Order Ahead for Pickup: Promise or Peril?

Problem definition: Mobile technologies have increasingly enabled remote customers to order ahead at quick-service restaurants. As customers travel to the service facility to pick up their orders, their orders also advance in the food preparation queue. It is widely believed that the ability to order ahead reduces customers' total delay and therefore allows restaurants to attract more orders and achieve higher throughput than if customers must order onsite. **Methodology/results**: We build a queueing-game-theoretic model to study a hybrid order-ahead scheme where some customers order ahead and some order onsite. Our analysis shows that the common practice of accepting all orders as they come in and requiring all orders to be irrevocable can cause a hybrid order-ahead scheme to surprisingly achieve lower throughput than a pure order-onsite scheme. The throughput shortfall can persist even when the service provider freely chooses whether to share queue-length information with remote customers. However, allowing remote customers to cancel orders not ready for pickup when they arrive at the service facility can restore the throughput advantage of ordering ahead over ordering onsite, but the cancellation scheme may nevertheless fall short of the non-cancellation scheme in throughput. We then study another mitigation strategy that rejects new remote orders if the number of outstanding orders reaches a certain threshold. When the service provider optimally sets the rejection threshold, the hybrid order-ahead-with-rejection scheme outperforms both its non-rejection counterpart and the pure-order-onsite scheme in throughput, but not necessarily the cancellation scheme. Finally, we propose an integrated mechanism that allows for both rejection and cancellation, subsuming all the previously considered hybrid order-ahead schemes as special cases. **Managerial implications**: Our paper highlights the unintended consequences of ordering ahead and provides prescriptive guidance for managing such a service system.

Key words: On-demand service, Omnichannel service, Order ahead, Travel, Reneging

1. Introduction

In today's on-demand economy, customers value *instant gratification* more than ever before and would like to have their demand fulfilled as promptly as possible. In response, quick-service restaurants are increasingly enabling customers to *order ahead* on demand and pick up their orders at the restaurant. Online food ordering is projected to be a $106 billion industry by 2031 (Business Research Insights 2024) and the key ingenuity of ordering ahead is that it allows customer orders to virtually advance in the order-processing queue while customers themselves physically travel to the restaurant. By the time they arrive, their orders will be near completion or even ready for pickup. This *parallel effect* contrasts with the tandem nature of a traditional *order-onsite* scheme in which customers must travel to the service facility and place orders only after they arrive onsite.

Hence, ordering ahead is believed to reduce customers' total delay, thereby attracting more orders and generating higher "transaction volume" (Pucci 2017) than the pure order-onsite scheme.

Ordering ahead of time usually also means commitment ahead of time. It is not uncommon for restaurants to require that all orders be final once they are placed. For example, Starbucks (2017) says in the FAQs about its Mobile Order & Pay that "once your order has been placed it cannot be delayed or canceled." Other restaurant chains (Peet's Coffee 2019, Subway 2020) have similar terms of use. While this *lock-in effect* is characteristic of ordering ahead, the pure order-onsite scheme does not require pre-travel commitment. Customers who order onsite can postpone their ordering decision until they arrive at the restaurant and see the status of the queue (i.e., how many people are waiting for their orders). Given that allowing customers to order ahead both attracts orders (with less delay) and retains orders (with a no-cancellation policy), one would naturally think that it would increase the restaurant's throughput. Our paper challenges this view.

We develop a queueing-game-theoretic model in which a service provider faces a mix of remote and local customers, both of whom are delay-sensitive. Remote customers are distanced from the service facility and it takes time for them to travel to the facility, whereas local customers are nearby and their travel time is negligible. Upon experiencing a need, a remote (local) customer decides whether to place an irrevocable order ahead (onsite). The service provider operates in a *hybrid* mode, taking both remote and onsite orders, and processing them following a first-come-first-served rule. Remote customers who order ahead travel to the service facility to pick up their orders. We compare the above hybrid order-ahead scheme with a pure order-onsite scheme in which all customers may only order onsite. In both schemes, customers see the number of outstanding orders onsite but not remotely. This is consistent with the practice that restaurants often put up digital screens in store to show the list of outstanding orders but nevertheless only share with remote customers a wait-time estimate at best rather than the length of the order-processing queue.

We find that contrary to conventional wisdom, the hybrid order-ahead scheme has lower throughput than the pure order-onsite scheme when the market size is intermediate and travel time is short (Theorem 1). The key culprit for this throughput shortfall is the lock-in effect of ordering ahead, which is a double-edged sword. On the one hand, customer demand is secured early on, which puts upward pressure on throughput. On the other hand, remote customers who order ahead may unknowingly place and commit to orders when the queue is already long. This mismatch can exacerbate congestion and make the actual wait time more unpredictable, which in turn, deters customers from placing orders in the first place. Hence, the lock-in effect also puts downward pressure on throughput. In fact, it outweighs the upward pressure from both the lock-in effect and the parallel effect combined when the market size is intermediate (which implies abundant orders and nontrivial congestion) and travel time is short (which implies a limited benefit from parallelization), causing the hybrid order-ahead scheme to fall short of the pure order-onsite scheme in throughput.

The argument above reveals two driving forces for the throughput shortfall: (1) remote customers place orders without knowing the real-time queue length; (2) remote customers commit to their orders even if they see a long queue upon arrival at the service facility. This begs the question of whether the throughput shortfall can be eliminated if the service provider has discretion in whether to share the real-time queue-length information with remote customers or if the service provider allows for order cancellation. We investigate the effectiveness of these two mitigation strategies. We find that the throughput shortfall can persist even after the first mitigation strategy is put in place (Theorem 2). The service provider's dilemma is that sharing such queue-length information remotely directly causes remote customers to stop ordering at the sight of a long queue but doing so also regulates congestion, which indirectly increases customers' willingness to order. When the market size is intermediate, the indirect benefit of regulating congestion can still give way to the direct fear of losing orders, prompting the service provider to prefer not sharing information. Hence, the throughput shortfall cannot be eliminated by merely adjusting the information policy.

By contrast, we find that the second mitigation strategy—allowing remote customers to cancel their orders when they are updated on their orders' queue positions upon arrival at the service facility—can eliminate the throughput shortfall, enabling the hybrid-order ahead scheme to outperform the pure-order onsite scheme (analytically shown in Theorem 3 for a small buffer system and numerically confirmed more broadly). Allowing cancellation is a self-regulating mechanism that does not deter remote customers with a long queue at the moment of ordering; rather, customers abandon on their own only if the queue is actually long when they have to wait onsite. However, cancellation addresses an existing problem by creating a new one: allowing cancellation reduces the throughput of the hybrid order-ahead scheme when the market size is small (and thus regulating congestion is a secondary concern) as it forgoes orders that could otherwise be captured (Theorem 4). Thus, allowing order cancellation is not an all-encompassing solution.

The limitations of the two proposed mitigation strategies motivate us to examine a third one: proactively rejecting new remote orders when the number of existing outstanding orders reaches a threshold that is optimally determined by the service provider to maximize throughput. Order rejection is practiced by some Starbucks stores that turn off point-of-sale systems used for mobile ordering when the stores are too busy (Dean 2021). We show that the service provider should not reject any incoming orders if the market size is small, but as the market size grows, the service provider may reject remote orders to tame the increased amount of congestion; in fact, when the market size is sufficiently large, the optimal rejection scheme will mimic the outcome of information sharing (Theorem 5). Thus, the service provider who can optimize the rejection threshold no longer has the incentive to share queue-length information remotely (Theorem 6). Importantly, the hybrid order-ahead-with-rejection scheme (with an optimal rejection threshold) achieves higher throughput than both the hybrid order-ahead scheme (without rejection or cancellation) and the pure order-onsite scheme when queue-length information is not shared remotely (Theorem 7). Like information sharing, the rejection scheme regulates congestion by forgoing orders at the outset, but unlike information sharing, the rejection threshold in the rejection scheme can be fine-tuned to ensure that inducing more customers to place orders does not come at the expense of letting go too many orders that have already been placed. We show that the rejection scheme beats the cancellation scheme in throughput when the market size is small or large (Theorem 8) but numerically find that the opposite can be true when the market size is intermediate. This speaks to the complementary strengths of the proactive approach of forgoing orders at the outset (i.e., rejection) and the reactive approach of forgoing orders in the process (i.e., cancellation).

This complementarity motivates us to propose an integrated mechanism that allows the throughput-maximizing service provider to control both the amount of order rejection and cancellation and thus subsumes all the previously considered hybrid order-ahead schemes as special cases. We show that this integrated mechanism reduces to the optimal rejection scheme (without cancellation) when the market size is extreme (Theorem 9) but numerically find that cancellation and rejection should be jointly used when the market size is intermediate. We also observe from our numerical study that overall, the hybrid order-ahead-with-rejection scheme has the smallest throughput gap from the integrated mechanism among all the simple schemes considered.

We study three model extensions. The first extension captures food quality degradation that may arise when remote orders are complete before customers arrive at the store. The second extension captures remote customers' channel choice on whether to order ahead, order onsite, or not order at all. The third extension allows the travel speed of remote customers to be heterogeneous. We demonstrate the robustness of many of our key insights.

2. Related Literature

Our paper contributes to the growing literature on omnichannel retail in general and omnichannel service operations in particular. Most relatedly, Gao and Su (2018) show that the adoption of online self-order technologies (such as mobile apps) increases restaurants' throughput. They argue that one driver of this result is the "advance order effect." However, instead of actually modeling the act of ordering in advance, they make the simplifying assumption that ordering online entails a lower waiting cost per unit time than ordering offline. This same assumption has been used by a series of papers that model customer behavior in settings broadly related to omnichannel services, including Baron et al. (2023), Cui et al. (2020), Feldman et al. (2023), Chen et al. (2022).

Our paper differs from this line of work in both modeling and insights. On the modeling front, our paper explicitly models customers' travel time and therefore the parallelism of travel time and waiting time (i.e., the system-state evolution during travel). Doing so allows us to capture the "advance order effect" with higher operational fidelity. Modeling travel time also naturally gives rise to two potential decision epochs spaced by a time lag (the moment a customer's need arises and the moment the customer arrives at the service facility) and enables us to build a particularly novel model of the hybrid-order-ahead-with-cancellation scheme that captures both strategic balking (not placing orders at the first decision epoch) and strategic reneging (canceling orders at the second), a rare combination in the literature. On the insight front, the extant literature (e.g., Gao and Su 2018, Baron et al. 2023) all points to online ordering as a means to increase throughput. In contrast, by carefully modeling the operational subtleties and customer incentives in such a service system, we find that the hybrid-order-ahead scheme can counterintuitively result in lower throughput than the pure-order-onsite scheme. We further propose mitigation strategies to overcome the throughput shortfall. Collectively, these new developments highlight the challenges in managing the order-ahead scheme and advance our theoretical understanding of such a system.

In the omnichannel-service-operations space, our paper complements Farahani et al. (2022), who also model travel time in ordering ahead—yet the focus is diametrically different. They study how to manage queues to best meet a pre-specified target of pickup time, balancing the tradeoff between earliness and tardiness of order readiness. As such, they abstract away from customers' strategic ordering decisions and focus on *supply-side* interventions. By contrast, our work carefully models customer incentives and examines the *demand-side* response to different order-ahead mechanisms.

Beyond the application area, our paper contributes to the queueing-economics literature that studies customers' strategic behavior in queueing systems, pioneered by Naor (1969). We refer to Hassin and Haviv (2003), Hassin (2016) for comprehensive reviews. In particular, Hassin and Roet-Green (2021) form the basis of our benchmark model of the pure-order-onsite scheme in which traveling and waiting are in tandem. They numerically show that to maximize throughput, the service provider should withhold queue-length information from remote customers when the market size is small but reveal such information when the market size is large. This insight is consistent with the earlier finding established in a simpler setting (Hassin 1986, Chen and Frank 2004).

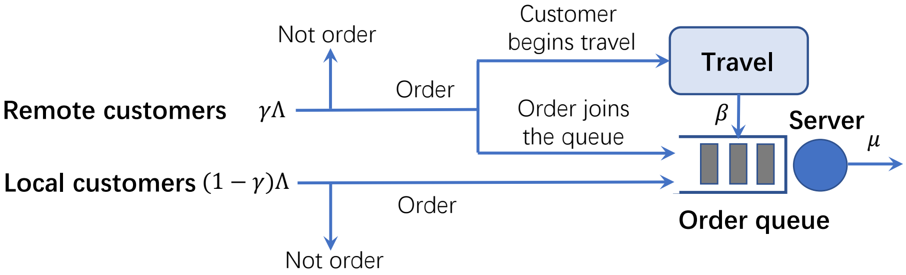
We not only extend the framework of Hassin and Roet-Green (2021) from homogeneous customers to heterogeneous customers (i.e., a mix of local and remote customers who differ in travel time) but also enrich this literature with various models of hybrid order-ahead schemes in which traveling and waiting are in parallel. Further, the ordering-onsite nature of Hassin and Roet-Green (2021) precludes strategic reneging, which nevertheless emerges in our cancellation model. The scant literature on strategic reneging has considered reneging triggered by either time-varying service rewards (Hassin and Haviv 1995), nonlinear waiting costs (Haviv and Ritov 2001), or random utility shocks (Ata and Peng 2018). By contrast, our cancellation model is novel in that reneging is driven by information about customers' updated queue position upon arrival at the service facility.

3. Model Setup

We model a service provider (e.g., a restaurant) as a single server that processes customer orders. Order processing times are independent and identically distributed (IID) random variables following an exponential distribution with mean $1/\mu$, where μ is referred to as the capacity of the service provider. Customer needs arise according to a Poisson process with rate Λ, where Λ is referred to as the market size. The market consists of two types of customers who differ in their physical location: remote customers and local customers. Let $\gamma \in (0, 1]$ and $1 - \gamma$ be the fraction of remote customers and local customers, respectively.

Remote customers are away from the service facility when their needs arise and their travel times to the service facility are IID random variables following an exponential distribution with mean $1/\beta$, where β is referred to as the travel speed. A remote customer is not entirely certain about her travel time before travel due to potentially unanticipated (elevator/road) traffic. Upon experiencing a need, a remote customer decides whether to place an (irrevocable) order online (from where she is located); if she places an order, she travels to the service facility to pick up her order. Local customers are near the service facility when their needs arise and their travel times to the service facility are negligible. Upon experiencing a need, a local customer decides whether to place an order onsite. Local customers can be thought of as those who stroll down the street and happen to pass by the store. The service provider processes orders on a first-come-first-served basis. See Figure 1 for an illustration of the process flow for this hybrid order-ahead scheme. We highlight two key features: (1) the service provider operates in a *hybrid* mode, with some customers ordering ahead and some ordering onsite (if at all); (2) ordering ahead has a *parallel* structure: as remote customers travel to the service facility, their orders are also advancing in the order-processing queue.

Figure 1 **A Hybrid Order-Ahead Scheme**



Customers receive a reward V for having their needs fulfilled on demand. Each customer incurs a delay cost c per unit time between the point she experiences a need and the point she receives her order. Customers are expected-utility maximizers. Consistent with practice, remote customers are provided with a wait-time estimate based on the historical average (e.g., on a mobile app), whereas local customers see the real-time queue length (i.e., the number of outstanding orders) as quick-service restaurants often set up in-store digital screens to display the status of outstanding orders. We will later investigate in §5 a case where the service provider can also share this real-time queue-length information with remote customers if doing so increases the system throughput.

To preclude trivial cases in which the service reward is too low for remote customers to ever place orders, we enforce Assumption 1 for the rest of the paper.

Assumption 1. $V > c\,[1/\mu + 1/\beta - 1/(\beta+\mu)]$.

4. Equilibrium and Comparison with Pure Order-Onsite

In this section, we first characterize customers' order-placing strategies in equilibrium (in which nobody can strictly increase her own expected utility through unilateral deviation). Next, we compare the equilibrium throughput of the hybrid order-ahead scheme with that of a pure order-onsite scheme in which both remote and local customers must order onsite (introduced in §4.3).

4.1. Preliminaries

This subsection derives the expected utility of a remote customer if she places an order for a given queue length. We first derive the probability distribution of the *queue position* for a remote customer's order after a random travel time. Suppose that when a remote customer's need arises, the initial queue length is $n \geq 0$ (i.e., n outstanding orders yet to be processed). If she places an order, she joins the back of the queue and her queue position is $n+1$. Let N_n denote her updated queue position upon arrival at the service facility. Thus, N_n is equal to $n+1$ less the number of service completions X up to her own order during the customer's travel. The support of N_n is $\{0, 1, \ldots, n+1\}$, where $N_n = 0$ means her order is complete and ready for pickup. Formally, the random variable $N_n \overset{\mathrm{d}}{=} [n+1-X]^+$, where $y^+ \equiv \max\{y,0\}$ and X is a geometric random variable with $\mathbb{P}(X=i) = [\beta/(\beta+\mu)][\mu/(\beta+\mu)]^i$ for $i=0,1,\cdots$. Denote $\sigma \equiv \mu/(\beta+\mu)$. We characterize the probability distribution of N_n in Lemma 1.

LEMMA 1 **(Updated queue-position distribution)**. *The probability distribution of N_n is:*

$$p_n(0) \equiv \mathbb{P}(N_n=0) = \sigma^{n+1}; \quad p_n(i) \equiv \mathbb{P}(N_n=i) = (1-\sigma)\sigma^{n-i+1}, \quad i=1,\cdots,n+1.$$

For a remote customer who places an order when the initial queue length is n, her expected utility conditioned on n, $\bar{U}(n) = V - c\sum_{i=0}^{(n+1)}(i/\mu)\cdot p_n(i) - c/\beta$. After simplification,

$$\bar{U}(n) \equiv V - \frac{c}{\beta}\left(\sigma^{n+1} + \frac{(n+1)\beta}{\mu}\right). \tag{1}$$

4.2. Equilibrium

This subsection characterizes customers' order-placing strategies in equilibrium. Local customers place an order if and only if the queue length they see is less than $n_e \equiv \lfloor \mu V/c \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. We refer to n_e as the Naor threshold (Naor 1969). Remote customers place an order with probability $q \in [0,1]$ (determined in equilibrium) based on the expected utility. Next, we characterize this equilibrium order-placing probability q.

Given remote customers' order-placing probability q and local customers' order-placing threshold n_e, let $\rho_T \equiv [\gamma\Lambda q + (1-\gamma)\Lambda]/\mu$, $\rho_R \equiv \gamma\Lambda q/\mu$. Thus, for $\rho_R < 1$, the steady-state probability of the number of outstanding orders being i is

$$\pi_i^u(q) = \begin{cases} \rho_T^i\left(\frac{1-\rho_T^{n_e}}{1-\rho_T} + \frac{\rho_T^{n_e}}{1-\rho_R}\right)^{-1}, & \text{for } i < n_e, \\ \rho_R^{i-n_e}\rho_T^{n_e}\left(\frac{1-\rho_T^{n_e}}{1-\rho_T} + \frac{\rho_T^{n_e}}{1-\rho_R}\right)^{-1}, & \text{for } i \geq n_e. \end{cases} \tag{2}$$

The expected utility for a remote customer who places an order is $U^u(q) = \sum_{n=0}^{\infty}\bar{U}(n)\pi_n^u(q)$. Thus, $q \in (0,1)$ is an equilibrium only if $U^u(q)=0$; $q=1$ is an equilibrium if $U^u(1)>0$, and $q=0$ is an equilibrium if $U^u(0)<0$. Proposition 1 characterizes the equilibrium strategy q_A^u.

PROPOSITION 1 **(Equilibrium)**. *There exist thresholds on market size Λ, $\underline{\lambda}_A^u$ and $\bar{\lambda}_A^u$, such that remote customers' equilibrium order-placing probability q_A^u is:*

$$q_A^u = \begin{cases} 1, & \text{if } \Lambda \leq \underline{\lambda}_A^u, \\ \hat{q} \in (0,1), & \text{if } \underline{\lambda}_A^u < \Lambda < \bar{\lambda}_A^u, \\ 0, & \text{if } \Lambda \geq \bar{\lambda}_A^u, \end{cases} \tag{3}$$

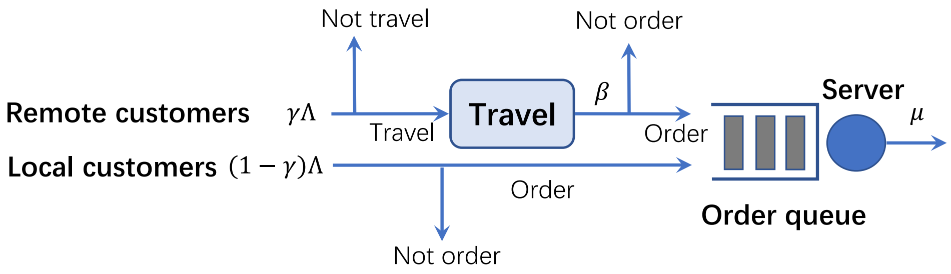
where $\hat{q}$ uniquely solves the equation $U^u(\hat{q})=0$. The resulting throughput $TH_A^u = \mu(1-\pi_0^u(q_A^u))$.

When the market size is sufficiently small ($\Lambda \le \underline{\lambda}^u_A$), the service system is not expected to be congested, and thus all remote customers place orders. Despite this, the resulting throughput is still lower than the market size Λ because local consumers do not always place orders if they see a long queue. When the market size is intermediate ($\underline{\lambda}^u_A < \Lambda \le \bar{\lambda}^u_A$), the system is expected to be somewhat congested, causing some remote customers to not place orders. When the market size is sufficiently large ($\Lambda > \bar{\lambda}^u_A$), all remote customers stop placing orders because the system is expected to be heavily congested even with local customers only.

4.3. Comparison with a Pure Order-Onsite Scheme

In this subsection, we compare the order-ahead scheme with a pure *order-onsite* scheme in which remote customers may place orders (i.e., join the queue) only after they travel to and arrive at the service facility. In the pure order-onsite scheme, as before, local customers decide whether to place orders based on the observed queue length; remote customers first decide whether to travel to the service facility and if they choose to travel, then upon arrival at the service facility, they further decide whether to place an order based on the observed queue length, just like local customers. See Figure 2 for an illustration of the process flow. Note that remote customers' decision process has a *tandem* structure: they need to travel to the service facility before placing an order.

Figure 2 **A Pure Order-Onsite Scheme**



When onsite, customers (local or remote) place an order if and only if the queue length they see is less than the Naor threshold n_e. Given the onsite order-placing threshold n_e and remote customers' travel probability $q \in [0,1]$ (determined in equilibrium), the system operates as an $M/M/1/n_e$ queue, and the steady-state probability of the number of outstanding orders being i is $\pi^u_{i,S}(q) = (\rho_T)^i / \sum_{j=0}^{n_e} (\rho_T)^j$, $i = 0, 1, \cdots, n_e$, where $\rho_T = [\gamma\Lambda q + (1-\gamma)\Lambda]/\mu$. Then a remote customer's expected utility of joining is $U^u_S(q) = \sum_{i=0}^{n_e-1} (V - c(i+1)/\mu)\, \pi^u_{i,S}(q) - c/\beta$. Thus, similar to Proposition 1, we can show that there exist thresholds on market size Λ, $\underline{\lambda}^u_S$ and $\bar{\lambda}^u_S$, such that a remote customer' equilibrium travel probability $q^u_S = \mathbf{1}_{\{\Lambda \le \underline{\lambda}^u_S\}} + \hat{q}_S \cdot \mathbf{1}_{\{\underline{\lambda}^u_S < \Lambda < \bar{\lambda}^u_S\}}$, where $\hat{q}_S \in (0,1)$ uniquely solves the equation $U^u_S(\hat{q}_S) = 0$. The resulting throughput $TH^u_S = \mu(1 - \pi^u_{0,S}(q^u_S))$. Notably, when the market size is large enough, remote customers stop traveling to the service facility, let alone placing orders. Theorem 1 compares the throughput of the pure order-onsite scheme TH_S^u with that of the hybrid order-ahead scheme TH_A^u.

Theorem 1 **(Hybrid order-ahead can be worse than pure order-onsite)**.

(*i*) *When travel speed β is sufficiently low or market size Λ is sufficiently small, the hybrid order-ahead scheme has higher throughput than the pure order-onsite scheme ($TH_A^u > TH_S^u$).*

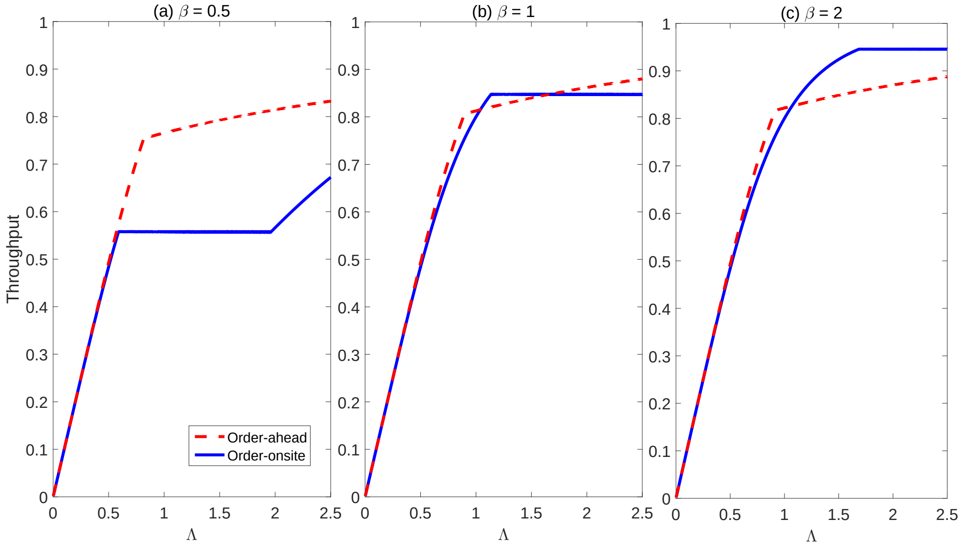
(*ii*) *When β is sufficiently high, for an intermediate range of the market size, the hybrid order-ahead scheme has lower throughput than the pure order-onsite scheme ($TH_A^u < TH_S^u$).*

Theorem 1 shows that whether the hybrid order-ahead scheme can achieve higher throughput than the pure order-onsite scheme depends on both remote customers' travel speed and the market size. Strikingly, if remote customers travel fast, then allowing remote customers to order ahead results in lower throughput than if they must order onsite when the market size is intermediate.

