L^* \approx qC$.

1.4 Platform Economics, Spatial Pricing, and Pass-Through

The theoretical foundations of cost pass-through under monopoly and imperfect competition originate from [7] and [22], with extensions to two-sided platform structures by [16] and fulfillment network pricing by [1]. Recent empirical and theoretical inquiries by [15] and [20] examined how platform commissions, spatial distance schedules, and driver remuneration impact fee pass-through and pooling efficiency.

While standard monopoly models yield a 50% pass-through rate under linear demand, we show that delay feedback attenuates pass-through to $\rho^* = 1/(2 + \varepsilon)$, where ε is the elasticity of slot delay to batch size. Furthermore, while prior literature advocates per-kilometre fee schedules to capture 28% higher slotting surplus, our empirical analysis of Olist data reveals that real-world metro delivery fees remain strictly flat, identifying a major gap between theoretical pricing mechanisms and market practice.

Table 1: Taxonomy of Literature vs. Present Paper

Study / Stream	Operational Setting	Pricing / Slotting Mechanism	Key Limitation / Distinction from Our Work
Beardwood et al. (1959); Daganzo (1984)	General TSP routing & continuum approximations	Asymptotic tour length ($k\sqrt{mA}$)	Distortion at $n = 1, 2$ (charges unphysical stem leg $k\sqrt{A}$); fixed by our $(m-1)^+$ formulation.
Agatz et al. (2011, 2021); Dynamic AHD	Multi-hour/day Attended Home Delivery from regional hubs	Dynamic slot availability; fixed window incentives	Assumes long delivery horizons; misses capacity-filling saturation (C) in sub-hour Q-commerce.
Ulmer et al. (2021); Stroh et al. (2022)	Dynamic same-day and instant delivery dispatching	Dynamic order batching & fleet assignment	Purely operational (cost-minimization); lacks endogenous customer adoption and fee pass-through economics.
Li & Wang (2022); Tong et al. (2023)	Two-sided delivery platforms & spatial pricing	Commission rates, per-km fees, courier piece rates	Assumes fixed/exogenous service delays; ignores customer delay feedback loops (ε) on adoption.
Bulow & Pfleiderer (1983); Weyl & Fabinger (2013)	General monopoly / oligopoly price theory	Standard cost pass-through ($\rho = \partial p / \partial c$)	Lacks service-quality feedback; misses Q-commerce pass-through reduction to $\rho^* = 1/(2 + \varepsilon)$.
Our Paper	Hyperlocal Quick Commerce from dark stores	Slot discount (d), slot length (Δ), flat vs. per-km fees	Integrates exact solo-trip tour geometry, endogenous delay feedback, capacity kinks (C), and empirical order-clock calibration.

2 Problem Statement

A quick-commerce platform quotes a delivery fee at checkout. If the customer wants the order now, a courier is dispatched with that order alone and the platform pays for a round trip to her door. If she is willing to wait, the platform can hold the order, accumulate others arriving in the same window and neighbourhood, and send one courier on a shared tour. The saving is real: the round-trip stem is shared and routing is dense. But it is a saving to the platform. Three questions follow.

- Q1** *Pass-through.* If a customer agrees to a delay, how much of the saving can the platform hand back as a lower fee, and does the answer depend on whether the instant fee is flat or per kilometre?
- Q2** *Value of patience.* How does the passable benefit grow with the slot length Δ ? Is there a point beyond which more patience is worth nothing to either side?
- Q3** *Scope of slotting.* Given the intraday order distribution, at which hours is there enough flow to fill a slot, how long must it be, and what should the platform do where there is not?

Total demand is fixed—groceries are bought regardless—so fees do not create or destroy orders; they steer each order between the instant and slotted options. Pricing reduces to a cost-side trade: discount dollars given up against courier cost saved. The instant fee schedule is an anchor set outside the model; the design objects are the slot discount and the slot length.

The answers, calibrated on open data (Sections 12–13):

- **Q1.** The platform can pay a waiting customer about \$0.50–0.90 per order, which is 40–55% of the marginal saving it makes by clubbing her; a competitive fee market would pass on all of it. Per-kilometre fees would grade the discount by distance (a customer at 1.6 km gets ten times the discount of one at 0.4 km) and raise the slotting surplus by 28%, but real fee schedules are flat within a city.
- **Q2.** The benefit saturates at one courier-load. Marginal tour time per added stop falls from 5.8 minutes at the second stop to under 2 minutes by the twelfth; the discount is flat in Δ from 15 minutes to three hours while the customer’s realised delay keeps growing. The best slot is the shortest one that fills a courier, and the optimum sits on multiples of capacity.
- **Q3.** Almost every busy hour. Grocery demand is a plateau from 10:00 to 16:00, and same-hour slots are viable from 08:00 to 21:00. Deferring off-peak orders to a peak wave is worthwhile only when the trough falls below about 15% of the peak, which grocery demand does not do.

What is established and what is assumed is set out in Section 14.

