Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Strategic Idleness and Dynamic Scheduling in an Open-Shop Service Network: Case Study and Analysis

Opher Baron, Oded Berman, Dmitry Krass, Jianfu Wang

informs.
http://pubsonline.informs.org/journal/msom/

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 19, No. 1, Winter 2017, pp. 52–71
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Strategic Idleness and Dynamic Scheduling in an Open-Shop Service Network: Case Study and Analysis

Opher Baron,[a] Oded Berman,[a] Dmitry Krass,[a] Jianfu Wang[b]

[a] Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada; [b] Nanyang Business School, Nanyang Technological University, Singapore 639798

Contact: opher.baron@rotman.utoronto.ca (OpB); berman@rotman.utoronto.ca (OdB); krass@rotman.utoronto.ca (DK); wangjf@ntu.edu.sg (JW)

Received: February 21, 2014
Revised: November 24, 2014; December 23, 2015; April 18, 2016
Accepted: May 3, 2016
Published Online in Articles in Advance: October 27, 2016

https://doi.org/10.1287/msom.2016.0591

Abstract. This paper, motivated by a collaboration with a healthcare service provider, focuses on stochastic open-shop service networks with two objectives: more traditional macrolevel measures (such as minimizing total system time or minimizing total number of tardy customers) and the atypical microlevel measure of reducing the incidents of excessively long waits at any workstation within the process. While work-conserving policies are optimal for macrolevel measures, scheduling policies with strategic idleness (SI) might be helpful for microlevel measures. Using the empirical data obtained from the service provider, we provide statistical evidence that SI is used by its schedulers to manage the macro- and microlevel measures. However, the company has no specific rules on implementing SI and the schedulers make decisions based on their own experience. Our primary goal is to develop a systematic framework for the joint usage of SI with dynamic scheduling policies (DSPs). We suggest to use *threshold-based policies* to intelligently combine SI and DSPs and show that the resulting policies provide an efficient way to simultaneously address both macro- and microlevel measures. We build two simulation models: one based on empirical data and one based on a randomly generated open-shop network. We use both models to demonstrate that an open-shop service network can be systematically and effectively managed to deliver improved service level by using SI.

Funding: The first three authors acknowledge financial support from their respective Natural Sciences and Engineering Research Council of Canada grants.

Supplemental Material: The online appendix is abailable at https://doi.org/10.1287/msom.2016.0591.

Keywords: strategic idleness • dynamic scheduling policy • open shop • healthcare • service network

1. Introduction

In this paper, we focus on stochastic open-shop processes operating under multiple objectives. The system objectives include a combination of the more traditional *macrolevel* measures (such as minimizing total system time and minimizing total number of tardy customers) with an atypical *microlevel* objective seeking to limit the number of incidents where a customer experiences an excessively long wait at a workstation within the process. This combination of objectives is motivated by the understanding that customers' perception of service quality is affected by both macro- and microlevel factors.

Open-shop service networks, where customers need to visit a set of stations without a specific service order, are common in modern service industry in both brick-and-mortar and virtual service operations. Examples of the former include retail stores, where customers may visit several departments before proceeding to the cashier, and hospitals, where patients often go through several diagnostic and treatment stations. Examples of virtual systems include contact centers, where calls may be served by a mixture of automated response units and human agents with different levels of expertise, and online websites, where customers often visit a number of pages before proceeding to the checkout page.

This paper was motivated by an open shop operation in the healthcare service industry. XYZ (the real name of the company is removed and relevant data has been disguised for confidentiality) is one of the leaders in preventive healthcare services in Canada. Their flagship service is composed of 10–20 different medical tests, each test performed in a different station. The order in which customers take most of these tests is mostly immaterial (there are only a few precedence constraints), so XYZ operates an open-shop service system. XYZ's service is primarily targeted at busy professionals who are willing, for a fee, to have a complete assessment of their health performed in just a few hours. We note that under the Canadian medical system, most of these tests can be done for free, but would likely take days or even months to schedule and complete. Thus, convenience is the main selling feature, and delivering excellent customer service is of paramount importance for XYZ. Tight management of customer waiting times in the system is deemed to be essential. While, because of the inherent variability of the service times, waiting time is unavoidable, the goal is to minimize waiting times and maximize customers' perception of service quality. To study the waiting times in XYZ, we interviewed their key personnel and collected 2 months of data, comprising 41 business days with just over 2,000 customers and about 20,000 station visits.

While there are many determinants of service quality in service networks, the link between customers' waiting times and the perceived service quality is well recognized (Friedman and Friedman 1997, Taylor 1994). Waiting times, which are easily quantifiable, have long been the focus of much of the queueing literature. The most common measure of waiting time is the expected overall waiting time for service. A related measure is the probability that the total system time or the total waiting time exceeds a certain predetermined threshold. XYZ uses both of these service-level measures (SLMs) with the following stated targets: mean system time is less than four hours, and the probability of system time longer than four hours is less than 50%.

The two SLMs described above take a "macro" view of the service network, essentially treating it as a one-stage system and looking at the overall system or waiting time. Such SLMs may be sufficient in the manufacturing context, where customers, after placing an order, remain outside the system and essentially view the system as a black box. However, in service context the customer is not just an outside observer: they experience the internal performance of the system as well, i.e., waiting times in front of individual workstations. A poor experience at a given workstation may lead to a poor perceived service quality, even when the macrolevel SLM (e.g., the overall waiting time) does not indicate a problem.

In fact, XYZ's senior management noticed that there have been two types of complaints about the service experience: the first is with respect to the total time customers spend in the system, and the second, more prevalent, is associated with a long wait for a particular station. From customer satisfaction surveys, XYZ found that customers whose wait for service at any station exceeded 20 minutes were substantially less satisfied with their overall service experience than others. This led XYZ to define a microlevel SLM: the waiting time in front of any station should not exceed 20 minutes. This measure is taken very seriously: when a customer's waiting time at a station reaches 15 minutes, an anxious *yellow face* appears on process schedulers' screens. Once the waiting time reaches 20 minutes, a *red face* (RF) appears, and a service breakdown is considered to have occurred. The total number of red faces is carefully tracked: the goal is to keep it below 100 per month.

The importance of similar microlevel measures has been observed in other settings. For example, Bouch et al. (2000) show that for an online service, when a single web page takes longer than eight seconds to load, customer service quality ratings fall off dramatically. Soman and Shi (2003) show that people prefer a process in which they are making progress toward their goal at a steady rate, with the presence of waits adversely impacting the perception of service quality. Similarly, the psychology of queueing literature, e.g., Larson (1987), shows that customers' perception of the queueing experience may vary nonlinearly with the length of the delay.

Nevertheless, the systematic treatment of microlevel SLMs in queueing or stochastic scheduling literature is relatively new. The only prior paper we are aware of is Baron et al. (2014), where it was demonstrated analytically for a two-station tandem queue network that a scheduling policy with *strategic idleness* (*SI*) might be helpful in reducing the probability of long waits at a station. The idea behind SI policies is that when a downstream station accumulates a queue, continuing to operate upstream stations at the normal rate may lengthen the queue downstream and increase the probability of long waits (i.e., the expected number of RFs). A better idea is to idle the upstream stations (or to temporarily reduce their service rates), allowing the downstream queue to dissipate. Such idling effectively redistributes the waiting times more evenly among the stations and reduces the number of red faces. We note that the scheduling policies employing SI violate one of the common assumptions in queueing analysis: the work conserving property, which states that a workstation should continue to operate as long as there are customers waiting to be served. However, the potential payoff of SI may be very attractive. Indeed the classical way of reducing probability of long waits is by adding capacity to the system (e.g., adding a doctor in the healthcare setting), which may be quite expensive. Using a dynamic scheduling policy with SI can potentially achieve the same objective at a negligible cost: by simply idling some resources in the system.