Here is the rationale. On the one hand, for remote customers, the hybrid order-ahead scheme parallelizes waiting (for order processing) and traveling (for order pickup). This *parallel effect* lures more remote customers and puts upward pressure on throughput. On the other hand, the hybrid order-ahead scheme requires remote customers to pre-commit to their order before they observe the real-time congestion, whereas remote customers in the pure order-onsite scheme can defer ordering decisions until they see the queue onsite. Pre-commitment is a double-edged sword for throughput. On the positive side, it secures customer orders early on. This *lock-in effect* puts upward pressure on the system throughput. On the flip side, remote customers may unknowingly place and commit to orders when the queue is already long. This exacerbates system congestion, which, in turn, deters both local and remote customers from placing orders. As a consequence, this lock-in effect also puts downward pressure on the system throughput. When travel is fast and the market size is intermediate, the parallel effect dwindles but the increased congestion due to the lock-in effect becomes formidable. The downward pressure from the lock-in effect on throughput overshadows the upward pressure from the parallel and lock-in effects combined, causing the hybrid order-ahead scheme to lag behind the pure order-onsite scheme in throughput.

We supplement Theorem 1 with a numerical trial illustrated by Figure 3. We observe from Figure 3-(a) that when the travel speed is low, the throughput of the hybrid order-ahead scheme always exceeds the throughput of the pure order-onsite scheme, consistent with Theorem 1-(i). We also observe from Figure 3-(b) and Figure 3-(c) that when the travel speed is not low, the hybrid order-ahead scheme has lower throughput than that of the pure order-onsite scheme for an intermediate

Figure 3 Throughput Comparison of Hybrid Order-Ahead vs. Pure Order-Onsite

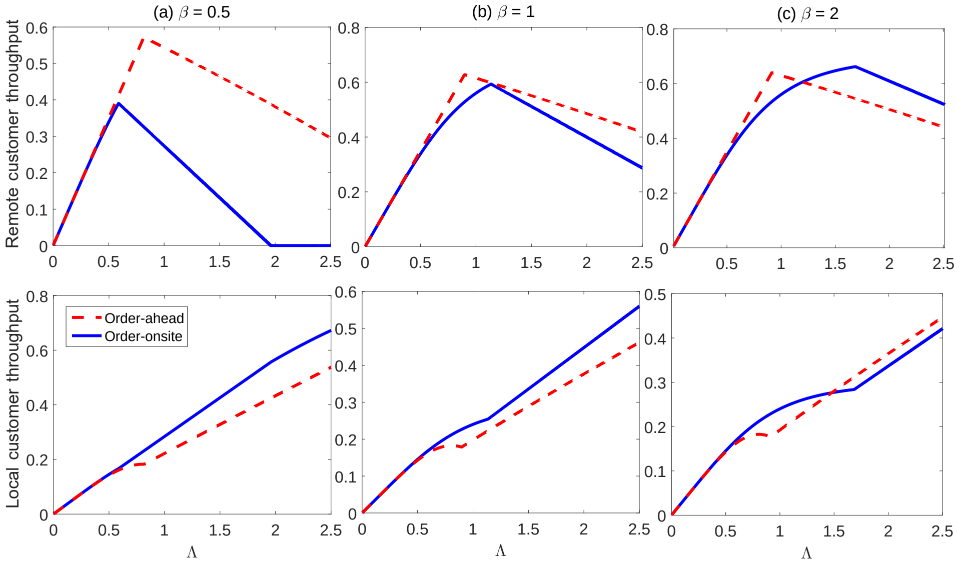


Note. $\mu = 1$, $V = 2$, $c = 0.5$, $\gamma = 0.7$.

market size, consistent with consistent with Theorem 1-(ii). Additional numerical studies reveal that the throughput shortfall does not necessarily hinge on the mean travel time being shorter than the mean processing time. In practice, the order processing time depends on the nature of the food: coffee may be quick to make but hot meals may take longer.

We further break down the total throughput into the throughput of remote customers and that of local customers. In the hybrid order-ahead scheme, define remote customer throughput to be $\gamma\Lambda q_A^u$ and local customer throughput to be $(1-\gamma)\Lambda\sum_{i=0}^{n_e-1}\pi_i^u(q_A^u)$. In the pure order-onsite scheme, define remote customer throughput to be $\gamma\Lambda q_S^u\sum_{i=0}^{n_e-1}\pi_{i,S}^u(q_S^u)$ and local customer throughput to be $(1-\gamma)\Lambda\sum_{i=0}^{n_e-1}\pi_{i,S}^u(q_S^u)$. We present our numerical findings in Figure 4. We observe from Figure 4-(a) and Figure 4-(b) that when travel is slow or intermediate, a change from the pure order-onsite scheme to the hybrid order-ahead scheme increases the throughput of remote customers but reduces the throughput of local customers. This observation is consistent with the popular belief that remote customers who order ahead crowd out local customers who order onsite (Buell 2020). It further reveals that the shortfall of the total throughput observed in Figure 3-(b) occurs as the increase in remote customers' orders does not make up for the loss of local customers' orders. Moreover, we observe from Figure 3-(c) that when travel is fast enough and the market size is large enough, remote customers' throughput falls while local customers' throughput rises, contrary to popular belief. Collectively, our analytical and numerical results demonstrate the intricacies of the hybrid order-ahead scheme. Neither the conventional wisdom that allowing remote customers to order ahead attracts more orders in total nor the popular belief that doing so at least attracts more orders from remote customers is always valid.

Figure 4 **Throughput Comparison of Hybrid Order-Ahead vs. Pure Order-Onsite by Customer Type**



Note. $\mu = 1$, $V = 2$, $c = 0.5$, $\gamma = 0.7$.

5. Mitigation Strategies

In this section, we consider three strategies that might mitigate the potential throughput shortfall of the hybrid order-ahead scheme: (i) in §5.1, we allow the service provider to share queue-length information with remote customers; (ii) in §5.2, we allow order cancellation from remote customers; and (iii) in §5.3, we allow order rejection from the service provider.

5.1. Information

Recall that a driver for the throughput shortfall of the hybrid order-ahead scheme discussed in §4.3 is that remote customers—who have access to wait time estimates based only on historical averages—may unknowingly place orders when the queue is already long. This begs the question of whether the throughput shortfall can be eliminated if the service provider has discretion in whether to share the real-time queue-length information with remote customers. This subsection explores remote queue-length information as a mitigation strategy.

5.1.1. Hybrid Order-Ahead Scheme We first study the hybrid order-ahead scheme. We start by characterizing remote customers' order-placing strategy when queue-length information is shared remotely, i.e., when remote customers are informed of the number of outstanding orders when deciding whether to place orders. Note that the expected utility of a remote customer who places an order after observing n outstanding orders, $\bar{U}(n)$, is derived in (1). Thus, a remote customer places an order if and only this expected utility is non-negative. Proposition 2 characterizes the order-placing strategy of a remote customer.

Proposition 2 **(Remote customer strategy with information)**. *When queue-length information is shared remotely, remote customers follow a threshold order-placing strategy:*

(i) A remote customer places an order if and only if she observes a queue length n less than threshold n_e^* (i.e., $n < n_e^*$), where n_e^* is uniquely determined by $n_e^* \equiv \min\{n \in \mathbb{N} : \bar{U}(n) < 0\}$.

(ii) Threshold n_e^* is no greater than the Naor threshold n_e, i.e., $n_e^* \leq n_e$. Specifically, if $\lfloor \frac{\mu V}{c} \rfloor = \frac{\mu V}{c}$, then $n_e^* < n_e$ for any $0 < \beta < \infty$; otherwise, $n_e^* < n_e$ if and only if travel speed β is low, i.e., $\beta < \underline{\beta}$, where $\underline{\beta}$ uniquely solves $V - \frac{cn_e}{\mu} - \frac{c}{\underline{\beta}}\sigma^{n_e} = 0$.

While the threshold structure of remote customers' order-placing strategy in Proposition 2-(i) is intuitive, Proposition 2-(ii) may be slightly less straightforward. Here is the rationale. The total delay remote customers experience is either (a) the travel time (if the order is ready before travel is complete) or (b) the delay between order generation and order completion (if travel is complete before the order is ready), whichever is longer. Hence, the total delay is expected to be longer than the delay in (b) alone, which is what local customers would endure for the same queue length. This implies that remote customers will generally be less receptive to a long queue than their local counterparts (i.e., $n_e^* \leq n_e$). In particular, when travel time is expected to be long, the total delay is also expected to be much longer than the delay in (b), inducing remote customers to adopt a strictly lower joining threshold than that of local customers (i.e., $n_e^* < n_e$).

Given local customers' order-placing threshold n_e and remote customers' order-placing threshold n_e^*, let $\rho \equiv \Lambda/\mu$ be the potential traffic intensity. Thus, the steady-state probability of the number of outstanding orders being i is

$$\pi_i^o = \begin{cases} \rho^i \left(\frac{1-\rho^{n_e^*}}{1-\rho} + \frac{\rho^{n_e^*}(1-((1-\gamma)\rho)^{n_e - n_e^* + 1})}{1-(1-\gamma)\rho} \right)^{-1}, & i = 0, 1, \cdots, n_e^*, \\ ((1-\gamma)\rho)^{i-n_e^*}\rho^{n_e^*} \left(\frac{1-\rho^{n_e^*}}{1-\rho} + \frac{\rho^{n_e^*}(1-((1-\gamma)\rho)^{n_e - n_e^* + 1})}{1-(1-\gamma)\rho} \right)^{-1}, & i = n_e^* + 1, \cdots, n_e. \end{cases}$$

The resulting system throughput is $TH_A^o = \mu(1 - \pi_0^o)$. Next, we take the perspective of a throughput-maximizing service provider who chooses whether to share queue-length information with remote customers, i.e., a service provider who solves the optimization problem: $\max\{TH_A^o, TH_A^u\}$, where throughput TH_A^u is defined in Proposition 1. Proposition 3 characterizes the service provider's information-sharing policy in the hybrid order-ahead scheme.

Proposition 3 **(Whether to share information)**. *There exists a unique threshold $\widetilde{\Lambda}$ such that the throughput-maximizing service provider should not share queue-length information with remote customers ($TH_A^u \geq TH_A^o$) if $\Lambda \leq \widetilde{\Lambda}$ and should share information otherwise.*

Proposition 3 generalizes the classical result of Chen and Frank (2004) to a setting of heterogeneous customers. When the market size is small, congestion is light, and withholding queue-length information from remote customers can induce all of them to place orders whereas if queue-length information is shared, remote customers who happen to see a long queue will refrain from ordering. Hence, not sharing information is preferred. By contrast, when the market size is large, congestion is nontrivial, and revealing queue-length information to remote customers induces them to place orders only when the queue is sufficiently short, which keeps the queue length in check and regulates congestion. This, in turn, entices more customers.

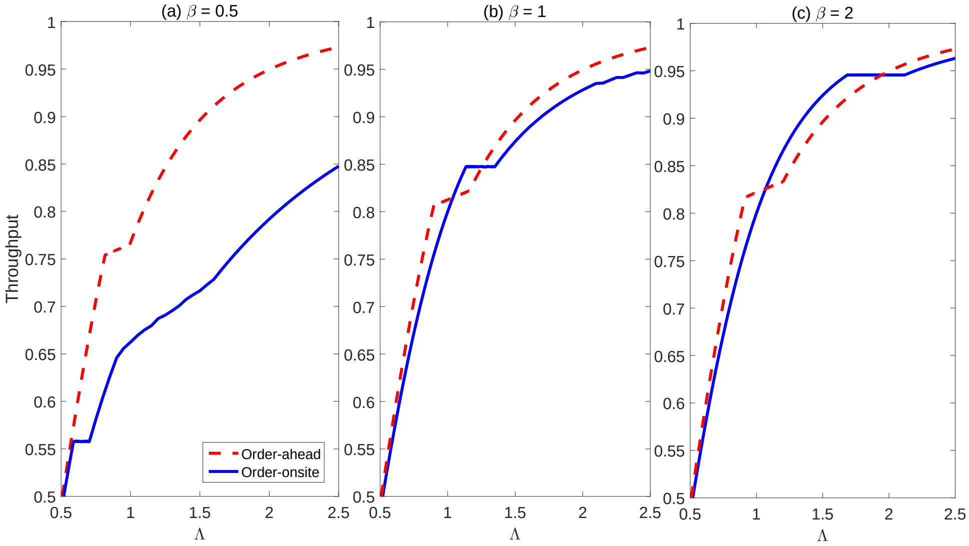
5.1.2. Pure Order-Onsite Scheme In the pure order-onsite scheme, customers follow a simple Naor joining threshold when onsite. However, if queue-length information is shared, then remote customers' strategy in deciding whether to travel after observing the queue length is highly complex. Hassin and Roet-Green (2021) study a simplified version of this problem that contains only (homogeneous) remote customers but not local customers. They show that finding remote customers' traveling equilibrium is analytically intractable and instead develop an algorithm to numerically search for the equilibrium. We extend their numerical procedure to our setting of heterogeneous customers and obtain the resulting throughput of the pure order-onsite scheme TH_S^o.

5.1.3. Throughput Comparison Let TH_A^* and TH_S^* denote the maximum throughputs achieved by the optimal remote-information-sharing policy in the hybrid order-ahead scheme and the pure order-onsite scheme, respectively. That is, $TH_i^* \equiv \max\{TH_i^o, TH_i^u\}$, $i \in \{A, S\}$. Theorem 2 compares TH_A^* with TH_S^*.

Theorem 2 **(Hybrid order-ahead can still backfire)**. *When travel speed β is sufficiently high, for an intermediate range of the market size, the hybrid order-ahead scheme has lower throughput than the pure order-onsite scheme even if the service provider optimally chooses whether to share queue-length information with remote customers in each respective scheme ($TH_A^* < TH_S^*$) .*

Theorem 2 shows that the potential throughput shortfall of the hybrid order-ahead scheme cannot be eliminated even when the service provider freely decides whether to share queue-length information with remote customers. The conundrum is that withholding queue-length information from remote customers causes a supply-demand mismatch that drives congestion (which is particularly problematic when congestion is already high, i.e., when the market size is large), but sharing the information turns customers away outright (which is particularly problematic when the service provider desperately needs customers, i.e., when the market size is small). Thus, when the market

Figure 5 **Throughput Comparison of Hybrid Order-Ahead vs. Pure Order-Onsite Under Optimal Information**



Note. $\mu = 1$, $V = 2$, $c = 0.5$, $\gamma = 0.7$.

size is neither too small nor too large, the service provider is pushed into a tight corner and cannot salvage the hybrid order-ahead scheme by merely adjusting information.

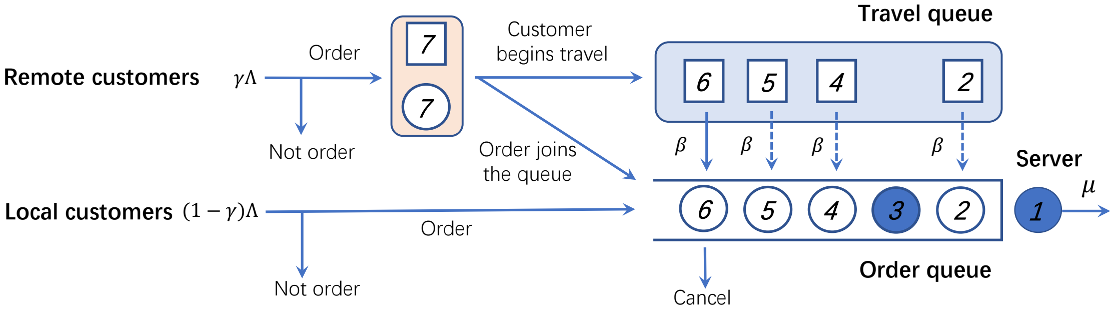
We supplement Theorem 2 with a numerical study that compares the throughput of the hybrid order-ahead scheme (TH_A^*) with that of the pure order-onsite scheme (TH_S^*) when the service provider optimally chooses whether to share queue-length information with remote customers in each respective scheme. The result is presented in Figure 5. Consistent with Theorem 2, we observe from Figure 5-(b) and Figure 5-(c) that when the travel speed is not too low and the market size is intermediate, the pure-order-onsite scheme outperforms the hybrid order-ahead scheme. Notably, comparing Figure 5-(b) and Figure 5-(c) with their counterparts of Figure 3 reveals instances in which the service provider switches to sharing queue information with remote customers in the hybrid order-ahead scheme while sticking to no-sharing in the pure order-onsite scheme, yet the throughput of the hybrid order-ahead scheme still falls behind that of the pure order-onsite scheme.

5.2. Cancellation

Recall that another driver for the throughput shortfall of the hybrid order-ahead scheme discussed in §4.3 is that remote customers commit to orders when the queue is already long. This begs the question of whether non-commitment (i.e., allowing remote customers to cancel orders when they arrive onsite) helps. This subsection explores order cancellation as a mitigation strategy.

In the hybrid order-ahead-with-cancellation scheme, the service provider informs onsite customers of the number of outstanding orders and the queue position of each order (e.g., by displaying a sequence of order ID's on in-store digital screens); further, (remote) customers who order ahead can freely cancel their unfinished orders when they arrive at the service facility. Upon cancellation, an order will be withdrawn from the order queue and will no longer be processed. Hence, not all orders initially placed will eventually be prepared. See Figure 6 for a process-flow illustration.

Figure 6 **Hybrid Order-Ahead with Cancellation**



Note. Customers and orders are depicted by squares and circles, respectively. The onsite order-placing threshold is $n_e = 4$. Customers 1 and 3 are onsite, waiting to pick up their orders (represented by the two solid circles); customers 2, 4, and 5 are still traveling; customer 6 is about to arrive at the service facility and cancel her order; customer 7 is about to place an order.

On the surface, this cancellation scheme bears a resemblance to the pure order-onsite scheme in that remote customers in both may choose not to stay after traveling to the service facility. Nevertheless, the key distinction is that in the pure order-onsite scheme, customers may choose to *balk* from the physical queue before committing to the service; while in the hybrid order-ahead scheme with cancellation, remote customers may choose to *renege* on a previously secured spot in the virtual order queue. As such, our model of the cancellation scheme captures both strategic balking (not placing orders) and strategic reneging (canceling orders).

We first characterize the cancellation strategy of remote customers who arrive at the service facility and observe their queue positions. One complication is that an arriving customer cannot tell whether those with order IDs ahead of hers have arrived or not. Customers who have not yet arrived may later cancel their orders upon arrival, thus affecting the focal customer's calculation about her own expected wait time if she does not cancel. Thus, each arriving customer needs to think strategically about the cancellation strategies of those who have not arrived. Yet, Lemma 2 shows that customers' cancellation strategy has a surprisingly simple threshold structure.

LEMMA 2 **(Cancellation strategy)**. *In the hybrid order-ahead-with-cancellation scheme, each remote customer cancels her order upon arrival at the service facility if and only if her queue position (the number of outstanding orders ahead of hers plus her own order) is greater than* n_e.

Here is the rationale behind Lemma 2. If a customer's queue position does not exceed the Naor threshold n_e, then her dominant strategy is to keep waiting for her order. This result further implies that the first n_e outstanding orders in the queue will not get canceled. Therefore, if a customer's queue position exceeds n_e upon arrival at the service facility, then she would definitely cancel her order because she knows that at least the first n_e orders will not get canceled and thus will be processed before hers (regardless of when their order-placing customers arrive), which implies that the reward from getting her own order is not worth the wait. Following the same logic, a local customer places an order if and only if the order queue length is less than n_e.

5.2.1. Queue-Length Information Not Shared Remotely We next divide our analysis by whether queue-length information is shared with remote customers when they decide whether to order. We start with the case in which such information is not shared.

Given remote customers' cancellation threshold n_e and local customers' joining threshold n_e at the service facility, we now derive remote customers' ordering placing probability q when they experience a need. Given q, the system state of the order queue, i.e., the number of outstanding orders i, evolves according to a birth-death process with a state-dependent birth rate $\lambda_i(q) = \gamma\Lambda q + (1-\gamma)\Lambda \cdot \mathbf{1}_{\{i \leq n_e - 1\}}$ for $i = 0, 1, \ldots$ and a state-dependent death rate μ_i: $\mu_i = \mu + \beta(i - n_e)^+$, $i = 0, 1, \cdots$. Birth rates $\lambda_i(q)$ correspond to order arrivals. When the queue is short $(i \leq n_e - 1)$, both remote and local customers place orders, hence the birth rate $\gamma\Lambda q + (1-\gamma)\Lambda$; otherwise, only remote customers place orders, hence the birth rate $\gamma\Lambda q$. Next, we explain death rate μ_i. An order cancellation occurs only when a remote customer completes traveling and finds a long order queue ahead of her order at the service facility. A remote customer who sees less than n_e outstanding orders ahead of hers will not cancel her order. Such a customer's arrival at the service facility will not trigger a "death" event in the system. Therefore, when all customers have a short queue ahead of their orders (i.e., the number of outstanding orders $i \leq n_e$), the order queue length can only be decremented by a service completion (which occurs at rate μ). On the other hand, if $i > n_e$, it must imply that $i - n_e$ remote customers have an order-queue position above n_e, and furthermore, these customers must all be traveling and have not yet arrived onsite (because otherwise, they would have already canceled their orders). Hence, in addition to a service-completion event, the order queue length can also be decremented by an order cancellation (which corresponds to one of those $i - n_e$ customers arriving at the service facility) at rate $\beta(i - n_e)$.

Given the birth and death rates, the steady-state probability of the number of the outstanding orders being i, $\pi^u_{i,C}(q)$, satisfies the flow balance equations $\lambda_i(q)\pi^u_{i,C}(q) = \mu_{i+1}\pi^u_{i+1,C}(q)$ for $i = 0, 1, \ldots$, from which, we obtain the following product-form steady-state probabilities:

$$\pi^u_{i,C}(q) = \left(\frac{1-\rho_T^{n_e+1}}{1-\rho_T} + \rho_T^{n_e}\sum_{j=1}^{\infty}\prod_{k=1}^{j}\frac{\gamma\Lambda q}{\mu+k\beta}\right)^{-1} \rho_T^{(i\wedge n_e)} \prod_{k=1}^{(i-n_e)^+}\frac{\gamma\Lambda q}{\mu+k\beta}, \quad i = 0, 1, \cdots, \tag{4}$$

where $\rho_T = [\gamma\Lambda q + (1-\gamma)\Lambda]/\mu$. Next, similar to Lemma 1, Lemma 3 characterizes remote customers' updated queue position upon arrival at the service facility (before she cancels, if at all), denoted by N^C_n, when her queue position is $n+1$ at the moment of ordering.

Lemma 3 **(Updated queue-position distribution under cancellation)**. *(i) If $n > n_e$, the probability distribution of N^C_n is $p^C_n(0) \equiv \mathbb{P}(N^C_n = 0) = \prod_{k=0}^{n}\frac{\mu_k}{\mu_k+\beta}$; $p^C_n(i) \equiv \mathbb{P}(N^C_n = i) = \frac{\beta}{\mu_{i-1}+\beta}\prod_{k=i}^{n}\frac{\mu_k}{\mu_k+\beta}$, $1 \le i \le n+1$, where $\mu_i = \mu + \beta(i-n_e)^+$, $i = 0, 1, \cdots$ and $\prod_{k=i}^{j} x_k = 1$ for $i > j$; (ii) If $n \le n_e$, N^C_n has the same distribution as N_n given by Lemma 1.*

Given q, the expected utility of a remote customer who places an order is

$$U^u_C(q) \equiv \sum_{i=0}^{n_e-1}\bar{U}(i)\pi^u_{i,C}(q) + \sum_{i=n_e}^{\infty}\left[\sum_{j=0}^{n_e}\left(V - \frac{cj}{\mu}\right)p^C_i(j) - \frac{c}{\beta}\right]\pi^u_{i,C}(q), \tag{5}$$

where $\bar{U}(i)$, $\pi^u_{i,C}(q)$, and $p^C_i(j)$ are given by Equations (1), (4), and Lemma 3, respectively. Proposition 4 characterizes remote customers' equilibrium order-placing probability q^u_C.

Proposition 4 **(Equilibrium in the cancellation model)**.
In the hybrid order-ahead-with-cancellation scheme, when queue-length information is not shared remotely, there exist thresholds on market size Λ, $\underline{\lambda}^u_C$ and $\bar{\lambda}^u_C$, such that remote customers' equilibrium order-placing probability $q^u_C = \mathbf{1}_{\{\Lambda \le \underline{\lambda}^u_C\}} + \widetilde{q}_C \cdot \mathbf{1}_{\{\underline{\lambda}^u_C < \Lambda < \bar{\lambda}^u_C\}}$, where $\widetilde{q}_C \in (0,1)$ uniquely solves $U^u_C(\widetilde{q}_C) = 0$, with $U^u_C(q)$ given in Equation (5). The resulting throughput is $TH^u_C = \mu[1 - \pi^u_{0,C}(q^u_C)]$.

5.2.2. Queue-Length Information Shared Remotely We next consider the case in which queue-length information is shared with remote customers when they decide whether to order. Recall from Proposition 2 that in such a case, remote customers place orders only when the observed queue length is less than n^*_e with $n^*_e \le n_e$. By the time they arrive onsite, their queue position will only be improved (at least no worse than n^*_e). Since Lemma 2 shows that a customer will only cancel if her queue position is worse than n_e, it implies that no customers have the incentive to cancel in the original order-ahead scheme. Hence, enabling cancellation does not make a difference when queue-length information is shared with remote customers. Hence, the system throughput in this case TH^o_C equals the throughput in its non-cancellation counterpart, TH^o_A, i.e., $TH^o_C = TH^o_A$.

5.2.3. Throughput Comparison Let $TH_C^* = \max\{TH_C^o, TH_C^u\}$ denote the maximum throughput achieved by optimal information sharing in the hybrid order-ahead-with-cancellation scheme. We next compare this throughput with that in the pure-order onsite scheme, TH_S^*.

