Prioritization in the Presence of Self-ordering Opportunities in Omni-channel Services

(Authors' names blinded for peer review)

Problem definition: Motivated by the popularity of mobile-order-and-pay applications, especially in fast-casual food restaurants and coffee shops, we study omni-channel service systems—where customers can employ mobile applications for self-ordering—with respect to sojourn times, throughput, and social welfare. **Methodology/results:** Our models are two-stage queues with two customer classes: walk-ins and mobiles. We identify Pareto efficient prioritization policies, highlighting trade-offs between each class's mean sojourn times. We allow customers to make strategic joining decisions based on their anticipated delays under an information structure where walk-ins observe partial queue length information. We draw from a wide array of techniques, including steady-state, transient, busy period, hitting-time analyses, and matrix analytic methods. We showcase the significance of prioritization on the system throughput and social welfare. We demonstrate settings where the (typically beneficial) transformation of a traditional service system to an omni-channel reduces throughput. **Managerial implications:** Our analysis highlights the importance of prioritization policy choice for an efficient transition to an omni-channel service system. The throughput-optimal policy choice is highly dependent on the operational parameters and on customer patience levels; implementing a wrong policy can yield a significant loss in throughput and, hence, profitability.

Key words: Omni-channel services; Self-ordering technology; Strategic queueing; Prioritization

1. Introduction

Millions spend time waiting for services at coffee shops, government offices, and medical clinics every day. Recent developments in mobile technologies aim to improve customers' waiting time experience. For example, in some fast-food restaurants and coffee shops, customers can use their mobile phones to place online orders and pay in advance, effectively skipping the in-store ordering line. The usage of such applications has been growing steadily. For example, the fraction of transactions conducted via Starbucks's Mobile Order & Pay application increased from 4% in 2016 to 24% in 2020 (Campbell 2020).

Despite the potential advantages of *self-processing*, introducing these mobile applications has also caused complications. Reports of "long lines that are being exacerbated by an uptick in mobile ordering... [that are causing] customers to walk out" at Starbucks (Ryan 2017) illustrate the need for proper system design to mitigate throughput loss due to unsatisfied customers. Leveraging detailed queueing models and analyses, we fulfill this

need by highlighting task prioritization as a crucial operational design lever impacting system throughput and customers' waiting experience in *omni-channel* services, in which customers can employ mobile applications for self-ordering.

We take the coffee shop as the paradigmatic case of an omni-channel service system, which has two *stages*: a customer waits in line to place and pay for their order and then waits for the order to be prepared. The major brands, including Starbucks, Dunkin' Donuts, and McDonald's, have launched online ordering applications in recent years, enabling *mobile customers* (*mobiles* for short)—those who use the application—to make and pay for their selections, skipping the cashier line and only waiting for preparation. Meanwhile, *walk-in customers* (*walk-ins* for short)—those who cannot or choose not to use the application—must first wait to place their order.

Under this omni-channel paradigm, the staff preparing orders for mobiles often take orders from—and prepare orders for—walk-ins; i.e., the service capacity is shared between both channels. Mobiles bypass the first stage by processing their ordering and payment tasks, which reduces their waiting times and frees up some service capacity. These benefits could result in reduced total service requirements, lower waiting times (potentially for walk-ins and mobiles), and possibly higher profits. It is crucial to note that this omni-channel paradigm is distinct from its long-existing predecessor, whereby customers can call in an order (e.g., pizza); the latter does not involve self-processing as the phone call engages an employee's time, reducing their ability to attend to other duties.

Introducing mobile self-processing applications, however, could result in inferior customer satisfaction despite the mentioned benefits, eventually leading to throughput and revenue loss (Ryan 2017). We show that part of these inefficiencies stems from higher task prioritization complications (compared to the single-channel services). The introduction of self-processing applications splits the homogeneous pool of customers (with respect to service requirements) into two *classes* (walk-ins and mobiles) with distinct service flows. In this case, an essential system design choice is how to prioritize the orders from the two customer classes. Popular and easy-to-implement service policies (e.g., the first-come-first-served (FCFS) policy) might not correctly differentiate the walk-ins' and mobiles' distinct service requirements and waiting time expectations.

We capture the complicated stochasticity in omni-channel services by modeling them as two-stage tandem queueing networks under single- and two-server settings (§3). We identify

and analyze Pareto efficient prioritization policies (with respect to the *class-specific mean sojourn times* of walk-ins and mobiles) in the case of *non-strategic* customers; we show that they generate the entire Pareto frontier (§4). Then, we allow *strategic* customers to join or balk based on their anticipated delays (§5). The challenge here is that walk-ins observe the first stage's state (based on which they draw inferences about the second stage) when constructing their delay anticipation, while mobiles observe nothing. To address this challenge, we draw from various techniques, including steady-state, transient, busy period, and hitting-time analyses and matrix analytic methods. We showcase the significance of the prioritization policy choice on system throughput and social welfare (§6).

We find a clear trend where throughput increases, on the *aggregate level*, with the rate at which customers adopt mobile ordering technology. However, such throughput gains are not homogeneous and heavily rely on implementing the optimal prioritization policy. Furthermore, such throughout gains are not universally achievable. We locate diverse settings where transforming to an omni-channel service reduces throughput, even under the optimal prioritization policy (among those we study). This observation runs counter to both the intuition on the benefits of offering a more efficient service stream and insights generated by some recent work on omni-channel services, which celebrate the advantages that the introduction of the mobile channel offers.

Our findings are driven by explicitly modeling previously abstracted queueing-theoretic and information-structural features of omni-channel services. One such crucial feature is the availability of self-service opportunities for mobile customers. In the absence of customers' strategic behavior, self-service opportunities can sometimes present challenges for service providers seeking to prioritize customers so that customers can anticipate the same end-to-end delay regardless of their channel. In our single-server model, imposing such strict "fairness" constraints comes at the cost of suboptimal mean sojourn times and, in some parameter settings, comes at the cost of strategically idling the server, resulting in artificial delays for at least some mobile customers. Meanwhile, when customers exhibit strategic behavior, the operational advantages of self-service opportunities (i.e., service requirement reductions) are not always sufficient to overcome inefficiencies introduced by information uncertainty, leading to degraded throughput and/or social welfare.

2. Literature Review

Methodologically, our work draws from several research streams. In exploring the Pareto efficient prioritization policies with respect to class-specific mean sojourn times, we take inspiration from the achievable regions methods developed in Bertsimas (1995) and further articulated in Dacre et al. (1999). While our single-server models can also be interpreted as polling models, our objectives, design choices, and analytic techniques are mainly unrelated to those found in the polling literature, as surveyed in Boon et al. (2011) and Borst and Boxma (2018). In terms of strategic customer behavior, we are indebted to Naor's classical paper on the subject (Naor 1969) and the long tradition of work on queueing games that it has inspired, as surveyed in Hassin and Haviv (2003) and Hassin (2016).

In our strategic models, walk-ins observe only the queue length in the first stage and infer a distribution on the second stage's queue length when deciding to join. Similarly D'Auria and Kanta (2015), Kim and Kim (2016), Kerner et al. (2017), and Ji and Roet-Green (2020) present models where arrivals make joining decisions while observing only partial queue-length information. In these papers, the unobserved information is a random variable with finite support. In our work, the support is unbounded; hence, we must contend with an infinite state space, necessitating distinct analytic techniques. One feature of our single-server model—the server alternation between the two stages—is shared with the model studied by Nimrod et al. (2020); however, our model differs significantly in that their work renders both queues unobservable. Most significantly, our model differs from those featured in the aforementioned papers in that we consider an *omni-channel* two-class system; the papers cited above study single-class single-channel systems.

The analytical modeling of omni-channel retailing has received significant attention from various aspects (examples include Chopra 2016, Gao and Su 2016, Bayram and Cesaret 2017, Gallino et al. 2017, Gao and Su 2017, Bell et al. 2018, Jin et al. 2018, Delasay et al. 2021). However, the queueing-theoretic study of omni-channel services remains in its infancy. In the remainder of this section, we discuss several related papers. To the best of our knowledge, these papers—together with ours—constitute the entirety of the analytical work on omni-channel services to date.

Gao and Su (2018) investigate the high-level impact of self-processing technologies on capacity planning (i.e., staffing). While—like our work—they model omni-channel service systems as tandem queues, in their model, they consider an unobservable queueing setting.

Consequently—unlike our work—the technical contributions of the paper are not queueing-theoretic. Under this model, Gao and Su find that self-ordering not only reduces mean sojourn times for those customers who opt to use the mobile service (as expected) but could also improve the mean sojourn time experienced by walk-ins. Although Gao and Su endogenize the arrival rate as a function of the waiting time, it does not explicitly model customers as rational. Considering rational customers in omni-channel services has been the primary focus of several recent papers (including ours). We discuss the three papers in this area most closely related to ours below. Baron et al. (2021) study customers' channel choice in a single-stage FCFS omni-channel system, and show that offering an online ordering channel will increase in the system throughput; this increase comes at the cost of a drop in social welfare due to the resulting information uncertainty. However, they find that prioritizing walk-ins can overcome social welfare loss. Our paper complements this line of investigation by highlighting prioritization as a primary design choice for an efficient transition from single-channel to omni-channel (although we show that such a transition is not always possible). Moreover, much of our paper is dedicated to addressing what Baron et al. (2021) identify as "an intriguing question and a promising future research direction." Namely, a model where "walk-in customers are aware of the availability of the online channel but only observe the physical queue... [which] increases the analysis complexity of walk-in customers' joining decisions."

Roet-Green and Yuan (2020) study omni-channel services in a way that can also be thought of as addressing the "intriguing question" posed in Baron et al. (2021). They treat *information settings*—in terms of the level of system occupancy observability—as the primary design choice. By contrast, our work treats *prioritization policies* as the primary design choice. Each approach is capable of obtaining fundamentally different insights. Furthermore, in all of their information settings, mobiles are also privy to some system state information. This induces a threshold joining behavior on the part of mobiles and thus yields finite state spaces, which—together with their restricted focus on single-stage models—precludes their need for much of the sophisticated queueing-theoretic analysis that forms an integral part of our paper. These differences have salient consequences: e.g., they prove that their model yields unique equilibria, whereas we find many cases where our model gives rise to multiple equilibria. Roet-Green and Yuan express interest in exploring

models of heterogeneous customers' patience levels; we explore such heterogeneity as it pertains to our models in Appendix EC.4.7.

Ghosh et al. (2020) explore a discrete-time model, primarily addressing the phenomenon of channel choice (like Baron et al. 2021). Their work considers some additional features (e.g., not all customers are given the opportunity to choose their channel) and also—as in the work of Roet-Green and Yuan (2020)—explores more than one information setting (either mobiles have full, but delayed, information, or no information at all). Unlike the models in Baron et al. (2021) and Roet-Green and Yuan (2020), the extra features present in the models considered in Ghosh et al. (2020) lead to the existence of settings where system throughput under an omni-channel structure falls below that of a single-channel system. In that respect, they draw conclusions that match ours, despite emphasizing different design aspects of omni-channel services. An important contribution of Ghosh et al. (2020) is the study of the possibility of quality degradation during a mobile customer's travel time. This feature connects the paper to another stream of research on omni-channel services with rational customers that focus on issues associated with travel (examples less closely related to our work include Baron et al. (2020) and Liu and Yang (2020)).

Among the papers studying omni-channel services in the presence of rational customer behavior discussed above, only our work considers a two-stage tandem queueing system. This consideration allows our models to capture the reduction in the need for service capacity when processing mobile customers due to their ability to self-order. Studying a partially observable two-stage queueing system under various prioritization policies introduces the need for substantial queueing analysis, which constitutes one of the key contributions of our paper. Together with the three papers discussed above, we view our paper as providing valuable complementary perspectives on the various quintessential features of omni-channel service systems. We posit that considering all perspectives at once allows one to grasp the bigger picture better than taking each perspective in isolation. That said, it may be infeasible to analyze a single model that fully incorporates and exhaustively explores all of these features (e.g., service requirement reduction from self-ordering, prioritization design, information design, channel choice, travel time, etc.) simultaneously, justifying the need for any given study to emphasize some of these features over others.

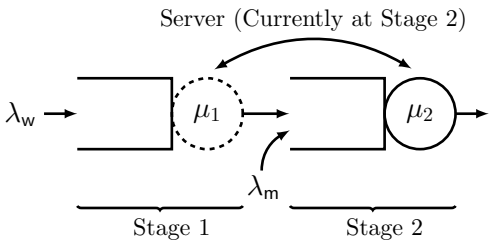


Figure 1 **Single-server model (server at Stage 2)**

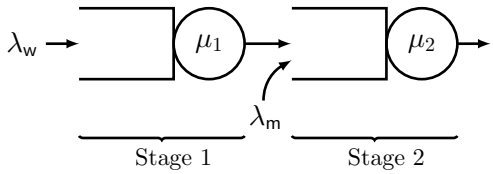


Figure 2 **Two-server model**

3. Model

We consider a family of queueing systems with two *service stages* and two *customer classes*. Each service stage consists of an infinite buffer queue. *Walk-ins* arrive to *Stage 1* according to a Poisson process with rate λ_w and proceed to *Stage 2* upon service completion at Stage 1. Meanwhile, *mobiles* bypass Stage 1 and arrive directly to Stage 2 according to a Poisson process with rate λ_m; let $\Lambda = \lambda_w + \lambda_m$ denote the total arrival rate to the system. A walk-in's (resp., mobile's) sojourn time, T_w (resp., T_m), is the duration of time from the moment of arrival to Stage 1 (resp., Stage 2) until the completion of service at Stage 2. We emphasize that while only walk-ins can be present at Stage 1, customers of both classes can be simultaneously present at Stage 2. For tractability, we assume that all service requirements are independent and exponentially distributed with rates μ_1 and μ_2 at Stages 1 and 2, respectively.

We consider two models: (i) In our *single-server* model (see Fig. 1), a single *flexible* server moves between the two stages instantaneously to serve customers according to a prioritization policy (which we define in §3.1). (ii) In our *two-server* model (see Fig. 2), each stage is served by its *dedicated* (inflexible) server; while the Stage 1 server serves walk-ins at Stage 1 in their arrival order, the Stage 2 server can make service order decisions; e.g., they could prioritize mobiles ahead of walk-ins. The single-server model allows us to highlight the sojourn time trade-offs between the walk-ins and mobiles, while the two-server model allows us to test the generalizability of our insights.

In a coffee shop setting, we can think of each walk-in as beginning their sojourn when they arrive at a physical waiting line (Stage 1) leading to a cashier who takes orders, while each mobile begins their sojourn as soon as they place their order via an app. A barista (who is also the cashier in the single-server model) prepares food and beverages from a virtual queue of orders (Stage 2) placed by walk-ins and mobiles. In practice, mobiles can

often place their orders before arriving at the coffee shop, in which case their travel time overlaps with their sojourn time. For simplicity, our models do not distinguish between the experience of traveling and that of waiting in the coffee shop (see Liu and Yang (2020) and Baron et al. (2020) for detailed models on this effect). We also do not allow for the possibility of a walk-in deciding to switch to using the app after their arrival.

3.1. Prioritization Policy Structure

At any time, the flexible server in the single-server model must choose at which stage to work. Furthermore, a server at Stage 2 (the flexible server in the single-server model or the Stage 2 dedicated server in the two-server model) must choose which customer class to serve. To this end, we introduce the notion of prioritization policies that dictate whom the flexible (or Stage 2 dedicated) server must serve at any time. It is helpful to add further granularity in how we view customers by breaking up each walk-in's service into two *tasks*. At any given time, each customer's service belongs to one of three *task classes*: walk-in tasks at Stage 1 (**O**s), walk-in tasks at Stage 2 (**W**s), and mobile tasks at Stage 2 (**M**s).

In the single-server model, we use the convention MWO, for example, to represent a specific *work-conserving preemptive class-based priority policy* in which the flexible server prioritizes tasks in the following order: (1) **M**s (mobiles), (2) **W**s (walk-ins in Stage 2), and (3) **O**s (walk-ins in Stage 1). We can construct $3! = 6$ policies by permuting the three task classes. We use a similar convention in the two-server model: Noting that Stage 1's dedicated server can only serve **O**s (and **O**s can only be served by this server), we only need to consider the relative prioritization between **W**s and **M**s at Stage 2. This yields only two work-conserving preemptive class-based priority policies: MW (where **M**s are prioritized) and WM (where **W**s are prioritized).

The families of policies discussed above are not exhaustive. Other feasible policies include those that are not work-conserving, non-preemptive policies, randomized mixtures of other policies, and policies that give two or more classes an equal priority. We note that in the single-server model, much of our work extends to non-preemptive policies with modest modifications to our analytic contributions. Still, we restrict attention to preemptive policies in the interest of brevity.

Given any policy P, we are primarily interested in the *class-specific mean (equivalently, expected) sojourn times*, $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}]$ and $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}]$, that emerge under that policy in *steady state*. We facilitate steady-state analysis by making the following assumption:

ASSUMPTION 1. *The parameters Λ, λ_{w}, μ_1, and μ_2 must ensure system stability; i.e., (a) $\lambda_{\mathsf{w}}/\mu_1 + \Lambda/\mu_2 < 1$ for the single-server model, and (b) $\lambda_{\mathsf{w}} < \mu_1$ and $\Lambda < \mu_2$ for the two-server model.*

3.2. Customer Behavior and Information Structure

We consider delay-sensitive customers who join (resp., balk) if their anticipated expected sojourn time upon arrival is less (resp., more) than their *patience levels* $T_{\mathsf{w}}^{\max}$ for walk-ins and $T_{\mathsf{m}}^{\max}$ for mobiles; they are *indifferent* when their anticipation matches their patience level exactly. We consider *homogeneous* patience levels within each customer class (i.e., $T_{\mathsf{w}}^{\max}$ and $T_{\mathsf{m}}^{\max}$ are constants). However, our approach and insights largely generalize to the heterogeneous case (see Appendix EC.4.7 for details). When anticipating their expected sojourn times, customers indirectly take into account the prioritization policy: they have become accustomed to the policy's steady-state mean sojourn time, e.g., from experience or word-of-mouth.

Walk-ins joining behavior. Walk-ins observe the number of customers in Stage 1, N_1, upon arrival, motivated by the fact that a customer walking into a coffee shop sees how many have lined up to place orders but cannot see how many outstanding orders are currently awaiting preparation (or which Stage the server is currently serving). While walk-ins cannot observe the number of customers in Stage 2, N_2, upon arrival, their observation of N_1 allows for inference on N_2.

In light of the above, under policy P, a walk-in joins if their *conditional expected sojourn time* is no greater than their patience level ($\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}|N_1 = i] \leq T_{\mathsf{w}}^{\max}$). This gives rise to a threshold, b, whereby walk-ins join if they observe $N_1 < b$ and balk otherwise; consequently, b acts as a finite buffer size for Stage 1. Here, we simplify exposition by implicitly considering that all indifferent walk-ins join. We can interpret our model so that the behavior of walk-ins fits within the standard framework of rational queueing for risk-neutral customers with linear waiting-time costs (see, e.g., Naor 1969); i.e., if we assume that walk-ins obtain a benefit R from receiving service and experience a cost C per unit of sojourn time, then they would behave exactly as described above if $T_{\mathsf{w}}^{\max} = R/C$.

Mobiles joining behavior. Unlike walk-ins, mobiles enter the system observing nothing: they place their order online before being present to witness the system occupancy. While hypothetically, a mobile application could provide real-time delay estimates, we do not

consider such features in our model. We concur with the following assessment of this issue provided in Baron et al. (2021): "The invisibility of the online channel also reflects industry practice. To the best of our knowledge, no omnichannel service provider offers real-time queue length information to online customers ... Yet, some providers, e.g., Starbucks, quote expected waiting times to online customers." Even in the absence of such announcements, mobiles can still behave strategically by employing a *mixed joining strategy*. Specifically, under prioritization policy P, each mobile joins with probability p_{m} (independently of other mobiles) and balks otherwise, where p_{m} is the highest probability for which $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}] \leq T_{\mathsf{m}}^{\max}$.

Strategy profiles. Based on the discussion above, the joining behavior of all customers is described by the *strategy profile* (b, p_{m}), where walk-ins join if and only if they observe $N_1 < b$ upon arrival and mobiles join with probability p_{m}. For any $b \in \mathbb{Z}_{\geq 0}$ and $p_{\mathsf{m}} \in [0,1]$, the strategy profile (b, p_{m}) results in a well-defined queueing system; we are most interested in *equilibrium* strategy profiles, i.e., consistent with the joining behavior outlined above (see §5 for details). For example, if (b, p_{m}) is an equilibrium, then $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}|N_1 = i] \leq T_{\mathsf{w}}^{\max}$ for all $i \in \{0, 1, \ldots, b-1\}$. However, the notation used in expressing the walk-in's expected sojourn time obfuscates a vital subtlety: $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}|N_1 = i]$ can depend on b and p_{m}. To this end, we write $\mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}$ and $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}$ to denote the probability and expectation operators, respectively, under the strategy profile (b, p_{m}) and priority policy P.

3.3. Throughput, Overall Mean Sojourn Time, Social Welfare

The *throughput rate* of walk-ins (resp., mobiles), χ_{w} (resp., χ_{m}), is the rate at which walk-ins (resp. mobiles) are served. When patience levels are infinite (i.e., $T_{\mathsf{w}}^{\max} = T_{\mathsf{m}}^{\max} = \infty$), we have throughput rates $\chi_{\mathsf{w}} = \lambda_{\mathsf{w}}$ and $\chi_{\mathsf{m}} = \lambda_{\mathsf{m}}$; otherwise, under the strategy profile (b, p_{m}) and priority policy P, we have $\chi_{\mathsf{w}} = \lambda_{\mathsf{w}} \mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(N_1 < b)$ and $\chi_{\mathsf{m}} = \lambda_{\mathsf{m}} p_{\mathsf{m}}$. We measure the *overall throughput rate* as $X \equiv \chi_{\mathsf{w}} + \chi_{\mathsf{m}}$, which can serve as a proxy for revenue if walk-ins and mobiles pay the same average price for service. The *overall mean sojourn time* is given by $\mathbb{E}^{\mathsf{P}}[T] \equiv \left(\lambda_{\mathsf{w}} \mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \lambda_{\mathsf{m}} \mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}]\right) / \Lambda$ when $T_{\mathsf{w}}^{\max} = T_{\mathsf{m}}^{\max} = \infty$, and is given by $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T] \equiv \left(\chi_{\mathsf{w}} \mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i] + \chi_{\mathsf{m}} \mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}]\right) / X$ when customers are strategic.

When customers are finitely patient, we define the *social welfare*—denoted by $\mathrm{SW}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}$ under strategy profile (b, p_{m}) and policy P—as the mean surplus experienced across all customers, where a customer's surplus is their patience level less their sojourn time (0 if they balk). Our definition corresponds to the standard one in the rational queueing

literature under the normalization that sets the waiting cost rate equal to one ($C = 1$). Conditioning appropriately, we have:

$$\mathrm{SW}^{\mathsf{P}}_{(b,p_{\mathsf{m}})} = \frac{\lambda_{\mathsf{w}}}{\Lambda} \sum_{i=0}^{b-1} \left(T^{\max}_{\mathsf{w}} - \mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}} | N_1 = i]\right) \mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(N_1 = i) + \frac{p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\Lambda} \left(T^{\max}_{\mathsf{m}} - \mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}]\right).$$

Sections 4 and 5 analyze the cases where customers have infinite and finite patience, respectively. In the infinite patience case ($T^{\max}_{\mathsf{w}} = T^{\max}_{\mathsf{m}} = \infty$), since all customers join (i.e., $(b, p_{\mathsf{m}}) = (\infty, 1)$), the primary metrics of interest are the class-specific and overall mean sojourn times. Meanwhile, in the finite patience case ($T^{\max}_{\mathsf{w}}, T^{\max}_{\mathsf{m}} < \infty$), we are most interested in the equilibrium throughput and social welfare values, requiring the computation of expected sojourn times.

4. Analysis: The Case of Customers with Infinite Patience

Customers always join when they have infinite patience, i.e., patience levels $T^{\max}_{\mathsf{w}} = T^{\max}_{\mathsf{m}} = \infty$. Thus, we do not need to consider any strategic joining decision on their part. Assumption 1 guarantees that the system is both stable and throughput-optimal under any work-conserving policy. In this setting, we aim to understand the trade-offs associated with helping one customer class over the other via prioritization in terms of the mean sojourn time experienced by each class. We assume throughout this section that $\lambda_{\mathsf{w}}, \lambda_{\mathsf{m}} > 0$.

We formalize our discussion of tradeoffs by writing $\mathcal{P}$ to denote the set of all possible policies P. For any $\mathsf{P} \in \mathcal{P}$, we write $a^{\mathsf{P}} \equiv (\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}], \mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}])$ to denote the *allocation* (i.e., the pair of class-specific mean sojourn times) under policy P; we write $\mathcal{O} \equiv \{a^{\mathsf{P}} \in \mathbb{R}^2_+ : \mathsf{P} \in \mathcal{P}\}$ to denote the *achievable region* of allocations. Given two policies P and P', we say that a customer class is "better off" under policy P as opposed to P' if the class experiences a *lower* mean sojourn time under P; if one class is better off under P and the other is *not* better off under P', then we say that P *dominates* P', writing $a^{\mathsf{P}} \succ a^{\mathsf{P}'}$. The relation '$\succ$' induces partial orders on both $\mathcal{P}$ and $\mathcal{O}$. We call a policy P *Pareto optimal*—writing $\mathsf{P} \in \mathcal{P}^*$—if no other policy dominates it; equivalently, in symbols: $(\mathsf{P} \in \mathcal{P}^*) \equiv (\nexists \mathsf{P}' \in \mathcal{P} : a^{\mathsf{P}'} \succ a^{\mathsf{P}})$. We call the set of allocations yielded by Pareto optimal policies the *Pareto frontier*, $\mathcal{O}^* \equiv \{a^{\mathsf{P}} : \mathsf{P} \in \mathcal{P}^*\}$.

4.1. Pareto Optimal Policies

Typically, we are interested in selecting a policy P that minimizes some function of the class-specific sojourn times (e.g., the overall mean sojourn time, either class's mean sojourn time, or any weighted average of these) that is *strictly monotone* with respect to the

ordering on allocations induces by the '$\succ$' relation. Consequently, a system designer seeking to minimize such a function need only consider Pareto-optimal policies.

We proceed by observing that given any pair of policies $\mathsf{P}, \mathsf{P}' \in \mathcal{P}$, we can construct a family of policies $\{\langle \mathsf{P}, \mathsf{P}' \rangle(\theta) \colon \theta \in [0,1]\} \subseteq \mathcal{P}$ parameterized by θ, where $\langle \mathsf{P}, \mathsf{P}' \rangle(\theta)$ is implemented as follows: at the start of each busy period, we choose to use either P or P' for the remainder of the busy period with probabilities θ and $1-\theta$ (independent of past choices), respectively. It follows that for any work-conserving policies P and P', $a^{\langle \mathsf{P}, \mathsf{P}' \rangle(\theta)} = \theta a^{\mathsf{P}} + (1-\theta) a^{\mathsf{P}'} \in \mathcal{O}$. Extending this reasoning can establish the achievability of convex combinations of achievable allocations. We identify several Pareto optimal policies (in both the single- and two-server models). We show that these policy sets can generate all other Pareto optimal policies (i.e., the entire Pareto frontier for the corresponding model) through random mixtures of the kind described above; we call such policies *Pareto generators*. Formally, for a given model, a set of Pareto optimal policies $\mathcal{G} \subseteq \mathcal{P}^*$ form a set of Pareto generators if the Pareto frontier $\mathcal{O}^*$ satisfies $\mathcal{O}^* \subseteq \operatorname{conv}\left(\left\{a^{\mathsf{P}} \colon \mathsf{P} \in \mathcal{G}\right\}\right)$.

Recall our notation for *work-conserving preemptive class-based priority policies* from §3.1, where, for example, MWO denotes the policy that prioritizes **M**s (mobiles) ahead of **W**s (walk-ins at Stage 2) and **W**s ahead of **O**s (walk-ins at Stage 1). Of the six policies in the *single-server model*, three—MOW, OMW, and OWM—prioritize **O**s over **W**s; it is straightforward to show that none of these three policies are Pareto optimal, so we disregard them, focusing instead on MWO, WMO, and WOM. Meanwhile, in the *two-server model*, the only relevant prioritization is between the two Stage 2 task classes, **M**s and **W**s (as the Stage 1 dedicated server only servers **O**s), yielding the MW and WM policies. We explicitly compute a^{P} under all five of these policies (see Appendix EC.1.1 for a presentation and proof of these results) and leverage these computations to establish that these policies are Pareto generators for their respective models.

Proposition 1 *The set* $\{\mathsf{MWO}, \mathsf{WMO}, \mathsf{WOM}\}$ *and the set* $\{\mathsf{MW}, \mathsf{WM}\}$ *form a set of Pareto generators for the single- and two-server models, respectively.*

We also study a third policy in the two-server model that (our informal inspection suggests) is the norm in practice: FCFS (first-come-first-served) gives the **M**s and **W**s equal priority and serves them in the order in which they enter Stage 2. The computation of a^{FCFS} (also presented in Appendix EC.1.1) yields the following result:

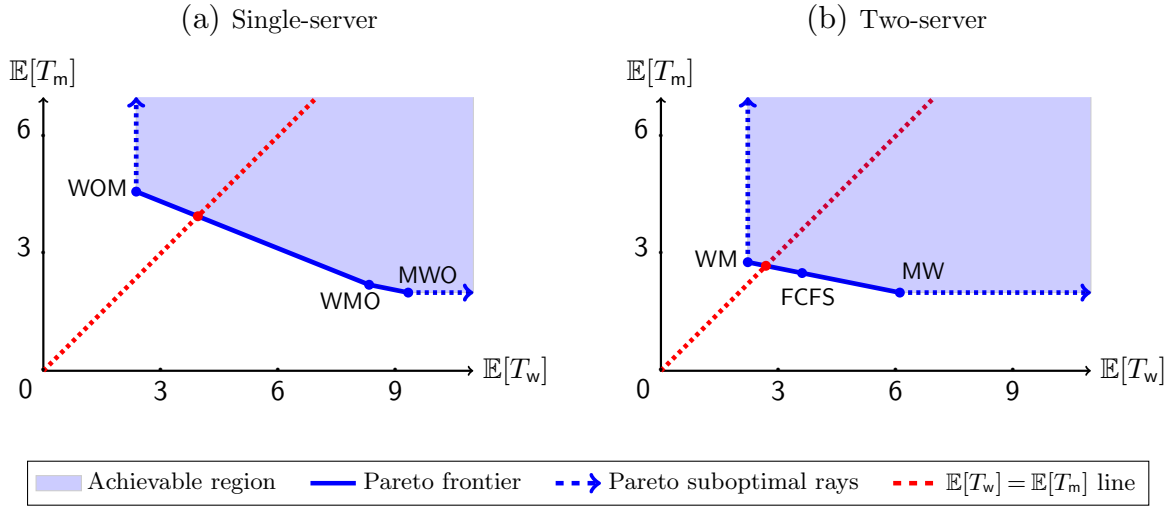


Figure 3 **Examples of the achievable region for the parameter setting** $\lambda_w = 0.1$, $\lambda_m = 0.5$, $\mu_1 = \mu_2 = 1$.

Proposition 2 *In the two-server model, we have* $a^{\mathsf{FCFS}} = \theta a^{\mathsf{MW}} + (1-\theta)a^{\mathsf{WM}}$ *where* $\theta = (\mu_2 - \lambda_m)/(2\mu_2 - \Lambda)$ *and* $\mathsf{FCFS} \in \mathcal{P}^*$.

As a result of this proposition, we can treat FCFS as an "extraneous" Pareto generator. That is, $\{\mathsf{MW}, \mathsf{FCFS}, \mathsf{WM}\}$ also forms a set of Pareto generators.

4.2. Fairness

Our performance analysis also allows us to address questions of fairness between the two customer classes. While many notions of fairness exist, one strict and straightforward definition would be to require all customers to experience the same sojourn time in expectation regardless of their class, i.e., P is "fair" (in this sense) if $\mathbb{E}[T_m] = \mathbb{E}[T_w]$. We facilitate our discussion of fairness by way of illustrations. Fig. 3 shows the achievable region and Pareto frontier for two example instances (one for each single- and two-server model). It follows from arguments presented in our proof of Proposition 1 that these illustrative examples are "representative" of what we would encounter in all instances. Specifically, in the single-server model, WOM, WMO, and MWO are connected (from the "northwest" to "southeast" in that order) by a pair of line segments, where the latter segment is steeper than the former. Meanwhile, in the case of the two-server model, FCFS will lie on the line segment running from WM southeast to MW, consistent with Proposition 2.

Now observe that in either model, a policy P is "fair" in the sense defined above if a^{P} lies on the line where $\mathbb{E}[T_w] = \mathbb{E}[T_m]$ (the red dotted line in Fig. 3). Furthermore, this line crosses

the boundary of the achievable region exactly once, so there is at most one Pareto optimal fair allocation. Moreover, it follows from our expressions for the allocations under WMO and FCFS (see Appendix EC.1.1) that $\mathbb{E}\left[T_{\mathsf{m}}^{\mathsf{WMO}}\right] < \mathbb{E}\left[T_{\mathsf{w}}^{\mathsf{WMO}}\right]$ and $\mathbb{E}\left[T_{\mathsf{m}}^{\mathsf{FCFS}}\right] < \mathbb{E}\left[T_{\mathsf{w}}^{\mathsf{FCFS}}\right]$ in the single- and two-server models, respectively. Consequently, in the single-server model, the $\mathbb{E}[T_{\mathsf{w}}] = \mathbb{E}[T_{\mathsf{m}}]$ line crosses the boundary of the achievable region on either (i) the line segment connecting WOM and WMO (including possibly at point a^{WOM}) or (ii) the ray extending "north" of WOM, i.e., $\{(0,y) + a^{\mathsf{WOM}} : y > 0\}$. An analogous observation—where we replace WOM and WMO by WM and FCFS, respectively—holds for the two-server model. In Fig. 3, case (i) holds for both policies (i.e., the "fairness" line crosses the Pareto frontier), however, one can easily find examples for either model where case (ii) holds. Under case (i), there is a unique Pareto optimal fair policy; whereas under case (ii), any fair policy is Pareto-suboptimal, i.e., fairness comes at the steep cost of strategically idling the server, resulting in artificial delays for at least some mobiles.