3 Notation

Symbol	Meaning
$A, r, \bar{r}$	service-zone area; customer distance from store; mean distance
κ, m, η	address (hotspot) intensity; distinct addresses visited; fill $\eta = n/\kappa A$
v, s_d, k, C	courier speed; per-drop service time; routing constant; courier capacity (orders)
$\bar{w}$	courier reservation wage (\$/h)
$N_{\text{tot}}, \lambda(t), N(t, \Delta)$	daily volume; arrival rate; expected arrivals in $[t, t + \Delta]$
n, L, q	realised batch; mean batch; tours needed $q = \lceil n/C \rceil$
$H_1(n), H(n)$	courier time of one tour of n orders; of a dispatch of n orders, all tours
$c_1(r), c_S(L), c_{\text{mg}}(L)$	solo cost at distance r ; average and marginal slot cost at mean batch L
$\theta(n)$	in-tour excess delay of a typical slotted customer
x, x_S	excess delay beyond the point-to-point time; slot excess delay
$\beta, G, \beta_{hi}, \bar{V}$	time value (\$/h), its distribution, its upper support; satiation cap
$p_1(r), p_S(r), d$	instant fee, slot fee, slot discount $d = p_1 - p_S$
a	slot adoption share; $L = aN$
$S(r, L), \rho$	marginal saving from switching a customer at r ; pass-through rate d/S
ε	elasticity of slot delay to batch size, $L \partial x_S / \partial L / x_S$

4 Primitives

4.1 Fixed demand on an order clock

Assumption 4.1 (Inelastic unit demand). N_{tot} orders are placed per day, each by a distinct customer at an exogenous time and location. Fees affect only which delivery option an order takes.

Assumption 4.2 (Order clock). Orders arrive as a nonhomogeneous Poisson process with rate $\lambda(t)$; the flow a slot opened at t can collect is

$$N(t, \Delta) = \int_t^{t+\Delta} \lambda(u) du. \quad (1)$$

No parametric shape is imposed on λ . A two-peak mixture is a natural description of meal delivery; grocery demand turns out to be a daytime plateau (Section 12). The integral in (1), not $\lambda(t)\Delta$, is what matters when λ moves within a window.

4.2 Customers: distance and satiable impatience

Customer locations are uniform on the zone, so r has density $2r/R^2$ on a disc of radius R and mean $\bar{r} = 2R/3$. A customer regards the point-to-point travel time as the natural lag and attaches disutility only to the excess x :

$$u(x; \beta) = \min(\beta x^+, \bar{V}), \quad \beta \sim G \text{ on } [0, \beta_{hi}], \quad (2)$$

with a common cap $\bar{V}$. Facing options (p_j, x_j) she minimises $p_j + u(x_j)$.

4.3 Spatial clustering and tour time

Orders cluster at addresses (buildings) that are themselves scattered over the zone; a dispatch of n orders on a zone with κA addresses visits $m = \kappa A(1 - e^{-\eta})$ distinct addresses in expectation, $\eta = n/\kappa A$.

Assumption 4.3 (Tour time). A single tour of n orders visiting m addresses on a zone of area A_z takes

$$H_1(n; A_z) = \frac{2\bar{r}}{v} + \frac{k}{v} \sqrt{(m-1)^+ A_z} + n s_d, \quad (3)$$

so that $H_1(1) = 2\bar{r}/v + s_d$ is exactly the expected solo round trip. Where an empirical tour-time curve is available, it replaces (3).

The $(m-1)$ under the root is a small but consequential departure from the textbook $k\sqrt{mA}$: the textbook form charges a one-stop “tour” a routing leg of $k\sqrt{A}$ that does not exist, which makes a two-order batch cost more than two solo trips and small slots appear to destroy value. In distance the corrected form fits the data well (Section 12); in travel time neither parametric form is exact for two- to six-stop tours, which is why the calibrated results below use the empirical curve directly.

Assumption 4.4 (Capacity and batch noise). A courier carries at most C orders. A dispatch of n orders is split into $q = \lceil n/C \rceil$ tours over equal sectors of area A/q , so

$$H(n) = q H_1(n/q; A/q). \quad (4)$$

A slot collects a random number of orders with mean L ; expected courier time is $\mathbb{E}[H(n)]$.

We compute expectations under a Poisson batch law. Simulation shows realised batches are overdispersed (variance/mean ≈ 3) because the arrival rate moves within a window, so the Poisson expectation slightly understates cost; a negative-binomial law would be more accurate.

4.4 Couriers

Assumption 4.5 (Courier pay). Couriers hold a reservation wage $\bar{w}$ and accept a dispatch iff pay per hour of tour time is at least $\bar{w}$. The platform pays a two-part piece rate— $\bar{w}s_d$ per drop plus $\bar{w}/v$ per routed kilometre—so pay on a dispatch equals $\bar{w}H(n)$.

Under Assumption 4.5 delivery cost is $\bar{w}$ times courier time, and batching savings accrue to the platform mechanically. A flat per-drop fee would instead make the market-clearing rate equal to the average cost curve $\bar{w}H(n)/n$, hand density gains to couriers at any posted rate, and require repricing as batches thicken; a distance component removes all of that, which is why platforms pay per drop plus per kilometre. Assumption 4.5 also omits idle time between dispatches; Section 13 quantifies when that omission bites.

5 Costs, delay, fees, and pass-through

Definition 5.1 (Cost objects). The cost of an instant delivery at distance r , its mean, and the slot cost objects at mean batch L are

$$c_1(r) = \bar{w} \left(s_d + \frac{2r}{v} \right), \quad \bar{c}_1 = \bar{w}H_1(1), \quad K(L) = \bar{w} \mathbb{E}[H(n)], \quad c_S(L) = \frac{K(L)}{L}, \quad c_{\text{mg}}(L) = K'(L) = \bar{w} \mathbb{E}[H(n+1) - H(n)] \quad (5)$$

the last equality being the Poisson identity $\frac{d}{dL} \mathbb{E}f(n) = \mathbb{E}[f(n+1) - f(n)]$.

$c_{\text{mg}} < c_S$ wherever c_S is decreasing: an added order shares the stem and thickens routing for everyone in the batch. The gap $c_S - c_{\text{mg}} = -Lc'_S(L)$ is the density externality.

