Submitted to *Manufacturing & Service Operations Management*
manuscript (Please, provide the manuscript number!)

Waiting Experience in Open-Shop Service Networks: Improvements via Flow Analytics & Automation

(Authors' names blinded for peer review)

Problem definition: We study open-shop service networks where customers go through multiple services. We were motivated by a partnering health screening clinic, where customers are routed by a dispatcher and operational performance is measured at two levels: micro-level, via waits for *individual* services, and macro-level, via *overall* wait. Both measures reflect customer experience and could support its management. Our analysis revealed that waits were long and increased along the service process. Such long waits give rise to negative waiting experience and the increasing shape is detrimental as it is known to create *perceived* waits that are even longer. Our goal is hence to analyze strategies that shape and improve customers' perceived experience. ***Methodology/results:*** Analytically, we use a stylized two-station open-shop network to show that prioritizing advanced customers, jointly with pooling (virtual) queues, can improve both macro- and micro-level performance. We validate these findings with a simulation model, calibrated with our clinic's data. Practically, we find that an automated routing system (ARS), recently implemented in the clinic, had a negligible impact on overall wait — it simply redistributed waiting among wait-for-routing and wait-for-service. Still ARS renders applicable sophisticated priority and routing policies (that were infeasible under the manual routing practice), specifically the ones arising from the present research. ***Managerial implications:*** Our study amplifies performance benefits of accounting for individual customers' system-status in addition to station-level load information. We offer insights into the implementation of new technologies: firms better plan for fundamental changes in their operation, rather than harness new technology to their existing operation, which may be sub-optimal due to past technical limitations.

Key words: service analytics, information technology, wait time management, open shop, priority policy

1. Introduction

Open-shop service networks, in which customers visit multiple service-stations in no specific order, prevail in the modern service industry. Examples include health screening clinics and large primary care physician practices, where customers complete different medical examinations in no particular service order; amusement parks, where customers visit various attractions in some arbitrary order; and retail stores, where customers visit several departments before paying at the cashiers. An operational characteristic of such service networks is that a customer's experience is affected by both *macro-level* measures, such as the total wait over all visited stations, and *micro-level* measures, such as the probability of excessive delays at individual stations or waits dynamic during the customer visit. This

paper focuses on improving customers' waiting experience in open-shop service networks, as captured by both their macro- and micro-level performance measures.

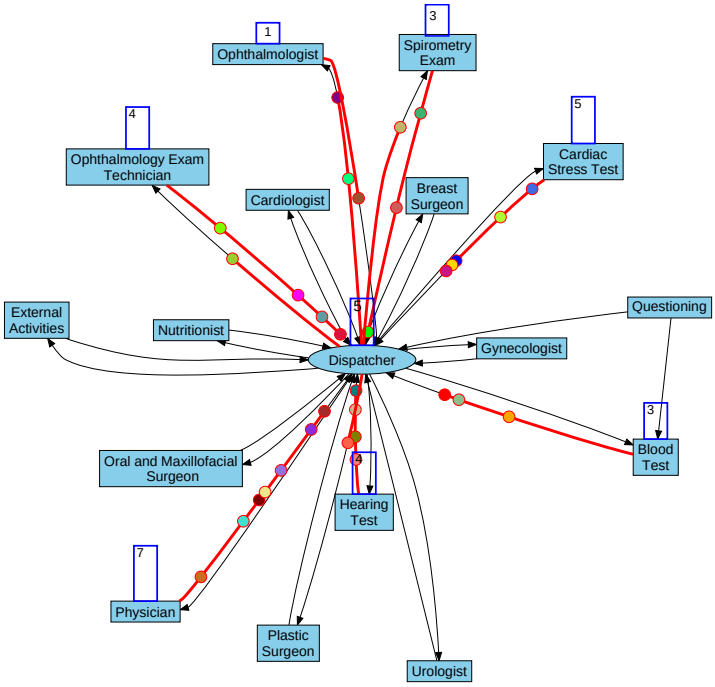
Our partner clinic: This research grew out of a collaboration with a health screening clinic, in a leading healthcare institution in Israel. The global health screening market is estimated to reach $338 billion by 2028, from $255 billion in 2021; see QYResearch Group 2022. The clinic offers health screening services, composing over ten medical tests. It operates as an open-shop service network, namely customers can take most tests in any order, with a few precedence constraints (e.g., cardiac stress test preceded by doctor exam). Following each test, a dispatcher routes customers to their next test, as depicted in Figure 1, which illustrates the dispatcher-centric network topology. Subsequently, customers wait for service in test-dedicated queues, and are served on First-Come-First-Served (FCFS) basis.

The Israeli public healthcare system provides most screening tests for free, while the clinic charges a substantial price for its services. Hence, customer satisfaction is of utmost importance to the clinic. Due to limited resources and the inherent variability in process components, waiting is inevitable. This gives rise to the challenge of controlling, over customer visits, both macro- and micro-level measures.

The clinic's practice is in concert with existing research. From the macro perspective, operational performance is quantified by the overall wait for the entire clinic visit. This is a standard measure of operational performance in service systems, which stems from the strong negative correlation between overall wait and customers' perception of service quality (see, e.g., Osuna 1985, Taylor 1994)—long waits lower perceived service quality. The micro perspective is applicable to open-shop service networks where customers experience several waits along their service paths—customers with poor experience at individual stations may perceive overall service quality to be low, even under the same macro-level overall wait (see, e.g., Baron et al. 2014, 2017).

Automated Routing System (ARS): In 2018, the clinic implemented an ARS to speed up routing decisions and reduce the dispatcher's workload. To assess ARS's effects, we empirically analyze data from the clinic in 2016 and 2019—pre- and post-ARS implementation. We discover that, by accelerating routing decisions, ARS has a negligible impact on customers' overall wait. Instead, ARS redistributes customers' wait times—it shortens

Figure 1 Snapshot of the Clinic's Open-Shop Service Network, under the Dispatcher's Manual Routing



Note. Light blue rectangles represent stations. The ellipse at the center represents the dispatcher. The rectangle above a station depicts the number of in-service plus waiting customers at that station, when the snapshot was taken. Dots on arrows from stations to the dispatcher represent customers who will wait for the dispatcher once their service is completed. Dots on arrows from the dispatcher to stations represent customers who will be routed to these stations. Section 4.1 gives a more detailed description. An animation created from our customer-flow data is clickable here.

wait-for-routing and lengthens wait-for-service times (see Observation 1 in §4.3.1). We further discover that ARS did not fundamentally change the clinic's operations: it merely automated the clinic's sub-optimal priority and routing policies.

Specifically, the clinic's priority policy assigns customers to tests according to FCFS, considering their station-level wait but failing to account for their system-status, such as service history or future service requests (see Observation 2 in §4.3.2). Consequently, customers who have completed most tests may have to wait after customers who just arrived at the clinic, which leads to long overall waits. Moreover, the speedy routing decisions of ARS forego information of tests becoming available in the near future, which lead to less informed routing decisions (see Observation 3 in §4.3.3). In other words, customers may be routed to tests, while more suitable tests become available soon after the decision, thus resulting in inefficient routing. Yet despite its drawbacks, ARS is not without benefits—it enables the implementation of more advanced priority and routing policies towards improv-

ing customers' waiting experience in the clinic; such policies, as we now describe, were less applicable, if not infeasible, under the dispatcher's manual routing practice.

ARS-enabled controls: To be concrete, we first propose the *advanced customer priority* (ACP) policy that prioritizes customers closer to completing their clinic visits by considering their system-status information. This ACP policy shall outperform the clinic's current station-level FCFS policy. Second, we propose a *buffer strategy* that postpones routing decisions. Customers are assigned to stations only when those stations' queue lengths fall below a certain threshold, while other available customers wait in a pooled queue. In extreme cases, a zero-buffer strategy acts as a fully pooled system, delaying routing until the last minute. Currently, the clinic routes customers as soon as they complete their previous tests, effectively operates as an infinite-buffer strategy. In reality, a buffer strategy with a finite buffer size may be more effective: it makes more informed routing decisions while leaving customers some time to react to routing announcements and travel to their next tests. We expect these two policies to improve performance over the clinic's practice.

To analytically establish the benefits of the ACP policy and buffer strategy, we first study a stylized two-station open-shop service network. We find that the ACP policy consistently reduces macro-level average overall wait and micro-level average wait at the last station compared to FCFS, with a reduction in the probability of station-level excessive delays in some cases. The buffer strategy further reduces the macro-level wait. Additionally, we validate these policies using a data-driven simulation model of the clinic while maintaining the integrity of the clinic's data and only modifying the priority and routing policies. It is shown that replacing the clinic's current station-level FCFS policy with ACP policy and setting a buffer for stations' queues improves *both* macro- and micro-level performance. For example, an ACP policy prioritizing the shortest-expected-remaining-processing-time (SERPT), combined with a buffer size of 5, reduces the average overall wait by 15.88% and lowers the probability of station-level delays exceeding 20 minutes by 46.98%.

Contribution: We introduce and analyze the combined use of the ACP policy and buffer strategy in open-shop service networks, aiming to improve customer experiences via operational steps. These improvements are validated through both analytical and simulation approaches, using a stylized two-station open-shop service network and real clinic data. The insights gained extend beyond health screening services. Additionally, our paper highlights

the importance of evolving service operations in tandem with new technology, rather than retrofitting technology as a point solution into existing operations, without re-optimizing in view of the capabilities of the new technology.

Outline: In Section 2, we review related research. In Section 3, we analyze a stylized two-station open-shop service network to understand the operational advantages of priority policies and buffer strategy. In Section 4, we introduce the clinic's operations and analyze, using the clinic's data, how ARS affected clinic performance. In Section 5, we test our proposed priority policies and buffer strategy via a data-driven simulation of the clinic. We conclude in Section 6 with summary remarks and suggestions for future research.

2. Literature Review

Our work is related to the following research streams: customer experience in multi-stage services, priority and routing policies in networks, and buffer strategies.

2.1. Customer Experience in Multi-Stage Services

Our health examination encompasses over ten stations, each with its own wait, resulting in a multi-stage waiting experience. We evaluate customer experience using both macro- and micro-level performance measures. The most commonly used metric is the macro-level overall wait, a widely recognized factor affecting customer experience negatively (see, e.g., Osuna 1985, Taylor 1994).

The importance of micro-level performance in service processes is underlined by studies on customer experience (see, e.g., Lee et al. 2012, Das Gupta et al. 2016, Bray 2020, Gallino et al. 2023). Researchers have demonstrated that customers evaluate their experience using gestalt characteristics rather than equally considering all attributes of a multi-stage service (see, e.g., Ariely and Carmon 2000, Bitran et al. 2008). Specifically, the peak-end effect claims that customer perceptions are heavily influenced by the most extreme (peak) and final leg (end) of their service experience (see, e.g., Kahneman 2000, Chase and Dasu 2001, Verhoef et al. 2004). This effect was shown to influence customer perceptions in multiple contexts, such as package delivery services (Bray 2020), online retail website design (Gallino et al. 2023), medical treatment (Ariely and Carmon 2000), and hospitality (Geng et al. 2013). Moreover, the trend or dynamic of the experience over the course of the service delivery process also significantly impacts customer perceptions. Customers prefer patterns that improve over time (see, e.g., Hansen and Danaher 1999, Ariely and Zauberman 2003),

and services that are of finer partitions provide a better sense of progress (see, e.g., Ariely and Zauberman 2000, 2003, Dixon and Verma 2013).

Recent analytical work considered service design in view of the above characteristics, aiming to optimize customers' experience from both macro and micro perspectives. Baron et al. (2014, 2017) explore the role of strategic idleness in smoothing workflow and balancing wait along the service process. We further explore the impact of priority policies and buffer strategies on macro- and micro-aspects, both analytically and empirically. Specifically, the priority policies reduce the increasing trends and restrict the peak-end effect. The buffer strategy postpones routing decisions, creating a pooling effect and enhancing the system's operational flexibility, thereby improving performance at both macro- and micro-levels.

2.2. Scheduling with Priorities

A *priority policy* specifies which customer will be served next when a server becomes idle, similar to scheduling policies in manufacturing. The literature on scheduling is vast and thus we only discuss work of direct relevance to us. Pinedo and Ross (1982) analyze a two-station open-shop, proving that length-of-stay is minimized by prioritizing the tasks that have not yet started their processing in either station. Adding due dates and allowing preemption, Pinedo (1984) shows shortest-expected-remaining-processing-time first (SERPT) minimizes the number of tardy jobs. Most open-shop scheduling problems consider macro-level performance only and they are NP-hard, even for simple two-station networks (see reviews by Anand and Panneerselvam 2015, Pinedo 2016). Another policy widely used in observable service systems is FCFS; its prevalence is due to it being perceived as fair to customers (see, e.g., Larson 1987). The multiple routes in open-shop service networks render fairness challenging to articulate and hence enforce (see, e.g., Raz et al. 2006).

Healthcare settings are fertile sources for papers on priorities and scheduling. Of most relevance are those that seek to improve patient flow in the emergency department (ED) networks (see, e.g., Huang et al. 2015); or to manage appointments with walk-ins, no-shows, and returns (see, e.g., Wang et al. 2019, Kong et al. 2020). Characteristics of ED operations are random arrivals at time-varying rates and a high frequency of feedback routes. In contrast, health screening networks serve a pre-defined finite population of patients who visit each station once and mostly in an arbitrary order.

Inspired by the above, in Section 5.1, we consider two advanced customer priority (ACP) policies using the following criteria: longest-system-time first (LST), prioritizing customers

who arrive at the clinic first and hence spend more time both in service and waiting; and SERPT, prioritizing customers with the shortest expected remaining service time in all remaining stations. We show that these policies effectively improve both macro- and micro-level service levels. In open-shop service networks, station-level FCFS policy ensures fairness at the micro-level, as customers are served in their arrival order at each station, but not at the macro-level, where customers are served based on their arrival order to the system. On the other hand, system-level priority policies like LST and SERPT account for system-level information, offering a different dimension of fairness at the macro-level.

2.3. Open-Shop Routing – Centralized and Customer-Driven

A *routing policy* specifies the station that a customer joins, both upon arrival and after service completion. In manufacturing, routing arises when machines travel among jobs located at different sites; Averbakh et al. (2006) show that the open-shop routing problem is NP-hard for makespan minimization, even for a 2-node network. In service systems, customers travel among stations. Examples include the traveling repairman problem (Afrati et al. 1986), museum visitors who go through all exhibits while avoiding congestion (Chou and Lin 2007), and army-recruitment screening processes (Shtrichman et al. 2001). Manufacturing and services differ in cost structures: manufacturing jobs typically do not incur waiting costs unless they are perishable or tardy relative to some due date; in contrast, delays of customers, while traversing their service routes, directly affect service evaluations.

Healthcare open-shops were first considered by Baron et al. (2014, 2017). Our research grows out of these works. Through customer interviews, they discover that the longest wait carries a heavy weight in forming a waiting experience; hence, they propose and analyze strategic idling as a means to reduce long(est) delays that, in turn, will improve macro- and micro-level performance. Baron et al. (2017), however, does not distinguish whether long(est) waits occur at the outset, in the middle, or towards the end of the service process. We fill this gap by exploring the benefits of priority policies and a buffer strategy, as far as macro- and micro-level performance and dynamics of station-level waits are concerned.

Further research studies strategic customers' self-interested routing behavior, i.e., customers choose the sequence of services by themselves. Parlaktürk and Kumar (2004) consider a two-station multi-class queueing network where customers make routing choices, and the dispatcher selects the scheduling policy. Arlotto et al. (2018) reveal strategic customer herding behavior to minimize system time. Our theoretical model, described in

Section 3, is closely related to their work. The main differences include the following: in our setting, customers follow the dispatcher's routing decisions; stations can follow priority protocols for advanced customers, rather than FCFS discipline; and attention is given to both macro- and micro-level performance.

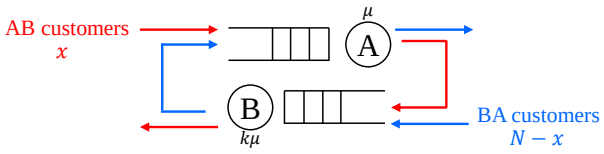
2.4. Buffer Strategies

In Sections 3.2 and 5.2, we propose a buffer strategy, which employs pooled queues by imposing limits on each station's queue length and postponing routing decisions. Pooled queues reduce waits (see, e.g., Kleinrock 1976, Mandelbaum and Reiman 1998). Nevertheless, pooling benefits could be hurt by server behaviors, as dedicated queues might give physicians greater ownership over patients (Song et al. 2015), or by generating social loafing (Wang and Zhou 2018). However, such mechanisms do not apply in our open-shop system, as servers are station-specific with fixed workloads; therefore, servers cannot offload their workload to other servers by adjusting their service rates.

3. Two-Station Open-Shop Network

To generate insights into the impacts of the advanced customer priority (ACP) policy and the buffer strategy on waits in open-shop service networks, we analyze a stylized two-station open-shop service network, as in Arlotto et al. (2018). As depicted in Figure 2, the system has two single-server stations, A and B. Customers need to visit both stations in an arbitrary order. An AB (resp. BA) customer denotes a customer who is routed first to station A (resp. B) and then to station B (resp. A). Our base model has deterministic service durations: $1/\mu$ and $1/(k\mu)$, in A and B, respectively. Without loss of generality, we assume that station A is the bottleneck that takes a longer service time, i.e., $k > 1$. Furthermore, we consider a scenario with N customers, who are all present in the system when the operations start and remain in the system until completing both services. This setup mirrors the clinic's operation, as customers are invited to arrive during the first hour of operation and are prepared to dedicate a substantial portion of their day to the service. The same was assumed in Arlotto et al. (2018).

Figure 2 **Two-Station Open-Shop Service Network**



The routing decision is represented by the number of AB customers, x, with the remaining $N-x$ as BA customers. These x AB customers and $N-x$ BA customers form queues at stations A and B, respectively, and queue positions are drawn uniformly at random.

We set the status quo at FCFS. Under this policy, station A (resp. B) serves the x AB customers (resp. $N-x$ BA customers) first and then the $N-x$ BA customers (resp. x AB customers). Station A never idles from its first service, and if station B becomes idle, its queue will never build up afterward (see Properties 1 and 2 in Arlotto et al. 2018). Therefore, at station A, all customers need to wait except the first AB customer in line. At station B, if it serves quickly enough or the number of BA customers is small, some AB customers from the end of station B's sequence may not need to wait at all.

As discussed in Section 2.1, to ultimately improve customers' perceived waiting experience, we adopt a comprehensive approach that encompasses both macro and micro aspects of customer waits. At the macro-level, we assess the average overall wait, a critical factor affecting customers' experience. We also consider the average wait at customers' last visited station and the probability of excessive waits along the service path. By addressing these various dimensions, we aim to improve customers' perceived experience.

In Section 3.1, we evaluate the ACP policy against FCFS. Next, in Section 3.2, we investigate the impact of the buffer strategy. Then, in Section 3.3, we examine the effectiveness of the ACP policy and buffer strategy through a stochastic simulation model.

3.1. Advanced Customer Priority Policy

In this subsection, we examine the Advanced Customer Priority (ACP) policy. ACP policy prioritizes customers closer to completing their visits by considering their system-status information, such as service history and pending services. In the context of the two-station open-shop service network, ACP policy prioritizes BA (resp. AB) customers over AB (resp. BA) customers at station A (resp. B). Since it is uncommon to interrupt ongoing services in a healthcare setting, we limit attention to the non-preemptive ACP policy. While here we establish the ACP policy in the context of two-station network, we extend its application to more complex open-shop networks in Section 5.1. The following proposition compares customers' macro-level average overall waits under ACP and FCFS policies, $t_P(x)$, for $P = A, F$.

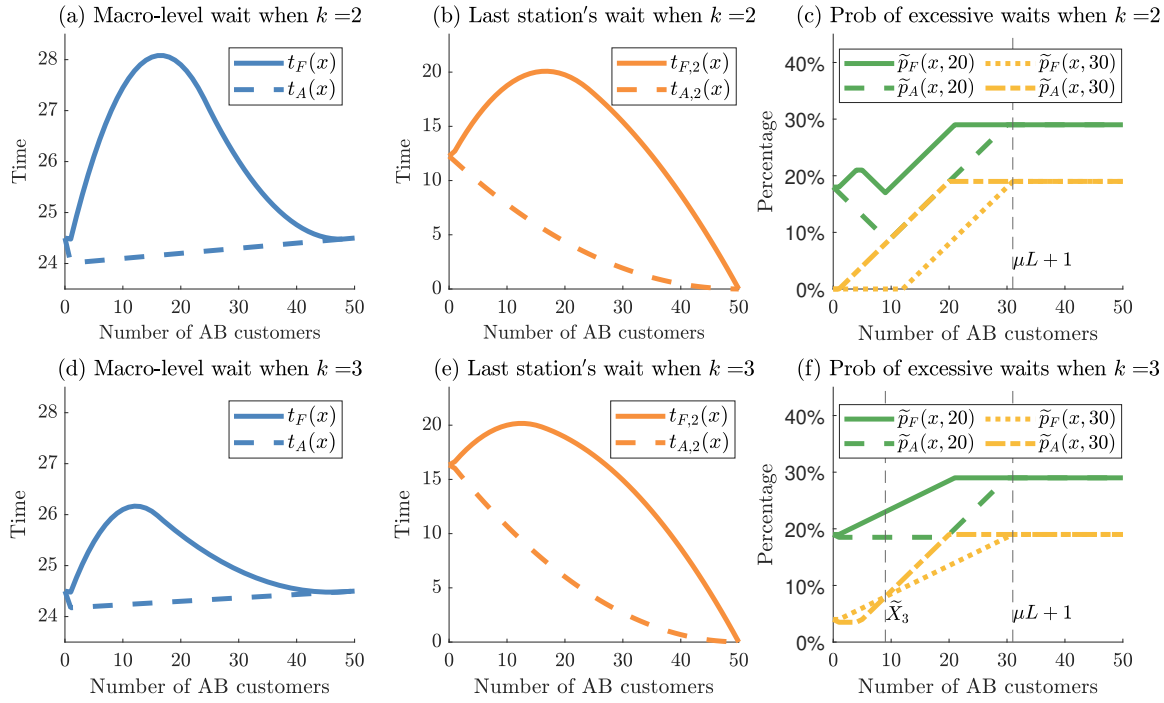
PROPOSITION 1. **[Macro-Level Measure: ACP vs. FCFS]**

(*i*) *Compared to FCFS, ACP policy weakly reduces the macro-level average wait; i.e.,* $t_A(x) \le t_F(x) \forall x$.