In the single-server model, fairness always comes with a price as it is inefficient (from the overall system perspective) to prioritize walk-ins in Stage 1 over mobiles in Stage 2 even under case (i). This is because the $\mathbb{E}[T_{\mathsf{w}}] = \mathbb{E}[T_{\mathsf{m}}]$ line crosses the line segment connecting a^{WOM} and a^{WMO} and this line segment descends with a slope of a greater magnitude than the line segment connecting a^{WMO} and a^{WOM}; only the points on the latter line segment are optimal with respect to overall mean sojourn time, and hence, the unique fair Pareto optimal policy is suboptimal in this respect. By contrast, in the two-server model, the unique Pareto policy is optimal with respect to the overall mean sojourn time. These observations are a consequence of the following general result:

Proposition 3 *(a) In the single-server model, a work-conserving prioritization policy minimizes the overall mean response time if it preemptively prioritizes customers in Stage 2 (i.e.,* **W***s and* **M***s) over those in Stage 1 (i.e.,* **O***s); consequently,* WMO *and* MWO *are optimal. Meanwhile,* WOM *is suboptimal with respect to the overall mean sojourn time despite being Pareto optimal. (b) In the two-server model, a work-conserving prioritization policy is optimal with respect to the overall mean response time if it is Pareto optimal.*

This section highlighted how each customer class affects the other. The "interaction" between the two classes becomes more complicated once we consider strategic behavior on the part of customers with finite patience levels, which is the focus of the next two sections.

5. Analysis: The Case of Customers with Finite Patience

In the setting where customers have finite patience levels, i.e., $T_{\mathsf{w}}^{\max}, T_{\mathsf{m}}^{\max} < \infty$, we need to consider customers' strategic joining behavior. We are interested in the equilibrium strategy profiles that emerge under policies we identified as Pareto generators in the infinite patience case (see §4). Recall that walk-ins observe the current Stage 1 occupancy N_1 upon arrival, while mobiles observe nothing. As implied by our choice of notation, each of $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i]$ and $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}]$ can depend on both b and p_{m}. Hence, given a policy P, we seek to find an equilibrium of the form (b^*, p^*_{m}) where (i) b^* is the equilibrium threshold such that $\mathbb{E}^{\mathsf{P}}_{(b^*,p^*_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i] \le T_{\mathsf{w}}^{\max}$ whenever $N_1 = i \le b^*$ and (ii) p^*_{m} is the highest probability with which mobiles can join while ensuring that $E^{\mathsf{P}}_{(b^*,p_{\mathsf{m}})}[T_{\mathsf{m}}] \le T_{\mathsf{m}}^{\max}$. Formalizing, we have the following necessary and sufficient conditions on equilibrium (b^*, p^*_{m}):

$$\begin{aligned} \mathbb{E}^{\mathsf{P}}_{(b^*,p^*_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i] &\le T_{\mathsf{w}}^{\max}, \qquad \forall i \in \{0, 1, \ldots, b^* - 1\}, \\ \mathbb{E}^{\mathsf{P}}_{(b^*,p^*_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = b^*] &> T_{\mathsf{w}}^{\max}, \\ \arg\max\{p_{\mathsf{m}} \in [0,1] \colon \mathbb{E}^{\mathsf{P}}_{(b^*,p_{\mathsf{m}})}[T_{\mathsf{m}}] \le T_{\mathsf{m}}^{\max}\} &= p^*_{\mathsf{m}}, \end{aligned}$$

where $\arg\max\{\emptyset\} \equiv 0$. While assumption 1 guarantees that $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i] < \infty$ and $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] < \infty$ for all policies P under consideration, $b \in \mathbb{Z}_{\ge 0}$, and $p_{\mathsf{m}} \in [0,1]$, we will see in §6 that neither the uniqueness nor the existence of equilibria is guaranteed.

5.1. Determining Equilibria in the Finite Patience Model

We proceed by discussing our method of finding equilibria, which applies to both the single- and two-server models with minimal differences. The method requires one to obtain $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i]$ and $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}]$. For now, we assume these expressions are given, deferring their derivations for the single- and two-server models to §§5.2 and 5.3, respectively. The following proposition simplifies the process of searching for equilibria by limiting the candidate values of threshold b and establishing that there exists a mobile joining probability p_{m} that is a "best response" to any threshold b.

Proposition 4 *For any fixed threshold b and any $\mathsf{P} \in \{\mathsf{MWO}, \mathsf{WMO}, \mathsf{WOM}\}$ (in the single-server model) or $\mathsf{P} \in \{\mathsf{MW}, \mathsf{WM}, \mathsf{FCFS}\}$ (in the two-server model):*

(a) If we take p_{m} to be a value such that (b, p_{m}) is an equilibrium under P, then the threshold $b < B \equiv \mu_1(T_{\mathsf{w}}^{\max} - 1/\mu_2)$.

(b) When we view $p_{\mathsf{m}} \in [0,1]$ as a variable, the expected sojourn time of mobiles $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}]$ is strictly increasing in p_{m}.

Proposition 4(a) simplifies the process of searching for an equilibrium threshold b, requiring us to consider only finitely many cases, $b \in \{0,1,\ldots,B-1\}$ (for all six policies of interest). Given a policy P, for each possible $b \in \{0,1,\ldots,B-1\}$ (where the bound $B \equiv \mu_1(T^{\max}_{\mathsf{w}} - 1/\mu_2)$ or some better bound if available), we compute

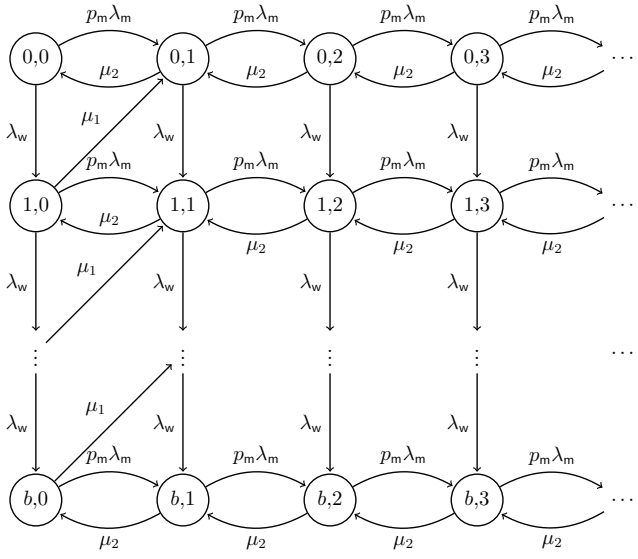
$$p_{\mathsf{m}}(b) \equiv \sup\{p_{\mathsf{m}} \in [0,1] \colon \mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] \le T^{\max}_{\mathsf{m}}\},$$

where $\sup\{\emptyset\} \equiv 0$. Meanwhile, Proposition 4(b) (together with the continuity of the mobiles' mean sojourn time in p_{m}) guarantees the existence of $p_{\mathsf{m}}(b)$. Specifically, $p_{\mathsf{m}}(b) = 1$ if $\mathbb{E}^{\mathsf{P}}_{(b,1)}[T_{\mathsf{m}}] \le T^{\max}_{\mathsf{m}}$, $p_{\mathsf{m}}(b) = 0$ if $\mathbb{E}^{\mathsf{P}}_{(b,0)}[T_{\mathsf{m}}] > T^{\max}_{\mathsf{m}}$, and $p_{\mathsf{m}}(b) = f_b^{-1}(T^{\max}_{\mathsf{m}})$ (letting the function $f_b(\cdot) \equiv \mathbb{E}^{\mathsf{P}}_{(b,\cdot)}[T_{\mathsf{m}}]$) in any other case. While $f_b^{-1}(T^{\max}_{\mathsf{m}})$ is well-defined, it may not be possible to compute it exactly, in which case we can resort to arbitrarily accurate numerical inversion techniques (e.g., the bisection method). Finally, we must check whether each $(b, p_{\mathsf{m}}(b))$ pair is an equilibrium; this is the case if and only if $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}}(b))}[T_{\mathsf{w}}|N_1 = i] \le T^{\max}_{\mathsf{w}}$, for each $i \in \{0,1,\ldots,b-1\}$), and $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}}(b))}[T_{\mathsf{w}}|N_1 = b] > T^{\max}_{\mathsf{w}}$.

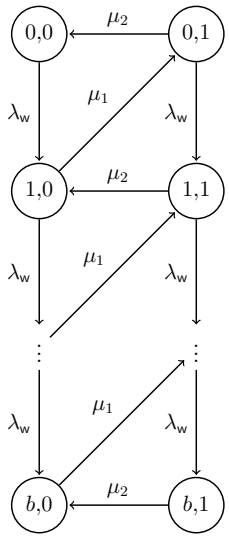
To complete our analysis, we obtain $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i]$ and $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}]$ for the single- and two-server policies of interest in §§5.2 and 5.3, respectively. We begin each discussion with an examination of the continuous-time Markov chain (CTMC) governing (N_1, N_2) and/or $(N_1, N_{2,\mathsf{w}})$, where $N_{2,\mathsf{w}}$ is the number of **W** tasks in Stage 2. In particular, we must find the steady-state limiting probability distributions of these chains, which we denote by $\pi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j) \equiv \mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(N_1 = i, N_2 = j)$ and $\phi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j) \equiv \mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(N_1 = i, N_{2,\mathsf{w}} = j)$ for any policy P and strategy profile (b, p_{m}).

5.2. Single-Server Finite Patience Model: Mean Sojourn Times

We derive *exact* expressions for the mean sojourn times in the single-server model under policies $\mathsf{MWO}, \mathsf{WMO}$, and WOM by analyzing their underlying CTMCs. We begin with MWO and WMO, under which $(N_1, N_2) \in \{0,1,\ldots,b\} \times \mathbb{Z}_{\ge 0}$ evolves according to the same CTMC (Fig. 4a). We use the limiting probabilities of this CTMC ($\pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j) \equiv \pi^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}(i,j)$) in terms of infinite series later to derive the mean sojourn times. For any specified value of $b \in \{0,1,\ldots,B-1\}$, these limiting probabilities—and hence, the expected sojourn times of interest—can be determined in closed form (see Appendix EC.3.1).



(a) (N_1, N_2) under MWO and WMO



(b) $(N_1, N_{2,w})$ under WOM

Figure 4 **Single-server CTMCs. In (a), N_1 increases by 1 with rate λ_w due to a walk-in arrival when $N_1 < b$; N_2 increases by 1 with rate $p_m\lambda_m$ due to a mobile arrival; N_1 decreases by 1 and N_2 increases from 0 to 1 with rate μ_1 due to a Stage 1 service when $N_1 > N_2 = 0$; N_2 decreases by 1 with rate μ_2 due to a Stage 2 departure when $N_2 > 0$. In (b), N_1 increases by 1 with rate λ_w due to a walk-in arrival when $N_1 < b$; N_1 decreases by 1 and N_2 increases from 0 to 1 with rate μ_1 due to a Stage 1 service when $N_1 > N_2 = 0$; N_2 decreases from 1 to 0 with rate μ_2 due to a Stage 2 departure when $N_2 = 1$.**

To find the mean sojourn times under WOM, rather than analyzing the CTMC governing (N_1, N_2), we analyze a chain with state variables $(N_1, N_{2,w}) \in \{0, 1, \ldots, b\} \times \{0, 1\}$ where $N_{2,w}$ is the number of **W**s in Stage 2. Note that under WOM, **O**s (i.e., walk-ins in Stage 1) receive service only when there are no **W**s in the system. Moreover, once a walk-in's **O** task completes service at Stage 1, their **W** task arrives to Stage 2 and immediately receives the highest priority, entering service, and precluding the service of any **O**s until its service completion. Hence, there can be at most one **W** in the system at any given time under WOM, resulting in the finite-state CTMC illustrated in Fig. 4b. The chain's finite state space allows for the straightforward determination of its exact limiting probabilities, $\phi^{\mathsf{WOM}}_{(b,p_m)}(i,j)$ (see Appendix EC.3.2). In the special case where $b = 0$, we have a degenerate chain where $\phi^{\mathsf{WOM}}_{(b,p_m)}(0,0) = 1$. In any case, the limiting probabilities allow us to express the conditional expected sojourn time $\mathbb{E}^{\mathsf{WOM}}_{(b,p_m)}[T_w | N_1 = i]$.

On the other hand, the $\phi^{\mathsf{WOM}}_{(b,p_m)}(i,j)$ values do not immediately lend themselves to determining $\mathbb{E}^{\mathsf{WOM}}_{(b,p_m)}[T_m]$. Instead, we express $\mathbb{E}^{\mathsf{WOM}}_{(b,p_m)}[T_m]$ in terms of the first and second moments

of two *hitting time* random variables, U and V, which depend on (b, p_{m}) (for the computation of these moments—which can be found in closed-form for any specified value of b—see Appendix EC.3.3): U represents the *waiting time* of a mobile (i.e., the duration from arrival time until service begins) who arrives when there are no other mobiles in the system, while V represents the *sojourn time* of a mobile who enters an empty system.

Carrying out the analysis described above for all three policies of interest, we obtain all of the desired expected sojourn times in the following proposition:

Proposition 5 *Under* MWO, WMO, *and* WOM *in the single-server model, we have*

$$\begin{cases} \mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] = \dfrac{1}{\mu_2 - p_{\mathsf{m}}\lambda_{\mathsf{m}}} \\ \mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i] = \left(\left(\dfrac{\mu_2}{\mu_1} + 1\right)(i+1) + \displaystyle\sum_{j=0}^{\infty} j\pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j) \Big/ \sum_{j=0}^{\infty} \pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j)\right) \mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] \end{cases}, \tag{1}$$

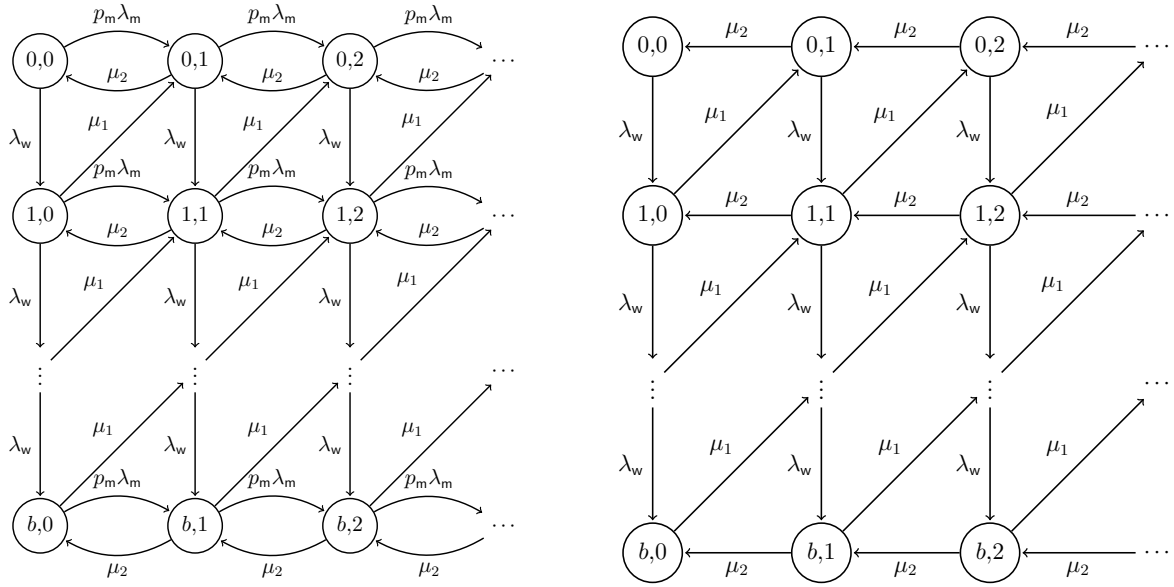
$$\begin{cases} \mathbb{E}^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] = \dfrac{1}{\mu_2}\left(1 + \displaystyle\sum_{i=0}^{b}\sum_{j=0}^{\infty} j\pi^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}(i,j)\right) \\ \mathbb{E}^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i] = \dfrac{1}{\mu_2} - \mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] + \mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i] \end{cases}, \tag{2}$$

$$\begin{cases} \mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i] = (i+1)\left(\dfrac{1}{\mu_1} + \dfrac{1}{\mu_2}\right) + \dfrac{\phi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,1)}{\mu_2\left(\phi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,1) + \phi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,0)\right)} \\ \mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] = \mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[V] + \dfrac{p_{\mathsf{m}}\lambda_{\mathsf{m}}\mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[V^2]}{2\left(1 - p_{\mathsf{m}}\lambda_{\mathsf{m}}\mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[V]\right)} + \dfrac{2\mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[U] + p_{\mathsf{m}}\lambda_{\mathsf{m}}\mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[U^2]}{2\left(1 + p_{\mathsf{m}}\lambda_{\mathsf{m}}\mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[U]\right)} \end{cases}. \tag{3}$$

The results presented in Proposition 5 can then be used to determine the equilibria of the form (b^*, p^*_{m}) in the single-server model under all three prioritization policies of interest.

5.3. Two-Server Finite Patience Model: Mean Sojourn Times

We proceed to seek expressions for the appropriate expected sojourn times in the two-server model—again with the ultimate goal of determining equilibria of the form (b^*, p^*_{m}). Determining such expected sojourn times for the two-server model will often necessitate analyzing intractable infinite-state CTMCs and computing infinite sums over recursively defined quantities. Consequently, unlike in the single-server model, the expected sojourn times in the two-server model cannot generally be expressed in closed form (with $\mathbb{E}^{\mathsf{MW}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}]$ being a notable exception). While we provide exact expressions for all sojourn times of interest, these expressions will be in terms of auxiliary quantities (e.g., infinite sums of limiting probabilities) that cannot be determined exactly; we provide methods for approximating these quantities throughout Appendix EC.3.



(a) (N_1, N_2) under any two-server policy (b) $(N_1, N_{2,w})$ under WM

Figure 5 **Two-server CTMCs. In (a) N_1 increases by 1 with rate λ_w due to a walk-in arrival when $N_1 < b$; N_2 increases by 1 with rate $p_m\lambda_m$ due to a mobile arrival; N_1 decreases by 1 and N_2 increases by 1 with rate μ_1 due to a walk-in service at Stage 1 when $N_1 > 0$; N_2 decreases by 1 with rate μ_2 due to a Stage 2 departure when $N_2 > 0$. The CTMC in (b) corresponds to that in (a) when $p_m = 0$.**

Under the two-server policies—MW, WM, and FCFS—the system occupancy $(N_1, N_2) \in \{0, 1, \ldots, b\} \times \mathbb{Z}_{\geq 0}$ evolves according to the Fig. 5a CTMC. Our analysis requires the limiting probabilities $\pi^{\mathsf{TS}}_{(b,p_m)}(i,j)$ of this CTMC (where TS stands for our three policies of interest in the T*wo*-S*erver* model), which can be approximated with arbitrary accuracy (see Appendix EC.3.4).

Prioritization plays a less critical role in the two-server model, as it only affects Stage 2 tasks. However, in this model, service can be provided at both stages simultaneously; this complicates system dynamics, leading to significant analytic challenges. For example, consider the FCFS policy: a tagged walk-in must infer the distribution of N_2 based on the observed value of N_1 upon arrival. Even if the tagged walk-in knows $N_2 = j$ with certainty when they arrive, by the time they finally reach Stage 2, the occupancy there may have varied significantly from j due to arrivals and departures. Hence, the tagged walk-in's conditional expected sojourn time is $\mathbb{E}[T_w | N_1 = i] = (i+1)/\mu_1 + Y(i,j)$, where $Y(i,j)$ is the *expected workload* that the tagged walk-in will encounter at Stage 2 *once it arrives there*, given that they initially observed $N_1 = i$ and $N_2 = j$ when first arriving at Stage 1. By

the workload at Stage 2, we mean the amount of time needed to clear all Stage 2 tasks—including the tagged walk-in's task—assuming no further arrivals to Stage 2. Determining the expected workload $Y(i,j)$ requires *transient* queueing analysis while determining the distribution of N_2 conditioned on $N_1 = i$ requires *steady-state* analysis. To allow for transient analysis, let $\{M_\rho(t)\}_{t\geq 0}$ denote the number of customers in an M/M/1 system under load $\rho \in (0,\infty)$ at time t and $\{t_n\}_{n\geq 1}$ denote the time of the n-th Poisson arrival to this system since time 0. Now consider Definition 1, adapted from Kaczynski et al. (2012):

Definition 1 *For integers $u \geq 0$, $v \geq 1$, and $w \in \{1,2,\ldots,u+v\}$, let the probability $P(u,v,w;\rho) \equiv \mathbb{P}(M_\rho(t_v) = w | M_\rho(0) = u)$; i.e., $P(u,v,w;\rho)$ is the probability that the system occupancy of an M/M/1 system under load $\rho > 0$ transitions from u to w after exactly v further arrivals.*

Lemma 1 expresses $Y(i,j)$ exactly in terms of infinite sums of these probabilities, which allows for $Y(i,j)$—and further infinite sums expressed in terms of $Y(i,j)$—to be approximated by using sum truncation together with a recursive method presented Kaczynski et al. (2012) for computing the $P(u,v,w;\rho)$ exactly (see Appendix EC.3.5 for details).

Lemma 1 *If a walk-in joins a two-server system when $(N_1,N_2) = (i,j)$, the expected Stage 2 workload upon arrival of this customer to Stage 2 (including the customer's own Stage 2 service requirement) under any work-conserving policy is given by:*

$$Y(i,j) = \left(\frac{\mu_1}{\mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}}}\right)^{i+1} \sum_{k=0}^{\infty} \sum_{\ell=1}^{i+j+k+1} \frac{\ell}{\mu_2} P\left(j, i+k+1, \ell; \frac{\mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_2}\right) \binom{k+i}{k} \left(\frac{p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}}}\right)^k. \quad (4)$$

The probabilities $P(u,v,w;\rho)$ are also instrumental in deriving the mean sojourn times under the WM policy. Under WM, (N_1,N_2) is again governed by the Fig. 5a CTMC, with limiting probabilities $\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)$. However, in this case (as in the case of WOM in the single-server model), we are also interested in the limiting probabilities of the CTMC governed by $(N_1,N_{2,\mathsf{w}}) \in \{0,1,\ldots,b\} \times \mathbb{Z}_{\geq 0}$, which we can approximate with arbitrary accuracy (see Appendix EC.3.7). This CTMC is depicted in Fig. 5b. We also need the expectation of the "hitting time" random variable $Z(i,j)$, which represents the time it takes to reach a state where $N_{2,\mathsf{w}} = 0$ from state $(N_1,N_{2,\mathsf{w}}) = (i,j)$ under WM, given the strategy profile (b,p_{m}), i.e., $Z(i,j) \sim \inf\{s \geq 0 \colon N_{2,\mathsf{w}}(t+s) = 0 | N_1(t) = i, N_{2,\mathsf{w}}(t) = j\}$ for all $t \geq 0$. Details on approximating $\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,j)]$ with arbitrary precision are given in Appendix EC.3.8.

Carrying out performance analysis for all three policies of interest in the two-server setting, we obtain the following results for the sojourn times of interest in terms of the problem parameters and sums involving $P(u,v,w;\rho)$, $Y(i,j)$, and/or $\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,j)]$.

Proposition 6 *Under* MW, FCFS, *and* WM *in the two-server model, we have*

$$\begin{cases} \mathbb{E}^{\mathsf{MW}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i] = \dfrac{i+1}{\mu_1} + \dfrac{1}{1-p_{\mathsf{m}}\lambda_{\mathsf{m}}/\mu_2}\displaystyle\sum_{j=0}^{\infty} Y(i,j)\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j) \Big/ \sum_{j=0}^{\infty}\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j) \\ \mathbb{E}^{\mathsf{MW}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] = \dfrac{1}{\mu_2 - p_{\mathsf{m}}\lambda_{\mathsf{m}}} \end{cases}, \tag{5}$$

$$\begin{cases} \mathbb{E}^{\mathsf{FCFS}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i] = \dfrac{i+1}{\mu_1} + \displaystyle\sum_{j=0}^{\infty} Y(i,j)\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j) \Big/ \sum_{j=0}^{\infty}\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j) \\ \mathbb{E}^{\mathsf{FCFS}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] = \dfrac{1}{\mu_2}\left(1 + \displaystyle\sum_{i=0}^{b}\sum_{j=0}^{\infty} j\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)\right) \end{cases}, \tag{6}$$

$$\begin{cases} \mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i] = \dfrac{i+1}{\mu_1} + \displaystyle\sum_{j=0}^{\infty}\sum_{\ell=1}^{i+j+1} \frac{\ell}{\mu_2} P\left(j, i+1, \ell; \frac{\mu_1}{\mu_2}\right)\phi^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}(i,j) \Big/ \sum_{j=0}^{\infty}\phi^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}(i,j) \\ \mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] = \displaystyle\sum_{i=0}^{b}\sum_{j=0}^{\infty}\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,j+1)]\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j) \end{cases}. \tag{7}$$

6. Results and Insights

This section employs our equilibrium determination methodology in the case of finite patience levels (outlined in §5) to explore the impact of our two service design choices on throughput and social welfare: (1) whether to offer a mobile ordering option and if so (2) the prioritization policy to be implemented. We investigate what happens if a single-channel walk-in only system transitions to an omni-channel system when an exogenous fraction $\alpha \in [0,1]$ of customers "adopt" the new technology once the app is introduced, switching from being walk-ins to mobiles (i.e., $\lambda_{\mathsf{w}} = (1-\alpha)\Lambda$ and $\lambda_{\mathsf{m}} = \alpha\Lambda$). We examine what occurs under the new steady-state equilibrium resulting from the adoption of the app. In reality, some new customers who were previously uninterested in the single-channel system may also adopt the service (allowing for an increase in Λ); while we do not consider this possibility in the interest of brevity, we can study such scenarios using the same methods by setting λ_{w} and λ_{m} to any desired values.

6.1. Illustration of the Adoption Rate Impact

This section demonstrates the possible impact of the adoption rate α on the normalized throughput rate X/Λ and social welfare $\mathrm{SW}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}$ using an illustrative problem instance. In this problem, we consider a single-server system in which walk-ins are less patient ($T^{\max}_{\mathsf{w}} = 5.2 < T^{\max}_{\mathsf{m}} = 8$). For illustration, we generate the plots in Figs. 6a and 6b, by computing

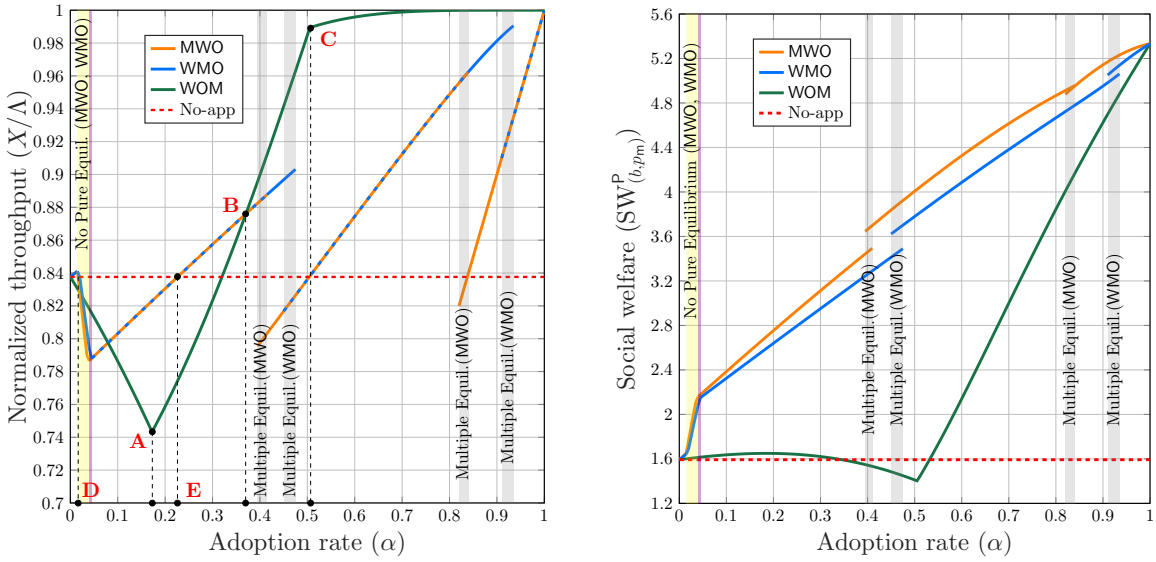


Figure 6 **Single-server model:** $\Lambda = 0.625$, $\mu_1 = 2$, $\mu_2 = 1$, $T_w^{\max} = 5.2$, $T_m^{\max} = 8$.

equilibria and the resulting metrics for the adoption rate $\alpha \in \{0, 0.005, \ldots, 0.995, 1\}$. At some α values, there are multiple equilibria for a policy (the gray regions in the figures). Furthermore, at some α values, no pure strategy on the part of walk-ins yields an equilibrium (the yellow regions); therefore, we plot the metrics associated with an equilibrium featuring a mixed strategy on walk-ins where the assumption that indifferent walk-ins join is relaxed. That is, walk-ins will randomly choose whether or not to join with some probability $\mathbf{p}_w$ (which is part of the description of their strategy) when they observe $b-1$ other customers in the Stage 1 queue upon their arrival (see Appendix EC.4.1).

We first discuss the impact of the adoption rate on throughput (Fig. 6a). We observe that although WMO always performs at least as well as MWO, the two policies yield the same throughput at most adoption rates (orange-blue dashed curves) due to the discrete nature of the walk-ins' equilibrium threshold b^*. As a higher mobile adoption alleviates the load at Stage 1—and because we are considering a case where mobiles are more patient—unsurprisingly, higher α *tends to* improve the throughput. Nevertheless, α's increase may trigger a discontinuous drop in throughput due to a shift of size one in b^*. As WOM aggressively favors walk-ins, mobiles do not join until the adoption rate crosses a point **A**, making λ_w low enough to allow a fraction of mobiles to join using a mixed strategy, $p_m \in$

$(0,1)$. At adoption rates beyond point **B**, enough mobiles opt to join such that WOM outperforms the other two policies with respect to throughput. Finally, when the adoption rate is beyond a threshold (point **C**), all mobiles join ($p_{\mathsf{m}} = 1$) as service interruptions due to walk-ins become sufficiently infrequent.

As expected, there is little throughput benefit in offering the app when the adoption rate is very low (below point **D**). Most surprisingly, the no-app benchmark outperforms all three policies for α between points **D** and **E**—even though mobiles are more patient in this setting—which suggests that the operational advantages from self-ordering technology (i.e., service requirement reductions) can be insufficient to overcome inefficiencies introduced by *information uncertainty*, which can be in two forms: (i) walk-ins have uncertainty regarding the queue length at Stage 2 once mobiles are also using the system, and (ii) mobiles lack access to queue length information that they would have had if they were walk-ins. Hence, *the omni-channel structure is not always beneficial*. We will revisit the information uncertainty issue in greater detail at the end of this subsection.

Turning our attention to social welfare (Fig. 6b), we observe that MWO and WMO outperform the no-app benchmark for the vast majority of α values. Social welfare tends to increase with α partially because the average system-wide patience level also increases with α (because $T_{\mathsf{m}}^{\max} > T_{\mathsf{w}}^{\max}$ in this problem instance). On the other hand, aggressive prioritization of walk-ins (i.e., WOM) often yields considerably lower social welfare than the no-app benchmark; by prioritizing walk-ins—who have higher service requirements than mobiles—WOM yields relatively poor mean sojourn times and hence lower social welfare.

We can observe qualitatively similar phenomena in the two-server model. Specifically, Fig. 7b shows that in a system with equal patience levels for walk-ins and mobiles, it is even possible for all three policies to underperform the no-app benchmark with respect to social welfare (even at $\alpha = 1$). The dominance of the no-app at the adoption rate $\alpha = 1$ is initially counter-intuitive: moving from a customer base of all walk-ins (no-app) to one of all mobiles ($\alpha = 1$)—who require less service, are equally patient, and equally numerous—may be expected to generate *higher* social welfare. It turns out that in the "all walk-in" (no-app) case, about 20% (see the dashed red line in Fig. 7a where $X/\Lambda \approx 0.8$) of customers balk. This throughput inefficiency in the no-app case has a beneficial side effect of reducing congestion and hence expected sojourn times. As a result, despite (in fact, *because of*) the lower throughput in the no-app case, reduced congestion allows the *average* customer to

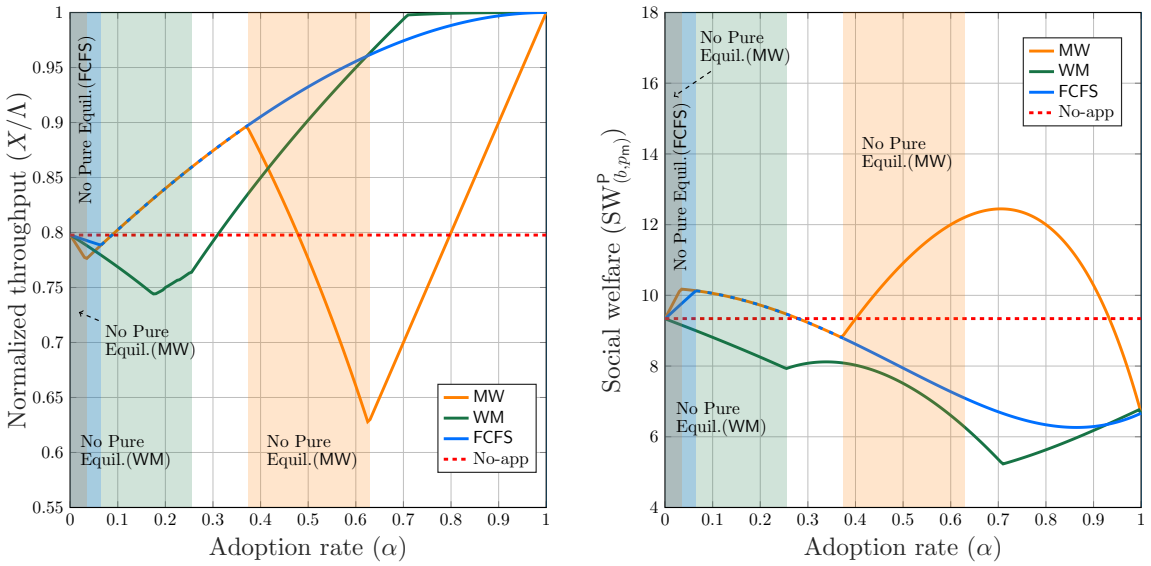


Figure 7 **Two-server model:** $\Lambda = 0.05$, $\mu_1 = 0.16$, $\mu_2 = 0.08$, $T_{\mathsf{w}}^{\max} = T_{\mathsf{m}}^{\max} = 40$.

experience a greater surplus (i.e., contribution to social welfare) than what they would experience in the omni-channel system at some α values, including the "all mobile" system. Apart from an intermediate region where α is roughly between 28%–40%, the policy that prioritizes mobiles (i.e., MW) performs very well with respect to social welfare until $\alpha >$ 90%, where the congestion effect sharply increases the overall mean sojourn time. These results suggest a rich space of trade-offs between throughput and social welfare.