Lemma 5.2 (In-tour excess delay). *For a customer in a uniformly random position on a tour of n orders, the expected excess delay attributable to the tour is $\theta(n) = \frac{1}{2}[H_1(n) - H_1(1)]$. With uniform arrivals in the window, the slot excess delay at mean batch L split into q tours is*

$$x_S(L, \Delta) = \frac{\Delta}{2} + \theta(L/q; A/q). \quad (6)$$

Proof. Inter-stop travel and service before a random stop average half of $H_1(n) - 2\bar{r}/v - s_d$, and $2\bar{r}/v + s_d = H_1(1)$. $\square$

Simulation confirms (6) to within a minute when L is the realised mean batch (Section 13). Because θ grows with L and L grows with adoption, a customer who accepts a “30-minute slot” experiences about 27 minutes of excess delay at a 13-order batch, and more if more people adopt.

Assumption 5.3 (Fee schedules). The instant fee is $p_I(r) = F_I + \gamma_I r$ and the slot fee $p_S(r) = F_S + \gamma_S r$. Two regimes: *flat* ($\gamma_I = \gamma_S = 0$, $F_I \geq 0$ possibly zero) and *per-kilometre* (p_I tracks c_I up to a markup, $\gamma_I = 2\bar{w}/v$). The instant schedule is a market datum; the platform designs $d(r) = p_I(r) - p_S(r) \geq 0$.

Because volume is fixed, the level of the instant schedule never enters the discount problem: revenue at p_I is a fixed base and the discount is a pure transfer against cost savings. A zero instant fee is admissible; the discount is then a basket credit.

Definition 5.4 (Saving and pass-through). When a customer at r moves from instant to a slot with mean batch L , the marginal saving is $S(r, L) = c_I(r) - c_{\text{mg}}(L)$ and the pass-through rate of a discount d is $\rho = d/S$.

6 Demand allocation

Proposition 6.1 (Adoption fixed point). *For a slot of length Δ at hour t with distance-independent discount $d < \bar{V}$ and $N = N(t, \Delta)$, a type- β customer adopts iff $\min(\beta x_S, \bar{V}) \leq d$, and adoption solves*

$$a = G\left(\frac{d}{x_S(aN, \Delta)}\right), \quad (7)$$

which has a unique solution because the right-hand side is nonincreasing in a . At $d \geq \bar{V}$ all types adopt.

Fixed demand removes the self-reference through market size that free-pricing slot models have, but not the one through service quality: more adopters lengthen the tour, lengthen the delay, and deter adopters. The loop is monotone and solved by bisection.

7 Optimal discount and pass-through (Q1)

In the flat regime the platform’s gain per window from offering the slot, relative to instant-only, is

$$\Pi(d) = aN(\bar{c}_I - d) - K(aN), \quad a = a(d) \text{ from (7)}. \quad (8)$$

Theorem 7.1 (Optimal discount). *At an interior optimum on a smooth segment of K , with $L^* = a(d^*)N$,*

$$d^* = S(\bar{r}, L^*) - \frac{a}{a_d}, \quad a_d = \frac{g/x_S}{1 + \varepsilon g d/(a x_S)}, \quad \varepsilon = \frac{L \partial x_S / \partial L}{x_S}, \quad (9)$$

with $g = G'$ at d/x_S . Under $\beta \sim U[0, \beta_{hi}]$,

$$\boxed{d^* = \frac{S(\bar{r}, L^*)}{2 + \varepsilon}, \quad \rho^* = \frac{1}{2 + \varepsilon}.} \quad (10)$$

Proof. $\Pi'(d) = a_d N [\bar{c}_1 - c_{\text{mg}}(L) - d] - aN$. Differentiating (7) gives $a_d = g[1/x_S - d x'_S N a_d / x_S^2]$ with $x'_S N = x'_S L/a$; solve for a_d . Under the uniform law $g/x_S = a/d$, so $a/a_d = d(1 + \varepsilon)$. $\square$

Without delay feedback ($\varepsilon = 0$) this is the half-of-saving rule of a monopolist facing linear demand. With feedback, each adopter lengthens the realised delay, recruiting is harder, and the platform discounts less; calibrated, $\varepsilon \approx 0.4$ and $\rho^* \approx 0.41$. If fees were set competitively, the slot fee would fall to c_{mg} and $\rho = 1$. The pass-through rate measures market power in the delivery-option market.

Proposition 7.2 (Capacity kinks). *$K(L)$ has downward kinks in its slope at $L \in \{C, 2C, \dots\}$: the marginal cost of an order that completes a courier-load is lower than that of the order which starts the next load, and the in-tour delay θ drops when a batch splits into one more tour. Consequently Π has local maxima at batch sizes $L^* \approx qC$, and the global optimum is generically at such a point rather than at a stationary point of (9). At a kink the pass-through rate exceeds $1/(2 + \varepsilon)$ because the relevant marginal cost is the low one just before the kink.*

In the calibration, L^* is 12, 24 or 36 in most hours (Table 3), with ρ of 0.50–0.58 at kinks against 0.37–0.41 on smooth segments. The design rule is therefore operational rather than marginal: set the discount that fills an integer number of couriers.

Proposition 7.3 (Full-adoption corner). *The interior optimum is compared with $d \downarrow \bar{V}$, at which all customers adopt and instant service is retired. The corner is preferred iff*

$$\bar{V} \leq a^* d^* + (1 - a^*) [\bar{c}_1 - c_S(N)] + a^* [c_S(L^*) - c_S(N)], \quad a^* = a(d^*). \quad (11)$$

A sufficient condition against the corner at every hour is $\bar{V} \geq \bar{c}_1 - \lim_{L \rightarrow \infty} c_{\text{mg}}(L)$. Because a jumps at $\bar{V}$, Π is discontinuous there; the supremum over $[0, \bar{V})$ is compared with the value at any d marginally above $\bar{V}$.

With the calibrated cap of \$4 against a maximal saving of about \$1.8, the corner is never chosen, and $\bar{V}$ is irrelevant to every interior optimum.