THEOREM 3 **(Hybrid order-ahead-with-cancellation v.s. pure order-onsite)**. *When queue-length information is optimally shared remotely, for $n_e = 1$, the hybrid order-ahead-with-cancellation scheme has higher throughput than the pure order-onsite scheme, i.e., $TH_C^* \geq TH_S^*$.*

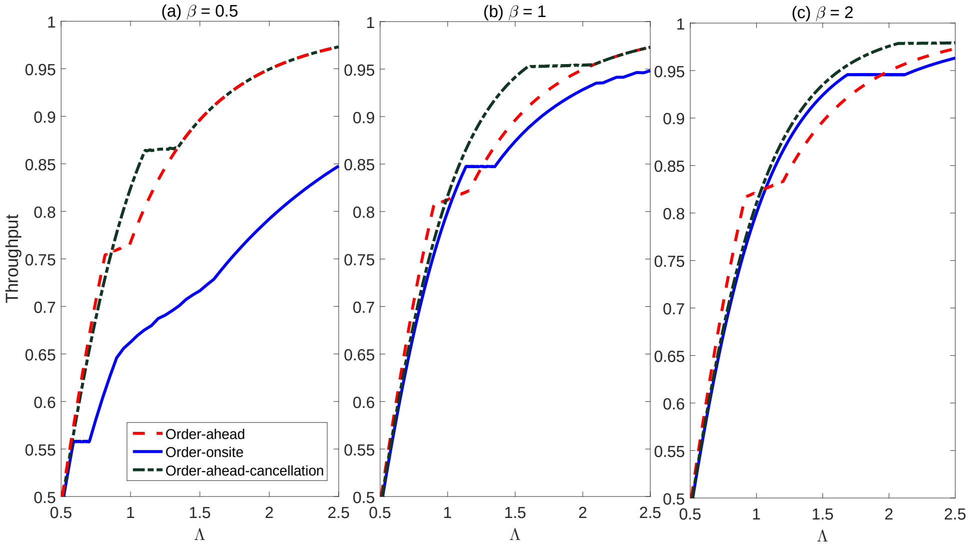
Theorem 3 shows that enabling cancellation mitigates the throughput shortfall under optimal information sharing; we prove the result analytically for $n_e = \lfloor V\mu/c \rfloor = 1$, but numerically, we find that it holds for all the problem instances tested (see §5.4 for details of our numerical study). Cancellation is a self-regulating mechanism that alleviates the issue of over-congestion created by the lock-in effect without turning remote customers away from the outset just because the queue is long initially (which would be the case if the service provider resorts to information sharing alone). Customers abandon only after they arrive at the service facility and actually expect a long wait onsite. Therefore, cancellation restores the advantage of the hybrid order-ahead scheme over the pure order-onsite scheme. Next, we investigate the impact of enabling cancellation on the throughput of the hybrid order-ahead scheme.

THEOREM 4 **(To cancel or not to cancel)**. *When queue-length information is optimally shared remotely, allowing cancellation in the hybrid order-ahead scheme results in lower throughput ($TH_C^* < TH_A^*$) if market size Λ is small.*

While Theorem 3 suggests that enabling cancellation is a promising solution that allows the hybrid order-ahead scheme to outperform the pure-order-onsite scheme, Theorem 4 shows that the cancellation scheme nevertheless falls short of the non-cancellation one when the market size is small. The cancellation scheme forgoes orders that the non-cancellation scheme would hold on to otherwise. This order loss is critical when there are not too many orders to begin with, i.e., when the market size is small. In this case, the cancellation scheme falls short of the non-cancellation scheme in retaining orders. Hence, cancellation addresses an existing problem (i.e., the throughput shortfall relative to the pure order-onsite scheme) by creating a new one (i.e., a potential throughput loss relative to the hybrid order-ahead scheme without cancellation).

Figure 7 illustrates a three-way throughput comparison of the hybrid order-ahead scheme (TH_A^*), the pure order-onsite scheme (TH_S^*), and the hybrid order-ahead-with-cancellation scheme (TH_C^*) when the service provider chooses the optimal information in each respective scheme. We observe that the hybrid order-ahead-with-cancellation scheme always outperforms the pure-order-onsite

Figure 7 Throughput Comparison of Hybrid Order-Ahead, Pure Order-Onsite, and Hybrid Order-Ahead-with-Cancellation Under Optimal Information



Note. $\mu = 1$, $V = 2$, $c = 0.5$, $\gamma = 0.7$.

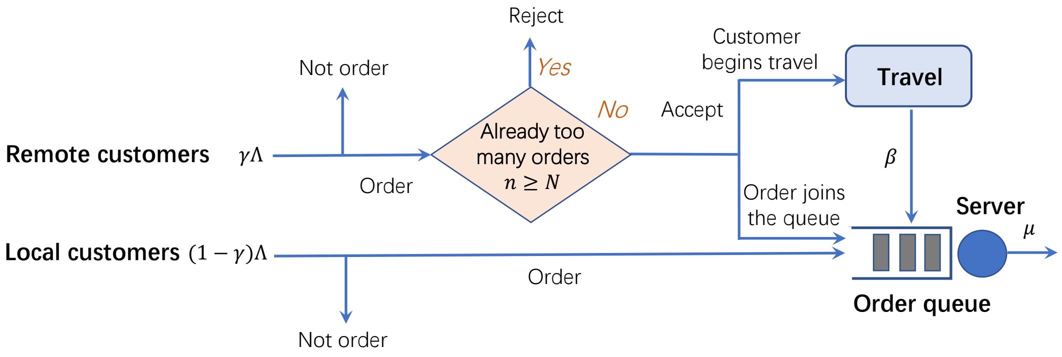
scheme, i.e., $TH_C^* > TH_S^*$ (confirming Theorem 3 for general n_e), mitigating the throughput shortfall of the non-cancellation scheme. Nevertheless, we also observe that the cancellation scheme falls short of its non-cancellation counterpart when the market size is small (confirming Theorem 4). Further, we observe that the cancellation scheme coincides with the non-cancellation scheme when the market size is large because both choose to share information and that the former achieves strictly higher throughput than the latter only when the market size is intermediate.

5.3. Rejection

The limitations of the previous two mitigation strategies motivate us to explore yet another alternative: a hybrid order-ahead-with-rejection scheme in which the service provider accepts new remote orders (placed by remote customers online) if the total number of outstanding orders is strictly less than a threshold $N \in \mathbb{N} \cup \{\infty\}$ and rejects any new remote orders otherwise. Threshold N is a decision variable of the service provider. Rejecting remote orders is in the spirit of Starbucks stores turning off point-of-sale systems for mobile ordering when the stores are too busy (Dean 2021). See Figure 8 for a process-flow illustration of the hybrid order-ahead-with-rejection scheme.

5.3.1. Queue-Length Information Not Shared Remotely We start with the case in which queue-length information is not shared with remote customers when they decide whether to order. We first characterize the equilibrium order-placing probability of remote customers for a

Figure 8 Hybrid Order-Ahead with Rejection



given rejection threshold N. If $N > n_e$, then given remote customers' order placing probability q, the steady-state probability of the number of outstanding orders being i is

$$\pi^u_{i,R}(q) = \begin{cases} \rho_T^i \left(\frac{1-\rho_T^{n_e}}{1-\rho_T} + \frac{\rho_T^{n_e}\left(1-\rho_R^{N-n_e+1}\right)}{1-\rho_R} \right)^{-1}, & \text{if} \quad i < n_e, \\ \rho_R^{i-n_e}\rho_T^{n_e} \left(\frac{1-\rho_T^{n_e}}{1-\rho_T} + \frac{\rho_T^{n_e}\left(1-\rho_R^{N-n_e+1}\right)}{1-\rho_R} \right)^{-1}, & \text{if} \quad i = n_e, \cdots, N, \end{cases} \tag{6}$$

where $\rho_T = [\gamma\Lambda q + (1-\gamma)\Lambda]/\mu$ and $\rho_R = \gamma\Lambda q/\mu$. If $N \leq n_e$, the steady-state probability of the number of outstanding orders being i is

$$\pi^u_{i,R}(q) = \begin{cases} \rho_T^i \left(\frac{1-\rho_T^{N}}{1-\rho_T} + \frac{\rho_T^{N}(1-\rho_L^{n_e-N+1})}{1-\rho_L} \right)^{-1}, & \text{if} \quad i < N, \\ \rho_L^{i-N}\rho_T^{N} \left(\frac{1-\rho_T^{N}}{1-\rho_T} + \frac{\rho_T^{N}(1-\rho_L^{n_e-N+1})}{1-\rho_L} \right)^{-1}, & \text{if} \quad i = N, \cdots, n_e, \end{cases} \tag{7}$$

where $\rho_T = [\gamma\Lambda q + (1-\gamma)\Lambda]/\mu$ and $\rho_L = (1-\gamma)\Lambda/\mu$. Thus, for a remote customer who places an order, with probability $\pi^u_{N,R}(q)$, her order will be rejected (from which she gets zero utility); with probability $1-\pi^u_{N,R}(q)$, her order will be accepted (which implies the queue length at the moment is less than N). Recall from (1) that a remote customer's expected utility conditioned on the queue length being n is $\bar{U}(n)$. Thus, her unconditional expected utility from ordering is $U^u_{R,N}(q) = \sum_{n=0}^{N-1} \bar{U}(n)\pi^u_{n,R}(q)$.

Proposition 5 characterizes remote customers' equilibrium order-placing probability.

PROPOSITION 5 **(Equilibrium in the N-rejection model)**. *In the hybrid-order-ahead scheme with a rejection threshold N, when queue-length information is not shared remotely, remote customers' equilibrium order-placing probability $q^u_{R,N}$ is as follows: if $\bar{U}(N-1) \geq 0$, $q^u_{R,N} = 1$; otherwise, there exist thresholds on market size Λ, $\underline{\Lambda}^u_{R,N}$ and $\bar{\Lambda}^u_{R,N}$, such that $q^u_{R,N} = \mathbf{1}_{\{\Lambda \leq \underline{\Lambda}^u_{R,N}\}} + \hat{q}_{R,N} \cdot \mathbf{1}_{\{\underline{\Lambda}^u_{R,N} < \Lambda < \bar{\Lambda}^u_{R,N}\}}$, where $\hat{q}^u_{R,N}$ uniquely solves $U^u_{R,N}(\hat{q}^u_{R,N}) = 0$. The resulting throughput is $TH^u_{R,N} = \mu[1-\pi^u_{0,R}(q^u_{R,N})]$.*

Building on Proposition 5, we characterize the optimal rejection threshold N^* that maximizes the system throughput. Theorem 5 establishes the structural properties of N^*. Let the maximum throughput achieved by N^* be TH^u_R, i.e., $TH^u_R = TH^u_{R,N^*} = \max_{N \in \mathbb{N} \cup \{\infty\}} \{TH^u_{R,N}\}$.

THEOREM 5 **(The optimal rejection threshold)**. *In the hybrid-order-ahead-with-rejection scheme, when queue-length information is not shared remotely, the optimal rejection threshold N^* satisfies $N^* \geq n^*_e$, with $N^* = n^*_e$ when Λ is large enough and $N^* = \infty$ when Λ is small enough.*

In setting the rejection threshold, the service provider faces the following tradeoff: a lower, more aggressive rejection threshold forgoes a bigger fraction of remote orders placed but prompts more remote consumers to place remote orders because it shortens the queue and better regulates congestion. When the market size is small enough, there is not much congestion anyway, so the service provider should not reject any orders (i.e., $N^* = \infty$). In this case, the rejection scheme degenerates into a non-rejection scheme. As the market size increases, congestion becomes a growing concern, so the service provider may reject orders to tame congestion. However, the optimal rejection threshold should never be set any lower than n^*_e, the order-placing threshold that remote customers employ in the non-rejection scheme when they have queue-length information at the time of ordering (see Proposition 2). Recall that $\bar{U}(n^*_e - 1) \geq 0$. Thus, if $N = n^*_e$, then from Proposition 5, $q^u_{R,N} = 1$, i.e., if $N = n^*_e$, the queue is guaranteed to be short enough that all remote customers place orders. Reducing the rejection threshold even further will only forgo a bigger fraction of the remote orders without prompting more remote customers to place orders and is therefore not fruitful. In fact, $N^* = n^*_e$ is indeed the optimal rejection threshold for a sufficiently large market size, in which case, the rejection scheme without information sharing mimics the information-sharing scheme without rejection (characterized in Proposition 2) in the induced customer behavior and throughput.

5.3.2. Queue-Length Information Shared Remotely

When queue-length information is shared with remote customers, let the maximum throughput achieved by the optimal rejection threshold in this case be TH^o_R. Theorem 6 shows that if the service provider can choose both whether to share information and the rejection threshold, then sharing information is not necessary.

THEOREM 6 **(No Incentive to share information in the rejection scheme)**. *In the hybrid order-ahead scheme with an optimized rejection threshold, not sharing queue-length information with remote customers achieves higher throughput than sharing, i.e., $TH^u_R \geq TH^o_R$.*

Sharing queue-length information with remote customers drives them away when the queue is long, which helps regulate congestion. Yet, the same effect can also be achieved by rejecting orders when the queue is long. Therefore, the service provider can replicate the order-placing outcome of information sharing with an appropriate rejection threshold without sharing information. More specifically, if queue-length information is shared, the service provider cannot improve throughput by imposing a rejection threshold. This is because a rejection threshold higher than n_e^* will not be reached (since remote customers will voluntarily stop ordering once the queue length reaches n_e^*), whereas a rejection threshold lower than n_e^* will only hurt throughput by turning down too many orders. Therefore, the service provider who does not share information can at least match the throughput of the information-sharing case by simply imposing a rejection threshold of n_e^*, which ensures all remote customers are willing to place orders. Optimizing over the rejection threshold will only increase the throughput further. Hence, the service provider will not bother to share queue-length information with remote customers. In other words, let $TH_R^* \equiv \max\{TH_R^o, TH_R^u\}$ be the maximum throughput achieved by optimizing over whether to share queue-length information with remote customers along with the rejection threshold; Theorem 6 establishes that $TH_R^* = TH_R^u$.

So far, we have implicitly assumed that customers rejected remotely will exit the system. Yet, one may naturally wonder if these customers, upon rejection, will instead travel to the service facility and order onsite. Proposition 6 shows that in the optimal rejection scheme, rejected customers indeed have no incentive to make such an attempt even when given the opportunity.

Proposition 6 **(Rejected customers will not order onsite)**. *In the optimal hybrid-order-ahead-with-rejection scheme, rejected customers will not choose to order onsite.*

Recall from Theorems 5 and 6 that the optimal rejection threshold is at least n_e^*. When customers are rejected remotely, they know the queue is too long for ordering ahead to be worth it anyway (recall from Proposition 2 that customers will not order ahead if there are n_e^* outstanding orders). We further show that ordering onsite will be even worse than ordering ahead and thus, rejected customers have no incentive to travel to the service facility and make a second ordering attempt.

5.3.3. Throughput Comparison We next compare the throughput of the hybrid order-ahead-with-rejection scheme (with a throughput-maximizing rejection threshold) with those of the three schemes introduced earlier: (1) the non-rejection scheme; (2) the pure order-onsite scheme; and (3) the cancellation scheme, when each scheme has its respective optimal remote information sharing policy. As for (1), the non-rejection scheme essentially has a rejection threshold of $N = \infty$ and is thus a special case of the rejection scheme with an optimized rejection threshold. Hence, it follows that the rejection scheme outperforms the non-rejection scheme, i.e., $TH_R^* \geq TH_A^*$. Theorems 7 and 8 address comparisons (2) and (3), respectively.

Theorem 7 **(Hybrid order-ahead with rejection vs. pure order-onsite)**. *When queue-length information is optimally shared remotely, for $n_e = 1$, the hybrid order-ahead-with-rejection scheme has higher throughput than the pure order-onsite scheme, i.e., $TH_R^* \geq TH_S^*$.*

Theorem 7 shows that the twist of order rejection mitigates the throughput shortfall of the hybrid order-ahead scheme, enabling it to capture more customers than the pure order-onsite scheme. We prove that this throughput dominance holds for general n_e if queue-length information is not shared remotely (i.e., $TH_R^u \geq TH_S^u$ for general n_e), but when information is shared in the pure order-onsite scheme, the underlying queueing system becomes analytically intractable and thus we can only analytically prove this result for $n_e = 1$, even though numerically, we find that it holds for all the problem instances tested (see §5.4 for details of our numerical study). The introduction of order rejection keeps the queue length in check. Thus, order rejection regulates congestion and enables customers who successfully order ahead to enjoy the benefit of the parallel effect without worrying about the longer-than-usual delay they might otherwise encounter in the non-rejection scheme. Therefore, customers are more willing to place orders. Moreover, the rejection threshold can be fine-tuned to strike the balance between acquisition (getting more customers to place orders) and retention (keeping more orders that have been placed) so that the hybrid order-ahead throughput will indeed be higher than the pure-order-onsite throughput.

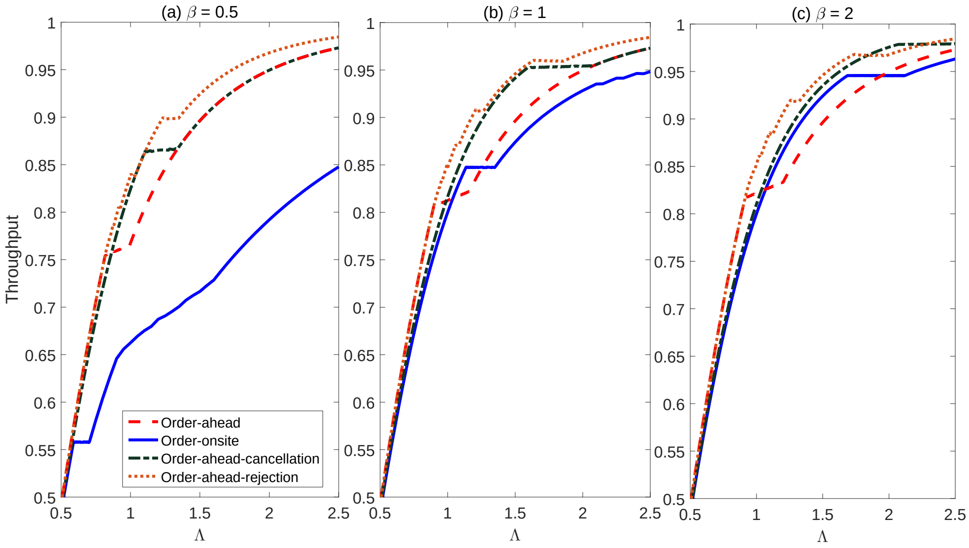
Theorem 8 compares the throughput of the rejection scheme with that of the cancellation scheme.

Theorem 8 **(To cancel or to reject)**. *When queue-length information is optimally shared remotely, the hybrid order-ahead-with-rejection scheme has higher throughput than the cancellation scheme ($TH_R^* \geq TH_C^*$) when the market size is sufficiently small or large.*

When the market size is small, recall from Theorem 5 that the rejection scheme reduces to the non-rejection scheme, which has higher throughput than the cancellation scheme, according to Theorem 4. Therefore, in this case, the rejection scheme outperforms the cancellation scheme. When the market size is large, the rejection scheme fends off orders at the outset, which more sharply regulates congestion that the cancellation scheme that lets customers voluntarily withdraw orders in the process. Hence, in this case, the rejection scheme again outperforms the cancellation scheme. However, when the market size is intermediate, it is unclear whether the rejection still beats the cancellation scheme. We explore this question numerically in Figure 9.

Figure 9 conducts a four-way throughput comparison of the hybrid order-ahead scheme (TH_A^*), the pure order-onsite scheme (TH_S^*), the hybrid order-ahead-with-cancellation scheme (TH_C^*), and the hybrid order-ahead-with-rejection scheme (TH_R^*) when the service provider optimally chooses

Figure 9 **Throughput Comparison of Hybrid Order-Ahead, Pure Order-Onsite, Hybrid Order-Ahead-with-Cancellation, Hybrid Order-Ahead-with-Rejection Under Optimal Information**



Note. $\mu = 1$, $V = 2$, $c = 0.5$, $\gamma = 0.7$.

whether to share queue-length information with remote customers in each respective scheme. We observe that the hybrid order-ahead-with-rejection scheme always outperforms both the hybrid order-ahead scheme ($TH_R^* \geq TH_A^*$), which is by construction, and the pure-order-onsite scheme ($TH_R^* \geq TH_S^*$), confirming Theorem 7 for general n_e. Hence, like the cancellation scheme, order rejection can be yet another approach to mitigate the throughput shortfall. Yet, unlike the cancellation scheme, order rejection does not have the unintended consequence of falling short of the basic scheme without cancellation or rejection (since the rejection threshold can be optimized). Further, in many instances, the rejection scheme also dominates the cancellation scheme (echoing Theorem 8), but this is not always the case. For example, Figure 9-(c) shows that when travel is fast and the market size is intermediate, the rejection scheme results in lower throughput than the cancellation scheme (although the magnitude of the difference seems small). The rationale is that order rejection is a more drastic measure of regulating congestion than allowing order cancellation and therefore can overshoot.

We have also numerically examined the impact of γ (the share of remote customers) on the throughput comparison. We observe that when γ is small, the throughputs of the various schemes considered are barely distinguishable from each other. This is not surprising since these schemes differ in how they affect remote customers; if the vast majority of customers are locals, then there clearly will not be any sizable differences across these schemes. As γ increases, the throughputs of the different schemes diverge, making our findings (e.g., the throughput shortfall of ordering ahead and the throughput restoration of the rejection scheme) more salient. This observation suggests that our insights are particularly relevant for restaurants with broad geographical coverage and a keen interest in remote ordering, which might be a growing industry trend.

In sum, the hybrid order-ahead-with-rejection scheme holds promise as it attains higher throughput than both the hybrid order-ahead scheme (by construction) and the pure order-onsite scheme (proved analytically in Theorem 7 for $n_e = 1$ and confirmed numerically for general n_e). While it is not guaranteed to outperform the hybrid order-ahead-with-cancellation scheme, it tends not to fall far behind (based on numerical observation). However, in order for the rejection scheme to work in practice, the rejection threshold must be carefully calibrated to the business characteristics and clearly communicated to remote customers, both of which are not without practical challenges.

5.4. An Integrated Mechanism

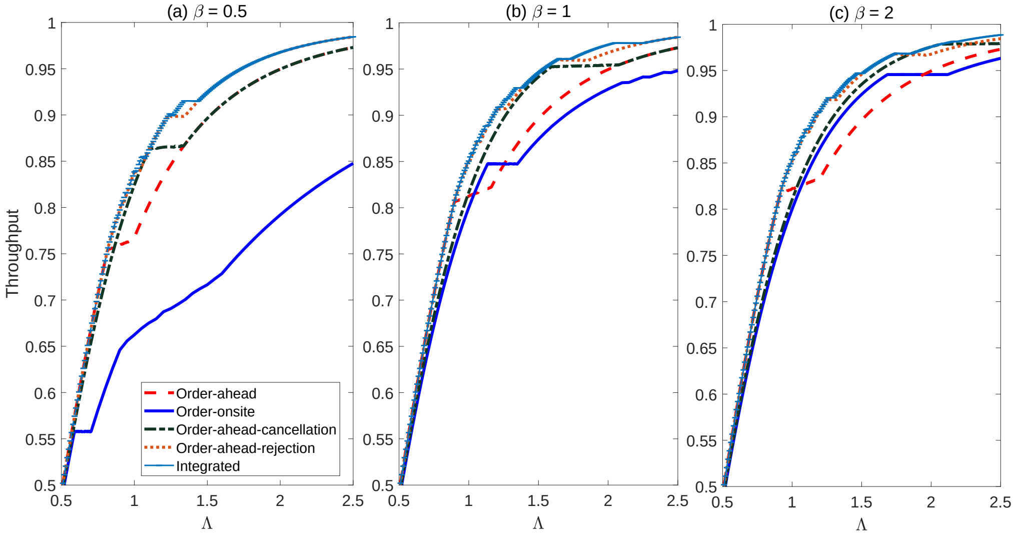
In this subsection, we consider an integrated mechanism that subsumes all the previously studied hybrid order-ahead schemes. In this integrated mechanism, the service provider (1) rejects remote orders if the total number of outstanding orders reaches or exceeds threshold $N_1 \in \mathbb{N} \cup \{\infty\}$ and (2) cancels online orders from remote customers when they arrive at the service facility and still have a queue position exceeding threshold $N_2 \in \mathbb{N} \cup \{\infty\}$. Both thresholds N_1 and N_2 are decision variables of the throughput-maximizing service provider. We focus on rejection thresholds with $N_1 \geq n_e^*$ to prevent creating perverse incentives for rejected customers to reorder onsite (Proposition 6); we focus on cancellation thresholds with $N_2 \geq n_e$ to prevent creating perverse incentives for canceled customers to reorder onsite. When queue-length information is shared remotely, this integrated mechanism reduces to the hybrid order-ahead scheme with information sharing in §5.1 and is throughput-dominated by the optimal rejection scheme without information sharing (Theorem 6). We henceforth focus on the case without information sharing (which is without loss of optimality). Further, when $N_1 = \infty$ and $N_2 = n_e$, this integrated mechanism degenerates into the cancellation scheme in §5.2; when $N_2 = \infty$, this integrated mechanism degenerates into the rejection scheme in §5.3. Theorem 9 (partially) characterizes the structure of the optimal integrated mechanism for the throughput-maximizing service provider.

Theorem 9 **(The Optimal Integrated Mechanism)**. *In the integrated mechanism, not rejecting or canceling orders ($N_1 = N_2 = \infty$) is optimal if the market size is sufficiently small; rejecting orders at threshold n_e^* ($N_1 = n_e^*, N_2 = \infty$) is optimal if the market size is sufficiently large.*

Theorem 9 shows that the integrated mechanism matches the optimal rejection scheme (without cancellation) when the market size is extreme (cf. Theorem 5). However, we observe from Figure

10 that when the market size is intermediate, the integrated mechanism can strictly outperform all of the four previously considered schemes (each under its respective optimal information sharing). This indicates that in such instances, both order rejection and cancellation are active in the optimal integrated mechanism ($N_2 < N_1 < \infty$), even though Figure 10 seems to suggest a small throughput gap between the rejection scheme and the integrated mechanism.

Figure 10 Throughput Comparison with the Integrated Mechanism



Note. $\mu = 1$, $V = 2$, $c = 0.5$, $\gamma = 0.7$.

We next quantify such a gap more systematically through a numerical study. We generate 625 representative instances from the following parameter combinations: $\mu = 1$, $c = 0.5$, $V \in \{1.5, 2.5, \cdots, 5.5\}$, $\gamma \in \{0.1, 0.3, \cdots, 0.9\}$, $\Lambda \in \{0.5, 1, \cdots, 2.5\}$, $\beta \in \{0.5, 1, \cdots, 2.5\}$ (all satisfying Assumption 1). In each instance, we compute the percentage throughput gap between the optimal integrated mechanism and each of the four previously considered schemes under optimal information sharing (the percentage throughput gap is the throughput difference divided by the throughput of the integrated mechanism). We present statistics of these percentage throughput gaps in Table 1, including the mean, median, maximum, minimum, first quartile, and third quartile.

Table 1 indicates that the hybrid order-ahead-with-rejection scheme is overall the closest to the integrated mechanism in throughput, with a maximum throughput loss of only 2.45% (even though it does not always dominate the cancellation scheme). Hybrid order-ahead-with-cancellation is also relatively effective but shows more variability in performance. While its throughput gap is substantially smaller than the hybrid order-ahead scheme (without cancellation) in worst-case scenarios (9.58% vs 17.26%), its median throughput gap turns out to be larger (0.11% vs 0.02%).