Given the results from Baron et al. (2014) discussed above and that the number of red faces is used as a performance measure at XYZ, one may expect that the schedulers employ SI. While our interviews with XYZ's management team indicate that such use of SI is not the company's policy, our analysis of the empirical data provide indirect statistical evidence that XYZ's schedulers effectively employ SI to manage the number of red faces—simulation results indicate that the current scheduling rule without the use of SI would likely result in more than twice the current number of red face incidents.

Our primary interest is to study the benefit of combining SI with dynamic scheduling policies (DSPs) for open-shop service networks, such as the one operated by XYZ. We start by representing general DSPs as simple scoring rules. A variety of different DSPs based on intuition and known theoretical results for stylized systems are proposed to address the macrolevel SLMs. We then show how a *threshold-based policy* approach can be used to intelligently inject SI into a given DSP, resulting in a policy that can potentially address both macro- and microlevel objectives simultaneously.

To test our policies we use simulation models. The need for the simulation-based approach is driven by both the complexity of stochastic open-shop networks, making analytical results very hard to obtain, and the transient nature of real-life systems: since the system starts each workday with an empty queue and ends it in the same state after seeing relatively few customers, the steady state may never set in.

In the first simulation model, we use the empirical data on arrivals and service times to first understand the service process at XYZ and then to calibrate a detailed simulation model. We investigate the performance of several DSPs with and without SI and compare them to the performance of the empirical scheduling policies used by XYZ. We show that the automated policies achieve very promising results: the best DSPs are able to significantly outperform the actual schedules with respect to the macrolevel measures. While without the use of SI the DSPs tend to perform poorly on the microlevel measures (the same effect is demonstrated for the actual scheduling policies), after the SI modification, DSPs are able to perform very competitively on the microlevel SLM, while maintaining their advantage with respect to macrolevel SLMs.

In the second simulation model, we evaluate our policies in randomly generated open-shop networks and demonstrate the benefits of using SI for these more general networks under both transient- and steady-state operations.

To summarize, this paper makes several contributions: (1) we narrow the gap between theory and practice on scheduling in open-shop service networks; (2) we develop principles for integrating SI with DSPs for a stochastic open-shop service system and, using simulation models, demonstrate that an open-shop service network can be managed in a *systematic fashion* to deliver improved service levels with respect to both macro- and microlevel SLMs; and (3) we provide statistical evidence for the usage of SI in practice.

This paper continues as follows. In Section 2, we provide a brief literature review, focusing on known results for open-shop systems. In Section 3, we introduce the framework for using DSPs and integrating SI into DSPs in a stochastic open-shop service network. We propose several DSPs and SI policies that are likely to be useful in practice. In Section 4, we analyze the empirical data provided by XYZ and present evidence for the usage of SI in XYZ. In Section 5, we use simulation models to demonstrate the effect of SI and evaluate several DSPs and SI policies. In Section 6, we present conclusions and suggestions for future research.

2. Literature Review

Open shops have been studied extensively in manufacturing settings (see, e.g., Roemer 2006), such as airplane maintenance, fire engine assembly, just-in-time systems, and supply chain assembly systems. A few papers consider service systems, such as accounting services, but they still consider common manufacturing measures as service objectives.

Open-shop problems are typically NP-hard and consider only macrolevel measures (e.g., see, Pinedo 2012 and reference therein). For the deterministic open shop with preemptions, polynomial time algorithms are available for the makespan objective and maximum lateness objective. Without preemptions, for the makespan objective, only an open shop with $m = 2$ stations has the polynomial-time optimal policy (the longest alternate processing time first policy), and an open shop with $m \geq 3$ stations is known to be NP-hard. For the maximum lateness objective, open shop with $m = 2$ stations is already strongly NP-hard. Furthermore, very little can be said about the total completion time objective; open shops with this objective are NP-hard for all $m \geq 2$ cases, with or without preemptions.

For the stochastic open shops, theoretical results are limited to the $m = 2$ case. Pinedo and Ross (1982) prove that the longest expected remaining processing time first policy minimizes the expected makespan of a stochastic two-station open shop. Pinedo (1984) show that the preemptive shortest expected remaining processing time first (SERP) policy minimizes the total expected completion time in a two-station open shop within the class of preemptive dynamic policies.

Alcaide et al. (2006) develop a predictive-reactive approach to minimize expected makespan in an open shop with $m \geq 3$ stations; they dynamically modify a heuristic schedule, based on Alcaide et al. (1997), whenever an unexpected event occurs.

A vast literature is focused on the analysis, design, and control of queueing networks; see Stidham (2002) and Chen and Yao (2001). However, a stochastic open-shop cannot be easily represented as a queueing network, as it requires customer-based history (keeping track of which stations have already been visited); queueing networks typically lack mechanisms for keeping track of such history of visits.

By far, the most popular objective in the queueing literature is the total system time (see, e.g., the survey by Gans et al. 2003). A related measure is the probability that the total system time or the total waiting time exceeds a certain predetermined threshold; see Baron and Milner (2009) and de-Véricourt and Jennings (2011) and references therein. While several authors have looked beyond the traditional measures, their objectives can also be classified as macrolevel SLMs (see de-Véricourt and Zhou 2005, Mehrotra et al. 2012, Saghafian et al. 2012).

The systematic study of the microlevel SLM focusing on the instances of excessive waits originates with Baron et al. (2014), who, in the context of serial queues, demonstrate the advantage of policies with SI by applying a *threshold-based policy* (*TBP*)—we will use these ideas in Section 3.

We also mention several papers on applied stochastic flow networks that have certain similarities to the present work. Huang et al. (2015) study a patient flow control problem with constraints on the initial waits in an emergency department. They demonstrate that in the asymptotical regime, the optimal policy to reduce waiting costs is to prioritize patients with the shortest remaining work, similar to our shortest expected remaining processing time policy introduced in Section 3.

Graves (1986) develops a tactical planning model that can be used to set production rates to achieve a smooth work flow. He suggests that smoothing work at upstream stations can reduce the variability of the arrivals to the downstream stations, and this helps the downstream stations in workforce planning (see also subsequent papers like Teo et al. 2012, Chhaochhria and Graves 2013). The model, although quite different from ours, is somewhat similar in spirit.

Shtrichman et al. (2001) considers an open-shop service network of an army recruitment office and tries to minimize the total system time and the probability that a customer does not finish visiting all stations in one day. These objectives are similar to our macrolevel SLMs. Using simulation, they show that a shortest expected waiting time routing policy (similar to our SERP policy) can significantly improve their two main service-level measures. The main differences between their study and ours lie in the presence of the microlevel SLM in our case, as well as in the much larger scale of the physical system (spanning several floors of a building) in their case. This makes it necessary to coordinate multiple physical queues and to account for interstation travel, while XYZ employs a single centralized waiting room and interstation travel times are not significant.

The service network of XYZ can be considered as a fork-joint network, but as discussed in Section 4, this network has different characteristics than the classic fork-joint queues. For reference on routing of such queues, see Atar et al. (2012) and references therein.

There are two other settings where intentionally idling a capacitated resource has been considered previously. Strategic delays were introduced by Afèche (2013), who showed how such delays can allow a service provider to differentiate between customer types and thus improve the overall profit. In the manufacturing process control literature, the most prominent example of using intentional idleness is the kanban manufacturing system where the total inventory between two stations is restricted to be lower than a threshold; for further details, see Masin et al. (2005) and references therein. In this case the motivation for idling is the need to control inventory and its cost without sacrificing too much capacity. In both cases, the motivation for intentional idling is significantly different from that in the current paper, which is improving the customer service experience. This difference leads to completely different analysis and implementation challenges.