8 Flat versus per-kilometre fees

Solo cost is linear in distance while $c_{\text{mg}}(L)$ does not depend on where the marginal customer lives, so $S(r, L)$ rises with r at slope $2\bar{w}/v$: the customers costliest to serve instantly are those for whom patience is worth most.

Proposition 8.1 (Distance-graded discounts). *In the per-kilometre regime the optimal discount is $d^*(r) = \rho^* S(r, L^*)$, increasing in distance, with adoption $a(r) = G(d^*(r)/x_S)$ increasing in r . In the flat regime $d^* = \rho^* S(\bar{r}, L^*)$. Under a uniform G at fixed L , the slotting surplus is proportional to $\mathbb{E}_r[S(r, L)^2]$ in the per-kilometre regime and to $S(\bar{r}, L)^2$ in the flat regime; the per-kilometre regime gains $\rho^*(1 - \rho^*)\text{Var}_r[S]/(x_S \beta_{hi})$ per reachable order.*

Proof. The first-order condition of Theorem 7.1 holds pointwise in r with a common ε ; under the uniform law surplus at r is $d(S - d)/(x_S \beta_{hi})$; substitute $d = \rho S$ and take expectations. $\square$

A flat discount overpays near customers and underpays far ones. Calibrated (Table 4), per-kilometre pricing discounts a 1.6-km customer ten times more than a 0.4-km customer, adopts 27% of the former against 3% of the latter, and raises the hourly gain by 28% at the same batch size. Section 12 shows that observed fee schedules are flat within a metro, so this is the value of a change the market has not made.

9 The value of patience (Q2)

Hold the hour fixed and let Δ vary. Reachable flow N grows, the batch $L = aN$ grows, and the delay $x_S = \Delta/2 + \theta$ grows. The passable benefit is $d^*(\Delta) = \rho^* S(r, L(\Delta))$.

Proposition 9.1 (Front-loaded and capped savings). *(i) $S(r, L)$ is increasing and concave in L on $[0, C]$; (ii) for $L > C$ the dispatch splits and c_{mg} is approximately flat, oscillating with L/C but not trending down; (iii) hence $d^*(\Delta)$ rises while $L(\Delta) < C$ and is essentially flat thereafter, while $\min(\beta x_S, \bar{V})$ keeps rising until it caps. The platform’s gain per hour is therefore decreasing in Δ once one courier-load can be filled, and the optimal slot is*

$$\Delta^* \approx \frac{C}{a \lambda(t)}. \quad (12)$$

Calibrated: marginal tour time per added stop is 5.8 minutes at the second stop, 3.3 at the fourth, 1.9 at the twelfth, 1.4 at the twenty-fourth; c_{mg} is \$0.89 at $L = 12$ and \$0.77 at $L = 24$; and d^* lies in \$0.50–0.90 for every Δ from 15 minutes to 3 hours (Table 5). At 150–250 orders per hour, (12) gives 15–30 minutes. Two qualifications: with noisy batches a short slot often falls short of C , which (8) prices; and (12) presumes a courier is free when the window closes, which requires fleet slack (Section 13).

10 Scope of slotting from the order clock (Q3)

Proposition 10.1 (Viability screen). *A same-hour slot of length Δ at hour t is worth offering iff*

$$\max_{d \in (0, \bar{V})} \left\{ a(d)N(t, \Delta) [\bar{c}_1 - d] - K(a(d)N(t, \Delta)) \right\} > 0. \quad (13)$$

Since $K(L)/L \rightarrow \bar{c}_1$ as $L \rightarrow 0$, the screen fails when reachable flow times attainable adoption cannot produce a mean batch materially above 2–3.

Proposition 10.2 (Deferral to a peak wave). *If hour t fails (13) and a later peak wave with batch L_p exists at t_p , a “delivered with the t_p wave” option inserts the order at marginal cost $c_{mg}(L_p)$ —the same marginal saving a peak adopter generates—at delay $x_w = t_p - t$ independent of trough flow. Its optimal discount is $\rho^*(x_w)S(r, L_p)$ and its adoption $G(d_w/x_w)$. Deferral dominates a same-hour slot when the window needed to pass (13) exceeds $t_p - t$.*

Under the two real order clocks the screen passes at essentially every operating hour (Tables 3–5), and deferral becomes relevant only when the trough falls below roughly 15% of the peak. For grocery demand, which sits on a daytime plateau, the platform should offer a short same-hour slot all day; the deferral product is a meal-delivery or deep-trough special case.

11 Menu length

Should several slot lengths be offered? Demand values granularity only below $\bar{V}/\beta$ type by type, and Proposition 9.1 shows supply gains almost nothing from patience beyond one courier-load—under the calibration, under \$0.12 per order from a second load. A longer second slot could screen patient types at a deeper discount (a Mussa–Rosen problem in delay), but its gain is bounded by that saving, so a single short slot is within cents of optimal. We state this as an argument, not a theorem.

12 Data and calibration

Four open datasets calibrate the model. None is grocery quick commerce, so they validate mechanisms and supply parameters rather than levels.

Amazon Last Mile Routing Research Challenge (2021). A 10% route sample (611 routes, 17 stations in Los Angeles, Boston, Chicago, Seattle, Austin) with stop coordinates, actual visiting sequences, packages per stop and planned service times; for 153 of these routes the actual stop-to-stop travel-time matrices. Each route serves a compact zone (convex-hull area 7.8 km², equivalent radius 1.4 km) from a station on average 18 km away. Three checks on the full routes: the stem equals $2\bar{r}$ (median ratio 0.99); stops receive one package 66% of the time, two 21%, three 7% (a zero-truncated Poisson fill of $\eta \approx 1.06$ at route scale, i.e. about 229 addresses per zone); planned service is 1.13 minutes per package.