(*ii*) *If* $N \ge \frac{k^2+5k+2}{2}$, $t_F(x)$ *is minimized at* $x=1$ *and all integers* $x \in [N-2k+1, N-k]$. $t_A(x)$ *is minimized at* $x=1$. *Moreover, their absolute difference* $\min t_F - \min t_A = \frac{N-k-1}{Nk\mu}$ *increases with the population size* N*; and their relative difference* $\frac{\min t_F - \min t_A}{\min t_F} = \frac{2(N-k-1)}{k(N+1)(N+2)}$ *decreases with the population size* N.

Proposition 1(i) shows that, by incorporating customers' system-status information, ACP outperforms FCFS from the macro-level perspective. Figures 3(a) and (d) depict $t_F(x)$ and $t_A(x)$ as functions of the number of AB customers x, for $\mu=1$, $N=50$, and $k=2,3$. We see that ACP consistently leads to a shorter macro-level average wait compared to FCFS; i.e., $t_A(x) \le t_F(x)$ for all x. To explain this result, we note that station A serves two types of customers: BA customers who will complete their visits after station A, and AB customers who need to wait for service at station B after being served by station A. To reduce customers' overall wait, station A can prioritize BA customers so they can complete the visit and leave the system, rather than letting them wait behind those AB customers who still need to wait at station B after finishing station A. Similarly, prioritizing AB customers over BA customers at station B reduces customers' average overall wait. Hence, ACP reduces customers' average overall wait compared with FCFS.

Figure 3 Comparison of Macro- and Micro-Level Measures Under FCFS and ACP Policies ($\mu=1, N=50$)



Proposition 1(ii) suggests that both ACP and FCFS policies minimizes the macro-level average wait when $x=1$. This optimization keeps bottleneck station A busy and lets non-bottleneck station B serve customers who would otherwise wait unproductively at station A. In addition, ACP policy outperforms FCFS by further reducing customers' average overall wait by $\frac{N-k-1}{Nk\mu}$. This result extends the findings of Arlotto et al. (2018) by considering the impact of priority policies. Moreover, under FCFS, we find that other optimal routing decisions, i.e., all integers $x \in [N-2k+1, N-k]$, minimize the average overall wait. Arlotto et al. (2018) show that $x=1$ is the minimum point of $[0,N] \setminus \left(\frac{N-1}{k}, N-k\right]$ and bound the difference between $t_F(1)$ and the minimum value in $\left(\frac{N-1}{k}, N-k\right]$. We extend their result by identifying all global minimum points with identical minimum values, $t_F(1)$. Under these alternative optimal x decisions, the last $x-1$ AB customers will find station B idle when joining and get served immediately. Proposition 1(ii) further finds that the absolute improvement, $\min t_F - \min t_A$, increases with the population size N, while the relative improvement, $\frac{\min t_F - \min t_A}{\min t_F}$, decreases with the population size N. These findings suggest that the improvement brought in by ACP policy diminishes because the congestion caused by the increase in population offsets the benefits of ACP policy.

We next compare ACP and FCFS policies in two micro-level measures—the average wait at customers' second (last) visited stations, $t_{P,2}(x)$, and the probability of excessive waits in the service path, $p_P(x,L) = \sum_{s=A,B} n_{s,L}/2N$, where $n_{s,L}$ is the number of waits at station s exceeding certain threshold L, for $P = A, F$. For a given routing decision x and threshold value L, a lower value of $p_P(x,L)$ indicates customers are less likely to wait for long at any station. Improving these micro-level measures aligns with the goal of improving customers' perceived waiting experience. The following proposition discusses approximated probabilities, $\widetilde{p}_P(x,L)$, instead of the exact probabilities, $p_P(x,L)$. The approximations are used to simplify integrality and exhibit qualitative behavior consistent with the exact probabilities (see Online Appendix B.2).

Proposition 2. **[Micro-Level Measures: ACP vs. FCFS]** *Suppose the population size is large (specifically $N > k^2+2k-3$). Compared to FCFS,*

(*i*) *ACP weakly reduces the average wait at the second (last) station; i.e., $t_{A,2}(x) \le t_{F,2}(x)$.*

(*ii*) *ACP weakly reduces the approximated probability of excessive waits; i.e., $\widetilde{p}_A(x,L) \le \widetilde{p}_F(x,L)$,*

- *under any routing decision when* $k=2$ *and* $L \le \frac{N-1}{2\mu}$, *or when* $k \ge 3$ *and* $\frac{k^2-2}{k\mu} \le L \le \frac{N-k}{k\mu}$ *or* $\frac{(k-1)N-k}{k\mu(k-1)} \le L \le \frac{N-1}{2\mu}$;
- *under the following routing decisions* x*:*
  - ◇ *when* $k=2$, $x \ge \mu L+1$ *for* $L > \frac{N-1}{2\mu}$;
  - ◇ *when* $k \ge 3$, *(i)* $x \le \widetilde{X}_1$ *and* $x \ge \widetilde{X}_2$ *for* $L < \frac{k-1}{\mu}$, *(ii)* $\widetilde{X}_0 \le x \le \widetilde{X}_1$ *and* $x \ge \widetilde{X}_2$ *for* $\frac{k-1}{\mu} < L < \frac{k^2-2}{k\mu}$, *(iii)* $x \ge \widetilde{X}_2$ *for* $\frac{N-k}{k\mu} < L < \frac{(k-1)N-k}{k\mu(k-1)}$, *(iv)* $x \le \widetilde{X}_3$ *and* $x \ge \mu L+1$ *for* $\frac{N-1}{2\mu} < L < \frac{(k-1)(N-1)}{k\mu}$, *(v)* $x \ge \mu L+1$ *for* $L \ge \frac{(k-1)(N-1)}{k\mu}$, *where* $X_1 \equiv N-k\mu L-1$, $\widetilde{X}_0 \in \left(1, \frac{X_1}{k}\right]$, $\widetilde{X}_1 \in \left[\frac{X_1}{k}, X_1\right)$, $\widetilde{X}_2 \in (X_1, N-\mu L]$, *and* $\widetilde{X}_3 \in (1, N-\mu L]$.

Proposition 2(i) demonstrates that the ACP policy consistently reduces customers' wait at their second (last) station; i.e., $t_{A,2}(x) \le t_{F,2}(x) \forall x$. As illustrated in Figures 3(b) and (e), the improvement is substantial in most cases, averaging 72.67% and 66.23% reductions, respectively. This is natural because customers gain priority at their second station.

Proposition 2(ii) suggests that the ACP policy helps reduce the probability of excessive waits under specific circumstances. This is illustrated in Figures 3(c) and (f) for two threshold values $L=20$ and $L=30$. For example, given a threshold $L=20 < {}^{N-1}/_{2\mu}$ (see the solid and dashed lines in Figures 3(c) and (f)), as implied in the first case of Proposition 2(ii), the probability of excessive waits under ACP policy is consistently lower or equal to that under the FCFS policy, i.e., $\widetilde{p}_A(x,L) \le \widetilde{p}_F(x,L)$. For $L=30 \ge {}^{N-1}/_{2\mu}$ (see the dotted and dash-dotted lines in Figures 3(c) and (f)), as implied in the second case of Proposition 2(ii), we have $\widetilde{p}_A(x,L) = \widetilde{p}_F(x,L)$ under certain conditions, such as when $k=2$ and $x \ge \mu L+1$, and when $k=3$ and $x \le \widetilde{X}_3$ or $x \ge \mu L+1$.

In summary, Propositions 1 and 2 demonstrate the superiority of ACP over FCFS policy in improving customers' macro- and micro-level waits and perceived experience. From the macro perspective, prioritizing advanced customers reduces the average overall wait, thus improving customers' perceptions of waiting (Osuna 1985, Taylor 1994). From the micro perspective, customers gain priority at their second (last) station, substantially reducing wait times at this station and resulting in a strong positive end effect (Chase and Dasu 2001). Moreover, the probability of excessive waits diminishes under specific circumstances, mitigating the negative peak effect (Verhoef et al. 2004).

3.2. Buffer Strategy

Next, we examine the buffer strategy that delays routing decisions until the number of customers at a station is below a certain *threshold* buffer size. This approach allows the

dispatcher collect more information on suitable stations for more informed routing decisions. The buffer strategy separates waiting customers into two groups: (i) those awaiting tests at various stations and (ii) those in a central pool awaiting routing. ACP policy is applied to both groups. For group (i), the server uses the ACP policy to select the next customer from the station's queue when it becomes available. For group (ii), when the number of customers at a station falls below the buffer size, the dispatcher routes the customer from the central pool using the ACP policy. Setting the buffer size to infinity results in immediate routing after completing previous tests, essentially resembling the ACP policy discussed in Section 3.1.

The optimal buffer size depends on the routing time, denoted as r, which encompasses the dispatcher's examination of system load and customer information, as well as the time taken to make a routing decision. When routing time is negligible, customers can proceed to their next station immediately upon completing their previous tasks. Intuitively, setting the buffer size to zero would be ideal, ensuring that customers are routed only when stations explicitly require their presence. This approach maximizes system flexibility. However, when the dispatcher needs time to make routing decisions, a zero-buffer strategy could introduce unnecessary idleness to the system. During the routing time, stations and customers would have to wait for each other unproductively. The following proposition characterizes the relationship between routing time and the optimal buffer size.

PROPOSITION 3. *Under ACP policy,*

(*i*) *When the routing time is* $r = 0$*, for any routing decision* x*, setting a* $y = 0$ *buffer (weakly) reduces the average overall wait. The optimal buffer that minimizes the average overall wait is* $y^* = 0$*.*

(*ii*) *When the routing time is* $0 < r < \frac{1}{k\mu}$*, the optimal buffer that minimizes the average overall wait is* $y^* = 1$ *for sufficiently large customer pool, i.e.,* $N > \overline{N}$*.*

Proposition 3(i) demonstrates the efficiency of a zero-buffer strategy in deterministic open-shop networks with zero routing time. This strategy optimizes system flexibility by routing customers only when stations become available, thus minimizing the average overall wait. Table 1 in Section 3.3 further illustrates the effectiveness of such a buffer strategy in a two-station open-shop service network with stochastic arrivals and service times. However, when routing involves time, the zero-buffer strategy loses its edge. Proposition 3(ii) gives

a sufficient condition under which a buffer size of $y = 1$ becomes optimal. This buffer size outperforms zero buffer because it allows the dispatcher time to route customers, reducing the likelihood of stations idling while awaiting customer arrivals. Meanwhile, larger buffer sizes ($y > 1$) prove less effective. As per Proposition 1, the optimal policy aims to keep the bottleneck station busy and make other customers complete the non-bottleneck station first. Setting a buffer size of $y > 1$ implies that some AB customers are held in queue A and delayed by the later arriving BA customers who have completed service B and get prioritized, which negatively affects the system's efficiency.

Proposition 3 considers a routing time between zero and the non-bottleneck station's service time. Cases where the routing time exceeds this service time (i.e., $r > \frac{1}{k\mu}$) introduce the possibility of the routing process itself becoming the bottleneck. While the optimal buffer size may increase with the routing time (to prevent situations where customers and stations wait unproductively for each other due to unfinished routing decisions), in practice, routing is typically faster than station services. Therefore, scenarios where routing times exceed service times might not be realistic, especially when employing advanced technology like the ARS for automated routing.

Moreover, similar to the routing time, the transfer time that customers need to physically move between stations can also affect the optimality of a zero-buffer strategy. When customers require time for transitions, a zero-buffer strategy can lead to unnecessary idleness as stations await transitioning customers. In real-world scenarios, the dispatcher has to strike a balance, delaying routing decisions to improve decision-making while preserving station capacity. In Section 3.3, we identify optimal buffer sizes in stochastic systems, where the optimal buffer size may exceed 1. In Section 5.2, we examine the buffer strategy in a more complex multi-station open-shop service network using the clinic's data.

3.3. Simulations of Stochastic Systems

In Sections 3.1 and 3.2, we have shown that in a two-station open-shop network with deterministic service times and all customers present when the operation starts, ACP policy combined with buffer strategy outperforms the standard FCFS policy in both macro- and micro-level performance measures. In this section, we examine the effectiveness of ACP policy and buffer strategy via simulation of the two-station open-shop network with stochastic inter-arrival and service times.

Following Arlotto et al. (2018), we consider N customers arrive stochastically. The arrival time of customer $i \in \{1, ..., N\}$ is uniformly distributed on the interval $[i\gamma - \phi, i\gamma + \phi]$ so that customer i's expected arrival time is $i\gamma$ and her realized arrival time uniformly falls into an interval of length 2ϕ. Customers' arrival times are mutually independent. Service times are exponentially distributed with service rates $\mu = 1$ for station A and $k\mu$ for station B. We simulate all combinations of $\gamma \in \{0.001, 0.1, 0.25, 0.5, 0.75, 1\}$, $\phi \in \{0, 0.25, 0.5, 0.75, 1\}$, and $k \in \{4/3, 3/2, 2\}$, for a total of 90 experiments. For each combination, we consider $N = 50$ customers and run 1000 independent trials. We measure the macro-level overall wait, t_P, the micro-level wait at the second (last) station, $t_{P,2}$, and the probability of station-level waits exceeding the threshold value L, $p_P(t_s > L)$, for $P = F, A$. Note that, with stochastic arrivals, the routing decision becomes more complicated than simply the number of AB customers x in the deterministic system in Sections 3.1 and 3.2. Here, we use routing policy that assigns available customers to the station with the shortest expected wait time, which is the product of queue length and the average service time of one customer.

Table 1 **Impact of ACP Policy and Buffer Strategy on Macro- and Micro-Level Measures ($r = 0$, $k = 3/2$, and $L = 20$)**

(i) FCFS Policy

ϕ	0			0.25			0.5			0.75			1		
γ	t_F	$t_{F,2}$	$p_F(t_s > L)$	t_F	$t_{F,2}$	$p_F(t_s > L)$	t_F	$t_{F,2}$	$p_F(t_s > L)$	t_F	$t_{F,2}$	$p_F(t_s > L)$	t_F	$t_{F,2}$	$p_F(t_s > L)$
0.001	29.84	20.37	27.15%	29.64	20.22	27.29%	29.56	20.11	26.93%	29.94	20.04	27.29%	29.03	19.56	26.85%
0.01	26.08	17.10	19.96%	25.86	16.81	19.50%	26.45	17.33	20.34%	26.44	17.41	19.97%	25.29	16.48	17.85%
0.25	21.17	13.49	11.90%	20.55	13.03	9.02%	21.09	13.40	10.27%	20.44	12.99	9.20%	20.09	12.73	9.15%
0.5	13.22	9.10	4.01%	12.56	8.51	2.34%	13.26	8.97	3.09%	12.69	8.85	3.20%	12.76	8.83	3.20%
0.75	7.38	6.28	2.10%	7.31	6.05	1.06%	7.18	5.91	1.12%	6.92	5.65	0.86%	7.25	5.86	1.14%
1	3.63	3.28	0.10%	3.22	2.86	0.00%	3.52	3.13	0.00%	3.50	3.10	0.00%	3.01	2.54	0.00%

(ii) ACP Policy

ϕ	0			0.25			0.5			0.75			1		
γ	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$
0.001	24.55	4.18	19.81%	24.08	4.14	20.03%	24.55	4.37	20.41%	24.36	4.26	20.29%	23.68	4.15	18.75%
0.01	22.07	4.68	17.04%	21.70	4.30	16.78%	22.45	4.68	16.97%	22.16	4.67	16.54%	21.18	4.23	16.21%
0.25	18.93	6.06	11.63%	18.21	5.60	11.36%	18.72	5.83	11.85%	18.21	5.70	11.92%	17.77	5.48	11.28%
0.5	12.33	6.32	6.10%	11.86	5.97	5.36%	12.46	6.15	5.95%	12.16	6.72	5.22%	12.20	6.61	5.82%
0.75	7.15	5.20	2.78%	7.26	5.18	1.91%	7.14	5.05	1.88%	6.89	4.72	1.96%	7.24	4.89	2.10%
1	3.60	2.92	0.38%	3.18	2.44	0.50%	3.50	2.72	0.52%	3.48	2.71	0.41%	2.96	2.11	0.32%

(iii) ACP Policy with Zero-Buffer Strategy

ϕ	0			0.25			0.5			0.75			1		
γ	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$
0.001	23.81	8.25	21.20%	23.45	7.47	21.44%	23.90	8.54	21.46%	23.95	7.94	21.69%	23.28	8.01	20.34%
0.01	21.68	8.07	17.40%	21.12	7.36	17.06%	22.17	8.45	17.49%	21.93	8.34	17.52%	20.65	7.23	15.69%
0.25	18.54	8.31	9.41%	17.93	7.76	8.05%	18.33	7.93	8.10%	17.98	8.00	8.84%	17.41	7.56	7.20%
0.5	12.26	7.38	3.64%	11.64	6.85	2.21%	12.23	7.11	2.92%	11.88	7.43	2.50%	11.95	7.47	3.19%
0.75	7.17	6.04	2.08%	7.06	5.69	1.01%	6.86	5.58	1.14%	6.66	5.34	0.85%	6.97	5.49	1.10%
1	3.53	3.23	0.10%	3.12	2.82	0.29%	3.40	3.07	0.23%	3.38	3.03	0.07%	2.86	2.48	0.00%

Tables 1(i) and (ii) report simulation results under FCFS and ACP policies when $r = 0$, $k = 3/2$, and $L = 20$. The ACP policy consistently outperforms FCFS policy in the

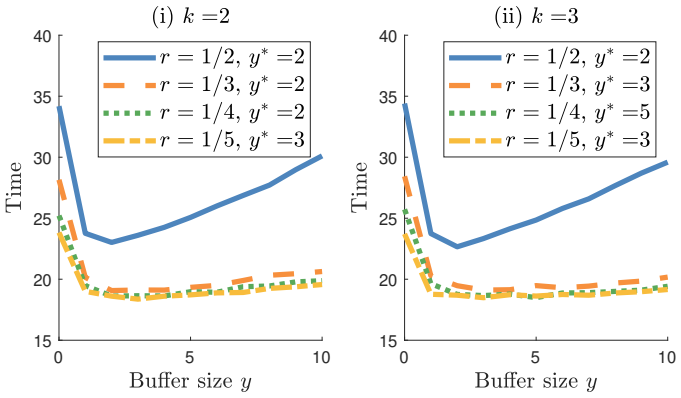
average overall wait, except when the inter-arrival time is long ($\gamma = 1$), in which case customers rarely wait. The average improvement in all 30 experiments is 10.92%. ACP policy also reduces the average wait at the last station, and, in most cases, the probability of excessive waits. These outcomes align with our analytical results in Propositions 1 and 2. Prioritizing advanced customers leads to shorter overall wait, a positive end effect, and potential mitigation of the negative peak effect, and thus benefits customers' waiting experience. Simulation results for $k = 4/3$ and $k = 2$ are in Online Appendix C.

In Table 1(iii), we simulate a system under the ACP policy with a zero-buffer strategy. A comparison between Tables 1(ii) and 1(iii) reveals that implementing a buffer strategy alongside the ACP policy further improves customers' macro-level waiting experience, in line with Proposition 3(i). This observation aligns with our intuition that a buffer strategy, by postponing routing decisions and pooling waiting customers, increases the system's efficiency and consequently reduces customers' average overall wait. The reason ACP policy without a buffer yields shorter average waits at customers' second stations compared to the scenario with zero-buffer is as follows: In the absence of a buffer, some earlier-arrived customers may be assigned to bottleneck station upon arrival (as AB customers) and get delayed by later-joining BA customers. These AB customers, after completing service A, are immediately served at station B with zero delay, significantly influencing the performance measure of $t_{F,2}$.

In Section 3.2, Proposition 3(ii) demonstrates that a buffer size of $y = 1$ surpasses the zero-buffer strategy in a deterministic system with routing time $0 < r < 1/k\mu$. We further extend our analysis to a system with stochastic arrivals and service times ($\mu = 1, N = 50, \gamma = \phi = 0.25, k = 2, 3$). Figure 4 illustrates the average overall wait as a function of the buffer size, with $r = 1/2, 1/3, 1/4, 1/5$. The wait time exhibits non-monotonic behavior relative to the buffer size. As discussed after Proposition 3(ii), a small buffer may lead to stations' unproductive idleness, reducing station utilization, while a large buffer sacrifices routing flexibility, causing some customers to wait unnecessarily in stations' queues, jeopardizing system efficiency. In stochastic systems, a buffer size of $y = 1$ cannot guarantee constant occupation of bottleneck station A, necessitating an optimal buffer size y^* greater than 1.

To conclude, we find that prioritizing customers with one finished service weakly reduces customers' macro-level average overall wait, micro-level average wait at the last station, and the probability of station-level excessive delays in some cases. Moreover, a buffer

Figure 4 **Impact of Routing Time r on Macro-Level Wait under ACP Policy and Buffer Strategy ($\mu = 1, N = 50, \gamma = \phi = 0.25$)**



strategy effectively pools customers so the dispatcher can make more informed routing decisions, improving customers' waiting experience. We verify these insights in a more complex multi-station open-shop service network via a simulation study built on the clinic's data in Section 5. Note that implementing ACP policy and buffer strategy requires support from sophisticated IT systems but is less applicable in manually routed service networks. We next introduce the operations of the clinic that motivated this paper.

4. Descriptive Analytics of the Clinic

Before 2019, the clinic provided executive health screening (EHS), covering over ten tests; see Table OA.1 in Online Appendix A. In 2019, the clinic added two new services: cancer prevention screening (CPS) and integrated prevention screening (IPS), combining EHS and CPS. These services are by appointment only. The clinic schedules customers on specific dates and advises them to arrive between 7:00 A.M. and 8:30 A.M. The clinic operates from 7:00 A.M. until services conclude.

In 2018, the clinic implemented an automated routing system (ARS) to expedite routing decisions. To study the impact of ARS, we examined data from year 2016 (pre-ARS) and year 2019 (post-ARS). Data from 2016 included 10,808 visits over 247 working days, while data from 2019 had 6,733 visits in 146 working days.

4.1. Initial Descriptive Analysis

In this section, we investigate the length of stay of each service type (i.e., EHS, CPS, and IPS) and how it is partitioned into different activities.

Upon arrival, customers register and undergo a fasting blood test. They subsequently take other tests in an arbitrary order, adhering to certain precedence rules. After a test, customers return to the lounge, awaiting routing to the next station by either the dispatcher

or ARS. Next, they wait for the station to call them for the test, with stations making multiple calls if necessary. If a customer does not show up after multiple calls, the station may proceed to call the next customer in line and try the previous customer again later. Customers may leave temporarily in exceptional cases, coordinating with the system as a service break. We focus on post-registration activities, excluding the registration step as it is unaffected by managerial initiatives, without impacting our analysis.