3. Dynamic Scheduling Policies and Strategic Idleness Modification

We start by introducing stochastic open shop with precedence constraints in Section 3.1. We next present completely reactive DSPs in terms of simple scoring rules in Section 3.2. Using this framework, we propose several simple DSPs that have the potential to perform well in practice with respect to the macrolevel SLMs. In Section 3.3, we show how any completely reactive DSP can be modified to inject SI, allowing the policy to take into account our microlevel RF SLM. We use mathematical notation in defining our DSPs and SI modification so that these definitions will be rigorous.

3.1. Stochastic Open Shop with Precedence Constraints

We consider a general stochastic open-shop problem with precedence constraints described as follows: a set of n customers $C = \{1, \dots, n\}$ with (possibly random) arrival times $r_1^c, \dots, r_n^c$, wish to finish service within T^s time units after arrival (i.e., their due dates are $r_1^c + T^s, \dots, r_n^c + T^s$). The customers need to obtain service from a set of m stations $S = \{1, \dots, m\}$ that open at prescheduled times $r_1^s, \dots, r_m^s$ and close when all customers have finished service. For simplicity, we assume that $r_i^c < r_{i+1}^c$, for $i = 1, \dots, n-1$, and $r_j^s < r_{j+1}^s$, for $j = 1, \dots, m-1$; i.e., this implies no batch arrivals or stations opening at the same time. Customer i requires service from some subset $S_i \subseteq S$ of stations, and she must visit every station in S_i exactly once. The services from stations in S_i can be received in any order that satisfies precedence constraints $U_i = \{(h, k), \dots \mid h, k, \dots \in S_i\}$, where constraint (h, k) means that customer i must visit station h before becoming eligible for station k. For example, $U_i =$

$\{(1,2),(1,3),\ldots,(1,m) \mid 1,,\ldots,m \in S_i\}$ means that customer i needs to visit station 1 before visiting any other stations. Note that if $U_i = \varnothing$ for all i, the problem becomes a classic open-shop problem (see, e.g., Pinedo 2012). Customer i's service time at station j (for $j \in S_i$), X_{ij}, is a *continuous* random variable with distribution G_j and mean $\bar{s}_j$. We assume that X_{ij} are independent and identically distributed for all j. The realization x_{ij} only becomes known upon service completion. We consider the problem *without* preemptions (i.e., a station is not allowed to accept new customers before finishing the service of the current customer). The time customer i finishes service and exits the system is denoted by F_i, which is also a random variable.

We treat the construction of an optimal schedule as a multiobjective problem. Specifically, we consider two *macrolevel SLMs* as our objectives: minimize *expected total system time*,

$$E\left[\sum_{i=1}^{n}(F_i - r_i^c)\right], \tag{1}$$

and minimize *the expected number of tardy customers*,

$$E\left[\sum_{i=1}^{n} 1(F_i > (r_i^c + T^s))\right]. \tag{2}$$

Note that since r_i^c and T^s are independent of the scheduling policy, the expected total system time objective is equivalent to the expected total lateness objective, $E[\sum_{i=1}^{n}(F_i - (r_i^c + T^s))]$, or the expected total completion time objective, $E[\sum_{i=1}^{n} F_i]$.

In addition to the macrolevel SLMs above, we also consider a *microlevel SLM* as our third objective: minimize *the expected number of red faces* (i.e., the number of instances of unacceptably long waits),

$$E\left[\sum_{i,\, j\in S_i} 1(W_{ij} > T^{RF})\right], \tag{3}$$

where T^{RF} is the threshold used to identify red faces, and W_{ij} is the random variable denoting the time customer i spent in the waiting room before entering station j.

3.2. Work-Conserving Dynamic Scheduling Policies in Open Shops

In the manufacturing literature, DSPs that consider uncertain real-time events (e.g., stochastic service times, station breakdown, defective material, job cancelation, etc.) have been classified into three categories (see, e.g., Ouelhadj and Petrovic 2009): (1) a *completely reactive* scheduling policy generates no firm schedule in advance and makes decisions locally in real time, (2) a *predictive–reactive* scheduling policy develops a schedule first and revises it in response to real-time stochastic events, and (3) a *robust proactive* scheduling policy follows a preset schedule that satisfies performance requirements predictively in a stochastic environment.

We focus our investigation on completely reactive DSPs: as discussed in the literature review, since polynomial time algorithms for deriving the optimal scheduling policy in open shops are not available, it is hard to generate any firm schedule in advance, and thus the advantage of predictive-reactive or robust proactive DSPs is difficult to assess. We will initially consider only work-conserving DSPs, i.e., policies under which a customer cannot be waiting for a currently idle station.

The completely reactive work-conserving DSPs take actions at three types of events: service completions, customer arrivals, and station openings. In fact, by introducing n dummy stations serving n customers with service completion times equal to customer arrival times $r_1^c,\ldots,r_n^c$, and m dummy customers with service completion times equal to station opening times $r_1^s,\ldots,r_m^s$, it suffices to only consider service completion events.

Following a common simplifying assumption in queueing and stochastic scheduling literature, we assume that no two service completions happen at the same time, i.e., there exists an $\epsilon > 0$ such that the time interval between any two service completions is at least ϵ. We also assume that all customer waits happen in the same (physical or virtual) "waiting room" from which the customer can join any station they still require. These assumptions fit the XYZ process and simplify the discussion that follows.

Consider a service completion involving customer i and station j occurring at some time t. Let $\Omega_j(t) \subseteq C$ be the set of free *eligible* customers, i.e., those satisfying all precedence constraints, that still require service from station j and who are in the waiting room at time t. Similarly, let $\Psi_i(t) \subseteq S_i$ be the set of idle stations at time t for which customer i is currently eligible. The DSP needs to perform at most two assignments: assign a waiting customer $h \in \Omega_j(t)$ to station j (assuming $\Omega_j(t) \neq \varnothing$) and assign an idle station $k \in \Psi_i(t)$ to customer i (assuming $\Psi_i(t) \neq \varnothing$).

These observations allow us to represent a DSP with two scoring rules, where higher is better. We assign a score, $PT_h^c(t) \geq 0$, to customers $h \in \Omega_j(t)$, and a score, $PT_k^s(t) \geq 0$, to stations $k \in \Psi_i(t)$. To make $PT_h^c(t)$ and $PT_k^s(t)$ general enough, we define them as arbitrary nonnegative functions of the history of the system up to time t.

Definition 1. Completely reactive and work-conserving *DSP* in a stochastic open-shop network is defined by scoring rules $PT_h^c(t)$ and $PT_k^s(t)$ as follows. Suppose customer i completes service on station j at time t:

(1) If the set of free eligible customers $\Omega_j(t)$ is not empty, we assign the highest-scoring customer from this set to be the next customer of station j; otherwise, station j stays idle.

(2) If the set of idle stations $\Psi_i(t)$ is not empty, we assign the highest-scoring station from this set as the next station for customer i; otherwise customer i joins the waiting customers.

Once these assignments are made, the service process continues until the next service completion event.

Observe that under this DSP, if no service completion happens at time t, then for any idle customer i and idle station j, no feasible assignments are possible; i.e., we must have $i \notin \Omega_j(t)$ and $j \notin \Psi_i(t)$.

To describe the scoring rules required to complete the definition of a particular DSP, let $w_i^{TS}(t)$ and $w_i^{TW}(t)$ be customer i's total system time and total waiting time at time t, respectively. Note that $w_i^{TS}(t) = t - r_i^c$ and, for a customer who is idle at time t, $w_i^{TW}(t) = w_i^{TS} - \sum_{j\in S_i^F(t)} x_{ij}$, where $S_i^F(t)$ is the set of all stations already completed by customer i. We also let $w_i(t)$ represent the current waiting time of customer i, i.e., time since her last service completion.