Table 1 The Percentage Throughput Gap of Different Schemes

Scheme	Mean	Median	Max.	Min.	1st Qu.	3rd Qu.
Hybrid order-ahead-with-rejection	0.07%	0	2.45%	0	0	0
Hybrid order-ahead-with-cancellation	0.71%	0.11%	9.68%	0	0.01%	0.72%
Hybrid order-ahead	1.12%	0.02%	17.26%	0	0	0.81%
Pure order-onsite	2.76%	0.44%	89.88%	0	0.08%	2.3%

$\mu = 1$, $c = 0.5$, $V \in \{1.5, 2.5, 3.5, 4.5, 5.5\}$, $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $\beta \in \{0.5, 1, 1.5, 2, 2.5\}$, $\Lambda \in \{0.5, 1, 1.5, 2, 2.5\}$.

The pure order-onsite scheme is overall the least effective in maintaining throughput, demonstrating the biggest gaps in all statistical measures (even though it outperforms the hybrid order-ahead scheme in some instances). In sum, our numerical study underscores the value of ordering ahead and points to order rejection within the hybrid order-ahead scheme as an effective control lever that balances simplicity and performance.

6. Extensions

6.1. Food-Quality Degradation

This extension incorporates the issue of food-quality degradation. Specifically, when remote customers order ahead, food can be ready before customers arrive, and thus may be "soggy" at the time of pickup. Our base model assumes away the disutility caused by "soggy" food and still finds that the hybrid order-ahead scheme may result in lower throughput than the pure order-onsite scheme. Incorporating such disutility in ordering ahead would imply that remote customers are even less inclined to place orders, leading to even lower throughput, thus only strengthening this key insight. In §EC.1.1, we formally model food deteriorating in quality over time after an order is complete. We find that our most interesting results that occur when travel is fast are particularly robust in that incorporating food-quality degradation hardly affects the system throughput of any scheme. This is because when travel is fast, customers are likely to arrive at the service facility before their order is complete, making food-quality degradation a secondary concern.

6.2. Channel Choice

This extension expands remote customers' strategy space and allows for channel choice in hybrid order-ahead schemes. When a need arises, remote customers decide whether to order ahead, order onsite, or not order at all. That is, remote customers not only choose whether to order (as in the base model), but also which channel to order from. When queue-length information is shared remotely, remote customers will not choose to order onsite, and therefore our results from the base model carry over. However, when queue-length information is not shared remotely, remote customers face a tradeoff in the channel choice: ordering ahead allows an order to join the queue earlier but ordering onsite prevents customers from unknowingly joining a long queue. Thus, a remote customer may choose to order onsite with a certain probability. Nevertheless, if remote customers can cancel their orders upon arrival at the service facility, then they again will not order onsite. Thus, even when queue-length information is not shared remotely, modeling the channel choice may only affect the hybrid order-ahead scheme (with or without rejection), but not the cancellation scheme or the pure order-onsite scheme. In §EC.1.2, we formally characterize the order-placing equilibrium in these affected schemes and demonstrate the robustness of our insights.

6.3. Heterogeneous Travel Speed of Remote Customers

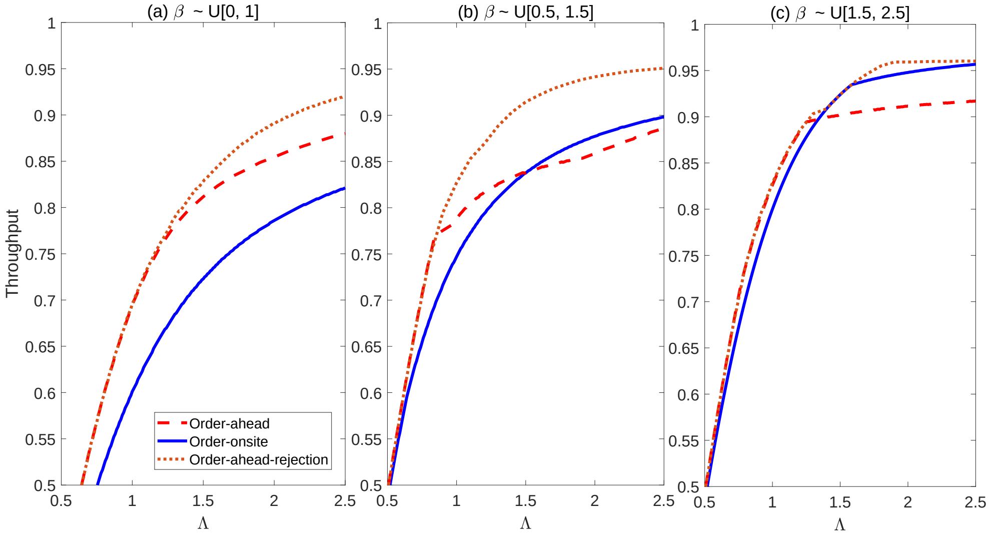
This extension allows remote customers' travel speed to be heterogeneous. Let remote customers' travel speed β be continuously distributed over support $[a, b]$, where $0 \leq a < b \leq \infty$. For a remote customer with travel speed β, her travel time is drawn from an exponential distribution with rate β. In the hybrid order-ahead scheme, each remote customer chooses whether to order ahead, order onsite, or not order, based on their own travel speed β. We set up the model in §EC.1.3 and characterize remote customers' ordering strategy in Proposition 7.

Proposition 7 **(A double-threshold strategy)**. *Under heterogeneous travel speed, in the hybrid order-ahead scheme without queue-length information shared remotely, there exist two thresholds β_1, β_2 with $a \leq \beta_1 \leq \beta_2 \leq b$ such that a remote customer with travel speed β does not order if $\beta < \beta_1$, orders ahead if $\beta_1 \leq \beta \leq \beta_2$, and orders onsite if $\beta > \beta_2$.*

Proposition 7 shows that customers adopt a double-threshold ordering strategy in the hybrid order-ahead scheme. Those with a high β (those who live near or travel fast, e.g., by car) order onsite because the benefit of ordering ahead (parallelization) for these customers is outweighed by the benefit of ordering onsite (queue-length information); those with an intermediate β order ahead because their time savings from ordering ahead (due to parallelization) is significant enough to prevail over the lack of information; and those with a low β (tho who live afar or travel slowly, e.g., by foot) do not place orders because they expect too much delay with either ordering mode.

We numerically compare the throughputs of three schemes: (i) pure order-onsite, (ii) hybrid order-ahead, and (iii) hybrid order-ahead with rejection (with the optimal rejection threshold) when queue-length information is not shared remotely. We observe from Figure 11 that consistent with the base model, the hybrid order-head scheme can have lower throughput than the pure-order-onsite scheme, yet introducing order rejection into the order-ahead scheme restores its throughput advantage. In fact, we analytically prove this result of throughput dominance in Theorem 10.

Figure 11 Throughput Comparison Under Heterogeneous Travel Speed of Remote Customers



Note. $\mu = 1$, $V = 2$, $c = 0.5$, $\gamma = 0.7$.

THEOREM 10. *Under heterogeneous travel speed, when queue-length information is not shared remotely, the hybrid order-ahead-with-rejection scheme (with the rejection threshold optimized) has higher throughput than the pure order-onsite scheme.*

We acknowledge that one limitation of this extension is the omission of the cases where queue-length information is shared remotely and those that permit order cancellations due to their intractability. Due to the heterogeneity in remote customers' travel speed, if information is shared or cancellation is allowed, then the computation of throughput requires deriving the steady-state distribution of a high-dimensional Markov chain that tracks the travel speed of every single traveling customer en route to the service facility, which would be best left for future research.

7. Conclusions

A key value proposition of letting customers order ahead is that doing so presumably attracts more orders and achieves higher throughput than if customers must order onsite. Our paper cautions that whether ordering ahead delivers this value hinges on the way it is operationalized. Specifically, a common practice in the field—all orders are final once placed—can generate orders that are placed and locked in when the queue is already long, burdening the service system and prolonging congestion-driven delay, which, in turn, deters customers from placing orders. As a result, a hybrid order-ahead scheme may alarmingly achieve lower throughput than a pure order-onsite scheme.

To overcome this throughput shortfall, we consider a variety of mitigation strategies. The first one is to let restaurants optimally choose whether to share queue-length information with remote customers. We find that the throughput shortfall can persist despite this intervention. The second strategy is to allow remote customers who order ahead to cancel orders after they arrive at the service facility. While this strategy is promising in eliminating the throughput shortfall, it triggers a new problem as allowing cancellation in the hybrid order-ahead scheme reduces throughput when the market size is small. The third strategy is to reject new remote orders at the outset in the event of too many outstanding orders. If the rejection threshold can be optimally determined, then the hybrid order-ahead-with-rejection scheme will outperform both the one without rejection and the pure order-onsite scheme, but not necessarily the hybrid-order-ahead-with-cancellation scheme. Finally, we consider an integrated hybrid order-ahead mechanism that allows for both rejection and cancellation of remote orders and subsumes all the previously considered order-ahead schemes as special cases. We numerically find that overall, the rejection scheme has the smallest throughput gap from the integrated mechanism among all the schemes considered.

We conclude by discussing the caveats of our model and future research directions. First, a practical concern of order rejection or cancellation is the loss of goodwill, which might hurt future business. Second, our paper focuses on on-demand services (Taylor 2018) in which customers value instant gratification and prefer to have their requests fulfilled as soon as possible. That is why our model assumes that customers incur the same unit delay cost regardless of the nature of the delay (onsite or during travel), consistent with Hassin and Roet-Green (2021). Thus, in such an on-demand setting, customers have no incentive to postpone their travel because delay is equally costly regardless of where it occurs. In settings without such a salient on-demand feature, one may argue that waiting at home is less annoying and thus less costly than waiting at the service facility and that customers can have an incentive to postpone their travel after placing their order. Such strategic postponement prolongs total delay and can be left for future research.

References

Ata B, Peng X (2018) An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. *Operations Research* 66(1):163–183.

Baron O, Chen X, Li Y (2023) Omnichannel services: The false premise and operational remedies. *Management Science* 69(2):865–884.

Buell RW (2020) Breakfast at the Paramount. Harvard Business School Case 617-011, (Revised Jan. 2020).

Business Research Insights (2024) Online food ordering market size, share, growth, and industry analysis regional forecast by 2031. `https://www.businessresearchinsights.com/market-reports/online-food-ordering-market-102644`.

Chen H, Frank M (2004) Monopoly pricing when customers queue. *IIE Transactions* 36(6):569–581.

Chen M, Hu M, Wang J (2022) Food delivery service and restaurant: Friend or foe? *Management Science* 68(9):6539–6551.

Cui S, Wang Z, Yang L (2020) The economics of line-sitting. *Management Science* 66(1):227–242.

Dean G (2021) Former Starbucks workers say the chain's mobile ordering is out of control. Business Insider (Jun 26), URL https://www.businessinsider.com/starbucks-mobile-ordering-app-barista-pandemic-coffee-customers-online-digital-2021-6.

Farahani MH, Dawande M, Janakiraman G (2022) Order now, pickup in 30 minutes: Managing queues with static delivery guarantees. *Operations Research* 70(4):2013–2031.

Feldman P, Frazelle AE, Swinney R (2023) Managing relationships between restaurants and food delivery platforms: Conflict, contracts, and coordination. *Management Science* 69(2):812–823.

Gao F, Su X (2018) Omnichannel service operations with online and offline self-order technologies. *Management Science* 64(8):3595–3608.

Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* 54(5):1185–1195.

Hassin R (2016) *Rational Queueing* (Boca Raton: CRC Press, Taylor and Francis Group).

Hassin R, Haviv M (1995) Equilibrium strategies for queues with impatient customers. *Operations Research Letters* 17(1):41–45.

Hassin R, Haviv M (2003) *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59 (Springer Science & Business Media).

Hassin R, Roet-Green R (2021) On queue-length information when customers travel to a queue. *Manufacturing & Service Operations Management* 23(4):989–1004.

Haviv M, Ritov Y (2001) Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions. *Queueing Systems* 38:495–508.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(6):15–24.

Peet's Coffee (2019) Can I cancel my mobile order after it's been placed? URL https://faq.peets.com/hc/en-us/articles/360025049112-Can-I-cancel-my-mobile-order-after-it-s-been-placed-.

Pucci R (2017) Mobile order and pay ahead: A new sales channel for restaurants and merchants. Mercator Advisory Group (March), URL https://www.mercatoradvisorygroup.com/Press_Releases/Mobile_Order_and_Pay_Ahead__a_New_Sales_Channel__Increases_Volume_for_Restaurants_and_Merchants/.

Starbucks (2017) When will my mobile order be ready? URL https://customerservice.starbucks.com/app/answers/detail/a_id/3041.

Subway (2020) Subway®website ordering and mobile app terms of use—USA ONLY. URL https://www.subway.com/en-us/legal/order_apptermsofuse.

Taylor TA (2018) On-demand service platforms. *M&SOM* 20(4):704–720.

Electronic Companion

Appendix EC.1: More Details about the Extensions in §6

EC.1.1. Food-Quality Degradation

Let $t \geq 0$ be the time elapsed after an order is complete. If a customer arrives at the service facility and picks up her order at time t, her reward from the order is $Ve^{-d \cdot t}$, where $d > 0$ is a parameter that measures the magnitude of quality degradation. The pure-order-onsite scheme is not affected by d since, by definition, orders will be complete only after customers arrive. Next, we re-derive the equilibria for various order-ahead schemes. To begin with, local customers still follow the Naor threshold. For remote customers, the expected utility of ordering given initial queue length n is $U_d(n) = V\mathbb{P}(N_n > 0) + \mathbb{E}[Ve^{-d(T-X)}|T > X]\mathbb{P}(N_n = 0) - c\sum_{i=0}^{(n+1)} \frac{i}{\mu} \cdot p_n(i) - \frac{c}{\beta}$, where T is a random variable denoting travel time and X is a random variable denoting the steady-state sojourn time in the order queue. Note that in computing the expected service reward, quality degradation only arises in the event that the order is ready for pickup upon arrival at the service facility, i.e., $\{N_n = 0\}$, which is equivalent to saying that travel time exceeds sojourn time, i.e., $\{T > X\}$. Since T follows an exponential distribution with rate β, due to its memoryless property, $\mathbb{E}[Ve^{-d(T-X)}|T > X] = \mathbb{E}[Ve^{-dT}] = \frac{\beta}{\beta+d}V$. Therefore, $U_d(n) = \left(1 - \sigma^{n+1}\frac{d}{\beta+d}\right)V - \frac{c}{\beta}\left(\sigma^{n+1} + \frac{(n+1)\beta}{\mu}\right)$.

EC.1.1.1. Hybrid Order-Ahead **(1) Queue-length information not shared remotely.** Suppose all remote customers place an order with probability $q \in [0,1]$ and local customers follow the Naor-threshold. The steady-state probability of the number of outstanding orders being i, $\pi_i^u(q)$, is the same as (2). The unconditional expected utility for a remote customer who places an order is $U_{A,d}(q) = \sum_{n=0}^{\infty} U_d(n)\pi_n^u(q)$, which is decreasing in q. Thus, $q \in (0,1)$ is an equilibrium only if $U_{A,d}(q) = 0$, $q = 1$ is an equilibrium if $U_{A,d}(1) > 0$ and $q = 0$ is an equilibrium if $U_{A,d}(0) < 0$. Hence, remote customers' equilibrium order-placing probability is: $q_{A,d} = 0, \text{if } \Lambda \geq \bar{\lambda}_{A,d}; q_{A,d} = \hat{q}_{A,d} \in (0,1), \text{if } \underline{\lambda}_{A,d} < \Lambda < \bar{\lambda}_{A,d}; q_{A,d} = 1, \text{if } \Lambda \leq \underline{\lambda}_{A,d}$, where $\hat{q}_{A,d}$ uniquely solves the equation $U_{A,d}(\hat{q}_{A,d}) = 0$. The resulting system throughput $TH_{A,d}^u = \mu(1 - \pi_0^u(\hat{q}_{A,d}))$. **(2) Queue-length information shared remotely.** Remote customers follow a threshold joining strategy. The joining threshold of remote customers $\hat{n}_e \leq n_e^* \leq n_e$ is uniquely solved by $\hat{n}_e^* \equiv \min\{n \in \mathbb{N} : U_d(n) < 0\}$. Accordingly, the steady-state probability of the number of outstanding orders being i is $\pi_i^o = \rho^i \pi_0^o,\ i = 0, 1, \cdots, \hat{n}_e; \pi_i^o = ((1-\gamma)\rho)^{i-\hat{n}_e}\rho^{\hat{n}_e}\pi_0^o,\ i = \hat{n}_e + 1, \cdots, n_e$, where $\hat{\pi}_0^o = \left(\frac{1-\rho^{\hat{n}_e}}{1-\rho} + \frac{\rho^{\hat{n}_e}(1-((1-\gamma)\rho)^{n_e-\hat{n}_e+1})}{1-(1-\gamma)\rho}\right)^{-1}$. with $\rho \equiv \Lambda/\mu$. The resulting system throughput is $TH_{A,d}^o = \mu(1 - \hat{\pi}_0^o)$.

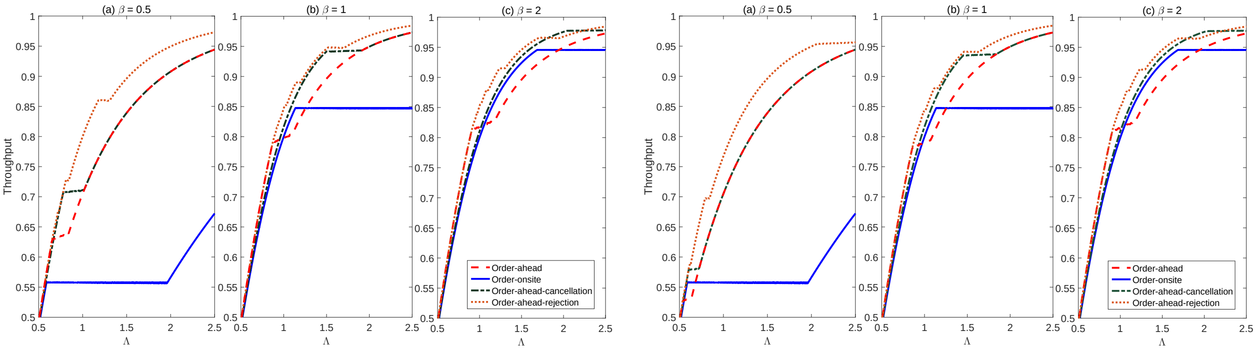
EC.1.1.2. Hybrid Order-Ahead with (Optimal) Rejection **(1) Queue-length information not shared remotely.** Suppose that the service provider accepts new remote orders if the number of outstanding orders is strictly less than threshold N and rejects any new remote orders otherwise. The unconditional expected utility from ordering is $U_{R,d}^N(q) = \sum_{n=0}^{N-1} U_d(n)\pi_{n,R}(q)$, The steady-state probability of the number of outstanding orders being i is given in (6) and (7). The remote customers' equilibrium order-placing probability $q_{R,d}^N$ is as follows: if $U_d(N-1) \geq 0$, $q_{R,d}^N = 1$; otherwise, $q_{R,d}^N = 0, \text{if } \Lambda \geq \bar{\lambda}_{R,d}^N; q_{R,d}^N = \hat{q}_{R,d}^N \in (0,1), \text{if } \underline{\lambda}_{R,d}^N < \Lambda < \bar{\lambda}_{R,d}^N; q_{R,d} = 1, \text{if } \Lambda \leq \underline{\lambda}_{R,d}^N$, where $\hat{q}_{R,d}^N$ uniquely solves the equation $U_{R,d}^N(\hat{q}) = 0$. The resulting system throughput is $TH_{R,d}^N = \mu[1 - \pi_{0,R}^u(q_{R,d}^N)]$.

(2) Queue-length information shared remotely. Suppose that the rejection threshold is N. Thus, a remote order effectively joins the queue according to a threshold $\min\{N, \hat{n}_e\}$. The steady-state probability of the number of outstanding orders being i is $\widetilde{\pi}_i^o = \rho^i \widetilde{\pi}_0^o,\ i = 0, 1, \cdots, x; \widetilde{\pi}_i^o = ((1-\gamma)\rho)^{i-x}\rho^x\widetilde{\pi}_0^o,\ i = x+1, \cdots, n_e$, where $x = \min\{N, \hat{n}_e\}$ and $\widetilde{\pi}_0^o = \left(\frac{1-\rho^x}{1-\rho} + \frac{\rho^x(1-((1-\gamma)\rho)^{n_e-x+1})}{1-(1-\gamma)\rho}\right)^{-1}$. The resulting system throughput is $TH_{R,d}^o = \mu(1 - \widetilde{\pi}_0^o)$.

EC.1.1.3. Hybrid Order-Ahead with Cancellation **(1) Queue-length information not shared remotely.** If a remote customer places an order when the queue length is n, then her expected utility is given by $U_{C,d}(n) = V\mathbb{P}(0 < N_n^C \le n_e) + \mathbb{E}[Ve^{-d(T-X)}|T > X]\mathbb{P}(N_n^C = 0) - cw_d^C(n) - \frac{c}{\beta}$, where $\mathbb{P}(0 < N_n^C \le n_e) = \begin{cases} 1 - p_n^C(0) = 1 - \prod_{k=0}^{n} \frac{\mu_k}{\mu_k+\beta}, & \text{if } n < n_e, \\ \sum_{i=1}^{n_e} p_n^C(i) = \sum_{i=1}^{n_e} \frac{\beta}{\mu_{i-1}+\beta} \prod_{k=i}^{n} \frac{\mu_k}{\mu_k+\beta}, & \text{otherwise,} \end{cases}$ where $\mu_n = \mu + (n - n_e)^+\beta$. The expected onsite delay is $w_{C,d}(n) \equiv \mathbb{E}[W(n)] = \sum_{i=0}^{(n+1)\wedge n_e} \mathbb{E}[W(n)|N_n = i] \cdot p_n^C(i) = \sum_{i=0}^{(n+1)\wedge n_e} \frac{i}{\mu} \cdot p_n^C(i) = \begin{cases} \frac{1}{\beta}\left(\sigma^{n+1} + \frac{(n+1)\beta}{\mu} - 1\right), & \text{if } n < n_e; \\ \frac{\beta}{\mu+\beta} \sum_{i=1}^{n_e} \frac{i}{\mu} \prod_{k=i}^{n} \frac{\mu_k}{\mu_k+\beta}, & \text{otherwise.} \end{cases}$ Therefore, the expected utility of a remote customer who places an order is $U_{C,d}(n) = \bar{U}_{C,d}(n)\mathbf{1}_{\{n<n_e\}} + \widetilde{U}_{C,d}(n)\mathbf{1}_{\{n\ge n_e\}}$, where $\mathbf{1}_A$ is the indicator of event A, and the two functions $\bar{U}_{C,d}(n)$ and $\widetilde{U}_{C,d}(n)$ are given by $\bar{U}_{C,d}(n) \equiv V\left(1 - \frac{d}{d+\beta}\prod_{k=0}^{n} \frac{\mu_k}{\mu_k+\beta}\right) - \frac{c}{\beta}\left(\sigma^{n+1} + \frac{(n+1)\beta}{\mu}\right), \widetilde{U}_{C,d}(n) \equiv V\frac{\beta}{\mu+\beta}\sum_{j=1}^{n_e}\prod_{k=j}^{n} \frac{\mu_k}{\mu_k+\beta} + V\frac{\beta}{\beta+d}\prod_{k=0}^{n}\frac{\mu_k}{\mu_k+\beta} - c\frac{\beta}{\mu+\beta}\sum_{j=1}^{n_e}\frac{j}{\mu}\prod_{k=i}^{n}\frac{\mu_k}{\mu_k+\beta} - \frac{c}{\beta}$. Given that all other remote customers place orders with probability q, the expected utility of a tagged remote customer who places an order is $U_{C,d}(q) = \sum_{i=0}^{n_e-1} \bar{U}_{C,d}(i)\pi_{i,C}^u(q) + \sum_{i=n_e}^{\infty} \widetilde{U}_{C,d}(i)\pi_{i,C}^u(q) = \sum_{i=0}^{n_e-1}\left[V\left(1 - \frac{d}{d+\beta}\prod_{k=0}^{i}\frac{\mu_k}{\mu_k+\beta}\right) - \frac{c}{\beta}\left(\sigma^{i+1} + \frac{(i+1)\beta}{\mu} - 1\right)\right]\pi_{i,C}^u(q) - \frac{c}{\beta} \quad + \sum_{i=n_e}^{\infty}\left[V\left(\frac{\beta}{\beta+d}\prod_{k=0}^{i}\frac{\mu_k}{\mu_k+\beta} + \frac{\beta}{\mu+\beta}\sum_{j=1}^{n_e}\prod_{k=j}^{i}\frac{\mu_k}{\mu_k+\beta}\right) - c\frac{\beta}{\mu+\beta}\sum_{j=1}^{n_e}\frac{j}{\mu}\prod_{k=j}^{i}\frac{\mu_k}{\mu_k+\beta}\right]\pi_{i,C}^u(q)$, where $\mathrm{t}\pi_{i,C}^u(q)$ is given by (4). Similar to what we show in the proof of Proposition 4, we can show that $U_{C,d}(q)$ is decreasing in q. Hence, remote customers' equilibrium order-placing probability is $q_{C,d} = \begin{cases} 0, & \text{if } \Lambda \ge \bar{\lambda}_{C,d}, \\ \hat{q}_{C,d} \in (0,1), & \text{if } \underline{\lambda}_{C,d} < \Lambda < \bar{\lambda}_{C,d}, \\ 1, & \text{if } \Lambda \le \underline{\lambda}_{C,d}, \end{cases}$ where $\hat{q}_{C,d}$ uniquely solves the equation $U_{C,d}(\hat{q}_{C,d}) = 0$. The resulting system throughput is $TH_{C,d} = \mu[1 - \pi_{0,C}^u(q_{C,d})]$.

(2) Queue-length information shared remotely. Customers follow the threshold $\hat{n}_e \le n_e^*$. The resulting system throughput is the same as in the case without cancellation.

Figure EC.1 Throughput Comparison with Food Quality Degradation



(a) $d = 0.3$

(b) $d = 0.5$

Note. $\mu = 1$, $V = 2$, $c = 0.5$, $\gamma = 0.7$.