The data encompasses test details for each customer, including the wait start time (i.e., when the routing decision was made), call start times, service completion time, and the corresponding server and examination room. Customers experience two stages of wait before each test. First, they wait for a routing decision, right after completing the prior tests. Then, customers wait for the next test that starts once they are called by the assigned test. As service start times are unspecified in the data, we define service time as the time elapsed from the last call start time to the service completion time, and the call time as the time elapsed from the first to the last call start time.

Table 2 breaks down the length of stay into five components: wait-for-routing, wait-for-service, call time, service time, and break time. The last two columns assess two station-level metrics: average wait-for-service and the probability of waits exceeding 20 minutes. In both years, customers faced substantial delays, and waiting for routing decisions or services consumed nearly 60% of the total stay. For instance, in 2019, on average, EHS and CPS customers visited around 7 stations and spent over 100 minutes waiting. IPS customers, with around 5 additional stations, waited about 42 minutes longer. All service types faced a probability exceeding 15% of waiting more than 20 minutes for services.

Table 2 Length of Stay Partition over Different Activities and Station-Level Waits

Year	Service type	Arrivals per day	Num of stations		Length of stay (minutes)												Station-level wait-for-service	
					Wait-for-routing		Wait-for-service		Call		Service		Break		Total			
			Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	>20 min (%)
2016	EHS	43.76	8.53	3.04	12.53	9.63	88.65	44.77	6.94	6.81	54.10	24.21	11.74	26.79	173.96	73.85	10.80	13.12%
2019	EHS	31.46	7.06	2.44	9.62	9.16	97.28	49.68	3.91	5.01	60.25	24.73	6.12	18.49	177.18	74.77	13.71	21.72%
	CPS	8.67	7.09	1.26	12.21	8.94	108.06	42.48	3.97	5.45	57.31	19.84	7.14	18.94	188.68	56.26	15.23	25.95%
	IPS	5.75	11.63	2.25	20.13	11.82	135.71	46.37	7.12	6.86	93.68	26.01	6.37	18.51	263.01	63.05	11.69	16.62%

4.2. Current Priority and Routing Policies

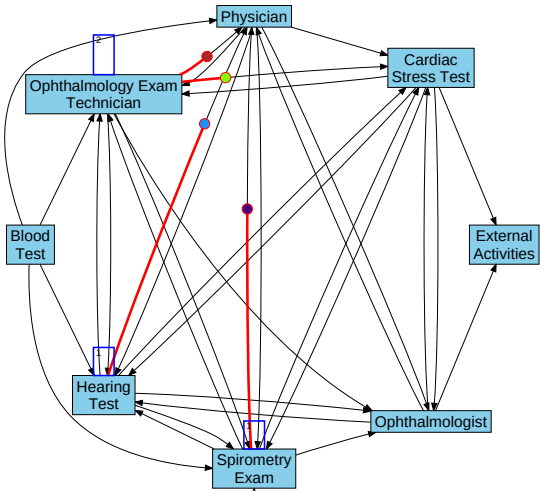
The clinic employs a *station-level FCFS priority policy*, calling the head-of-line (HOL) customer when a station becomes available. By prioritizing customers who arrive at the station early and wait for the longest time, this policy reduces excess wait. The dispatcher's routing decisions consider static long-term congestion levels (each station's average service

time) and dynamic real-time congestion levels (queue length and HOL customer's wait). Further routing policy details are provided in Online Appendix D. The ARS adheres to this routing policy, as the data confirm.

4.3. Impacts of ARS

The clinic implemented ARS to reduce customers' wait-for-routing and streamline clinic operations. The ARS exclusively operates at seven stations (see Figure 5) that are open throughout the entire day. Consequently, ARS was exclusively utilized for routing male EHS customers, enabling them to move efficiently between stations, as depicted in Figure 5, bypassing the need for dispatcher's routing station as shown in Figure 1. All other customer groups necessitating visits to other stations that are only available during specific time windows within the workday continued to be routed manually by the dispatcher.

Figure 5 **Snapshot of the Clinic's Open-Shop Service Network, under ARS for male EHS Customers**



4.3.1. Overall Wait

We investigate the effects of ARS by first looking at the descriptive statistics of male EHS customers on days with and without ARS:

Observation 1 *ARS does not reduce but rather redistributes customers' overall wait—it shortens wait-for-routing and lengthens wait-for-service.*

Table 3 **Length of Stay Partition of Male EHS Customers in Manually- and Auto-Routed Days**

Length of stay	Wait-for-routing		Wait-for-service		Call		Service		Break		Total	
(Minutes)	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Manually-routed	12.89	10.05	101.55	35.95	3.54	4.04	68.18	14.69	2.22	8.29	188.38	41.20
Auto-routed	5.07	6.02	107.45	36.78	4.17	5.29	66.71	15.52	1.34	5.65	184.75	39.96

Table 3 compares activity times for male EHS customers in 15 manually-routed days (no ARS) and 15 auto-routed days (on average, 80% of male EHS customers were auto-routed) in 2019. These 30 days had similar customer compositions and workloads (averaging 50 customers visiting eight stations each), ensuring the system performance analysis is less influenced by workload fluctuations. Table 3 supports Observation 1. ARS significantly reduces wait-for-routing by 7.82 minutes ($p < 0.001$) and increases wait-for-service by 5.90 minutes ($p = 0.05$). Auto-routed customers still experience some wait-for-routing because the dispatcher occasionally overrides ARS (e.g., accommodating real-time test list changes). However, overall waits, call times, and service times do not significantly differ between groups ($p > 0.1$). ARS reallocates waits between wait-for-routing and wait-for-service but does not reduce overall wait, as the routing process is not the clinic's bottleneck; wait-for-service is considerably longer than wait-for-routing.

4.3.2. Dynamics of Waits

We next investigate how ARS affects the dynamics of two micro-level measures, the station-level wait-for-service and the probability of this wait exceeding 20 minutes. Figure 6 presents the mean and the 95% confidence interval of both measures as functions of the visited station order. Figures 6(a)–(b) include visits of all service types (EHS, CPS, and IPS) in 2019, while Figures 6(c)–(d) include only male EHS customers' visits in these 30 days (15 with and 15 without ARS operating).

Figure 6 Micro-Level Measures: Dynamics of Wait-for-Service and Probability of Excessive Waits

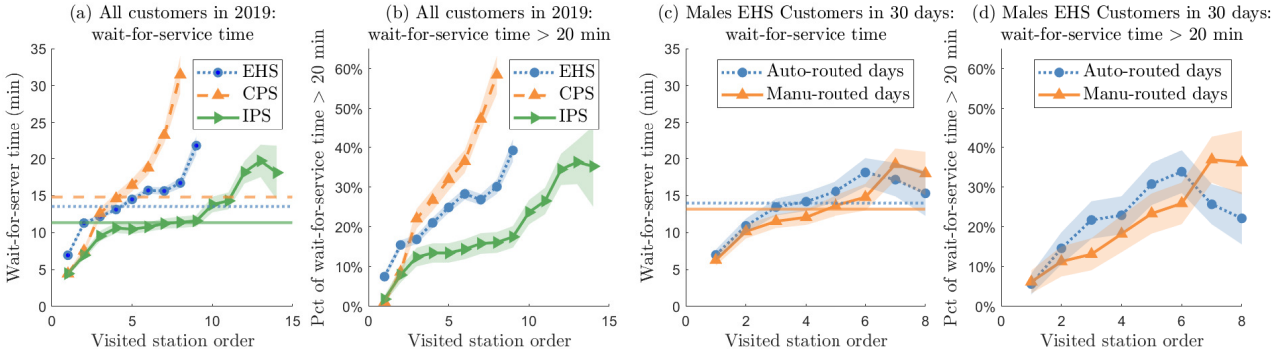


Figure 6(a) illustrates that customers experience longer delays as they progress towards the end of their visits—delays during the latter half of the visit are above the average level (horizontal lines). Such a trend persists in the station-level probability of excessive waits (see Figure 6(b)) and regardless of ARS operation (see Figures 6(c)–(d)). (For auto-routed

customers in Figure 6(d), there is some reduction in the last two stations, but the reduction is not significant and the overall trend is increasing.) Due to the end and trend effects, these dynamics can negatively affect customers' perceived waiting experience (see Section 2.1). The following observation identifies two factors leading to such increasing dynamics.

Observation 2 *Under the station-level FCFS policy, the micro-level performance measures, i.e., wait-for-service and the probability of excessive wait per station, increase as the service process progresses. ARS worsens these increasing trends.*

First, the clinic uses FCFS policy to govern the service order in each station. This policy prioritizes customers according to their *station-level* wait and ignores their *system-status* information, such as arrival time, service history, and expected future service requests. Overlooking such system-status information may contribute to the long waits at the end of the process and lead to a long macro-level overall wait. For example, customers who have finished most tests typically arrive at their last stations in heavily loaded condition and may be delayed by customers who have recently arrived at the clinic for long time. Second, from Figures 6(c)–(d), we observe that at the 3rd to 6th stations, auto-routed customers experience significantly *longer* wait-for-service and a *higher* probability of long delays than manually-routed customers ($p < 0.05$ at 3rd, 4th, and 6th stations, $p < 0.1$ at 5th station), indicating that the automatic routing maintained and even worsened these increasing trends. Notably, ARS has a minimal and statistically insignificant impact on wait-for-service at the first and last stations ($p > 0.1$ at 1st, 2nd, 7th, and 8th stations). The first two stations follow a strict order of tests, and the last few routing decisions offer limited or no choices for subsequent tests.

4.3.3. Quality of Routing Decisions

Here, we investigate the reason for Observation 2 that ARS worsens the increasing trends of micro-level performance measures, the wait-for-service and the probability of this wait exceeding 20 minutes. We suspect it is because the *immediacy* of routing decisions creates a mismatch between customers' demand and the system's service capacity. For instance, when a customer is already assigned to one station and a required test becomes available, both the customer and the available station remain idle due to the assignment to the busy station. We capture this mismatch using *chance of regret*, measuring the fraction of customers whose waiting could be reduced by delaying routing decisions by t_D. Suppose

customer i is routed to station s_1 at time $t+t_D$, where $t_D=0$ under ARS and $t_D>0$ under manual routing. We calculate the probabilities that customers will be routed to a non-bottleneck station s_2, s.t. $s_2 \neq s_1$, with a shorter queue. When we delay routing decisions by $t_D=3$ minutes, these probabilities on the 15 manually- and auto-routed days are 2.92% and 3.86%, respectively ($p<0.01$). This increase of 32.2% in the chance of regret indicates that ARS exacerbates the mismatch between customers' demand and the clinic's service capacity, leading to longer customer waits. Therefore:

Observation 3 *ARS's immediacy may lead to information loss for routing decision-making, which reduces system's flexibility, and causes a high chance of regret and long overall waits for customers.*

Observation 3 signals an opportunity to improve routing decisions by postponing them to accumulate more information. Intuitively, the benefit of postponement is that it facilitates the pooling of queues, which provides more flexibility than systems with dedicated queues, and thus reduces waits. In our case, postponing routing decisions until some stations become available can capture some benefits of pooling queues.

4.4. ARS: Summary and Opportunities

To conclude, ARS effectively accelerates routing processes but has limited impact on improving customers' waits as the routing is not the system's bottleneck. The station-level FCFS policy leads to longer waits and higher chances of excessive delays as customers get closer to the end of their visits, indicating that these last few stations play a significant role in the long average overall wait. Ineffective routing decisions have contributed to these stations' poor micro-level service performance, exacerbated by ARS's immediacy, which overlooks critical information during routing.

Under the clinic's current implementation, ARS has not yielded significant benefits, and in some cases, it may have worsened the situation. Nevertheless, the availability of ARS provides opportunities to apply sophisticated priority and routing policies, which were impractical with manual routing. We propose two such policies: the *advanced customer priority* (ACP) policy, which prioritizes customers closer to completing their clinic visits by considering their system-status information, and the *buffer strategy*, which limits station queue lengths with a buffer size, postponing routing decisions until some stations have fewer customers. These two policies bring fundamental changes to the clinic's operations

and have the potential to fully leverage ARS to improve the customer waiting experience. We conjecture that combining ACP policy with a buffer strategy should outperform FCFS policy in macro- and micro-level measures. In Section 5, we examine this conjecture in a simulation model built on the clinic's data.

5. Evaluation of ACP Policy and Buffer Strategy

In this section, we use a simulation model calibrated with the clinic's data to examine the effectiveness of the ACP policy and the buffer strategy in a general service network. We note that both policies are extremely difficult, if not impossible to implement under the manual routing. The implementation of ARS makes both policies practical.

We first build a data-driven simulation model that represents the clinic's operations by keeping the station-level FCFS policy and the routing policy (see Section 4.2) and all other elements—customers' arrival times, service needs, routing times, call and service times, and servers' break times after services—the same as in the data collected from the clinic.

Table 4 rows 1–2 present the performance measures of the empirical data and the simulation, including the mean and the standard deviation of the macro-level measures of wait-for-routing, wait-for-service, and overall wait, plus micro-level measures of wait at the last station and the probability of station-level wait-for-service exceeding 20 minutes. The difference between the wait-for-routing in rows 1 and 2 is because the dispatcher may occasionally violate FCFS discipline—she may intentionally keep some customers for stations opening soon. Therefore, when we simulate the dispatcher as an FCFS service station without delays, customers' average wait-for-routing is reduced. And this reduction is captured by the increase in wait-for-service. Comparing the first two rows shows that our simulation model closely mimics the clinic's operations. Hence, we use the simulation results in row 2 as our baseline: from the macro perspective, customers' average overall wait is 114.37 minutes, with a standard deviation of 57.47 minutes; and from the micro perspective, on average, the wait at customers' last station is 19.88 minutes and the probability of waits exceeding 20 minutes is 21.36%.

5.1. ACP Policies

Propositions 1 and 2 prove that the ACP policy could simultaneously improve macro- and micro-level performance measures compared to the FCFS policy. We next extend the concept of ACP policy to more practical scenarios with two generalizations:

Table 4 Macro- and Micro-Level Performance Measures under FCFS and ACP Policies

Priority policies	Macro-level wait (min)						Micro-level wait			
	Wait-for-routing		Wait-for-service		Total		Last wait		>20 min (%)	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
1 Empirical data	11.38	10.10	103.62	50.04	115.00	53.96	19.87	19.92	22.20%	18.56%
2 FCFS	8.48	5.84	105.89	55.15	114.37	57.47	19.88	20.97	21.36%	17.34%
3 ACP-LST	7.82	5.24	101.01	56.04	108.83	57.45	17.72	28.53	18.85%	17.21%
4 ACP-SERPT	7.07	4.88	92.44	73.88	99.51	75.69	10.19	15.55	13.35%	13.22%

- *ACP-Longest-system-time first (LST) policy* prioritizes the customer who experiences the longest system time. This policy can be thought of as a clinic-level FCFS policy. In contrast to the standard, station-level FCFS policy that considers arrival time at the station, LST priority policy uses the arrival time at the clinic: customers who arrive earlier at the clinic have priority over customers who arrive later.
- *ACP-Shortest-expected-remaining-processing-time first (SERPT) policy* prioritizes the customer with the shortest expected remaining service time in all remaining stations in order to move customers out of the system faster. This policy forecasts the processing time (i.e., call and service times) required to complete customers' unfinished tests.

We note that in the two-station model in Section 3, all customers arrive at time 0 and visit both stations. Then, ACP-LST and ACP-SERPT policies are identical. However, when the number of stations exceeds two and customers have varying arrival times or service needs, distinctions between these two policies arise. While both policies, as the core ACP, prioritize customers closer to exiting the system, ACP-LST makes decisions based on customers' service history, whereas ACP-SERPT looks ahead in time.

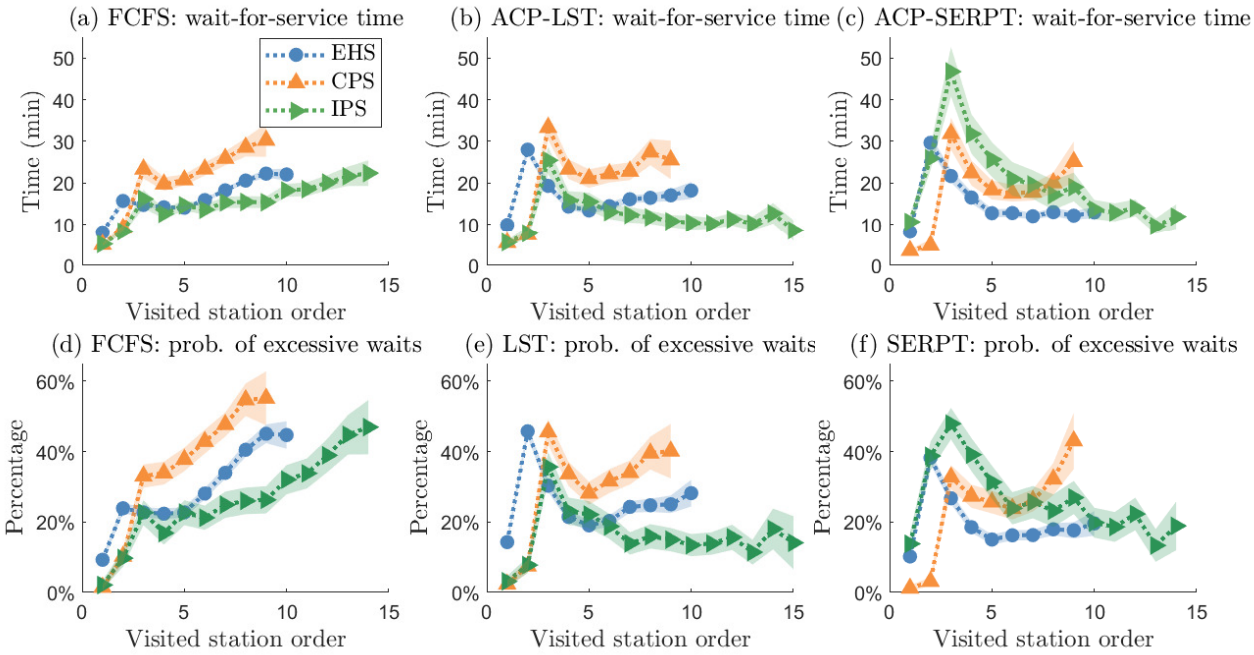
Simulation results are provided in Table 4 rows 3–4. As in the baseline in row 2, we use the clinic routing policy and the clinic's data on the customers' arrival times, service needs, routing times, call and service times. We only change the priority policy, which determines stations' service orders (i.e., the order in which customers are served) and thus affects customers' service orders (i.e., the order in which stations are visited). We expect customers' average overall wait and the last station's wait to be reduced.

As conjectured in the discussion of Observation 2, the impact of priority policies on waits is significant. From the macro perspective, implementing ACP policies reduces the overall wait. Compared to FCFS policy, ACP-LST and ACP-SERPT policies shorten the average overall wait from 114.37 minutes to 108.83 minutes (4.84% reduction) and 99.51 minutes (12.99% reduction), respectively. From the micro perspective, ACP-LST and ACP-SERPT

policies effectively reduce the average wait at the last station from 19.88 minutes to 17.72 minutes (10.87% reduction) and 10.19 minutes (48.74% reduction), and the probability of excessive delays from 21.36% to 18.85% (11.75% reduction) and 13.35% (37.50% reduction). All these improvements are statistically significant compared with the baseline, with p-values smaller than 0.01. ACP policies not only reduce the average overall wait but also create a positive end effect and mitigate the potential negative peak effect, improving the overall waiting experience for customers.

These two ACP policies also reduce waits as customers progress toward the end of their visits, generating a positive trend effect for customers. In Figure 7, we plot the station-level wait dynamics over times (the mean and the 95% confidence interval of wait-for-service in Figure 7(a)–(c) and the probability of waits exceeding 20 minutes in Figure 7(d)–(f)) for three service types (i.e., EHS, CPS, and IPS) under FCFS, ACP-LST, and ACP-SERPT policies. Similar to Figures 6(a)–(b), Figures 7(a) and (d) show that customers face increasing micro-level waits under FCFS policy. As shown in Figures 7(b)–(c) and (e)–(f), replacing FCFS policy with ACP policies moderates these increasing trends. ACP-SERPT policy significantly reduces the waits at the last few stations, especially for EHS and IPS customers. Customers wait shorter as their visits proceed to the end.

Figure 7 Dynamics of Wait-for-service and Probability of Excessive Waits Under FCFS and ACP Policies



Our final observation from Table 4 is that ACP-SERPT policy minimizes all performance measures and has an advantage over ACP-LST policy. ACP-SERPT policy is a forward-looking policy that prioritizes customers closer to the end of their visits. Figures 7(c) and (f) strengthen this conclusion and show customers have shorter waits at their last few stations. Nevertheless, this comes at the cost of longer waits at the beginning of the process (stations 2–3) and results in a higher probability of excessive wait peaks. In contrast, ACP-LST policy is backward-looking. It prioritizes customers with the longest system time to avoid customers staying in the system for too long. This policy seems to balance wait over the service process better than the other policies do (see Figures 7(b) and (e)), but it does not optimize the macro- or micro-level performance measures.

In Online Appendix E, we obtain similar observations when using the Shortest-Expected-Wait-Time first routing policy to make routing decisions.

5.2. Buffer Strategy

As discussed in Observation 3, ARS's immediacy may lead to system inefficiency—chance of regret (stations become idle while customers are waiting for assigned busy stations), causing excessive waits. We propose a buffer strategy that increases operational flexibility by postponing routing decisions and forming a pooling effect. Customers are routed to stations only if stations' queue lengths are below a threshold buffer size. All other customers wait in a centralized virtual pooled queue. When some station completes one service, and the queue length decreases by one, an available customer will be selected to join that station's queue. Such a pooling of queues uses the station's capacity more efficiently. Stations are less likely to be idle when there are potential customers waiting in the pool.

Proposition 3 demonstrates that, under ACP, the optimal buffer size depends on routing time. While a zero-buffer strategy is ideal for a two-station network when routing times are minimal and arrival and service times are deterministic as in Proposition 3(i), real-world situations are more complex. A positive buffer should be considered for the clinic for the following reasons: First, similar to Proposition 3(ii), the dispatcher requires time to assign customers to their next stations, so making routing decisions in advance prevents unnecessary servers' idleness. Second, customers need to travel from the waiting lounge to examination rooms, which means routing decisions should be made before the next station becomes available. Third, customers may not show up after being called for service. The no-show rates of service calls were 20.70% and 9.83% in 2016 and 2019, respectively. A

non-zero buffer thus reduces idleness caused by no-shows. By considering these factors, a positive buffer size can balance the trade-off between routing flexibility at individual stations and the pool, ultimately improving overall system efficiency.