Using our intuition and known results on open-shop scheduling, we next define a number of different DSPs. We consider five different customer scoring rules—their definitions and the intuition behind them are described as follows:

1. The *longest system time first* (*LS*) *policy* assigns to station j the customer in Ω_j who has the longest system time among all waiting customers who still require service from this station, i.e., $PT_i^c(t) = w_i^{TS}(t)$. This rule is motivated by the idea that the customer who has already accumulated a long system time (because of waits or long processing times) is more likely to be tardy, and thus should be prioritized.

2. The *longest accumulated waiting time first* (*LAW*) *policy* assigns the waiting customer in Ω_j who has accumulated the longest waiting time, i.e., $PT_i^c(t) = w_i^{TW}(t)$. This policy is also motivated by prioritizing customers who have a higher risk of being tardy. By counting only waiting time we are giving preference to customers who have already been "victimized" by long waits.

3. The *longest current waiting time first* (*LCW*) *policy* assigns the waiting customer in Ω_j who has the longest current waiting time, i.e., $PT_i^c(t) = w_i(t)$. This policy follows the spirit of first-come-first-serve policy and prioritizes customers who enter the centralized waiting room earlier.

4. The *longest mean overage processing time first* (*LMOP*) *policy* assigns to station j the waiting customer in Ω_j who has the longest mean overage service time, i.e., $PT_i^c(t) = (1/|S_i^F(t)|)\sum_{j\in S_i^F(t)}(x_{ij} - \bar{s}_j)$. Similar to the LS policy, the LMOP policy prioritizes the customer who experienced longer than usual service times (represented by a longer mean overage service time). This customer thus has a higher risk of being tardy.

5. The *shortest expected remaining processing time first* (*SERP*) *policy* assigns the waiting customer in Ω_j who has the shortest total expected remaining processing time, i.e., $PT_i^c(t) = (\sum_{j\in S_i - S_i^F(t)} \bar{s}_j)^{-1}$. This policy is motivated by the optimality of preemptive SERP policy in a two-station open shop with the total expected completion time objective (see, e.g., Pinedo 1984).

We also considered several station scoring rules. Since stations with the higher remaining average workload at time t can be thought of as bottlenecks over the remainder of the process, perhaps the most natural rule is to perform customer assignment so as not to keep more valuable resources idle. Let n_j be the number of servers at station j, and let $u_j(t)$ be the number of customers who still require service from station j. Then a natural station scoring rule is $PT_j^s(t) = (u_j(t)\bar{s}_j)/n_j$, which measures the *remaining expected workload* (*REW*) of station j. Intuitively, this scoring rule prioritizes stations that are likely to become bottlenecks over the remainder of the process. Indeed, this rule demonstrated good and robust performance in our computational tests reported in Section 5.

Of course, many alternative station scoring rules can be developed. The general rule in scheduling is to avoid idling bottleneck resources. However, a "bottleneck station" may not be well defined in an open shop environment; as discussed in Section 4, we prefer the term "problematic station." If some prior data for the process are available, they can be used to prioritize such stations; we call this the *problematic station priority* (*PSP*) rule. Under the PSP rule, a problematic station is prioritized over a nonproblematic one; in case more than one assignment is possible under this rule, ties are broken using the REW rule above.

Note that one may attempt to consider service time distributions of different stations in DSPs. However, to the best of our knowledge, there are no scheduling policies in the literature that incorporate distributional information.

Again, many more rules, for both customer and station scoring, could be defined. Our goal is not to perform an exhaustive evaluation of every reasonable scoring rule, but rather to define a framework that can accommodate many different DSP types.

3.3. Strategic Idleness Modification—Generalized TBP

Recall that in addition to the two macrolevel SLMs, we are also interested in the microlevel SLM: the total number of RFs (incidents of long waits). Since such incidents often occur at bottleneck stations, one strategy to reduce this number is to intentionally delay service at stations that are upstream from the bottlenecks when there is already a long queue in front of the bottleneck station. We call such intentional delays strategic idleness. The key idea is to modify a given work-conserving DSP so that when the DSP assigns a free customer to a free station, instead of starting the service immediately, they both stay idle for a certain time period (possibly zero) prior to the service start. A nonidle policy can be thought of as a special SI modification where the idling times are all zero.

Note that under this modification, the customer–station assignment produced by the DSP remains valid during the SI period; the TBP modification simply introduces a possible delay to the start of the service. This makes the modification easy to add to any DSP.

Note that under SI, both customer i and station j are idled, which may lead to a waste of resource capacity. Indeed, it is possible that while customer i is idled, station j may serve some other customer k, allowing customer k to "overtake" customer i. While this "overtaking TBP" modification may have intuitive appeal, it introduces additional complexity as it "overrules" the customer–station assignment selected by the DSP (indeed, at the next service completion event, customer i may no longer be assigned to station j). Thus, instead of first selecting a well-performing DSP and then finding a suitable TBP modification for it (essentially treating DSP and TBP as separate components, as per our overtake-free framework defined above), one has to optimize the combined DSP/TBP policy, which may be quite difficult computationally. We also note that the "overtaking" version is less flexible with respect to the exact amount of strategic delay to be introduced for customer i—while under the "overtake-free" version the service for customer i starts as soon as the delay conditions are lifted, this may not be the case in the overtaking version: customer i may wait until (at least) the next service completion at station j. Our extensive computational experiments reported in Table 6 (see Section 5.1) reveal no performance advantages for the overtaking option.

SI policies can be defined in a variety of ways. We focus on the family of TBPs introduced by Baron et al. (2014) in the context of a two-station tandem queue. The TBP idles the upstream station whenever the difference between the queue lengths of the upstream and downstream stations reaches a threshold $TH \geq 0$.

Clearly this definition needs to be generalized to the open-shop setting we are considering where each customer may take a unique path through the network and the notions of upstream and downstream stations are not available. To extend the definition of the TBP to this setting, we note that two components are required to define a particular TBP policy: the differencing rule δ_{ij}, which evaluates the state of the current queue at station j with respect to customer i who has just been assigned to this station, and the station-specific threshold $TH_j \geq 0$, leading to the following definition:

Definition 2 (Modified Policy DSP + TBP). When customer i is assigned to station j under a certain DSP, the service starting time is delayed as long as $\delta_{ij}(t) \geq TH_j$.

Recall that $u_k(t)$ is the number of customers who still need service from station k. We let the differencing rule $\delta_{ij}(t)$ be a nondecreasing function of $u_k(t)$ for $k \neq j$ and a strictly decreasing function of $u_j(t)$. Moreover, we assume that $\delta_{ij}(t) = 0$ when all queues are empty, i.e., $u_k(t) = 0$ for all k. Moreover, we also allow $\delta_{ij}(t)$ to depend on the set $S_i^U(t) = S_i - S_i^F(t)$ of stations still required by customer i at time t. While the threshold value can also be made customer specific, this substantially complicates the policy. Thus, we assume the threshold TH_j is defined at the station level and applies to all customers assigned to this station.

Definition 2 above requires a specification of the functions δ_{ij} and the thresholds TH_j, though the latter can be treated as a model parameters, with the best values chosen during the model calibration stage. We thus equate a particular TBP with the specification of the differencing function δ_{ij}. The following specification will be used in the numerical experiments in Section 5:

Maximum workload TBP (*MW-TBP*). $\delta_{ij}(t)$ is the difference between the number who still need service from the station with the largest number of unfinished customers and is still required by customer i and the number of customers who still need service from station j, i.e., $\delta_{ij}(t) = \max_{k \in S_i^U(t)} u_k(t) - u_j(t)$.

A natural modification of the "eligibility" version of the MW-TBP where $u_k(t)$ and $u_j(t)$ are redefined to refer to only customers who could actually use these stations at time t (i.e., who have completed all stations required by the precedence constraints). This version may be particularly appropriate in systems with many precedence constraints.