The tour-time curve is built by *subsampling*: for 141 zones, random subsets of $m \in \{1, 2, \dots, 96\}$ stops are drawn, a store is placed at the stop nearest the zone centroid, and a nearest-neighbour + 2-opt tour is computed from the real travel-time matrix (Figure 1, left). The heuristic is 6% faster than the drivers’ actual sequences on full routes, so the curve is slightly optimistic in level but not in shape. Effective speed falls from 15 km/h on a 12-stop tour to 10 km/h on a 96-stop tour because short hops are slow, so travel time is not proportional to straight-line distance. In straight-line kilometres the $2\bar{r} + k\sqrt{(m-1)A}$ form fits with $k = 0.667$ and RMSE 0.33 km against 0.60 km for $s + k\sqrt{mA}$, which overstates the solo trip by 65%. In travel time the corrected form over-predicts two- to six-stop tours by 2–3 minutes (RMSE 1.8 min) and the unconstrained form fits better (0.4 min) only with a negative stem. The calibration therefore uses the empirical time curve itself. Fitted k_t is 3.7–4.4 min/km across cities (bootstrap 95% CI 3.74–3.98 pooled); out-of-sample RMSE on held-out zones is 4.4 minutes.

Grubhub Meal Delivery Routing instance. One day of 880 orders with placement times and coordinates (the public instances sharing an order code differ only in courier supply). Spatially, 33% of orders have another order within 100 m against 4% under complete spatial randomness at the same density; mean nearest-neighbour distance is 245 m against a CSR value of 435 m; 1-km quadrat counts have variance/mean 6.2. Demand clusters as the model requires. Temporally the profile is bimodal (lunch, dinner) with the afternoon trough at 45% of the dinner peak.

Instacart (2017). 3.4M grocery orders with hour of placement. The profile is a plateau: 57% of orders between 10:00 and 16:00 at a nearly constant 8% of daily volume per hour, a slow evening decline, no dinner peak (Figure 1, right). This is the clock used for the grocery results.

Olist (2016–18). 94,391 delivered single-seller Brazilian e-commerce orders with delivery fee, seller and customer geolocation, and realised and promised delivery times. Used only for the shape of observed fee schedules (Section 13).

Calibrated parameters. Zone area 7.77 km^2 , $\bar{r} = 1.09 \text{ km}$, solo round trip 9.6 min travel, 229 addresses, $s_d = 2 \text{ min}$ (a grocery handoff; Amazon’s parcel drop is 1.1), $\bar{w} = \$12/\text{h}$, $C = 12 \text{ orders}$, $N_{\text{tot}} = 2000 \text{ orders/day}$ on the Instacart clock (Grubhub clock reported alongside). Demand side, not identified by any dataset: $\beta \sim U[0, 8] \text{ \$/h}$ and $\bar{V} = \$4$ as the base case, with sensitivity in Section 13. Then $\bar{c}_1 = \$2.31$ and $c_1(r) = 0.40 + 1.76r$.

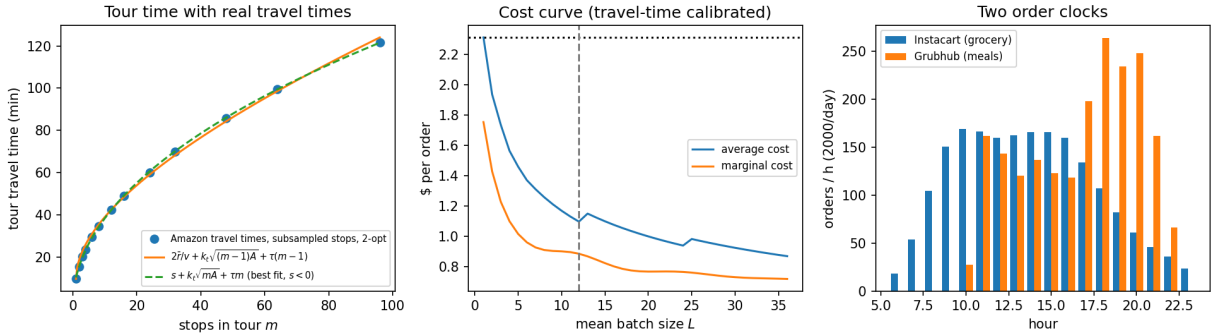


Figure 1: Left: tour travel time by number of stops from subsampled Amazon zones with real travel-time matrices, against the two parametric forms. Centre: calibrated average and marginal cost per order (dotted: solo cost; dashed: capacity). Right: the Instacart (grocery) and Grubhub (meal) order clocks, both scaled to 2,000 orders per day.

13 Results

13.1 The cost curve

Table 2: Calibrated cost curve (Poisson batches, $C = 12$). θ is the in-tour excess delay.

L	$\bar{w}H(L)/L$ (\$)	average $c_S(L)$ (\$)	marginal $c_{\text{mg}}(L)$ (\$)	θ (min)
1	2.31	2.01	1.75	0
2	1.94	1.79	1.43	4
3	1.74	1.64	1.23	7
5	1.46	1.43	1.02	12
8	1.26	1.25	0.91	19
12	1.10	1.13	0.89	27
16	1.08	1.06	0.80	16 (two tours)
24	0.94	0.97	0.77	22

One courier-load realises about 90% of the saving reachable at two. The marginal saving from clubbing an order, $S = \bar{c}_I - c_{mg}$, is \$1.42 at $L = 12$ and \$1.54 at $L = 24$.