6.2. Discussion of Information Uncertainty

This section numerically illustrates the impact of information uncertainty on walk-in and mobile customers' joining decisions as the adoption rate varies. We focus on the single-server model (as we can compute all expected sojourn times exactly) and consider the impact of information uncertainty on a tagged customer's decision—and the impact on the throughput—while fixing the behavior of all other customers in equilibrium.

We can classify customers' decisions into four categories: (i) those who join given the information they observed at the time of arrival, who would have joined anyway had they observed the full system state $N_1 + N_2$ at the time of their arrival, (ii) those who join but would not have done so with full information, (iii) those who balked and would have done the same with full information, and (iv) those who balked but would not have done

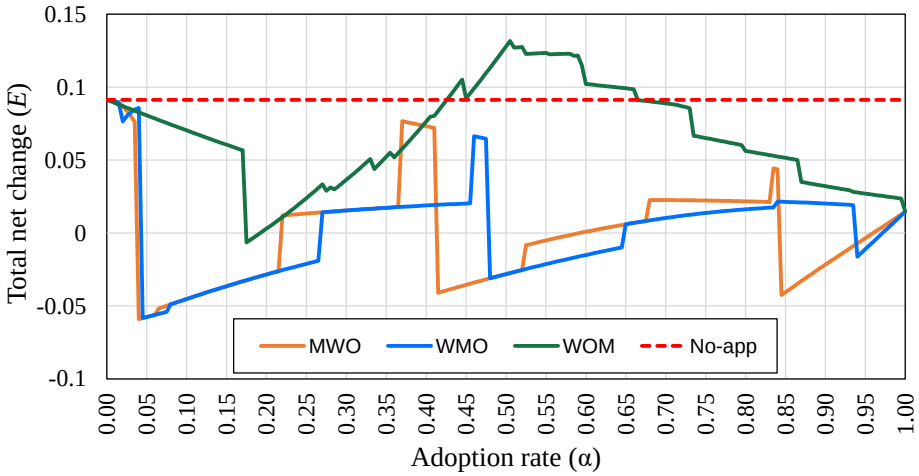


Figure 8 **Total net change in throughput due to information uncertainty under single-server model:** $\Lambda = 0.625$, $\mu_1 = 2$, $\mu_2 = 1$, $T_{\mathsf{w}}^{\max} = 5.2$, $T_{\mathsf{m}}^{\max} = 8$.

so with full information. For any given instance and prioritization policy, let E_{w} quantify the net change to the system's throughput due to each individual walk-in's information uncertainty compared to how they would have made their join/balk decision had they been given full information regarding the system's current state (N_1, N_2). We must have $E_{\mathsf{w}} \equiv \lambda_{\mathsf{w}}(\upsilon_{\mathsf{w}}^{+} - \upsilon_{\mathsf{w}}^{-})$ denote this net change, where $\upsilon_{\mathsf{w}}^{+}$ and $\upsilon_{\mathsf{w}}^{-}$ are the probabilities with which a walk-in's decision falls into categories (ii) and (iv), respectively. We can define E_{m} (and the probabilities $\upsilon_{\mathsf{m}}^{+}$ and $\upsilon_{\mathsf{m}}^{-}$ for mobiles analogously (although we note that walk-ins are privy to N_1 when making their decision, while mobiles decide based on no information about the current system state). We then let $E \equiv E_{\mathsf{w}} + E_{\mathsf{m}}$ denote the total net change to the throughput due to the individuals' information uncertainty—while noting that E does not fully capture the role information uncertainty can play in the equilibrium that emerges.

It is straightforward to compute E in the single-server model by leveraging our performance analysis. Leveraging this computation, we plot E as a function of the adoption rate α in Fig. 8 (for the same scenario considered in Fig. 6). This plot leads us to conclude that there is no easily interpretable relationship between the net change E and the adoption rate α. In particular, we note that this figure exhibits many oscillations inherited from oscillations in the $\upsilon_{\mathsf{w}}^{+}$ and $\upsilon_{\mathsf{w}}^{-}$—and especially the $\upsilon_{\mathsf{m}}^{+}$ and $\upsilon_{\mathsf{m}}^{-}$—values as the set of states (N_1, N_2) where a walk-in or mobile would prefer to join the system can frequently change with α and those variables that change with α (e.g., the equilibrium (b^*, p_{m}^*), the limiting probability distribution over (N_1, N_2), etc.).

It is clear from the plot, however, that when measuring the benefit of information uncertainty on throughput in this way, offering the app rarely increases this benefit under this example; only the policy that prioritizes walk-ins (i.e., WOM) "outperforms" No-app with respect to E, and even then only for about a quarter of the range of α values. Hence, information uncertainty plays a subtle role in driving the impact of α on throughput (and a similar role in driving social welfare), giving rise to the counter-intuitive observation that we have discussed in §6.1: that is, sometimes the detrimental impact of the nature of information uncertainty introduced by offering an app offsets the benefit of self-ordering opportunities provided by the addition of the mobile stream.

6.3. Full Factorial Experiment

In §6.1, we showed through illustrative examples that introducing the self-ordering technology may sometimes hurt throughput and social welfare. To explore the generality of this observation and other discussions provided in §6.1, we design an extensive problem set by setting $\lambda = 1$ and varying the other parameters as follows: $\mu_2 \in \{1.5, 2, 2.5, 3\}$, $\mu_1/\mu_2 \in \{0.25, 0.5, 1, 2, 4\}$, $\alpha \in \{0.05, 0.15, \ldots, 0.95\}$, and $T_{\mathsf{m}}^{\max} \in \{0.5, 1, 2, 4\}$, $T_{\mathsf{w}}^{\max}/T_{\mathsf{m}}^{\max} \in \{0.8, 1, 1.25\}$. We focus on the single-server model under which we can obtain all expected sojourn times exactly. Of the 2400 possible combinations, we remove 980 instances where customers of at least one class are too impatient to join even an empty system (i.e., $b = 0$ is the best response to $p_{\mathsf{m}} = 0$ or vice-versa; such cases occur precisely when $T_{\mathsf{w}}^{\max} \leq 1/\mu_1 + 1/\mu_2$ or $T_{\mathsf{m}}^{\max} \leq 1/\mu_2$). We do not remove cases where Assumption 1 is violated; such violations merely limit the space of feasible b and p_{m} that yield finite sojourn times and do not preclude the existence of equilibria.

For each problem instance, we record the policy that yields the highest throughput (including the no-app scenario with $\alpha = 0$). Occasionally, there will be a tie for the highest throughput between MWO and WMO; where possible, we break such ties in favor of the policy with the higher social welfare, while in the remaining cases—where the systems behave identically—we report a tie. In summary, we list our key observations below:

- *In most settings (93.2% of problem instances), introducing the app using the optimal prioritization policy increases the throughput. Under the optimal policy, throughput increases almost linearly with the adoption rate (see Fig. 10).*
- *In some settings (6.8% of problem instances), introducing the app, even using the optimal policy, reduces the throughput substantially (on average, 12.4%).*

Table 1 Policies and associated regrets

	No-app	MWO	WMO	WOM	Tie
Optimality freq.	96	4	63	867	390
Optimality prop. (%)	6.8	0.3	4.4	61.1	27.5
Regret prop. (%)	93.2	72.3	68.1	38.9	
Regret magn. (%)	15.9	5.4	3.3	8.4	

Table 2 Throughput loss of suboptimally offering the app

Average	Std. dev.	Median	Max
12.4%	11.0%	10.6%	40.3%

- *Prioritizing walk-ins (i.e.,* WOM*) is often the best policy (61.1% of problem instances), but the regret from suboptimally employing it is the highest (on average, 8.4%).*

We elaborate on these and other observations in the remainder of this section.

When should an omni-channel structure be employed? According to Table 1, transitioning to an omni-channel setting reduces the throughput in 96 (6.8% of the) experiments. This suggests that the detrimental effect of app introduction is not so unlikely that it can be safely dismissed out of hand. Across these 96 *no-app cases*, the throughput loss resulting from suboptimally offering the app (compared to the policy that generates the highest throughput) can be as high as 40.3%, with a mean of 12.4% (see Table 2).

Based on Fig. 9, the incidence of no-app cases initially increases with the adoption rate, peaking at $\alpha = 0.25$, after which the frequency of these cases drops monotonically; two-thirds of no-app cases occur in the lower half of the α values examined (i.e., between 0.05 and 0.45). As expected, the likelihood of these cases decreases as $T_{\mathsf{m}}^{\max}$ grows: more patience among mobiles is favorable for the app introduction. Note that for a fixed $T_{\mathsf{w}}^{\max}/T_{\mathsf{m}}^{\max}$, $T_{\mathsf{w}}^{\max}$ grows along with $T_{\mathsf{m}}^{\max}$, but more patience among walk-ins is also favorable for app introduction as walk-ins will be willing to wait behind mobiles, under say MWO. Similarly, the likelihood of such cases drops as μ_2 rises (and μ_1 with it): faster service rates play a similar role to that of higher patience levels. On the other hand, there is no such clear trend associated with $T_{\mathsf{w}}^{\max}/T_{\mathsf{m}}^{\max}$, although we note that the no-app cases are more likely to arise when $T_{\mathsf{w}}^{\max} > T_{\mathsf{m}}^{\max}$. Meanwhile, cases where app introduction is detrimental rise sharply with μ_1/μ_2. The faster the walk-in's service at Stage 1 (relative to that at Stage 2), the less significant the advantage of bypassing Stage 1; consequently, the operational advantage of offering a mobile-ordering option diminishes as μ_1/μ_2 increases.

What prioritization policy should be implemented? Based on Table 1, WOM outperforms the other policies in 61.1% of our experiments. Table 1 also quantifies the *regret* associated with choosing a policy and implementing it across *all* experiments in terms of the "proportion" of experiments where another policy would yield either greater

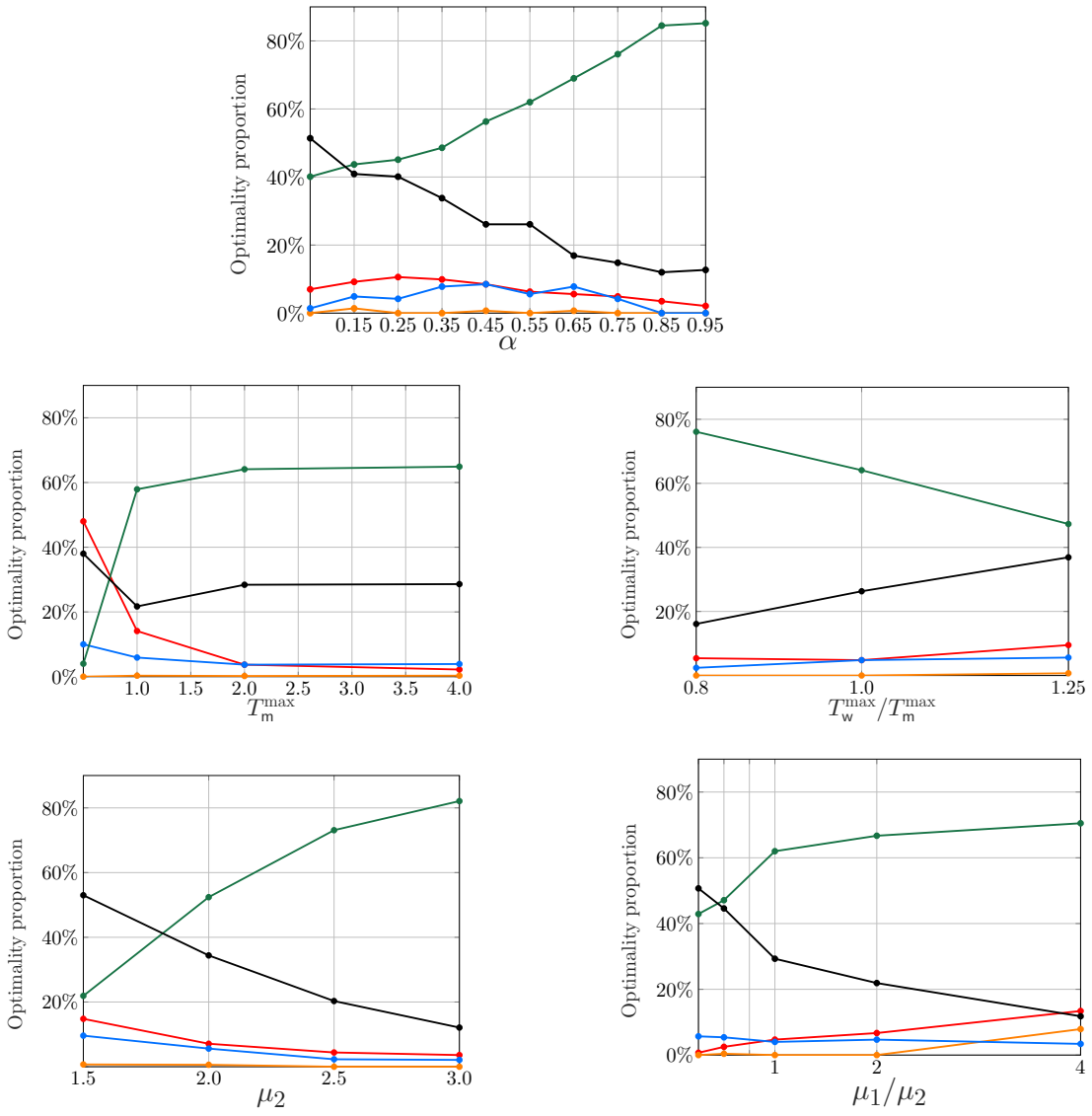


Figure 9 **Impact of parameters on the optimality proportion. (No app: —●—, MWO: —●—, WMO: —●—, WOM: —●—, MWO & WMO tie: —●—)**

throughput or the same throughput (but greater social welfare) and the "magnitude" of this regret (average throughput loss relative to the optimal policy). Prioritizing walk-ins (i.e., WOM) generates regret in the fewest experiments by far. However, it performs quite poorly when suboptimal. This observation is corroborated by Fig. 10, which plots the average throughput change as a function of α relative to the no-app case.

We attribute the widespread dominance of WOM (and the lesser success of the other two policies) to the fact that it is possible to achieve mobile throughput optimality (i.e., $p_{\mathsf{m}} = 1$) in many experiments, even when prioritizing walk-ins. As long as the full participation of

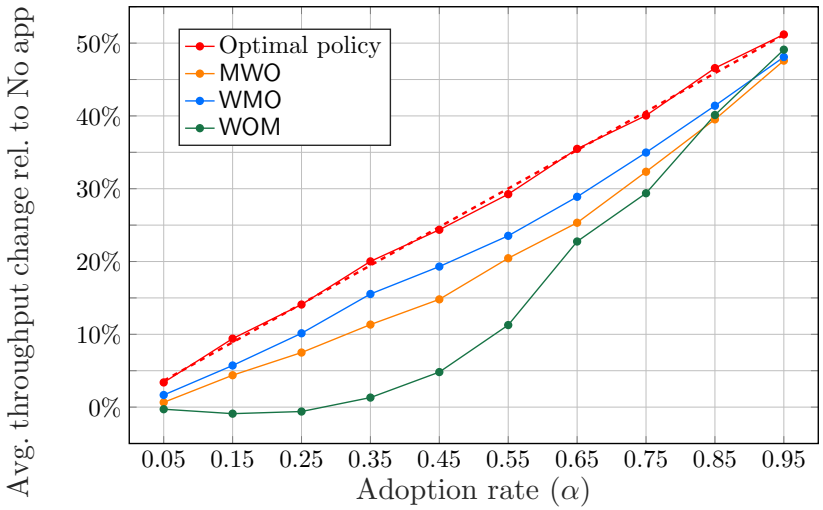


Figure 10 **Average throughput gain; the average for "optimal policy" is calculated by averaging the throughput across each problem instance for whichever of** MWO, WMO, **and** WOM **attain the highest throughput. The "optimal policy" throughput is nearly linear in α, as seen from the dotted regression line based on the plotted data points.**

mobiles can be guaranteed, the problem of maximizing the overall throughput reduces to maximizing that of walk-ins, which is achieved through WOM. As (i) faster service, (ii) a reduction in the share of customers that are walk-ins (i.e., increased adoption rate), and (iii) more mobile patience all tend to reduce the effect of the negative externality imposed on mobiles by the prioritization of walk-ins, the number of instances in which WOM is optimal increases with (i) μ_2 (and μ_1/μ_2), (ii) α, and (iii) $T_{\mathsf{m}}^{\max}$ (Fig. 9). On the other hand, these instances become more rare as $T_{\mathsf{w}}^{\max}/T_{\mathsf{m}}^{\max}$ increases: when the ratio of walk-in patience to mobile patience grows—and the latter is not high enough to guarantee $p_{\mathsf{m}} = 1$ under WOM—the alternative policies (i.e., MWO and WMO) tend to become more favorable. We can explain this tendency by observing that while prioritizing mobiles can lead to both a mobile throughput gain and a walk-in throughput loss, as $T_{\mathsf{w}}^{\max}/T_{\mathsf{m}}^{\max}$ grows, it becomes increasingly likely that the gain will outweigh the loss.

The discussion above illustrates the conditions that make full mobile participation favorable even under WOM. In the absence of these conditions, giving the mobiles top priority can yield the opposite: *no participation of mobiles* ($p_{\mathsf{m}} = 0$), leading to a significant loss in throughput compared to MWO, WMO, and even the single-channel system.

7. Conclusion

This paper utilizes queueing-theoretic techniques to evaluate single- and two-server omni-channel service models in the presence of non-strategic customers with infinite patience levels and strategic customers with finite patience levels. We highlight the importance of prioritization for an efficient transition to an omni-channel service with a finitely-patient customer base. The throughput-optimal policy choice is highly dependent on the operational parameters and on customer patience levels; implementing a wrong policy can yield a significant loss in throughput and, hence, profitability. We uncover a non-negligible number of settings where offering the app under any of the policies we have studied (i.e., Pareto optimal in the setting where customers are infinitely patient) would be detrimental. Such settings arise in the single- and two-server models (and in the former case, even when patience levels are heterogeneous). Such settings also exist (at least in single-server models) when customers exhibit heterogeneous patience levels within each class (see Appendix EC.4.7).

We believe that our contributions in this paper open up ample room for future work in game-theoretic queueing models of omni-channel services. First, our results implicitly feature the occasional existence of throughput-welfare trade-offs, suggesting a rich space of problems that would emerge from introducing a (channel-specific) pricing design lever and the objective of profit maximization. Second, mobile apps are beginning to provide delay estimates to customers, suggesting a real-world need for future work to explore models like those in this paper that assume different information structures. As mentioned in §2, Roet-Green and Yuan have already begun an exploration of this space, and we are optimistic that a detailed examination of a richer model incorporating features of both our models and theirs can shed further light on the dynamics of omni-channel services. Third, in reality, mobiles will often place an order before they are present at the service facility; by incorporating their travel time into our model (possibly incorporating ideas from Baron et al. 2020, Hassin and Roet-Green 2020), we may extract additional insights.

Additionally, future work on customers with finite patience could introduce dynamic (state-dependent) policies and new techniques for analyzing their performance and the equilibria they yield. We are particularly curious about how alternate information structures and dynamic policies can avoid or mitigate the potential harm associated with omni-channel services that this paper highlights. We could also extend our model by endogenizing

the adoption rate α. One approach would be to leverage our expected sojourn time expressions to incorporate strategic channel choice behavior (as in Baron et al. (2021) and Ghosh et al. (2020)) in our models; we suspect this modeling change will suppress adoption rates under WOM and WM, thus making these policies less attractive.

Finally, the blend of queueing-theoretic methods that we have employed in evaluating expected sojourn times may have implications beyond omni-channel services. Specifically, our performance evaluation methods may be seen as the first steps in analyzing a rich space of queueing network models where some—but not all—service stations are buffered.

References

Baron O, Berman O, Wang L (2020) Synchronizing travelling and waiting processes: Customer strategy with an online reservation system. *Available at SSRN 3536517* .

Baron O, Chen X, Li Y (2021) Omnichannel services: The false premise and operational remedies. *Available at SSRN 3444772* .

Bayram A, Cesaret B (2017) Ship-from-store operations in omni-channel retailing. *IIE Annual Conference. Proceedings*, 1181–1186 (Institute of Industrial and Systems Engineers (IISE)).

Bell DR, Gallino S, Moreno A (2018) Offline showrooms in omnichannel retail: Demand and operational benefits. *Management Science* 64(4):1629–1651.

Bertsimas D (1995) The achievable region method in the optimal control of queueing systems; formulations, bounds and policies. *Queueing Systems* 21(3-4):337–389.

Boon MAA, van der Mei RD, Winands EMM (2011) Applications of polling systems. *Surveys in Operations Research and Management Science* 16:67–82.

Borst S, Boxma O (2018) Polling: past, present, and perspective. *TOP* 26:335–369.

Campbell IC (2020) Starbucks says nearly a quarter of all US retail orders are placed from a phone. *The Verge* URL `https://www.theverge.com/2020/10/30/21540908/starbucks-app-q4-earnings-mobile-payments`, accessed: 2021-06-08.

Chopra S (2016) How omni-channel can be the future of retailing. *Decision* 43(2):135–144.

Dacre M, Glazebrook K, Niño-Mora J (1999) The achievable region approach to the optimal control of stochastic systems. *Journal of the Royal Statistical Society. Series B (Methodological)* 61(4):747–791.

D'Auria B, Kanta S (2015) Pure threshold strategies for a two-node tandem network under partial information. *Operations Research Letters* 43:467–470.

Delasay M, Jain A, Kumar S (2021) Impacts of the covid-19 pandemic on grocery retail operations: An analytical model. URL `https://tinyurl.com/ms4vf54r`.

Gallino S, Moreno A, Stamatopoulos I (2017) Channel integration, sales dispersion, and inventory management. *Management Science* 63(9):2813–2831.

Gao F, Su X (2016) Omnichannel retail operations with buy-online-and-pick-up-in-store. *Management Science* 63(8):2478–2492.

Gao F, Su X (2017) Online and offline information for omnichannel retailing. *Manufacturing & Service Operations Management* 19(1):84.

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

Ghosh A, Bassamboo A, Lariviere M (2020) The queue behind the curtain: Information disclosure in omnichannel services. *Available at SSRN 3730482* .

Harchol-Balter M (2013) *Performance Modeling and Design of Computer Systems: Queueing Theory in Action* (Cambridge University Press).

Hassin R (2016) *Rational Queueing* (CRC Press).

Hassin R, Haviv M (2003) *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems* (USA: Kluwer Academic Publishers).

Hassin R, Roet-Green R (2020) On queue-length information when customers travel to a queue. *Manufacturing & Service Operations Management (forthcoming)* .

Ji J, Roet-Green R (2020) On queue-length information in a tandem queueing system. *Available at SSRN 3728894* .

Jin M, Li G, Cheng T (2018) Buy online and pick up in-store: Design of the service area. *European Journal of Operational Research* 268(2):613–623.

Kaczynski WH, Leemis LM, Drew JH (2012) Transient queueing analysis. *INFORMS Journal on Computing* 24(1):10–28.

Kerner Y, Sherzer E, Yanco MA (2017) On non-equilibria threshold strategies in ticket queues. *Queueing Systems* 86:419–431.

Kim B, Kim J (2016) Equilibrium strategies for a tandem network under partial information. *Operations Research Letters* 44:532–534.

Latouche G, Ramaswami V (1999) *Introduction to Matrix Analytic Methods in Stochastic Modeling* (Philadelphia: ASA-SIAM).

Liu Y, Yang L (2020) Order ahead for pickup: Promise or peril? *Available at SSRN 3673617* .

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

Nimrod D, Hassin R, Yechiali U (2020) Strategic behaviour in a tandem queue with alternating server. *Queueing Systems* 96:205–244.

Roet-Green R, Yuan Y (2020) Information visibility in omnichannel queues. *Available at SSRN 3485810* .

Ryan T (2017) Starbucks mobile ordering is working too well. *Forbes* URL `https://www.forbes.com/sites/retailwire/2017/04/12/starbucks-mobile-ordering-is-working-too-well/\#6e03818cea28`.

Prioritization in the Presence of Self-ordering Opportunities in Omni-channel Services—Technical Appendices.

The following six technical appendices are provided as a supplement the body of the paper "Prioritization in the Presence of Self-ordering Opportunities in Omni-channel Services."

First, Appendix EC.1 provides the supplemental results and proofs while Appendix EC.2 provides proofs of the results presented throughout the body of the paper. Second, several quantities discussed in the paper (such as a variety of limiting probability distributions) appear in formulas of the key results, but details on how to compute these quantities (either exactly or approximately) are omitted from the main body of the paper. A discussion on how to obtain these values exactly or approximately is provided in Appendix EC.3. Third, Appendix EC.4 provides a discussion of mixed strategies on the part of walk-ins. Building off of this discussion, this appendix also provides the analysis of the case where patience levels are heterogeneous. Next, Appendix EC.5 presents tables of results associated with the pruned full factorial experiment presented in Section 6 of the body of the paper. Finally, in the interest of aiding the reader, we provide a near-exhaustive table of the notation used throughout the body of the paper and/or these appendices in Appendix EC.6.

EC.1. Supplemental Results

EC.1.1. Allocations under Pareto Generators and Proofs

The following proposition provides the allocations under Pareto generators (MWO, WMO, and WOM in the single-server model; MW, FCFS, and WM in the two-server model).

Proposition EC. 1 *We summarize the class-specific mean sojourn times as follows:*

(a) for the single-server model:

$$\begin{cases} a^{\mathsf{MWO}} = \left(\dfrac{\mu_2\left(\mu_1+\mu_2-\Lambda\right)}{\left(\mu_2-\lambda_{\mathsf{m}}\right)\left(\mu_1\mu_2-\mu_1\Lambda-\mu_2\lambda_{\mathsf{w}}\right)}, \dfrac{1}{\mu_2-\lambda_{\mathsf{m}}} \right) & \text{(EC.1)} \\ a^{\mathsf{WMO}} = \left(\dfrac{\mu_2^3+\mu_2^2\left(\mu_1-\Lambda\right)-\mu_2\lambda_{\mathsf{m}}\left(\mu_1-\lambda_{\mathsf{w}}\right)+\mu_1\Lambda\lambda_{\mathsf{m}}}{\mu_2\left(\mu_2-\lambda_{\mathsf{m}}\right)\left(\mu_1\mu_2-\mu_1\Lambda-\mu_2\lambda_{\mathsf{w}}\right)}, \dfrac{\mu_2+\lambda_{\mathsf{w}}}{\mu_2\left(\mu_2-\lambda_{\mathsf{m}}\right)} \right) & \text{(EC.2)} \\ a^{\mathsf{WOM}} = \left(\dfrac{\mu_1+\mu_2-\lambda_{\mathsf{w}}}{\mu_1\mu_2-\left(\mu_1+\mu_2\right)\lambda_{\mathsf{w}}}, \dfrac{\mu_2\left(\mu_1^2+\mu_2\lambda_{\mathsf{w}}\right)}{\left(\mu_1\mu_2-\lambda_{\mathsf{w}}\left(\mu_1+\mu_2\right)\right)\left(\mu_1\mu_2-\mu_1\Lambda-\mu_2\lambda_{\mathsf{w}}\right)} \right) & \text{(EC.3)} \end{cases}$$

(b) for the two-server model:

$$\begin{cases} a^{\mathsf{MW}} = \left(\dfrac{\mu_2}{(\mu_2-\Lambda)(\mu_2-\lambda_{\mathsf{m}})}, \dfrac{1}{\mu_2-\lambda_{\mathsf{m}}} \right) & \text{(EC.4)} \\ a^{\mathsf{FCFS}} = \left(\dfrac{\mu_1+\mu_2-\lambda_{\mathsf{w}}-\Lambda}{(\mu_1-\lambda_{\mathsf{w}})(\mu_2-\Lambda)}, \dfrac{1}{\mu_2-\Lambda} \right) & \text{(EC.5)} \\ a^{\mathsf{WM}} = \left(\dfrac{\mu_1+\mu_2-2\lambda_{\mathsf{w}}}{(\mu_1-\lambda_{\mathsf{w}})(\mu_2-\lambda_{\mathsf{w}})}, \dfrac{\mu_2}{(\mu_2-\Lambda)(\mu_2-\lambda_{\mathsf{w}})} \right) & \text{(EC.6)} \end{cases}$$

Proofs of Proposition EC. 1

Proof for MWO (Eq. (EC.1)). We can view a system under MWO as operating like a two-class M/G/1 system under preemptive-priority scheduling with class-specific service requirement distributions. Under MWO the mobiles (resp. walk-ins) form the high-priority (resp. low-priority) class, and are conventionally designated as class 1 (resp. class 2). Therefore, we can obtain the desired sojourn times by using the formula (see Chapter 32.2 in Harchol-Balter (2013)),

$$\mathbb{E}[T_k] = \frac{\mathbb{E}[S_k]}{1-\sum_{i=1}^{k-1}\rho_i} + \frac{\sum_{i=1}^{k}\rho_i \mathbb{E}[S_i^2]/(2\mathbb{E}[S_i])}{\left(1-\sum_{i=1}^{k-1}\rho_i\right)\left(1-\sum_{i=1}^{k}\rho_i\right)}, \tag{EC.7}$$

where $\mathbb{E}[T_k]$ is the sojourn time associated with class k, $[S_i]$ and $\mathbb{E}[S_i^2]$ are the first and second moments of the class i service requirement distribution, and $\rho_i = \lambda_i \mathbb{E}[S_i]$ is the contribution to the load due to class i (with λ_i the class i arrival rate).

By observing that under MWO mobiles (resp. walk-ins) require service only at Stage 2 (resp. both Stages 1 and 2), we see that their service requirements are distributed $\text{Exp}(\mu_2)$ (resp., like the sum of an $\text{Exp}(\mu_1)$ and an independent $\text{Exp}(\mu_2)$ random variable). It then follows that

$$\lambda_1 = \lambda_{\mathsf{w}}, \quad \mathbb{E}[S_1] = \frac{1}{\mu_2}, \quad \mathbb{E}[S_1^2] = \frac{2}{\mu_2^2}, \quad \lambda_2 = \lambda_{\mathsf{m}}, \quad \mathbb{E}[S_2] = \frac{1}{\mu_2} + \frac{1}{\mu_1}, \quad \mathbb{E}[S_2^2] = \frac{2}{\mu_1\mu_2} + \frac{2}{\mu_1^2} + \frac{2}{\mu_2^2}. \tag{EC.8}$$

Substituting the values given in display (EC.8) into (EC.7) readily yields (EC.1).

Proof for WMO (Eq. (EC.2)). Under WMO, once a walk-in finishes service in Stage 1, they will be served with the highest priority and without interruption in Stage 2 until his service is completed; i.e., the mean sojourn time of walk-ins in Stage 2 is $1/\mu_2$. Therefore, we can represent the walk-in mean sojourn time as:

$$\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}] = \mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w},1}] + \frac{1}{\mu_2}, \tag{EC.9}$$

where $T_{\mathsf{w},1}$ represents a walk-in's sojourn time in Stage 1. A walk-in's Stage 1 sojourn time consists of a busy period with initial work equal to the amount of work the walk-in finds in the system (at both stages) upon its arrival, W, in addition to its own contribution to work in Stage 1—distributed $\text{Exp}(\mu_1)$—and interruptions due to mobile arrivals (which arrive according to a Poisson process with rate λ_{m}, where each interruption contributes an average of $1/\mu_2$ additional work). Hence, standard busy period analysis yields

$$\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w},1}] = \frac{\mathbb{E}[W] + 1/\mu_1}{1-\lambda_{\mathsf{m}}/\mu_2}. \tag{EC.10}$$

We proceed to determine $\mathbb{E}[W]$. First observe that W has the same distribution under any work-conserving service policy, and therefore corresponds to the distribution of the sojourn time *in queue*, T_Q, associated with an M/G/1 system under first-come-first-serve scheduling with two independent arrival streams: the first (resp. second) stream corresponds to that of walk-ins (resp. mobiles) in the original setting and has an arrival rate of λ_{w} (resp. λ_{m}); meanwhile, service requirements are distributed like $\text{Exp}(\mu_1)+\text{Exp}(\mu_2)$ (resp. $\text{Exp}(\mu_2)$). By "merging" these arrival streams, we find that this M/G/1 system has a total arrival rate of $\Lambda=\lambda_{\mathsf{w}}+\lambda_{\mathsf{m}}$, with the first and second moments of the service requirement distribution—denoted by $\mathbb{E}[S]$ and $\mathbb{E}[S^2]$, respectively—given by

$$\mathbb{E}[S]=\frac{\lambda_{\mathsf{w}}}{\Lambda}\left(\frac{1}{\mu_1}+\frac{1}{\mu_2}\right)+\frac{\lambda_{\mathsf{m}}}{\Lambda}\left(\frac{1}{\mu_2}\right),\qquad \mathbb{E}\left[S^2\right]=\frac{\lambda_{\mathsf{w}}}{\Lambda}\left(\frac{2}{{\mu_2}^2}+\frac{2}{{\mu_1}^2}+\frac{2}{\mu_1\mu_2}\right)+\frac{\lambda_{\mathsf{m}}}{\Lambda}\left(\frac{2}{{\mu_2}^2}\right). \tag{EC.11}$$

Letting $\rho\equiv\Lambda\mathbb{E}[S]$ denote the load associated with this M/G/1 system, the Pollaczek-Khinchine formula yields the following:

$$\mathbb{E}[W]=\mathbb{E}[T_Q]=\frac{\rho}{1-\rho}\frac{\mathbb{E}\left[S^2\right]}{2\mathbb{E}[S]}. \tag{EC.12}$$

Substituting the equations in display (EC.11) into Eq. (EC.12), and the result into Eq. (EC.9), we obtain the $\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]$ expression in Eq. (EC.2) as desired.