Another option is to normalize the MW-TBP to account for the service rate and the number of servers at station k, which leads to $\delta_{ij}(t) = \max_{k \in S_i^U(t)} (u_k(t)\bar{s}_k)/n_k - (u_j(t)\bar{s}_j)/n_j$. This *normalized maximum workload TBP* has the advantage of taking into account the differences in processing rates and manpower of stations j and k, and intuitively they should perform better than the MW-TBP. However, they complicate the search for the optimal value of the threshold TH_j: while for the original MW-TBP the differencing function, $\delta_{ij}(t)$, can only take integer values, this is no longer the case for the normalized MW-TBP, and thus the search for the best value TH_j is more computationally involved. We tested the normalized MW-TBP and searched for the optimal TH in $\{0, 1/2, 1, 3/2, 2, \ldots\}$, but the optimal policy does not perform better than the original simpler TBP. We conjecture that searching over a denser set of TH values will identify a better policy. However, the search effort grows exponentially with the performance improvement.

One other alternative that should be mentioned is the *maximum workload kanban TBP*: the number of customers who need service from the station other than j that is still required by customer i and has the most unfinished customers, i.e., $\delta_{ij}(t) = \max_{\{k \in S_i^u(t) \setminus j\}} u_k(t)$. While this can be viewed as a simplified version of the MW-TBP, it shows the similarity of our approach to SI and the classical kanban-based methods in assembly-line scheduling.

The effectiveness of our DSP + MW-TBP approach will be illustrated in the context of the XYZ study in Section 5.3 and for random networks in Online Appendix A. We next describe the case study which served as the inspiration for the current paper.

4. Case Study—Preventative Healthcare Assessment at XYZ

As discussed in the introduction, preventative healthcare assessment (PHA) is one of the main services offered by XYZ. This service consists of up to 21 different medical tests, including blood and urine lab tests, chest X-ray, abdominal ultrasound, a fitness test, a treadmill test, a physician exam, an audiovisual (AV) test, and nutrition. Once all the tests are completed, the results are reviewed with the doctor, who obtains a comprehensive snapshot of the customer's state of health.

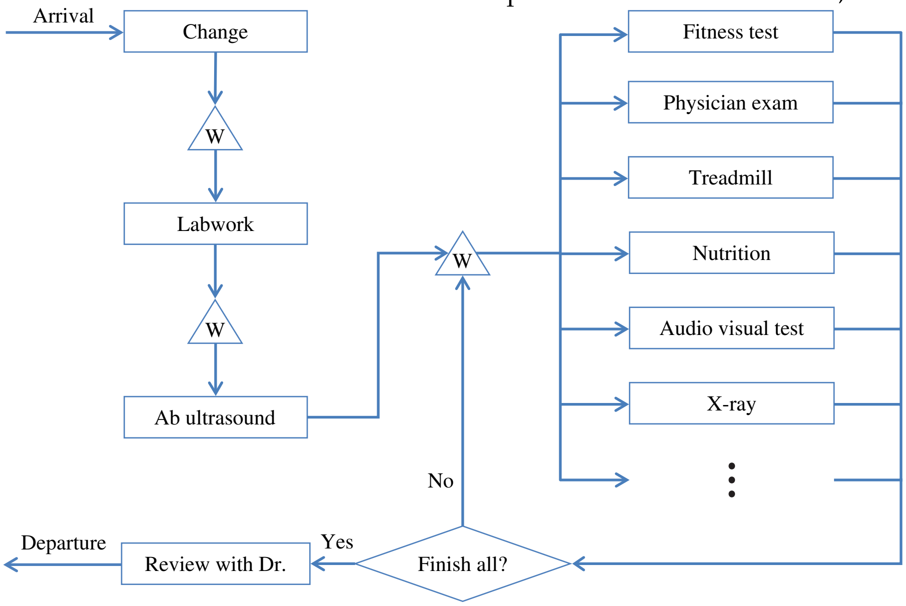
While the tests above are required by almost every customer, there are add-on assessments that can also be requested (such as optometry, echocardiogram, genetic risk assessment, etc.). Some customers may choose to opt out of some standard PHA tests (e.g., chest X-ray may be skipped if recent results are available).

The main PHA tests are performed at nine different stations. Each station handles only one test, except for doctors who need to perform both the physician exam and review. The typical PHA process flow is depicted in Figure 1. Arriving customers first change into gym clothes and then proceed to the lab (for blood and urine samples), followed by abdominal ultrasound. These procedures are scheduled first since they have to be done on an empty stomach. Once these tests are completed, the remaining stations, except for the doctor review, can be visited in arbitrary order. After visiting each station, the customer is brought back to a central waiting lounge. On average, each customer visits 10 stations, including all nine routine diagnostic tests and one add-on. The overall process is controlled by process schedulers (usually two are on duty at any given time) based on their experience, rather than some automated techniques.

Each test requires exactly one server per customer. Most stations have multiple (up to eight) servers, allowing them to process several customers in parallel. In general, customers are indifferent with respect to which server performs a particular test. The exceptions are the physician exam and doctor review, both of which are performed by the specific doctor the customers are preassigned to. Thus, even though multiple doctors may be available, the customer needs a specific doctor for these stations.

It should be noted that the process flow in Figure 1 is what XYZ management would *ideally* like to see for each customer, as it ensures that the doctor has all results at hand before the review. However, for a variety of reasons, mostly concerned with minimizing customer waits and using the available resources (especially doctors) efficiently, the actual flows may depart from this ideal. In fact, in our observed data,

Figure 1. (Color online) Typical Service Order at XYZ



Note. The three triangles marked with "W" represent the same centralized waiting room.

just under 50% of customers actually had the doctor review as their last station.

The company's goals are to keep the average system time given in (1) under four hours, to have at least 50% of customers complete their PHA in four hours or less (i.e., in (2), $T^s = 4$ hours), and to minimize the number of RF incidents defined as a wait exceeding 20 minutes between any two sequential tests (in (3), $T^{RF} = 20$ minutes). In fact, all RF incidents are regarded as significant service failures and are carefully tracked. The IT system used by the process schedulers records departures and arrivals at each station and keeps track of customer waiting times. When a customer's waiting time reaches 15 and 20 minutes, yellow and red faces appear, respectively, on the schedulers' screen. The number of red faces that occur during each day is tracked and used as part of performance reviews for process schedulers and their supervisors.

We obtained two months of data from XYZ. For each customer visit, the data contain the customer's appointment time, actual arrival time, and departure time and the tests performed during the PHA. For each specific test and each customer, the data also contain the starting, ending, and service times. During the two months for which data were collected, there were 41 business days, with just over 2,000 total customer visits and about 20,000 tests performed. The number of customers who visited the clinic each business day ranged from 25 to 61 (with a mean of 49 and a standard deviation of 7.2).

XYZ schedules the arrivals of its customers at different times throughout the morning of each day. The clinic opens at 7:00 A.M. and closes when all customers leave, typically around 16:00 P.M. Figure 2 illustrates the histograms of average daily scheduled/realized arrivals alongside the histogram of average daily departures. Note that the resulting arrival pattern is not stationary (contrary to the assumptions of traditional queueing models). On average, XYZ schedules five customers every half hour starting from 7:00 A.M. until (typically) 10:30 A.M. Between 10:30 A.M. and 12:00 P.M., the number of scheduled arrivals is gradually reduced. Most customers leave the clinic after 11:00 A.M., at a rate of four every half hour, until around 16:00 P.M. The clinic is busier after 11:00 A.M. compared to before 11:00 A.M., and 400 out of 456 total RFs occur after 11:00 A.M. in the empirical data.

4.1. Performance Analysis

In this section, we analyze the performance of XYZ's open-shop network, focusing primarily on the nine routine stations (the ones listed in Figure 1 except for "Change"). To calculate the utilization of each station, we first focus on each of its servers and calculate the average starting time (when the first customer enters service) and the average closing time (when the last customer leaves). Then, for each station, we derive the average time span (the time between starting and closing) and the average busy time. Each station's utilization is obtained as the ratio of its average busy time and its average time span. The number of servers, mean service time (averaged among servers at the same station) and coefficient of variation (CV), and utilization ratios are given in Table 1 in the second, third, fourth, and fifth columns, respectively. The stations are sorted by the utilization ratio.