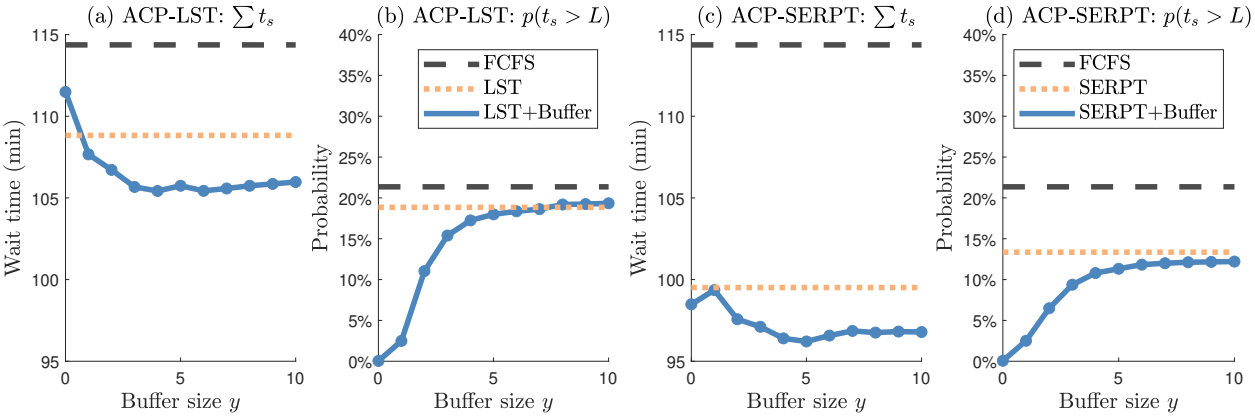
We next conduct simulations under various buffer sizes. Similar to simulations conducted in Table 4, we keep customers' arrival times, service needs, routing times, and call and service times the same as in the clinic's data. The only difference between our simulation and the clinic's operations is the priority and routing policies: we replace the station-level FCFS policy with ACP policies and cap stations' queue lengths by a certain buffer size.

Figures 8(a) and (c) illustrate simulation results of average overall wait, $\sum t_s$, as a function of buffer size under ACP-LST and ACP-SERPT policies. Dashed and dotted lines represent results obtained under FCFS and ACP policies without the buffer strategy (i.e., rows 2–4 in Table 4). Similar to Figure 4, the simulation results highlight a non-monotonic relationship between the buffer size and average overall wait in such a complex multi-station open-shop network. This non-monotonic relationship arises due to the trade-off between station utilization and flexibility in dispatching from the pool. With a small buffer, the dispatcher can make more informed routing decisions owing to the presence of more awaiting customers with potentially higher priorities in the pool. However, this may adversely impact station utilization, as stations may occasionally need to wait for customers to arrive upon completing all customers in line. Conversely, as the buffer size increases, fewer customers wait in the pool, resulting in a loss of flexibility for the dispatcher in routing customers to stations. On the other hand, stations experience less idle time and can select customers according to the specific priority policy.

Figures 8(a) and (c) demonstrate that introducing finite buffers reduces average overall waits. For example, with the ACP-LST policy, a buffer size of 4 minimizes overall wait, resulting in a 7.82% reduction from 114.37 minutes to 105.42 minutes. Similarly, employing the ACP-SERPT policy with a buffer size of 5 reduces the average overall wait from 114.37 minutes to 96.21 minutes, achieving a 15.88% reduction.

In addition to the macro-level improvement, the buffer strategy also reduces the probability of station-level excessive waits, $p(t_s > L)$, to a great extent, as seen in Figures 8(b) and (d) for $L = 20$ minutes. Setting a buffer size of 5 and replacing FCFS with ACP-LST (resp. ACP-SERPT) policy reduces the probability $p(t_s > 20)$ by 15.78% (resp. 46.98%), from 21.36% to 17.98% (resp. 11.32%). The micro-level outperforming is partly attributed to

Figure 8 Macro- and Micro-Level Performance Measures Under ACP Policies and Buffer Strategy



that the buffer strategy enables the dispatcher to use more system information for decision-making, and some of the wait-for-service becomes wait-for-routing. Moreover, from the customers' perspective, the buffer strategy reduces the chance of regret. Setting a buffer size of 5 and replacing FCFS with ACP-LST and ACP-SERPT policies reduce the chance of regret from 2.66% to 2.40% and 2.04%, respectively.

To conclude, the simulation results verify our insights from Sections 3 and 4 that the ACP policy accounting for individual customers' system-status information can reduce customers' macro-level overall wait, micro-level wait at the last station, and the probability of excessive waits in an open-shop service network. Furthermore, these findings underscore the significance of considering buffer size when optimizing the efficiency of a multi-station open-shop network. They highlight the benefits of integrating the buffer strategy with the ACP policy to further improve system performance. This indicates that if the clinic and the management of open-shop service networks could apply some advanced customer priority policy and implement an appropriate buffer strategy, both the macro- and micro-level performance measures could be significantly improved.

6. Conclusion

This paper examines the impact of information technology—an automated routing system (ARS)—on an open-shop service network's performance, which is evaluated using the macro-level measure of average overall wait and micro-level measures of the probability of excessive delays at stations, the wait at the last station, and the dynamics of waits along the service process. We find that the implemented technology fails to significantly improve the clinic's operations because these were not fundamentally changed. However, we recognize that ARS is a necessary enabler for implementing operational interventions which can

lead to improvements. We propose practical, theory-supported, and simulation-validated priority and routing policies to improve customers' waiting experience.

First, we discover that the myopic station-level FCFS policy, neglecting customer system-status profiles, may lead to inferior system performance. Customers who have stayed in the network for a long time, or are expected to finish their visit in a short time, often experience excessive delays at their last few stations. Through theoretical and simulation analyses, we show that ACP policy outperforms FCFS and improves waits at both macro- and micro-levels. Second, we notice that ARS makes routing decisions once customers finish their previous tests. This sometimes caused customers to wait at congested stations while other stations were idle. We propose a buffer strategy that limits each station's queue length below a certain buffer and further improves micro- and macro-level performance.

We provide several insights for technology implementations and customer waiting experience management in open-shop service networks. First, routing decisions should consider customer profiles such as arrival time and left-to-visit stations. Second, counter-intuitively, immediate dispatching doesn't necessarily improve system performance. Delaying routing decisions can increase system flexibility, balance waits throughout visits, and improve both macro and micro-level performance. Moreover, technologies may not always benefit the system. There is often a gap between the technology's inherent value and its practical effectiveness. To fully leverage new technologies, management must rethink and possibly redesign operations. This could lead to fundamental changes, rather than trying to fit the technology into existing practices.

Our paper gives rise to several questions for future work. While ACP improves both macro- and micro-level performance, this hinges on maintaining the workload at the bottleneck station. When not all customers require services from both stations, designated as A-only and B-only customers, a logical routing policy involves directing all A-only customers to station A while allocating the remaining customers to station B. In this case, prioritizing B-only customers over BA customers at station B may cause idling at the bottleneck station A, leading to system inefficiency. Insights obtained from our analytical model indicate that when station A is nearing the completion of its queue, station B ought to prioritize serving some BA customers to replenish queue A and sustain the bottleneck station A's productivity. Future studies should explore ACP's adaptability to

varying service needs among customers. We have done initial investigation into this direction in Online Appendix F. Besides, determining optimal buffer sizes tailored to individual stations is an intriguing problem. Proposition 3 suggests that optimal buffers should consider routing and transfer times. Yet, we applied a uniform buffer size in our analysis and simulations. Stations may benefit from station-specific buffer sizes, considering factors like throughput, service rate, room distances, and customer behavior. From a fairness perspective, the ACP policy, while prioritizing certain customers, ensures macro-level fairness. Our proposed ACP-LST policy operates as a macro system-level FCFS policy, aligning with established fairness principles (see, e.g., Maister 1985, Larson 1987). The ACP-SERPT policy, essentially a stochastic adaptation of ACP-LST, assumes that all customers have identical service needs, promoting equitable customer treatment. Within the open-shop environment, managers have the flexibility to deviate from FCFS at the micro station-level. The complexity introduced by multiple routes makes it challenging for customers to discern fairness in specific orders or the absence thereof. This intriguing scenario prompts a future research question on balancing macro- and micro-level fairness within such operational settings. Finally, with technological advances, it would be natural to explore and implement hybrid policies that dynamically shift among various priority policies depending on system state and managerial goals.

References

Afrati F, Cosmadakis S, Papadimitriou CH, Papageorgiou G, Papakostantinou N (1986) The complexity of the travelling repairman problem. *RAIRO - Theoretical Informatics and Applications - Informatique Théorique et Applications* 20(1):79–87.

Anand E, Panneerselvam R (2015) Literature review of open shop scheduling problems. *Intelligent Information Management* 7(1):33–52.

Ariely D, Carmon Z (2000) Gestalt characteristics of experiences: the defining features of summarized events. *Journal of Behavioral Decision Making* 13(2):191–201.

Ariely D, Zauberman G (2000) On the making of an experience: The effects of breaking and combining experiences on their overall evaluation. *Journal of Behavioral Decision Making* 13(2):219–232.

Ariely D, Zauberman G (2003) Differential partitioning of extended experiences. *Organizational Behavior and Human Decision Processes* 91(2):128–139.

Arlotto A, Frazelle E, Wei Y (2018) Strategic open routing in service networks. *Management Sci.* 65(2):735–750.

Averbakh I, Berman O, Chernykh I (2006) The routing open-shop problem on a network: Complexity and approximation. *Eur. J. Oper. Res.* 173(2):531–539.

Baron O, Berman O, Krass D, Wang J (2014) Using strategic idleness to improve customer service experience in service networks. *Oper. Res.* 62(1):123–140.

Baron O, Berman O, Krass D, Wang J (2017) Strategic idleness and dynamic scheduling in an open-shop service network: Case study and analysis. *Manufacturing & Service Oper. Management* 19(1):52–71.

Bitran G, Ferrer J, Oliveira P (2008) Managing customer experiences: Perspectives on the temporal aspects of service encounters. *Manufacturing & Service Oper. Management* 10(1):61–83.

Bray RL (2020) Operational transparency: Showing when work gets done. *Manufacturing & Service Oper. Management* 25(3):812–826.

Chase RB, Dasu S (2001) Want to perfect your company's service? Use behavioral science. *Harvard Business Review* 79(6):78–84.

Chou S, Lin S (2007) Museum visitor routing problem with the balancing of concurrent visitors. Loureiro G, Curran R, eds., *Complex Systems Concurrent Engineering*, 345–353 (Springer).

Das Gupta A, Karmarkar US, Roels G (2016) The design of experiential services with acclimation and memory decay: Optimal sequence and duration. *Management Sci.* 62(5):1278–1296.

Dixon M, Verma R (2013) Sequence effects in service bundles: Implications for service design and scheduling. *Journal of Operations Management* 31(3):138–152.

Gallino S, Karacaoglu N, Moreno A (2023) Need for speed: The impact of in-process delays on customer behavior in online retail. *Oper. Res.* 71(3):876–894.

Geng X, Chen Z, Lam W, Zheng Q (2013) Hedonic evaluation over short and long retention intervals: The mechanism of the peak-end rule. *Journal of Behavioral Decision Making* 26(3):225–236.

Hansen DE, Danaher PJ (1999) Inconsistent performance during the service encounter: What's a good start worth? *Journal of Service Research* 1(3):227–235.

Huang J, Carmeli B, Mandelbaum A (2015) Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Oper. Res.* 63(4):892–908.

Ibrahim R, Whitt W (2010) Real-time delay estimation based on delay history in many-server service system with time-varying arrivals. *Prod. and Oper. Management* 20(5):654–667.

Kahneman D (2000) Evaluation by moments: Past and future. Kahneman D, Tversky A, eds., *Choices, Values and Frames*, 693–708 (Cambridge University Press, New York).

Kleinrock L (1976) *Queueing systems, Volume 2: Computer applications* (Wiley-Interscience, New York).

Kong Q, Li S, Liu N, Teo CP, Yan Z (2020) Appointment scheduling under time-dependent patient no-show behavior. *Management Sci.* 66(8):3480–3500.

Larson RC (1987) Perspectives in queues: Social justice and the psychology of queueing. *Oper. Res.* 35(6):895–905.

Lee HH, Pinker EJ, Shumsky RA (2012) Outsourcing a two-level service process. *Management Sci.* 58(8):1569–1584.

Maister DH (1985) The psychology of waiting lines. Czepiel JA, Solomon MR, Suprenant CF, eds., *The Service encounter*, 113–123 (Lexington Books).

Mandelbaum A, Reiman MI (1998) On pooling in queueing networks. *Management Sci.* 44(7):971–981.

Osuna EE (1985) The psychological cost of waiting. *Journal of Mathematical Psychology* 29(1):82–105.

Parlaktürk K, Kumar S (2004) Self-interested routing in queueing networks. *Management Sci.* 50(7):949–966.

Pinedo L (1984) A note on the flow time and the number of tardy jobs in stochastic open shops. *Eur. J. Oper. Res.* 18(1):81–85.

Pinedo L (2016) *Scheduling: Theory, Algorithms, and Systems*. SpringerLink : Bücher (Springer New York).

Pinedo L, Ross SM (1982) Minimizing expected makespan in stochastic open shops. *Advances in Applied Probability* 14(4):898–911.

QYResearch Group (2022) Global physical examination market size, status and forecast 2022-2028, https://www.marketresearch.com/LP-Information-Inc-v4134/Global-Physical-Examination-Growth-Status-31591895/.

Raz D, Avi-Itzhak B, Levy H (2006) Fairness considerations of scheduling in multi-server and multi-queue systems. *Proceedings of the 1st International Conference on Performance Evaluation Methodolgies and Tools*, 39–es (New York, NY, USA: Association for Computing Machinery).

Shtrichman O, Ben-Haim R, Pollatschek MA (2001) Using simulation to increase efficiency in an army recruitment office. *Interfaces* 31(4):61–70.

Song H, Tucker AL, Murrell KL (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Sci.* 61(12):3032–3053.

Taylor S (1994) Waiting for service: The relationship between delays and evaluations of service. *Journal of Marketing* 58(2):56–69.

Verhoef P, Antonides G, de Hoog A (2004) Service encounters as a sequence of events: The importance of peak experiences. *Journal of Service Research* 7(1):53–64.

Wang J, Zhou YP (2018) Impact of queue configuration on service time: Evidence from a supermarket. *Management Sci.* 64(7):3055–3075.

Wang S, Liu N, Wan G (2019) Managing appointment-based services in the presence of walk-in customers. *Management Sci.* 66(2):667–686.

Westphal M, Yom-Tov GB, Parush A, Rafaeli A (2022) Reducing abandonment and improving attitudes in emergency departments: Integrating delay announcements into operational transparency to signal service quality, working paper, Technion.

Online Appendix to
"Waiting Experience in Open-Shop Service Networks: Improvements via Flow Analytics & Automation"

Appendix A: Station List and Station-Level Performance of the Clinic

Table OA.1: Change of Stations and Stations' Performance in the Empirical Data

Stations[a]	Year 2016								Year 2019							
	#Servers	Utilization[b]	Wait-for-service		Call		Service		#Servers	Utilization[b]	Wait-for-service		Call		Service	
			Mean	>20min	Mean	CV	Mean	CV			Mean	>20min	Mean	CV	Mean	CV
EKG	1.82	69.90%	06:34	5.25%	01:05	1.67	04:50	0.51								
Blood Test	2.17	67.34%	09:07	11:07%	00:30	2.74	05:18	0.68	2.62	72.66%	07:12	8.07%	00:17	3.13	05:37	0.50
Gynecological Ultrasound	1.02	73.55%	10:25	14.07%	00:58	2.66	07:24	0.56								
Gynecologist	1.05	59.99%	14:06	24.15%	00:21	3.16	02:58	2.31	1.04	73.45%	19:20	40.21%	00:36	3.38	08:33	0.67
Brest Surgeon	1.08	68.63%	17:59	36.88%	00:36	3.15	05:23	1.89	1.00	77.34%	21:31	40.17%	00:25	3.87	06:26	0.81
Cardiac Stress Test	4.69	71.82%	08:35	12.95%	00:48	2.16	21:11	0.32	3.99	83.77%	18:37	36.50%	00:43	2.75	24:21	0.30
Spirometry	1.04	52.03%	06:51	4.18%	00:44	2.17	03:01	0.50	1.09	71.36%	11:37	14.06%	00:36	2.82	04:07	0.40
Ophthalmology	1.21	61.39%	07:24	5.07%	00:51	2.26	02:37	0.74	1.04	59.14%	11:39	15.16%	00:33	2.81	02:54	0.85
Opthalmologist	1.32	53.41%	09:55	12.98%	00:34	3.13	02:53	2.13	1.05	67.41%	09:58	12.12%	00:30	2.99	03:41	0.78
Cardiologist	2.08	61.89%	08:43	9.82%	01:17	2.19	06:18	0.91	1.16	30.60%	07:28	8.71%	00:54	2.89	06:01	1.04
Physician 1	3.82	68.04%	12:31	20.42%	01:15	2.65	11:16	0.71	3.44	81.42%	18:54	37.54%	00:31	3.16	15:05	0.68
Hearing Test	1.11	70.43%	09:19	12.73%	00:59	2.21	04:41	0.99	1.77	74.16%	13:48	21.02%	00:49	2.35	07:41	0.44
Urine Test	1.00	52.26%	05:32	0.31%	01:43	1.44	02:29	1.40								
Review with Cardiologist	1.95	60.35%	10:48	15.85%	00:38	2.51	04:08	1.72	1.19	30.87%	08:13	10.15%	00:36	3.82	04:54	1.55
Chest X-Ray	1.04	66.54%	09:55	10.03%	00:56	1.96	01:50	1.56								
Stomatology									1.03	60.70%	12:00	19.40%	00:45	3.22	03:55	0.95
Plastic Surgeon									1.11	72.28%	14:08	26.25%	00:42	2.44	05:43	0.51
Survey									1.06	45.06%	03:53	0.31%	00:18	2.85	02:46	0.68
Nutritionist									1.07	69.65%	16:27	32.50%	00:39	2.46	13:30	0.43
Urologist									1.01	72.81%	15:15	25.11%	00:38	3.00	06:50	0.55
Physician 2									1.05	83.51%	24:09	47.25%	00:48	3.58	15:42	0.70

[a] **Change of Stations:** Three stations (EKG, urine test, and chest X-ray) provided in 2016 were no longer available in 2019 and five stations (stomatology, plastic surgeon, health history survey, nutritionist, and urologist) were newly added in 2019. Moreover, the gynecological ultrasound was merged with the gynecologist examination. In addition, in 2019, a new physician examination (Physician 2) was provided for IPS customers; EHS and CPS customers can visit either Physician 1 or 2.

[b] **Station Utilization:** For each station and server, we first identify the server's working hours (starting when the server calls the first customer and ending when the last customer leaves the examination room), exclude the break time (we assume a server leaves for a break if she idles for more than 15 minutes while customers are waiting), and then derive the server's busy time (sum of call and service times). Each station's utilization is obtained as the average ratio of servers' busy time and working hours.

[c] **Bottleneck Stations:** We identify three bottleneck stations in 2019 are the cardiac stress test and two types of physician examinations, with utilizations over 80%. Three female-only stations (gynecological ultrasound, gynecologist examination, and breast surgeon) also had high utilizations of over 70% and average wait around 15–20 minutes due to a staffing issue—specialists working at these stations also had shifts outside the clinic and were available for in-clinic service for a limited period.

Appendix B: Proofs of Theoretical Results in Section 3

B.1. Proof of Proposition 1

To prove Proposition 1, we first derive the macro-level average wait under FCFS and ACP policies in Propositions OA.1 and OA.2, respectively, then make the comparison in Section B.1.3.

B.1.1. Macro-level measure under FCFS policy

PROPOSITION OA.1. [**Macro-Level Measure: FCFS**] *Given a routing decision of x AB customers, customers' average overall wait is*

$$t_F(x)=\begin{cases}\frac{N-1}{2\mu} & \text{if } x=0;\\ \frac{-(1+k)x^2+(1-k+2N)x-N(2-(N-1)k)}{2N\mu k} & \text{if } 1\le x\le \frac{N-1}{k};\\ \frac{x^2+(3k-2N-1)x+(k^2-k+1)N^2+(1-2k-k^2)N}{2N\mu k(k-1)}+\frac{\varepsilon}{N} & \text{if } \frac{N-1}{k}<x\le N-k;\\ \frac{kN^2-(2+k)N+2x}{2Nk\mu} & \text{if } N-k<x\le N,\end{cases}$$

where $\varepsilon=\frac{a(2N-2x-k-1-(k-1)a)}{2k\mu}-\frac{(N-x)(N-x-k-1)}{2k\mu(k-1)}$ *and* $a=\left\lfloor\frac{N-x-1}{k-1}\right\rfloor$. *If* $N\ge\frac{k^2+5k+2}{2}$, *then the optimal routing decisions are* $\arg\min t_F(x)=\{1$ *and all integers in* $[N-2k+1,N-k]\}$.

Proof of Proposition OA.1. The average overall wait $t_F(x)$ in Proposition OA.1 is calculated by combining the average waits at stations A and B, $t_1(x)$ and $t_2(x)$. We first derive these in Lemma 1.

LEMMA 1. [**Average Waits at Station A and B**] *Average waits at station A and station B are*

$$t_1(x)=\begin{cases}\frac{(N-1)(k-1)}{2k\mu} & \text{if } x=0;\\ \frac{(k-1)N^2+(2x-k-1)N-x^2+x}{2Nk\mu} & \text{if } 1\le x\le N;\end{cases}\quad\text{and } t_2(x)=\begin{cases}\frac{N^2-N-kx^2-kx}{2Nk\mu} & \text{if } 0\le x\le\frac{N-1}{k};\\ \frac{N^2-2N(x+1)+x^2+2x}{2N\mu(k-1)}+\frac{\varepsilon}{N} & \text{if } \frac{N-1}{k}<x\le N-k;\\ \frac{x^2-(2N-1)x+N^2-N}{2Nk\mu} & \text{if } N-k<x\le N,\end{cases}$$

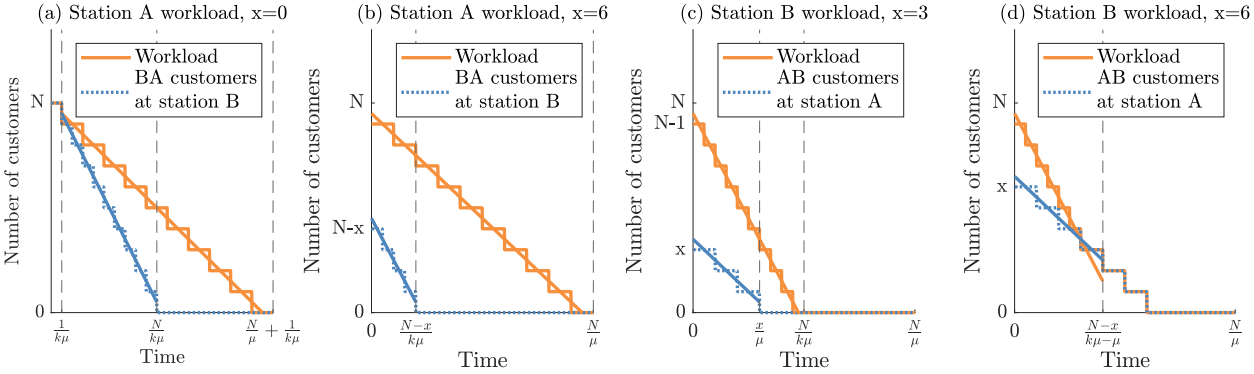
where $\varepsilon=\frac{a(2N-2x-k-1-(k-1)a)}{2k\mu}-\frac{(N-x)(N-x-k-1)}{2k\mu(k-1)}$ *and* $a=\left\lfloor\frac{N-x-1}{k-1}\right\rfloor$.