EC.1.1.4. Throughput Comparison We replicate the four-way throughput comparison in Figure 9 by incorporating the effect of food-quality degradation. The results are presented in Figure EC.1. We observe that when travel is fast, the system throughput of any of three order-ahead schemes is hardly affected by incorporating quality degradation. When travel is fast, quality degradation rarely occurs because customers are unlikely to arrive at the service facility after their order is complete. However, when travel is slow, quality degradation is more likely to arise and quality degradation clearly takes its toll on all three order-ahead schemes when travel is slow. In particular, when quality degradation becomes more intense, as shown in Figure EC.1-(b)-(a), the hybrid order-ahead scheme can fall short of the pure order-onsite scheme. This issue can be addressed by strategic idleness as suggested by Farahani et al. (2022) and geo-location technology that alerts the kitchen when customers are getting close so it does not start preparing orders too early. It is worth emphasizing that these supply-side interventions to mitigate quality degradation cannot address the throughput shortfall of the hybrid order-ahead scheme when travel is fast because the throughput shortfall arises even in the absence of quality degradation.

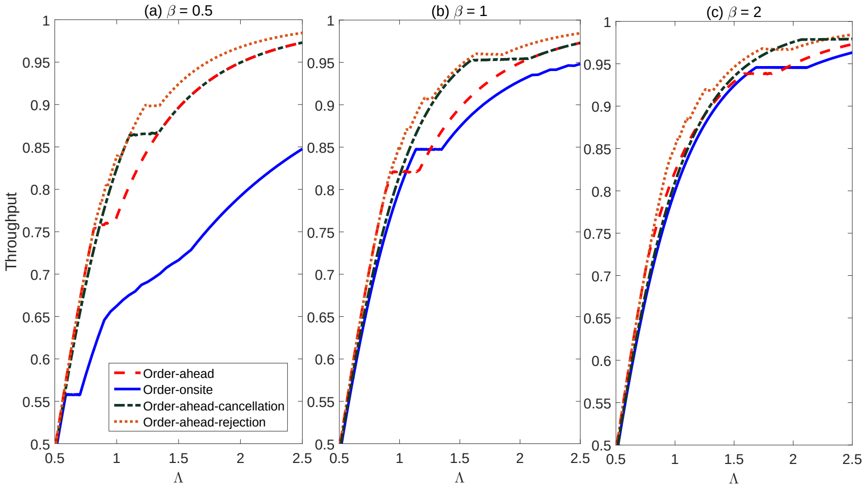
EC.1.2. Channel Choice

As argued in 6.2, modeling the channel choice will not affect the pure order-onsite scheme or the hybrid order-ahead-with-cancellation scheme. It will not affect the hybrid order-ahead scheme (with or without rejection) when queue-length information is shared remotely. The only cases in which modeling the channel choice will make a difference are the hybrid order-ahead scheme and that with rejection when queue-length information is not shared remotely. We rederive the equilibria below for these two cases.

EC.1.2.1. Hybrid Order-Ahead All customers who order onsite follow a threshold strategy n_e. We suppose that remote customers order ahead with probability $q_A \in [0,1]$ and order onsite with probability $q_S \in [0,1]$, and do not order with probability $1-q_A-q_S$, where $q_A+q_S \leq 1$. The corresponding steady-state probability of the number of outstanding orders being i is $\pi_i^u(q_A,q_S) = \begin{cases} \rho_{T1}^i \pi_0^u(q_A,q_S), & i<n_e, \\ \rho_{R1}^{i-n_e} \rho_{T1}^{n_e} \pi_0^u(q_A,q_S), & i \geq n_e, \end{cases}$ for $\rho_{R1}<1$, and $\rho_{T1} = \frac{\gamma\Lambda(q_A+q_S)+(1-\gamma)\Lambda}{\mu}$ and $\rho_{R1} = \frac{\gamma\Lambda q_A}{\mu}$, and $\pi_0^u(q_A,q_S) = \left(\frac{1-\rho_{T1}^{n_e}}{1-\rho_{T1}} + \frac{\rho_{T1}^{n_e}}{1-\rho_{R1}}\right)^{-1}$. The expected utility for a remote customer who orders ahead is $U_A^R(q_A,q_S) = \sum_{n=0}^{\infty} \bar{U}(n)\pi_n^u(q_A,q_S)$. The expected utility for a remote customer who orders onsite is $U_S^R(q_A,q_S) = \sum_{n=0}^{n_e-1}\left(V - \frac{(n+1)c}{\mu}\right)\pi_n^u(q_A,q_S) - \frac{c}{\beta}$. Thus, remote customers' equilibrium order-placing probability is: $(q_A^e,q_S^e) = \begin{cases} (q_A^R,0), & \text{if } U_A^R(q_A^R,0) \geq 0 \text{ and } U_A^R(q_A^R,0) > U_S^R(q_A^R,0), \\ (0,q_S^R), & \text{if } U_S^R(0,q_S^R) \geq 0 \text{ and } U_S^R(0,q_S^R) > U_A^R(0,q_S^R), \\ (q_{A1}^R, 1-q_{A1}^R), & \text{if } U_A^R(q_{A1}^R,1-q_{A1}^R) = U_S^R(q_{A1}^R,1-q_{A1}^R) \geq 0, \\ (q_{A2}^R,q_{S2}^R), & \text{if } U_A^R(q_{A2}^R,q_{S2}^R) = U_S^R(q_{A2}^R,q_{S2}^R) = 0. \end{cases}$ When $q_A^R \in (0,1)$, it is solved by $U_A^R(q_A^R,0)=0$; when $q_S^R \in (0,1)$, it is solved by $U_S^R(0,q_S^R)=0$. The resulting system throughput is $TH^u = \mu[1-\pi_0^u(q_A^e,q_S^e)]$.

EC.1.2.2. Hybrid Order-Ahead with (Optimal) Rejection Given rejection threshold N and the order-placing probabilities (q_A,q_S) of remote customers, we first give the steady-state probability of the number of outstanding orders being i. If $N>n_e$ the steady-state probability of the number of outstanding orders being i is $\pi_{i,R}^u(q_A,q_S) = \begin{cases} \rho_{T1}^i \pi_{0,R}^u(q_A,q_S), & i<n_e, \\ \rho_{R1}^{i-n_e}\rho_{T1}^{n_e}\pi_{0,R}^u(q_A,q_S), & i=n_e,\cdots,N, \end{cases}$ where $\pi_{0,R}^u(q_A,q_S) = \left(\frac{1-\rho_{T1}^{n_e}}{1-\rho_{T1}} + \frac{\rho_{T1}^{n_e}(1-\rho_{R1}^{N-n_e+1})}{1-\rho_{R1}}\right)^{-1}$. If $N \leq n_e$, the steady-state probability of the number of outstanding orders being i is $\pi_{i,R}^u(q_A,q_S) = \begin{cases} \rho_{T1}^i \pi_{0,R}^u(q_A,q_S), & i<N, \\ \rho_L^{i-N}\rho_{T1}^N \pi_{0,R}^u(q_A,q_S), & i=N,\cdots,n_e, \end{cases}$ where $\pi_{0,R}^u(q) = \left(\frac{1-\rho_{T1}^N}{1-\rho_{T1}} + \frac{\rho_{T1}^N(1-\rho_{L1}^{n_e-N+1})}{1-\rho_{L1}}\right)^{-1}$, and $\rho_{T1} = \frac{\gamma\Lambda(q_A+q_S)+(1-\gamma)\Lambda}{\mu}$ and $\rho_{R1} = \frac{\gamma\Lambda q_A}{\mu}$, and $\rho_{L1} = \frac{\gamma\Lambda q_S+(1-\gamma)\Lambda}{\mu}$. Thus, for a remote customer who places an order ahead, with probability $\pi_{N,R}^u(q_A,q_S)$, her order will be rejected (from which she gets zero utility); with probability $1-\pi_{N,R}^u(q_A,q_S)$, her order will be accepted (which implies the queue length at the moment is less than N). Thus, her unconditional expected utility from ordering ahead is $U_{R,N}^u(q_A,q_S) = \sum_{n=0}^{N-1}\bar{U}(n)\pi_{n,R}^u(q_A,q_S)$. For a remote customer who places an onsite order, her unconditional expected utility from ordering is $U_{S,R}^R(q_A,q_S) = \sum_{n=0}^{n_e-1}\left(V-\frac{(n+1)c}{\mu}\right)\pi_{n,R}^u(q_A,q_S) - \frac{c}{\beta}$. Thus, remote customers' equilibrium order-placing probability is: $(q_A^e,q_S^e) = \begin{cases} (q_A^R,0), & \text{if } U_{A,R}^R(q_A^R,0) \geq 0 \text{ and } U_{A,R}^R(q_A^R,0) > U_{S,R}^R(q_A^R,0), \\ (0,q_S^R), & \text{if } U_{S,R}^R(0,q_S^R) \geq 0 \text{ and } U_S^R(0,q_S^R) > U_{A,R}^R(0,q_S^R), \\ (q_A^R,1-q_A^R), & \text{if } U_{A,R}^R(q_A^R,1-q_A^R) = U_{S,R}^R(q_A^R,1-q_A^R) \geq 0, \\ (q_A^R,q_S^R), & \text{if } U_{A,R}^R(q_A^R,q_S^R) = U_{S,R}^R(q_A^R,q_S^R) = 0. \end{cases}$
When $q_A^R \in (0,1)$, it is solved by $U_{A,R}^R(q_A^R,0)=0$; when $q_S^R \in (0,1)$, it is solved by $U_{S,R}^R(0,q_A^R)=0$. The resulting system throughput is $TH^u = \mu[1-\pi_0^u(q_A^e,q_S^e)]$.

Figure EC.2 **Throughput Comparison with Channel Choice**



Note. $\mu = 1$, $V = 2$, $c = 0.5$, $\gamma = 0.7$.

EC.1.2.3. Throughput Comparison We replicate the four-way throughput comparison in Figure 9 by incorporating the channel choice. The result is presented in Figure EC.2. We observe that Figure EC.2 is qualitatively similar to Figure 9, suggesting that the insights from our base model carry over. When travel is slow, incorporating the channel choice barely affects the equilibrium outcome, as judged by the similarly between Figure EC.2-(a) and Figure 9-(a). In such a case, ordering ahead has such a substantial advantage that remote customers will not forgo it. When travel is fast, some remote customers switch to ordering onsite, making the hybrid order-ahead scheme more similar to the pure-order-onsite scheme, as illustrated by Figure EC.2-(c). Still, we find that providing the order-ahead option can result in lower throughput, but such shortfall can be addressed through cancellation or rejection.

EC.1.3. Heterogeneous Travel Speed of Remote Customers

Denote the cumulative distribution function of β by F. In the hybrid order-ahead scheme, when queue-length information is not shared remotely, we represent remote customers' equilibrium by (λ_A, λ_S), where λ_A and λ_S denote the arrival rates of remote customers who choose to order ahead and order onsite, respectively, By definition, $\lambda_A + \lambda_S \leq \gamma\Lambda$. Further, the arrival rate of local customers is $\lambda_L = (1-\gamma)\Lambda$. Given the triplet $\boldsymbol{\lambda} \equiv (\lambda_A, \lambda_S, \lambda_L)$, let $\hat{\rho}_T = (\lambda_A + \lambda_S + \lambda_L)/\mu$ and $\hat{\rho}_R = \lambda_A/\mu$; for $\hat{\rho}_R < 1$, the corresponding steady-state probability of the number of outstanding orders being i is: $\hat{\pi}_0^u(\boldsymbol{\lambda}) = \left(\frac{1-\hat{\rho}_T^{n_e}}{1-\hat{\rho}_T} + \frac{\hat{\rho}_T^{n_e}}{1-\hat{\rho}_R}\right)^{-1}$; $\hat{\pi}_i^u(\boldsymbol{\lambda}) = \hat{\rho}_T^i \hat{\pi}_0^u(\boldsymbol{\lambda}), i < n_e$; $\hat{\pi}_i^u(\boldsymbol{\lambda}) = \hat{\rho}_R^{i-n_e} \hat{\rho}_T^{n_e} \hat{\pi}_0^u(\boldsymbol{\lambda}), i \geq n_e$. For a remote customer with travel speed β, let $U_A(\boldsymbol{\lambda}, \beta)$ and $U_S(\boldsymbol{\lambda}, \beta)$ be her expected utility of ordering ahead and that of ordering onsite, respectively. Thus, $U_A(\boldsymbol{\lambda}, \beta) = \sum_{n=0}^{\infty} \bar{U}(n) \hat{\pi}_n^u(\boldsymbol{\lambda})$ and $U_S(\boldsymbol{\lambda}, \beta) = \sum_{n=0}^{n_e-1} \left(V - \frac{(n+1)c}{\mu}\right) \hat{\pi}_n^u(\boldsymbol{\lambda}) - \frac{c}{\beta}$. Therefore, the equilibrium (λ_A, λ_S) solves the following set of fixed-point equations: $\lambda_A = \Lambda\gamma \int_a^b \mathbf{1}_{\{U_A(\boldsymbol{\lambda},\beta) > [U_S(\boldsymbol{\lambda},\beta)]^+\}} dF(\beta)$; $\lambda_S = \Lambda\gamma \int_a^b \mathbf{1}_{\{U_S(\boldsymbol{\lambda},\beta) > [U_A(\boldsymbol{\lambda},\beta)]^+\}} dF(\beta)$. The equilibrium for the hybrid order-ahead-with-rejection scheme and that for the pure order-onsite scheme can be similarly defined.

Appendix EC.2: Proofs

We first give the following two technical lemmas that will be repeatedly used in the subsequent proofs.

LEMMA EC.2.1. *For a strictly decreasing function $f(n)$ and two probability distributions $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$, supported over $\{0, 1, \cdots, \bar{n}\}$, where $\bar{n}$ can possibly be ∞, if $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$ cross each other only once (there exists an n^* such that $\pi_n^1 \geq \pi_n^2$ when $n \leq n^*$ and $\pi_n^1 < \pi_n^2$ when $n > n^*$), then for two random variables X_1 and X_2 following $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$, we have $\mathbb{E}[f(X_1)] = \sum_{n=0}^{\bar{n}} f(n)\pi_n^1 > \sum_{n=0}^{\bar{n}} f(n)\pi_n^2 = \mathbb{E}[f(X_2)]$.*

Proof of Lemma EC.2.1 We write $\sum_{n=0}^{\bar{n}} f(n)\pi_n^1 - \sum_{n=0}^{\bar{n}} f(n)\pi_n^2 = \sum_{n=0}^{n^*} f(n)(\pi_n^1 - \pi_n^2) + \sum_{n=n^*+1}^{\bar{n}} f(n)(\pi_n^1 - \pi_n^2) > f(n^*)\left(\sum_{n=0}^{n^*}(\pi_n^1 - \pi_n^2) + \sum_{n=n^*+1}^{\bar{n}}(\pi_n^1 - \pi_n^2)\right) = 0$, where the inequality holds due to the single-crossing property, and the last equality holds because both $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$ are well defined probability distributions over $\{0, 1, \cdots, \bar{n}\}$, i.e., $0 = 1 - 1 = \sum_{n=0}^{\bar{n}}(\pi_n^1 - \pi_n^2) = \sum_{n=0}^{n^*}(\pi_n^1 - \pi_n^2) + \sum_{n=n^*+1}^{\bar{n}}(\pi_n^1 - \pi_n^2)$. □

Lemma EC.2.2. *Consider a birth-and-death process with birth rate λ_{i-1} and death rate μ_i for state i and denote $\rho_i = \lambda_{i-1}/\mu_i$. We consider two systems indexed by (1) and (2) with stationary distributions $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$, respectively. If $\rho_i^{(1)} \geq \rho_i^{(2)}$ for all $i = 1, \cdots$, we must have $\pi_0^{(1)} \leq \pi_0^{(2)}$, and the queue length of the first system is stochastically larger than that of the second system, i.e., $Q^{(1)} \geq_{st} Q^{(2)}$.*

Proof of Lemma EC.2.2 In a birth-and-death system, $\pi_0 = \frac{1}{1+\rho_1+\rho_1\rho_2+\cdots+\prod_{i=1}^n \rho_i+\cdots}$. Therefore, a larger ρ_i induces a smaller π_0. Hence, $\pi_0^{(1)} \leq \pi_0^{(2)}$. The steady-state probability of the number of customers being i in the two systems are $\pi_i^{(1)} = \rho_1^{(1)} \cdots \rho_i^{(1)} \pi_0^{(1)}$, and $\pi_i^{(2)} = \rho_1^{(2)} \cdots \rho_i^{(2)} \pi_0^{(2)}$. Thus, $\pi_i^{(1)}/\pi_i^{(2)}$ is increasing in i. Hence, $Q^{(1)} \geq_{st} Q^{(2)}$ in the likelihood ratio order. □

Proof of Lemma 1 Consider a remote customer with n existing orders upon her arrival. Let $T \sim \text{Exp}(\beta)$ be her travel time, and let I.I.D. $\text{Exp}(\mu)$ r.v.'s $S_1, \cdots, S_n$ denote the service times for the n outstanding orders, with S_i corresponding to the i^{th} order to be processed in the order queue. Let S_0 be the service time of the tagged customer. We next derive the distribution of N_n, the tagged remote customer' updated queue position (including herself) when she arrives at the service facility. Note that $N_n \in \{0, 1, \cdots, n+1\}$. We denote the probabilities by $p_n(0), \ldots, p_n(n+1)$. It is straightforward to see that the updated queue position is i if and only if there are exactly $n - i + 1$ service completions when the tagged customer arrives at the service facility; this corresponds to the event $\{S_1 + \cdots + S_{n-i+1} < T < S_1 + \cdots + S_{n-i+1} + S_{n-i+2}\}$. Let $\sigma \equiv \mu/(\mu + \beta)$. By the memoryless property of the exponential distribution, we have $p_n(n+1) \equiv \mathbb{P}(N_n = n+1) = \mathbb{P}(T < S_1) = 1 - \sigma; p_n(i) \equiv \mathbb{P}(N_n = i) = \mathbb{P}(T > S_1) \times \cdots \times \mathbb{P}(T > S_{n-i+1}) \times \mathbb{P}(T < S_{n-i+2}) = (1-\sigma)\sigma^{n-i+1}, i = 1, 2, \cdots, n; p_n(0) \equiv \mathbb{P}(N_n = 0) = \mathbb{P}(T > S_1) \times \cdots \times \mathbb{P}(T > S_n) \times \mathbb{P}(T > S_0) = \sigma^{n+1}$. □

Proof of Proposition 1 The proof of Proposition 1 will be based on Lemmas EC.2.1 and EC.2.2. Recall that when the queue-length information is not shared remotely, the expected utility for a remote customer who places an order is $U^u(q) = \sum_{n=0}^{\infty} \bar{U}(n)\pi_n^u(q)$. The next lemma establishes properties of $U_\rho^u(q)$.

Lemma EC.2.3 **(Property of U^u function)**. *The utility function U^u has the following properties: (i)$U_\infty^u(1) = -\infty$, $U_0^u(1) > 0$, $U_\infty^u(0) = \bar{U}(n_e) < 0$ and $U_0^u(0) > 0$; (ii) $U_\rho^u(q)$ is continuous and strictly decreasing in q for a fixed ρ; (iii) $U_\rho^u(1)$ and $U_\rho^u(0)$ are continuous and strictly decreasing in ρ.*

Proof of Lemma EC.2.3 To prove Part (i), first, we have $U_0^u(1) = U_0^u(0) = \bar{U}(0) > 0$ by Assumption 1. When the system load goes to infinity and if all remote customers place orders (i.e., $q = 1$), the birth-death process becomes unstable so that $U_\infty^u(1) = -\infty$. When the system load goes to infinity and if no other remote customers join (i.e., $q = 0$), the steady-state probability $\pi_{n_e}^u = 1$ because there are infinite local customers keeping the queue size at its capacity n_e. Hence, we have $U_\infty^u(0) = \bar{U}(n_e) < 0$. For Part (ii), continuity is obvious. To prove monotonicity, we pick $q_1 < q_2$ and consider the two corresponding steady state distributions $\{\pi_n^u(q_1)\}$ and $\{\pi_n^u(q_2)\}$, we have that $\pi_0^u(q_1) > \pi_0^u(q_2)$ since the system 2 is busier than system 1 according to Lemma EC.2.2. In addition, there must exist some integers n' such that $\pi_{n'}^u(q_1) < \pi_{n'}^u(q_2)$. Otherwise, we cannot have $\sum_{n=0}^{\infty} \pi_n^u(q_1) = \sum_{n=0}^{\infty} \pi_n^u(q_2) = 1$. We then claim that for any n', if $\pi_{n'}^u(q_1) < \pi_{n'}^u(q_2)$, then $\pi_n^u(q_1) < \pi_n^u(q_2)$ for $n \geq n' + 1$ due to the geometric structure of $\boldsymbol{\pi}$ distribution. Then it is straightforward to see these two distributions satisfy Condition (ii) of Lemma EC.2.1. And because $\bar{U}(n)$ is decreasing in n, it follows from Lemma EC.2.1 that $U_\rho^u(q_1) > U_\rho^u(q_2)$. Hence, $U_\rho^u(q)$ is decreasing in q. The proof of (iii) is similar to that of (ii). □

Finishing the proof of Proposition 1. From Parts (i) and (iii) of Lemma EC.2.3 and $U_0^u(1) = \bar{U}(0) > 0$ by Assumption 1, there must be a unique solution $\underline{\rho}_A^u$ to equation $U_\rho^u(1) = 0$ for $\rho \in (0,1)$. Similarly, there must be a unique solution $\bar{\rho}_A^u$ to equation $U_\rho^u(0) = 0$. Denote $\underline{\lambda}_A^u = \underline{\rho}_A^u \mu$ and $\bar{\lambda}_A^u = \bar{\rho}_A^u \mu$. By Lemma EC.2.3, $U^u(1) > 0$ when $\Lambda < \underline{\lambda}_A^u$, which implies that $q_A^u = 1$. On the other hand, when $\Lambda \geq \bar{\lambda}_A^u$, $q_A^u = 0$. Otherwise, q_A^u must satisfy $U^u(q) = 0$. □

Proof of Theorem 1 We first consider the small β case. When $\beta = \beta_0 \equiv c/(V - c/\mu)$, the cost of travel and undergoing a single service time becomes too high so no remote customer will join in the order-onsite model. Hence, the order-onsite model reduces to a standard $M/M/1/n_e$ model with arrival rate $(1-\gamma)\Lambda$ and service rate μ. On the other hand, the parallelization effect achieves some waiting time reduction so some remote customers may still join the system which makes the order-ahead model stochastically busier than the order-onsite model (to see this we again invoke Lemma EC.2.2). Hence, the order-ahead model yields higher throughput than the order-onsite model. We next consider the large β case. We define $\Delta\lambda \equiv \underline{\lambda}_S^u - \bar{\lambda}_A^u$, where $\underline{\lambda}_S^u$ is defined in the order-onsite model and $\bar{\lambda}_A^u$ is defined in (3). Our strategy is to study the asymptotic behavior of $\underline{\lambda}_S^u$ and $\bar{\lambda}_A^u$ when β is sufficiently large. We will show that, when β grows large, $\underline{\lambda}_S^u$ increases without bound whereas $\bar{\lambda}_A^u$ does not (it approaches a finite number). Then, we will have an interval $[\bar{\lambda}_A^u, \underline{\lambda}_S^u]$ such that $q_A^u(\Lambda) = 0$ and $q_S^u(\Lambda) = 1$ for all $\Lambda \in [\bar{\lambda}_A^u, \underline{\lambda}_S^u]$, as long as β is sufficiently large. This result will ensure that the order-onsite Continuous Time Markov Chain (CTMC) is stochastically busier than the order-ahead CTMC when $\Lambda \in [\bar{\lambda}_A^u, \underline{\lambda}_S^u]$. To see this, note that the two models have an equal death rate but the former has a strictly larger birth rate than the latter. Hence, invoking Lemma EC.2.2, we must have $\pi_0^S(\Lambda) < \pi_0^A(\Lambda)$, so that the order-onsite model yields strictly higher throughput for all $\Lambda \in [\bar{\lambda}_A^u, \underline{\lambda}_S^u]$. For $\bar{\lambda}_A^u$, we let $\beta \to \infty$, so that a remote customer's utility in the order-ahead model with $q = 0$ is $U^u(0) = \sum_{i=0}^{n_e} \left(V - \frac{(i+1)c}{\mu} - \frac{c}{\beta}\sigma^{i+1}\right) \pi_i^u(0) \to \sum_{i=0}^{n_e} \left(V - \frac{(i+1)c}{\mu}\right) \pi_i^u(0) = \sum_{i=0}^{n_e-1} \underbrace{\left(V - \frac{(i+1)c}{\mu}\right)}_{\geq 0} \pi_i^u(0) + \underbrace{\left(V - \frac{(n_e+1)c}{\mu}\right)}_{<0} \pi_{n_e}^u(0)$, where $\pi_i^u(0) = \frac{\rho_L^i(1-\rho_L)}{1-\rho_L^{n_e+1}}$, $i = 0, 1, \ldots, n_e$, and $\rho_L = (1-\gamma)\Lambda/\mu$. As Λ increases, $\pi_{n_e}^u(0)$ will have a bigger weight so the negative term in $U^u(0)$ will dominate the positive terms, so that $U^u(0)$ will become negative when $\Lambda > \bar{\Lambda}$ for some finite $\bar{\Lambda}$. Recall that $\bar{\lambda}_A^u$ is the root of $U^u(0) = 0$, $\bar{\lambda}_A^u$ must be a finite number as β grows large. On the other hand, recall that the utility function of remote customers in the order-onsite model is $U_S^u(q) = \sum_{i=0}^{n_e-1} \left(V - \frac{(i+1)c}{\mu}\right) \pi_{i,S}^u(q) - \frac{c}{\beta}$. When β is sufficiently large, $U_S^u(q) > 0$ for any $q \in [0,1]$ (because $V - (i+1)c/\mu > 0$ for all $i = 0, \ldots, n_e - 1$). So we must have $\underline{\lambda}_S^u = \infty$ and $\Delta\lambda = -\infty$. Hence, $\Delta\lambda$ is sufficiently negative when β is large. Because $\Delta\lambda$ is continuous in β, there exists a $\bar{\beta}$ such that $\Delta\lambda < 0$ for all $\beta > \bar{\beta}$. By the definition of $\underline{\lambda}_S^u$ and $\bar{\lambda}_A^u$, we now have $q_A^u(\Lambda) = 0$ and $q_S^u(\Lambda) = 1$ for all $\Lambda \in [\bar{\lambda}_A^u, \underline{\lambda}_S^u]$, as long as $\beta > \bar{\beta}$. Thus, Lemma EC.2.2 implies that the order-ahead scheme has lower throughput than the order-onsite model. □