We next want to compute waiting times and the number of RFs associated with each station—this will indicate the bottlenecks in the system. In a serial process, the waits can be unambiguously attributed to the

Figure 2. (Color online) The Histogram of Average Daily Arrivals and Departures

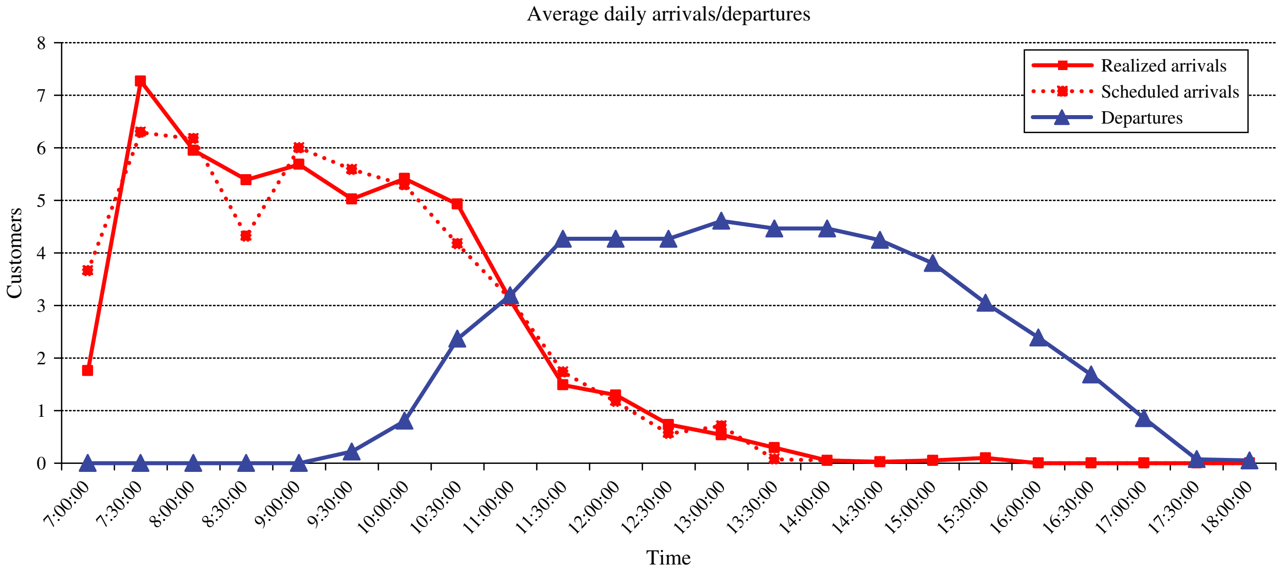


Table 1. Summary of Key Performance Indicators for PHA Stations in the Empirical Data

		Service time			Direct attribution		Proportional attribution	
Stations/Ave.	#Servers	Mean	CV	Utilization (%)	Wait	RFs	Wait	RFs
Fitness test	7.4	20:28	0.35	78	5:46	45	6:40	44.9
Phys. exam	7.7	32:45	0.39	73	4:50	59	4:26	40.8
Doc review		14:36	0.61		7:06	173	15:06	210.7
Treadmill	4	21:50	0.24	72	4:48	20	3:59	12.7
Ab ultrasound	4	16:16	0.36	71	5:23	10	1:31	1.9
Nutrition	3.9	19:49	0.33	69	4:42	18	4:50	11.3
AV test	3.9	16:51	0.32	63	6:57	55	7:03	44.1
X-ray	1	6:14	0.60	50	5:23	6	5:23	12.4
Lab work	1	3:23	0.48	48	4:19	2	0:33	1.0
Ave. system time	4:04:26							
System time ≥ 4 hrs	52.5%							
Ave. total wait time	1:01:08							
Total RFs	456							

station the customer entered at the end of the waiting period. However, for an open-shop process, because of the lack of service order, the time customer i spends in the waiting room before entering station A may not accurately reflect the wait for station A: this customer may have several unvisited stations, and station A just happened to be the first station that freed up. Thus, attributing the wait (and an RF if the wait exceeded 20 minutes) to this station may be somewhat misleading. On the other hand, in the long run, this "direct attribution" probably does indicate which stations are busier than others. It may be argued that a more accurate rule is to attribute a certain proportion of this wait to all stations the customer could have entered, since the wait was due to the unavailability of all these stations. However, which proportion should be attributed to what stations is less clear.

We report the results for two attribution rules: the "direct attribution," where the wait is attributed to the station the waiting customer entered next, and the "proportional attribution," where the wait is evenly distributed to all stations the customer was eligible to enter (i.e., if the customer was eligible to enter k stations and incurred a wait of W, then each of these stations would be allocated a waiting time of W/k). The same rules are applied for allocating RF incidents. The results for both rules are presented in the last four columns in Table 1. The lower part of the table also provides the three main SLMs and average total waiting time.

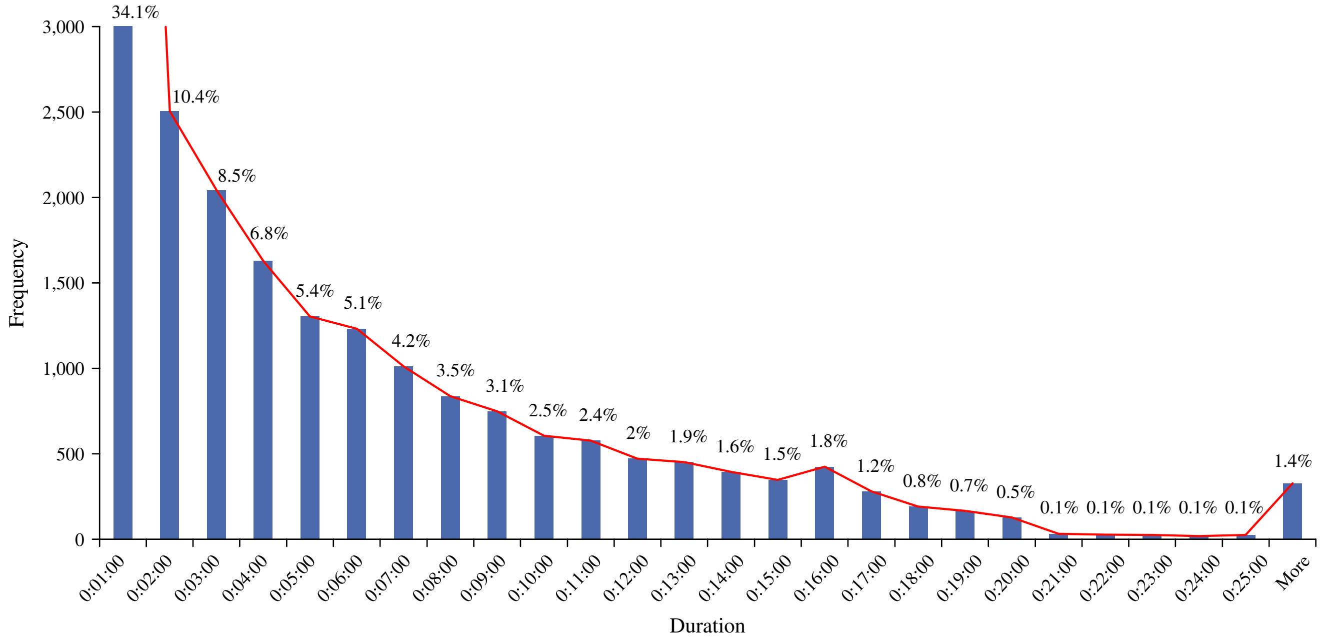
From the summary results at the bottom of Table 1, we observe that the service levels at XYZ are already quite good. The average total system time exceeds the four-hour goal by just over four minutes. Moreover, the average total waiting time is just over one hour, i.e., less than a quarter of the total time a customer spends in the system. The total number of RF incidents is 456. With just under 10 stations per visit, on average, this corresponds to less than 2.5% of all tests and about 25% of customers experiencing a red face. The performance with respect to the total number of tardy customers is reasonably good: 52.5% of customers end up spending more than four hours in the system—not far from the target value of 50% for this SLM.