Proof of Lemma 1. We first derive customers' average wait at station A, $t_1(x)$. If $x=0$, all customers route BA. Starting at $t=\frac{1}{k\mu}$ (i.e., when the first BA customer finishes service B), customers arrive at station A every $\frac{1}{k\mu}$ time unit. Station A serves all N customers during $\frac{1}{k\mu}\le t<\frac{N}{\mu}+\frac{1}{k\mu}$. Customers' average wait at station A is $t_1(x)=\frac{1}{N}(\int_{\frac{1}{k\mu}}^{\frac{N}{\mu}+\frac{1}{k\mu}}(\frac{1}{2}+N-1-\mu(t-\frac{1}{k\mu}))dt-\int_{\frac{1}{k\mu}}^{\frac{N}{k\mu}}(\frac{1}{2}+N-k\mu t)dt)=\frac{(N-1)(k-1)}{2k\mu}$. The first integral accumulates all customers' waits at station A, assuming they arrive at time $\frac{1}{k\mu}$, while the second integral accumulates customers' time spent at station B; see Figure OA.2(a).

If $1\le x\le N$, as station A is the bottleneck, all BA customers find station A busy. The $N-x$ BA customers join queue A during $0<t\le\frac{N-x}{k\mu}$. Station A serves all N customers during $0\le t<\frac{N}{\mu}$. In this case, $t_1(x)=\frac{1}{N}(\int_0^{\frac{N}{\mu}}(\frac{1}{2}+N-1-\mu t)dt-\int_0^{\frac{N-x}{k\mu}}(\frac{1}{2}+N-x-k\mu t)dt)=\frac{(k-1)N^2+(2x-k-1)N-x^2+x}{2Nk\mu}$. Again, the first integral accumulates all customers' waits at station A, assuming they arrive at time 0, while the second integral accumulates customers' time spent at station B; see Figure OA.2(b).

We next derive customers' average wait at station B, $t_2(x)$. If $0\le x\le\frac{N-1}{k}\Leftrightarrow\frac{x}{\mu}\le\frac{N-x-1}{k\mu-\mu}$, the last AB customer joins station B before the queue length at B reduces to 1; i.e., all AB customers find station B busy. AB customers join queue B at $0<t\le\frac{x}{\mu}$, and station B is busy during $0\le t\le\frac{N}{k\mu}$. Customers' average wait at station B is $t_2(x)=\frac{1}{N}(\int_0^{\frac{N}{k\mu}}(\frac{1}{2}+N-1-k\mu t)dt-\int_0^{\frac{x}{\mu}}(\frac{1}{2}+x-\mu t)dt)=\frac{N^2-N-kx^2-kx}{2Nk\mu}$. The first integral

Figure OA.2 Workload at station A and B under different routing decisions ($N=10, \mu=1, k=2$)



accumulates all customers' waits at station B, assuming they arrive at time 0, while the second integral accumulates customers' time spent at station A; see Figure OA.2(c).

If $\frac{N-1}{k} < x \le N-k \Leftrightarrow \frac{1}{\mu} \le \frac{N-x}{k\mu}$ and $\frac{x}{\mu} > \frac{N-x-1}{k\mu-\mu}$, the first AB customer finds station B busy when joining, and the last AB customer arrives at station B after it starts serving the last customer in line; i.e., earlier AB customers encounter station B as busy, while later ones might find it idle. Station B remains busy during $0 < t \le \frac{N-x}{k\mu-\mu}$. After this point, arriving AB customers immediately get served. In this case, $t_2(x) = \frac{1}{N}\int_0^{\frac{N-x}{k\mu-\mu}} (N-x-1-k\mu t+\mu t)dt + \frac{\varepsilon}{N} = \frac{N^2-2N(x+1)+x^2+2x}{2N\mu(k-1)} + \frac{\varepsilon}{N}$, see Figure OA.2(d). Note that this derivation ignores integrality and thus has an error term $\frac{\varepsilon}{N}$. An AB customer starts service B only after completing service A, which means station B may become idle before $t = \frac{N-x}{k\mu-\mu}$. Whether the ith AB customer finds station B busy depends on the time station B takes to serve $N-x+i-1$ customers and the time station A takes to serve i customers. Set $a = \left\lfloor \frac{N-x-1}{k-1} \right\rfloor$, if $i \le a$, the ith AB customer waits for service B; if $i > a$, the ith AB customer is served immediately. Therefore, the error term is $\varepsilon = \sum_{i=1}^{a}(\frac{N-x+i-1}{k\mu} - \frac{i}{\mu}) - (\int_0^{\frac{N-x}{k\mu-\mu}} (\mu t - \frac{1}{2})dt - \int_0^{\frac{N-x}{k\mu-\mu}-\frac{N-x}{k\mu}} (\frac{1}{2} + \frac{N-x}{k-1} - k\mu t)dt) = \frac{a(2N-2x-k-1-(k-1)a)}{2k\mu} - \frac{(N-x)(N-x-k-1)}{2k\mu(k-1)}$.

If $N-k < x \le N \Leftrightarrow \frac{N-x}{k\mu} < \frac{1}{\mu}$. This implies station B completes all BA customers before the first AB customer arrives; i.e., all AB customers are served immediately at station B. Customers' average wait at station B is $t_2(x) = \frac{1}{N}\int_0^{\frac{N-x}{k\mu}} (\frac{1}{2} + N - x - 1 - k\mu t)dt = \frac{x^2-(2N-1)x+N^2-N}{2Nk\mu}$. □

The average overall wait, $t_F(x)$, is the sum of $t_1(x)$ and $t_2(x)$. We then derive $\arg\min t_F$ as follows.

For $1 \le x \le \frac{N-1}{k}$, $t_F(x)$ is concave and maximized at $\frac{2N-k+1}{2(1+k)}$. Under the condition that $N \ge \frac{k^2+5k+2}{2}$, we have $1 < \frac{2N-k+1}{2(1+k)} < \left\lfloor \frac{N-1}{k} \right\rfloor$ and $\left\lfloor \frac{N-1}{k} \right\rfloor - \frac{2N-k+1}{2(1+k)} < \frac{2N-k+1}{2(1+k)} - 1$, implying that $t_F(x)$ is minimized at $x=1$. Moreover, we have $t_F(1) < t_F(0)$.

For $\frac{N-1}{k} < x \le N-k$, we have $t_F(x) - t_F(1) = \frac{1-a}{2Nk\mu}((k-1)a - 2(N-k-x))$. Let $f(a) = (k-1)a - 2(N-k-x)$, we have $f(a) \le f(\frac{N-x-1}{k-1}) = 2k - N + x - 1$. For $x < N-2k+1$, we have $f(a) \le f(\frac{N-x-1}{k-1}) < 0$, and thus, $t_F(x) > t_F(1)$. For $N-2k+1 \le x < N-k$, we have $1 < \frac{N-x-1}{k-1} < 2$, $a = \left\lfloor \frac{N-x-1}{k-1} \right\rfloor = 1$, and thus, $t_F(x) = t_F(1)$.

For $N-k < x \le N$, $t_F(x)$ linearly increases with x and thus $t_F(x) \ge t_F(N-k)$. □

B.1.2. Macro-level measure under ACP policy

PROPOSITION OA.2. **[Macro-level Measure: ACP]** *Given a routing decision of x AB customers, for $k \in \{2,3,...\}$, customers' average overall wait is* $t_A(x) = \begin{cases} \frac{N-1}{2\mu} & \text{if } x=0; \\ \frac{kN^2-(2+k)N+2x}{2Nk\mu} & \text{if } 1 \le x \le N. \end{cases}$ *. Moreover, we have* $\arg\min t_A(x) = 1$.

Proof of Proposition OA.2. Denote $T_s^R(x)$ as the overall wait of route R customers at station s, $R \in \{AB, BA\}, s \in \{A, B\}$. If $x = 0$, all customers are routed in BA order, the priority policy does not affect service orders at stations. Thus, from Proposition OA.1, the customers' average overall wait is $t_A(0) = t_F(0) = \frac{N-1}{2\mu}$.

If $1 \le x \le N-k \Leftrightarrow \frac{1}{\mu} \le \frac{N-x}{k\mu}$, station B cannot finish all BA customers before the first AB customer joins. Station A starts serving the first BA customer at $t = \frac{1}{\mu}$ and later BA customers join station A before it finishes serving earlier arriving BA customers. Thus station A serving all BA customers in $\frac{1}{\mu} < t \le \frac{N-x+1}{\mu}$. At station A, the first AB customer gets served during $0 \le t \le \frac{1}{\mu}$, while the later $x-1$ AB customers are delayed by BA customers until $t = \frac{N-x+1}{\mu}$. Thus, $T_A^{AB}(x) = \int_0^{\frac{x}{\mu}} (\frac{1}{2} + x - 1 - \mu t)dt + (x-1)(\frac{N-x}{\mu}) = \frac{(2N-x)(x-1)}{2\mu}$. At station B, the first AB customer arrives at $t = \frac{1}{\mu}$ and immediately gets served. The later $x-1$ AB customers arrive after $t = \frac{N-x+1}{\mu}$ and get served upon arrival as well, because station B is idle. Therefore, we have $T_B^{AB}(x) = 0$ and $T_B^{BA}(x) = \int_0^{\frac{N-x}{k\mu}} (\frac{1}{2} + N - x - 1 - k\mu t)dt + \frac{N-x-k}{k\mu} = \frac{N^2-(2x-1)N+x^2-x-2k}{2k\mu}$. Regarding BA customers' waits at station A, BA customers join queue A in two intervals $\left[0, \frac{1}{\mu}\right] \cup \left[\frac{1}{\mu} + \frac{1}{k\mu}, \frac{N-x+1}{k\mu}\right]$. They are served during $\frac{1}{\mu} < t \le \frac{N-x+1}{\mu}$. Therefore, BA customers' overall wait at station A is $T_A^{BA}(x) = \int_0^{\frac{N-x}{k\mu}} (-\frac{1}{2} + k\mu t)dt + (N-x)(\frac{N-x+1}{\mu} - \frac{N-x+1}{k\mu}) + \frac{1}{k\mu}\frac{k\mu}{\mu} - \int_0^{\frac{N-x+1}{\mu}} (-\frac{1}{2} + \mu t)dt = \frac{(k-1)N^2+(k+2x-2kx-3)N+(k-1)x^2+(3-k)x+2k}{2k\mu}$. Customers' average overall wait is $t_A(x) = \frac{\sum_{s=A,B} T_s^{AB}(x)+T_s^{BA}(x)}{N} = \frac{kN^2-(2+k)N+2x}{2Nk\mu}$.

If $x > N-k \Leftrightarrow \frac{1}{\mu} > \frac{N-x}{k\mu}$, station B completes all BA customers before station A completes the first AB customer. Thus, the priority policy only affects station A's service order: it prioritizes BA customers over the remaining $x-1$ AB customers. This change in service order reduces BA customers' waits by the same amount as it increases AB customers' waits, resulting in no change in overall wait. From Proposition OA.1, customers' average overall wait is $t_A(x) = \frac{kN^2-(2+k)N+2x}{2Nk\mu}$.

For $x \ge 1$, $t_A(x)$ linearly increases with x and thus is minimized at $x = 1$. And we have $t_A(1) < t_A(0)$. □

B.1.3. Comparison We compare average overall waits, $t_F(x)$ and $t_A(x)$. Set $\Delta t(x) = t_A(x) - t_F(x)$. If $x = 0$ or $N-k < x \le N$, we have $\Delta t(x) = 0$. If $1 \le x \le \frac{N-1}{k}$, we have $\Delta t(x) = \frac{-x(2N-(x+1)(k+1))}{2Nk\mu} \le \frac{-(N-1)(k-1)(N-k-1)}{2Nk^3\mu} < 0$. If $\frac{N-1}{k} < x \le N-k$, we have $\Delta t(x) = -\frac{(N-x)(N-x-1-k)}{2Nk(k-1)\mu} - \frac{\varepsilon}{N}$. If $x \le N-k-1$, clearly $\Delta t(x) \le -\frac{\varepsilon}{N} \le 0$. If $x = N-k$, we have $a = \left\lfloor \frac{N-x-1}{k-1} \right\rfloor = 1$ given $k \ge 2$, and thus, $\varepsilon = \frac{-x^2+(2N-3k+1)x-N^2+(3k-1)N+(2k-2k^2)}{2k\mu(k-1)}$ and $\Delta t(x) = -\frac{N-k-x}{Nk\mu} \le 0$. Therefore, we have $t_A(x) \le t_F(x) \forall x$.

For the optimal $x = 1$ routing decision minimizing average overall wait, we have $t_F(1) - t_A(1) = \frac{k+1}{N^2k\mu}$ and $\frac{t_F(1)-t_A(1)}{t_F(1)} = \frac{2(N-k-1)}{k(N+1)(N-2)}$. Taking derivatives with respect to population size N gives $\frac{\partial(t_F(1)-t_A(1))}{\partial N} = \frac{k+1}{N^2k\mu} > 0$ and $\frac{\partial(\frac{t_F(1)-t_A(1)}{t_F(1)})}{\partial N} = -\frac{2(k+3+N(N-2k-2))}{k(N^2-N-2)^2}$. Given that $N \ge \frac{k^2+5k+2}{2}$, we have $\frac{\partial(\frac{t_F(1)-t_A(1)}{t_F(1)})}{\partial N} < 0$. We have $\lim_{N\to\infty} \frac{t_F(1)-t_A(1)}{t_F(1)} = \lim_{N\to\infty} \frac{2(N-k-1)}{k(N+1)(N-2)} = \lim_{N\to\infty} \frac{\frac{\partial(2(N-k-1))}{\partial N}}{\frac{\partial(k(N+1)(N-2))}{\partial N}} = \lim_{N\to\infty} \frac{2}{(2N-1)k} = 0$. □

B.2. Proof of Proposition 2

To prove Proposition 2, we first derive the micro-level measures under FCFS and ACP policies in Propositions OA.3 and OA.4, respectively, then make comparisons in Section B.2.3.

B.2.1. Micro-level measures under FCFS policy

PROPOSITION OA.3. [**Micro-level Measures: FCFS**] *Given a routing decision of x AB customers,*

(*i*) *average station-level wait at the second station is*

$$t_{F,2}(x)=\begin{cases}\frac{(N-1)(k-1)}{2k\mu} & \text{if } x=0;\\ \frac{-2(k+1)x^2+4Nx+(N^2k-N-Nk-N^2)}{2Nk\mu} & \text{if } 1\le x\le \frac{N-1}{k};\\ \frac{(N-x)((k^2-2)x-(1-N)k^2-(2N+1)k+2N)}{2Nk(k-1)\mu}+\frac{\varepsilon}{N} & \text{if } \frac{N-1}{k}<x\le N-k;\\ \frac{N-x}{2Nk\mu}((k+1)x-(N+k-Nk+1)) & \text{if } N-k<x\le N.\end{cases}.$$

(*ii*) *The probability of station-level excessive waits, $p_F(x,L)$, is nondecreasing in x for $x<\frac{X_1}{k}$ or $x\ge X_1$, and nonincreasing in x for $\frac{X_1}{k}\le x<X_1$, where $X_1\equiv N-k\mu L-1$.*

Proof of Proposition OA.3. To prove part (i), let $t_{F,1}(x)$ and $t_{F,2}(x)$ denote customers' average wait at their first and second stations, respectively. Part (i) follows Proposition OA.1. We have $t_{F,1}(x)=\frac{\int_0^{\frac{x}{\mu}}(\frac{1}{2}+x-1-\mu t)dt+\int_0^{\frac{N-x}{k\mu}}(\frac{1}{2}+N-x-1-k\mu t)dt}{N}=\frac{(k+1)x^2+(1-k-2N)x+N(N-1)}{2Nk\mu}$ and $t_{F,2}(x)=t_F(x)-t_{F,1}(x)$.

To prove part (ii), let $t_s^R(i)$ denote the waits of route R customers at station s, $R\in\{AB,BA\}, s\in\{A,B\}$. Let n_s^R count the waits exceeding a threshold value L at station s for route R customers at station s, so that $n_s=n_s^{AB}+n_s^{BA}$ and $p_F(x,L)=\frac{n_A+n_B}{2N}$. Due to integrality, we will approximate the counts as $\widetilde{n}_s^R$ and $\widetilde{n}_s$, and the probability as $\widetilde{p}_F(x,L)=\frac{\widetilde{n}_s^{AB}+\widetilde{n}_s^{BA}}{2N}$. At the end of the proof, we show that the accurate probability $p_F(x,L)$ behaves qualitatively as the approximation $\widetilde{p}_F(x,L)$.

If $x=0$, all customers are routed BA. We have $t_B^{BA}(j)=\frac{j-1}{k\mu}$ and $t_A^{BA}(j)=\frac{1}{k\mu}+\frac{j-1}{\mu}-\frac{j}{k\mu}$.

If $x\ge 1$, we have $t_A^{AB}(i)=\frac{i-1}{\mu}$, $t_A^{BA}(j)=\frac{x+j-1}{\mu}-\frac{j}{k\mu}$, $t_B^{AB}(i)=\frac{N-x+i-1}{k\mu}-\frac{i}{\mu}$, and $t_B^{BA}(j)=\frac{j-1}{k\mu}$. Thus the approximated counts of long waits are $\widetilde{n}_A=\widetilde{n}_A^{AB}+\widetilde{n}_A^{BA}=\begin{cases}N-L\mu-1 & \text{if } L<\frac{x-1}{\mu};\\ N-x-\frac{k\mu L-k(x-1)}{k-1} & \text{if } \frac{x-1}{\mu}\le L<\frac{(k-1)N-k+x}{k\mu};\\ 0 & \text{if } L\ge\frac{(k-1)N-k+x}{k\mu}.\end{cases}$

and $\widetilde{n}_B=\widetilde{n}_B^{AB}+\widetilde{n}_B^{BA}=\begin{cases}N-Lk\mu-1 & \text{if } L<\frac{N-kx-1}{k\mu};\\ \frac{k(N-x-Lk\mu-1)}{k-1} & \text{if } \frac{N-kx-1}{k\mu}\le L<\frac{N-x-1}{k\mu};\\ 0 & \text{if } L\ge\frac{N-x-1}{k\mu}.\end{cases}$

Note that compared to $x=0$, the $x=1$ routing decision does not affect delays at station A and reduces the last customer's wait at station B from $\frac{N-1}{k\mu}$ to $\frac{N-k-1}{k\mu}$, which implies that $p_F(0,L)\ge p_F(1,L)$.

Next, we discuss how the probability of excessive waits $\widetilde{p}_F(x,L)=\frac{\widetilde{n}_A+\widetilde{n}_B}{2N}$ changes with the routing decision x. This change depends on the relationship between the threshold value L and four functions that determine $\widetilde{n}_A$ and $\widetilde{n}_B$: $\frac{x-1}{\mu}$, $\frac{(k-1)N-k+x}{k\mu}$, $\frac{N-kx-1}{k\mu}$, and $\frac{N-x-1}{k\mu}$. Figure OA.3 depicts these functions, dividing the graph into regions, which can be at most nine (see Figure OA.3(a)), and sometimes fewer (see Figure OA.3(b)).

In Table OA.2, we summarize the regions, compute the corresponding probability $\widetilde{p}_F(x,L)$, and establish monotonicity concerning the routing decision x. For $\frac{N-kx-1}{k\mu}<L<\frac{N-x-1}{k\mu}$ (regions ②⑤⑧, blue shaded area in Figure OA.3), the probability $\widetilde{p}_F(x,L)$ increases with the routing decision x. For $\frac{x-1}{\mu}<L\le\min(\frac{N-kx-1}{k\mu},\frac{(k-1)N-k+x}{k\mu})$ or $\max(\frac{x-1}{\mu},\frac{N-x-1}{k\mu})\le L\le\frac{(k-1)N-k+x}{k\mu}$ (regions ④⑥, green shaded area in Figure OA.3), the probability $\widetilde{p}_F(x,L)$ decreases with the routing decision x. Otherwise, the probability $\widetilde{p}_F(x,L)$ is independent of the routing decision x. For analytical simplicity, set $X_1:=N-k\mu L-1$. We have $\widetilde{p}_F(x,L)$ is nondecreasing in x for $x<\frac{X_1}{k}$ and $x\ge X_1$, and nonincreasing in x for $\frac{X_1}{k}\le x<X_1$.