13.2 Pass-through and the optimum by hour (Q1)

Table 3: Optimal 30-minute slot by hour, flat fee, $\beta \sim U[0, 8]$. $\rho = d^*/S$; x_S is realised excess delay; gain is profit per hour over instant-only.

clock	hour	λ	d^* (\$)	S (\$)	ρ	adoption	L^*	x_S (min)	gain (\$/h)
Instacart	08:00	104	0.83	1.43	0.58	23.0%	12.0	27	8.5
	10:00	169	0.59	1.45	0.41	15.7%	13.2	28	16.3
	13:00	163	0.58	1.44	0.41	15.8%	12.8	28	15.7
	16:00	159	0.58	1.44	0.41	15.9%	12.6	28	15.5
	19:00	82	0.55	1.34	0.41	14.1%	5.8	29	4.6
	21:00	46	0.48	1.16	0.42	15.4%	3.5	24	1.8
Grubhub	11:00	161	0.58	1.44	0.41	15.8%	12.8	28	15.6
	13:00	120	0.72	1.43	0.50	19.9%	12.0	27	11.2
	16:00	118	0.73	1.43	0.51	20.3%	12.0	27	10.9
	18:00	264	0.78	1.55	0.50	19.4%	25.6	30	29.9
	20:00	248	0.83	1.56	0.53	20.7%	25.6	30	27.3

The pass-through rate is 0.41 on smooth segments ($\varepsilon \approx 0.42$, $1/(2 + \varepsilon) = 0.41$) and 0.50–0.58 where the optimum sits on a capacity kink (Proposition 7.2); L^* is 12, 24 or 36 in eight of eleven cells. A patient customer can be paid \$0.48–0.83 per order. The corner $d = \bar{V}$ is never chosen.

13.3 Flat versus per-kilometre fees

Table 4: Flat versus per-kilometre fees, Instacart 13:00, 30-minute slot. Both regimes settle at $L \approx 13$ and $\rho^* \approx 0.42$.

	$r = 0.4$ km	$r = 0.8$	$r = 1.2$	$r = 1.6$
$c_I(r)$ (\$)	1.10	1.81	2.51	3.22
$S(r, L^*)$ (\$)	0.24	0.94	1.65	2.35
per-km $d^*(r)$ (\$)	0.11	0.40	0.70	1.00
per-km adoption	2.9%	10.8%	18.7%	26.6%
flat d^* (\$)	0.58	0.58	0.58	0.58
flat adoption	15.8%	15.8%	15.8%	15.8%
gain (\$/h): per-km 20.2, flat 15.8 (+28%)				

Against this, Olist shows how fees are actually set. Fees are strongly concave in distance (log-log elasticity 0.21; linearly R\$13.84 + R\$1.06 per 100 km per item, $R^2 = 0.16$), and item price explains more of the fee than distance. Within the São Paulo metropolitan range (19,096 orders under 150 km) the fee is flat: R\$10.79 + R\$2.30 per 100 km with $R^2 = 0.014$, a median of R\$9.42 and an interquartile range of R\$7.91–12.88 that distance does not explain (Figure 2). Paying more does not buy speed—median delivery time rises with the fee quintile because both track distance. The flat regime is thus the empirical baseline at intra-city scale, and the near-customer cross-subsidy of Section 8 is present in the data.

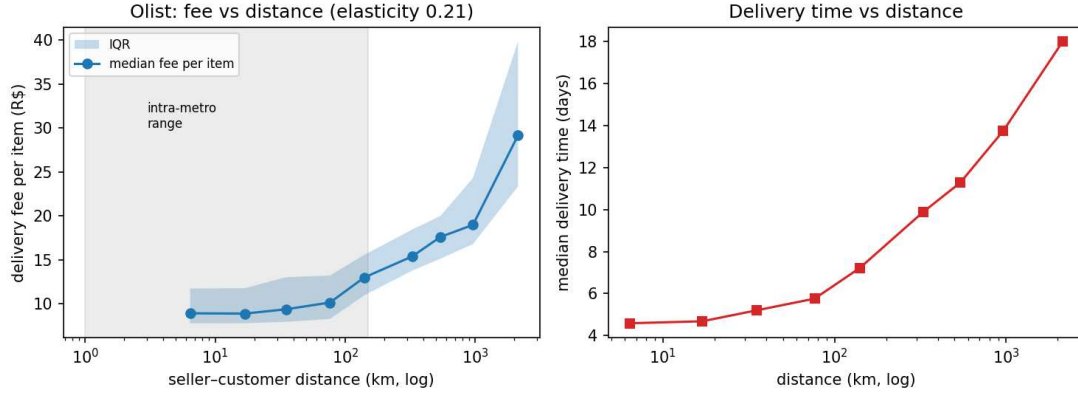


Figure 2: Olist: median delivery fee per item (with interquartile band) and median delivery time against seller–customer distance, log scale. Shaded: the intra-metro range, where the fee is flat.

13.4 Slot length (Q2) and hours (Q3)

Table 5: Optimal discount and gain by slot length and hour, Instacart clock, flat fee. N is reachable flow.

hour	Δ (min)	N	d^* (\$)	adoption	L^*	x_S (min)	gain (\$/h)
10:00	15	42	0.84	30.9%	13.0	20	18.7
	30	84	0.59	15.7%	13.2	28	16.3
	60	169	0.84	14.2%	24.0	44	12.2
	120	335	0.71	7.2%	24.0	74	7.7
13:00	15	41	0.52	16.3%	6.6	24	12.8
	30	81	0.58	15.8%	12.8	28	15.7
	60	163	0.87	14.8%	24.0	44	11.5
	120	328	0.72	7.3%	24.0	74	7.5
19:00	15	21	0.46	19.7%	4.1	18	5.5
	30	41	0.55	14.1%	5.8	29	4.6
	60	82	0.82	14.6%	12.0	42	4.4
	120	143	0.80	8.4%	12.0	72	2.3
21:00	15	11	0.38	21.6%	2.5	13	2.2
	30	23	0.48	15.4%	3.5	24	1.8
	60	46	0.55	10.0%	4.6	41	1.4
	120	82	0.60	6.2%	5.1	73	0.7

Three readings. The discount is flat in Δ once a courier can be filled (\$0.52–0.87 from 15 minutes to two hours at plateau hours), while realised delay climbs from 20 to 74 minutes. Hourly gain falls with Δ wherever a 15–30 minute slot already fills a courier (10:00, 13:00), so the short slot is the right slot. The screen (13) passes at every hour shown; the 21:00 shoulder, with batches of 2–5, is the only period where slotting is worth under \$2/h. Deferral to a peak wave never dominates on either real clock.