We also see that the overall utilization rates, between 48% and 78%, are not high. However, the coefficients of variation of different stations' service times are substantial (i.e., mostly over 0.33, up to 0.61). In a service network with significant variability, we expect relatively low utilizations to be required to maintain short waiting times. This confirms the customer-centric emphasis of XYZ.

Figure 3 illustrates the histogram of customers' waiting times between two sequential station visits with the bin interval of one minute. The shape of the histogram follows the classical exponential-like pattern. The mean and standard deviation of the interstation waiting times are five and seven minutes, respectively.

To determine which stations are system bottlenecks (in term of accounting for long wait times and a large number of RFs), the conclusions are largely unaffected by the attribution rule used. From Table 1, the direct attribution rule identifies the review with the doctor, audiovisual test, and fitness test as accounting for the longest waiting times; the proportional attribution identifies the same stations (and in the same order). Both rules indicate that the review with the doctor is undoubtedly the system bottleneck—with the longest average waiting time and the largest number of RFs (38%–47% of all RF incidents, depending on the attribution rule used). The main cause is that a customer's review with a doctor has to be completed by the same doctor who performed the physician exam, resulting in low effective capacity; the service time also has the

Figure 3. (Color online) The Histogram of Customers' Waiting Times Between Two Sequential Station Visits



Note. The label above each bar represents the relative frequency of waiting times.

highest coefficient of variation among all stations, coupled with the second-highest utilization rate.

The fitness test and audiovisual test together generate 89–100 RFs (again, depending on the attribution rule used). A key problem at these two stations is the capacity loss due to late starting times. On average, the fitness test and the audiovisual test start 2.5 and 1.5 hours, respectively, after the clinic opens, with some operators arriving even later.

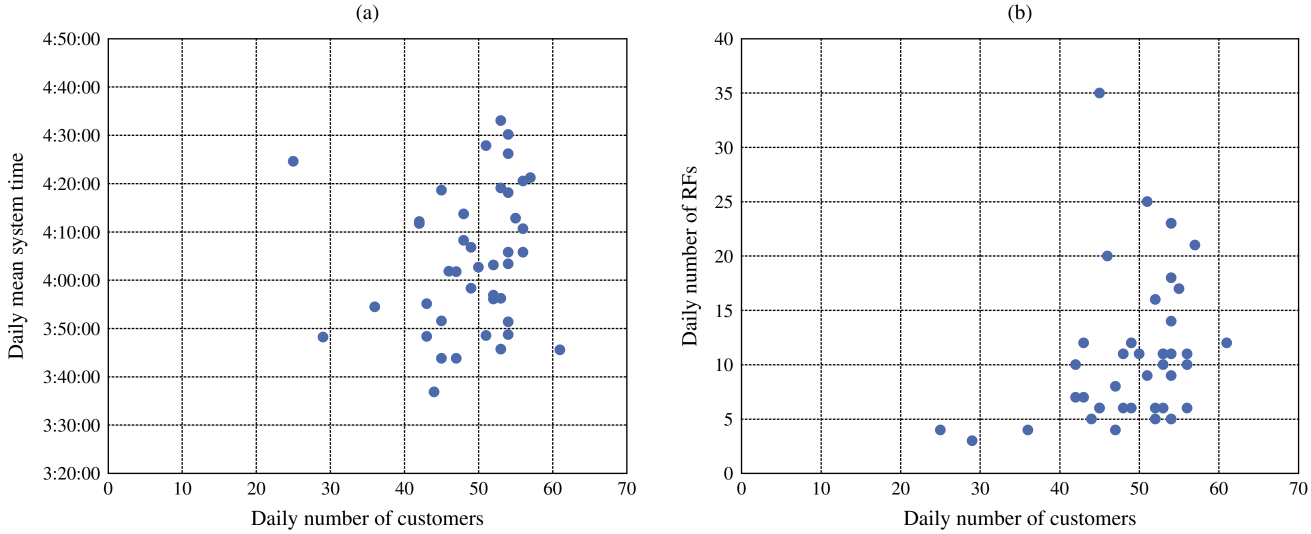
We note that the fitness test and the review with the doctor may be considered bottleneck stations of the network according to the classical definition: they both have high capacity and utilization and long waiting times (under either attribution rule). On the other hand, the audiovisual test, with a relatively low utilization, does not meet the standard definition of a bottleneck. Therefore, for lack of a better name, we call the fitness test, review with doctor, and audiovisual test *problematic* stations and other stations *nonproblematic*.

The staffing level at each PHA station varies on a daily or hourly basis (except X-ray and lab work tests that have only one server), which leads to the fractional average number of servers at some PHA stations, as seen in the second column of Table 1.

One natural question is whether the staffing levels are sufficiently adjusted for the different customer volumes on different days; if not, understaffing may account for days with larger than average system time and large number of RFs. To investigate the impact of daily staffing levels, we compute the correlations between the daily number of servers at each PHA station and (i) daily number of customers, (ii) daily number of RFs, and (iii) daily average system time; these are presented for each station (excluding lab work and X-ray) in Table 2. In addition, in the last row of the table we compute the correlations between the daily number of customers and two SLMs: (iv) the daily average system time and (v) the daily number of RFs. The scatter plots of the data for (iv) and (v) are presented in Figure 4.

First, we observe (from the second column of Table 2) that the correlations between the daily number of servers at each station and the daily number of customers are fairly strong (mostly around 0.70), indicating that XYZ does a fairly good job of adjusting staffing levels to customer volumes. The next two columns of the table indicate that the correlations between staffing levels and both SLMs are quite weak: since there are 41 days in the data, the standard errors corresponding to the correlation values in the last two columns are in the 0.15–0.16 range, and thus none of the observed values are significantly different from 0 at the 95% level. This indicates that below-average staffing levels are unlikely to explain either large number of RFs or long system times in a given day. The values in the last row of Table 2 as well as plots in Figure 4 indicate that the same conclusion applies to customer volumes as well: neither the number of RFs nor the average system time is significantly correlated with the number of customers. Indeed, days with the same number of customers exhibit large variability with respect to both SLMs. This suggests that the system performance with respect to these measures is driven by intraday vari-

Figure 4. (Color online) The Daily Number of Customers vs. (a) the Daily Mean System Time and (b) the Daily Number of RFs



ability rather than inadequate daily staffing or large customer volumes.

We next turn our attention to the variability in service times. The key question here is whether station-level CV's presented in Table 2 are a good representation of the intrinsic variability in service times. It has been observed in many systems that service times tend to shrink when the system gets busy, and thus it is possible that the average measures of variability (such as CVs) are in fact more reflective of the differences between service times during idle and busy periods rather than intrinsic variability of service times.

To analyze whether service times indeed shrink when the clinic is more congested, we formulate and test two statistical hypotheses. First, recalling that the clinic is consistently more busy in the afternoon leads to the following hypothesis.

Hypothesis 1 (Morning vs. Afternoon (H_i^{MA})). *The mean service times for station i tend to be higher before* 11:00 A.M. *versus after* 11:00 A.M.