Now, we derive the mobiles mean sojourn time. Under WMO, when a mobile begins service, we know that there are no walk-ins currently at Stage 2, and hence the mobile's service cannot be interrupted. Let $\mathbb{E}^{\mathsf{WMO}}[N_{\mathsf{m},Q}]$ and $\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m},Q}]$ denote the mean *queue* length (ignoring the server) and mean sojourn time *in queue* (ignoring the service time) associated with mobiles. We have:

$$\begin{aligned}\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m},Q}]&=\mathbb{E}^{\mathsf{WMO}}[\text{Time to serve orders in queue}]\\&\qquad+\mathbb{P}^{\mathsf{WMO}}(\mathbf{M}\text{ arrival finds server busy with }\mathbf{W}\text{ or }\mathbf{M})\cdot\mathbb{E}^{\mathsf{WMO}}[\text{Time to finish current service}]\\&=\mathbb{E}^{\mathsf{WMO}}[N_{\mathsf{m},Q}]\cdot\frac{1}{\mu_2}+\mathbb{P}^{\mathsf{WMO}}(\mathbf{M}\text{ arrival finds server busy with }\mathbf{W}\text{ or }\mathbf{M})\cdot\frac{1}{\mu_2}\\&=\frac{\lambda_{\mathsf{m}}}{\mu_2}\cdot\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m},Q}]+\frac{\Lambda}{\mu_2}\cdot\frac{1}{\mu_2}\qquad\text{(according to the Little's law.)}\end{aligned} \tag{EC.13}$$

From Eq. (EC.13), we derive $\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m},Q}]=\Lambda/(\mu_2(\mu_2-\lambda_{\mathsf{m}}))$; using $\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]=\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m},Q}]+1/\mu_2$, we derive the $\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]$ expression in Eq. (EC.2).

Proof for WOM (Eq. (EC.3)). WOM prioritizes the walk-ins in both stages as opposed to MWO, which prioritizes the mobiles over all walk-ins. Therefore, applying the same

procedure presented in the case of MWO—with the modification that walk-ins are now designated as class 1 and mobiles as class 2—yields the desired result.

Viewing the two-server model as a tandem Jackson network, we see that Stage 1 is an M/M/1 queue with only walk-in customers. We observe that Stage 2 receives two independent arrival streams: new **W**s which are former **O**s departing Stage 1 (with rate λ_{w}) and external **M** arrivals (with rate μ_2). The former arrival stream is also the departure process of an M/M/1, and hence, by Burke's Theorem (see Harchol-Balter (2013) Ch. 16.3), it constitutes a Poisson process, while the latter arrival stream is a Poisson process by assumption. As the two arrival streams are independent, the resulting merged process—and hence, the overall arrival process to Stage 2—is a Poisson process with rate $\lambda_{\mathsf{w}} + \lambda_{\mathsf{m}}$. It follows that Stage 2 is also an M/M/1 queue.

MW **Policy:** Under MW, mobiles have the higher priority in Stage 2, so they experience an M/M/1 with arrival rate λ_{m} and service rate μ_2, and hence $\mathbb{E}^{\mathsf{MW}}[T_{\mathsf{m}}] = 1/(\mu_2 - \lambda_{\mathsf{m}})$. Meanwhile, we determine the mean sojourn time of walk-ins under WM by summing their Stage 1 mean sojourn time (which is that of an M/M/1 system with arrival rate λ_{w} and service rate μ_1) with their Stage 2 mean sojourn time; this latter mean sojourn time is obtained from Eq. (EC.7), by noting that under WM walk-ins have lower priority than mobiles in Stage 2. Simplifying the result yields the following:

$$\mathbb{E}^{\mathsf{MW}}[T_{\mathsf{w}}] = \frac{1}{\mu_1 - \lambda_{\mathsf{w}}} + \frac{1}{(\mu_2 - \Lambda)(1 - \lambda_{\mathsf{m}}/\mu_2)} = \frac{\mu_2}{(\mu_2 - \Lambda)(\mu_2 - \lambda_{\mathsf{m}})}.$$

WM **Policy:** Under WM, walk-ins have the higher priority in Stage 2, so they experience two successive M/M/1 sojourn times (one for each stage); summing the resulting mean sojourn times yields the following:

$$\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{w}}] = \frac{1}{\mu_1 - \lambda_{\mathsf{w}}} + \frac{1}{\mu_2 - \lambda_{\mathsf{w}}} = \frac{\mu_1 + \mu_2 - 2\lambda_{\mathsf{w}}}{(\mu_1 - \lambda_{\mathsf{w}})(\mu_2 - \lambda_{\mathsf{w}})}.$$

Meanwhile, as mobiles have lower priority than walk-ins in Stage 2 under WM, we determine the mean mobile sojourn time by applying Eq. (EC.7):

$$\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{m}}] = \frac{1}{(\mu_2 - \Lambda)(1 - \lambda_{\mathsf{w}}/\mu_2)} = \frac{\mu_2}{(\mu_2 - \Lambda)(\mu_2 - \lambda_{\mathsf{w}})}.$$

FCFS **Policy:** Under FCFS, both walk-ins and mobiles have the same mean sojourn time at Stage 2, so we have $\mathbb{E}^{\mathsf{FCFS}}[T_{\mathsf{m}}] = 1/(\mu_2 - \Lambda)$ and

$$\mathbb{E}^{\mathsf{FCFS}}[T_{\mathsf{w}}] = \frac{1}{\mu_1 - \lambda_{\mathsf{w}}} + \frac{1}{\mu_2 - \Lambda} = \frac{\mu_1 + \mu_2 - \lambda_{\mathsf{w}} - \Lambda}{(\mu_1 - \lambda_{\mathsf{w}})(\mu_2 - \Lambda)}.$$

EC.1.2. The Statement and Proof of the Deconditioning Lemma

The following lemma—which we call the deconditioning lemma—is helpful in proving a number of this paper's propositions:

Lemma EC. 1 *For any policy* P*, we have*

$$\begin{aligned}\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i] &= \sum_{j=0}^{\infty}\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i,N_2=j]\pi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j)\Bigg/\sum_{j=0}^{\infty}\pi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j)\\ &= \sum_{j=0}^{\infty}\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i,N_{2,\mathsf{w}}=j]\phi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j)\Bigg/\sum_{j=0}^{\infty}\phi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j).\end{aligned}$$

Proof. The first equality follows from "deconditioning" on $N_2=j$—along with the implicit use of the PASTA (Poisson Arrivals See Time Averages) property—as follows:

$$\begin{aligned}\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i] &= \sum_{j=0}^{\infty}\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i,N_2=j]\mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(N_2=j|N_1=i)\\ &= \sum_{j=0}^{\infty}\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i,N_2=j]\mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(N_1=i,N_2=j)\Bigg/\mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(N_1=i)\\ &= \sum_{j=0}^{\infty}\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i,N_2=j]\mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(N_1=i,N_2=j)\Bigg/\sum_{j=0}^{\infty}\mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(N_1=i,N_2=j)\\ &= \sum_{j=0}^{\infty}\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i,N_2=j]\pi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j)\Bigg/\sum_{j=0}^{\infty}\pi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j).\end{aligned}$$

The second equality follows in a similar fashion by deconditioning on $N_{2,\mathsf{w}}=j$:

$$\begin{aligned}\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i] &= \sum_{j=0}^{\infty}\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i,N_{2,\mathsf{w}}=j]\mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(N_{2,\mathsf{w}}=j|N_1=i)\\ &= \sum_{j=0}^{\infty}\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i,N_{2,\mathsf{w}}=j]\phi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j)\Bigg/\sum_{j=0}^{\infty}\phi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j).\end{aligned}$$

EC.2. Proofs of Results

Here we provide the proofs of the Propositions and Theorems presented in body of the paper.

EC.2.1. Proof of Proposition 1

Proof outline. We first prove the set $\{\mathsf{MWO},\mathsf{WMO},\mathsf{WOM}\}$ forms a set of Pareto generators for the single-server model in section EC.2.1.1, then we proceed to prove the set $\{\mathsf{MW},\mathsf{WM}\}$ also forms a set of Pareto generators for the two-server model in section EC.2.1.2.

EC.2.1.1. Proof for the single-server model

Preliminaries. To prove the statement, it is sufficient to show that the achievable region $\mathcal{O} = \text{conv}\{a^{\mathsf{MWO}}, a^{\mathsf{WMO}}, a^{\mathsf{WOM}}\} + \text{cone}\{(0,1),(1,0)\} \subseteq \mathbb{R}^2$ is equivalent to the unbounded convex polygon defined by all pairs a^{P}=$(\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}], \mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}])$ satisfying the following four inequality constraints (equivalently, all such points lying in the intersection of the four half-planes defined by these affine inequality constraints), which correspond (at equality) to the rays and line segments, which together make up the boundary of the achievable region, $\mathbf{bd}(\mathcal{O})$, as captured by the example illustrated in Fig. 3a (from leftmost to rightmost):

1. $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] \geq \mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}]$
2. $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}] \geq \left(\dfrac{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{m}}] - \mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]}\right)\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \dfrac{\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]}$
3. $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}] \geq -\dfrac{\lambda_{\mathsf{w}}}{\lambda_{\mathsf{m}}}\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \dfrac{\Lambda}{\lambda_{\mathsf{m}}}\mathbb{E}^{\mathsf{WMO}}[T]$
4. $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}] \geq \mathbb{E}^{\mathsf{MWO}}[T_{\mathsf{m}}]$

The first and fourth inequalities are readily apparent from the formulation of $\mathcal{O}$ given above, with the second corresponding to the line that runs through both a^{WOM} and a^{WMO}, and the third—which corresponds to the line running through a^{WMO} and a^{MWO}—following directly from the fact that for all $\mathsf{P} \in \mathcal{P}$, we have $\mathbb{E}^{\mathsf{P}}[T] \geq \mathbb{E}^{\mathsf{MWO}}[T]$, where the overall mean sojourn time is given by $\mathbb{E}^{\mathsf{P}}[T] = \left(\lambda_{\mathsf{w}}\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \lambda_{\mathsf{m}}\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}]\right)/\Lambda$ (see Proposition 3). It can be verified in a straightforward manner that—consistent with what we observe from Fig. 3a—the line corresponding to the first inequality is vertical (i.e., parallel to the $\mathbb{E}[T_{\mathsf{m}}]$-axis), those corresponding to the second and third inequalities are negatively sloped (with the second steeper than the third), while that corresponding to the forth is horizontal (i.e., parallel to the $\mathbb{E}[T_{\mathsf{w}}]$-axis). Moreover, the first and second lines intersect at a^{WOM}, the second and third at a^{WMO}, and the last two at a^{MWO}, establishing that $\mathcal{O}$ will always qualitatively resemble that in Fig. 3a, although the locations of—and thus the angles and distances between—a^{WOM}, a^{WMO}, and a^{MWO} are parameter-dependent. If we can show that these four inequalities define the achievable region, then we have proved the first claim of the theorem, and the second claim follows from straightforward observation that the only Pareto allocations are those that in addition to satisfying all four inequalities, satisfy the second and/or third with equality.

It remains only to prove that the constraints defined by these four inequalities are both necessary (i.e., for any policy $\mathsf{P} \in \mathcal{P}$, a^{P} satisfies these four inequalities) and sufficient (i.e.,

any allocation a satisfying these four inequalities can be achieved by implementing some feasible policy $\mathsf{P} \in \mathcal{P}$, or equivalently for all such a there exists $\mathsf{P} \in \mathcal{P}$ such that $a = a^{\mathsf{P}}$) in order to establish that an allocation $a \in \mathcal{O}$. In referring to these four, we use the terms inequality and constraint interchangeably.

Proof for sufficiency. We first address the case where a lies on one of the four lines corresponding to the inequalities that we claim define the achievable region (i.e., if a satisfies one or more of these inequalities strictly). If a lies on the line corresponding to the first inequality, then we can achieve allocation $a = (\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}], r_{\mathsf{m}})$ for some $r_{\mathsf{m}} > \mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{m}}]$ by implementing a modification of WOM where we slow down the rate at which we serve mobiles (but not walk-ins) at Stage 2 from μ_2 to some specific $\mu_2' < \mu_2$ that would cause the mean sojourn time of mobiles to rise from $\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{m}}]$ to r_{m} while keeping that of walk-ins fixed at $\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}]$. Such a a value of μ_2' must exist as the mean sojourn time of walk-ins under such modifications of WOM will continuously vary over the interval $(\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{m}}], \infty)$ as we vary the new service rate of walk-ins at Stage 2 over the interval $(\lambda_m/(1 - \lambda_w/\mu_1 - \lambda_w/\mu_2), \mu_2)$. If a lies on the line corresponding to the second or third inequalities, i.e., if $a \in \text{conv}\{a^{\mathsf{WOM}}, a^{\mathsf{WMO}}\} \cap \text{conv}\{a^{\mathsf{WMO}}, a^{\mathsf{MWO}}\}$, then we can achieve this allocation by implementing $\langle \mathsf{WOM}, \mathsf{WMO} \rangle(\theta)$ or $\langle \mathsf{WMO}, \mathsf{MWO} \rangle(\theta)$ for the appropriately chosen θ. Next, see that if a lies on the line corresponding to the fourth inequality, then $a = (r_{\mathsf{w}}, \mathbb{E}^{\mathsf{MWO}}[T_{\mathsf{m}}])$ can be achieved by implementing a modification of MWO analogous to the modification of WMO considered for a lying on the line corresponding to the second inequality; in this case, we slow down the service rate of walk-ins—rather than that of mobiles—at Stage 2.

We now address the remaining case where a satisfies all four inequalities, but does not satisfy any of them strictly. We consider two sub-cases: first, if a lies in (the interior or boundary) of the triangle $\text{conv}\{a^{\mathsf{WOM}}, a^{\mathsf{WMO}}, a^{\mathsf{MWO}}\}$ (shaded in red in Fig. EC.1), then we can achieve a by implementing a policy that randomly uses WOM, WMO, and MWO at the start of each busy period with the appropriate probabilities. The only case that remains is when a satisfies all of the inequalities and also lies above and to the right of the line segment connecting WOM and MWO. In this case, as shown in Fig. EC.1, we can take achieve a by implementing a policy that randomizes between two specific policies, P_1 and P_4 with appropriate probabilities. These two policies are chosen so that they yield allocations a^{P_1} and a^{P_4} that uniquely satisfy the following: (i) a^{P_1} and a^{P_4} satisfy the

first and fourth inequalities with equality, respectively, (ii) the line segment connecting a^{P_1} and a^{P_4} is parallel to the line segment connecting a^{WOM} to a^{MWO}, and (iii) a lies on the aforementioned line segment. Recall from the preceding paragraph that any policy, such as P_1 (resp. P_4), that satisfies the first (resp. fourth) inequality with equality can be implemented by modifying WOM (resp. MWO) through a service rate reduction for mobiles (resp. walk-ins) at Stage 2. Note that in this case, while we can still implement all of the policies $\{\langle \mathsf{P}_1, \mathsf{P}_4\rangle(\theta)\colon \theta \in [0,1]\}$, it may not be the case that $a^{\langle \mathsf{P}_1,\mathsf{P}_4\rangle(\theta)} = \theta a^{\mathsf{P}_1} + (1-\theta)a^{\mathsf{P}_4}$, as P_1 and P_4 are not work-conserving. Nevertheless, there must exist some value of $\theta \in (0,1)$ for which $a = a^{\langle \mathsf{P}_1,\mathsf{P}_4\rangle(\theta)}$ (for P_1 and P_4 chosen appropriately) because $\{a^{\langle \mathsf{P}_1,\mathsf{P}_4\rangle(\theta)}\colon \theta \in (0,1)\} = \{\theta a^{\mathsf{P}_1} + (1-\theta)a^{\mathsf{P}_4}\colon \theta \in (0,1)\}$. This completes the proof that the four inequalities provide constraints on allocations a, that are sufficient for establishing that $a \in \mathcal{O}$.

Proof for necessity. We proceed by showing that for any $a \in \mathcal{O}$ (or equivalently, for any $\mathsf{P} \in \mathcal{P}$), each of the four constraints must hold. Addressing the first constraint, observe that WOM achieves the minimum possible mean walk-in sojourn time, as this policy strictly prioritizes walk-ins over mobiles, while also prioritizing those walk-ins with the least remaining expected service requirements (the latter follows from the fact that **W**s are prioritized over **O**s), and hence, the first constraint must hold. We can address the fourth constraint

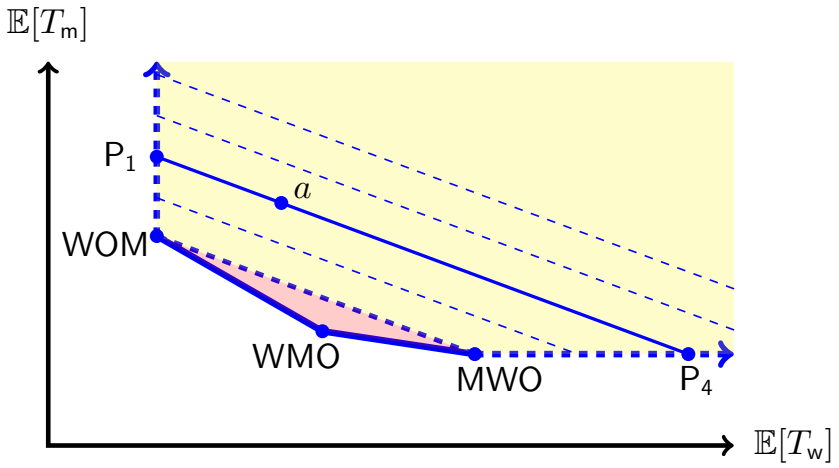


Figure EC.1 **Allocations lying in the red triangle can be implemented by considering a probabilistic mixture of the WOM, WMO, and MWO policies, while allocations within the blue region, such as the example a illustrated here, all lie on a line that runs parallel to that connecting a^{WOM} and a^{MWO} and intersects the vertical and horizontal boundaries at a pair of allocations that can be implemented through the policies P_1 and P_4. Implementing an appropriate random mixture of these two policies will allow for achieving allocation a.**

in a similar manner: MWO achieves the minimum possible mean mobile sojourn time, so the fourth constraint must also hold. Meanwhile, as alluded to earlier in this proof, the necessity of the third constraint follows directly from Proposition 3, which establishes that for all $\mathsf{P} \in \mathcal{P}$, we have $\mathbb{E}^{\mathsf{P}}[T] \geq \mathbb{E}^{\mathsf{MWO}}[T]$, where the overall mean sojourn time is given by $\mathbb{E}^{\mathsf{P}}[T] = \left(\lambda_{\mathsf{w}}\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \lambda_{\mathsf{m}}\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}]\right)/\Lambda$.

We now turn to addressing the only remaining item: the necessity of the second inequality:

$$\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}] \geq \left(\frac{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{m}}] - \mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]}\right)\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \frac{\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]}.$$

We begin by examining the overall mean *work* in the system under P, which we denote by $\mathbb{E}^{\mathsf{P}}[W]$. Clearly, each **W** and **M** task contributes an average of $1/\mu_2$ work each. Meanwhile, each **O** task contributes an average of $1/\mu_1$ work *by itself*; if we also account for the **W** task that must be served after serving each **O** task (in order to serve a walk-in in its entirety), we can view each **O** currently in the system as contributing an average of $1/\mu_1 + 1/\mu_2$ work to the system. Before using the observations above to derive the total work in the system, we recall that N_1 and N_2 denote the number of customers at Stages 1 (all of which are walk-ins) and 2, respectively; we further let N_{w}, $N_{2,\mathsf{w}}$, and N_{m} denote the number of walk-in customers in the system as a whole, the number of walk-ins at Stage 2 specifically, and the number of mobile customers (all of whom are at Stage 2), respectively, and note that $N_1 + N_{2,\mathsf{w}} = N_{\mathsf{w}}$, while $N_{\mathsf{m}} + N_{2,\mathsf{w}} = N_2$. We can now decompose $\mathbb{E}^{\mathsf{P}}[W]$ as follows:

$$\begin{aligned}\mathbb{E}^{\mathsf{P}}[W] &= \left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right)\mathbb{E}^{\mathsf{P}}[N_1] + \left(\frac{1}{\mu_2}\right)\mathbb{E}^{\mathsf{P}}[N_{2,\mathsf{w}}] + \left(\frac{1}{\mu_2}\right)\mathbb{E}^{\mathsf{P}}[N_{\mathsf{m}}] \\ &= \left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right)\mathbb{E}^{\mathsf{P}}[N_{\mathsf{w}}] - \left(\frac{1}{\mu_1}\right)\mathbb{E}^{\mathsf{P}}[N_{2,\mathsf{w}}] + \left(\frac{1}{\mu_2}\right)\mathbb{E}^{\mathsf{P}}[N_{\mathsf{m}}] \qquad \text{(EC.14)}\end{aligned}$$

Applying Little's Law to Eq. (EC.14), and rearranging terms, we have:

$$\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}] = -\left(\frac{\rho_{\mathsf{w}}}{\rho_{\mathsf{m}}}\right)\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \left(\frac{1}{\rho_{\mathsf{m}}}\right)\mathbb{E}^{\mathsf{P}}[W] + \left(\frac{1}{\mu_1\rho_{\mathsf{m}}}\right)\mathbb{E}^{\mathsf{P}}[N_{2,\mathsf{w}}], \qquad \text{(EC.15)}$$

where $\rho_{\mathsf{w}} \equiv \lambda_{\mathsf{w}}(1/\mu_1 + 1/\mu_2)$ and $\rho_{\mathsf{m}} \equiv \lambda_{\mathsf{m}}/\mu_2$ are the fractions of the time spent serving walk-ins and mobiles, respectively (and hence, $1 - \rho_{\mathsf{w}} - \rho_{\mathsf{m}}$ is the fraction of time in which the server is idle).

We rewrite Eq. (EC.15) in terms of $\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}]$, $\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{m}}]$, $\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]$, $\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]$ (all of which are provided explicitly in Proposition 3), and use the resulting expression to bound $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}]$ as follows:

$$\begin{aligned}\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}] &= \left(\frac{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{m}}]-\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}]-\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]}\right)\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}]+\left(\frac{1}{\rho_{\mathsf{m}}}\right)\mathbb{E}^{\mathsf{P}}[W]+\left(\frac{1}{\mu_1\rho_{\mathsf{m}}}\right)\mathbb{E}^{\mathsf{P}}[N_{2,\mathsf{w}}]\\ &\geq \left(\frac{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{m}}]-\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}]-\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]}\right)\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}]+\left(\frac{1}{\rho_{\mathsf{m}}}\right)\mathbb{E}^{\mathsf{WOM}}[W]+\left(\frac{1}{\mu_1\rho_{\mathsf{m}}}\right)\mathbb{E}^{\mathsf{WOM}}[N_{2,\mathsf{w}}] \qquad \text{(EC.16)}\\ &= \left(\frac{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{m}}]-\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}]-\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]}\right)\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}]+\frac{\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}]-\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WOM}}[T_{\mathsf{w}}]-\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]}. \qquad \text{(EC.17)}\end{aligned}$$

Hence, $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}]$ is bounded below by the expression to the right of the equals sign in Eq. (EC.17), which yields precisely the second constraint, and so it only remains to justify Ineq. (EC.16) and Eq. (EC.17). We justify Ineq. (EC.16) by showing that $\min_{\mathsf{P}\in\mathcal{P}}\mathbb{E}^{\mathsf{P}}[W]=\mathbb{E}^{\mathsf{WOM}}[W]$ and $\min_{\mathsf{P}\in\mathcal{P}}\mathbb{E}^{\mathsf{P}}[N_{2,\mathsf{w}}]=\mathbb{E}^{\mathsf{WOM}}[N_{2,\mathsf{w}}]$. Moreover, we provide explicit expressions for these two expectations; Eq. (EC.17) follows from these expressions directly after straightforward (if lengthy) calculations.

We first show that $\min_{\mathsf{P}\in\mathcal{P}}\mathbb{E}^{\mathsf{P}}[W]=\mathbb{E}^{\mathsf{WOM}}[W]$. This follows directly from the fact that WOM is work-conserving; indeed, $\mathbb{E}^{\mathsf{P}}[W]$ must attain its minimum value under all work-conserving policies $\mathsf{P}\in\mathcal{P}$. We proceed to compute $\mathbb{E}^{\mathsf{WOM}}[W]$, noting that this is the same as the time average work under any work-conserving policy. In fact, we can view $\mathbb{E}^{\mathsf{WOM}}[W]$ as the average work in an ordinary M/G/1 system (under any work-conserving scheduling policy) with two streams of Poisson arrivals, exactly like those described in the proof of Eq. (EC.2) in Appendix EC.1.1; i.e., the first (resp. second) stream corresponds to that of walk-ins (resp. mobiles) in the original setting and has an arrival rate of λ_{w} (resp. λ_{m}); meanwhile, service requirements are distributed like $\mathrm{Exp}(\mu_1)+\mathrm{Exp}(\mu_2)$ (resp. $\mathrm{Exp}(\mu_2)$), and so by standard M/G/1 analysis, we have

$$\mathbb{E}^{\mathsf{WOM}}[W]=\left(\lambda_{\mathsf{w}}\left(\frac{1}{\mu_1^2}+\frac{1}{\mu_2^2}+\frac{1}{\mu_1\mu_2}\right)+\frac{\rho_{\mathsf{m}}}{\mu_2}\right)\Big/(1-\rho_{\mathsf{w}}-\rho_{\mathsf{m}}). \qquad \text{(EC.18)}$$

Finally, we justify $\min_{\mathsf{P}\in\mathcal{P}}\mathbb{E}^{\mathsf{P}}[N_{2,\mathsf{w}}]=\mathbb{E}^{\mathsf{WOM}}[N_{2,\mathsf{w}}]$. In fact $\mathbb{E}^{\mathsf{P}}[N_{2,\mathsf{w}}]$ is minimized by any policy $\mathsf{P}\in\mathcal{P}$ that give **W**s priority over all other tasks. Such policies (including WOM), allow only one **W** task to be in the system at any given time, as they would not serve an

O (allowing it to become a **W**) if there is already a **W** present in the system. Hence, under such policies, $N_{2,\mathsf{w}} = 1$ whenever there is a **W** in service and $N_{2,\mathsf{w}} = 0$ otherwise. Since each **W** spends the minimum average amount of time possible (i.e., $1/\mu_2$) in service, the claim is justified. Furthermore, **W**s arrive to the system at the same rate at which **O**s complete service, and since the system is throughput-optimal, we know that the arrival rate of **W**s is λ_{w}. Meanwhile, we have already argued that under WOM and the other **W**-prioritizing policies, **W**s spend an average of $1/\mu_2$ time in the system, and so by Little's law, we have $\mathbb{E}^{\mathsf{WOM}}[N_{2,\mathsf{w}}] = \lambda_{\mathsf{w}}/\mu_2$.

With the explicit computation of $\mathbb{E}^{\mathsf{WOM}}[W]$ as given in Eq. (EC.18) and the fact that we have $\mathbb{E}^{\mathsf{WOM}}[N_{2,\mathsf{w}}] = \lambda_{\mathsf{w}}/\mu_2$, we can readily verify Eq. (EC.17), which completes the proof.

EC.2.1.2. Proof for the two-server model

We follow the same approach that we used in proving the statement for single-server model (see Appendix EC.2.1.1); we opt for less expository precision and shorter justifications in the interest of brevity. The achievable region $\mathcal{O} = \mathrm{conv}\{a^{\mathsf{MW}}, a^{\mathsf{WM}}\} + \mathrm{cone}\{(0,1),(1,0)\} \subseteq \mathbb{R}^2$ (for allocations in the two-server model) is equivalent the region expressed by the conjunction of the following inequalities (also referred to as constraints):

1. $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] \geq \mathbb{E}^{\mathsf{WM}}[T_{\mathsf{w}}]$
2. $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}] \geq \left(\dfrac{\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{m}}] - \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{w}}]}\right)\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \dfrac{\mathbb{E}^{\mathsf{MW}}[T_{\mathsf{m}}]\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{w}}]\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{w}}]}$
3. $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}] \geq \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{m}}]$

Note that the allocation of FCFS policy, a^{FCFS}, is located on the line segment generated by a^{WM} and a^{MW}. Applying an analogous argument to that deployed in Appendix EC.2.1.1, we can deduce that any allocation satisfying these three constraints can be implemented by a feasible two-server prioritization policy $\mathsf{P} \in \mathcal{P}$. It remains to show that these three constraints are also necessary.

It is straightforward to see the first and the third inequalities are satisfied by any service policy since WM and MW achieve the minimum possible mean sojourn time for walk-ins or mobiles respectively. It remains only to prove the second inequality for all $\mathsf{P} \in \mathcal{P}$.

For any given set of parameters λ_{w}, λ_{m}, μ_1, and μ_2 satisfying Assumption 1(b), it follows from Burke's Theorem (see Section 16.3 in Harchol-Balter (2013)) that the departure process at Stage 1 (and hence the arrival rate of walk-ins to Stage 2) is a Poisson process with rate $\chi_{\mathsf{w}} = \lambda_{\mathsf{w}}$. Hence, we focus on Stage 2, which we view as an M/M/1 system with

arrival rate $\Lambda = \lambda_{\mathsf{w}} + \lambda_{\mathsf{m}}$ and service rate μ_2. For any two-server prioritization policy $\mathsf{P} \in \mathcal{P}$, we can decompose $\mathbb{E}^{\mathsf{P}}[W_2]$, the mean work at Stage 2, and apply Little's Law to obtain the following:

$$\begin{aligned}\mathbb{E}^{\mathsf{P}}[W_2] &= \left(\frac{1}{\mu_2}\right)\mathbb{E}^{\mathsf{P}}[N_{2,\mathsf{w}}] + \left(\frac{1}{\mu_2}\right)\mathbb{E}^{\mathsf{P}}[N_{\mathsf{m}}] \\ &= \left(\frac{\lambda_{\mathsf{w}}}{\mu_2}\right)\left(\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] - \frac{1}{\mu_1 - \lambda_{\mathsf{w}}}\right) + \left(\frac{\lambda_{\mathsf{m}}}{\mu_2}\right)\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}]. \qquad \text{(EC.19)}\end{aligned}$$

We rearrange terms and write Eq. (EC.19) in terms of $\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{w}}]$, $\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{m}}]$, $\mathbb{E}^{\mathsf{MW}}[T_{\mathsf{w}}]$, $\mathbb{E}^{\mathsf{MW}}[T_{\mathsf{m}}]$ (all of which are provided explicitly in Proposition EC. 1 (b)), and use the resulting expression to bound $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}]$ as follows:

$$\begin{aligned}\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}] &= -\left(\frac{\lambda_{\mathsf{w}}}{\lambda_{\mathsf{m}}}\right)\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \left(\frac{\mu_2}{\lambda_{\mathsf{m}}}\right)\mathbb{E}^{\mathsf{P}}[W_2] + \frac{\lambda_{\mathsf{w}}}{\lambda_{\mathsf{m}}(\mu_1 - \lambda_{\mathsf{w}})} \qquad \text{(EC.20)} \\ &= \left(\frac{\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{m}}] - \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{w}}]}\right)\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \left(\frac{\mu_2}{\lambda_{\mathsf{m}}}\right)\mathbb{E}^{\mathsf{P}}[W_2] + \frac{\lambda_{\mathsf{w}}}{\lambda_{\mathsf{m}}(\mu_1 - \lambda_{\mathsf{w}})} \\ &\geq \left(\frac{\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{m}}] - \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{w}}]}\right)\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \left(\frac{\mu_2}{\lambda_{\mathsf{m}}}\right)\mathbb{E}^{\mathsf{WM}}[W_2] + \frac{\lambda_{\mathsf{w}}}{\lambda_{\mathsf{m}}(\mu_1 - \lambda_{\mathsf{w}})} \qquad \text{(EC.21)} \\ &= \left(\frac{\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{m}}] - \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{w}}]}\right)\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \frac{\mathbb{E}^{\mathsf{MW}}[T_{\mathsf{m}}]\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{w}}]\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{m}}]}{\mathbb{E}^{\mathsf{WM}}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{MW}}[T_{\mathsf{w}}]}. \qquad \text{(EC.22)}\end{aligned}$$

Hence, $\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}]$ is bounded below by the expression to the right of the equals sign in Eq. (EC.22), which yields precisely the second constraint, and so it only remains to justify Ineq. (EC.21) and Eq. (EC.22). We justify Ineq. (EC.21) by showing that $\min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}^{\mathsf{P}}[W_2] = \mathbb{E}^{\mathsf{WM}}[W_2]$. Moreover, we provide an explicit expression for $\mathbb{E}^{\mathsf{WM}}[W_2]$, from which we can obtain Eq. (EC.22) directly after straightforward (if lengthy) calculations.

We first show that $\min_{\mathsf{P}\in\mathcal{P}} \mathbb{E}^{\mathsf{P}}[W_2] = \mathbb{E}^{\mathsf{WM}}[W_2]$. This follows directly from the fact that WM is work-conserving; indeed, $\mathbb{E}^{\mathsf{P}}[W_2]$ must attain its minimum value under all work-conserving policies $\mathsf{P} \in \mathcal{P}$. Then we proceed to determine $\mathbb{E}^{\mathsf{WM}}[W_2]$. Once more, we view Stage 2 as an M/M/1 queueing system, but this time we are considering the system under WM; leveraging the fact that WM is a work-conserving policy we can apply standard M/M/1 analysis together with Little's Law to obtain the following:

$$\mathbb{E}^{\mathsf{WM}}[W_2] = \left(\frac{1}{\mu_2}\right)\mathbb{E}^{\mathsf{WM}}[N_2] = \frac{\Lambda}{\mu_2(\mu_2 - \Lambda)}. \qquad \text{(EC.23)}$$

With the explicit computation of $\mathbb{E}^{\mathsf{WM}}[W_2]$ as given in Eq. (EC.23), we can readily verify Eq. (EC.22), which completes the proof.