Figure OA.3 Probability of Station-Level Excessive Waits $\widetilde{p}_F(x,L)$ under FCFS Policy ($N=50, \mu=1$)

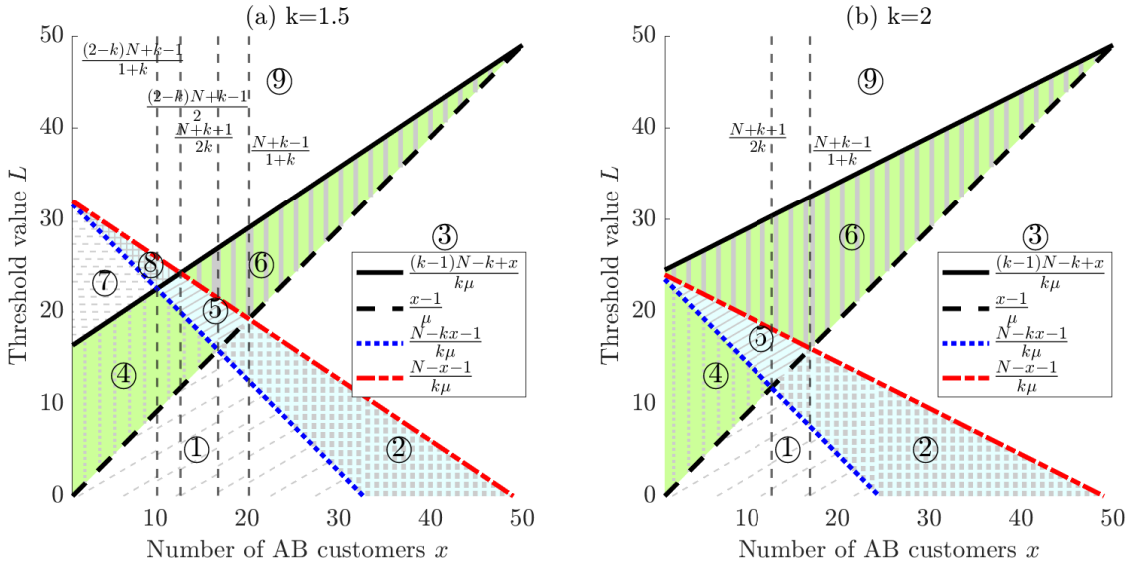


Table OA.2 Relationship Between Threshold L and Probability of Station-Level Excessive Waits $\widetilde{p}_F(x,L)$

Region number and definition	$\widetilde{p}_F(x,L)$	Monotonicity
(1) $L < \min(\frac{x-1}{\mu}, \frac{(k-1)N-k+x}{k\mu}, \frac{N-kx-1}{k\mu}, \frac{N-x-1}{k\mu})$	$1-\frac{(1+k)\mu L+2}{2N}$	Independent of x
(2) $\frac{N-kx-1}{k\mu} \le L < \frac{N-x-1}{k\mu} \le \frac{x-1}{\mu} \le \frac{(k-1)N-k+x}{k\mu}$	$1-\frac{(k^2+k-1)\mu L+2k-N+kx-1}{2N(k-1)}$	Decreases with x
(3) $\frac{N-kx-1}{k\mu} < \frac{N-x-1}{k\mu} \le L < \frac{x-1}{\mu} \le \frac{(k-1)N-k+x}{k\mu}$	$\frac{N-L\mu-1}{2N}$	Independent of x
(4) $\frac{x-1}{\mu} \le L < \min(\frac{(k-1)N-k+x}{k\mu}, \frac{N-kx-1}{k\mu}) \le \frac{N-x-1}{k\mu}$	$1-\frac{L\mu k^2+2k-x-1}{2N(k-1)}$	Increases with x
(5) $\max(\frac{x-1}{\mu}, \frac{N-kx-1}{k\mu}) \le L < \min(\frac{(k-1)N-k+x}{k\mu}, \frac{N-x-1}{k\mu})$	$1-\frac{(k-1)x+2k+Lk\mu(k+1)-N}{2N(k-1)}$	Decreases with x
(6) $\frac{N-kx-1}{k\mu} \le \max(\frac{N-x-1}{k\mu}, \frac{x-1}{\mu}) \le L < \frac{(k-1)N-k+x}{k\mu}$	$\frac{(k-1)N+x-k-Lk\mu}{2N(k-1)}$	Increases with x
(7) $\frac{x-1}{\mu} \le \frac{(k-1)N-k+x}{k\mu} \le L < \frac{N-kx-1}{k\mu} \le \frac{N-x-1}{k\mu}$	$1-\frac{N+Lk\mu+1}{2N}$	Independent of x
(8) $\frac{x-1}{\mu} \le \max(\frac{(k-1)N-k+x}{k\mu}, \frac{N-kx-1}{k\mu}) \le L < \frac{N-x-1}{k\mu}$	$\frac{k(N-x-Lk\mu-1)}{2N(k-1)}$	Decreases with x
(9) $\max(\frac{x-1}{\mu}, \frac{(k-1)N-k+x}{k\mu}, \frac{N-kx-1}{k\mu}, \frac{N-x-1}{k\mu}) \le L$	0	Independent of x

In our derivation, we need to consider the effects of errors resulting from the approximation process. We denote these errors as ε_1, ε_2, ε_3, and ε_4: (i) ε_1 arises from counting ith AB customer's wait at station A as exceeding L (for $L < \frac{x-1}{\mu}$) if $i > \mu L+1$ instead of $i \ge \lfloor \mu L+1 \rfloor$, leading to $\varepsilon_1 = n_A^{AB} - \widetilde{n}_A^{AB} = (x - \lfloor \mu L+1 \rfloor) - (x-(\mu L+1)) = \mu L - \lfloor \mu L \rfloor$, independent of x. (ii) ε_2 occurs when counting jth BA customer's wait at station A as exceeding L (for $\frac{x-1}{\mu} \le L < \frac{(k-1)N-k+x}{k\mu}$) if $j > \frac{k\mu L-k(x-1)}{k-1}$ instead of $j > \left\lfloor \frac{k\mu L-k(x-1)}{k-1} \right\rfloor$, leading to $\varepsilon_2 = n_A^{BA} - \widetilde{n}_A^{BA} = (N-x-\left\lfloor \frac{k\mu L-k(x-1)}{k-1} \right\rfloor) - (N-x-\frac{k\mu L-k(x-1)}{k-1}) = \frac{k\mu L-k(x-1)}{k-1} - \left\lfloor \frac{k\mu L-k(x-1)}{k-1} \right\rfloor$. (iii) ε_3 arises from counting BA customers' waits at station B (for $L < \frac{N-x-1}{k\mu}$): $\varepsilon_3 = n_B^{BA} - \widetilde{n}_B^{BA} = (N-x-\lfloor k\mu L+1 \rfloor) - (N-x-k\mu L-1) = k\mu L - \lfloor k\mu L \rfloor$, independent of x. (iv) ε_4 arises from counting AB customers' waits at station B (for $\frac{N-kx-1}{k\mu} \le L < \frac{N-x-1}{k\mu}$): $\varepsilon_4 = n_B^{AB} - \widetilde{n}_B^{AB} = \left\lfloor \frac{N-x-1-k\mu L}{k-1} \right\rfloor - \frac{N-x-1-k\mu L}{k-1}$. Note that if ε_4 exists, ε_3 must exist as well. The accurate count of excessive waits at station B for $\frac{N-kx-1}{k\mu} \le L < \frac{N-x-1}{k\mu}$ is

$n_B = N - x - \lfloor k\mu L + 1 \rfloor + \left\lfloor \frac{N-x-1-k\mu L}{k-1} \right\rfloor$, the approximation is $\widetilde{n}_B = N - x - k\mu L - 1 + \frac{N-x-1-k\mu L}{k-1}$. Both are nonincreasing in x. To conclude, ε_1 and ε_3 are independent of x; they clearly do not affect the monotonicity property derived above. If ε_2 exists alone or exists with ε_1 or ε_3, the property remains unaffected. If ε_4 exists, we have shown that ε_3 exists as well, and their combined effect does not affect the property. Last, if ε_2, ε_3, ε_4 exist, the approximated count $\widetilde{n}(x) = \frac{-(k-1)x+(2k-1)N-(k+1)\mu kL-2k}{k-1}$ decreases with x; the accurate count is $n(x) = 2N - 2x - \left\lfloor \frac{k\mu L-k(x-1)}{k-1} \right\rfloor - \lfloor k\mu L + 1 \rfloor + \left\lfloor \frac{N-x-1-k\mu L}{k-1} \right\rfloor$, and we have $n(x+1) \le n(x)$. Again, the error does not affect the monotonicity property. Therefore, given the threshold value L, the approximation error is $\frac{\sum_{i=1}^{4}\varepsilon_i}{2N}$ and does not affect the relationship between $p_F(x, L)$ and x derived above. □

B.2.2. Micro-level measures under ACP policy

PROPOSITION OA.4. **[Micro-level Measures: ACP]** *Given a routing decision of x AB customers, for $k \in \{2, 3, ...\}$,*

(i) average station-level wait at the second station is

$$t_{A,2}(x) = \begin{cases} \frac{(N-1)(k-1)}{2k\mu} & \text{if } x = 0; \\ \frac{(k-1)x^2+(2N-k-2Nk+3)x+(2k-3N+N^2k+Nk-N^2)}{2Nk\mu} & \text{if } 1 \le x \le N-k; \\ \frac{(k-1)(N^2-2Nx+N+x^2-x)}{2Nk\mu} & \text{if } N-k+1 \le X \le N. \end{cases}$$

(ii) the probability of station-level excessive waits, $p_A(x, L)$, has the following properties:

- *For $L \le \frac{N-1}{k\mu}$, $p_A(x, L)$ is weakly decreasing in x for $x \le X_1 + 1$ and weakly increasing in x for $x > X_1 + 1$.*
- *For $L > \frac{N-1}{k\mu}$, $p_A(0, L) \ge p_A(1, L)$, and $p_A(x, L)$ is weakly increasing in x for $x \ge 1$.*

Proof of Proposition OA.4. Part (i) follows Proposition OA.2. We have $t_{A,1}(x) = \frac{T_A^{AB}(x)+T_B^{BA}(x)}{N}$ and $t_{A,2}(x) = t_A(x) - t_{A,1}(x)$.

We next prove part (ii). For $x = \{0, N\}$, the priority policy does not affect stations' service orders; the results follow Proposition OA.3(ii). Our focus here is on $1 \le x \le N-1$ and $\frac{k+1}{k\mu} \le L < \frac{N-1}{\mu}$. Following a similar approach as Proposition OA.3(ii), we use the approximated counts to establish properties for the approximated probability and then note that the accurate probability exhibits similar qualitative behavior.

Under ACP, the ith AB customer's station-level waits are $t_A^{AB}(i) = \begin{cases} 0 & \text{if } i = 1; \\ \frac{i-1+N-x}{\mu} & \text{if } 2 \le i \le x; \end{cases}$ and $t_B^{AB}(i) = 0$. Hence, the approximated counts of excessive wait at station A and B are $\widetilde{n}_A^{AB}(x = 1) = 0$ and $\widetilde{n}_A^{AB}(x \ge 2) = \begin{cases} x - 1 & \text{if } L < \frac{N-x}{\mu}; \\ N - \mu L - 1 & \text{if } L \ge \frac{N-x}{\mu}; \end{cases}$ and $n_B^{AB} = \widetilde{n}_B^{AB} = 0$.

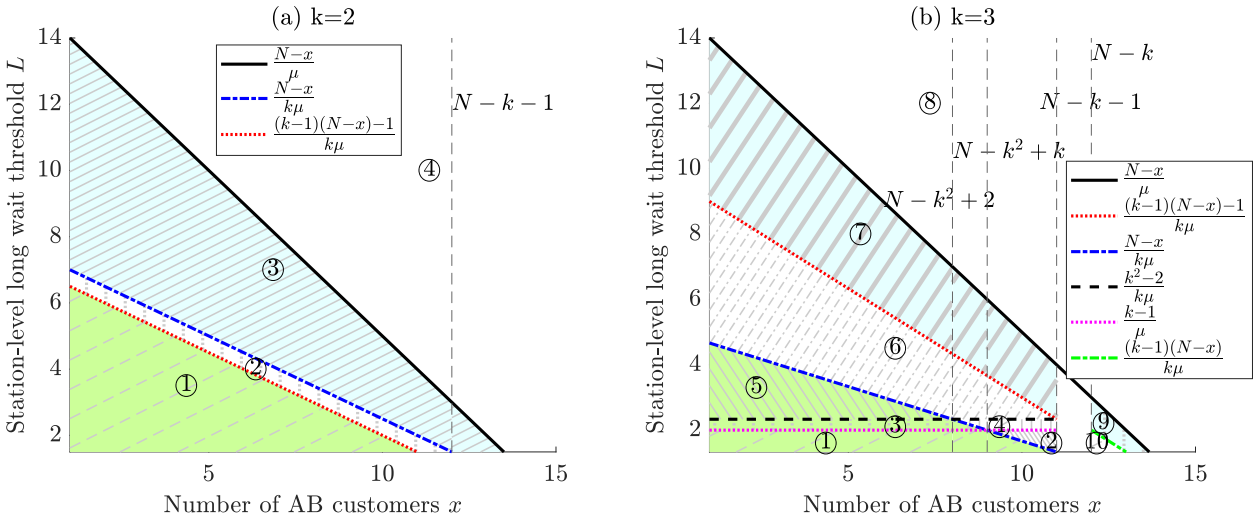
The jth BA customer's station-level waits are $t_A^{BA}(j) = \begin{cases} \frac{(k-1)j}{k\mu} & \text{if } j \le k; \\ \frac{(k-1)j-1}{k\mu} & \text{if } k+1 \le j \le N-x; \end{cases}$ and $t_B^{BA}(j) = \begin{cases} \frac{j-1}{k\mu} & \text{if } j \le k; \\ \frac{j}{k\mu} & \text{if } k+1 \le j \le N-x. \end{cases}$ If $x \le N-k-1$, the approximated counts of BA customers' excessive wait at stations are $\widetilde{n}_A^{BA} = \begin{cases} N - x - \frac{k\mu L}{k-1} & \text{if } L < \frac{k-1}{\mu}; \\ N - x - k & \text{if } \frac{k-1}{\mu} \le L < \frac{k^2-2}{k\mu}; \\ N - x - \frac{k\mu L+1}{k-1} & \text{if } \frac{k^2-2}{k\mu} \le L < \frac{(k-1)(N-x)-1}{k\mu}; \\ 0 & \text{if } L \ge \frac{(k-1)(N-x)-1}{k\mu}. \end{cases}$ and $\widetilde{n}_B^{BA} = \begin{cases} N - x - k\mu L & \text{if } L < \frac{N-x}{k\mu}; \\ 0 & \text{if } L \ge \frac{N-x}{k\mu}. \end{cases}$

Otherwise if $x \ge N-k$, all BA customers complete service B before the first AB customer joins, and thus, $\widetilde{n}_A^{BA} = \begin{cases} N - x - \frac{k\mu L}{k-1} & \text{if } L < \frac{(k-1)(N-x)}{k\mu}; \\ 0 & \text{if } L \ge \frac{(k-1)(N-x)}{k\mu}. \end{cases}$ Note that for $k = 2$, we have $t_A^{BA}(k) = t_A^{BA}(k+1)$. If $N - x > k + 1$,

we have $\widetilde{n}_A^{BA} = \begin{cases} N-x-\frac{k\mu L+1}{k-1} & \text{if } L < \frac{(k-1)(N-x)-1}{k\mu}; \\ 0 & \text{if } L \geq \frac{(k-1)(N-x)-1}{k\mu}. \end{cases}$ If $N-x \leq k+1$, we have $\widetilde{n}_A^{BA}=0$ given $L \geq \frac{k+1}{k\mu}$. And we have $\max t_B^{BA}(j) = t_B^{BA}(N-x) = \frac{N-x-1}{k\mu} \leq \frac{k-1}{k\mu}$ so $\widetilde{n}_B^{BA}=0$ given $L \geq \frac{k+1}{k\mu}$.

We next see how the probability $\widetilde{p}_A(x,L) = \frac{n_A+n_B}{2N}$ changes with respect to the routing decision x. First, compared to $x=0$ (derived in Appendix OA.3), the $x=1$ decision does not affect delays at station B and reduces BA customers' waits at station A by $\frac{1}{k\mu}$ for $k+1 \leq j \leq N-1$, which implies $p_A(1,L) \leq p_A(0,L)$. For $1 \leq x \leq N$, the change depends on the relationship between the threshold value L and six functions that determine $\widetilde{n}_s$: $\frac{N-x}{\mu}$, $\frac{k-1}{\mu}$, $\frac{k^2-2}{k\mu}$, $\frac{(k-1)(N-x)-1}{k\mu}$, $\frac{(k-1)(N-x)}{k\mu}$, and $\frac{N-x}{k\mu}$. For $k=2$, we have $\frac{k-1}{\mu} < \frac{k+1}{k\mu}$ and $\frac{(k-1)(N-x)-1}{k\mu} < \frac{N-x}{k\mu} = \frac{(k-1)(N-x)}{k\mu} < \frac{N-x}{\mu}$. For $k \geq 3$, we have $\frac{k+1}{k\mu} < \frac{k-1}{\mu} < \frac{k^2-2}{k\mu}$ and $\frac{N-x}{k\mu} < \frac{(k-1)(N-x)-1}{k\mu}$. We plot these six functions with respect to the routing decision x in Figure OA.4. These six functions partition the graph into four regions when $k=2$ (see Figure OA.4(a)) or ten regions when $k \geq 3$ (see Figure OA.4(b)).

Figure OA.4 Probability of Station-Level Excessive Waits $\widetilde{p}_A(x,L)$ under ACP Policy ($N=15, \mu=1$)



We list the regions, calculate the corresponding probability $\widetilde{p}_A(x,L)$, and derive monotonicity results with respect to x in Table OA.3. If $k=2$ (see Figure OA.4(a)), for $L < \frac{(k-1)(N-x)-1}{k\mu}$ (region ①), the probability $\widetilde{p}_A(x,L)$ decreases with x; for $\frac{N-x}{k\mu} \leq L < \frac{N-x}{\mu}$ (region ③), the probability $\widetilde{p}_A(x,L)$ increases with x; otherwise, the probability $\widetilde{p}_A(x,L)$ is independent of the routing decision x. If $k \geq 3$ (see Figure OA.4(b)), for $x \leq N-k-1$ and $L > \frac{(k-1)(N-x)-1}{k\mu}$ and for $x \geq N-k$ and $L > \frac{(k-1)(N-x)}{k\mu}$ (blue shaded area, regions ⑦⑨), $\widetilde{p}_A(x,L)$ increases with x; for $L < \frac{N-x}{k\mu}$ (green shaded area, regions ①③⑤), $\widetilde{p}_A(x,L)$ decreases with x. Otherwise, $\widetilde{p}_A(x,L)$ is independent of the routing decision x. Hence, we have $\widetilde{p}_A(x,L)$ is nonincreasing in x for $x < X_1+1$ (i.e., $L < \frac{N-x}{k\mu}$), and nondecreasing in x for $X_1+1 \leq x \leq N-1$.

Last, we compare $p_A(N-1,L)$ and $p_A(N,L)$. When $x=N-1$, the later $N-2$ AB customers and the only BA customer wait at station A; i.e., $t_A^{AB}(i \geq 2) = \frac{i}{\mu}$ and $t_A^{BA}(1) = \frac{1}{2\mu}$. When $x=N$, AB customers' waits are $t_A^{AB}(i \geq 2) = \frac{i-1}{\mu}$. We have $p_A(N-1,L) = p_A(N,L)$ given $L \geq \frac{k+1}{k\mu}$. Therefore, $p_A(x)$ is nonincreasing in x for $x < X_1+1$ and nondecreasing in x for $X_1+1 \leq x$.

Table OA.3 Relationship between Threshold L and Probability of Station-Level Excessive Waits $\widetilde{p}_A(x,L)$

$k=2$				$k\geq 3$		
Region no. and definition	$\widetilde{p}_A(x,L)$	Monotonicity		Region no. and definition	$\widetilde{p}_A(x,L)$	Monotonicity
(A1) $L<\frac{(k-1)(N-x)-1}{k\mu}$	$1-\frac{x+L\mu k^2+k}{2N}$	Decreases with x		(B1) $L<\min(\frac{k-1}{\mu},\frac{N-x}{\mu})$	$\frac{(2N-x)(k-1)-L\mu k^2-k+1}{2N(k-1)}$	Decreases with x
(A2) $\frac{(k-1)(N-x)-1}{k\mu}\leq L<\frac{N-x}{k\mu}$	$\frac{N-Lk\mu-1}{2N}$	Independent of x		(B2) $\frac{N-x}{k\mu}\leq L<\frac{k-1}{\mu}$	$\frac{(k-1)(N-1)-Lk\mu}{2N(k-1)}$	Independent of x
(A3) $\frac{N-x}{k\mu}\leq L<\frac{N-x}{\mu}$	$\frac{x-1}{2N}$	Increases with x		(B3) $\frac{k-1}{\mu}\leq L<\min(\frac{k^2-2}{k\mu},\frac{N-x}{k\mu})$	$\frac{2N-x-Lk\mu-k-1}{2N}$	Decreases with x
(A4) $\frac{N-x}{\mu}\leq L$	$\frac{N-\mu L-1}{2N}$	Independent of x		(B4) $\max(\frac{k-1}{\mu},\frac{N-x}{k\mu})\leq L<\frac{k^2-2}{k\mu}$	$\frac{N-k-1}{2N}$	Independent of x
				(B5) $\frac{k^2-2}{k\mu}\leq L<\frac{N-x}{k\mu}$	$\frac{(2N-x)(k-1)-L\mu k^2-k}{2N(k-1)}$	Decreases with x
				(B6) $\max(\frac{N-x}{\mu},\frac{k^2-2}{k\mu})\leq L<\frac{(k-1)(N-x)}{k\mu}$	$\frac{(k-1)N-Lk\mu-k}{2N(k-1)}$	Independent of x
				(B7) $\frac{(k-1)(N-x)-1}{k\mu}\leq L<\frac{N-x}{\mu}$	$\frac{x-1}{2N}$	Increases with x
				(B8) $\frac{N-x}{\mu}\leq L$	$\frac{N-\mu L-1}{2N}$	Independent of x
				(B9) $\frac{(k-1)(N-x)}{k\mu}\leq L<\frac{N-x}{\mu}$	$\frac{x-1}{2N}$	Increases with x
				(B10) $L<\frac{(k-1)(N-x)}{k\mu}$	$\frac{(k-1)N-Lk\mu-k+1}{2N(k-1)}$	Independent of x

Therefore, for $L\leq\frac{N-1}{k\mu}$, we have $X_1+1\geq 1$ so that $\widetilde{p}_A(x,L)$ is nonincreasing in x for $x\leq X_1+1$ and nondecreasing in x for $x>X_1+1$. For $L>\frac{N-1}{k\mu}$, we have $X_1+1<1$, $\widetilde{p}_A(0,L)\geq p_A(1,L)$, and $\widetilde{p}_A(x,L)$ is nondecreasing in x for $x\geq 1$. Similar to Appendix OA.3, this derivation ignores integrality. Note that the error is independent of the routing decision x and thus does not qualitatively affect the results derived above.

B.2.3. Comparison First, basic algebra operation gives part (i): $t_{F,2}(x)\geq t_{A,2}(x)\forall x$.

We next prove part (ii). Set $\Delta p(x,L)=\widetilde{p}_A(x,L)-\widetilde{p}_F(x,L)$. Given $k\geq 2$, we have $\frac{N-kx-1}{k\mu}<\frac{N-x-1}{k\mu}<\frac{(k-1)N-k+x}{k\mu}$ for $x\geq 1$, which rules out Table OA.2 (7)–(8). Following Proposition OA.4(ii), we limit the discussion to $\frac{k+1}{k\mu}<L<\frac{N-1}{\mu}$ and assume the population size is large; i.e., $N\geq k^2+2k-3$.