Proof of Proposition 2 Recall that the expected utility of a remote customer who sees a queue length n and places an order is $\bar{U}(n) = V - \frac{c}{\beta}\left(\sigma^{n+1} + \frac{(n+1)\beta}{\mu}\right)$. We leverage the following properties of $\bar{U}(n)$: (i) $\lim_{n\to\infty} \bar{U}(n) = -\infty$; (ii) $\bar{U}(n)$ is strictly decreasing in n since $\bar{U}(n) - \bar{U}(n-1) = c(\sigma^{n+1} - 1)/\mu < 0$. Since $\bar{U}(0) > 0$ (Assumption 1), the equilibrium strategy must be of a threshold type, given by $n_e^* \equiv \min\{n \geq 0 : \bar{U}(n) < 0\}$, and the threshold n_e^* is at most n_e. Since $\bar{U}(n_e) = V - \frac{c}{\beta}\left(\sigma^{n_e+1} + \frac{(n_e+1)\beta}{\mu}\right) = V - \frac{c(n_e+1)}{\mu} - \frac{c}{\beta}\sigma^{n_e+1} < 0$, and $\bar{U}(n)$ is decreasing in n, the joining threshold n_e^* must be attained in $\{0, 1, \ldots, n_e\}$, i.e., $n_e^* = \min\{n \geq 0 : \bar{U}(n) < 0\}$. When $\lfloor \frac{\mu V}{c} \rfloor = \frac{\mu V}{c}$, for any finite β, $\bar{U}(n_e - 1) = -\frac{c}{\beta}\sigma^{n_e} < 0$, so that the joining threshold $n_e^* \leq n_e - 1 < n_e$. Otherwise, from Equation (1), $\bar{U}(n)$ decreases in n and increases in β. Note that $n_e^* = n_e$ if and only if $\bar{U}(n_e - 1) = V - \frac{cn_e}{\mu} - \frac{c}{\beta}\sigma^{n_e} \geq 0$, which requires that $\beta \geq \underline{\beta}$, where $\underline{\beta}$ is given by $V - \frac{cn_e}{\mu} - \frac{c}{\underline{\beta}}\sigma^{n_e} = 0$. □

Proof of Proposition 3 Given any Λ, there exists a unique order-placing probability $\widetilde{q} \in (0,1)$, which induces the same throughput regardless of whether queue-length information is shared directly or not remotely, i.e.,

$TH_A^u(\widetilde{q}) = TH_A^o$. It holds since $TH_A^u(0) < TH_A^o < TH_A^u(1)$ by Lemma EC.2.3 and the throughput when queue-length information is not shared remotely increases in the order-placing probability q given the market size. Recall from Proposition 1 that the equilibrium order-placing probability of remote customers under the hybrid order-ahead scheme decreases in the market size, and $q_A^u = 1$ when the market size $\Lambda < \underline{\lambda}_A^u$. Hence, $\widetilde{q} < q_A^u = 1$ when $\Lambda < \underline{\lambda}_A^u$. Pick two market sizes Λ_1, Λ_2 $(\underline{\lambda}_A^u < \Lambda_1 < \Lambda_2)$, there exist two order-placing probabilities $\widetilde{q}_1(\Lambda_1), \widetilde{q}_2(\Lambda_2)$ that satisfy $TH_A^u(\widetilde{q}_1(\Lambda_1)) = TH_A^o(\Lambda_1)$ and $TH_A^u(\widetilde{q}_2(\Lambda_2)) = TH_A^o(\Lambda_2)$. We next prove $U^u(\widetilde{q}_1(\Lambda_1)) - U^u(\widetilde{q}_2(\Lambda_2)) = \sum_{i=0}^{\infty} \bar{U}(n)(\pi_i^u(\widetilde{q}_1(\Lambda_1)) - \pi_i^u(\widetilde{q}_2(\Lambda_2)) > 0$. Since $TH_A^u(\widetilde{q}_1(\Lambda_1)) = TH_A^o(\Lambda_1)$ and $TH_A^u(\widetilde{q}_2(\Lambda_2)) = TH_A^o(\Lambda_2)$, we also have $\pi_0^u(\widetilde{q}_1(\Lambda_1)) = \pi_0^o(\Lambda_1) > \pi_0^o(\Lambda_2) = \pi_0^u(\widetilde{q}_2(\Lambda_2))$, and there must exists some $\tilde{n}$ such that $\pi_{\tilde{n}}^u(\widetilde{q}_1(\Lambda_1)) < \pi_{\tilde{n}}^u(\widetilde{q}_2(\Lambda_2))$. Otherwise, we cannot have $\sum_{i=0}^{\infty} \pi_i^u(\widetilde{q}_1(\Lambda_1)) = \sum_{i=0}^{\infty} \pi_i^u(\widetilde{q}_2(\Lambda_2)) = 1$. Recall that $\pi_0^u(\widetilde{q}_1) = \left(\frac{1-\rho_{T1}^{n_e}}{1-\rho_{T1}} + \frac{\rho_{T1}^{n_e}}{1-\rho_{R1}}\right)^{-1}$ and $\pi_0^u(\widetilde{q}_2) = \left(\frac{1-\rho_{T2}^{n_e}}{1-\rho_{T2}} + \frac{\rho_{T2}^{n_e}}{1-\rho_{R2}}\right)^{-1}$, where $\rho_{T1} = \frac{\gamma\Lambda_1\widetilde{q}_1+(1-\gamma)\Lambda_1}{\mu}, \rho_{T2} = \frac{\gamma\Lambda_2\widetilde{q}_2+(1-\gamma)\Lambda_2}{\mu}, \rho_{R1} = \frac{\gamma\Lambda_1\widetilde{q}_1}{\mu}, \rho_{R2} = \frac{\gamma\Lambda_2\widetilde{q}_2}{\mu}$. First, we aim to show that $\rho_{T1} < \rho_{T2}$. To see this, assume $\rho_{T1} \geq \rho_{T2}$, then because $\Lambda_1 < \Lambda_2$, we must have $\rho_{R1} > \rho_{R2}$. The geometric structure of the steady-state probability in (2), along with $\pi_0^u(\widetilde{q}_1) > \pi_0^u(\widetilde{q}_2)$ implies that $\pi_i^u(\widetilde{q}_1) > \pi_i^u(\widetilde{q}_2)$ for all i. Hence, a contradiction. Next, we aim to show that $\rho_{R1} < \rho_{R2}$. To see this, assume that $\rho_{R1} \geq \rho_{R2}$ which is equivalent to $\Lambda_1\tilde{q}_1 \geq \Lambda_2\tilde{q}_2$. Consider the special case $\gamma = 1$, where the steady-state distribution follows an exact geometric structure with $\rho_{T1} = \rho_{R1} \geq \rho_{T2} = \rho_{R2}$. Because $\pi_0^u(\widetilde{q}_1) > \pi_0^u(\widetilde{q}_2)$, similarly, we have a contradiction. Hence, we conclude that $\rho_{T1} < \rho_{T2}$ as well as $\rho_{R1} < \rho_{R2}$. Note that the two probability distributions have the same structure (both are geometric-like, with the former having a bigger probability mass at 0 than the latter. Hence, it is straightforward to see that there must exists some $\tilde{n}$ such that $\pi_n^u(q_1(\Lambda_1)) > (\leq)\pi_n^u(q_2(\Lambda_2))$ when $n < \tilde{n}$ $(n \geq \tilde{n})$. We shall show that the probability distribution $\{\pi_n^u(\widetilde{q}_2)\}_{n=0}^{\infty}$ stochastically dominates the probability distribution $\{\pi_n^u(\widetilde{q}_1)\}_{n=0}^{\infty}$. Hence, the distribution of $\{\pi_n^u(\widetilde{q})\}_{n=0}^{\infty}$ satisfies the condition (ii) of technical Lemma EC.2.1 and $U^u(\widetilde{q}_1(\Lambda_1)) - U^u(\widetilde{q}_2(\Lambda_2)) > 0$. is proved. Further, the utility function of remote customers under equilibrium is $U^u(\Lambda) = 0$ when $\underline{\lambda}_A^u \leq \Lambda < \bar{\lambda}_A^u$. Then, $U^u(\widetilde{q}(\Lambda)) - U^u(q_A^u)$ is decreasing in Λ, and this indicates $\widetilde{q} - q_A^u$ increases in $\Lambda \in (\underline{\lambda}_A^u, \bar{\lambda}_A^u)$. Since we have know that the equilibrium order-placing probability $q_A^u = 0 < \widetilde{q} \in (0,1)$ when the market size $\Lambda \geq \bar{\lambda}_A^u$ and $q_A^u = 1 > \widetilde{q} \in (0,1)$ when the market size $\Lambda \leq \bar{\lambda}_A^u$, this reminds us that there exists a unique market size $\widetilde{\Lambda} \in (\underline{\lambda}_A^u, \bar{\lambda}_A^u)$ that enables $q_A^u > (\leq)\widetilde{q}$ when $\Lambda < (\geq)\widetilde{\Lambda}$. Therefore, the throughput under two information provision policies are equal when the market size $\Lambda = \widetilde{\Lambda}$, and when $\Lambda < \widetilde{\Lambda}$, $TH_A^u(q_A^u) > TH_A^u(\widetilde{q}) = TH_A^o$; when $\Lambda > \widetilde{\Lambda}$, $TH_A^u(q_A^u) < TH_A^u(\widetilde{q}) = TH_A^o$. □

Proof of Theorem 2 To compare the throughput of the two models under optimal information, we first consider the order-onsite model. Recall that when the market size $\Lambda \leq \underline{\lambda}_S^u$, the equilibrium order-placing probability of remote customers is $q_S^u = 1$ of the order-onsite model. We then focus on the specific market size $\Lambda = \underline{\lambda}_S^u$. In this case, $TH_S^* = TH_S^u$ by using the result in technical Lemma EC.2.2. We compare the throughput under optimal information of the two models by considering two cases: First, when $TH_A^* = TH_A^o$, the order-onsite model outperforms the order-ahead model by achieving a higher throughput $(TH_S^u > TH_A^* = TH_A^o)$ by using the result in technical Lemma EC.2.2. Otherwise, when the maximum throughput of the order-ahead model is $TH_A^* = TH_A^u$, the equilibrium order-placing probability $q_S^u = 1$ while q_A^u is solved by Proposition 1. The rest of the proof is similar to the last part of the proof of Theorem 1, where we showed that for a sufficiently large β, we must have $TH_S^u > TH_A^u$ when $\Lambda = \underline{\lambda}_S^u$. □

Proof of Lemma 2 Consider an arriving customer with N outstanding orders ahead of hers. (i) If $N < n_e$, her expected utility if she keeps on waiting is no less than $V - c(N+1)/\mu \geq 0$. This lower bound $V - c(N+1)/\mu$ would be attained if no customers ahead of her were to cancel their order. Since even the lower bound is nonnegative, any customer seeing $N < n_e$ keeps on waiting. (ii) If $N \geq n_e$, the arriving customer knows (based on the preceding argument) that the first n_e outstanding orders will not be canceled. Thus, her expected utility if she keeps on waiting is no greater than $V - c(n_e+1)/\mu < 0$. This upper bound $V - c(n_e+1)/\mu$ would be attained if all customers ahead of her beyond the first n_e were to cancel their order. Since even the upper bound is negative, any customer seeing $N \geq n_e$ abandons and cancels her order. □

Proof of Lemma 3 Consider a tagged remote customer who observes n existing orders upon her arrival. Let $T \sim \text{Exp}(\beta)$ denote her travel time, and $S_0 \sim \text{Exp}(\mu)$ denote her service time. Let r.v.'s $Y_1, \ldots, Y_n$ denote inter-departure time of the order for the n outstanding orders (excluding the tagged customer), either by service completion or by order cancellation, with Y_i corresponding to the i^{th} order in the order queue. The corresponding departure rate is $\mu_i = \mu + (i - n_e)^+\beta, \quad i = 0, 1, \cdots$. We next derive the distribution for N_n^C, the tagged remote customer' updated queue position (including herself) when she arrives at the service facility. Note that $N_n^C \in \{0, 1, \cdots, n+1\}$. We denote the probabilities by $p_n^C(0), \ldots, p_n^C(n+1)$. It is straightforward to see that the updated queue position is i if and only if there are exactly $n - i + 1$ orders removed (either for service completion or cancellation) from the order queue when the tagged customer arrives at the service facility; this corresponds to the event $\{Y_1 + \cdots + Y_{n-i+1} < T < Y_1 + \cdots + Y_{n-i+1} + Y_{n-i+2}\}$. (i) If $n > n_e$, the probability distribution of N_n^C is given by $p_n^C(n+1) \equiv \mathbb{P}(N_n^C = n+1) = \mathbb{P}(T < Y_{n+1}) = \frac{\beta}{\mu_n + \beta}; p_n^C(i) \equiv \mathbb{P}(N_n^C = i) = \mathbb{P}(T > Y_{n+1}) \times \cdots \times \mathbb{P}(T > Y_{i+1}) \times \mathbb{P}(T < Y_i) = \frac{\beta}{\mu_{i-1}+\beta} \prod_{k=i}^{n} \frac{\mu_k}{\mu_k+\beta}, 1 \leq i \leq n; p_n^C(0) \equiv \mathbb{P}(N_n^C = 0) = \mathbb{P}(T > Y_{n+1}) \times \cdots \times \mathbb{P}(T > Y_1) \times \mathbb{P}(T > S_0) = \prod_{k=0}^{n} \frac{\mu_k}{\mu_k+\beta}$; (ii) If $n \leq n_e$, the departure rate of outstanding orders is $\mu_n = \mu$ and then N_n^C has the same distribution as N_n given by Lemma 1. □

Proof of Proposition 4 First, we characterize remote customers' expected utility function. Consider a tagged remote customer who observes n outstanding orders upon experiencing a need and places an order. The probability she will continue to wait upon arriving onsite is $\vartheta_C(n) = \mathbb{P}(N_n^C \leq (n+1) \wedge n_e) = \begin{cases} 1, & \text{if} \quad n < n_e \\ \sum_{i=0}^{n_e} p_n^C(i) = \prod_{k=0}^{n} \frac{\mu_k}{\mu_k+\beta} + \frac{\beta}{\mu+\beta} \sum_{i=1}^{n_e} \prod_{k=i}^{n} \frac{\mu_k}{\mu_k+\beta}, & \text{otherwise} \end{cases}$ where the probabilities $p_n^C(i)$ are given in Lemma 3, and $x \wedge y \equiv \min\{x, y\}$. Hence, the mean remaining onsite waiting time if joining is $w_C(n) \equiv \sum_{i=0}^{(n+1)\wedge n_e} \frac{i}{\mu} \cdot p_n^C(i) = \begin{cases} \frac{1}{\beta}\left(\sigma^{n+1} + \frac{(n+1)\beta}{\mu} - 1\right), & \text{if} \quad n < n_e; \\ \frac{\beta}{\mu+\beta} \sum_{i=1}^{n_e} \frac{i}{\mu} \prod_{k=i}^{n} \frac{\mu_k}{\mu_k+\beta}, & \text{otherwise.} \end{cases}$ Let $U_C(n)$ denote the expected utility of a remote customer who observes a queue length n and places an order to join the queue at Stage 1. Thus, $U_C(n) = V\vartheta_C(n) - cw_C(n) - \frac{c}{\beta} = \bar{U}_C(n)\mathbf{1}_{\{n<n_e\}} + \widetilde{U}_C(n)\mathbf{1}_{\{n \geq n_e\}}$, where the indicator function $\mathbf{1}_A$ is equal to 1 if condition A holds and 0 otherwise, and the two functions $\bar{U}_C(n)$ and $\widetilde{U}_C(n)$ are given by $\bar{U}_C(n) \equiv V - \frac{c}{\beta}\left(\sigma^{n+1} + \frac{(n+1)\beta}{\mu} - 1\right) - \frac{c}{\beta} = \bar{U}(n), \widetilde{U}_C(n) \equiv V\left(\prod_{k=0}^{n} \frac{\mu_k}{\mu_k+\beta} + \frac{\beta}{\mu+\beta} \sum_{j=1}^{n_e} \prod_{k=j}^{n} \frac{\mu_k}{\mu_k+\beta}\right) - c\frac{\beta}{\mu+\beta} \sum_{j=1}^{n_e} \frac{j}{\mu} \prod_{k=j}^{n} \frac{\mu_k}{\mu_k+\beta} - \frac{c}{\beta}$. Therefore, when the queue-length information is not shared remotely in the cancellation model, given that all other remote customers place orders with probability q, the expected utility for a tagged customer to place an order is $U_C^u(q) = \sum_{i=0}^{n_e-1} \bar{U}_C(i)\pi_{i,C}^u(q) + \sum_{i=n_e}^{\infty} \widetilde{U}_C(i)\pi_{i,C}^u(q) = \sum_{i=0}^{n_e-1}\left(V - \frac{c}{\beta}\left(\sigma^{i+1} + \frac{(i+1)\beta}{\mu} - 1\right)\right)\pi_{i,C}^u(q) - \frac{c}{\beta} + \sum_{i=n_e}^{\infty}\left[V\left(\prod_{k=0}^{i} \frac{\mu_k}{\mu_k+\beta} + \frac{\beta}{\mu+\beta} \sum_{j=1}^{n_e} \prod_{k=j}^{i} \frac{\mu_k}{\mu_k+\beta}\right) - c\frac{\beta}{\mu+\beta} \sum_{j=1}^{n_e} \frac{j}{\mu} \prod_{k=j}^{i} \frac{\mu_k}{\mu_k+\beta}\right]\pi_{i,C}^u(q)$, where the steady-state probabilities are given by $\pi_{i,C}^u(q) = \left(\frac{1-\rho_T^{n_e+1}}{1-\rho_T} + \rho_T^{n_e} \sum_{j=1}^{\infty} \prod_{k=1}^{j} \frac{\gamma\Lambda q}{\mu+k\beta}\right)^{-1} \rho_T^{(i\wedge n_e)} \prod_{k=1}^{(i-n_e)^+} \frac{\gamma\Lambda q}{\mu+k\beta}, \quad i = 0, 1, \cdots$, and $\rho_T = [\gamma\Lambda q + (1-\gamma)\Lambda]/\mu$. To show $U_C(n)$ is strictly decreasing in n, we have $\bar{U}_C(n) - \bar{U}_C(n-1) = \frac{c}{\mu}\left(\sigma^{n+1} - 1\right) < 0, \widetilde{U}_C(n) - \widetilde{U}_C(n-1) = \left(\frac{\mu_n}{\mu_n+\beta} - 1\right)\left[V\prod_{k=0}^{n-1} \frac{\mu_k}{\mu_k+\beta} + \frac{\beta}{\mu+\beta} \sum_{i=1}^{n_e}\left(V - \frac{ic}{\mu}\right)\prod_{k=i}^{n-1} \frac{\mu_k}{\mu_k+\beta}\right] < 0$. In addition, $\widetilde{U}_C(n_e) - \bar{U}_C(n_e - 1) = (\sigma - 1)\left(V - \frac{n_e c}{\mu}\right) + \frac{c}{\mu+\beta}\left(\sigma^{n_e} - 1\right) < 0$, which implies that $U_C(n)$ is decreasing in n. To give customers' equilibrium joining strategy, we first give some properties of $U_C^u(q)$ in (5) in the following Lemma. Let $U_{C,\rho}^u(q)$ denote the expected utility of a joining remote customer when $\Lambda/\mu = \rho$ and other remote customers join with probability q.

Lemma EC.2.4 **(Property of U_C^u function)**. *The utility function U_C^u has the following properties: (i) $U_{C,\infty}^u(1) = U_{C,\infty}^u(0) = -c/\beta < 0$ and $U_{C,0}^u(1) = U_{C,0}^u(0) = \bar{U}(0) > 0$. (ii) $U_C^u(q)$ is continuous and strictly decreasing in q. (iii) $U_C^u(1)$ is continuous and strictly decreasing in ρ, where $U_C^u(1)$ is given by $U_C^u(1) = \sum_{i=0}^{n_e-1}\left(V - \frac{c}{\beta}\left(\sigma^{i+1} + \frac{(i+1)\beta}{\mu} - 1\right)\right)\pi_{i,C}^u(1) - \frac{c}{\beta} + \sum_{i=n_e}^{\infty}\left[V\left(\prod_{k=0}^{i} \frac{\mu_k}{\mu_k+\beta} + \frac{\beta}{\mu+\beta} \sum_{j=1}^{n_e} \prod_{k=j}^{i} \frac{\mu_k}{\mu_k+\beta}\right) - c\frac{\beta}{\mu+\beta} \sum_{j=1}^{n_e} \frac{j}{\mu} \prod_{k=j}^{i} \frac{\mu_k}{\mu_k+\beta}\right]\pi_{i,C}^u(1)$.*

The proof of Lemma EC.2.4 is similar to that of Lemma EC.2.3 and thus omitted due to the page limit.

Finishing the proof of Proposition 4. Similar to the proof of Proposition 1, and by the property of the U_C^u function, there must be a unique solution $\underline{\rho}_C^u$ to equation $U_C^u(1)=0$. Similarly, there must be a unique solution $\bar{\rho}_C^u$ to equation $U_C^u(0)=0$. Denoting $\underline{\lambda}_C^u=\mu\underline{\rho}_C^u$ and $\bar{\lambda}_C^u=\mu\bar{\rho}_C^u$ completes the proof. □

Proof of Theorem 3 The proof proceeds in two steps. **Step 1**: We first prove $TH_C^u \geq TH_S^u$. We compare the system throughput of the order-ahead-with-cancellation (OAC) model and the order-onsite model in the following cases specified by the equilibrium order-placing probabilities of remote customers (q_C^u and q_S^u): **Case 1:** $q_C^u \geq q_S^u$. This case includes three subcases: (a) $q_C^u=1, q_S^u \in [0,1]$, (b) $q_S^u=q_C^u=0$, and (c) $q_C^u \in (0,1), q_S^u=0$. Note that $\pi_{0,C}^u(q)=\left(\frac{1-\rho_T^{n_e+1}}{1-\rho_T}+\rho_T^{n_e}\sum_{j=1}^{\infty}\prod_{k=1}^{j}\frac{\gamma\Lambda q}{\mu+k\beta}\right)^{-1} \leq \left(\frac{1-\rho_T^{n_e+1}}{1-\rho_T}\right)^{-1}=\pi_{0,S}^u(q)$. Therefore, $\pi_{0,C}^u(q_C^u) \leq \pi_{0,C}^u(q_S^u) \leq \pi_{0,S}^u(q_S^u)$, which implies that $TH_C^u \geq TH_S^u$. **Case 2:** $q_C^u < q_S^u$. In this case, the equilibrium order-placing probability of remote customers must be strictly positive with $q_S^u>0$ and we must have $q_C^u<1$. When $n_e=1$, in the order-onsite model, the remote customer's expected utility from traveling is $\pi_{0,S}^u(q_S^u)(V-c/\mu)-c/\beta \geq 0$, where $\pi_{0,S}^u$ is the idle probability. We then consider the following two subcases specified by the value of q_C^u. **Case 2a:** We first consider the case $q_C^u \in (0,1)$. In the OAC model, the remote customer's expected utility from ordering is $p_0V+p_1(V-c/\mu)-c/\beta=0$, where p_0 is the probability that the order is ready when the customer arrives at the service facility, p_1 is the probability that the order is not ready when the customer arrives at the service facility, and $1-p_0-p_1$ is the cancellation probability. Let $\pi_{0,C}^u$ be the idle probability in the OAC model. Thus, $p_0+p_1>\pi_{0,C}^u$, because if the system is idle when a customer places the order, the customer will surely not cancel the order. Moreover, we have $\underbrace{\pi_{0,S}^u(V-c/\mu)-c/\beta}_{\geq 0} \geq \underbrace{p_0V+p_1(V-c/\mu)-c/\beta}_{=0} > (p_0+p_1)(V-c/\mu)-c/\beta$, which implies $\pi_{0,S}^u>p_0+p_1$. Hence $\pi_{0,S}^u>\pi_{0,C}^u \Longleftrightarrow \mu(1-\pi_{0,C}^u)>\mu(1-\pi_{0,S}^u)$. Therefore, the OAC model yields higher throughput than the order-onsite model. **Case 2b:** We next consider the case $q_C^u=0$. First, we must have $p_0V+p_1(V-c/\mu)-c/\beta \leq 0$. The OAC system with $q_C^u=0$ is equivalent to an $M/M/1/1$ queue. Thus, the steady-state probability that the system has exactly one outstanding order is $\pi_{1,C}^u=\frac{\Lambda(1-\gamma)}{\mu+\Lambda(1-\gamma)}$, and the steady-state probability that the system is empty is $\pi_{0,C}^u=1-\pi_{1,C}^u$. The cancellation probability is $(1-p_0-p_1)$, i.e., the probability of seeing exactly two outstanding orders (including one's own order) ahead after arriving onsite. Hence, $1-p_0-p_1=\pi_{1,C}^u\frac{\beta}{\beta+\mu}<\pi_{1,C}^u$. Thus, $p_0+p_1>\pi_{0,C}^u$. Note that in the order-onsite system, $\pi_{0,S}^u(q)$ is decreasing in q and $\pi_{0,S}^u(0)=\pi_{0,C}^u(0)$. Because $q_S^u>0$, we have $\pi_{0,C}^u(0)>\pi_{0,S}^u(q_S^u)$. Thus, $p_0+p_1>\pi_{0,C}^u(0)>\pi_{0,S}^u(q_S^u)$, which implies that $p_0V+p_1(V-c/\mu)-c/\beta>\pi_{0,S}^u(q_S^u)(V-c/\mu)-c/\beta \geq 0$. That is, $p_0V+p_1(V-c/\mu)-c/\beta>0$. However, this contradicts $p_0V+p_1(V-c/\mu)-c/\beta \leq 0$. Therefore, the case $q_C^u=0$ and $q_S^u>0$ does not exist. In summary, when $n_e=1$, $TH_C^u \geq TH_S^u$.

Step 2: We next prove $TH_C^o \geq TH_S^o$ for $n_e=1$. Note that $TH_C^o=TH_A^o$. Hence, we need to prove $TH_A^o \geq TH_S^o$ for $n_e=1$. This is done later in Theorem EC.1. □

To facilitate our proofs, we introduce an order-ahead model variant when queue information is shared with remote customers. In this variant, remote customers can cancel their orders upon arrival at the store, based on the updated queue position of their order at that time. We refer to this variant as "order-ahead with onsite balking" (OAOB).