EC.2.2. Proof of Proposition 3

In the single-server model, under MWO, mobiles experience an M/M/1 queue as they have the highest priority, while walk-ins will be preempted by mobile arrivals. Meanwhile, under WOM, walk-ins experience an M/G/1 queue with the highest priority, while mobiles are preempted by walk-in arrivals (to Stage 1). Since customers at Stage 2—which are (expected to be) closer to service completion than those at Stage 1—receive the highest priority under both MWO and WMO, we have the following lowest overall mean sojourn time:

$$\mathbb{E}^{\mathsf{MWO}}[T] = \mathbb{E}^{\mathsf{MWO}}[T_{\mathsf{w}}]\frac{\lambda_{\mathsf{w}}}{\Lambda} + \mathbb{E}^{\mathsf{MWO}}[T_{\mathsf{m}}]\frac{\lambda_{\mathsf{m}}}{\Lambda} = \mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]\frac{\lambda_{\mathsf{w}}}{\Lambda} + \mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]\frac{\lambda_{\mathsf{m}}}{\Lambda} = \mathbb{E}^{\mathsf{WMO}}[T] \tag{EC.24}$$

Replacing $\mathbb{E}^{\mathsf{MWO}}[T_{\mathsf{w}}]$ and $\mathbb{E}^{\mathsf{MWO}}[T_{\mathsf{m}}]$ (or $\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{w}}]$ and $\mathbb{E}^{\mathsf{WMO}}[T_{\mathsf{m}}]$) in Eq. (EC.24) results in the exact formula of the lowest overall mean sojourn time.

In the two-server model, for any Pareto optimal policy, the statement directly follows from Eq. (EC.20) after rearranging terms.

EC.2.3. Proof of Proposition 4(a)

Let (b, p_{m}) be an equilibrium and assume by way of contradiction that $b > B$. If we can show that $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = b-1] \geq b/\mu_1 + 1/\mu_2 > T_{\mathsf{w}}^{\max}$ holds, the proof is complete as we have contradicted the equilibrium conditions. The first inequality follows from the fact that the assumption on P dictates that a walk-in that sees $N_1 = b-1$ upon arrival must wait behind $b-1$ other walk-ins at Stage 1, in addition to their own service at each stage. The second inequality follows from the definition of B, i.e., $B \equiv \mu_1(T_{\mathsf{w}}^{\max} - 1/\mu_2)$, and straightforward arithmetic.

EC.2.4. Proof of Proposition 4(b)

First we address the cases of MWO and MW in the single- and two-server settings, respectively. In both cases, mobiles have preemptive priority over all others, and so mobiles experience an M/M/1 system with arrival rate $p_{\mathsf{m}}\lambda_{\mathsf{m}}$ and service rate μ_2. Therefore, $\mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] = \mathbb{E}^{\mathsf{MW}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] = 1/(\mu_2 - p_{\mathsf{m}}\lambda_{\mathsf{m}})$, which is clearly increasing in the arrival rate $p_{\mathsf{m}}\lambda_{\mathsf{m}}$, and hence in p_{m}.

Next, we examine the case of WOM in the single-server model. We observe that the arrival process of **W**s to Stage 2 is the same as the departure process of **O**s at Stage 1, and

since these have priority over mobiles and b is fixed, this arrival process does not depend on $p_{\mathsf{m}}\lambda_{\mathsf{m}}$. Hence, if we examine only mobiles, we note that they experience a queue with Poisson arrivals and an exponential service process with (exogenous) Markov-modulated service interruptions. In such a system, the higher the arrival rate, the longer the sojourn time, so the desired result holds.

The case of WM and FCFS in the two-server model is similar to that of WOM in the single-server model. As in that case, the arrival process of **W**s to Stage 2 does not depend on $p_{\mathsf{m}}\lambda_{\mathsf{m}}$, which again allows us to view mobiles as experiencing a queue that is a modified M/M/1 with (exogenous) Markov-modulated service interruptions. Again, the desired result follows.

Finally, we consider the case of WMO in the single-server model. Unlike the previous cases, the arrival process of **W**s to Stage 2 can depend on $p_{\mathsf{m}}\lambda_{\mathsf{m}}$ under WMO; so, the same type of argument that we used for the previous cases does not suffice. Instead, we consider the *overall* mean sojourn time, observing that $\mathbb{E}^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}[T] = \mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T]$. This observation follows from the fact that b and p_{m} are fixed, which ensures that the evolution of (N_1, N_2)—and hence $N = N_1 + N_2$—is the same under both policies. Naturally, $\mathbb{E}[N]$ is also the same under both policies, as is $\mathbb{E}[T]$ (by Little's Law). With this observation in mind, we can break up these overall sojourn times into class-specific sojourn times, yielding:

$$\frac{\chi_{\mathsf{w}}}{X}\mathbb{E}^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}] + \frac{\chi_{\mathsf{m}}}{X}\mathbb{E}^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] = \frac{\chi_{\mathsf{w}}}{X}\mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}] + \frac{\chi_{\mathsf{m}}}{X}\mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}], \tag{EC.25}$$

where χ_w, χ_m, and χ can depend on one or more of $\lambda_{\mathsf{m}}p_{\mathsf{m}}$, λ_{w}, μ_1, μ_2, and b, but do not depend on the choice of MWO versus WMO (recall that we are not considering an equilibrium, but a fixed value of b that is the same under both policies). From Eq. (EC.25), we obtain:

$$\begin{aligned}\mathbb{E}^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] &= \frac{\chi_{\mathsf{w}}}{\chi_{\mathsf{m}}}\left(\mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}] - \mathbb{E}^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}]\right) + \mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] \\ &= \frac{\chi_{\mathsf{w}}}{p_{\mathsf{m}}\lambda_{\mathsf{m}}}\left(\frac{1}{\mu_2 - p_{\mathsf{m}}\lambda_{\mathsf{m}}} - \frac{1}{\mu_2}\right) + \frac{1}{\mu_2 - p_{\mathsf{m}}\lambda_{\mathsf{m}}} = \frac{\chi_{\mathsf{w}} + \mu_2}{\mu_2(\mu_2 - p_{\mathsf{m}}\lambda_{\mathsf{m}})},\end{aligned}$$

where the difference in the mean sojourn times for walk-ins under the two policies is computed by considering only the sojourn times in Stage 2 (as those in Stage 1 are identical for both policies): under MWO the walk-in sojourn time in Stage 2 is distributed like an

M/M/1 busy period, while under WMO it would be a single exponential distributed service time. From the computation above, together with the fact that χ_{w} is constant in $p_{\mathsf{m}}\lambda_{\mathsf{m}}$, we conclude that $\mathbb{E}^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}[T_m]$ is increasing in $p_{\mathsf{m}}\lambda_{\mathsf{m}}$ as desired.

EC.2.5. Proof of Proposition 5

Proof for MWO (Eq. (1))

Proof for mobiles. Since mobiles have preemptive priority over all others under MWO, they will experience an M/M/1 queue with arrival rate $p_{\mathsf{m}}\lambda_{\mathsf{m}}$ and service rate μ_2. Consequently, we have $\mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] = 1/(\mu_2 - p_{\mathsf{m}}\lambda_{\mathsf{m}})$.

Proof for walk-ins. We first find $\mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i, N_2 = j]$, and then apply Lemma 1. Under MWO walk-ins are preempted by mobiles in both stages; therefore, we can think of a walk-in's sojourn time as being distributed as a particular kind of busy period. When a walk-in joins the system seeing $N_1 = i$ other customers in Stage 1 and $N_2 = j$ customers in Stage 2, the total work in the system—which is equal to the sojourn time of the newly arrived walk-in under MWO assuming no further arrivals—consists of $i+1$ independent Stage 1 services and $i+j+1$ independent Stage 2 services (as the customers in Stage 1 will all also require service at Stage 2). However, the walk-in will be preempted by any mobile arrivals, with each contributing its service requirement to the walk-in's sojourn time. The aforementioned preemptions occur according to a Poisson process with rate $p_{\mathsf{m}}\lambda_{\mathsf{m}}$. Hence, the standard busy period analysis—together with the fact that all services are i.i.d. and consume $\mathrm{Exp}(\mu_k)$ time at Stage k—yields:

$$\mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i, N_2 = j] = \left(\frac{i+1}{\mu_1} + \frac{i+j+1}{\mu_2}\right) \Bigg/ \left(1 - \frac{p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_2}\right) = \frac{(i+1)(\mu_1+\mu_2) + j\mu_1}{\mu_1(\mu_2 - p_{\mathsf{m}}\lambda_{\mathsf{m}})}. \tag{EC.26}$$

Applying Lemma 1 to the above, we obtain the sojourn time of walk-ins:

$$\begin{aligned}
\mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i] &= \left(\sum_{j=0}^{\infty} \mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i, N_2 = j]\pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j)\right) \Bigg/ \sum_{j=0}^{\infty} \pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j) \\
&= \left(\sum_{j=0}^{\infty} \frac{(i+1)(\mu_1+\mu_2) + j\mu_1}{\mu_1(\mu_2 - p_{\mathsf{m}}\lambda_{\mathsf{m}})} \pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j)\right) \Bigg/ \sum_{j=0}^{\infty} \pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j) \\
&= \left((i+1)(\mu_1+\mu_2) + \mu_1 \sum_{j=0}^{\infty} j\pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j) \Bigg/ \sum_{j=0}^{\infty} \pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j)\right) \Bigg/ \left(\mu_1(\mu_2 - p_{\mathsf{m}}\lambda_{\mathsf{m}})\right) \\
&= \left(\left(\frac{\mu_2}{\mu_1} + 1\right)(i+1) + \sum_{j=0}^{\infty} j\pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j) \Bigg/ \sum_{j=0}^{\infty} \pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j)\right) \mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}].
\end{aligned}$$

Proof for WMO (Eq. (2))

Proof for mobiles. We determine the sojourn time of mobiles under WMO, by seeing that a mobile arriving to a system where $N_2 = j$ experiences a sojourn time consisting of $j+1$ services, each distributed $\text{Exp}(\mu_2)$. We note that when a mobile enters the system when Stage 2 is nonempty (i.e., when $N_2 = j \geq 1$), the job currently in service may be a mobile or a walk-in, while all customers in the Stage 2 queue are mobiles. Which of these is the case, however, is immaterial, however, as the remaining service requirement of the customer in service is distributed $\text{Exp}(\mu_2)$, regardless of whether they are an **M** or **W**. So in any case, we have $\mathbb{E}^{\mathsf{WMO}}_{(b,p_\mathsf{m})}[T_\mathsf{m}|N_2 = j] = (j+1)/\mu_2$.

Deconditioning on $N_2 = j$, observing that the mobile-sojourn time is independent of N_1 under WMO, and applying the PASTA property proves the claimed result for the mobiles in Eq. (2):

$$\begin{aligned}\mathbb{E}^{\mathsf{WMO}}_{(b,p_\mathsf{m})}[T_\mathsf{m}] &= \sum_{i=0}^{b}\sum_{j=0}^{\infty}\mathbb{E}^{\mathsf{WMO}}_{(b,p_\mathsf{m})}[T_\mathsf{m}|N_1 = i, N_2 = j]\mathbb{P}^{\mathsf{WMO}}_{(b,p_\mathsf{m})}(N_1 = i, N_2 = j)\\ &= \sum_{i=0}^{b}\sum_{j=0}^{\infty}\mathbb{E}^{\mathsf{WMO}}_{(b,p_\mathsf{m})}[T_\mathsf{m}|N_2 = j]\pi^{\mathsf{WMO}}_{(b,p_\mathsf{m})}(i,j) = \sum_{i=0}^{b}\sum_{j=0}^{\infty}\left(\frac{j+1}{\mu_2}\right)\pi^{\mathsf{WMO}}_{(b,p_\mathsf{m})}(i,j) = \frac{1}{\mu_2}\left(1+\sum_{i=0}^{b}\sum_{j=0}^{\infty}j\pi^{\mathsf{WMO}}_{(b,p_\mathsf{m})}(i,j)\right).\end{aligned}$$

Proof for walk-ins. We follow an approach similar to that used to prove the walk-in's equation in Eq. (1). Under WMO, walk-ins are preempted by mobiles while they are in Stage 1, but then they receive priority once they are in Stage 2. Consequently, since walk-ins can complete Stage 1 service only when Stage 2 is unoccupied (i.e., when $N_2 = 0$), upon completion of Stage 1 service, they move to Stage 1, where they will be served uninterrupted (as there are no other walk-ins already present at Stage 1, nor can they be preempted by mobiles). Hence, the time a walk-in spends in Stage 1 is distributed like a busy period (discussed below), while the time spent in Stage 2 is simply distributed $\text{Exp}(\mu_2)$.

Now recall that under MWO and based on the busy period analysis, we expressed $\mathbb{E}^{\mathsf{MWO}}_{(b,p_\mathsf{m})}[T_\mathsf{w}|N_1 = i, N_2 = j]$ by the first equality in Eq. (EC.26).

By contrast under WMO, the initial workload (seen by a walk-in that arrives when $N_1 = i$ and $N_2 = j$) that contributes to possible preemptions by mobiles (from the perspective of this walk-in) consists of one fewer Stage 2 service, since the walk-in's own Stage 2 service is "immune" to interruptions. The arrival rate and service requirement of these interruptions remain unchanged. Hence, we have:

$$\mathbb{E}^{\mathsf{WMO}}_{(b,p_\mathsf{m})}[T_\mathsf{w}|N_1 = i, N_2 = j] = \left(\frac{i+1}{\mu_1} + \frac{i+j}{\mu_2}\right)\Big/\left(1 - \frac{p_\mathsf{m}\lambda_\mathsf{m}}{\mu_2}\right) + \frac{1}{\mu_2}.$$

The application of Lemma 1 again resembles that featured in the proof of Eq. (1), and yields the claimed result for the walk-ins in Eq. (2).

Proof for WOM (Eq. (3))

Proof for walk-ins. Consider a walk-in that sees $N_1 = i$ **O**s in Stage 1 and $N_{2,\mathsf{w}} = j$ **W**s in Stage 2 upon arrival in the system under WOM; the presence of any **M**s in Stage 2 will not concern a walk-in as walk-ins have preemptive priority over mobiles under WOM. Moreover, recall that $j \in \{0, 1\}$ as there can be at most one **W** in Stage 2 under WOM (as soon as a walk-in advances to Stage 2, they receive uninterrupted service in Stage 2 until completion). Since walk-ins cannot be preempted, it follows that the walk-in's sojourn time consists of $i + 1$ services at each Stage, plus an additional service at Stage 2 if $j = 1$, so that

$$\mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i, N_{2,\mathsf{w}} = j] = \frac{i+1}{\mu_1} + \frac{i+1}{\mu_2} + \frac{j}{\mu_2},$$

which results in the expression for the walk-ins in Eq. (3) by deconditioning on $N_{2,\mathsf{w}} = j$.

Proof for mobiles. We start by tagging a mobile arrival under WOM. Consider two cases: (i) the tagged mobile arrives to an empty system with no other mobiles, and (ii) the mobile arrives to a system with at least one other mobile present in Stage 2. These cases are mutually exclusive and exhaustive and neither case stipulates anything regarding the presence or absence of walk-ins at either stage at the arrival time. In case (i), the tagged mobile's sojourn time is clearly distributed like $U + V$, as the mobile will initiate service after a duration of time distributed like U, after which its remaining sojourn time is distributed like that of a mobile that arrives to an empty system (i.e., like V). In case (ii) the tagged mobile begins service precisely when there are no mobiles in the system that arrived before it and no walk-ins in the system (at either stage); this will necessarily be a point in time at which the last mobile to arrive before the tagged mobile has just completed service. At this point, the remaining sojourn time of the tagged mobile is distributed like that of a mobile that arrives to an empty system, i.e., it is distributed like V.

Now imagine that we view the "service time" of the tagged mobile—and in fact of any mobile—as the time from when it first enters service until its completion time. That is, we view the service time as consisting of the ordinary service time of the mobile in addition to the service time of all walk-ins (originally **O**s and later **W**s) that interrupt this service time. Note that we cannot think of V as an ordinary busy period with Poisson arrivals,

because walk-ins do not effectively arrive according to a Poisson arrival process (walk-ins attempting to arrive when $N_1 = b$ will balk). Viewed like this, the system is always "serving" mobiles (if there are any in the system), as the system is either actually serving a mobile, or "serving" a mobile in the new view by actually serving walk-ins that are interrupting the service of a mobile. Hence, the system can be viewed as an M/G/1 system with i.i.d. service requirements distributed like V, however, the first mobile at the start of each mobile busy period (i.e., mobiles that arrive to a system with no other mobiles) must first wait for a duration of time distributed like U before service begins. Therefore, this system is an M/G/1/setup system with arrivals following a Poisson process with rate $p_{\mathsf{m}}\lambda_{\mathsf{m}}$, services distributed like V, and setups distributed like U. It follows from the discussion of such systems in Harchol-Balter (2013) (Section 27.3; Eq. (27.14)) that we have the claimed result for mobiles in Eq. (3). The calculation of the moments of U and V are provided in Appendix EC.3.3.

EC.2.6. Proof of Lemma 1

We consider a "tagged" walk-in who arrives to the system seeing $N_1 = i$ customers in Stage 1 and $N_2 = j$ customers in Stage 2. Now consider the time interval $\mathcal{I}(i)$ from when the tagged customer first arrived to Stage 1 (equivalently, arrived to the system) until they first arrived to Stage 2 (equivalently, finished service at Stage 1). As our notation suggests, $\mathcal{I}(i)$ depends on i. Observe that $Y(i,j)$ must be the expected Stage 2 workload at the end of $I(i)$. The length of $\mathcal{I}(i)$ is distributed Erlang$(i+1, \mu_1)$.

Now let $K(i)$ (which depends on i) be the random quantity of mobile customers that arrived during $\mathcal{I}(i)$. It follows that Stage 2 would have received $i+K+1$ arrivals—including the tagged customer—during $\mathcal{I}(i)$: the i walk-ins who were already present in Stage 1, the aforementioned $K(i)$ mobiles, and the walk-in that arrived to Stage 2 at the end of $\mathcal{I}$.

We will determine the distribution of $K(i)$ shortly, but for now let us assume that we are given that $K(i) = k$. Given this, let $L(i,j,k)$ (which depends on i, j, and k) be the number of customers present in Stage 2 at the end of $\mathcal{I}(i)$; the tagged customer will find anywhere between 0 and $i+j+k$ other customers in Stage 2 depending on the number of Stage 2 service completions during $\mathcal{I}(i)$; so, $L(i,j,k) \in \{1, 2, \ldots, i+j+k+1\}$. Moreover, note that $Y(i,j)$ is the expectation of the sum of $L(i,j,k)$ independent service requirements, $S_1, S_2, \ldots$, each of which is distributed Exp(μ_2). In order to compute $Y(i,j)$, we now turn to determining the distribution of $L(i,j,k)$.

Now observe that $L(i,j,k)=\ell$ precisely when the Stage 2 occupancy—which starts at j at the start of $\mathcal{I}(i)$—reaches ℓ at the end of $\mathcal{I}(i)$; i.e., $L(i,j,k)=\ell$, when N_2 goes from i to j after exactly $i+j+k+1$ arrivals. Since a customer is in service in Stage 1 during the entirety of $\mathcal{I}(i)$ (except possibly at the last moment), Stage 2 functions like an M/M/1 queue with an arrival rate of $\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}$ and a service rate of μ_2, and hence a load of $\rho=(\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}})/\mu_2$. Therefore, using the notation $P(\cdot,\cdot,\cdot;\cdot)$ as defined in Def. 1, we have:

$$\mathbb{P}(L(i,j,k)=\ell)=P\left(j,i+k+1,\ell;\frac{\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_2}\right).$$

We now return to determining the distribution of $K(i)$. Note that $K(i)\in\{0,1,\ldots\}$ is the number of arrivals during $\mathcal{I}(i)$, where the arrivals follow a Poisson process with rate $p_{\mathsf{m}}\lambda_{\mathsf{m}}$. Recall that the length of $\mathcal{I}(i)$ is distributed Erlang$(i+1,\mu_1)$, and note that it is independent of the aforementioned Poisson process. Consequently $K(i)$ can also be thought of as sum of $i+1$ independent copies of a random variable, X, corresponding to the number of arrivals in a duration of time that is distributed Exp(μ_1). Elementary techniques yield $X\sim\text{Geo}(p_{\mathsf{m}}\lambda_{\mathsf{m}}/(\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}))$, and hence $K(i)\sim\text{NB}(i+1,p_{\mathsf{m}}\lambda_{\mathsf{m}}/(\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}))$ (where both of these distributions are of the kind where the support consists of all non-negative integers, including zero). It follows that

$$\mathbb{P}(K(i)=k)=\binom{k+i}{k}\left(\frac{p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}}\right)^k\left(1-\frac{p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}}\right)^{i+1}.$$

Putting everything together, and recalling that $S_1,S_2,\ldots$ are i.i.d. Exp(μ_2) random variables representing (remaining) Stage 2 service requirements, we can prove our claim:

$$\begin{aligned}
Y(i,j)&=\mathbb{E}\left[\sum_{m=1}^{L(i,j,K)}S_m\right]=\sum_{k=0}^{\infty}\mathbb{E}\left[\sum_{m=1}^{L(i,j,k)}S_m\right]\mathbb{P}(K(i)=k)=\sum_{k=0}^{\infty}\mathbb{E}[L(i,j,k)]\mathbb{E}[S_1]\mathbb{P}(K(i)=k)\\
&=\sum_{k=0}^{\infty}\sum_{\ell=1}^{i+j+k+1}\frac{\ell}{\mu_2}\mathbb{P}(L(i,j,k)=\ell)\mathbb{P}(K(i)=k)\\
&=\sum_{k=0}^{\infty}\sum_{\ell=1}^{i+j+k+1}\frac{\ell}{\mu_2}P\left(j,i+k+1,\ell;\frac{\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_2}\right)\mathbb{P}(K(i)=k)\\
&=\left(1-\frac{p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}}\right)^{i+1}\sum_{k=0}^{\infty}\sum_{\ell=1}^{i+j+k+1}\frac{\ell}{\mu_2}P\left(j,i+k+1,\ell;\frac{\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_2}\right)\binom{k+i}{k}\left(\frac{p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}}\right)^k.
\end{aligned}$$

EC.2.7. Proof of Proposition 6

Proof for MW (Eq. (5))

Proof for walk-ins. Under MW, a walk-in seeing $N_1 = i$ customers in Stage 1 and $N_2 = j$ customers in Stage 2 upon arrival spends $i+1$ services in Stage 1 (each distributed $\text{Exp}(\mu_1)$) before advancing to Stage 2. The walk-in arrives to Stage 2 and spends an amount of time in Stage 2 that is distributed like a busy period initiated by $Y(i,j)$ workload (see Lemma 1) and interrupted by mobile arrivals (with rate $p_{\mathsf{m}}\lambda_{\mathsf{m}}$, with each interruption requiring $\text{Exp}(\mu_2)$ service). Hence, we have $\mathbb{E}^{\mathsf{MW}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i, N_2 = j] = (i+1)/\mu_1 + Y(i,j)/(1 - p_{\mathsf{m}}\lambda_{\mathsf{m}}/\mu_2)$, which with a straightforward application of Lemma 1 yields the result for the walk-ins in Eq. (5).

Proof for mobiles. As in the case of MWO in the single-server setting—by having preemptive priority over all others, mobiles experience an M/M/1 queue, and so $\mathbb{E}^{\mathsf{MW}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] = 1/(\mu_2 - p_{\mathsf{m}}\lambda_{\mathsf{m}})$ as claimed.

Proof for FCFS (Eq. (6))

Proof for walk-ins. Under FCFS a walk-in seeing $N_i = i$ and $N_2 = j$ waits for $i+1$ services in Stage 1, which takes on average $(i+1)/\mu_1$ time, and then waits for a number of services in Stage 2, which takes on average $Y(i,j)$ time (see Lemma 1). Hence, we have $\mathbb{E}^{\mathsf{FCFS}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i, N_2 = j] = (i+1)/\mu_1 + Y(i,j)$, and applying Lemma 1 yields the result for the walk-ins in Eq. (6).

Proof for mobiles. Mobiles are treated under FCFS in a similar fashion as they were under WMO in the single-server setting: they are not preempted, but have to wait behind any pre-existing **M**s or **W**s in Stage 2 when they arrive. Hence, if mobiles arrive seeing $N_2 = j$, their sojourn time will consist of $j+1$ Stage 2 services. Following an approach similar to that in the proof of Eq. (1) from Proposition 5, which gives $\mathbb{E}^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}]$, we readily have the claimed result:

$$\mathbb{E}^{\mathsf{FCFS}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}] = \sum_{i=0}^{b}\sum_{j=0}^{\infty}\frac{j+1}{\mu_2}\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j) = \frac{1}{\mu_2}\left(1 + \sum_{i=0}^{b}\sum_{j=0}^{\infty} j\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)\right).$$

Proof for WM (Eq. (7)) **Proof for walk-ins.** Recall that walk-ins can preempt mobiles under WM; so, they need only care about other walk-ins in the system upon arrival. Consider a tagged walk-in under WM that sees upon arrival $N_1 = i$ **O**s in Stage 1 and $N_{2,\mathsf{w}} = j$ **W**s in Stage 2. Let $L(i,j)$ be the number of customers in Stage 2 (including the tagged

walk-in) at time of the tagged walk-in's arrival to Stage 2, given that $N_1 = i$ and $N_2 = j$. It then readily follows that the tagged walk-in's mean sojourn time is $(i+1)/\mu_1 + \ell/\mu_2$, given that $L(i,j) = \ell$. Now we turn our attention to determining the distribution of $L(i,j)$.

The distribution of $L(i,j)$ is analogous to the distribution of $L(i,j,k)$ from the proof of Lemma 1 (see Appendix EC.2.6), with the key difference that we ignore mobile arrivals entirely (that is, $K(i) = k = 0$ and we can view $p_{\mathsf{m}} = 0$ when determining the arrival rate to Stage 2). That is, we view the queue of **W**s in Stage 2 as an M/M/1 system with arrival rate μ_1—as Stage 1 is occupied during the entirety of the tagged walk-in's sojourn there—and service rate μ_2, so that Stage 2 is under a load of $\rho = \mu_1/\mu_2$. It follows that $L(i,j) = \ell$ precisely with the probability that an M/M/1 system with load $\rho = \mu_1/\mu_2$ starting with j customers will have ℓ customers after $i+1$ additional arrivals (as arrival $i+1$ is the tagged walk-in), so that $\mathbb{P}(L(i,j) = \ell) = P(j, i+1, \ell, \mu_1/\mu_2)$ (see Def. 1). Therefore, it follows that:

$$\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{w}} | N_1 = i, N_{2,\mathsf{w}} = j] = \frac{i+1}{\mu_1} + \frac{\mathbb{E}[L(i,j)]}{\mu_2} = \frac{i+1}{\mu_1} + \sum_{\ell=1}^{i+j+1} \left(\frac{\ell}{\mu_2}\right) P\left(j, i+1, \ell, \frac{\mu_1}{\mu_2}\right).$$

Now recall that the probability of an arrival finding $N_1 = i$ and $N_{2,\mathsf{w}} = j$ is given by $\phi^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}(i,j)$; so, by deconditioning on $N_{2,\mathsf{w}} = j$ (in a fashion similar to Lemma 1), we have the claimed result for walk-ins in Eq. (7).

Proof for mobiles. Consider a tagged mobile arrival that enters a system under WM. Observe that any mobiles arriving after the tagged mobile have no impact on the sojourn time of the tagged mobile as they are of lower priority. Hence, we can carry out our analysis while imagining that no further mobiles arrive after the tagged mobile.

Under the view described above, the tagged mobile completes service precisely when Stage 2 is next empty, as the tagged arrival has the absolute lowest priority among all customers who will be present in Stage 2 at any point in its sojourn because (i) the tagged mobile is preempted by all **W**s, and (ii) the tagged mobile arrived after all other **M**s (given our modified view of the system). Hence the sojourn time of the tagged mobile is the time to clear Stage 2 of all its contents; alternatively, it is a busy period initiated by an amount of work equal to $j+1$ Stage 2 services (including the service of the tagged mobile), where the only other arrivals are walk-ins, given that there are currently i of them in Stage 1.

The exotic arrival process of **W**s through the tandem queue complicates using standard M/G/1 busy period analysis, so we use Markov chain analysis instead. To this end, we

observe that it does not matter how many of the $j+1$ Stage 2 services are **W**s and how many are **M**s, as this does not affect service times. So, let us think of all of them as being **W**s (note that this is clearly false as we know at least one of the $j+1$ Stage 2 customers is the tagged mobile, which is of course an **M** and not a **W**). It follows that $\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}]$ coincides with the time to clear a "mobile-less" system (i.e., one where $p_{\mathsf{m}}=0$) of all Stage 2 customers given that we start with $N_1=i$ and $N_2=N_{2,\mathsf{w}}=j+1$. In other words, we are interested in the time until we go from stage $(i,j+1)$ in the Markov chain governing $(N_1,N_{2,\mathsf{w}})$ (see Fig. 5b) to any state in the initial column, i.e., $(k,0)$ for some $k\in\{0,1,\ldots,b\}$. Hence, $(T_m|N_1=i,N_2=j)\sim Z(i,j+1)$, where

$$Z(i,j)\sim \inf\{s\geq 0\colon N_{2,\mathsf{w}}(t+s)=0|N_1(t)=i,N_{2,\mathsf{w}}(t)=j\}.$$

A method for approximating the expectation of $Z(i,j)$ with arbitrary accuracy is given in Appendix EC.3.8.

To complete the proof of the claim we condition on the event that $N_1=i$ and $N_2=j$. Recall that although earlier in our argument we chose to treat all Stage 2 customers as **W**s, when conditioning we condition on the event that $(N_1=i,N_2=j)$ and not on $(N_1=i,N_{2,\mathsf{w}}=j)$, because the tagged mobile arrival is concerned with the total number of Stage 2 customers at the arrival time of the tagged mobile, as the pre-existing mobiles still have a higher priority. Hence, the probabilities of the events of interest are given by $\pi^{\mathsf{MW}}_{(b,p_{\mathsf{m}})}(i,j)$ (equivalently, $\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)$) rather than $\phi^{\mathsf{MW}}_{(b,p_{\mathsf{m}})}(i,j)$. Finally, carrying out the appropriate conditioning step, we can establish the claimed result for mobiles in Eq. (7).

EC.3. Computational details

We provide details for calculating various quantities of interest.

EC.3.1. The Limiting Probabilities $\pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j)$ and $\pi^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}(i,j)$ and their Associated Series

Recall that (N_1,N_2) is governed by the same CTMC under both MWO and WMO (see Fig. 4a), which has finitely many phases (rows) and infinitely many levels (columns). We notice that phase transitions are *unidirectional* throughout the infinite repeating portion of the chain (but bidirectional in the initial non-repeating portion). We use $\pi_{(b,p_{\mathsf{m}})}(i,j)$ to denote the limiting probabilities under both MWO and WMO, and we let $\vec{\pi}_j=(\pi_{(b,p_{\mathsf{m}})}(0,j),\ldots,\pi_{(b,p_{\mathsf{m}})}(b,j))$, $j\geq 0$. We define the five square matrices $\mathbf{F_0},\mathbf{F},\mathbf{L_0},\mathbf{L}$,

and $\mathbf{B} \in \mathbb{R}^{(b+1)\times(b+1)}$ such that (using zero-based indexing so that the upper left element of any matrix $\mathbf{M}$ is denoted by $\mathbf{M}(0,0)$) for the repeated portion of the Markov chain, $\mathbf{F}(\ell,k)$, $\mathbf{L}(\ell,k)$, and $\mathbf{B}(\ell,k)$ "generally" correspond to the transition rates from states $(\ell, j-1)$, (ℓ,j), and $(\ell,j+1)$, respectively, to state (k,j) for any $\ell,k \in \{0,1,\ldots,b\}$ and $j \geq 1$. The only exceptions to this correspondence are the diagonal entries of $\mathbf{L}$, which are equal to the negative of the sum of the outflow rates from any state (ℓ,j). Meanwhile, the matrices $\mathbf{F}_0$ and $\mathbf{L}_0$ play the similar role as $\mathbf{F}$ and $\mathbf{L}$ (respectively) for the initial non-repeating portion of the chain. We now write the balance equations as matrix equations as follows:

$$\begin{cases} \vec{0} = \vec{\pi}_0 \cdot \mathbf{L_0} + \vec{\pi}_1 \cdot \mathbf{B} & \\ \vec{0} = \vec{\pi}_0 \cdot \mathbf{F_0} + \vec{\pi}_1 \cdot \mathbf{L} + \vec{\pi}_2 \cdot \mathbf{B} & , \\ \vec{0} = \vec{\pi}_j \cdot \mathbf{F} + \vec{\pi}_{j+1} \cdot \mathbf{L} + \vec{\pi}_{j+2} \cdot \mathbf{B} \qquad j = 1,2,\cdots & \end{cases} \tag{EC.27}$$

where

$$\mathbf{F_0} = \begin{pmatrix} p_{\mathsf{m}}\lambda_{\mathsf{m}} & & & & \\ \mu_1 & p_{\mathsf{m}}\lambda_{\mathsf{m}} & & & \\ & \ddots & \ddots & & \\ & & \mu_1 & p_{\mathsf{m}}\lambda_{\mathsf{m}} & \\ & & & \mu_1 & p_{\mathsf{m}}\lambda_{\mathsf{m}} \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} p_{\mathsf{m}}\lambda_{\mathsf{m}} & & & & \\ & p_{\mathsf{m}}\lambda_{\mathsf{m}} & & & \\ & & \ddots & & \\ & & & p_{\mathsf{m}}\lambda_{\mathsf{m}} & \\ & & & & p_{\mathsf{m}}\lambda_{\mathsf{m}} \end{pmatrix},$$

$$\mathbf{L_0} = \begin{pmatrix} -\gamma_0 & \lambda_{\mathsf{w}} & & & \\ & -\gamma_1 & \lambda_{\mathsf{w}} & & \\ & & \ddots & \ddots & \\ & & & -\gamma_{b-1} & \lambda_{\mathsf{w}} \\ & & & & -\gamma_b \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} -\xi_0 & \lambda_{\mathsf{w}} & & & \\ & -\xi_1 & \lambda_{\mathsf{w}} & & \\ & & \ddots & \ddots & \\ & & & -\xi_{b-1} & \lambda_{\mathsf{w}} \\ & & & & -\xi_b \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mu_2 & & & & \\ & \mu_2 & & & \\ & & \ddots & & \\ & & & \mu_2 & \\ & & & & \mu_2 \end{pmatrix},$$

$$\gamma_i = \begin{cases} p_{\mathsf{m}}\lambda_{\mathsf{m}} + \lambda_{\mathsf{w}} & i = 0 \\ \mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}} + \lambda_{\mathsf{w}} & 1 \leq i \leq b-1 \\ \mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}} & i = b \end{cases}, \qquad \xi_i = \begin{cases} p_{\mathsf{m}}\lambda_{\mathsf{m}} + \lambda_{\mathsf{w}} + \mu_2 & 0 \leq i \leq b-1 \\ p_{\mathsf{m}}\lambda_{\mathsf{m}} + \mu_2 & i = b \end{cases}. \tag{EC.28}$$

We aim to find a matrix $\mathbf{R} \in \mathbb{R}^{(b+1)\times(b+1)}$ such that $\vec{\pi}_j = \vec{\pi}_1 \mathbf{R}^{j-1}$ $\forall j \geq 1$. Following the standard theory of matrix analytic methods, this matrix satisfies the following matrix-

quadratic equation, which we proceed to solve in $\mathbf{R}$:

$$\mathbf{F}+\mathbf{R}\mathbf{L}+\mathbf{R}^2\mathbf{B}=\mathbf{0}, \tag{EC.29}$$

where $\mathbf{0}$ is the $(b+1)\times(b+1)$ square zero matrix.