Case 1: For $L\leq\frac{N-1}{2\mu}$, from Propositions OA.3(ii) and OA.4(ii), we know that $\widetilde{p}_F(x,L)$ is nondecreasing in x for $x<\frac{X_1}{k}$ or $x\geq X_1$ and nonincreasing in x for $\frac{X_1}{k}\leq x<X_1$, and $\widetilde{p}_A(x,L)$ is nonincreasing in x for $x\leq X_1+1$ and nondecreasing in x for $x>X_1+1$.

1.1 For $x\leq\frac{X_1}{k}$ (possible for $L\leq\frac{N-k-1}{k\mu}$, which implies $1\leq\frac{X_1}{k}$), we know that $\widetilde{p}_F(x,L)$ is nondecreasing in x and $\widetilde{p}_A(x,L)$ is nonincreasing in x. At $x=1$, we have $\frac{x-1}{\mu}\leq L<\min(\frac{(k-1)N-k+x}{k\mu},\frac{N-kx-1}{k\mu})$, and thus, $\widetilde{p}_F(1,L)=1-\frac{L\mu k^2+2k-2}{2N(k-1)}$ from Table OA.2 (4). For $k=2$, given $L\leq\frac{N-1}{2\mu}$ and $x=1$, we have $L\leq\frac{N-1}{2\mu}<\frac{(k-1)(N-x)-1}{\mu}$, $\widetilde{p}_A(1,L)=1-\frac{x+L\mu k^2+k}{2N}$ from Table OA.3 (A1), and thus, $\Delta p(1,L)=-\frac{1}{2N}<0$, which implies $\Delta p(x,L)\leq 0$ for $x\leq\frac{X_1}{k}$. For $k\geq 3$, we have $\frac{k-1}{\mu}<\frac{k^2-2}{k\mu}$. At $x=1$, we have $x=1<N-k^2+2$ and $L<\frac{N-x}{k\mu}<\frac{(k-1)(N-x)-1}{k\mu}$; $\widetilde{p}_A(1,L)$ depends on the threshold value L as discussed below.

1.1.1 For $L\leq\frac{k-1}{\mu}$, we have $\widetilde{p}_A(1,L)=1-\frac{L\mu k^2+2k-2}{2N(k-1)}$ from Table OA.3 (B1), implying that $\Delta p(1,L)=0$ and $\Delta p(x,L)\leq 0$ for $1\leq x\leq\frac{X_1}{k}$ as $\widetilde{p}_F(x,L)$ is nondecreasing and $\widetilde{p}_A(x,L)$ is nonincreasing in x.

1.1.2 For $\frac{k-1}{\mu}<L\leq\frac{k^2-2}{k\mu}$, we have $\widetilde{p}_A(1,L)=\frac{2N-x-L\mu k-k-1}{2N}$ from Table OA.3 (B3), and thus, $\Delta p(1,L)=\frac{\mu k(L-\frac{k-1}{\mu})}{2(k-1)N}>0$. At $x=\frac{X_1}{k}<N-k^2+2$, we have $\widetilde{p}_F(\frac{X_1}{k},L)=1-\frac{L\mu k^2+2k-\frac{X_1}{k}-1}{2N(k-1)}$ from Table OA.2 (4), $\widetilde{p}_A(\frac{X_1}{k},L)=\frac{2N-\frac{X_1}{k}-Lk\mu-k-1}{2N}$ from Table OA.3 (B3), and thus, $\Delta p(\frac{X_1}{k},L)=\frac{k\mu(L-\frac{N+k^2-2k-1}{2k\mu})}{N(k-1)}<0$. So there exists a point $\widetilde{X}_0\in\left(1,\frac{X_1}{k}\right]$ such that $\Delta p(x,L)\leq 0$ for $x\geq\widetilde{X}_0$.

1.1.3 For $\frac{k^2-2}{k\mu}<L\leq\frac{N-k-1}{k\mu}$, we have $\widetilde{p}_A(1,L)=\frac{(2N-1)(k-1)-L\mu k^2-k}{2N(k-1)}$ from Table OA.3 (B5), and thus, $\Delta p(1,L)=\frac{-1}{2N(k-1)}<0$, which implies $\Delta p(x,L)<0$ for $1\leq x\leq\frac{X_1}{k}$.

1.2 For $\frac{X_1}{k} < x \le X_1$ (possible for $L \le \frac{N-2}{k\mu}$, which implies $1 \le X_1$), we know that $\widetilde{p}_F(x,L)$ and $\widetilde{p}_A(x,L)$ are nonincreasing in x. From Table OA.2 (3) and (6), we have $\widetilde{p}_F(X_1,L) = \begin{cases} \frac{N-L\mu-1}{2N} & \text{for } L \le \frac{N-2}{(1+k)\mu}; \\ \frac{(k-1)N+x-k-Lk\mu}{2N(k-1)} & \text{for } L > \frac{N-2}{(1+k)\mu}. \end{cases}$ For $k=2$, we have $\widetilde{p}_A(X_1,L) = \frac{N-Lk\mu-1}{2N}$ from Table OA.3 (A2), so $\Delta p(X_1,L) = \begin{cases} \frac{-\mu(k-1)L}{2N} & \text{for } L \le \frac{N-2}{(1+k)\mu}; \\ \frac{\mu(L-\frac{N-2}{2\mu})}{N} & \text{for } L > \frac{N-2}{(1+k)\mu}, \end{cases}$ which is non-positive. For $k \ge 3$, we have $\Delta p(\frac{X_1}{k},L) < 0$ from Case 1.1, $\widetilde{p}_A(X_1,L)$ depends on the threshold value L as discussed below.

1.2.1 For $L \le \frac{k-1}{\mu}$, we have $L < \frac{N-x}{\mu}$, $N-k^2+2 < X_1 \le N-k-1$, and $\widetilde{p}_A(X_1,L) = \frac{(2N-x)(k-1)-L\mu k^2-k+1}{2N(k-1)}$ from Table OA.3 (B1), which implies that $\Delta p(X_1,L) = \frac{\mu(\frac{k-1}{\mu}-L)}{2N(k-1)} \ge 0$. So there exists a point $\widetilde{X}_1 \in \left[\frac{X_1}{k}, X_1\right]$ such that $\Delta p(x,L) \le 0$ for $x \le \widetilde{X}_1$ and $\Delta p(x,L) > 0$ for $x > \widetilde{X}_1$.

1.2.2 For $\frac{k-1}{\mu} < L \le \frac{k^2-2}{k\mu}$, we have $X_1 \le N-k^2+k$ and $\widetilde{p}_A(X_1,L) = \frac{2N-x-L\mu k-k-1}{2N}$ from Table OA.3 (B3), which implies that $\Delta p(X_1,L) = \frac{\mu(L-\frac{k-1}{\mu})}{2N} > 0$. So there exists a point $\widetilde{X}_1 \in \left[\frac{X_1}{k}, X_1\right)$ such that $\Delta p(x,L) \le 0$ for $x \le \widetilde{X}_1$ and $\Delta p(x,L) > 0$ for $x > \widetilde{X}_1$.

1.2.3 For $\frac{k^2-2}{k\mu} < L \le \frac{N-2}{(1+k)\mu}$, we have $L < \frac{N-x}{k\mu}$, $X_1 \le N-k^2+2$, and $\widetilde{p}_A(X_1,L) = \frac{(2N-x)(k-1)-L\mu k^2-k}{2N(k-1)}$ from Table OA.3 (B5), which implies that $\Delta p(X_1,L) = \frac{-(L\mu+1)}{2(k-1)N} < 0$. So we have $\Delta p(x,L) \le 0$ on $\frac{X_1}{k} < x \le X_1$.

1.2.4 For $\frac{N-2}{(1+k)\mu} < L \le \frac{N-k-1}{k\mu}$, we have $\Delta p(X_1,L) = \frac{2N-x-Lk\mu-k-1}{2N} - \frac{(k-1)N+x-k-Lk\mu}{2N(k-1)} = \frac{k\mu(L-\frac{N-k}{k\mu})}{2N(k-1)} < 0$, which implies that $\Delta p(x,L) \le 0$ for $\frac{X_1}{k} \le x < X_1$.

1.2.5 For $\frac{N-k-1}{k\mu} < L \le \frac{N-2}{k\mu}$, we have $\frac{X_1}{k} \le 1$ and $\Delta p(x,L) = \frac{k\mu(L-\frac{N-k}{k\mu})}{2N(k-1)}$, which implies that $\Delta p(x,L) \le 0$ for $L \le \frac{N-k}{k\mu}$ and $\Delta p(x,L) > 0$ for $L > \frac{N-k}{k\mu}$.

1.3 For $x > X_1$, we know that $p_F(x,L)$ is nondecreasing in x; $p_A(x,L)$ is nonincreasing in x for $x \le X_1+1$, and nondecreasing in x for $x > X_1+1$. Furthermore, given $L \le \frac{N-1}{2\mu}$, we have $\mu L+1 \le N-\mu L$.

1.3.1 For $L \le \frac{N-2}{(1+k)\mu}$, we have $\mu L+1 \le X_1$, $\frac{N-x-1}{k\mu} < L < \frac{x-1}{\mu}$, and thus, $\widetilde{p}_F(x,L) = \frac{N-L\mu-1}{2N}$ from Table OA.2 (3). And we know that $\widetilde{p}_A(x,L)$ increases to $\frac{N-L\mu-1}{2N}$ at $x = N-\mu L > X_1$. For $k \ge 3$ and $L < \frac{k^2-2}{k\mu}$, from Cases 1.2.1–1.2.2, we have $\Delta p(X_1,L) > 0$ so there exists a point $\widetilde{X}_2 \in (X_1, N-\mu L]$ such that $\Delta p(x,L) = 0$ for $x \ge \widetilde{X}_2$. For $k=2$ or $k \ge 3$ and $L \ge \frac{k^2-2}{k\mu}$, from Case 1.2.3, we have $\Delta p(X_1,L) \le 0$, and thus, $\Delta p(x,L) \le 0$ for $x > X_1$.

1.3.2 For $\frac{N-2}{(1+k)\mu} < L \le \frac{N-2}{k\mu}$, we have $L < \frac{(k-1)N-k+x}{k\mu}$, $X_1 < \mu L+1$, and $\widetilde{p}_F(x,L) = \begin{cases} \frac{N-1-L\mu}{2N} & \text{if } L \le \frac{x-1}{\mu}; \\ \frac{(k-1)N+x-k-Lk\mu}{2N(k-1)} & \text{if } L > \frac{x-1}{\mu}, \end{cases}$ from Table OA.2 (3) and (6). And we know that $\widetilde{p}_A(x,L)$ increases to $\frac{N-L\mu-1}{2N}$ at $x = N-\mu L > \mu L+1$. For $k=2$ or for $k \ge 3$ and $\frac{N-2}{(1+k)\mu} < L \le \frac{N-k}{k\mu}$, from Cases 1.2.4–1.2.5, we have $\Delta p(X_1,L) \le 0$, and thus, $\Delta p(x,L) \le 0$ for $x > X_1$. For $k \ge 3$ and $\frac{N-k}{k\mu} < L \le \frac{N-2}{k\mu}$, from Case 1.2.5, we have $\Delta p(X_1,L) > 0$. Hence, there exists a point $\widetilde{X}_2 \in (X_1, N-\mu L]$ such that $\Delta p(x,L) \le 0$ for $x \ge \widetilde{X}_2$.

1.3.3 For $L > \frac{N-2}{k\mu}$ (possible for $k \ge 3$, which implies $X_1 \le 1$), we have $\widetilde{p}_F(1,L) = \frac{(k-1)N+1-k-Lk\mu}{2N(k-1)}$, $\widetilde{p}_A(1,L) = \begin{cases} 1-\frac{L\mu k^2+2xk-2x+1}{2N(k-1)} & \text{if } L \le \frac{N-1}{k\mu}; \\ \frac{(k-1)(N-x)-Lk\mu-1}{2N(k-1)} & \text{if } L > \frac{N-1}{k\mu}, \end{cases}$ so $\Delta p(1,L) = \begin{cases} \frac{k\mu}{2N}\left(\frac{(k-1)N-k}{k\mu(k-1)} - L\right) & \text{if } L < \frac{N-1}{k\mu}; \\ \frac{-1}{2N(k-1)} & \text{if } L \ge \frac{N-1}{k\mu}. \end{cases}$ For $L < \frac{(k-1)N-k}{k\mu(k-1)}$, we have $\Delta p(1,L) > 0$, which implies that there exists a point $\widetilde{X}_2 \in (1, N-\mu L]$ such that $\Delta p(x,L) \le 0$ for $x \ge \widetilde{X}_2$. For $L \ge \frac{(k-1)N-k}{k\mu(k-1)}$, we have $\Delta p(x,L) \le 0 \forall x$.

Figure OA.5 **Probabilities of Station-Level Excessive Waits $\widetilde{p}_F(x,L)$ and $\widetilde{p}_A(x,L)$ for $L<\frac{N-1}{2\mu}$ $(N=50,\mu=1)$**

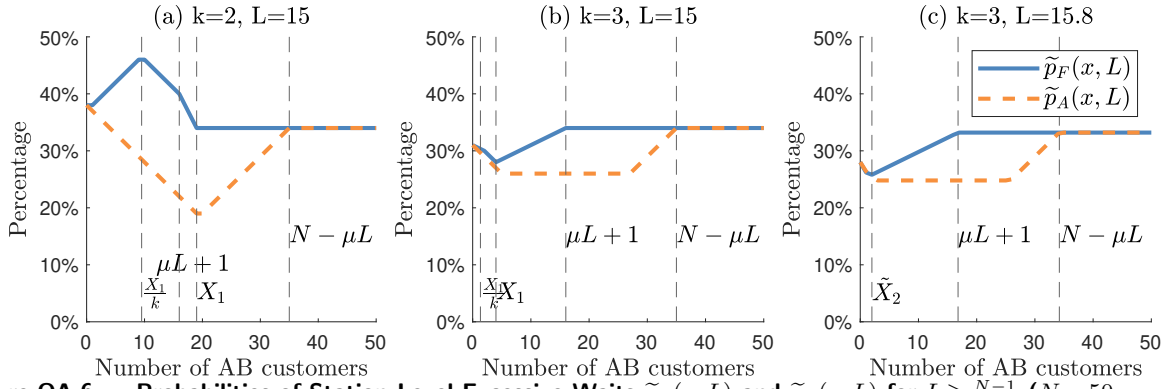
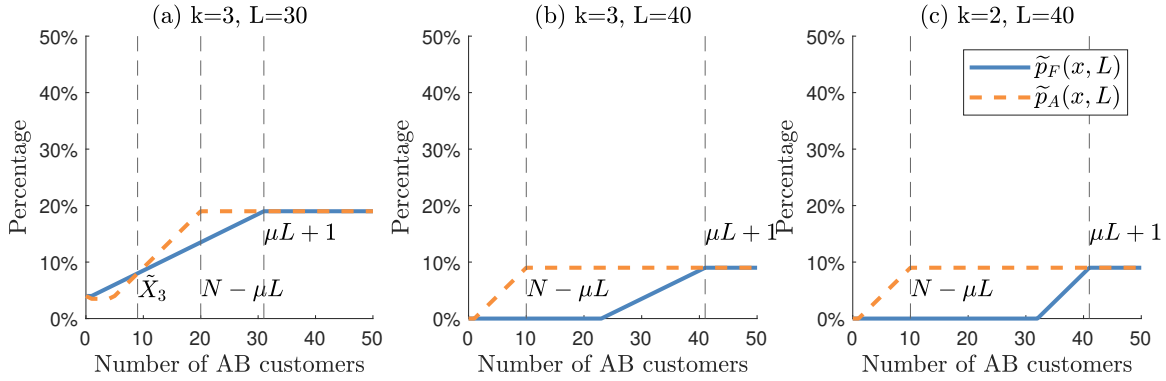


Figure OA.6 **Probabilities of Station-Level Excessive Waits $\widetilde{p}_F(x,L)$ and $\widetilde{p}_A(x,L)$ for $L\geq\frac{N-1}{2\mu}$ $(N=50,\mu=1)$**



Case 2: For $L>\frac{N-1}{2\mu}$, we have $X_1\leq 0$, $N-\mu L<\mu L+1$, $\widetilde{p}_F(x,L)$ and $\widetilde{p}_A(x,L)$ are nondecreasing in x.

2.1 For $L\leq\frac{(k-1)(N-1)}{k\mu}$ (possible for $k\geq 3$), we have (i) $\Delta p(1,L)<0$ from Case 1.3.3, (ii) $\widetilde{p}_A(x,L)$ is nondecreasing in x and reaches $\frac{N-\mu L-1}{2N}$ at $x=N-\mu L$ before $\widetilde{p}_F(x,L)$ increases to $\frac{N-\mu L-1}{2N}$ at $x=\mu L+1$, implying that there exists a point $\widetilde{X}_3\in(1,N-\mu L]$ such that $\Delta p(x,L)>0$ for $\widetilde{X}_3<x<\mu L+1$, and $\Delta p(x,L)\leq 0$ for $x\leq\widetilde{X}_3$ and $x\geq\mu L+1$.

2.2 For $L>\frac{(k-1)(N-1)}{k\mu}$, we have $\widetilde{p}_F(1,L)=\widetilde{p}_A(1,L)=0$. Therefore, we have $\Delta p(x,L)>0$ for $x<\mu L+1$, and $\Delta p(x,L)=0$ for $x\geq\mu L+1$.

Figures OA.5 and OA.6 numerically illustrate Cases 1 and 2, respectively. In Figure OA.5(a)–(b), for $k=2$ and $L=15<\frac{N-1}{2\mu}$, and for $k=3$ and $L=15\in(\frac{k^2-2}{k\mu},\frac{N-k}{k\mu})$, we have $\Delta p(x,L)\leq 0\forall x$. In Figure OA.5(c), for $k=3$ and $L=15.8\in(\frac{N-k}{k\mu},\frac{(k-1)N-k}{k\mu(k-1)})$, as derived in Case 1.3.2, there exists a point $\widetilde{X}_2$ such that $\Delta p(x,L)\leq 0$ if $x\geq\widetilde{X}_2$. In Figure OA.6(a), for $k=2$ and $L=30<\frac{(k-1)(N-1)}{k\mu}$, as derived in Case 2.1, we have $\Delta p(x,L)\leq 0$ for $x\leq\widetilde{X}_3$ and $x\geq\mu L+1$, where $\widetilde{X}_3\in(1,N-\mu L]$. In Figure OA.6(b)–(c), for $k=\{2,3\}$ and $L=40>\frac{(k-1)(N-1)}{k\mu}$, we have $\Delta p(x,L)=0$ for $x\geq\mu L+1$ as derived in Case 2.2.

B.3. Proof of Proposition 3

In the zero routing time case, i.e., $r=0$, with buffer $y=0$, stations A and B start serving the 1st AB and BA customer at $t=0$. All other customers wait in the pooled queue. BA customers return to the pooled queue and later get prioritized for service A. The 1st AB customer is the only AB customer, and the remaining $N-1$ customers are BA customers. Thus, the buffer $y=0$ is equivalent to the $x=1$ routing decision, leading to the shortest average overall wait. Therefore, the optimal buffer is $y^*=0$.

For routing time below the non-bottleneck station's service time, denoted as $r = 1/m\mu < 1/k\mu (\Leftrightarrow k < m)$, we claim it is optimal to assign customers first to the bottleneck station's queue until it reaches a certain buffer. To prove that, let's first assume that the buffer $y = 0$, and derive the average overall wait's lower bound, denoted by $\underline{t}^{Buff}(0)$ so that $t^B(0) \geq \underline{t}^{Buff}(0)$. The dispatcher routes the 1st AB customer and the 1st BA customer to stations A and B at $t = r$ and $t = 2r$. Given $m > k \geq 2$, we have $r + \frac{1}{\mu} \geq 2r + \frac{1}{k\mu}$, suggesting that station B finishes its first customer earlier than A, so the dispatcher routes the next customer (the 2nd BA customer) to station B at $t = 3r$. There is only 1 AB customer and $N-1$ BA customers. Assuming no queue at the dispatcher, the jth BA customer starts service A at $t = r + j(r + \frac{1}{\mu})$. The one AB customer starts service B earliest at $t = 2r + 2(r + \frac{1}{k\mu})$. Therefore, we have $t^B(0) \geq \underline{t}^{Buff}(0) = \frac{1}{N}(\sum_{j=1}^{N-1}(r + j(r + \frac{1}{\mu}) - \frac{1}{k\mu}) + 2r + 2(r + \frac{1}{k\mu}) - \frac{1}{\mu}) = \frac{(m+1)N}{2m\mu} + \frac{k-2m-km}{2km\mu} + \frac{3k+3m-km}{Nkm\mu}$.

When the buffer is set to $y = 1$, we first note that station A never idles from its first service, as the dispatcher can fill positions before A finishes its current service: given $m > k \geq 2$, we have $2r = \frac{2}{m\mu} \leq \frac{1}{\mu}$. We then investigate the number of AB customers, i.e., those starting with service A. The dispatcher first routes 2 AB customers to station A at $t = r$ and $t = 2r$, then routes 2 BA customers to station B at $t = 3r$ and $t = 4r$. Given that $m > k \geq 2$, we have $r + \frac{1}{\mu} \geq 3r$, suggesting that the dispatcher starts the 4th routing decision before station A finishes its first service.

Next, based on the comparison of three time points: (i) station A finishes the 1st AB customer at $t = r + \frac{1}{\mu}$; (ii) station B finishes the 1st BA customer at $t = 3r + \frac{1}{k\mu}$; and (iii) the dispatcher is ready to route the next customer to queue A/B at $t = 4r$, we have three different situations. Given that $r < \frac{1}{k\mu}$, we have $4r < 3r + \frac{1}{k\mu}$.