Lemma EC.2.5. *In the OAOB model, a remote customer orders if and only if the queue length $n<n_e^*$.*

Proof of Lemma EC.2.5 Consider a tagger remote customer who observes n orders upon arriving, N_n^{OB} denotes her updated queue position (including herself) upon arrival at the service facility if she places an order. If she places an order and travels to the service facility, she will keep the order with probability $\vartheta(n) \equiv \mathbb{P}(N_n^{OB} \leq n_e)$. Because $N_n^{OB} \leq n+1$, we have $\vartheta(n)=\mathbb{P}(N_n^{OB} \leq (n+1)\wedge n_e)=\begin{cases}1, & \text{if } \; n<n_e,\\ \sum_{i=0}^{n_e}p_n(i)=\sigma^{n-n_e+1}, & \text{otherwise,}\end{cases}$ where probabilities $p_n(i)$ are given in Lemma 1, and $x \wedge y \equiv \min\{x,y\}$. Her expected onsite waiting time, $w(n)$, is $w(n)=$

$\sum_{i=0}^{(n+1)\wedge n_e} \frac{i}{\mu} \cdot p_n(i) = \begin{cases} \frac{1}{\beta}\left(\sigma^{n+1} + \frac{(n+1)\beta}{\mu} - 1\right), & \text{if} \quad n < n_e, \\ \frac{1}{\beta}\left(\sigma^{n+1} - \left(1 - \frac{n_e\beta}{\mu}\right)\sigma^{n-n_e+1}\right), & \text{otherwise.} \end{cases}$ Let $U^{OB}(n)$ denote her expected utility. Then $U^{OB}(n) = V\vartheta(n) - cw(n) - \frac{c}{\beta} = \bar{U}(n)\mathbf{1}_{\{n<n_e\}} + \widetilde{U}(n)\mathbf{1}_{\{n\geq n_e\}}$, where $\bar{U}(n)$ is given by (1) and $\widetilde{U}(n) = V\sigma^{n-n_e+1} - \frac{c}{\beta}\left(\sigma^{n+1} - \left(1 - \frac{n_e\beta}{\mu}\right)\sigma^{n-n_e+1} + 1\right)$. It remains to show that $U^{OB}(n) < 0$ for all $n \geq n_e^*$. If $n_e^* < n_e$, $U^{OB}(n_e^*) = \bar{U}(n_e^*) < 0$; otherwise, $U^{OB}(n_e^*) = \widetilde{U}(n_e)$ and $\widetilde{U}(n_e) = \sigma\left(V - \frac{n_e c}{\mu}\right) + \frac{c}{\beta}\left(\sigma\left(1 - \sigma^{n_e}\right) - 1\right) = \sigma\left[\left(V - \frac{c(n_e+1)}{\mu}\right) - \frac{c}{\beta}\sigma^{n_e}\right] < 0$. We then conclude that $U^{OB}(n_e^*) < 0$. To show the monotonicity of $U^{OB}(n)$ in n, we establish the monotonicity for both $\bar{U}(n)$ and $\widetilde{U}(n)$, namely, $\bar{U}(n) - \bar{U}(n-1) = \frac{c}{\mu}\left(\sigma^{n+1} - 1\right) < 0$ and $\widetilde{U}(n) - \widetilde{U}(n-1) = -\sigma^{n-n_e}(1-\sigma)\left(V - \frac{n_e c}{\mu} + \frac{(1-\sigma^{n_e})c}{\beta}\right) < 0$. Further, $\widetilde{U}(n_e) - \bar{U}(n_e - 1) = -(1-\sigma)\left(V - \frac{n_e c}{\mu} + \frac{(1-\sigma^{n_e})c}{\beta}\right) < 0$. Hence, $U^{OB}(n)$ is decreasing in n. We then have $U^{OB}(n) < 0$ for all $n \geq n_e^*$. □

Theorem EC.1. *When queue-length information is shared remotely and $n_e = 1$, the hybrid order-ahead scheme has higher throughput than the pure order-onsite scheme, i.e., $TH_A^o \geq TH_S^o$.*

Proof of Theorem EC.1 In the pure order-onsite system, given that remote customers observe $n \leq n_e$ orders in the onsite queue, we will show that the maximum onsite queue length under which a remote customer is willing to travel with a positive probability must be less than n_e^*. We draw on the results of the auxiliary OAOB model given in Lemma EC.2.5. Consider a tagged remote customer who observes n outstanding orders upon arrival, let N_n^S denote her updated queue position (including herself) if she places an onsite order upon arrival at the service facility. We then have $U_S(n) = \mathbb{E}\left[V - \frac{cN_n^S}{\mu}\right]^+ - \frac{c}{\beta} \leq \mathbb{E}\left[V - \frac{cN_n^{OB}}{\mu}\right]^+ - \frac{c}{\beta} = U^{OB}(n)$, where N_n^{OB} and $U^{OB}(n)$ are the updated queue position and expected utility function defined in the OAOB model, and the inequality holds because $N_n^S \geq_{st} N_n^{OB}$. Because remote customers in the OAOB model will not place an order when $n \geq n_e^*$ (as shown in Lemma EC.2.5), neither will the aforementioned tagged customer in the present order-onsite model when $n \geq n_e^*$.

Now consider the case $n_e = 1$. Since we have proved $n_e^* \leq n_e$ in Proposition 2. Also, we have that $n_e^* \geq 1$ according to Assumption 1. Hence, $n_e^* = 1$ in this case. The throughput in the hybrid order-ahead system is $TH_A^o = \Lambda\mu/(\Lambda + \mu)$. We consider the following two cases for the pure order-onsite system: (1) No remote customers travel in the order-onsite system. Then clearly, the order-ahead throughput is higher. (2) Remote customers in the order-onsite system travel with a positive probability $p > 0$ if and only if they see an empty onsite queue. Let $\lambda_L = \Lambda(1-\gamma)$ be the arrival rate of local customers and $\lambda_R = \Lambda\gamma p$ be the travel rate of remote customers, and $\lambda_L + \lambda_R \leq \Lambda$. Let $\pi_{k,i}$ be the steady-state probability of state (k,i), where $k \in \{0,1\}$ is the number of customers in the onsite queue and $i \in \{0,1,...\}$ is the number of traveling customers. The balance equations are

$$(i\beta + \mu)\pi_{1,i} = \lambda_L\pi_{0,i} + (i+1)\beta\pi_{0,i+1} + (i+1)\beta\pi_{1,i+1}, \quad i = 0,1,... \tag{EC.1}$$

$$(\lambda_R + \lambda_L + i\beta)\pi_{0,i} = \mu\pi_{1,i} + \lambda_R\pi_{0,i-1}, \quad i = 1,2,... \tag{EC.2}$$

$$(\lambda_R + \lambda_L)\pi_{0,0} = \mu\pi_{1,0}. \tag{EC.3}$$

We prove the following: $\lambda_R\pi_{0,i} = (i+1)\beta(\pi_{0,i+1} + \pi_{1,i+1})$, $i = 0,1,...$ We first show that it holds for $i = 0$. Equation (EC.1) gives $\mu\pi_{1,0} = \lambda_L\pi_{0,i} + \beta\pi_{0,1} + \beta\pi_{1,1}$. Combining this with (EC.3) gives $(\lambda_R + \lambda_L)\pi_{0,0} = \lambda_L\pi_{0,i} + \beta\pi_{0,1} + \beta\pi_{1,1}$. Hence $\lambda_R\pi_{0,0} = \beta(\pi_{0,1} + \pi_{1,1})$. Next, we prove that $\lambda_R\pi_{0,i} = (i+1)\beta(\pi_{0,i+1} + \pi_{1,i+1})$, $i = 0,1,...$ holds for $i \geq 1$.

Equation (EC.1) gives $\mu\pi_{1,i} - \lambda_L\pi_{0,i} = (i+1)\beta(\pi_{0,i+1} + \pi_{1,i+1}) - i\beta\pi_{1,i}$ for $i \geq 0$. Equation (EC.2) gives $\mu\pi_{1,i} - \lambda_L\pi_{0,i} = (\lambda_R + i\beta)\pi_{0,i} - \lambda_R\pi_{0,i-1}$ for $i \geq 1$. Hence, for $i \geq 1$, $(i+1)\beta(\pi_{0,i+1} + \pi_{1,i+1}) - i\beta\pi_{1,i} = (\lambda_R + i\beta)\pi_{0,i} - \lambda_R\pi_{0,i-1}$, $(i+1)\beta(\pi_{0,i+1} + \pi_{1,i+1}) - i\beta(\pi_{1,i} + \pi_{0,i}) = \lambda_R(\pi_{0,i} - \pi_{0,i-1})$, $\sum_{i=1}^{j}\left[(i+1)\beta(\pi_{0,i+1} + \pi_{1,i+1}) - i\beta(\pi_{1,i} + \pi_{0,i})\right] = \sum_{i=1}^{j}\lambda_R(\pi_{0,i} - \pi_{0,i-1})$, $(j+1)\beta(\pi_{0,j+1} + \pi_{1,j+1}) - \beta(\pi_{1,1} + \pi_{0,1}) = \lambda_R(\pi_{0,j} - \pi_{0,0})$. Since we have proven $\lambda_R\pi_{0,0} = \beta(\pi_{0,1} + \pi_{1,1})$, it follows that $\lambda_R\pi_{0,i} = (i+1)\beta(\pi_{0,i+1} + \pi_{1,i+1})$ for $i \geq 0$. Hence, $i\beta\pi_{0,i} \leq \lambda_R\pi_{0,i-1}$ for $i \geq 1$. Combining this inequality with (EC.2) shows that $(\lambda_R + \lambda_L)\pi_{0,i} \geq \mu\pi_{1,i}$. Thus, $\mu\sum_{i=0}^{\infty}\pi_{1,i} \leq (\lambda_L + \lambda_R)\sum_{i=0}^{\infty}\pi_{0,i}$. Since $\sum_{i=0}^{\infty}\pi_{1,i} + \sum_{i=0}^{\infty}\pi_{0,i} = 1$, the throughput $TH_S^o = \mu\sum_{i=0}^{\infty}\pi_{1,i} \leq \mu(\lambda_L + \lambda_R)/(\lambda_L + \lambda_R + \mu)$. Since $\lambda_R + \lambda_L \leq \Lambda$ and $TH_A^o = \Lambda\mu/(\Lambda + \mu)$, it follows that $TH_S^o \leq TH_A^o$. □

Proof of Theorem 4 When we compare the order-ahead-with-cancellation (OAC) model and the plain order-ahead model, we note that given any fixed q, the birth rates of the two models are equal and the death rate of the OAC model is higher ($\mu_i > \mu$ when $i > n_e$). Hence, invoking Lemma EC.2.2 gives $\pi^u_{0,C}(q) > \pi^u_0(q)$ and then $\pi^u_{i,C}(q) = \rho^i_T \pi^u_{0,C}(q) > \pi^u_i(q) = \rho^i_T \pi^u_0(q)$ for $i = 0, 1, \cdots, n_e$. There must exist some $n' > n_e$ such that $\pi^u_{n',C}(q) < \pi^u_{n'}(q)$. Otherwise, we cannot have $\sum_{n=0}^{\infty} \pi^u_{n,C}(q) = \sum_{n=0}^{\infty} \pi^u_n(q) = 1$. We then claim that for any $\hat{n} \geq n_e$, if $\pi^u_{\hat{n},C}(q) > \pi^u_{\hat{n}}(q)$ and $\pi^u_{\hat{n}+1,C}(q) \leq \pi^u_{\hat{n}+1}(q)$, then $\pi^u_{n,C}(q) < \pi^u_n(q)$ for all $n > \hat{n}+1$. To show this claim, note that $\pi^u_{\hat{n}+1,C}(q) = \pi^u_{\hat{n},C}(q)\frac{\gamma\Lambda q}{\mu+(\hat{n}-n_e)\beta}$, $\pi^u_{\hat{n}+1}(q) = \pi^u_{\hat{n}}(q)\frac{\gamma\Lambda q}{\mu}$. Hence, $\frac{\gamma\Lambda q}{\mu+(\hat{n}-n_e)\beta} < \frac{\gamma\Lambda q}{\mu}$. This implies $\frac{\gamma\Lambda q}{\mu+(n-n_e)\beta} < \frac{\gamma\Lambda q}{\mu}$ for any $n \geq \hat{n}+1$, which further implies $\pi^u_{n,C}(q) < \pi^u_n(q)$ for $n > \hat{n}+1$. And this satisfies Part (ii) of Lemma EC.2.1, where $\boldsymbol{\pi}^C$ and $\boldsymbol{\pi}^A$ cross each other only once: there exists an $\hat{n}$ such that $\pi^u_{n,C} \geq \pi^u_n$ when $n \leq \hat{n}$ and $\pi^u_{n,C} < \pi^u_n$ when $n > \hat{n}$. In addition, $\bar{U}_C(n) = \bar{U}(n)$, and $\widetilde{U}_C(n) > \bar{U}(n)$ and thus $U_C(n) \geq \bar{U}(n)$ for all n. Both $\bar{U}(n)$ and $U_C(n)$ are decreasing in n. Hence, $U^u_C(q) = \sum_{n=0}^{\infty} U_C(n)\pi^u_{n,C}(q) \geq \sum_{n=0}^{\infty} \bar{U}(n)\pi^u_{n,C}(q) > \sum_{n=0}^{\infty} \bar{U}(n)\pi^u_n(q) = U^u(q)$, and the second inequality holds by Lemma EC.2.1. The above ranking of the utility functions implies the ranking of the solutions for $U^u_C(q) = 0$ and $U^u(q)$ due to the properties established for the two utility functions in Lemmas EC.2.3 and EC.2.4. That is, $\underline{\lambda}^u_A < \underline{\lambda}^u_C$ and $\bar{\lambda}^u_A < \bar{\lambda}^u_C$, hence the equilibrium joining probabilities exhibit $q^u_A \leq q^u_C$. Note that in the plain order-ahead model, $q^u_A = 1$ for $\Lambda \leq \underline{\lambda}^u_A$. In the OAC model, $q^u_C = 1$ for $\Lambda \leq \underline{\lambda}^u_C$. The birth-rate of the two systems are equal while the death-rate in the OAC model is larger, which implies (based on Lemma EC.2.2) that the plain order-ahead system is busier, and thus $TH^u_A > TH^u_C$ for $\Lambda < \min\{\underline{\lambda}^u_A, \underline{\lambda}^u_C\} = \underline{\lambda}^u_A$. It remains to show that for sufficiently small Λ, $TH^*_A = TH^u_A$ (which has already been proved in Proposition 3) and $TH^*_C = TH^u_C$, which can be similarly proved. □

Proof of Proposition 5 The expected utility of placing a remote order in the order-ahead-with-rejection (OAR) model is $U^u_{R,N}(q) = \sum_{n=0}^{N-1} \bar{U}(n)\pi^u_{n,R}(q)$. If $\bar{U}(N-1) \geq 0$, because $\bar{U}(n)$ decreases in n, we have $U^u_{R,N}(q) \geq 0$ for all $q \in [0,1]$, so that the equilibrium order-placing probability is $q^u_{R,N} = 1$. Next, we consider the case $\bar{U}(N-1) < 0$. Define $\tilde{U}^u_{R,N}(q) \equiv \frac{U^u_{R,N}(q)}{\sum_{j=0}^{N-1}\pi^u_{j,R}(q)} = \sum_{n=0}^{N-1} \bar{U}(n)\frac{\pi^u_{n,R}(q)}{\sum_{j=0}^{N-1}\pi^u_{j,R}(q)} \equiv \sum_{n=0}^{N-1} \bar{U}(n) f^u_{n,R}(q)$. We first consider $q = 1$ (the case $q = 0$ is similar). To establish that there exists a unique $\rho > 0$ such that $U^u_{R,N}(q) = 0$, it suffices to show that there exists a unique $\rho > 0$ such that $\tilde{U}^u_{R,N}(q) = 0$ because the latter is the former scaled by a positive term $\sum_{j=0}^{N-1}\pi^u_{j,R}$.

Lemma EC.2.6 **(Property of $\tilde{U}^u_{R,N}$ function)**. *The function $\tilde{U}^u_{R,N}$ exhibits the following properties: (i) $\tilde{U}^u_{R,N}(q)$ is continuous and strictly decreasing in q for a fixed ρ. (ii) $\tilde{U}^u_{R,N}(1)$ and $\tilde{U}^u_{R,N}(0)$ are continuous and strictly decreasing in ρ. (iii) When $\rho = 0$, $\tilde{U}^u_{R,N}(0) = \tilde{U}^u_{R,N}(1) = \bar{U}(0) > 0$; when $\rho \to \infty$, $\tilde{U}^u_{R,N}(0) < 0$ and $\tilde{U}^u_{R,N}(1) < 0$.*

Proof of Lemma EC.2.6 To prove Part (i), we consider two cases: (1) If $N > n_e$. Define $\bar{f}^u_{i,R}(\rho) \equiv f^u_{i,R}(q) = \begin{cases} \frac{\rho^i_T}{\sum_{j=0}^{n_e-1}\rho^i_T + \sum_{j=n_e}^{N-1}\rho_R^{i-n_e}\rho_T^{n_e}}, & \text{if} \quad i < n_e, \\ \frac{\rho_R^{i-n_e}\rho_T^{n_e}}{\sum_{j=0}^{n_e-1}\rho^j_T + \sum_{j=n_e}^{N-1}\rho_R^{j-n_e}\rho_T^{n_e}}, & \text{if} \quad i = n_e, \cdots, N-1, \end{cases}$ where $\rho_R = \gamma\rho q$ and $\rho_T = \gamma\rho q + (1-\gamma)\rho$. For two traffic intensities $\rho_1 < \rho_2$, we have $\bar{f}^u_{0,R}(\rho_2) < \bar{f}^u_{0,R}(\rho_1)$. We find that $\bar{f}^u_{i,R}(\rho_2)/\bar{f}^u_{i,R}(\rho_1)$ increases in i since $\gamma\rho_1 q < \gamma\rho_2 q$ and $\gamma\rho_1 q + (1-\gamma)\rho_1 < \gamma\rho_2 q + (1-\gamma)\rho_2$. This satisfies condition (ii) of Lemma EC.2.1. Because $\bar{U}(n)$ decreases in n, we conclude that $\tilde{U}^u_{R,N}(q)$ decreases in ρ. Next, when the traffic intensity ρ goes to infinity, it is evident that $\lim_{\rho\to\infty}\bar{f}^u_{N-1,R}(\rho) = 1$ when $q \in (0,1]$, and $\lim_{\rho\to\infty}\bar{f}^u_{n_e,R}(\rho) = 1$ when $q = 0$. Hence, the joining utility of a tagged remote customer approaches $\bar{U}(N-1))$ when $q > 0$ and $\bar{U}(n_e)$ when $q = 0$; in both cases, it converges to a negative value. (2) If $N \leq n_e$, from equation (7) we have $\bar{f}^u_{i,R}(\rho) \equiv f^u_{i,R}(q) = \frac{\rho^i_T}{\sum_{j=0}^{N-1}\rho^j_T}$, $\quad i = 0, \cdots, N-1$. For two traffic intensities $\rho_1 < \rho_2$, we have $\bar{f}^u_{0,R}(\rho_2) < \bar{f}^u_{0,R}(\rho_1)$. We find that $\bar{f}^u_{i,R}(\rho_2)/\bar{f}^u_{i,R}(\rho_1)$ increases in i for $i = 0, 1, \cdots, N-1$ since $\gamma\rho_1 q + (1-\gamma)\rho_1 < \gamma\rho_2 q + (1-\gamma)\rho_2$. Hence, $\tilde{U}^u_{R,N}(q)$ decreases in q for a fixed ρ. When $\rho \to \infty$, the joining utility of a tagged remote customer is $\tilde{U}^u_{R,N}(q)$ approaches $\bar{U}(N-1) < 0$ for all $q \in [0,1]$ because $\lim_{\rho\to\infty}\bar{f}^u_{N-1,R}(\rho) = 1$. The proof of Part (ii) is similar to Part (i). To prove Part (iii), $\tilde{U}^u_{R,N}(0) = \tilde{U}^u_{R,N}(1) = \bar{U}(0) > 0$ by Assumption 1. Combined with Part (ii), we have $\tilde{U}^u_{R,N}(0) < 0$ and $\tilde{U}^u_{R,N}(1) < 0$ when $\rho \to \infty$. □

Finishing the proof of Proposition 5. Similar to the proof of Proposition 1, and by the property of the utility function $\tilde{U}^u_{R,N}$, both $U^u_{R,N}(0)=0$ and $U^u_{R,N}(1)=0$ have a unique solution, denoted by $\bar{\rho}^u_{R,N}, \underline{\rho}^u_{R,N}$, respectively. Furthermore, let $\bar{\lambda}^u_{R,N}=\mu\bar{\rho}^u_{R,N}$, $\underline{\lambda}^u_{R,N}=\mu\underline{\rho}^u_{R,N}$. □

Proof of Theorem 5 First, if the service provider uses a rejection threshold $N \le n_e^*$, then all remote customers have nonnegative utilities so that their order-placing probability is $q=1$. In addition, the system throughput under rejection threshold $N < n_e^*$ is lower than that under threshold n_e^*, because the former model rejects customers who observe $i, i=n_e^*,\cdots,N-1$ outstanding orders, who are supposed to place an order in the latter model. Therefore, the optimal rejection threshold N^* must satisfy $N^* \ge n_e^*$. Hence, it suffices to focus on the case $N \ge n_e^*$ below.

Next, we prove that $N^*=n_e^*$ for sufficiently large Λ. For any rejection threshold $N > n_e^*$, $\bar{U}(N-1)<0$, and Proposition 5 shows that the order-placing probability $q=0$ for sufficiently large Λ. By contrast, $q=1$ for $N=n_e^*$. Hence, by Lemma EC.2.2, the throughput under rejection threshold n_e^* is higher than that under any rejection threshold $N > n_e^*$ for sufficiently large Λ.

Next, we prove that $N^*=\infty$ for sufficiently small Λ. Let $\rho=\Lambda/\mu$, $\rho_T=(\gamma\Lambda q+(1-\gamma)\Lambda)/\mu$, $\rho_R=\gamma\Lambda q/\mu$, $\rho_L=(1-\gamma)\Lambda/\mu$. Define steady-state probabilities: $\pi^u_{i,R}(q)=\begin{cases}\rho_T^{i\wedge n_e}\rho_R^{(i-n_e)^+}\pi^u_{0,R}(q), & \text{if} \quad N>n_e,\\ \rho_T^{i\wedge N}\rho_L^{(i-N)^+}\pi^u_{0,R}(q), & \text{if} \quad N\le n_e,\end{cases}$, $i=0,1,\ldots,N\vee n_e$, where $\pi^u_{0,R}(q)=\begin{cases}\left(\frac{1-\rho_T^{n_e}}{1-\rho_T}+\frac{\rho_T^{n_e}\left(1-\rho_R^{N-n_e+1}\right)}{1-\rho_R}\right)^{-1}, & \text{if} \quad N>n_e,\\ \left(\frac{1-\rho_T^{N}}{1-\rho_T}+\frac{\rho_T^{N}\left(1-\rho_L^{n_e-N+1}\right)}{1-\rho_L}\right)^{-1}, & \text{if} \quad N\le n_e.\end{cases}$ Let $\pi^u_{n,R}(q;N)$ be the steady-state probability of n orders in the OAR model with rejection threshold N, where remote customers place orders with probability q. Also, let $U^u_R(q;N)(q;N)$ denote the joining utility of a remote customer in the OAR model with rejection threshold N, where remote customers place orders with probability q. **(1)** We first prove that the expected utility when all customers join $U^u_R(1;N)=\sum_{n=0}^{N-1}\bar{U}(n)\pi^u_{n,R}(1;N)$ is decreasing in rejection threshold N for $N\ge n_e^*$. To prove this claim, recognize that $\pi^u_{n,R}(1;N) > \pi^u_{n,R}(1;N+1)$ for $n=0,1,\ldots,N$. Hence, distribution $\{\pi^u_{n,R}(1;N+1)\}$ stochastically dominates distribution $\{\pi^u_{n,R}(1;N)\}$. That is, we can stochastically rank the steady-state queue length $Q(N)$ and $Q(N+1)$ under the two thresholds N and $N+1$ as: $Q(N)\le_{\text{st}} Q(N+1)$. Define the function $f(x)\equiv\bar{U}(x)\mathbf{1}_{\{x\le N\}}$, we have $U^u_R(1;N)=\sum_{n=0}^{N-1}\bar{U}(n)\pi^u_{n,R}(1;N) > \sum_{n=0}^{N}\bar{U}(n)\pi^u_{n,R}(1;N)=\mathbb{E}[f(Q(N))]\ge\mathbb{E}[f(Q(N+1))]=\sum_{n=0}^{N}\bar{U}(n)\pi^u_{n,R}(1;N+1)=U^u_R(1;N+1)$, where the first inequality holds because $\bar{U}(n)<0$ for $n\ge n_e^*$, and the second inequality holds because the function $f(x)$ decreases in x. **(2)** Given that all customers join, the throughput $\mu(1-\pi^u_{0,R}(1))$ is increasing in rejection threshold N due to a larger birth-rate (see Lemma EC.2.2). **(3)** It follows from (1) and (2) that if $U^u_R(1;\infty)\ge 0$, then the optimal rejection threshold is ∞. Further $U^u_R(1;\infty)$ is decreasing in $\Lambda\in(0,\mu)$ because distribution $\{\pi^u_{n,R}(1;\infty,\Lambda)\}$ stochastically increases with Λ. Hence, $\exists\underline{\Lambda}$ such that if $\Lambda\le\underline{\Lambda}$, the optimal rejection threshold is ∞. Note that when $N=\infty$, this model reduces to the plain order-ahead model, hence $\underline{\Lambda}=\underline{\lambda}^u_A$. □

Proof of Theorem 6 When queue-length information is shared remotely, we define N_o^* to be the optimal rejection threshold in the rejection model. First, we consider the case where $N_o^*\ge n_e^*$. In this case, the joining behavior of remote customers coincides with the behavior of the plain order-ahead model when queue-length information is shared remotely (balk with threshold n_e^*), that is $TH^o_{R,N_o^*}=TH^o_A$. We next consider the case where $N_o^*<n_e^*$. In this case, customers will always join since $\bar{U}(N_o^*)>0$, which further implies that the joining probability of remote customers is $q^o_R=1$. Compared to the case where $N_o^*=n_e^*$, in which all remote customers place orders, the birth rate in the former case is smaller. Lemma EC.2.2 implies that the latter system is busier, which consequently results in a higher system throughput, i.e., $TH^o_{R,N_o^*}<TH^o_A$. This implies that system throughput in the OAR model when queue-length information is shared remotely will not exceed that of the plain order-ahead model when queue-length information is shared remotely, that is, $TH^o_R\le TH^o_A$. Recall from Theorem 5 that when the queue-length information is not shared

remotely, the optimal rejection threshold satisfies $N^* \geq n_e^*$, thus resulting in a higher system throughput than that of the plain order-ahead model when queue-length information is shared remotely. In summary, $TH_R^u \geq TH_R^o$. □