We let $\mathbf{R}(i,j)$ denote the (i,j)-th element of $\mathbf{R}, \forall i,j\in\{0,1,\ldots,b\}$ and observe that $\mathbf{R}$ is an upper triangular matrix (as all phase-transitions in the infinite repeating portion of the CTMC of interest are unidirectional). Consequently, $\mathbf{R}(i,j)=0$, whenever $0\le j<i\le b$. By rewriting the matrix-quadratic Eq. (EC.29) into the corresponding system of component-wise (scalar) quadratic equations, we observe that for all $i\in\{0,1,\ldots,b\}$, the i-th diagonal element of $\mathbf{R}$, $\mathbf{R}(i,i)$, is the (lesser) solution to a the single (scalar) quadratic equation $\mu_2\mathbf{R}(i,i)^2-\xi_i\mathbf{R}(i,i)+p_{\mathsf{m}}\lambda_{\mathsf{m}}=0$ (we discard the greater solution as it exceeds 1). Hence, $\mathbf{R}(i,i)=\left(\xi_i-\sqrt{\xi_i^2-4p_{\mathsf{m}}\lambda_{\mathsf{m}}\mu_2}\right)\Big/2\mu_2$. We note that all elements of the diagonal of $\mathbf{R}$ are actually the same except for the last, $\mathbf{R}(b,b)$.

After determining all the elements on the diagonal of $\mathbf{R}$, let e_i denote the i-th unit vector. We can compute each value of the super-diagonal of $\mathbf{R}$ by solving the following system of linear equations:

$$\begin{cases}\lambda_{\mathsf{w}}\mathbf{R}(i,j-1)-\xi_j\mathbf{R}(i,j)+\mu_2(e_i^T\mathbf{R}^2e_j)=0 & 1\le j\le b-1\\ \lambda_{\mathsf{w}}\mathbf{R}(i,j-1)-(p_{\mathsf{m}}\lambda_{\mathsf{m}}+\mu_2)\mathbf{R}(i,j)+\mu_2(e_i^T\mathbf{R}^2e_j)=0 & j=b\end{cases}.$$

As long as the values of this super-diagonal are determined, we can compute the "super-diagonal" of this super-diagonal following the same procedure; finally, all other elements of $\mathbf{R}$ can be determined recursively in closed form.

Example of finding the closed form solution of R when $b=2$. We first solve the diagonal element of the matrix R, which gives us $\mathbf{R}(i,i)=\left(\xi_i-\sqrt{\xi_i^2-4p_{\mathsf{m}}\lambda_{\mathsf{m}}\mu_2}\right)\Big/2\mu_2$, for $i=0,1,2$. Note that $\mathbf{R}(0,0)=\mathbf{R}(1,1)$ and $\mathbf{R}(2,2)$ can be further simplified as $\mathbf{R}(2,2)=p_{\mathsf{m}}\lambda_{\mathsf{m}}/\mu_2$. Then using the linear equations described above, we solve the super-diagonal elements ($\mathbf{R}(0,1)$ and $\mathbf{R}(1,2)$), finally, we derive $\mathbf{R}(0,2)$ in the closed form as well. We summarize the closed form solution of each element of the matrix $\mathbf{R}$ in this specific case as follows:

$$\begin{cases}\mathbf{R}(0,0)=\frac{\left(\xi_0-\sqrt{\xi_0^2-4p_{\mathsf{m}}\lambda_{\mathsf{m}}\mu_2}\right)}{2\mu_2}\\ \mathbf{R}(1,1)=\frac{\left(\xi_1-\sqrt{\xi_1^2-4p_{\mathsf{m}}\lambda_{\mathsf{m}}\mu_2}\right)}{2\mu_2}\\ \mathbf{R}(2,2)=p_{\mathsf{m}}\lambda_{\mathsf{m}}/\mu_2\end{cases}\quad\begin{cases}\mathbf{R}(0,1)=\frac{\lambda_{\mathsf{w}}\left(\xi_1-\sqrt{\xi_1^2-4p_{\mathsf{m}}\lambda_{\mathsf{m}}\mu_2}\right)}{2\mu_2\sqrt{\xi_1^2-4p_{\mathsf{m}}\lambda_{\mathsf{m}}\mu_2}}\\ \mathbf{R}(1,2)=\frac{\lambda_{\mathsf{w}}\left(\xi_1-\sqrt{\xi_1^2-4p_{\mathsf{m}}\lambda_{\mathsf{m}}\mu_2}\right)}{2\mu_2^2-\mu_2\left(\xi_1-\sqrt{\xi_1^2-4p_{\mathsf{m}}\lambda_{\mathsf{m}}\mu_2}\right)}\\ \mathbf{R}(0,2)=\frac{2\lambda_{\mathsf{w}}^2\left(\xi_1-\sqrt{\xi_1^2-4p_{\mathsf{m}}\lambda_{\mathsf{m}}\mu_2}\right)}{\sqrt{\xi_1^2-4p_{\mathsf{m}}\lambda_{\mathsf{m}}\mu_2}\left(2\mu_2-\xi_1+\sqrt{\xi_1^2-4p_{\mathsf{m}}\lambda_{\mathsf{m}}\mu_2}\right)^2}\end{cases}\quad\begin{cases}\mathbf{R}(1,0)=0\\ \mathbf{R}(2,0)=0\\ \mathbf{R}(2,1)=0\end{cases}$$

where

$$\xi_i = \begin{cases} p_{\mathsf{m}}\lambda_{\mathsf{m}} + \lambda_{\mathsf{w}} + \mu_2 & \forall i \in \{0,1\} \\ p_{\mathsf{m}}\lambda_{\mathsf{m}} + \mu_2 & i = 2 \end{cases}.$$

Finally, from the first two equations in Eq. (EC.27) we have that

$$\left[\vec{\pi}_0 \ \vec{\pi}_1\right] \begin{bmatrix} \mathbf{L}_0 & \mathbf{F}_0 \\ \mathbf{B}_0 & \mathbf{L}+\mathbf{RB} \end{bmatrix} = \vec{0},$$

which we can combine with the normalizing equation (i.e. the sum of all the limiting probabilities is equal to one) to find the initial limiting probabilities $\vec{\pi}_0$ and $\vec{\pi}_1$ (see Eq. 21.5 in Harchol-Balter 2013). Hence, the limiting probabilities $\pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j)$ and $\pi^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}(i,j)$ are all determined and their associated series such as $\sum_{j=0}^{\infty}\pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j)$ and $\sum_{j=0}^{\infty} j\pi^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}(i,j)$ can all be computed as follows (for any policy $\mathsf{P} \in \{\mathsf{MWO}, \mathsf{WMO}\}$):

$$\sum_{j=0}^{\infty}\pi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j) = \left(\vec{\pi}^{\mathsf{P}}_0 + \sum_{j=1}^{\infty}\vec{\pi}^{\mathsf{P}}_1\mathbf{R}^{j-1}\right)e_i = \left(\vec{\pi}^{\mathsf{P}}_0 + \sum_{j=0}^{\infty}\vec{\pi}^{\mathsf{P}}_1\mathbf{R}^{j}\right)e_i = \left(\vec{\pi}^{\mathsf{P}}_0 + \vec{\pi}^{\mathsf{P}}_1(\mathbf{I}-\mathbf{R})^{-1}\right)e_i,$$

$$\sum_{j=0}^{\infty} j\pi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j) = \vec{\pi}^{\mathsf{P}}_1\sum_{j=1}^{\infty} j\mathbf{R}^{j-1}e_i = \vec{\pi}^{\mathsf{P}}_1\frac{d}{d\mathbf{R}}\left(\sum_{j=0}^{\infty}\mathbf{R}^{j}\right)e_i = \vec{\pi}^{\mathsf{P}}_1(\mathbf{I}-\mathbf{R})^{-2}e_i.$$

EC.3.2. The Limiting Probabilities $\phi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,j)$

The quantities $\phi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,j)$, $i \in \{0,1,\ldots,b\}$ and $j \in \{0,1\}$, are the limiting probabilities of a finite state CTMC (see Fig. 4b), so we can find them by solving the balance equations below (where for simplicity we use the notation $\phi_{i,j} \equiv \phi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,j)$):

$$\begin{cases} \lambda_{\mathsf{w}}\phi_{0,0} = \mu_2\phi_{0,1} \\ (\lambda_{\mathsf{w}}+\mu_1)\phi_{i,0} = \lambda_{\mathsf{w}}\phi_{i-1,0} + \mu_2\phi_{i,1} & \forall i \in \{1,2,\ldots,b-1\} \\ \mu_1\phi_{b,0} = \lambda_{\mathsf{w}}\phi_{b-1,0} + \mu_2\phi_{b,1} \\ (\lambda_{\mathsf{w}}+\mu_2)\phi_{0,1} = \mu_1\phi_{1,0} \\ (\lambda_{\mathsf{w}}+\mu_2)\phi_{i,1} = \lambda_{\mathsf{w}}\phi_{i-1,1} + \mu_1\phi_{i+1,0} & \forall i \in \{1,2,\ldots,b-1\} \\ \mu_2\phi_{b,1} = \lambda_{\mathsf{w}}\phi_{b-1,1} \\ \sum_{i=0}^{b}(\phi_{i,0}+\phi_{i,1}) = 1 \end{cases}. \tag{EC.30}$$

EC.3.3. The Laplace Transforms and Moments of U and V

In this section we give a procedure for determining the Laplace transforms of U and V in closed form. We denote the transforms by $\widetilde{U}(s) \equiv \mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}\left[e^{-sU}\right]$ and $\widetilde{V}(s) \equiv \mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}\left[e^{-sV}\right]$ (all transforms in this appendix implicitly depend on the strategy profile (b, p_{m}), but we omit the reference to strategy profile in our notation in the interest of brevity). One can determine the first and second moments of U and V from the Laplace transforms readily from standard formulas (given at the end of this section). Alternatively, similar techniques used for computing the Laplace transforms can be used to compute the first moments directly, and subsequently, the second moments as well.

Finding $\widetilde{U}(s)$. Recall that U is the time until a mobile arrival that enters a *mobile-less* system (i.e., a system that has no mobiles)—but a steady-state number of walk-ins in each stage conditioned on the fact that there are no mobiles in the system—will ultimately leave the system. It follows that U depends on the system state at the time of the mobile's arrival. There are $2(b+1)$ such states, as the number of $\mathbf{O}s$ in the system, $N_1 \in \{0, 1, \ldots, b\}$, while the number of $\mathbf{W}s$ in the system, $N_{2,\mathsf{w}} \in \{0, 1\}$. Therefore, we can define random variables $U_{i,j} \sim (U \,|N_1 = i, N_{2,\mathsf{w}} = j\,)$. If we can find the probability that $N_1 = i$ and $N_{2,\mathsf{w}} = j$ at the time of a mobile's arrival to a mobile-less system, and the distribution of $\widetilde{U_{i,j}}(s)$ for all $(i, j) \in \{0, 1, \ldots, b\} \times \{0, 1\}$, then we can determine $\widetilde{U}(s)$ by taking a standard mixture of transforms.

We first address the probability that $N_1 = i$ and $N_2 = N_{2,\mathsf{w}} = j$ at the time of a mobile's arrival to a mobile-less system. We can determine such probabilities as the limiting probabilities—which we denote by $\psi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i, j)$—of a CTMC. Consider the stochastic process that governs $(N_1, N_{2,\mathsf{w}})$ during the union of all time intervals (epochs) in which the system is mobile-less. As soon as a mobile would enter the system, we immediately "jump ahead" in time until the first moment in which the system is again memory-less; so the time intervals in question are closed on the left (i.e., at their lower bound in time) and open at the right (i.e., at their upper bound in time). That is, if a mobile would arrive, we instead transition directly to state $(0, 0)$, as the next time that the system is again mobile-less, there would not be any walk-ins of any kind in the system (as all walk-ins have preemptive priority over mobiles under WOM). Since mobiles arrive with rate $p_{\mathsf{m}}\lambda_{\mathsf{m}}$, the stochastic process governing $(N_1, N_{2,\mathsf{w}})$ during mobile-less epochs is a CTMC, which corresponds to the one depicted in Fig. 4b with the key difference that there is an additional transition (or

increased transition rate) from each non-$(0,0)$ state to state $(0,0)$ with rate (or increase in rate equal to) $p_{\mathsf{m}}\lambda_{\mathsf{m}}$; mobiles cannot of course arrive when we are in state $(0,0)$ as well, but in that case we would be back at $(0,0)$ at the start of the next mobile-less time epoch; so no transition is necessary as CTMCs do not have "self-loops" by standard convention.

The limiting probability distribution of the CTMC corresponds to $\psi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,j)$, as mobile arrivals are governed by a Poisson process that is independent of the state of this chain, and so the likelihood of a mobile arriving to a mobile-less system in a state where $N_1 = i$ and $N_{2,\mathsf{w}} = j$ is given by the corresponding limiting probability of this CTMC. These limiting probabilities can be computed by solving a system of linear equations that greatly resemble those corresponding to the system of linear equations that we solve to obtain $\phi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,j)$ probabilities (Eqs. (EC.30) in Appendix EC.3.2) with several differences: (i) The variable symbols contain a ψ rather than a ϕ, and more crucially in that (ii) the balance equations take into account the outgoing rate from each non-$(0,0)$ state (i,j) equal to $p_{\mathsf{m}}\lambda_{\mathsf{m}}\psi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,j)$, and (iii) there is an increased incoming rate to state $(0,0)$ equal to the some of all those rates; taking the normalization equation into account, this increase is equal to $p_{\mathsf{m}}\lambda_{\mathsf{m}}\left(1-\psi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(0,0)\right)$. Hence, we can obtain the limiting probabilities of interest by solving the system equations below (where for simplicity we use the notation $\psi_{i,j} \equiv \psi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,j)$):

$$\begin{cases} \lambda_{\mathsf{w}}\psi_{0,0} = \mu_2\psi_{0,1} + p_{\mathsf{m}}\lambda_{\mathsf{m}}\left(1-\psi_{0,0}\right) & \\ \left(\lambda_{\mathsf{w}}+\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}\right)\psi_{i,0} = \lambda_{\mathsf{w}}\psi_{i-1,0} + \mu_2\psi_{i,1} & \forall i \in \{1,2,\ldots,b-1\} \\ \left(\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}\right)\psi_{b,0} = \lambda_{\mathsf{w}}\psi_{b-1,0} + \mu_2\psi_{b,1} & \\ \left(\lambda_{\mathsf{w}}+\mu_2+p_{\mathsf{m}}\lambda_{\mathsf{m}}\right)\psi_{0,1} = \mu_1\psi_{1,0} & \\ \left(\lambda_{\mathsf{w}}+\mu_2+p_{\mathsf{m}}\lambda_{\mathsf{m}}\right)\psi_{i,1} = \lambda_{\mathsf{w}}\psi_{i-1,1} + \mu_1\psi_{i+1,0} & \forall i \in \{1,2,\ldots,b-1\} \\ \left(\mu_2+p_{\mathsf{m}}\lambda_{\mathsf{m}}\right)\psi_{b,1} = \lambda_{\mathsf{w}}\psi_{b-1,1} & \\ \sum_{i=0}^{b}\left(\psi_{i,0}+\psi_{i,1}\right) = 1 & \end{cases} . \tag{EC.31}$$

Next, we turn to the task of finding $\widetilde{U_{i,j}}(s)$, which we shall also present as the solution to a linear system of equations (with symbolic coefficients). First, see that $U_{0,0} = 0$, as if a mobile arrives to an empty system, it immediately goes into service. In all other cases, $U_{i,j}$ corresponds to the time it takes for a system currently in a state where $N_1 = i$ and $N_{2,\mathsf{w}} = j$

to be empty of all its walk-ins, without regard for any mobile arrivals (since any mobile arrivals will have lower priority than the original mobile arrival). That is, $U_{i,j}$ is distributed like the time it takes to enter state $(0,0)$ of the Markov chain depicted in Fig. 4b, given that we initially start in state (i,j). Note that this is the original Markov chain and not the modified one with additional transitions to state $(0,0)$ that we described earlier in our procedure for finding $\psi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,j)$.

Now that we can interpret the $U_{i,j}$ random variables as the hitting times of a finite state Markov chain, it is straightforward to write a system of linear equations for the transforms of interest using first-step analysis. Recall that the Laplace transform of an exponential random variable with rate κ is $\kappa/(\kappa+s)$ and that the minimum of two exponential random variables $\text{Exp}(\eta)$ and $\text{Exp}(\kappa)$ is distributed as $\text{Exp}(\kappa+\eta)$. Then, we have:

$$\begin{cases} \widetilde{U_{0,0}}(s) = 1 \\ \widetilde{U_{i,0}}(s) = \dfrac{\lambda_{\mathsf{w}}+\mu_1}{s+\lambda_{\mathsf{w}}+\mu_1}\left(\dfrac{\lambda_{\mathsf{w}}}{\lambda_{\mathsf{w}}+\mu_1}\widetilde{U_{i+1,0}}(s)+\dfrac{\mu_1}{\lambda_{\mathsf{w}}+\mu_1}\widetilde{U_{i-1,1}}(s)\right) & \forall i \in \{1,2,\ldots,b-1\} \\ \widetilde{U_{b,0}}(s) = \dfrac{\mu_1}{s+\mu_1}\widetilde{U_{b-1,1}}(s) & \\ \widetilde{U_{i,1}}(s) = \dfrac{\lambda_{\mathsf{w}}+\mu_2}{s+\lambda_{\mathsf{w}}+\mu_2}\left(\dfrac{\lambda_{\mathsf{w}}}{\lambda_{\mathsf{w}}+\mu_2}\widetilde{U_{i+1,1}}(s)+\dfrac{\mu_2}{\lambda_{\mathsf{w}}+\mu_2}\widetilde{U_{i,0}}(s)\right) & \forall i \in \{0,1,\ldots,b-1\} \\ \widetilde{U_{b,1}}(s) = \dfrac{\mu_2}{s+\mu_2}\widetilde{U_{b,0}}(s) & \end{cases} . \tag{EC.32}$$

Solving the above system of equations will yield all of the $\widetilde{U_{i,j}}(s)$ in closed form. Together with the $\psi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,j)$ values, we can determine $\widetilde{U}(s)$ by taking the appropriate weighted sum:

$$\widetilde{U}(s) = \sum_{i=0}^{b} \widetilde{U_{i,0}}(s)\psi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,0) + \widetilde{U_{i,1}}(s)\psi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,1). \tag{EC.33}$$

Finding $\widetilde{V}(s)$. Recall that $V \sim (T_{\mathsf{m}} | N_1 = 0, N_2 = 0)$ under WOM. Once service begins on a mobile, we know that there are currently no walk-ins in the system. One of two events will happen, either (i) a walk-in will arrive to Stage 1 interrupting the service of the mobile until there are again no walk-ins in the system, or (ii) the mobile will be served before any walk-ins arrive. Under case (i), the process that interrupts the mobile will be distributed like $U_{1,0}$, and once the mobile resumes service its expected remaining service time is again

distributed like an independent copy of V (due to the memoryless property). Formalizing the first-step analysis described above, we have:

$$\begin{aligned}\widetilde{V}(s) &= \frac{\lambda_{\mathsf{w}}+\mu_2}{s+\lambda_{\mathsf{w}}+\mu_2}\left(\frac{\lambda_{\mathsf{w}}}{\lambda_{\mathsf{w}}+\mu_2}\widetilde{U_{1,0}}(s)\widetilde{V}(s)+\frac{\mu_2}{\lambda_{\mathsf{w}}+\mu_2}\right)\\ \implies \widetilde{V}(s) &= \frac{\mu_2}{s+\lambda_{\mathsf{w}}\left(1-\widetilde{U_{1,0}}(s)\right)+\mu_2}.\end{aligned} \tag{EC.34}$$

Finally, the moments of U and V can be obtained by using the standard technique which gives the first and second moments of a random variable X—with well defined Laplace transform $\widetilde{X}(s)$—to be $\lim_{s\to 0^+} X'(s) = -\mathbb{E}[X]$ and $\lim_{s\to 0^+} X''(s) = \mathbb{E}\left[X^2\right]$, respectively.

A computationally efficient technique for finding the first and second moments of U and V. Rather than compute $\widetilde{U}(s)$ and $\widetilde{V}(s)$, if we are only interested in the first and second moments of of U and V (which is the case for finding the sojourn times of interest in this paper), we can use the standard technique for finding moments from transforms (described above) to each equation in the system (EC.32) directly, yielding a new system (where we use the shorthand $\mathbb{E}[U_{i,j}]$ for $\mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[U_{i,j}]$:

$$\begin{cases}
\mathbb{E}[U_{0,0}] = 1 & \\
\mathbb{E}[U_{i,0}] = \dfrac{1+\lambda_{\mathsf{w}}\mathbb{E}[U_{i+1,0}]+\mu_1\mathbb{E}[U_{i-1,1}]}{\lambda_{\mathsf{w}}+\mu_1} & \forall i\in\{1,2,\ldots,b-1\}\\
\mathbb{E}[U_{b,0}] = \dfrac{1}{\mu_1}+\mathbb{E}[U_{b-1,1}] & \\
\mathbb{E}[U_{i,1}] = \dfrac{1+\lambda_{\mathsf{w}}\mathbb{E}[U_{i+1,1}]+\mu_2\mathbb{E}[U_{i,0}]}{\lambda_{\mathsf{w}}+\mu_2} & \forall i\in\{0,1,\ldots,b-1\}\\
\mathbb{E}[U_{b,1}] = \dfrac{1}{\mu_2}+\mathbb{E}[U_{b,0}] & \\
\mathbb{E}\left[U_{0,0}^2\right] = 1 & \\
\mathbb{E}\left[U_{i,0}^2\right] = \dfrac{2+2\lambda_{\mathsf{w}}\mathbb{E}[U_{i+1,0}]+2\mu_1\mathbb{E}[U_{i-1,1}]}{(\lambda_{\mathsf{w}}+\mu_1)^2}+\dfrac{\lambda_{\mathsf{w}}\mathbb{E}\left[U_{i+1,0}^2\right]+\mu_1\mathbb{E}\left[U_{i-1,1}^2\right]}{\lambda_{\mathsf{w}}+\mu_1} & \forall i\in\{1,2,\ldots,b-1\}\\
\mathbb{E}\left[U_{b,0}^2\right] = \dfrac{2+2\mu_1\mathbb{E}[U_{b-1,1}]}{{\mu_1}^2}+\mathbb{E}\left[U_{b-1,1}^2\right] & \\
\mathbb{E}\left[U_{i,1}^2\right] = \dfrac{2+2\lambda_{\mathsf{w}}\mathbb{E}[U_{i+1,1}]+2\mu_2\mathbb{E}[U_{i,0}]}{(\lambda_{\mathsf{w}}+\mu_2)^2}+\dfrac{\lambda_{\mathsf{w}}\mathbb{E}\left[U_{i+1,1}^2\right]+\mu_2\mathbb{E}\left[U_{i,0}^2\right]}{\lambda_{\mathsf{w}}+\mu_2} & \forall i\in\{0,1,\ldots,b-1\}\\
\mathbb{E}\left[U_{b,1}^2\right] = \dfrac{2+2\mu_2\mathbb{E}[U_{b,0}]}{{\mu_2}^2}+\mathbb{E}\left[U_{b,0}^2\right] &
\end{cases}. \tag{EC.35}$$

After solving this system, we can find the first and second moments via standard conditioning:

$$\mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[U^n] = \sum_{i=0}^{b} \mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[U^n_{i,0}]\psi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,0) + \mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[U^n_{i,1}]\psi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,1),$$

where we are interested in the cases where $n \in \{1,2\}$. Similar methods yield:

$$\mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[V] = \frac{1+\lambda_{\mathsf{w}}\mathbb{E}[U_{1,0}]}{\mu_2}, \quad \mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}\left[V^2\right] = \frac{2\left(1+\lambda_{\mathsf{w}}\mathbb{E}[U_{1,0}]\right)^2 + \lambda_{\mathsf{w}}\mu_2\mathbb{E}\left[U^2_{1,0}\right]}{{\mu_2}^2}. \tag{EC.36}$$

EC.3.4. Approximating the Limiting Probabilities $\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)$ and an Associated Series

To determine the limiting probabilities of the CTMC of Fig. 5a, $\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)$, we first observe that the chain has finitely many phases (rows) and infinitely many levels (columns). Moreover, phase transitions are *bidirectional* throughout the infinite portion of the chain, that is, we can transition to a higher row and a lower row from any phase. Such chains do not often lend themselves to exact analysis; so, we opt to approximate the probabilities via numerical matrix analytic methods.

We first define the three square matrices $\mathbf{F}, \mathbf{L}$, and $\mathbf{B} \in \mathbb{R}^{(b+1)\times(b+1)}$ such that (using zero-based numbering so that the upper left element of any matrix $\mathbf{M}$ is denoted by $\mathbf{M}(0,0)$) $\mathbf{F}(\ell,k)$, $\mathbf{L}(\ell,k)$, and $\mathbf{B}(\ell,k)$ "generally" correspond to the transition rates from states $(\ell, j-1)$, (ℓ,j), and $(\ell,j+1)$, respectively, to state (k,j) for any $\ell,k \in \{0,1,\ldots,b\}$ and $j \geq 1$. The only exceptions to this correspondence are the entries $\mathbf{L}(\ell,k)$ when $\ell = k$. In these cases, $\mathbf{L}(\ell,k) = \mathbf{L}(\ell,\ell)$ is equal to the negative of the sum of the outflow rates from state (ℓ,j). Thus, for the CTMC of Fig. 5a, $\mathbf{B}$ and $\mathbf{F}$ has the same structures as $\mathbf{B}$ and $\mathbf{F_0}$ in Eq. (EC.28), respectively, and $\mathbf{L}$ follows:

$$\mathbf{L} = \begin{pmatrix} -\nu_0 & \lambda_{\mathsf{w}} & & & \\ & -\nu_1 & \lambda_{\mathsf{w}} & & \\ & & \ddots & \ddots & \\ & & & -\nu_{b-1} & \lambda_{\mathsf{w}} \\ & & & & -\nu_b \end{pmatrix}, \quad \text{where} \quad \nu_i = \begin{cases} p_{\mathsf{m}}\lambda_{\mathsf{m}} + \lambda_{\mathsf{w}} + \mu_2 & i = 0 \\ \mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}} + \lambda_{\mathsf{w}} + \mu_2 & 1 \leq i \leq b-1 \\ \mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}} + \mu_2 & i = b \end{cases}. \tag{EC.37}$$

We would like to express the limiting probabilities of interest in terms of a square matrix $\mathbf{R} \in \mathbb{R}^{(b+1)\times(b+1)}$ that satisfies Eq. (EC.29). In general, we cannot find $\mathbf{R}$ in closed form, so we resort to a procedure where we iteratively calculate $\mathbf{R}_{n+1} = -(\mathbf{R}_n^2\mathbf{B} + \mathbf{F})\mathbf{L}^{-1}$ (here $\mathbf{R}_n$

denotes the n-th iteration of $\mathbf{R}$) until $||\mathbf{R}_{n+1}-\mathbf{R}_n|| < \epsilon$ (here we define the metric $||\cdot||$ to be the maximum of all the elements in the matrix), for any arbitrary given ϵ. The associated series can be computed in the similar way as in Appendix EC.3.1.

EC.3.5. The Transient Probabilities $P(u,v,w;\rho)$

Individual probabilities of the form $P(u,v,w;\rho)$ can be computed exactly in a recursive fashion from the following relations due to Kaczynski et al. (2012):

$$\begin{cases} P(u,u,u+v;\rho) = \left(\dfrac{\rho}{\rho+1}\right)^{v} & u \geq 1, v \geq 1 \\ P(0,v,v) = \left(\dfrac{\rho}{\rho+1}\right)^{v-1} & v \geq 1 \\ P(u,1,w;\rho) = \dfrac{\rho}{(\rho+1)^{u-w+2}} & 2 \leq w \leq u \\ P(u,v,w;\rho) = \dfrac{\rho}{\rho+1} \displaystyle\sum_{j=w-1}^{u+v-1} \left(\dfrac{1}{\rho+1}\right)^{j-w+1} P(u,v-1,j;\rho) & v \geq 2 \text{ and } 2 \leq w \leq u+v-1 \end{cases}.$$

In the interest of computational efficiency, it is advisable to use a "memoization" approach when computing a set of probabilities.

EC.3.6. Approximating $Y(i,j)$ and an Associated Series

We cannot determine $Y(i,j)$ in closed form so we rely on truncation. Truncating the first summation (by summing from $k=0$ to K instead of $k=0$ to ∞) in the expression giving $Y(i,j)$ (i.e., Eq. (4)) at K and denoting the result by $Y_K(i,j)$ given by

$$Y_K(i,j) \equiv \left(1-\frac{p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}}\right)^{i+1} \sum_{k=0}^{K} \sum_{\ell=1}^{i+j+k+1} \frac{\ell}{\mu_2} P\left(j, i+k+1, \ell; \frac{\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_2}\right) \binom{k+i}{k} \left(\frac{p_{\mathsf{m}}\lambda_{\mathsf{m}}}{\mu_1+p_{\mathsf{m}}\lambda_{\mathsf{m}}}\right)^{k},$$

it follows that $Y_K(i,j) \to Y(i,j)$ as $K \to \infty$, and so $Y(i,j) \approx Y_K(i,j)$ for sufficiently large K values. Based on our exploration of different parameters, it appears that $|Y_{K+1}(i,j) - Y_K(i,j)|$ is negligible for values of K on the order of 20, suggesting that the approximation is adequate when K is on that order.

Similarly, we approximate the following series involving $Y(i,j)$ via "double truncation" for sufficiently large J and K values (where it may or may not be appropriate to set $J=K$ based on the parameters). We have:

$$\sum_{j=0}^{\infty} Y(i,j)\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j) \approx \sum_{j=0}^{J} Y_K(i,j)\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j).$$

Of course, since we generally do not know $\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)$ exactly, we compute the above approximation in terms of the approximated (rather than exact) $\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)$ values.

EC.3.7. Approximating the Limiting Probabilities $\phi^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}(i,j)$ and an Associated Series

We can approximate the limiting probabilities of the CTMC shown in Fig. 5b, $\phi^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}(i,j)$, by using the same approach we used to determine the $\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)$ values (see Appendix EC.3.4), with the only difference that we set $p_{\mathsf{m}}=0$ everywhere (regardless of the actual value of p_{m}, which $\phi^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}(i,j)$ does not depend on) as the Fig. 5b chain is a special case of the Fig. 5a chain where $p_{\mathsf{m}}=0$. As a result, we start with a modified $\mathbf{F}$ matrix with zero entries for its main diagonal. We follow the rest of the procedure in the exact same way. The limiting probabilities yielded by this procedure will be (an approximation of) $\phi^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}(i,j)$, and the series computed will be (an approximation of) $\sum_{j=0}^{\infty} j\phi^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}(i,j)$.

EC.3.8. Approximating $\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,j)]$ and an Associated Series

First observe that $Z(i,j)$ corresponds to the "hitting time" associated with reaching a state of the form $(k,0)$ (for any value of $k\in\{0,1,\ldots,b\}$) starting at an initial state (i,j) in the CTMC shown in Fig. 5b. Now, assume that we are in some state (ℓ,m) where $m\geq 1$, and consider the first time we reach $(k,m-1)$ for any $k\in\{0,1,\ldots,b\}$ (i.e., the first time $N_{2,\mathsf{w}}$ drops from its initial value of m). Let τ_ℓ be the expected "hitting time" (duration) associated with this trip from (ℓ,m) to some $(k,m-1)$, and for any specific value of $k\in\{0,1,\ldots,b\}$, let $p_{\ell\to k}$ be the probability with which we specifically end up in $(k,m-1)$ at the conclusion of this trip (i.e., we reach $(k,m-1)$ before reaching $(k',m-1)$ for any $k'\neq k$). As our notation suggests, these quantities are well-defined for any $m\geq 1$, and do not otherwise depend on the particular value of m (i.e., the initial level or $N_{2,\mathsf{w}}$ value is irrelevant); this fact is easily confirmed by considering the repeating nature of the CTMC of Fig. 5b

With the τ_ℓ and $p_{\ell\to k}$ quantities, we can determine $\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,j)]$ via first-step analysis. First, note from the definition of $Z(i,j)$ and τ_ℓ that we readily have $\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,1)]=\tau_i$, $\forall i\in\{0,1,\ldots,b\}$. Meanwhile, when examining $Z(i,j)$ for any value of $j>1$, we string together trips that drop the phase number ($N_{2,\mathsf{w}}$) by one while taking into account the distribution over the level (N_1) that we are in at the conclusion of each phase drop. Hence, we have the following relations:

$$\begin{cases} \mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,1)]=\tau_i & 0\leq i\leq b, \\ \mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,j)]=\tau_i+\sum_{k=0}^{b}(p_{i\to k})\,\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,j-1)] & 0\leq i\leq b,\ 1\leq j \end{cases}. \tag{EC.38}$$

We can solve for any $\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,1)]$ values in closed form in terms of the τ_ℓ and $p_{\ell\to k}$ values; a "memoization" approach is advisable. We note that this can become cumbersome for large values of j, so from a computational efficiency perspective it is preferable to have numerical values for τ_ℓ and $p_{\ell\to k}$. We now address how to derive these values.