1. If $k \geq \frac{m}{m-2}$, we have $4r < 3r + \frac{1}{k\mu} \leq r + \frac{1}{\mu}$, implying that the dispatcher idles, then station B finishes the 1st BA customer before station A finishes the 1st AB customer. So the dispatcher routes the next customer to queue B to fill the position. In this case, there are 2 AB customers and $N-2$ BA customers. Knowing that station A never idles from $t = r$, the jth BA customer's overall wait is $t^{BA}(j) = r + \frac{j+1}{\mu} - \frac{1}{k\mu}$. Denote I_1 and I_2 as the 1st and 2nd AB customers' service order at station B. We have $I_1 \leq k + 2$ and $I_2 \leq 2k + 2$. We next derive an upper bound for these AB customers' overall wait. Assuming the routing decisions always cause a queue at the dispatcher, so that these two AB customers' overall waits at most are $t^{AB}(1) = 2r + r + \frac{2}{k\mu} + (\frac{1}{k\mu} + r)(k-1) + r - \frac{1}{\mu}$ and $t^{AB}(2) = 2r + r + \frac{2}{k\mu} + (\frac{1}{k\mu} + r)(2k-1) + r - \frac{1}{\mu}$. Therefore, the average overall wait should be at most $\bar{t}^{Buff}(1) = \frac{1}{N}(\sum_{j=1}^{N-2} t^{BA}(j) + t^{AB}(1) + t^{AB}(2)) = \frac{1}{2Nkm\mu}(8k + 8m + 2Nk - 2Nm + 6k^2 + N^2km - Nkm)$. Comparing with the zero-buffer case, we have $\bar{t}^{Buff}(1) < \underline{t}^{Buff}(0)$ when $N > \frac{1}{2k}(k + \sqrt{8k^2m + 8km + 9k^2 + 24k^3})$.

2. If $k < \frac{m}{m-2}$ and $m > 3$, we have $4r < r + \frac{1}{\mu} < 3r + \frac{1}{k\mu}$, implying that the dispatcher idles, then station A finishes the 1st AB customer before station B finishes the 1st BA customer. So the dispatcher routes the next customer (the 3rd AB customer) to queue A to fill the position at $t = 2r + \frac{1}{\mu}$. In this case, there are 3 AB customers and $N-3$ BA customers. Denote I_3 as the 3rd AB customer's service order at station B. We have $I_3 \leq 3k + 2$. The 3rd AB customer's overall wait at most is $t^{AB}(3) = 2r + r + \frac{2}{k\mu} + (\frac{1}{k\mu} + r)(3k-1) + r - \frac{1}{\mu}$. Therefore, the average overall wait should be at most $\bar{t}^{Buff}(1) = \frac{1}{N}(\sum_{j=1}^{N-3} t^{BA}(j) + t^{AB}(1) + t^{AB}(2) + t^{AB}(3)) = \frac{1}{2Nkm\mu}(12k + 12m + 2Nk - 2Nm + 6km + 12k^2 + N^2km - 3Nkm)$. Comparing with the zero-buffer case, we have $\bar{t}^{Buff}(1) < \underline{t}^{Buff}(0)$ when $N > \frac{1}{2k}(k + \sqrt{28k^2m + 4k^2m^2 + 24km + 25k^2 + 48k^3} - 2km)$.

3. If $k < \frac{m}{m-2}$ and $m \leq 3$ (which is $m = 3$ and $k = 2$), we have $r + \frac{1}{\mu} \leq 4r < 3r + \frac{1}{k\mu}$, implying that station A finishes the 1st AB customer, then the dispatcher becomes available before station B finishes the 1st BA customer. So the dispatcher routes next customer (the 3rd AB customer) to queue A to fill the position at $t = 5r$. Similar to Case 2, there are 3 AB customers and $N-3$ BA customers. The 1st BA customer is served at station A after the 3 AB customers, so $t^{BA}(j) = r + \frac{j+2}{\mu} - \frac{1}{k\mu}$. We have $\bar{t}^{Buff}(1) = \frac{1}{N}(\sum_{j=1}^{N-3} t^{BA}(j) + t^{AB}(1) + t^{AB}(2) + t^{AB}(3)) = \frac{1}{2Nkm\mu}(12k + 12m + 2Nk - 2Nm + 12k^2 + N^2km - Nkm)$. Comparing with the zero-buffer case, we have $\bar{t}^{Buff}(1) < \underline{t}^{Buff}(0)$ when $N > \frac{1}{2k}(k + \sqrt{k^2m + 24km + 25k^2 + 48k^3})$.

When the buffer is set to $y > 1$, similar to the case when $y = 1$, we know that station A never idles since $t = r$. From Proposition 1, the optimal policy is to let the bottleneck station busy while letting other customers finish the non-bottleneck station first. Increasing the buffer size y means routing more AB customers at the start and these AB customers will be postponed service A after joining BA customers who finished service B, which brings inefficiency. Therefore, we have $t^B(1) < t^B(y > 1)$.

In summary, our analysis shows that the optimal buffer size depends on the relationship between routing time and the stations' service rates. When the buffer is set to $y = 1$, the system can achieve a more efficient performance compared to $y = 0$ or $y > 1$, under certain conditions.

Appendix C: Simulation of Stochastic Two-Station Open-Shop Network

Tables OA.4 and OA.5 report the simulation results when $k = 4/3$ and $k = 2$, respectively.

Table OA.4 Impact of ACP Policy and Buffer Strategy on Macro- and Micro-Level Measures ($r = 0$, $k = 4/3$, and $L = 20$)

(i) FCFS Policy

ϕ	0			0.25			0.5			0.75			1		
γ	t_F	$t_{F,2}$	$p_F(t_s > L)$	t_F	$t_{F,2}$	$p_F(t_s > L)$	t_F	$t_{F,2}$	$p_F(t_s > L)$	t_F	$t_{F,2}$	$p_F(t_s > L)$	t_F	$t_{F,2}$	$p_F(t_s > L)$
0.001	31.71	21.46	62.81%	31.11	20.96	59.39%	30.83	20.59	58.17%	30.27	20.09	56.14%	29.88	19.78	54.24%
0.01	27.31	17.50	42.64%	27.21	17.42	42.13%	27.36	17.59	43.03%	26.87	17.17	41.55%	26.97	17.36	42.61%
0.25	21.80	13.45	21.12%	21.91	13.54	21.41%	22.12	13.66	22.25%	21.63	13.39	20.22%	21.51	13.26	20.06%
0.5	13.81	8.86	5.01%	13.81	8.87	5.10%	13.72	8.74	4.90%	14.01	8.96	5.54%	13.83	8.84	5.43%
0.75	7.35	5.53	1.45%	7.59	5.75	1.60%	7.73	5.82	1.63%	7.49	5.56	1.71%	7.70	5.75	1.40%
1	3.28	2.78	0.08%	3.62	3.06	0.17%	3.49	2.92	0.14%	3.46	2.88	0.08%	3.32	2.74	0.05%

(ii) ACP Policy

ϕ	0			0.25			0.5			0.75			1		
γ	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$
0.001	24.89	3.39	45.98%	24.87	3.60	44.74%	24.25	3.31	44.21%	23.96	3.41	43.60%	23.96	3.64	42.18%
0.01	22.21	3.65	37.73%	22.15	3.61	37.56%	22.36	3.71	37.27%	22.06	3.70	36.81%	22.26	3.81	36.54%
0.25	18.68	4.36	26.32%	18.67	4.35	26.80%	18.68	4.17	27.14%	18.47	4.26	26.34%	18.35	4.17	25.92%
0.5	12.67	5.31	12.19%	12.60	5.24	12.76%	12.55	5.14	12.21%	12.82	5.21	12.92%	12.61	5.13	12.09%
0.75	7.06	4.33	4.07%	7.31	4.56	4.14%	7.50	4.58	4.30%	7.26	4.22	4.54%	7.44	4.47	3.96%
1	3.25	2.45	0.67%	3.61	2.72	0.85%	3.50	2.52	0.92%	3.46	2.46	0.81%	3.36	2.36	0.77%

(iii) ACP Policy with Buffer Strategy $y = 0$

ϕ	0			0.25			0.5			0.75			1		
γ	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$
0.001	24.24	6.52	44.20%	24.09	6.70	42.99%	23.65	6.31	42.23%	23.23	6.13	41.58%	23.29	6.47	40.97%
0.01	21.64	6.14	36.06%	21.64	6.22	35.90%	21.69	6.25	36.18%	21.37	6.18	35.37%	21.71	6.63	35.20%
0.25	18.14	6.19	20.68%	18.22	6.16	21.26%	18.40	6.37	22.28%	18.10	6.36	20.42%	18.00	6.11	20.87%
0.5	12.46	6.26	3.86%	12.42	6.19	4.41%	12.33	6.12	3.88%	12.53	6.17	4.54%	12.44	6.19	4.15%
0.75	6.85	4.78	1.41%	7.12	5.07	1.52%	7.19	5.06	1.50%	6.95	4.77	1.61%	7.20	5.04	1.34%
1	3.05	2.57	0.08%	3.38	2.84	0.19%	3.23	2.68	0.13%	3.17	2.62	0.08%	3.07	2.53	0.05%

Table OA.5 Impact of ACP Policy and Buffer Strategy on Macro- and Micro-Level Measures ($r=0$, $k=2$, and $L=20$)

(i) FCFS Policy

ϕ	0			0.25			0.5			0.75			1		
γ	t_F	$t_{F,2}$	$p_F(t_s>L)$	t_F	$t_{F,2}$	$p_F(t_s>L)$	t_F	$t_{F,2}$	$p_F(t_s>L)$	t_F	$t_{F,2}$	$p_F(t_s>L)$	t_F	$t_{F,2}$	$p_F(t_s>L)$
0.001	27.81	19.80	46.24%	27.81	19.77	46.73%	27.17	19.23	45.43%	26.99	19.03	44.35%	26.28	18.40	41.42%
0.01	23.89	16.53	34.58%	24.05	16.63	36.02%	23.59	16.35	35.13%	23.94	16.62	35.94%	23.31	16.05	32.62%
0.25	19.01	13.52	22.98%	18.97	13.57	22.89%	18.78	13.32	22.46%	18.87	13.38	22.90%	18.65	13.34	22.58%
0.5	12.40	10.83	14.44%	12.49	11.02	15.31%	12.22	10.57	13.97%	12.35	10.66	14.28%	12.52	10.83	14.30%
0.75	6.78	6.48	2.74%	6.68	6.36	2.66%	6.95	6.58	2.89%	6.86	6.48	2.42%	6.72	6.30	2.50%
1	2.82	2.72	0.11%	2.80	2.69	0.16%	2.96	2.83	0.16%	3.27	3.10	0.19%	3.29	3.11	0.09%

(ii) ACP Policy

ϕ	0			0.25			0.5			0.75			1		
γ	t_A	$t_{A,2}$	$p_A(t_s>L)$	t_A	$t_{A,2}$	$p_A(t_s>L)$	t_A	$t_{A,2}$	$p_A(t_s>L)$	t_A	$t_{A,2}$	$p_A(t_s>L)$	t_A	$t_{A,2}$	$p_A(t_s>L)$
0.001	24.22	5.98	34.10%	24.10	6.15	33.52%	23.56	6.14	32.92%	23.74	6.53	32.57%	23.22	6.58	32.10%
0.01	21.82	7.63	29.01%	21.88	7.61	28.93%	21.46	7.48	29.03%	21.81	7.91	28.90%	21.16	7.49	27.51%
0.25	18.27	10.11	23.14%	18.18	10.18	22.87%	17.92	9.68	22.64%	18.13	9.94	22.68%	17.92	9.89	22.14%
0.5	12.31	10.10	14.80%	12.45	10.36	16.03%	12.16	9.85	14.80%	12.28	9.85	14.88%	12.45	10.09	14.83%
0.75	6.80	6.25	3.05%	6.66	6.10	2.95%	6.97	6.33	3.30%	6.87	6.18	2.90%	6.76	6.04	3.02%
1	2.81	2.65	0.23%	2.81	2.64	0.24%	2.97	2.75	0.32%	3.29	3.02	0.36%	3.30	3.00	0.31%

(iii) ACP Policy with Buffer Strategy $y=0$

ϕ	0			0.25			0.5			0.75			1		
γ	t_A	$t_{A,2}$	$p_A(t_s>L)$	t_A	$t_{A,2}$	$p_A(t_s>L)$	t_A	$t_{A,2}$	$p_A(t_s>L)$	t_A	$t_{A,2}$	$p_A(t_s>L)$	t_A	$t_{A,2}$	$p_A(t_s>L)$
0.001	23.83	11.73	27.88%	23.94	12.04	29.00%	23.37	11.69	27.63%	23.52	12.03	28.07%	22.94	11.64	26.68%
0.01	21.67	11.92	23.06%	21.79	11.96	23.54%	21.30	11.73	22.70%	21.66	12.06	24.20%	20.99	11.38	21.82%
0.25	18.16	11.78	19.12%	18.15	11.93	18.94%	17.85	11.50	18.64%	18.00	11.66	18.76%	17.82	11.70	18.72%
0.5	12.25	10.61	14.28%	12.33	10.78	15.10%	12.05	10.34	13.86%	12.14	10.37	14.02%	12.34	10.60	14.18%
0.75	6.67	6.37	2.67%	6.56	6.24	2.60%	6.80	6.45	2.76%	6.73	6.35	2.37%	6.58	6.17	2.40%
1	2.76	2.67	0.11%	2.73	2.63	0.16%	2.88	2.76	0.16%	3.16	3.01	0.19%	3.18	3.03	0.09%

Appendix D: The Clinic's Routing Policy

The *routing policy* determines the station that a newly freed customer should be routed to. Assume customer i completes a test at time t and needs to be routed to the next station. Let $\Omega_i(t)$ be the set of *eligible* stations for customer i at time t, i.e., stations that customer i needs to visit without violating precedence constraints. Let $l_s(t)$ be the number of customers at station s at time t, including those in the queue and being served, and let c_s be the number of servers at station s. The shortest queue that customer i observes among all eligible stations at time t is $L=\min_{s\in\Omega_i(t)} l_s(t)$. For the subset of non-bottleneck stations, customer i observes the shortest queue as $L_{NB}=\min_{s\in\Omega_i(t)\cap NB} l_s(t)$.

- If some eligible stations have idle servers, i.e., $l_s(t)<c_s$ for some $s\in\Omega_i(t)$, the dispatcher will route customer i to the idle station with the longest average service time.
- If no eligible stations have idle servers, i.e., $l_s(t)\geq c_s$ for $\forall s\in\Omega_i(t)$, the dispatcher routes customer i to the station with the longest HOL wait in a station set. This station set includes mostly the shortest queue length stations within $\Omega_i(t)$, with some exceptions:
  - —If some non-bottleneck stations have the shortest queue (i.e., $L_{NB}=L$), the dispatcher will add these stations into the station set.
  - —If non-bottleneck stations' shortest queues are of length $L+1$ (i.e., $L_{NB}=L+1$), while some bottleneck stations have queue length L, the dispatcher will only select the non-bottleneck stations with queue length $L+1$ into the station set.

—If all non-bottleneck stations have more than $L+1$ customers waiting (i.e., $L_{NB} \geq L+2$) or there are no eligible non-bottleneck stations (i.e., $\Omega_i(t) \cap NB = \emptyset$), the dispatcher will select all stations (either bottleneck or non-bottleneck) with queue length not exceeding $L+2$ into the station set.

Appendix E: Simulation of ACP Policies and Shortest-Expected-Wait-Time (SEWT) First Routing Policy

Routing decisions in simulations in Section 5 are made according to the routing policy stated in Section 4.2 and Appendix D. Here, we propose to estimate the congestion by the station's expected wait, based on the classical queue-length Markovian (QLM) delay-announcement estimator (Ibrahim and Whitt 2010). The expected wait at station j at moment t is $\frac{QL_{jt} \times AvgProcT_j}{NumServers_{jt}}$, where QL_{jt} is the number of customers waiting in queue for station j at moment t, $AvgProcT_j$ is the average processing (call and service) time of station j, and $NumServers_{jt}$ is the number of servers at station j at time t. A newly freed customer will be assigned to the eligible station with the Shortest-Expected-Wait-Time (SEWT), regardless of whether it is a bottleneck.

Table OA.6 Macro- and Micro-Level Measures under FCFS and ACP Policies and SEWT Routing Policy

Priority policies	Macro-level wait (min)						Micro-level wait			
	Wait-for-routing		Wait-for-service		Total		Last wait		> 20 min (%)	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
FCFS	8.74	6.02	105.55	53.75	114.29	56.11	20.99	20.17	21.56%	17.70%
ACP-LST	8.10	5.48	100.70	55.73	108.80	57.21	18.72	29.11	19.52%	17.71%
ACP-SERPT	7.22	5.11	91.84	71.69	99.05	73.66	10.91	16.05	13.34%	13.46%

Table OA.6 presents the macro- and micro-level performance measures under SEWT routing policy and different priority policies. Our observations are similar to those we made based on Table 4 in Section 5.1. First, compared to FCFS, ACP policies shorten the average overall wait, reduce the average wait at the last station, and the probability of station-level excessive waits. We also plot the dynamics of station-level waits and the probability of waits exceeding 20 minutes in Figure OA.7. In Figures OA.7(b–c) and (e–f), replacing FCFS with ACP policies moderates the increasing trends in Figures OA.7(a–b). Specifically, ACP-LST policy leads to a relatively balanced wait over the service process, while ACP-SERPT policy has higher peaks and lower ends. To conclude, we see that our observations on the advantages of the proposed ACP policies are robust to the specific congestion estimator used for routing.

Appendix F: Simulation of Customers with Varying Service Needs

In Section 3, we consider all customers need to visit both stations. We now relax this assumption and consider some customers only require service from one station. Denote N_A, N_B, and N_{AB} as the number of customers who visit only station A, only station B, both stations, respectively. Using simulation, we observe that, if the number of customers who only need one station's service is relatively small, then compared to FCFS, ACP policy still improves both macro- and micro-level performance measures. The simulation results are detailed in Table OA.7 below.

Figure OA.7 Dynamics of Wait-for-Service and Probability of Excessive Waits under SEWT Routing Policy

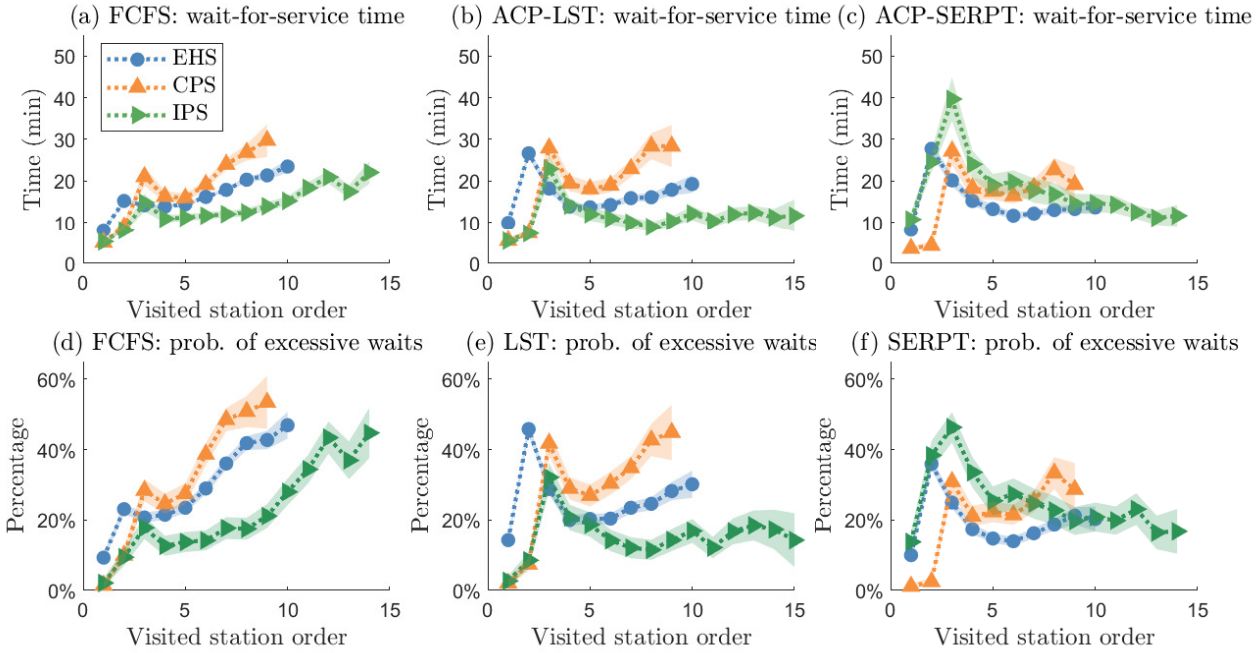


Table OA.7 Impact of ACP Policy on Macro- and Micro-Level Measures ($k = 3/2$, $L = 20$, $N_A = N_B = 10$, $N_{AB} = 50$)

(i) FCFS Policy

ϕ	0			0.25			0.5			0.75			1		
γ	t_F	$t_{F,2}$	$p_F(t_s > L)$	t_F	$t_{F,2}$	$p_F(t_s > L)$	t_F	$t_{F,2}$	$p_F(t_s > L)$	t_F	$t_{F,2}$	$p_F(t_s > L)$	t_F	$t_{F,2}$	$p_F(t_s > L)$
0.001	32.74	22.75	27.12%	32.26	22.38	32.70%	31.89	22.12	32.67%	31.58	21.90	31.97%	31.16	21.46	31.44%
0.01	27.68	18.76	18.26%	27.74	18.84	18.62%	27.82	18.92	18.67%	27.61	18.78	18.59%	27.29	18.54	18.17%
0.25	21.07	14.29	7.91%	21.34	14.45	8.18%	21.29	14.43	8.16%	21.04	14.31	7.97%	20.88	14.08	7.50%
0.5	12.03	9.37	1.48%	11.94	9.26	1.27%	11.88	9.18	1.24%	12.23	9.55	1.58%	12.22	9.58	1.62%
0.75	5.19	4.64	0.10%	5.21	4.68	0.18%	5.27	4.67	0.20%	5.24	4.63	0.16%	5.49	4.85	0.13%
1	1.88	1.72	0.03%	1.87	1.69	0.00%	1.93	1.73	0.00%	2.05	1.83	0.00%	2.01	1.78	0.00%

(ii) ACP Policy

ϕ	0			0.25			0.5			0.75			1		
γ	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$	t_A	$t_{A,2}$	$p_A(t_s > L)$
0.001	26.41	5.62	14.05%	26.20	5.64	17.94%	25.79	5.50	17.88%	25.75	5.66	17.59%	25.45	5.49	17.65%
0.01	23.19	5.54	11.05%	23.10	5.52	10.95%	23.13	5.52	11.07%	23.06	5.60	10.96%	22.91	5.65	10.89%
0.25	18.41	6.53	7.37%	18.55	6.52	7.11%	18.48	6.46	7.18%	18.38	6.54	7.08%	18.16	6.25	6.96%
0.5	11.33	7.11	2.90%	11.27	7.03	2.71%	11.14	6.83	2.80%	11.53	7.28	3.05%	11.54	7.30	3.04%
0.75	5.15	4.08	0.64%	5.15	4.10	0.73%	5.27	4.08	0.80%	5.19	4.01	0.70%	5.44	4.19	0.76%
1	1.89	1.56	0.18%	1.88	1.54	0.13%	1.93	1.56	0.14%	2.06	1.65	0.15%	2.02	1.58	0.17%