13.5 Sensitivity to the unobserved demand side

Holding Instacart 13:00 and $\Delta = 30$ fixed: $\beta \sim U[0, 4]$ gives $\rho = 0.57$, adoption 45%, gain \$38/h; $U[0, 16]$ gives $\rho = 0.41$, adoption 7%, gain \$4.5/h; a lognormal with median \$4/h gives $\rho = 0.52$, adoption 8.5%, gain \$4.2/h; a 10% mass of fully patient customers lowers ρ to 0.28 (the platform need not pay them) and raises the gain to \$26/h. $\bar{V} \in \{2, 4, 8\}$ changes nothing. Robust: $S \in [\$1.34, \$1.59]$, $d^* \in [\$0.43, \$0.91]$, and the optimum at or just above one courier-load. Not robust: adoption (7–45%) and the level of hourly gain (\$4–38).

13.6 Day-long simulation

A discrete-event simulation—nonhomogeneous Poisson arrivals on each clock, a 30-minute slot grid at $d = \$0.58$, capacity-12 loads, empirical tour times, a fleet of K couriers shared between instant and slot work, five simulated days—checks the stationary model against dynamics.

1. *Delay.* With an ample fleet, slot customers’ mean excess delay is 33–34 minutes against a formula value of 34 at the realised mean batch of 8 (90th percentile 42–47). Quoting (6) at the planned batch of 12.8 understates realised delay by 6 minutes because thin hours pull the day’s mean batch down.
2. *Batch noise.* Realised batch counts have variance/mean 2.6–3.9, not 1; the Poisson expectation in (5) slightly understates cost.
3. *Instant needs slack.* Instant orders wait for a courier with a 90th-percentile delay under 10 minutes only when utilisation is below about 0.5–0.65. At utilisation 0.9 (Grubhub clock, $K = 30$) instant orders wait 31 minutes on average and slot customers 68. A fleet sized for instant service at 50% utilisation carries an effective wage near $2\bar{w}$ per productive hour unless couriers are paid strictly per dispatch; Assumption 4.5 omits this.

14 Discussion: what is established and what is assumed

Q1. The platform can pass the benefit on, but as the monopolist over the customer’s option choice it optimally passes 40–55% of the marginal saving—\$0.50–0.90 per order—and keeps the rest; a competitive fee market would pass on all of it. The saving itself ($S \approx \$1.4$ –\$1.6 against a solo cost of \$2.31) is a cost-side fact supported by the Amazon travel-time data and robust to the demand side. Per-kilometre fees would grade the discount by distance and raise the surplus by 28%; Olist shows fees are flat within a metro, so this is a proposal, not a description.

Q2. The passable benefit saturates at one courier-load. This rests on the empirical tour curve (marginal time per stop $5.8 \rightarrow 1.9$ minutes by the twelfth stop) and on carrying capacity, and is robust to everything on the demand side. Longer slots lower adoption without raising the discount, so the optimal slot is the shortest that fills a courier, and the optimum lands on multiples of C .

Q3. Slotting is viable at almost every busy hour for grocery, whose demand is a daytime plateau (Instacart). The trough-and-deferral logic is a meal-delivery special case. This conclusion depends on the order clock, which is now data, and on attainable adoption, which is not.

Established. The tour-time curve and its front-loading; the exactness of the solo-trip stem; spatial clustering of demand; the grocery order clock; the delay formula (6) at realised batch sizes; the flatness of observed fee schedules; and the dominance of the capacity kink in deciding the optimum.

Assumed. The time-value distribution G and the cap $\bar{V}$. Across plausible alternatives, S and d^* barely move, but adoption ranges from 7% to 45% and hourly gain from \$4 to \$38. A discrete-choice survey with fee–delay pairs is the missing measurement; nothing in open data substitutes for it.

15 Limitations

(i) Idle time is not charged in the model; simulation shows instant service needs fleet utilisation below 0.5–0.65, which roughly doubles the effective wage for a held fleet. (ii) Batch sizes are treated as Poisson; realised batches are overdispersed by a factor of about 3. (iii) The tour curve is calibrated on parcel zones served from a distant station, with the store relocated to the zone centroid; this mimics a dark store but is a modelling choice. (iv) The 2-opt heuristic runs 6% faster than real drivers, so tour times are slightly optimistic in level. (v) The satiation cap is common and order timing is exogenous; a discount that shifted *when* people order would reopen a demand-shaping instrument. (vi) The single-slot menu is argued, not proven. (vii) Olist informs how fee schedules are set, not instant delivery; Grubhub and Instacart inform the clock and clustering, not grocery tour costs.