We proceed by deriving a system of equations relating the τ_ℓ values to one another in terms of the $p_{\ell\to m}$ values via a straightforward application of the first step analysis:

$$\begin{cases} \tau_0 = \dfrac{1}{\lambda_{\mathsf{w}}+\mu_2} + \dfrac{\lambda_{\mathsf{w}}}{\lambda_{\mathsf{w}}+\mu_2}\tau_1 \\ \tau_\ell = \dfrac{1}{\lambda_{\mathsf{w}}+\mu_1+\mu_2} + \dfrac{\lambda_{\mathsf{w}}}{\lambda_{\mathsf{w}}+\mu_1+\mu_2}\tau_{\ell+1} + \dfrac{\mu_1}{\lambda_{\mathsf{w}}+\mu_1+\mu_2}\left(\tau_{\ell-1} + \displaystyle\sum_{k=0}^{b}\left(p_{(\ell-1)\to k}\right)\tau_k\right) & 1 \le \ell \le b-1 \\ \tau_b = \dfrac{1}{\mu_1+\mu_2} + \dfrac{\mu_1}{\mu_1+\mu_2}\left(\tau_{b-1} + \displaystyle\sum_{k=0}^{b}\left(p_{(b-1)\to k}\right)\tau_k\right) \end{cases}. \tag{EC.39}$$

It turns out that Eq. (EC.39) is a finite system of equations that are linear in the τ_ℓ values, which we can easily solve for in closed form, this time in terms of the $p_{\ell\to k}$ probabilities. Unfortunately, the $p_{\ell\to k}$ probabilities cannot generally be determined in closed-form in terms of elementary functions, as writing a system of equations relating these values to one another will involve nonlinear terms and solving the system will require solving higher ordered polynomials (the order of which can be arbitrary high based on the value of b). Therefore, we resort to approximating the $p_{\ell\to k}$ probabilities numerically.

In order to approximate the $p_{\ell\to k}$ probabilities numerically, we invoke the notion of the $\mathbf{G}$ matrix from the literature on matrix analytic methods (for a comprehensive discussion of the $\mathbf{G}$ matrix, see the chapter 6 of the standard textbook (Latouche and Ramaswami (1999)). The $\mathbf{G}$ matrix associated with a quasi-birth–death process (such as those depicted in Fig. 5) is a square matrix with a number of rows and columns equal to the number of phases and levels of the chain in question such that (using zero-based numbering so that we start with row 0) the entry in row ℓ and column k of $\mathbf{G}$ corresponds precisely to $p_{\ell\to k}$ as we have defined it above. That is, $p_{\ell\to k} = \mathbf{G}(\ell,k)$, so it remains to approximate $\mathbf{G}$. There are a variety of ways to carry out the task in the literature, but for the purpose of our discussion the most straightforward (although not necessarily most efficient) approach is likely to use the relation:

$$\mathbf{G} = -\mathbf{F}^{-1}\left(\mathbf{R}^{-1}\mathbf{F} - \mathbf{L}\right), \tag{EC.40}$$

where $\mathbf{F}$ and $\mathbf{L}$ are matrices associated with the Markov chain of interest and $\mathbf{R}$ is the rate matrix. $\mathbf{F}$ and $\mathbf{L}$ are given in Eq. (EC.37) for the more general CTMC of Fig. 5a; we need only modify $\mathbf{F}$ for the CTMC of Fig. 5b by replacing the its main diagonal entries with zeroes. Approximating $\mathbf{G}$ turns out to be straightforward once we identify $\mathbf{F}$, $\mathbf{L}$, and $\mathbf{B}$ and use them to approximate the $\mathbf{R}$ matrix (on this, see Appendices EC.3.4 and EC.3.7).

Finally, putting everything together and proceeding in roughly reverse order of the presentation of our discussion in this section, we can find the $\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,j)]$ as follows:

1. Identify $\mathbf{F}$, $\mathbf{B}$, and $\mathbf{L}$, as given in Eq. (EC.37), with the modification that the main diagonal of $\mathbf{F}$ should be replaced with zeros.
2. Use $\mathbf{F}$, $\mathbf{B}$, and $\mathbf{L}$ to compute $\mathbf{R}$ following the procedure given in Appendix EC.3.4.
3. Use Eq. (EC.40) to compute $\mathbf{G}$.
4. Solve the linear system Eq. (EC.39) to obtain all of the τ_ℓ values based on $\mathbf{G}$ (recall that $p_{\ell\to k} = \mathbf{G}(\ell,k)$).
5. Use the recursive relations given in Eq. (EC.38), to compute any of the $\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,j)]$ of interest (say $\forall i \in \{0,1,\ldots,b\}$ and $j \in \{1,2,\ldots,J\}$ for some J).

We note that step 2 is the only step that is not based on one or more exact relations, i.e., it introduces an approximation, resulting in an inexact value for $\mathbf{R}$. Consequently, all calculations based directly or indirectly on $\mathbf{R}$—namely, $\mathbf{G}$, the τ_ℓ values, and the $\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,j)]$ values—will also all be approximations. We also note that $\mathbb{E}^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}[Z(i,j)]$ is actually constant in p_{m} as mobile arrivals do not affect this quantity.

Finally, in the absence of better alternatives, the following series (which depends on the index i) can be approximated by truncation:

$$\sum_{j=0}^{\infty} \pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)\mathbb{E}_{(b,p_{\mathsf{m}})}[Z(i,j+1)] \approx \sum_{j=0}^{J} \pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)\mathbb{E}_{(b,p_{\mathsf{m}})}[Z(i,j+1)],$$

for sufficiently large J (where the right-hand side converges to the left-hand side as $J \to \infty$). We also note that as we generally do not know $\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)$ exactly, we compute the approximation for this series in terms of the approximated (rather than exact) $\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)$ values.

EC.4. Mixed walk-in strategies and heterogeneous patience levels

In this section we relax the assumption that indifferent walk-ins will always join, by allowing walk-ins to pursue a mixed strategy. This generalization will be essential in addressing the

case of walk-ins with heterogeneous patience levels. In this section, we address both the single- and two-server settings, but we will primarily focus on the former, where we provide a systematic method for determining such equilibria, although in some problem instances we can only find approximate equilibria under WMO.

EC.4.1. Mixed Walk-In Strategies

Throughout Section 5, we assume that the strategy (behavior) employed by walk-ins is described by a single integer value, b. Specifically, in that section, we assume that if a walk-in observes $N_1 = i < b$ other walk-ins in Stage 1 upon arrival they will join, and otherwise they will balk. We now consider a more general mixed strategy on the part of walk-ins, where for each non-negative integer i, we denote by p_i, the fraction (probability) of walk-in customers who opt to join given that they observe $N_1 = i$ other walk-ins in Stage 1 upon arrival. Letting $b \equiv \arg\min_{i \in \mathbb{Z}_{\geq 0}} p_i = 0$, we once again have b as a threshold on N_1 at which *no* walk-ins join. A pure walk-in strategy described by b corresponds precisely to the mixed walk-in strategy where $p_0 = p_1 = \cdots = p_{b-1} = 1$ and $p_b = 0$; i.e., the strategy b corresponds to a $\mathbf{p}_\mathsf{w}$ that is a vector of length b, with each entry is equal to 1. It follows that the space of walk-in strategies is formally given by

$$\mathcal{S} \equiv \bigcup_{b=0}^{\infty} \prod_{i=0}^{b-1} (0,1],$$

where $\prod$ denotes the generalized Cartesian product. Note that $\mathbf{p}_\mathsf{w}$ is the "empty vector" (which we can denote by $\varnothing$) when $b = 0$. We could equivalently consider strategies coming from the space $\prod_{i=0}^{\infty}[0,1]$, which would include "redundant" strategies as whenever $p_i = 0$, the values of p_k where $k > i$ are inconsequential.

We use $(\mathbf{p}_\mathsf{w}, p_\mathsf{m})$ to denote the strategy profile describing the behavior of both walk-ins and mobiles, where the interpretation of p_m remains unchanged (i.e., p_m is the fraction of mobile arrivals who join). Note that the strategy profile $(\mathbf{p}_\mathsf{w}, p_\mathsf{m})$ implies a value for b as well (i.e., b is the number of entries in the vector $\mathbf{p}_\mathsf{w}$). In this setting, given a policy P, an *equilibrium* is a joint-strategy, $(\mathbf{p}^*_\mathsf{w}, p^*_\mathsf{m})$, which satisfies

$$\begin{aligned}
\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}^*_\mathsf{w}, p^*_\mathsf{m})}[T_\mathsf{w} | N_1 = i] &\leq T^{\max}_\mathsf{w} \quad \forall i \in \{0,1,\ldots,b^*-1\} \text{ s.t. } p_i = 1 \\
\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}^*_\mathsf{w}, p^*_\mathsf{m})}[T_\mathsf{w} | N_1 = i] &= T^{\max}_\mathsf{w} \quad \forall i \in \{0,1,\ldots,b^*-1\} \text{ s.t. } p_i < 1 \\
\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}^*_\mathsf{w}, p^*_\mathsf{m})}[T_\mathsf{w} | N_1 = b^*] &\geq T^{\max}_\mathsf{w} \\
\arg\max\{p_\mathsf{m} \in [0,1] \colon \mathbb{E}^{\mathsf{P}}_{(\mathbf{p}^*_\mathsf{w}, p_\mathsf{m})}[T_\mathsf{m}] \leq T^{\max}_\mathsf{m}\} &= p^*_\mathsf{m},
\end{aligned}$$

where b^* is the number of entries in $\mathbf{p}_{\mathsf{w}}^*$. Meanwhile, we also have a slightly revised formula for social welfare in this setting:

$$\mathrm{SW}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})} = \frac{1}{\Lambda}\left(\lambda_{\mathsf{w}}\sum_{i=0}^{b-1} p_i\left(T^{\max}_{\mathsf{w}} - \mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i]\right)\mathbb{P}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}(N_1=i) + p_{\mathsf{m}}\lambda_{\mathsf{m}}\left(T^{\max}_{\mathsf{m}} - \mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}]\right)\right).$$

EC.4.2. General Approach for Finding Equilibria with Mixed Walk-in Strategies

We proceed to discuss how we can find equilibria under mixed walk-in strategies in the single-server model, taking $\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i]$ and $\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[T_{\mathsf{m}}]$ as given; the computation of these sojourn times is deferred to Appendix EC.4.6. There are challenges associated with determining equilibria in the two-server setting, so we avoid that case.

A key distinguishing feature of equilibria determination in this setting as compared with the setting of pure walk-in strategies (where the walk-in behavior depends on an integer value, b), is that we can no longer examine a finite number of cases $b \in \{0,1,\ldots,B\}$. In fact the space of mixed walk-in strategies, $\mathcal{S}$, is unaccountably large, spanning a union of hypercubes of different dimensionalities.

In an effort to make the equilibria determination problem tractable, we use a different approach depending on the policy under consideration. We make a couple of observations. First, under MWO, we note that mobiles have priority over all walk-ins. As a result, $\mathbb{E}^{\mathsf{MWO}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[T_{\mathsf{m}}]$ does not depend on $\mathbf{p}_{\mathsf{w}}$, so we can determine p^*_{m} first (via the final equilibrium constraint) and then, given this value of p^*_{m}, we find those vectors $\mathbf{p}^*_{\mathsf{w}} \in \mathcal{S}$ that satisfy the equilibrium constraints for walk-ins. Second, under WOM, we have the opposite situation: walk-ins have priority over mobiles and so $\mathbb{E}^{\mathsf{WOM}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i]$ does not depend on p_{m}. This allows us to determine $\mathbf{p}^*_{\mathsf{w}}$ based on the equilibrium constraints for walk-ins, and then, given this vector for $\mathbf{p}^*_{\mathsf{w}}$, we find the value of $p_{\mathsf{m}} \in [0,1]$ that satisfies the final equilibrium constraint (i.e., we find the "best response" of mobiles to the strategies adopted by the walk-ins). Such straightforward situations do not necessarily arise in the case of WMO, and so we differ discussion equilibria determination under WMO to Appendix EC.4.5.

EC.4.3. Finding p^*_{m} in the setting with mixed walk-in strategies

We now directly address the method for finding equilibria, $(\mathbf{p}^*_{\mathsf{w}}, p^*_{\mathsf{m}})$. We first discuss the method of determining p^*_{m} under MWO and WOM, noting that this is the first step we use in finding equilibria for MWO, but the second step (following the determination of $\mathbf{p}^*_{\mathsf{w}}$, as this value is required) for WOM. For $\mathsf{P} \in \{\mathsf{MWO},\mathsf{WOM}\}$, we must simply compute

$p^*_{\mathsf{m}} = \arg\max_{p_{\mathsf{m}}\in[0,1]}\{\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}^*_{\mathsf{w}},p_{\mathsf{m}})}[T_{\mathsf{m}}] \le T^{\max}_{\mathsf{m}}\}$, taking $\mathbf{p}^*_{\mathsf{w}}$ to be as already found (using the method discussed below) under WOM and taking the choice of $\mathbf{p}^*_{\mathsf{w}}$ to be inconsequential for MWO (as $\mathbb{E}^{\mathsf{MWO}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[T_{\mathsf{m}}]$ is constant in $\mathbf{p}_{\mathsf{w}}$). Under both policies, if it is neither the case that $p^*_{\mathsf{m}} = 0$ or $p^*_{\mathsf{m}} = 1$ (both of which can be readily checked), then p^*_{m} is the unique value of p_{m} satisfying $\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[T_{\mathsf{m}}] = T^{\max}_{\mathsf{m}}$, which can either be determined exactly, if possible, or approximated with arbitrary precision using a bisection search, as $\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[T_{\mathsf{m}}]$ is continuous and monotone in p_{m} (we assert continuity without proof, while monotonicity follows from a slight modification of the proof of Proposition 4, which establishes the monotonicity of $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}[T_{\mathsf{m}}]$).

EC.4.4. Finding $\mathbf{p}^*_{\mathsf{w}}$ in the setting with mixed-walk in strategies

We now address the determination of $\mathbf{p}^*_{\mathsf{w}}$, noting that this is the second step when p^*_{m} is required, as is the case under MWO (and sometimes under WMO, see Appendix EC.4.5), and the first step otherwise (i.e., under WOM). We use the notation $\mathbf{x}^\frown y$ (resp. $\mathbf{x}^\frown\mathbf{y}$) to denote the concatenation of the vector $\mathbf{x}$ and the scalar y (resp. the vector $\mathbf{y}$); e.g., if $\mathbf{x} = (1, 1/2)$, $y = 1/3$, and $\mathbf{y} = (1/3, 1/4)$,then $\mathbf{x}^\frown y = (1, 1/2, 1/3)$, while $\mathbf{x}^\frown y = (1, 1/2, 1/3, 1/4)$. This notation allows us to present the following crucial result, which plays a key role in determination of $\mathbf{p}^*_{\mathsf{w}}$:

PROPOSITION EC.1. *For any policy* $\mathbf{P}$ *in the one-server setting, any* $\mathbf{p}_{\mathsf{w}} \in \mathcal{S}$ *with at least* i *entries, and any* $\mathbf{q} \in [0,1]^k$*, we have*

$$\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i] = \mathbb{E}_{(\mathbf{p}_{\mathsf{w}}^\frown\mathbf{q},p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1 = i].$$

That is, the response time of a walk-in seeing i *customers in the system upon arrival does not depend on the strategies of those walk-ins who observe at least* $i+1$ *customers upon arrival.*

Proof of Proposition EC.1. This observation follows readily from examining the relevant Markov chains (i.e., those depicted in Fig. 4), by noting that once one leaves state $(N_1, N_2) = (i, j)$, to enter phase $i+1$, the next time one will enter phase i will always be in state $(N_1, N_2) = (i, 1)$ (and analogously for $(N_1, N_{2,\mathsf{w}})$ in the WOM case), from which it follows that the limiting distribution of N_2 and $N_{2,\mathsf{w}}$ conditioned on $N_1 = i$ is the same under both $\mathbf{p}_{\mathsf{w}}$ and $\mathbf{p}_{\mathsf{w}}^\frown\mathbf{q}$; i.e., $\pi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j)\Big/\sum_{j=0}^{\infty}\pi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j)$ and $\phi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j)\Big/\sum_{j=0}^{\infty}\pi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j)$ do not change if we replace $\mathbf{p}_{\mathsf{w}}$ with $\mathbf{p}_{\mathsf{w}}^\frown\mathbf{q}$, which is sufficient to yield the desired claim (see Proposition 5). □

Unfortunately, Proposition EC.1 not hold in the two-server model (as phase transitions are bidirectional), hence analysis is not tractable in that setting; nonetheless, an adaptation of this technique was able to give approximate equilibria that were used in generating Fig. 7 for the two-server model. With this proposition in mind, we provide the following "algorithm sketch" for determining at least one equilibrium strategy, $\mathbf{p}_{\mathsf{w}}^*$:

1. Set $i \leftarrow 0$ and $\mathbf{p}_{\mathsf{w}} \leftarrow \varnothing$, where $\varnothing$ represents the empty vector. Then, continue on to step 2.
2. If $\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}}^*)}[T_{\mathsf{w}}|N_1 = i] \geq T_{\mathsf{w}}^{\max}$, then report that $(\mathbf{p}_{\mathsf{w}}^*, p_{\mathsf{m}}^*)$ is an equilibrium where $\mathbf{p}_{\mathsf{w}}^* = \mathbf{p}_{\mathsf{w}}$ and end the algorithm. Otherwise, continue on to step 3.
3. If $\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}}\frown 1, p_{\mathsf{m}}^*)}[T_{\mathsf{w}}|N_1 = i] \leq T_{\mathsf{w}}^{\max}$, set $i \leftarrow i+1$ and $\mathbf{p}_{\mathsf{w}} \leftarrow \mathbf{p}_{\mathsf{w}}\frown 1$; then, return to step 2. Otherwise, continue on to step 4
4. Consider the following function, g, of $p \in [0,1]$: $g(p) \equiv \mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}}\frown p, p_{\mathsf{m}}^*)}[T_{\mathsf{w}}|N_1 = i] - T_{\mathsf{w}}^{\max}$. Based on the results of steps 2 and 3, the fact that we have reached this step indicates that $g(0) < 1$ and $g(0) > 1$. So, by the continuity of g (which we state without proof), we know that g has at least one root. Find such a root—or approximate one to arbitrary accuracy via a bisection search—and call it p^*. Now set $i \leftarrow i+1$ and $\mathbf{p}_{\mathsf{w}} \leftarrow \mathbf{p}_{\mathsf{w}}\frown p^*$. Then, return to step 2.

Note that this algorithm will terminate in finite time (as long as the length of bisection searches are limited) as i increments by 1 through each loop of the algorithm, and the algorithm will terminate without i exceeding B. Further, note that this algorithm will find only one equilibrium value of $\mathbf{p}_{\mathsf{w}}$. We know of no method for systematically and exhaustively finding all such equilibria (although we have observed that multiple may exist as step 4 may have more than one solution), although one can "search" for additional equilibria in an exploratory manner by developing variants of this algorithm that permute (with appropriate modifications) steps 2, 3, and 4, and introduce some degree of randomization in initializing the bisection search.

We note that the mixed equilibria discussed in our results (see Section 6) are all of the form $(1, 1, \ldots, 1, p)$. We conjecture that equilibria of this form (allowing for $p = 0$, yielding a non-mixed threshold strategy) always exist. In obtaining the results presented in Section 6, we attempt to find equilibria with mixed walk-in strategies whenever we fail to find any equilibria with a pure walk-in strategy. In all such cases, we have observed that there exists some $b \in \mathbb{Z}_{\geq 0}$ such that $b+1$ is a best-response for a walk-in when all other walk-ins are

employing threshold b and vice versa. In such cases, we set $\mathbf{p}_{\mathsf{w}} \leftarrow (1,1,\ldots,1)$ with length b and set $i \leftarrow b$ and start running through the above algorithm at step 4; in all cases, we observe that the algorithm next terminates when reaching step 2, yielding an equilibrium walk-in strategy of the form $(1,1,\ldots,1,p)$.

EC.4.5. Determining equilibria with mixed walk-in strategies under WMO

The case of determining equilibria with mixed walk-in strategies under WMO is more challenging as compared to finding such equilibria under the other two single-server policies. This is because in the case of WMO, we must determine $\mathbf{p}^*_{\mathsf{w}}$ and p^*_{m} jointly, since walk-in strategies affect the "best response" of mobiles, and vice-versa. The following proposition highlights a restricted case, where we can circumvent this problem:

PROPOSITION EC.2. *If $T^{\max}_{\mathsf{m}} \geq 1/(\mu_2 - \lambda_{\mathsf{m}}) + 1/\mu_2$, then under any equilibrium $(\mathbf{p}^*_{\mathsf{w}}, p^*_{\mathsf{m}})$, we must have $p^*_{\mathsf{m}} = 1$ under* WMO.

Proof of Proposition EC.2. We first observe that we can view the subsystem of mobiles at Stage 2 under WMO as behaving like an M/M/1 with setup. Mobiles arrive according to a Poisson process with rate $p_{\mathsf{m}}\lambda_{\mathsf{m}}$. Once the system begins serving mobiles, it will continue serving mobiles (who have exponential service requirements) without interruption, at rate μ_2. However, when a mobile arrives to this system, they may not immediately begin service. Specifically, immediate service does not begin if a **W** is already present at Stage 2, in which case they will be in service. Say that whenever a mobile arrives into the system with no other mobiles, the event where the mobile cannot go into service immediately occurs with probability q (note that successive events are not necessarily independent, and q depends on $\mathbf{p}_{\mathsf{w}}$). When such an event occurs, we can view the time to serve this walk-in as a setup time that is distributed $\text{Exp}(\mu_2)$, after which time we can serve mobiles until the completion of the mobile busy period without any interruptions. It is easy to see that mobile sojourn times are upper-bounded by the special case where we always have setups, i.e., $q = 1$. In this case the mobiles experience an M/M/1 with exponentially distributed setup times with an arrival rate $p_{\mathsf{m}}\lambda_{\mathsf{m}}$ and both service and setup rates equal to μ_2, which is known to have a mean sojourn time equal to that of the corresponding M/M/1 plus the mean setup time (see Harchol-Balter (2013) Section 27.3), and so we have the upper bound $\mathbb{E}^{\mathsf{WMO}}_{(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}})}[T_{\mathsf{m}}] \leq 1/(\mu_2 - p_{\mathsf{m}}\lambda_{\mathsf{m}}) + 1/\mu_2$, which guarantees $p_{\mathsf{m}} = 1$ is a best response to any $\mathbf{p}_{\mathsf{w}} \in \mathcal{S}$, so long as $T^{\max}_{\mathsf{m}} \geq 1/(\mu_2 - \lambda_{\mathsf{m}}) + 1/\mu_2$, which establishes the claim. □

Proposition EC.2 tells us that by restricting attention to settings where $T_{\mathsf{m}}^{\max} \geq 1/(\mu_2 - \lambda_{\mathsf{m}}) + 1/\mu_2$ (under WMO), we know that $p_{\mathsf{m}}^* = 1$, and can thus proceed to determining $\mathbf{p}_{\mathsf{w}}^*$ in accordance with the method presented in Appendix EC.4.4. This condition is satisfied by 1232 out of (86.76%) the 1420 problem instances in our pruned full-factorial experiment. Of the remaining 188 instances, we find an equilibrium with a pure walk-in strategy in an additional 131 instances, leaving 54 remaining cases.

We now sketch an iterative technique for approximating equilibria with mixed walk-in strategies when $T_{\mathsf{m}}^{\max} < 1/(\mu_2 - \lambda_{\mathsf{m}}) + 1/\mu_2$ and no pure strategy equilibria are found to exist:

1. Set $\mathbf{p}_{\mathsf{w}} \leftarrow \varnothing$ and $p_{\mathsf{m}} \leftarrow 1$ (or better initial "guesses" if available based on the failed process of attempting to determine pure strategy equilibria). Continue on to step 2.
2. Apply the method presented in Appendix EC.4.4 (without overwriting $\mathbf{p}_{\mathsf{w}} \leftarrow \varnothing$ in step 1 of that algorithm and taking p_{m}^* to be the current value of p_{m}), updating $\mathbf{p}_{\mathsf{w}}$ based on the value of $\mathbf{p}_{\mathsf{w}}^*$ returned (note that this need not be an equilibrium strategy, rather it is merely a best response to the current of value of p_{m}). Note the change (in terms of an appropriate metric, e.g., the infinity-norm after adding zeros to the tail of a shorter vector where appropriate) in $\mathbf{p}_{\mathsf{w}}$ as a result of this entire step and call it Δ_{w}, then continue on to step 3.
3. Apply the method presented in Appendix EC.4.3 to find an updated value of p_{m} that is a best response to the current $\mathbf{p}_{\mathsf{w}}$. Note the change in p_{m} as a result of this step and call it Δ_{m}, then continue on to step 4.
4. If $\max(\Delta_{\mathsf{w}}, \Delta_{\mathsf{m}})$ falls below a desired precision threshold, then terminate the algorithm here and report $(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}})$ as an approximate equilibrium. Otherwise, return to step 2.

While we cannot prove that this technique is guaranteed to converge, it yielded adequate results in the aforementioned 54 cases where other methods did not suffice. A similar technique can be used to find mixed equilibria in the two-server setting.

EC.4.6. Sojourn Time Computation Under Mixed Walk-in Strategies

We now turn to the question of how to compute the sojourn times of interest under strategy profiles of the form $(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}})$ in both the single- and two-server settings. One can show without difficulty that $\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}})}[T_{\mathsf{w}} | N_1 = i]$ and $\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}})}[T_{\mathsf{m}}]$ follow the same forms given for $\mathbb{E}^{\mathsf{P}}_{(b, p_{\mathsf{m}})}[T_{\mathsf{w}} | N_1 = i]$ and $\mathbb{E}^{\mathsf{P}}_{(b, p_{\mathsf{m}})}[T_{\mathsf{m}}]$ (respectively), as given in Propositions 5 and

6. More precisely, under all policies of interest, P, the aforementioned propositions continue to hold when all instances of the operator $\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}$—and all implicit references to the operator $\mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}$—in their statements are replaced with $\mathbb{E}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}$ and $\mathbb{P}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}$, respectively. Such "implicit references" to $\mathbb{P}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}$ appear in the limiting probabilities $\pi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j)$ and $\phi^{\mathsf{P}}_{(b,p_{\mathsf{m}})}(i,j)$, where reference to the strategy profile has been suppressed in the interest of brevity.

In order to compute $\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[T_{\mathsf{w}}|N_1=i]$ and $\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[T_{\mathsf{m}}]$ for all policies of interest (exactly in the single-server setting and approximately in the two-server setting), we must compute the following under the strategy profile $(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})$: (i) the first and second moments of U and V under WOM, (ii) the mean value of $Z(i,j)$ under WM, and (iii) the limiting probabilities $\pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j)$ (equivalently, $\pi^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}(i,j)$), $\phi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,j)$, $\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)$, and $\phi^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}(i,j)$ (and where appropriate, one or more series associated with these limiting probabilities). The determination of these quantities under the strategy profile $(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})$ requires only a minor modification of the methods given throughout Appendix EC.3 for determination of their analogues under the strategy profile (b,p_{m}). These modifications result by observing that the only consequence of generalizing from strategy profiles of the form (b,p_{m}) to those of the form $(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})$ on all quantities of interest is an alteration of the Markov chains governing (N_1,N_2) and $(N_1,N_{2,\mathsf{w}})$. This is also why the aforementioned adaptation of Propositions 5 and 6 to the setting with mixed walk-in strategies is possible.

Specifically, all four chains of interest under the strategy profile $(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})$ are identical to their counterparts under (b,p_{m}) (these are illustrated in Figs. 4 and 5) with one crucial change: the transition rate from phase (row) i to phase (row) $i+1$ should be $p_i\lambda_{\mathsf{w}}$ rather than λ_{w}, $\forall i\in\{0,1,\ldots,b\}$. As a result, we can obtain the values of interest using the following modifications of the methods presented throughout Appendix EC.3, all of which essentially require replacing each instance of λ_{w} by $p_i\lambda_{\mathsf{w}}$ for the appropriately chosen value of i:

1. The limiting probabilities $\pi^{\mathsf{MWO}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}(i,j)$ and $\pi^{\mathsf{WMO}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}(i,j)$ are equal to one another (as were their analogues, $\pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j)$ and $\pi^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}(i,j)$). These quantities, together with their associated series, can be computed exactly via the same method given in Appendix EC.3.1 for computing $\pi^{\mathsf{MWO}}_{(b,p_{\mathsf{m}})}(i,j)$ and $\pi^{\mathsf{WMO}}_{(b,p_{\mathsf{m}})}(i,j)$, by using the following

revised matrices $\mathbf{L_0}$ and $\mathbf{L}$ and values $\gamma_0, \gamma_1, \ldots, \gamma_b$ and $\xi_0, \xi_1, \ldots, \xi_b$, in place of those given in display (EC.28):

$$\mathbf{L_0} = \begin{pmatrix} -\gamma_0 & p_0\lambda_{\mathsf{w}} & & & \\ & -\gamma_1 & p_1\lambda_{\mathsf{w}} & & \\ & & \ddots & \ddots & \\ & & & -\gamma_{b-1} & p_{b-1}\lambda_{\mathsf{w}} \\ & & & & -\gamma_b \end{pmatrix}, \mathbf{L} = \begin{pmatrix} -\xi_0 & p_0\lambda_{\mathsf{w}} & & & \\ & -\xi_1 & p_1\lambda_{\mathsf{w}} & & \\ & & \ddots & \ddots & \\ & & & -\xi_{b-1} & p_{b-1}\lambda_{\mathsf{w}} \\ & & & & -\xi_b \end{pmatrix},$$

$$\gamma_i = \begin{cases} p_{\mathsf{m}}\lambda_{\mathsf{m}} + p_0\lambda_{\mathsf{w}} & i = 0 \\ \mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}} + p_i\lambda_{\mathsf{w}} & 1 \le i \le b-1 \\ \mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}} & i = b \end{cases}, \qquad \xi_i = \begin{cases} p_{\mathsf{m}}\lambda_{\mathsf{m}} + p_i\lambda_{\mathsf{w}} + \mu_2 & 0 \le i \le b-1 \\ p_{\mathsf{m}}\lambda_{\mathsf{m}} + \mu_2 & i = b \end{cases}.$$

Note that the statement "all elements of the diagonal of $\mathbf{R}$ are actually the same except for the last, $\mathbf{R}(b,b)$," no longer holds, but this holds no consequences for the method in general.

2. The limiting probabilities $\phi^{\mathsf{WOM}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}(i,j)$ can be computed exactly via the same method given in Appendix EC.3.2 for computing the limiting probabilities $\phi^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}(i,j)$, by using the following revised system of equations in place of system (EC.30):

$$\begin{cases} p_0\lambda_{\mathsf{w}}\phi_{0,0} = \mu_2\phi_{0,1} \\ (p_i\lambda_{\mathsf{w}} + \mu_1)\phi_{i,0} = p_{i-1}\lambda_{\mathsf{w}}\phi_{i-1,0} + \mu_2\phi_{i,1} & \forall i \in \{1,2,\ldots,b-1\} \\ \mu_1\phi_{b,0} = p_{b-1}\lambda_{\mathsf{w}}\phi_{b-1,0} + \mu_2\phi_{b,1} \\ (p_0\lambda_{\mathsf{w}} + \mu_2)\phi_{0,1} = \mu_1\phi_{1,0}, \\ (p_i\lambda_{\mathsf{w}} + \mu_2)\phi_{i,1} = p_{i-1}\lambda_{\mathsf{w}}\phi_{i-1,1} + \mu_1\phi_{i+1,0} & \forall i \in \{1,2,\ldots,b-1\} \\ \mu_2\phi_{b,1} = p_{b-1}\lambda_{\mathsf{w}}\phi_{b-1,1} \\ \sum_{i=0}^{b}(\phi_{i,0} + \phi_{i,1}) = 1 \end{cases}.$$

3. The transforms of U and V under the strategy profile $(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}})$—from which one can find the quantities of interest $\mathbb{E}^{\mathsf{WOM}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[U]$, $\mathbb{E}^{\mathsf{WOM}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[U^2]$, $\mathbb{E}^{\mathsf{WOM}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[V]$, and $\mathbb{E}^{\mathsf{WOM}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}[V^2]$—can be computed exactly via the same method given in Appendix EC.3.3, by making the following modifications:

(a) System (EC.31) should be revised as follows:

$$\begin{cases} p_0\lambda_{\mathsf{w}}\psi_{0,0} = \mu_2\psi_{0,1} + p_{\mathsf{m}}\lambda_{\mathsf{m}}\left(1-\psi_{0,0}\right) \\ \left(p_i\lambda_{\mathsf{w}} + \mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}}\right)\psi_{i,0} = p_{i-1}\lambda_{\mathsf{w}}\psi_{i-1,0} + \mu_2\psi_{i,1}, & \forall i \in \{1,2,\ldots,b-1\} \\ \left(\mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}}\right)\psi_{b,0} = p_{b-1}\lambda_{\mathsf{w}}\psi_{b-1,0} + \mu_2\psi_{b,1}, \\ \left(p_0\lambda_{\mathsf{w}} + \mu_2 + p_{\mathsf{m}}\lambda_{\mathsf{m}}\right)\psi_{0,1} = \mu_1\psi_{1,0}, \\ \left(p_i\lambda_{\mathsf{w}} + \mu_2 + p_{\mathsf{m}}\lambda_{\mathsf{m}}\right)\psi_{i,1} = p_{i-1}\lambda_{\mathsf{w}}\psi_{i-1,1} + \mu_1\psi_{i+1,0}, & \forall i \in \{1,2,\ldots,b-1\}, \\ \left(\mu_2 + p_{\mathsf{m}}\lambda_{\mathsf{m}}\right)\psi_{b,1} = p_{b-1}\lambda_{\mathsf{w}}\psi_{b-1,1}, \\ \sum_{i=0}^{b}\left(\psi_{i,0}+\psi_{i,1}\right) = 1. \end{cases}$$

(b) System (EC.32) should be revised as follows:

$$\begin{cases} \widetilde{U_{0,0}}(s) = 1, \\ \widetilde{U_{i,0}}(s) = \dfrac{p_i\lambda_{\mathsf{w}}+\mu_1}{s+p_i\lambda_{\mathsf{w}}+\mu_1}\left(\dfrac{p_i\lambda_{\mathsf{w}}}{p_i\lambda_{\mathsf{w}}+\mu_1}\widetilde{U_{i+1,0}}(s) + \dfrac{\mu_1}{p_i\lambda_{\mathsf{w}}+\mu_1}\widetilde{U_{i-1,1}}(s)\right), & \forall i \in \{1,2,\ldots,b-1\}, \\ \widetilde{U_{b,0}}(s) = \dfrac{\mu_1}{s+\mu_1}\widetilde{U_{b-1,1}}(s), \\ \widetilde{U_{i,1}}(s) = \dfrac{p_i\lambda_{\mathsf{w}}+\mu_2}{s+p_i\lambda_{\mathsf{w}}+\mu_2}\left(\dfrac{p_i\lambda_{\mathsf{w}}}{p_i\lambda_{\mathsf{w}}+\mu_2}\widetilde{U_{i+1,1}}(s) + \dfrac{\mu_2}{p_i\lambda_{\mathsf{w}}+\mu_2}\widetilde{U_{i,0}}\right), & \forall i \in \{0,1,\ldots,b-1\}, \\ \widetilde{U_{b,1}}(s) = \dfrac{\mu_2}{s+\mu_2}\widetilde{U_{b,0}}(s). \end{cases} \tag{EC.41}$$

(c) Eq. (EC.34) should be revised as follows:

$$\begin{aligned} \widetilde{V}(s) &= \frac{p_0\lambda_{\mathsf{w}}+\mu_2}{s+p_0\lambda_{\mathsf{w}}+\mu_2}\left(\frac{p_0\lambda_{\mathsf{w}}}{p_0\lambda_{\mathsf{w}}+\mu_2}\widetilde{U_{1,0}}(s)\widetilde{V}(s) + \frac{\mu_2}{p_0\lambda_{\mathsf{w}}+\mu_2}\right) \\ &\Longrightarrow \widetilde{V}(s) = \frac{\mu_2}{s+p_0\lambda_{\mathsf{w}}\left(1-\widetilde{U_{1,0}}(s)\right)+\mu_2}. \end{aligned} \tag{EC.42}$$

If one opts to use the more efficient method discussed at the end of Appendix EC.3.3, system (EC.35) and display (EC.36), should be revised to be consistent with system (EC.41) and display (EC.42), respectively.