Proof of Proposition 6 From Theorem 5, we have proved the optimal rejection threshold $N^* \geq n_e^*$. Consider a tagged rejected remote customer. The number of outstanding orders satisfies $n \geq n_e^*$. Let N_n^S denote her updated queue position (including herself) upon arrival at the service facility if she travels to the service facility. Let $U_S(n)$ be her expected utility if she travels, and we then have $U_S(n) = \mathbb{E}\left[V - \frac{cN_n^S}{\mu}\right]^+ - \frac{c}{\beta} \leq \mathbb{E}\left[V - \frac{cN_n^{OB}}{\mu}\right]^+ - \frac{c}{\beta} = U^{OB}(n) < 0$, where N_n^{OB} and $U^{OB}(n)$ are the updated queue size and utility function defined in the OAOB model (see proof of Lemma EC.2.5), the first inequality holds because $N_n^S \geq_{st} N_n^{OB}$, and the last inequality holds by Lemma EC.2.5. □

Proof of Theorem 7 First, consider the case of queue-length information not being shared remotely. We suppose the rejection threshold $N = n_e$. Recall that the steady-state probability of the number of outstanding orders in OAR model is $\pi_{i,R}^u(q) = \frac{\rho_T^i}{\sum_{j=0}^{n_e} \rho_T^j}, i = 0, 1, \cdots, n_e$. and the steady-state probability of the number of outstanding orders in the order-onsite model is $\pi_{i,S}^u(q) = \frac{\rho_T^i}{\sum_{j=0}^{n_e} \rho_T^j} = \pi_{i,R}^u(q), i = 0, 1, \cdots, n_e$. We next compare the utilities of a tagged remote customer who decides to join the order-onsite model and the OAR model. We have $U_{R,N}^u(q) = \sum_{i=0}^{n_e-1} \bar{U}(i)\pi_{i,R}^u(q) = \sum_{i=0}^{n_e-1}\left(V - \frac{(i+1)c}{\mu}\right)\pi_{i,R}^u(q) - \frac{c}{\beta}\sum_{i=0}^{n_e-1}\sigma^i\pi_{i,R}^u(q) > \sum_{i=0}^{n_e-1}\left(V - \frac{(i+1)c}{\mu}\right)\pi_{i,S}^u(q) - \frac{c}{\beta} = U_S^u(q)$, for any given q, where the strict inequality holds because $\sigma < 1$ and $\pi_{0,R}^u(q) < 1$ under a rejection threshold n_e. We next use the normalized utility $\tilde{U}_{R,N}^u(q)$. According to the above inequality, we have $\tilde{U}_{R,N}^u(1) \geq U_{R,N}^u(1) > U_S^u(1)$, whenever $\tilde{U}_{R,N}^u(1) \geq 0$ because the normalization factor of $\tilde{U}_{R,N}^u(1)$ is $\sum_{j=0}^{N-1} \pi_{j,R}^u(1) \in (0,1)$. Also, since $\tilde{U}_{R,N}^u(q)$ decreases in q, the two solutions of $\tilde{U}_{R,N}^u(1) = 0$ and $U_S^u(1) = 0$ must satisfy $\underline{\lambda}_{R,N}^u > \underline{\lambda}_S^u$. Similarly, we have $\tilde{U}_{R,N}^u(0) \geq U_{R,N}^u(0) > U_S^u(0)$, whenever $\tilde{U}_{R,N}^u(0) \geq 0$, so that the two solutions of $\tilde{U}_{R,N}^u(0) = 0$ and $U_S^u(0) = 0$ satisfy that $\bar{\lambda}_{R,N}^u > \bar{\lambda}_S^u$. By Proposition 5 and the order-onsite model, the equilibrium order-placing probability in the OAR model and travel probability in the order-onsite model must satisfy the ordering $q_{R,N}^u \geq q_S^u$. Finally, the steady-state probabilities of $\pi_{i,R}^u(q)$ and $\pi_{i,S}^u(q)$ imply that $\pi_{0,R}^u \leq \pi_{0,S}^u$, showing that the OAR model achieves higher throughput under rejection threshold $N = n_e$. The OAR model under the optimal rejection threshold only achieves even higher throughput and thus, $TH_R^u \geq TH_S^u$.

Next, consider the case of queue-length information being shared remotely. We know that $TH_R^o = TH_A^o$. We further know from Theorem EC.1 that $TH_A^o \geq TH_S^o$ for $n_e = 1$. Hence, $TH_R^o \geq TH_S^o$ for $n_e = 1$. □

Proof of Theorem 8 Recall from Theorem 6 that the service provider has no incentive to share queue-length information in the OAR model. Moreover, when the queue-length information is shared in the OAC model, customers behave the same as they would in the hybrid order-ahead model, which represents a special case of the OAR model with an optimal rejection threshold. Hence, under optimal information, when it is optimal for the OAC model to share information ($TH_C^* = TH_C^o$), we always have $TH_R^* = TH_R^u \geq TH_A^o = TH_C^o$; when it is optimal for the OAC model not to share information ($TH_C^* = TH_C^u$), since $TH_R^* = TH_R^u$, it suffices to prove that $TH_R^u \geq TH_C^u$ when the market size is sufficiently small or large. The rest of the proof focuses on the case where queue-length information is not shared remotely. First, according to Theorem 4, we know that the throughput of the plain order-ahead model dominates that of the OAC model when the market size is small (i.e., when $\Lambda \leq \underline{\lambda}_A^u$). In addition, under the optimal rejection threshold, the OAR model yields higher throughput than the plain order-ahead model. Hence, the OAR model dominates the OAC model when the market size is sufficiently small. When the market size is large, we consider the OAR model with rejection threshold $N = n_e$ and show that this model already yields higher throughput than the OAC model. **Case 1:** If $\bar{U}(n_e - 1) \geq 0$, the joining probability of remote customers is $q_R^u = 1$ in the OAR model (Proposition 5). On the other hand, in the OAC model, if $\Lambda \geq \bar{\lambda}_C^u$, the joining probability of remote customers is $q_C^u = 0$, so that the OAR model is stochastically more congested than the OAC model (to see this, we again invoke Lemma EC.2.2). Hence, OAR yields higher throughput than OAC under a large market size ($\Lambda \geq \bar{\lambda}_C^u$). **Case 2:** Suppose $\bar{U}(n_e - 1) < 0$. Assume that the order-placing probability of remote customers in the OAC model is 0, so that the joining utility for a remote customer is $U_C^u(0) = \sum_{n=0}^{n_e-1} \left(\bar{U}(n) + \frac{c}{\beta}\right) \pi_{n,C}^u(0) + \left(\widetilde{U}(n_e) + \frac{c}{\beta}\right) \pi_{n_e,C}^u(0) - \frac{c}{\beta}$, where the corresponding steady-state probabilities are given by: $\pi_{i,C}^u(0) = \frac{\rho_T^i(1-\rho_T)}{1-\rho_T^{n_e+1}}, \quad i = 0, 1, \cdots, n_e$, where $\rho_T = (1-\gamma)\rho$. On the other hand, the utility of a remote customer in the OAR model is $U_{R,n_e}^u(0) = \sum_{n=0}^{n_e-1} \bar{U}(n)\pi_{n,R}^u(0)$, where the steady-state probability of the number of outstanding orders being i is $\pi_{i,R}^u(0) = \frac{\rho_T^i(1-\rho_T)}{1-\rho_T^{n_e+1}} = \pi_{i,C}^u(0), \quad i = 0, \cdots, n_e$. A straightforward comparison of the above the two utility functions reveals $U_{R,n_e}^u(0) - U_C^u(0) = \pi_{n_e,C}^u(0)\frac{c}{\beta} - \left(\widetilde{U}(n_e) + \frac{c}{\beta}\right) \pi_{n_e,C}^u(0) = -\pi_{n_e,C}^u(0)\widetilde{U}(n_e) > 0$, where the inequality holds because $\widetilde{U}(n_e) < \bar{U}(n_e - 1) < 0$. Consequently, we must have that $\bar{\lambda}_C^u < \bar{\lambda}_R^u$. When the market size $\Lambda \in (\bar{\lambda}_C^u, \bar{\lambda}_R^u)$, the OAR model is stochastically more congested than the OAC model (Lemma EC.2.2), so the former yields higher throughput than the latter. When the market size $\Lambda \geq \bar{\lambda}_R^u$, the order-placing probabilities under the two models are $q_C^u = q_R^u = 0$, which yields identical system throughput. Hence, the OAR model under the rejection threshold n_e already dominates the OAC model by giving higher throughput when the marker size $\Lambda > \bar{\lambda}_C^u$. We conclude that the OAR model has higher throughput than the OAC model in this case. □

Proof of Theorem 9 We first characterize the queueing system for a given rejection threshold N_1 and cancellation threshold N_2. Given remote customers' order-placing probability q, the number of outstanding orders i evolves according to a birth-death process with a state-dependent birth rate $\lambda_i(q)$ and death rate μ_i:

$$\lambda_i(q) = \gamma\Lambda q \cdot \mathbf{1}_{\{i \leq N_1 - 1\}} + (1-\gamma)\Lambda \cdot \mathbf{1}_{\{i \leq n_e - 1\}} \quad \text{and} \quad \mu_i = \mu + \beta(i - N_2)^+, \quad i = 0, 1, \cdots, \tag{EC.4}$$

Given the birth and death rates, the steady-state probability of the number of the outstanding orders X being i, $\hat{\pi}_{i,C}^u(q) \equiv \mathbb{P}(X = i)$, satisfies the balance equations below:

$$[\gamma\Lambda q \cdot \mathbf{1}_{\{i \leq N_1 - 1\}} + (1-\gamma)\Lambda \cdot \mathbf{1}_{\{i \leq n_e - 1\}}]\hat{\pi}_{i,C}^u(q) = [\mu + (i - N_2)^+\beta]\hat{\pi}_{i+1,C}^u(q), \quad i = 0, 1, \cdots, \max\{N_1, n_e\}.$$

Denote X_a as the steady-state number of outstanding orders when a remote customer's order is accepted. Thus, given q, $\mathbb{P}(X_a = i) = \mathbb{P}(X = i | X < N_1) = \hat{\pi}_{i,C}^u(q)/(1 - \hat{\pi}_{N_1,C}^u(q))$. The expected utility of a remote customer who places an order is

$$\begin{aligned} U_{R,C}^u(q) &\equiv \mathbb{E}\left[\left(V - \frac{c \cdot N_{X_a}}{\mu}\right) \mathbf{1}_{\{N_{X_a} < N_2\}}\right] - \frac{c}{\beta} \\ &= \sum_{i=0}^{N_2-1} \bar{U}(i)\hat{\pi}_{i,C}^u(q) + \sum_{i=N_2}^{N_1-1} \left[\sum_{j=0}^{N_2} \left(V - \frac{cj}{\mu}\right) p_i^{R,C}(j) - \frac{c}{\beta}\right] \hat{\pi}_{i,C}^u(q), \end{aligned} \tag{EC.5}$$

where $p_i^{R,C}(j)$ represents the probability distribution of the updated queue position N_{X_a} in the rejection-cancellation system, which is the same as in Lemma 3, except that n_e is replaced by N_2.

We next derive the optimal (N_1, N_2) for a sufficiently small market size Λ. In the integrated scheme: the birth rate $\lambda_i(q) = \gamma\Lambda q \cdot \mathbf{1}_{\{i \leq N_1\}} + (1-\gamma)\Lambda \cdot \mathbf{1}_{\{i \leq n_e - 1\}}$ is maximized when $N_1 = \infty$ and $q = 1$, and the death rate $\mu_i = \mu + \beta(i - N_2)^+$ is minimized when $N_2 = \infty$. Hence, from Lemma EC.2.2, the throughput is indeed maximized by setting $N_1 = N_2 = \infty$ and $q = 1$. On the other hand, According to Proposition 1, when Λ is sufficiently small, the hybrid order-ahead model (with $N_1 = N_2 = \infty$) induces the remote customers' order-placing probability $q_A^u = 1$. Hence, when Λ is sufficiently small, $N_1 = N_2 = \infty$ achieves the maximum throughput.

We next derive the optimal (N_1, N_2) when the market size Λ is sufficiently large. Suppose that the service provider adopts a rejection threshold $N_1 \leq n_e^*$, then cancellation will not kick in since the cancellation threshold satisfies $N_2 \geq n_e$ and the queue position of each remote arrival at the service facility will never exceed n_e^*. Thus, according to Theorem 5, any rejection threshold $N_1 < n_e^*$ is throughput-dominated by rejection threshold n_e^*.

Suppose that the service provider adopts a rejection threshold $N_1 > n_e^*$. We next show that if $N_1 > n_e^*$, then for any $N_2 \geq n_e$, remote customers' order-placing probability $q_{R,N_1}^u = 0$ for sufficiently large Λ. To see this, first, for a birth-death process with birth rates $\{\lambda_i, 0 \leq i \leq N_1 - 1\}$ and death rates $\{\mu_i, 1 \leq i \leq N_1\}$, we know its steady-state probability $\pi_i = \prod_{j=0}^{i-1} \rho_i / (\sum_{k=0}^{N_1} \prod_{l=0}^{k-1} \rho_l)$ for $i = 0, \ldots, N_1$, where $\bar{\rho}_i \equiv \lambda_i / \mu_{i+1}$. We prove by contradiction. Now suppose $q_{R,N_1}^u > 0$, we know from (EC.4) and (EC.5) that as $\Lambda \to \infty$, we have $\rho_i \equiv \lambda_{i-1}(q_{R,N_1}^u)/\mu_i(q_{R,N_1}^u) \to \infty$, and the probability that a remote customer is accepted at state $N_1 - 1$ is

$$\frac{\hat{\pi}_{N_1-1,C}^u(q_{R,N_1}^u)}{1 - \hat{\pi}_{N_1,C}^u(q_{R,N_1}^u)} = \frac{\prod_{j=0}^{N_1-2} \rho_i / (\sum_{k=0}^{N_1} \prod_{l=0}^{k-1} \rho_l)}{1 - \prod_{j=0}^{N_1-1} \rho_i / (\sum_{k=0}^{N_1} \prod_{l=0}^{k-1} \rho_l)} = \frac{\prod_{j=0}^{N_1-2} \rho_i}{1 + \rho_0 + \rho_0\rho_1 + \cdots + \prod_{j=0}^{N_1-2} \rho_i} \to 1,$$

so that

$$\begin{aligned} U_{R,C}^u(q_{R,N_1}^u) &\to \mathbb{E}\left[\left(V - \frac{c \cdot N_{N_1-1}}{\mu}\right) \mathbf{1}_{\{N_{N_1-1} < N_2\}}\right] - \frac{c}{\beta} \\ &= \underbrace{\left(\mathbb{E}\left[\left(V - \frac{c \cdot N_{N_1-1}}{\mu}\right)\right] - \frac{c}{\beta}\right)}_{\bar{U}(N_1-1)} \cdot \mathbb{P}(N_{N_1-1} < N_2) + \left(-\frac{c}{\beta}\right) \cdot \mathbb{P}(N_{N_1-1} \geq N_2). \end{aligned}$$

Because $\bar{U}(N_1 - 1) \leq \bar{U}(n_e^*) < 0$, the above limit is strictly negative regardless of the cancellation threshold N_2. Thus, there must exist a sufficiently large $\bar{\Lambda}_{R,C}$ such that, for any $\Lambda > \bar{\Lambda}_{R,C}$, a remote customer's joining expected utility is negative if $q_{R,N_1}^u > 0$, which leads to a contradiction. Hence, $q_{R,N_1}^u = 0$ for sufficiently large Λ when $N_1 > n_e^*$.

Thus, for sufficiently large Λ, the model with $N_1 > n_e^*$ and the model with $N_1 = n_e^*$ have identical states $i = 0, 1, \cdots, n_e$ and death rates, while the former has smaller birth rates (i.e., $(1-\gamma)\Lambda$) than those of the latter (Λ) at states $i = 0, 1, \cdots, n_e - 1$, thus resulting in lower throughput (according to Lemma EC.2.2).

Based on the above analysis, when the market size is sufficiently large, $N_1 = n_e^*, N_2 = \infty$ is optimal. □

Proof of Proposition 7 We first establish Lemma EC.2.7, where $U_A(\boldsymbol{\lambda}, \beta)$ and $U_S(\boldsymbol{\lambda}, \beta)$ are defined in §EC.1.3.

LEMMA EC.2.7. *Given arrival rate $\boldsymbol{\lambda}$, the utility functions $U_A(\boldsymbol{\lambda}, \beta)$ and $U_S(\boldsymbol{\lambda}, \beta)$ are increasing in β. In addition, $\Delta U(\boldsymbol{\lambda}, \beta) \equiv U_A(\boldsymbol{\lambda}, \beta) - U_S(\boldsymbol{\lambda}, \beta)$ is decreasing in β, and the two utility functions have exactly one intersection.*

Proof of Lemma EC.2.7. Note that $\bar{U}(n)$ increases in β, and $\tilde{\pi}_n^u(\boldsymbol{\lambda})$ is independent of β. It is straightforward to see that both $U_A(\boldsymbol{\lambda}, \beta)$ and $U_S(\boldsymbol{\lambda}, \beta)$ increase in β.

Let N be the steady-state queue length of an $M/M/1$ queue with birth rate $\lambda_i = \Lambda$ if $i < n_e$ and $\lambda_i = \lambda_A$ otherwise, and death rate $\mu_i = \mu$ for $i > 0$. The expected utility of a tagged customer with β who places a remote order is

$$U_A(\boldsymbol{\lambda}, \beta) = V - c\mathbb{E}[\max\{T(\beta), X_{N+1,\mu}\}],$$

where $T(\beta) \sim \text{Exp}(\beta)$ is the random travel time with speed β, and $X_{N+1,\mu}$ denotes an Erlang random variable with $(N+1)$ phases and rate μ. The expected utility of a tagged customer with β who places an onsite order is $U_S(\boldsymbol{\lambda}, \beta) = V\mathbb{P}(N < n_e) - c\mathbb{E}\left[T(\beta) + X_{N+1,\mu} \cdot \mathbf{1}_{\{N<n_e\}}\right]$. The difference between the two utility functions is

$$\triangle U(\boldsymbol{\lambda}, \beta) = V\mathbb{P}(N \geq n_e) - c\mathbb{E}[\max\{0, X_{N+1,\mu} - T(\beta)\} - X_{N+1,\mu} \cdot \mathbf{1}_{\{N<n_e\}}].$$

To show that $\triangle U(\boldsymbol{\lambda}, \beta)$ decreases in β, it suffices to show that the random variable $\max\{0, X_{N+1,\mu} - T(\beta)\}$ stochastically increases in β. To see this, let $\beta_1 \leq \beta_2$; it is straightforward to see that $T(\beta_1) \sim \text{Exp}(\beta_1) \geq_{st} \text{Exp}(\beta_2) \sim T(\beta_2)$.

Next, because $T(\beta) \sim \text{Exp}(\beta)$, we have $\lim_{\beta \to 0} \Delta U(\boldsymbol{\lambda}, \beta) = \infty$, and

$$\begin{aligned} \lim_{\beta \to \infty} \Delta U(\boldsymbol{\lambda}, \beta) &= V\mathbb{P}(N \geq n_e) - c\mathbb{E}[X_{N+1,\mu} - X_{N+1,\mu} \cdot \mathbf{1}_{\{N<n_e\}}] \\ &= V\mathbb{P}(N \geq n_e) - c\mathbb{E}[X_{N+1,\mu} \cdot \mathbf{1}_{\{N \geq n_e\}}] \\ &= \mathbb{E}[(V - c \cdot X_{N+1,\mu}) \cdot \mathbf{1}_{\{N \geq n_e\}}] < 0, \end{aligned}$$

where the inequality holds by the definition of n_e. Hence, the above analysis ensures that there exists only one intersection between utility functions $U_A(\boldsymbol{\lambda},\beta)$ and $U_S(\boldsymbol{\lambda},\beta)$. □

Finishing the proof of Proposition 7. From Lemma EC.2.7, there exist thresholds β_1 and β_2 with $a \le \beta_1 \le \beta_2 \le b$ such that $0 > \max\{U_A(\boldsymbol{\lambda},\beta), U_S(\boldsymbol{\lambda},\beta)\}$ (and hence customers do not order) if $\beta < \beta_1$; $U_A(\boldsymbol{\lambda},\beta) > \max\{0, U_S(\boldsymbol{\lambda},\beta)\}$ (and hence customers order ahead) if $\beta \in (\beta_1,\beta_2)$; $U_S(\boldsymbol{\lambda},\beta) > \max\{0, U_A(\boldsymbol{\lambda},\beta)\}$ (and hence customers order onsite) if $\beta > \beta_2$. □

Proof of Theorem 10 We prove the throughput dominance under rejection threshold n_e, which would imply throughput dominance under the optimal rejection threshold.

In the OAR model, let the arrival rate for remote customers who choose to order ahead and order onsite be λ_A and λ_S respectively, with $\lambda_A + \lambda_S \le \gamma\Lambda$; let the arrival rate for local customers be $\lambda_L = (1-\gamma)\Lambda$. The order queue evolves as a birth-death process with state-dependent birth rate $\lambda_i = (\lambda_A + \lambda_S + \lambda_L)\cdot \mathbf{1}_{\{i<n_e\}}$ and death rate $\mu_i = \mu$ for $i > 0$. The steady-state probability of the number of outstanding orders being $i \le n_e$ is $\hat{\pi}_i^u(\lambda_A + \lambda_S + \lambda_L) = (1-(\lambda_A+\lambda_S+\lambda_L)/\mu)((\lambda_A+\lambda_S+\lambda_L)/\mu)^i/(1-(\lambda_A+\lambda_S+\lambda_L)^{n_e+1})$. Let the expected utility for a remote customer who orders ahead be U_A^R and that for a remote customer who orders onsite be U_S^R:

$$U_A^R(\lambda_A,\lambda_S,\lambda_L,\beta) = \sum_{i=0}^{n_e-1}\left(V - \frac{(i+1)c}{\mu} - \frac{c}{\beta}\sigma^{i+1}\right)\hat{\pi}_i^u(\lambda_A+\lambda_S+\lambda_L),$$

$$U_S^R(\lambda_A,\lambda_S,\lambda_L,\beta) = \sum_{i=0}^{n_e-1}\left(V - \frac{(i+1)c}{\mu}\right)\hat{\pi}_i^u(\lambda_A+\lambda_S+\lambda_L) - \frac{c}{\beta},$$

where $\sigma = \mu/(\beta+\mu)$. It is straightforward that $U_A^R(\lambda_A,\lambda_S,\lambda_L,\beta) \ge U_S^R(\lambda_A,\lambda_S,\lambda_L,\beta)$. Thus, no remote customers place onsite orders ($\lambda_S = 0$) in equilibrium. Moreover, since $U_A^R(\lambda_A,\lambda_S,\lambda_L,\beta)$ increases in β, there exists a threshold for travel speed $\bar{\beta}_A$ that uniquely solves $U_A^R(\lambda_A,0,\lambda_L,\bar{\beta}_A) = \sum_{i=0}^{n_e-1}\left(V - \frac{(i+1)c}{\mu} - \frac{c}{\bar{\beta}_A}\sigma^{i+1}\right)\hat{\pi}_i^u(\lambda_A+\lambda_L) = 0$ with $\lambda_A = \gamma\Lambda(1-F(\bar{\beta}_A))$ such that customers place orders (ahead) if and only if $\beta > \bar{\beta}_A$.

In the order-onsite model, let the arrival rate for remote customers be $\lambda'_S \le \gamma\Lambda$ and the arrival rate for local customers be $\lambda_L = (1-\gamma)\Lambda$. The order queue state is a birth-death process with state-dependent birth rate $\lambda_i = (\lambda'_S + \lambda_L)\cdot \mathbf{1}_{\{i<n_e\}}$ and death rate $\mu_i = \mu$ for $i > 0$. The corresponding steady-state probability of the number of outstanding orders being $i \le n_e$ is $\hat{\pi}_i^u(\lambda'_S+\lambda_L) = (1-(\lambda'_S+\lambda_L)/\mu)((\lambda'_S+\lambda_L)/\mu)^i/(1-(\lambda'_S+\lambda_L)^{n_e+1})$. Thus, the expected utility of a joining remote customer with travel speed β is $U_S^S(\lambda'_S,\lambda_L,\beta) = \sum_{i=0}^{n_e-1}\left(V - \frac{(i+1)c}{\mu}\right)\hat{\pi}_i^u(\lambda'_S+\lambda_L) - \frac{c}{\beta}$. It is straightforward to see that $U_S^S(\lambda'_S,\lambda_L,\beta)$ increases in β. Hence, remote customers join when their travel speed β exceeds a threshold $\bar{\beta}_S$, which uniquely solves $U_S^S(\lambda'_S,\lambda_L,\bar{\beta}_S) = \sum_{i=0}^{n_e-1}\left(V - \frac{(i+1)c}{\mu}\right)\hat{\pi}_i^u(\lambda'_S+\lambda_L) - \frac{c}{\bar{\beta}_S} = 0$ with $\lambda'_S = \gamma\Lambda(1-F(\bar{\beta}_S))$.

Next, we prove $\lambda_A > \lambda'_S$, i.e., $\bar{\beta}_A < \bar{\beta}_S$. Note that $0 = U_A^R(\lambda_A,0,\lambda_L,\bar{\beta}_A) = \sum_{i=0}^{n_e-1}\left(V - \frac{(i+1)c}{\mu} - \frac{c}{\bar{\beta}_A}\sigma^{i+1}\right)\hat{\pi}_i^u(\lambda_A + \lambda_L) > \sum_{i=0}^{n_e-1}\left(V - \frac{(i+1)c}{\mu}\right)\hat{\pi}_i^u(\gamma\Lambda(1-F(\bar{\beta}_A))+\lambda_L) - \frac{c}{\bar{\beta}_A} = U_S^S(\gamma\Lambda(1-F(\bar{\beta}_A)),\lambda_L,\bar{\beta}_A)$. Since $U_S^S(\gamma\Lambda(1-F(\beta)),\lambda_L,\beta)$ is increasing in β and $U_S^S(\gamma\Lambda(1-F(\bar{\beta}_A)),\lambda_L,\bar{\beta}_A) < 0$ and $U_S^S(\gamma\Lambda(1-F(\bar{\beta}_S)),\lambda_L,\bar{\beta}_S) = 0$, we have $\bar{\beta}_A < \bar{\beta}_S$. Hence, $\lambda_A > \lambda'_S$, which implies that $\hat{\pi}_0(\lambda_A+\lambda_L) < \hat{\pi}_0(\lambda'_S+\lambda_L)$. Therefore, the throughput of the OAR model with rejection threshold n_e surpasses that of the order-onsite model. The throughput of the OAR model under the optimal rejection threshold will only be even higher. □