Summary of main formulas

$$\begin{aligned}
\text{Tour time (one tour): } & H_1(n; A_z) = \frac{2\bar{r}}{v} + \frac{k}{v} \sqrt{(m-1)^+ A_z} + n s_d, & H_1(1) &= \frac{2\bar{r}}{v} + s_d \\
\text{Capacity: } & H(n) = q H_1(n/q; A/q), & q &= \lceil n/C \rceil \\
\text{Costs: } & c_I(r) = \bar{w}(s_d + \frac{2r}{v}), & c_{\text{mg}}(L) &= \bar{w} \mathbb{E}[H(n+1) - H(n)] \\
\text{Slot delay: } & x_S = \frac{\Delta}{2} + \frac{1}{2} [H_1(L/q) - H_1(1)] \\
\text{Adoption: } & a = G(d/x_S(aN, \Delta)), & N(t, \Delta) &= \int_t^{t+\Delta} \lambda \\
\text{Saving, discount, pass-through: } & S(r, L) = c_I(r) - c_{\text{mg}}(L), & d^*(r) &= \rho^* S(r, L^*), & \rho^* &= \frac{1}{2+\varepsilon} \ (\beta \sim U, \text{ smooth seg}) \\
\text{Capacity rule: } & L^* \approx qC, & \Delta^* &\approx C/(a\lambda(t)) \\
\text{Viability: } & \max_d \{aN(\bar{c}_I - d) - K(aN)\} > 0; & \text{defer to a wave only if this fails before } t_p
\end{aligned}$$

References

- [1] Jason Acimovic and Stephen C. Graves. Making better fulfillment decisions on the fly in an online retail network. *Manufacturing & Service Operations Management*, 17(1):34–51, 2015.
- [2] Niels Agatz, Ann Campbell, Moritz Fleischmann, and Martin Savelsbergh. Time slot management in attended home delivery. *Transportation Science*, 45(3):435–449, 2011.
- [3] Niels Agatz, Ann Melissa Campbell, and Martin Savelsbergh. Incentive schemes and delivery slotting for same-day home delivery. *Operations Research*, 69(3):845–862, 2021.

- [4] Ali Aouad, Omer Saritac, and Daniela Saban. Market thickness in online food delivery platforms: The impact of food processing times and driver batching. *Manufacturing & Service Operations Management*, 26(2):512–531, 2024.
- [5] Jillian Beardwood, John H. Halton, and John M. Hammersley. The shortest path through many points. *Mathematical Proceedings of the Cambridge Philosophical Society*, 55(4):299–327, 1959.
- [6] Elena Belavina, Karan Girotra, and Ashish Kabra. Online grocery retail: Revenue management and environmental implications. *Management Science*, 63(5):1545–1563, 2017.
- [7] Jeremy I. Bulow and Paul Pfleiderer. A note on the effect of cost changes on prices. *Journal of Political Economy*, 91(1):182–185, 1983.
- [8] Ann Melissa Campbell and Martin W. P. Savelsbergh. Efficient incentives for attended home deliveries. *Transportation Science*, 38(3):371–388, 2004.
- [9] Alejandro A. Castillo, Nicholas Knoepfle, and Serguei Netessine. Surge pricing, order batching, and delivery delays in on-demand logistics platforms. *Management Science*, 68(8):5890–5910, 2022.
- [10] Catherine Cleophas, Caitlin Cottrill, Jan Fabian Ehmke, and Kevin Tierney. Collaborative urban transportation: Models and insights. *European Journal of Operational Research*, 273(3):819–833, 2019.
- [11] Carlos F. Daganzo. The length of tours in vehicle routing problems. *Transportation Research Part B: Methodological*, 18(4-5):315–316, 1984.
- [12] Jan Fabian Ehmke and Ann Melissa Campbell. Customer acceptance mechanisms for attended home deliveries in urban areas. *Transportation Science*, 48(4):560–579, 2014.
- [13] Omar El Housni and Huseyin Topaloglu. Joint assortment and inventory management for quick-commerce with front warehouses. *Management Science*, 69(11):6701–6721, 2023.
- [14] Charlotte Köhler, Maximilian Haferkamp, and Jan Fabian Ehmke. Dynamic customer admission and slot incentive pricing in on-demand urban delivery. *Transportation Science*, 56(4):1012–1032, 2022.
- [15] Chun Li and Yanzhi Wang. Service pricing, commission rates, and fee pass-through in on-demand delivery platforms. *Manufacturing & Service Operations Management*, 24(5):2410–2428, 2022.
- [16] Jean-Charles Rochet and Jean Tirole. Two-sided markets: A progress report. *RAND Journal of Economics*, 37(3):645–667, 2006.
- [17] Martin W. P. Savelsbergh and Marc Sol. The general pickup and delivery problem. *Transportation Science*, 29(1):17–29, 1995.
- [18] André Snoeck, Daniel Merchán, and Matthias Winkenbach. Designing urban micro-fulfillment networks for on-demand and same-day delivery. *Transportation Research Part B: Methodological*, 155:234–256, 2022.

- [19] Robert Stroh, Alan L. Erera, and Alejandro Toriello. Fleet sizing, service region partitioning, and dynamic batching for same-day delivery systems. *Transportation Science*, 56(5):1327–1347, 2022.
- [20] Zhaowei Tong, Xiao Wang, and Yidong Xu. Spatial pricing, courier remuneration, and delivery pooling in hyperlocal logistics. *Operations Research*, 71(4):1250–1268, 2023.
- [21] Marlin W. Ulmer, Barrett W. Thomas, Ann Melissa Campbell, and Nicholas Woyak. The instant delivery problem: Dynamic dispatching with order bundling. *Transportation Science*, 55(3):678–696, 2021.
- [22] E. Glen Weyl and Michal Fabinger. Pass-through as an economic tool: Principles and applications. *Journal of Political Economy*, 121(3):528–583, 2013.
- [23] Xiaojun Yang, Arne Karsten Strauss, Christine S. M. Currie, and Richard Eglese. Choice-based demand management in attended home delivery. *European Journal of Operational Research*, 250(1):208–220, 2016.
- [24] Xinjie Yang, Moritz Straub, and Tom Van Woensel. Differentiated time-slot pricing and delay tolerance in on-demand last-mile logistics. *European Journal of Operational Research*, 308(2):715–730, 2023.