4. The limiting probabilities $\pi^{\mathsf{TS}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}(i,j)$ and $\phi^{\mathsf{WM}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}(i,j)$ and their associated series can be approximated via the methods given in Appendices EC.3.4 and EC.3.7 for computing the limiting probabilities $\pi^{\mathsf{TS}}_{(b,p_{\mathsf{m}})}(i,j)$ and $\phi^{\mathsf{WM}}_{(b,p_{\mathsf{m}})}(i,j)$, respectively, by using the

following revised matrix $\mathbf{L}$ and values $\nu_0, \nu_1, \ldots, \nu_b$, in place of those given in display (EC.37):

$$\mathbf{L} = \begin{pmatrix} -\nu_0 & p_0\lambda_{\mathsf{w}} & & & \\ & -\nu_1 & p_1\lambda_{\mathsf{w}} & & \\ & & \ddots & \ddots & \\ & & & -\nu_{b-1} & p_{b-1}\lambda_{\mathsf{w}} \\ & & & & -\nu_b \end{pmatrix}, \quad \nu_i = \begin{cases} p_{\mathsf{m}}\lambda_{\mathsf{m}} + p_0\lambda_{\mathsf{w}} + \mu_2 & i = 0 \\ \mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}} + p_i\lambda_{\mathsf{w}} + \mu_2 & 1 \le i \le b-1 \\ \mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}} + \mu_2 & i = b \end{cases}.$$

5. The quantity $\mathbb{E}^{\mathsf{WM}}_{(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}})}[Z(i,j)]$ can be approximated via the methods given in Appendix EC.3.8 by using the following revised system of equations in place of system (EC.39):

$$\begin{cases} \tau_0 = \dfrac{1 + p_0\lambda_{\mathsf{w}}\tau_1}{p_0\lambda_{\mathsf{w}} + \mu_2} \\ \tau_\ell = \dfrac{1 + p_\ell\lambda_{\mathsf{w}}\tau_{\ell+1}}{p_\ell\lambda_{\mathsf{w}} + \mu_1 + \mu_2} + \dfrac{\mu_1}{p_\ell\lambda_{\mathsf{w}} + \mu_1 + \mu_2}\left(\tau_{\ell-1} + \displaystyle\sum_{k=0}^{b}\left(p_{(\ell-1)\to k}\right)\tau_k\right) & 1 \le \ell \le b-1 \\ \tau_b = \dfrac{1}{\mu_1 + \mu_2} + \dfrac{\mu_1}{\mu_1 + \mu_2}\left(\tau_{b-1} + \displaystyle\sum_{k=0}^{b}\left(p_{(b-1)\to k}\right)\tau_k\right). \end{cases}$$

EC.4.7. Heterogeneous Patience Levels in the Single-Server Setting

We turn our attention to the case where patience levels are *heterogeneous* and consider the case where for each walk-in (resp. mobile), $T_{\mathsf{w}}^{\max}$ (resp., $T_{\mathsf{m}}^{\max}$) is a random variable that is independently drawn from a bounded continuous distribution with c.d.f. F_{w} (resp., F_{m}). For the discussion that follows, it will be helpful to recall that a bounded distribution with c.d.f. F has lower and upper bounds that can be expressed by $F^{-1}(0)$ and $F^{-1}(1)$, respectively.

The theory developed in Appendix EC.4.1 makes it quite simple to address this extension, which also explains why we again necessarily restrict attention to the single-server setting. This is because we again have action profiles of the form $(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}})$ with $\mathbf{p}_{\mathsf{w}} \in \mathcal{S}$, although there is additional meaning carried in this action profile that was absent in the case of constant (homogeneous) patience levels: specifically, p_i (where p_i is determined by $\mathbf{p}_{\mathsf{w}} = (p_0, p_1, \ldots, p_{b-1})$) denotes that a walk-in with patience level $T_{\mathsf{w}}^{\max}$ will join when $N_1 = i$ if and only if $T_{\mathsf{w}}^{\max}$ is at or above the $(1 - p_i)$ quantile (of the patience level distribution for

all walk-ins), i.e., $T_{\mathsf{w}}^{\max} \geq F_{\mathsf{w}}^{-1}(1-p_i)$. Similarly, p_{m} denotes that a mobiles with patience level $T_{\mathsf{m}}^{\max}$ will join if and only if $T_{\mathsf{m}}^{\max}$ is at or above the $(1-p_{\mathsf{m}})$ quantile (of the patience level distribution for all mobiles), i.e., $T_{\mathsf{w}}^{\max} \geq F_{\mathsf{m}}^{-1}(1-p_{\mathsf{m}})$. Of course, as a consequence of these interpretations, the original meanings of p_i and p_{m} still hold true as well: an arbitrary walk-in joins at $N_1 = i$ with probability p_i, while an arbitrary mobile joins with probability p_{m}.

In light of the above, in this setting $(\mathbf{p}_{\mathsf{w}}^*, p_{\mathsf{m}}^*)$ is an equilibrium if it satisfies the following revised equilibrium conditions:

$$\begin{aligned} \mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}}^*, p_{\mathsf{m}}^*)}[T_{\mathsf{w}} | N_1 = i] &= F_{\mathsf{w}}^{-1}(1-p_i) \quad \forall i \in \{0,1,\ldots,b^*-1\} \\ \mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}}^*, p_{\mathsf{m}}^*)}[T_{\mathsf{w}} | N_1 = b^*] &\geq F_{\mathsf{w}}^{-1}(1) \\ \arg\max\{p_{\mathsf{m}} \in [0,1] \colon \mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}}^*, p_{\mathsf{m}})}[T_{\mathsf{m}}] \leq F_{\mathsf{m}}^{-1}(1-p_{\mathsf{m}})\} &= p_{\mathsf{m}}^*, \end{aligned} \tag{EC.43}$$

where b^* is again the number of entries in $\mathbf{p}_{\mathsf{w}}^*$. Meanwhile, social welfare takes on the following form under heterogeneous patience levels:

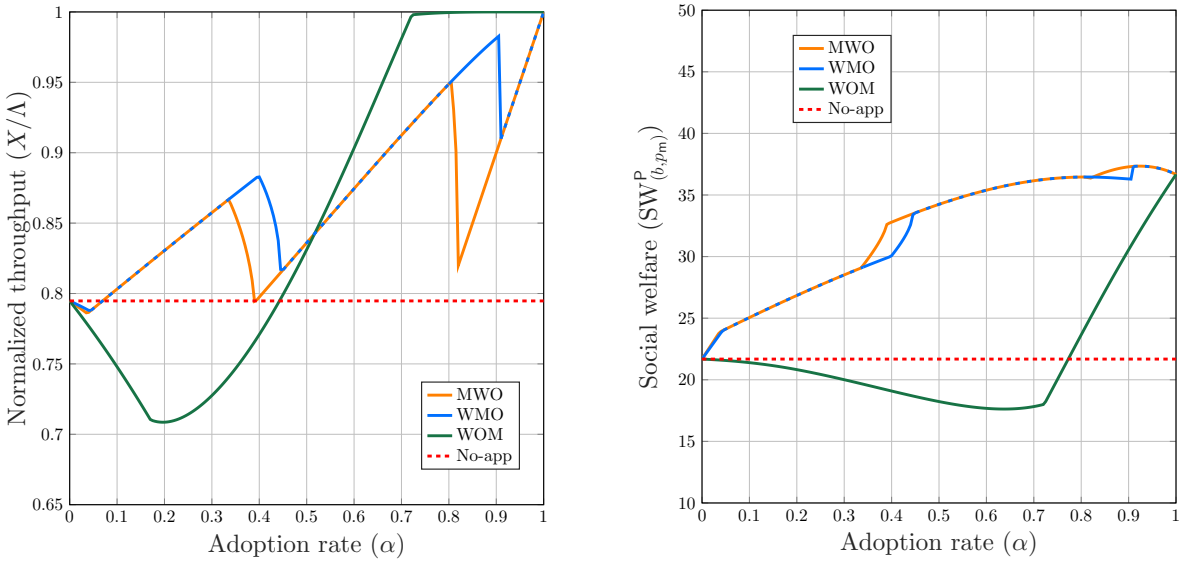
$$\mathrm{SW}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}})} = \frac{1}{\Lambda}\left(\lambda_{\mathsf{w}} \sum_{i=0}^{b-1} p_i \left(\overline{T_{\mathsf{w}}^{\max}}(i) - \mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}})}[T_{\mathsf{w}} | N_1 = i] \right) \mathbb{P}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}})}(N_1 = i) + p_{\mathsf{m}} \lambda_{\mathsf{m}} \left(\overline{T_{\mathsf{m}}^{\max}} - \mathbb{E}^{\mathsf{P}}_{(b, p_{\mathsf{m}})}[T_{\mathsf{m}}] \right) \right),$$

where $\overline{T_{\mathsf{w}}^{\max}}(i)$ is the average patience level of walk-ins who join when $N_1 = i$ and $\overline{T_{\mathsf{m}}^{\max}}$ is the average patience level of mobiles who join. Specifically, these quantities are given by

$$\overline{T_{\mathsf{w}}^{\max}}(i) = \frac{1}{p_i} \int_{F_{\mathsf{w}}^{-1}(1-p_i)}^{F_{\mathsf{w}}^{-1}(1)} t \, dF_{\mathsf{w}}(t) \qquad \text{and} \qquad \overline{T_{\mathsf{m}}^{\max}} = \frac{1}{p_{\mathsf{m}}} \int_{F_{\mathsf{m}}^{-1}(1-p_{\mathsf{m}})}^{F_{\mathsf{m}}^{-1}(1)} t \, dF_{\mathsf{m}}(t).$$

The method described in Appendix EC.4.2 can then be modified in order to find one or more equilibria in this setting. This modification is straightforward as one needs only change equations being solved and the inequalities being checked on the basis of the new equilibrium conditions given in display (EC.43). Note that the assumption $T_{\mathsf{m}}^{\max} \geq 1/(\mu_2 - \lambda_{\mathsf{m}}) + 1/\mu_2$ that facilities tractable analysis becomes $F_{\mathsf{m}}^{-1}(0) \geq 1/(\mu_2 - \lambda_{\mathsf{m}}) + 1/\mu_2$ in this setting.

Applying this method to the case where patience thresholds follow a truncated normal distribution, we find results that are in line with those presented in the body of the paper in the setting where customers have homogeneous patience level (see Fig. EC.2); i.e., we observe a region of adoption rates, α, where all three policies underperform the no-app benchmark with respect to throughput.



(a) Normalized throughput (X/Λ)

(b) Social welfare ($\mathrm{SW}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}$)

Figure EC.2 **Single-server with heterogeneous customers:** $\Lambda = 0.05, \mu_1 = 0.16, \mu_2 = 0.08$, $T_{\mathsf{w}}^{\max} \sim \mathbf{TrN}(62.5, 10, 60, 65)$, $T_{\mathsf{m}}^{\max} \sim \mathbf{TrN}(70, 8, 60, 80)$; $\mathbf{TrN}(\mu, \sigma, \mathrm{LB}, \mathrm{UB})$ **is a truncated Normal distribution with mean** μ**, std. dev.** σ**, and lower and upper bounds** LB **and** UB**.**

EC.5. Experiments

In Table EC.1, we specify the non-degenerate parameter combinations using the "×" symbol. Each cell in Table EC.1 includes 10 experiments (values of α). Our discussions in Section 6 of the paper, which are based on the tables provided in this section, are all based on the non-degenerate parameter combinations.

Table EC.1 Non-degenerate parameter combinations (specified by "×")

		$\mu_2 = 1.5$					$\mu_2 = 2$					$\mu_2 = 2.5$					$\mu_2 = 3$				
		μ_1/μ_2					μ_1/μ_2					μ_1/μ_2					μ_1/μ_2				
$T_{\mathsf{m}}^{\max}$	$\frac{T_{\mathsf{w}}^{\max}}{T_{\mathsf{m}}^{\max}}$	.25	.5	1	2	4	.25	.5	1	2	4	.25	.5	1	2	4	.25	.5	1	2	4
0.5	80%																				
	100%																				×
	125%														×	×				×	×
1	80%									×	×				×	×			×	×	×
	100%					×				×	×			×	×	×			×	×	×
	125%				×	×			×	×	×		×	×	×	×		×	×	×	×
2	80%			×	×	×		×	×	×	×		×	×	×	×		×	×	×	×
	100%			×	×	×		×	×	×	×		×	×	×	×	×	×	×	×	×
	125%		×	×	×	×		×	×	×	×	×	×	×	×	×	×	×	×	×	×
4	80%		×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	100%	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	125%	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×

Table EC.2 presents the descriptive statistics for the percentage of throughput loss due to introduction of a mobile-ordering application in experiments in which opting not to offer an app outperforms the three omni-channel prioritization policies in the single-server setting (MWO, WMO, and WOM) with respect to throughput. The percentage of throughput loss is calculated as:

$$\%\text{ throughput loss} = \frac{\text{No-app throughput} - \text{Maximum throughput of omni-channel policies}}{\text{No-app throughput}} \times 100.$$

Table EC.2 Summary statistics for % throughput loss

Average	Std. dev.	Min	1st quartile	Median	3rd quartile	Max
12.41%	10.95%	0.00%	2.60%	10.62%	19.85%	40.27%

Table EC.3 provides the number and percentage of cases in which each policy is optimal with respect to throughput (with ties broken in favor of highest social welfare, whenever possible), at all fixed levels of the five parameters μ_2, μ_1/μ_2, α, $T_{\mathsf{m}}^{\max}$, and $T_{\mathsf{w}}^{\max}/T_{\mathsf{m}}^{\max}$.

Table EC.3 Effect of parameters on the optimal policy

	μ_2				μ_1/μ_2				
	1.5	**2**	**2.5**	**3**	**0.25**	**0.5**	**1**	**2**	**4**
No app	40	24	17	15	1	6	14	24	51
	14.8%	7.1%	4.4%	3.6%	0.7%	2.5%	4.7%	6.7%	13.4%
MWO	2	2	0	0	0	1	0	0	3
	0.7%	0.6%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	0.79%
WMO	26	19	9	9	8	13	12	17	13
	9.6%	5.6%	2.3%	2.1%	5.7%	5.4%	4.0%	4.7%	3.4%
WOM	59	178	285	345	60	113	186	240	268
	21.9%	52.4%	73.1%	82.1%	42.9%	47.1%	62.0%	66.7%	70.5%
MWO&WMO	143	117	79	51	71	107	88	79	45
(Tie)	53.0%	34.4%	20.3%	12.1%	50.7%	44.6%	29.3%	21.9%	11.8%
# of instances	270	340	390	420	140	240	300	360	380

	α									
	0.05	**0.15**	**0.25**	**0.35**	**0.45**	**0.55**	**0.65**	**0.75**	**0.85**	**0.95**
No app	10	13	15	14	12	9	8	7	5	3
	7.0%	9.2%	10.6%	9.9%	8.5%	6.3%	5.6%	4.9%	3.5%	2.1%
MWO	0	2	0	0	1	0	1	0	0	0
	0.0%	1.4%	0.0%	0.0%	0.7%	0.0%	0.7%	0.0%	0.0%	0.0%
WMO	2	7	6	11	12	8	11	6	0	0
	1.4%	4.9%	4.2%	7.8%	8.5%	5.6%	7.8%	4.2%	0.0%	0.0%
WOM	57	62	64	69	80	88	98	108	120	121
	40.1%	43.7%	45.1%	48.6%	56.3%	62.0%	69.0%	76.1%	84.5%	85.2%
MWO&WMO	73	58	57	48	37	37	24	21	17	18
(Tie)	51.4%	40.9%	40.1%	33.8%	26.1%	26.1%	16.9%	14.8%	12.0%	12.7%
# of instances	142	142	142	142	142	142	142	142	142	142

	$T_{\mathsf{m}}^{\max}$				$T_{\mathsf{w}}^{\max}/T_{\mathsf{m}}^{\max}$		
	0.5	**1**	**2**	**4**	**0.8**	**1**	**1.25**
No app	24	41	18	13	22	22	52
	48.0%	14.1%	3.7%	2.2%	5.4%	4.8%	9.5%
MWO	0	1	1	2	0	0	4
	0.0%	0.3%	0.2%	0.3%	0.0%	0.0%	0.7%
WMO	5	17	18	23	10	22	31
	10.0%	5.9%	3.7%	3.9%	2.4%	4.8%	5.6%
WOM	2	168	314	383	312	295	260
	4.0%	57.9%	64.1%	64.9%	76.1%	64.1%	47.3%
MWO&WMO	19	63	139	169	66	121	203
(Tie)	38.0%	21.7%	28.4%	28.6%	16.1%	26.3%	36.9%
# of instances	50	290	490	590	410	460	550

EC.6. Notation Table

Table EC.4: Notation

α	$\triangleq$	Adoption rate; $\alpha \equiv \lambda_{\mathsf{m}}/\Lambda$
a^{P}	$\triangleq$	Allocation (class-specific mean sojourn time pair) under service policy P; $a \equiv (\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}], \mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}])$
a^{P^*}	$\triangleq$	An arbitrary Pareto optimal allocation; more precisely, the allocation under (an arbitrary Pareto optimal policy) P^*
b	$\triangleq$	Buffer size at Stage 1; queue length of Stage 1 at which all walk-ins balk
b^*	$\triangleq$	Equilibrium threshold for walk-ins
B	$\triangleq$	Strict upper bound on the buffer size at Stage 1
$\mathbf{B}$	$\triangleq$	Repeated backward transition matrix used in matrix-analytic methods
$\mathbf{bd}(\mathcal{O})$	$\triangleq$	Boundary of the achievable region
C	$\triangleq$	Cost per unit time spent waiting in the system
χ_{w}, χ_{m}	$\triangleq$	Throughput rate for walk-in(χ_{w}) and mobile (χ_{m}) customers
Δ_{w}, Δ_{m}	$\triangleq$	Change in $\mathbf{p}_{\mathsf{w}}$ (Δ_{w}) and p_{m} (Δ_{m}) in one step of the algorithm for determining equilibria with mixed walk-in strategies under WMO
e_i	$\triangleq$	i-th unit vector
E	$\triangleq$	Total net change of throughput due to information uncertainty
E_w, E_m	$\triangleq$	Net change of throughput due to individual walk-in (E_w) and mobile (E_m) information uncertainty
$\mathbb{E}^{\mathsf{P}}$	$\triangleq$	Expectation operator under policy P
$\mathbb{E}^{\mathsf{P}}_{(b,p_{\mathsf{m}})}$	$\triangleq$	Expectation operator under strategy profile (b, p_{m}) and policy P
$\mathbb{E}^{\mathsf{P}}_{(\mathbf{p}_{\mathsf{w}},p_{\mathsf{m}})}$	$\triangleq$	Expectation operator under strategy profile $(\mathbf{p}_{\mathsf{w}}, p_{\mathsf{m}})$ and policy P
$\mathbb{E}^{\mathsf{P}}[T]$	$\triangleq$	Overall mean response time; $\mathbb{E}^{\mathsf{P}}[T] \equiv (\lambda_{\mathsf{w}}\mathbb{E}^{\mathsf{P}}[T_{\mathsf{w}}] + \lambda_{\mathsf{m}}\mathbb{E}^{\mathsf{P}}[T_{\mathsf{m}}])/\Lambda$
$\mathbb{E}^{\mathsf{P}}[W]$	$\triangleq$	Mean value of overall work in the system
$\mathbb{E}^{\mathsf{P}}[W_2]$	$\triangleq$	Mean value of overall work in Stage 2
$\mathbb{E}^{\mathsf{P}}[W_{\mathsf{w}}]$, $\mathbb{E}^{\mathsf{P}}[W_{\mathsf{m}}]$	$\triangleq$	Mean values of the work due to walk-ins ($\mathbb{E}^{\mathsf{P}}[W_{\mathsf{w}}]$) and work due to mobiles ($\mathbb{E}^{\mathsf{P}}[W_{\mathsf{m}}]$) in the system
$\mathbf{F}_0$, $\mathbf{F}$	$\triangleq$	Initial ($\mathbf{F}_0$) and repeated ($\mathbf{F}$) forward transition matrices used in matrix-analytic methods
$f_b(\cdot)$	$\triangleq$	Mobiles' mean sojourn time as a function of p_{m} with index b; $f_b(\cdot) \equiv \mathbb{E}^{\mathsf{P}}_{(b,\cdot)}[T_{\mathsf{m}}]$
F_{w}, F_{m}	$\triangleq$	CDF of patience levels for walk-ins (F_{w}) and mobiles (F_{m})
$\mathbf{G}$	$\triangleq$	G-matrix used in matrix-analytic methods; $\mathbf{G}(\ell, k) \equiv p_{\ell \to k}$
γ_i	$\triangleq$	an auxiliary value defined by $p_{\mathsf{m}}\lambda_{\mathsf{m}} + \lambda_{\mathsf{w}}$, if $i = 0$; $\mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}} + \lambda_{\mathsf{w}}$, if $1 \le i \le b-1$; $\mu_1 + p_{\mathsf{m}}\lambda_{\mathsf{m}}$, if $i = b$
$\mathbf{I}$	$\triangleq$	Identity matrix
$\mathcal{I}(i)$	$\triangleq$	Time interval corresponding to a tagged walk-in's sojourn at Stage 1 until they arrive to Stage 2
$K(i)$	$\triangleq$	Random quantity of mobile customers that arrived during $\mathcal{I}(i)$
$\mathbf{L}_0$, $\mathbf{L}$	$\triangleq$	Initial ($\mathbf{L}_0$) and repeated ($\mathbf{L}$) local transition matrices used in matrix-analytic methods
Λ	$\triangleq$	Total arrival rate for all customers; $\Lambda \equiv \lambda_{\mathsf{w}} + \lambda_{\mathsf{m}}$
λ_{w}, λ_{m}	$\triangleq$	Arrival rates of walk-in (λ_{w}) and mobile (λ_{m}) customers
$L(i,j)$	$\triangleq$	Number of customers in Stage 2 (including the tagged walk-in) at time of the tagged walk-in's arrival to Stage 2, given that $N_1 = i$ and $N_2 = j$

$L(i,j,k)$	$\triangleq$	Number of customers present in Stage 2 at the end of $\mathcal{I}(i)$, given that $K(i)=k$ and initially $N_2=j$
$\mathbf{M}$	$\triangleq$	Mobile task at Stage 2
$M_\rho(t)$	$\triangleq$	Number of customers in an M/M/1 system under load $\rho\in(0,\infty)$ at time t
μ_1	$\triangleq$	Service rate at Stage 1
μ_2	$\triangleq$	Service rate at Stage 2
N_1, N_2	$\triangleq$	Number of customers in Stage 1 (N_1) and Stage 2 (N_2)
$N_{2,\mathsf{w}}$	$\triangleq$	Number of walk-ins at Stage 2 (i.e., number of $\mathbf{W}$ tasks)
N_w, N_m	$\triangleq$	Number of walk-ins (N_w) and mobiles (N_m)
ν_i	$\triangleq$	an auxiliary value defined by $p_\mathsf{m}\lambda_\mathsf{m}+\lambda_\mathsf{w}+\mu_2$, if $i=0$; $\mu_1+p_\mathsf{m}\lambda_\mathsf{m}+\lambda_\mathsf{w}+\mu_2$, if $1\le i\le b-1$; $\mu_1+p_\mathsf{m}\lambda_\mathsf{m}+\mu_2$, if $i=b$
$\mathcal{O}$	$\triangleq$	Achievable region of allocations; $\mathcal{O}\equiv\{t^\mathsf{P}\in\mathbb{R}_+^2\colon \mathsf{P}\in\mathcal{P}\}$
$\mathcal{O}^*$	$\triangleq$	Pareto frontier; $\mathcal{O}^*\equiv(\mathcal{O}\setminus\mathcal{V}_i)\cap\mathbf{bd}(\mathcal{O})$, for $i\in\{1,2\}$
$\mathbf{O}$	$\triangleq$	Walk-in task at Stage 1
P	$\triangleq$	Arbitrary service policy
P^*	$\triangleq$	Arbitrary Pareto optimal policy
$\langle\mathsf{P}_1,\mathsf{P}_2\rangle(\theta)$	$\triangleq$	Random mixture of policies P_1 (w.p. θ) and P_2 (w.p. $1-\theta$)
$\mathcal{P}$	$\triangleq$	Policy space: the set of all possible policies
$\mathcal{P}^*$	$\triangleq$	Pareto space; $\mathcal{P}^*\equiv\{\mathsf{P}^*\mid \nexists\mathsf{P}\in\mathcal{P}\colon t^\mathsf{P}\succ t^{\mathsf{P}^*}\}$
$\mathbb{P}^\mathsf{P}_{(b,p_\mathsf{m})}$	$\triangleq$	Probability operator under strategy profile (b,p_m) and policy P
$\phi^\mathsf{P}_{(b,p_\mathsf{m})}(i,j)$	$\triangleq$	Limiting probability associated with state (i,j) in the $(N_1,N_{2,\mathsf{w}})$ CTMC under strategy profile (b,p_m) and policy P; $\phi^\mathsf{P}_{(b,p_\mathsf{m})}(i,j)\equiv\mathbb{P}^\mathsf{P}_{(b,p_\mathsf{m})}(N_1=i,N_{2,\mathsf{w}}=j)$
$\pi^\mathsf{P}_{(b,p_\mathsf{m})}(i,j)$	$\triangleq$	Limiting probability associated with state (i,j) in the (N_1,N_2) CTMC under strategy profile (b,p_m) and policy P; $\pi^\mathsf{P}_{(b,p_\mathsf{m})}(i,j)\equiv\mathbb{P}^\mathsf{P}_{(b,p_\mathsf{m})}(N_1=i,N_2=j)$
$\pi^{\mathsf{TS}}_{(b,p_\mathsf{m})}(i,j)$	$\triangleq$	Two-server limiting probability $\pi^\mathsf{P}_{(b,p_\mathsf{m})}(i,j)$ under $\mathsf{P}\in\{\mathsf{WM},\mathsf{FCFS},\mathsf{MW}\}$
$\vec{\pi}_j$	$\triangleq$	Vector of limiting probabilities when for $N_1=i\in\{0,\dots,b\}$ when $N_2=j$
$\pi_{(b,p_\mathsf{m})}(i,j)$	$\triangleq$	Limiting probabilities under both MWO and WMO
$p_{\ell\to k}$	$\triangleq$	Probability that we specifically end up in state $(k,m-1)$ before reaching state $(k',m-1)$ for any $k'\neq k$ from state (l,m) under WM
p_m	$\triangleq$	Mobiles' joining probability
p^*_m	$\triangleq$	Mobiles' joining probability under equilibrium
$\psi^{\mathsf{WOM}}_{(b,p_\mathsf{m})}(i,j)$	$\triangleq$	Steady-state probability that a mobile arriving to a mobile-less system under WOM sees $(N_1=i,N_{2,\mathsf{w}}=j)$; $\psi^{\mathsf{WOM}}_{(b,p_\mathsf{m})}(i,j)=\mathbb{P}^{\mathsf{WOM}}_{(b,p_\mathsf{m})}(N_1=i,N_2=N_{2,\mathsf{w}}=j)$
$P(u,v,w;\rho)$	$\triangleq$	$\mathbb{P}\left(M_\rho(t_v)=w\mid M_\rho(0)=u\right)$, probability that the system occupancy of an M/M/1 system under load $\rho>0$ transitions from u to w after exactly v further arrivals
$\mathbf{p}_\mathsf{w}$	$\triangleq$	Walk-ins' mixed joining strategy which is a vector of length b
$\mathbf{p}^*_\mathsf{w}$	$\triangleq$	Walk-ins' mixed joining strategy under equilibrium
R	$\triangleq$	Benefit obtained by walk-ins from receiving service
$\mathbf{R}$	$\triangleq$	Rate matrix (R-matrix) used in matrix-analytic methods
ρ	$\triangleq$	Load in an M/M/1 system
ρ_w, ρ_m	$\triangleq$	fractions of the time spent serving walk-ins (ρ_w) and mobiles (ρ_m)
$\mathcal{S}$	$\triangleq$	Generalized walk-in strategy space; $\mathcal{S}\equiv\bigcup_{b=0}^\infty\prod_{i=0}^{b-1}(0,1]$
$\mathrm{SW}^\mathsf{P}_{(b,p_\mathsf{m})}$	$\triangleq$	Social welfare under policy P under strategy profile (b,p_m)
$\mathrm{SW}^\mathsf{P}_{(\mathbf{p}_\mathsf{w},p_\mathsf{m})}$	$\triangleq$	Revised social welfare formula when walk-ins applying mixed strategies
τ_ℓ	$\triangleq$	Expected "hitting time" associated with the trip from state (ℓ,m) where

		$m \geq 1$, until the first time we reach state $(k, m-1)$ for any $k \in \{0, 1, \ldots, b\}$
t_n	$\triangleq$	Time of the n-th Poisson arrival to an M/M/1 system since time 0
T_{w}, T_{m}	$\triangleq$	Sojourn time of a walk-in (T_{w}) or mobile (T_{m}) customer
$T_{\mathsf{w}}^{\max}$, $T_{\mathsf{m}}^{\max}$	$\triangleq$	Patience level for walk-in ($T_{\mathsf{w}}^{\max}$) and mobile ($T_{\mathsf{m}}^{\max}$) customers
$\overline{T_{\mathsf{w}}^{\max}}(i)$	$\triangleq$	Average patience level of walk-in customers who join when $N_1 = i$
$\overline{T_{\mathsf{m}}^{\max}}$	$\triangleq$	Average patience level of mobiles who join
U	$\triangleq$	Waiting time of a mobile who arrives when there are no other mobiles
$\widetilde{U}(s)$	$\triangleq$	The Laplace transform of the random variable U; $\widetilde{U}(s) \equiv \mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[e^{-sU}]$
$U_{i,j}$	$\triangleq$	The time it takes for a system currently in a state $(N_1, N_{2,\mathsf{w}}) = (i, j)$ to be empty of all its walk-ins, without regard for any mobile arrivals; $U_{i,j} \sim (U \,\vert N_1 = i, N_{2,\mathsf{w}} = j\,)$
$\widetilde{U_{i,j}}(s)$	$\triangleq$	Laplace transform of $U_{i,j}$; $\widetilde{U_{i,j}}(s) \equiv \mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[e^{-sU_{i,j}}]$
V	$\triangleq$	Sojourn time of a mobile who enters an empty system; $V \sim (T_{\mathsf{m}} \vert N_1 = N_2 = 0)$
$\widetilde{V}(s)$	$\triangleq$	Laplace transform of V; $\widetilde{V}(s) \equiv \mathbb{E}^{\mathsf{WOM}}_{(b,p_{\mathsf{m}})}[e^{-sV}]$
$\mathbf{W}$	$\triangleq$	Walk-in task at Stage 2
W	$\triangleq$	The work in the system
ξ_i	$\triangleq$	An auxiliary value defined by $p_{\mathsf{m}}\lambda_{\mathsf{m}} + \lambda_{\mathsf{w}} + \mu_2$, if $0 \leq i \leq b-1$; $p_{\mathsf{m}}\lambda_{\mathsf{m}} + \mu_2$, if $i = b$
X	$\triangleq$	Total throughput
$Y(i,j)$	$\triangleq$	Expected workload that a walk-in will encounter at Stage 2 once it arrives there given that $N_1 = i$ and $N_2 = j$ when it arrived to Stage 1
$Y_K(i,j)$	$\triangleq$	Truncation of the first summation (by summing from $k = 0$ to K instead of $k = 0$ to ∞) in the expression of $Y(i,j)$
$Z(i,j)$	$\triangleq$	Time it takes to reach a state where $N_{2,\mathsf{w}} = 0$ from state $(N_1, N_{2,\mathsf{w}}) = (i, j)$ under WM, given (b, p_{m}); $Z(i,j) \sim \inf\{s \geq 0 \colon N_{2,\mathsf{w}}(t+s) = 0 \vert N_1(t) = i, N_{2,\mathsf{w}}(t) = j\}$, $\forall t \geq 0$
$\succ$	$\triangleq$	Dominance relation on allocations
$\frown$	$\triangleq$	Vector concatenation operator