Table 2. Correlations Between the Daily Number of Servers at Each PHA Station and Daily Number of Customers, Daily Number of RFs, and Daily Average System Time

Daily staffing level at station	#Customers/day	#RFs/day	Ave. system time
Ab. ultrasound	0.69	0.26	−0.07
Treadmill	0.48	0.20	0.14
AV test	0.71	0.33	−0.05
Physician exam	0.72	0.19	−0.02
Review	0.72	0.19	−0.02
Nutrition	0.42	0.14	0.07
Fitness	0.71	0.12	−0.08
#Customers/day	1.00	0.32	0.14

Note. The last row contains correlations between daily customer volumes and corresponding columns.

Second, we say that a congested state (CS) at a station exists when there is at least one patient in the waiting room who still needs service from this station. This leads to the following hypothesis.

Hypothesis 2 (Congested State (H_i^{CS})). *The mean service times for station i tend to be higher under the CS versus the non-CS condition.*

The comparison of mean service times and the p-values for the respective statistical tests are provided in Table 3. We observe that the mean service times are generally *longer* under the supposedly more congested conditions represented by the two hypotheses above, leading to insignificant p-values for most stations. In fact, the only station for which H^{MA} may be plausible (p-value of 0.003) is the treadmill station, where the afternoon mean service time is slightly shorter than its morning counterpart (22:09 vs. 21:29). Similarly, there are only two stations (nutrition and fitness) where the service times under CS condition is smaller than under the non-CS conditions; in both cases the differences in service times are less than 1.5 minutes. Note that we can generalize the CS definition to when there are j, $j = 1, 2, 3$, patients in the waiting room who still need to visit this station, where the cases $j = 1, 2, 3$ correspond to the 49.96, 70.06, and 83.68 percentiles of the queue length distribution in the system; however, the conclusion stays the same for different j's.

We conclude that there is little evidence that service times are consistently reduced when the process is more congested. Note that this conclusion is fully in line with the company's customer-focused philosophy and our own process observations: since each customer pays a substantial amount of money for the service, they are unlikely to be treated any differently during any station visits—irrespective of the congestion level of the process.

Table 3. Results of Applying Standard t-Tests to H^{MA} and H^{CS}

	H^{MA}		H^{CS}	
PHA station	Mean (morning, afternoon)	p-value (one-sided)	Mean (no one waiting, at least one patient waiting)	p-value (one-sided)
X-ray	(06:03, 06:25)	0.968	(06:23, 06:09)	0.129
Ab. ultrasound	(16:21, 15:58)	0.138	(16:10, 16:20)	0.738
Lab work	(03:25, 03:15)	0.090	(03:22, 03:27)	0.861
Treadmill	(22:09, 21:29)	0.003	(21:56, 21:48)	0.308
AV test	(16:54, 16:48)	0.340	(16:35, 16:56)	0.886
Physician exam	(30:52, 35:44)	1.000	(32:37, 32:49)	0.628
Review	(12:59, 15:05)	1.000	(15:12, 14:27)	0.108
Nutrition	(19:25, 20:13)	0.995	(20:30, 19:33)	0.004
Fitness	(20:19, 20:35)	0.787	(21:24, 20:11)	0.001

While the results discussed above do not support the hypothesis that schedulers expedite service by reducing station service time, it is clear that some form of response from the schedulers is expected, as they receive yellow face warnings. One form of response is departing from the ideal process depicted in Figure 1, e.g., routing the customers to the doctor review station before all tests have been completed (as noted earlier, this occurs for over 50% of all patients). As discussed in the following section, it appears that another more subtle expediting strategy is also used—based on introducing strategic idleness and waits in the system.

4.2. Overlapped Waits: Evidence of Use of SI in Practice?

From the data provided by XYZ, we discovered a significant number of instances where a customer was waiting for a station that was free and waiting for this customer, i.e., a certain part of a customer's waiting time, was spent waiting for an ostensibly idle station; we call such periods overlapped waits (OWs). Note that each OW has a specific station associated with it, so the issue of attribution does not arise here. OWs occur in 78.7% of service instances and are thus too prevalent to be data errors. The average OW is 2.5 minutes, with a standard deviation of 4.5 minutes (note that a CV of 1.8 is quite large). While about 50% of the OWs are less than one minute, approximately 16.3% are longer than five minutes, i.e., longer than the average wait time per station.

There are two alternative explanations for OWs: they could result from regular service-supporting (SS) activities or they could be due to the *deliberate* use of SI by process schedulers. We discuss each of these alternatives in more detail below and develop statistical hypotheses that are then tested with the data. Of course, both SS and SI factors may be present, but we are interested in uncovering the primary driver of OWs.

A variety of regular SS activities could account for OWs. The most likely culprit may be setups: if the data system records the station as idle when the previous patient departs, the various setup steps (cleaning the room, changing the sheets, etc.) that are necessary before the new patient can be processed would be recorded as idle time for both the station and the patient that is waiting, thus resulting in an OW. Some other random factors may also lead to OWs, e.g., the station may be ready for the patient but the patient is finishing a meal or a phone call and thus the service cannot start immediately, the patient is ready but the server is occupied with a phone call or a question related to previous service, etc.

An alternative explanation is that OWs are generated by process schedulers intentionally delaying service starts under certain conditions to avoid RFs. Baron et al. (2014) demonstrate analytically for a two-station tandem queue network that a scheduling policy incorporating SI can be very effective in reducing the number of RF incidents. In Section 5 we will show that adding SI to DSPs is quite effective in reducing the number of such incidents in XYZ's process as well. The idea behind SI—of idling an upstream station when there is already a long queue in front of a downstream station—is quite intuitive. However, we are not aware of any empirical evidence of the use of SI in practice.

Of course, the most direct way to distinguish between the SS and SI alternatives would be to simply ask the service providers. Unfortunately, no clear answer was received. Process and IT managers were surprised by the presence of OWs and claimed to be unaware of any intentional use of SI. Because of personnel changes and the time gap between data collection and analysis, it was not possible to interview process schedulers present when the data were collected. As demonstrated below, the data seem to provide a fairly clear answer in this case (we note that our approach is in line with *data-driven research*; see, e.g., Simchi-Levi 2014). To this end, we describe a series of conditions that are more plausible under SS or SI alternatives, convert these conditions to statistical hypotheses, and test them using the process data.

We start by discussing conditions that we would expect to hold if the primary driver of OWs was SS activities:

Congested process (CP). OWs should be relatively stable over time or become shorter after 11:00 A.M., when the clinic becomes congested and stations may have pressures to expedite SS activities such as setups. This point should also hold for OWs of each station. On the other hand, if SI is present, one would expect to see more frequent use precisely when the system is more congested, i.e., after 11 A.M.

Near-zero idle times (NZIs). If OWs are largely due to the the IT system recording setups as part of station idle times, there should be a low incidence of zero or near-zero station idle times, i.e., instances where a server goes from treating one customer immediately to treating the next one, with minimal intervening station idle time (and, of course, no OW). We use 10 seconds as the threshold for near-zero idle times. On the other hand, if setups are recorded as part of station service times and the stations are recorded as idle only after setups are completed, near-zero idle times should be quite frequent in the data.

Problematic stations (PSs). Recall that our earlier analysis identified three problematic stations as being responsible for the bulk of RFs. Certainly, intentional idling of such stations under SI is (usually) not helpful. Therefore, if OWs are primarily due to SI, then there should be a significant difference in OWs between problematic stations and nonproblematic stations. If OWs are not due to SI, this difference should be small.

We next turn our attention to conditions consistent with SI. Generally, we would expect to see longer OWs in situations where SI would be most useful in reducing RF incidents; that is, we would expect the occurrence of longer OWs to be strategically timed. For example, when station A is ready to serve customer i and customer i's next station is station B, which is congested, the process scheduler could use SI to balance customer i's waiting time at both stations. The congestion at station B can be indicated either by that station B is one of the problematic stations or that station B's previous customer (served just before customer i) experienced a long waiting time (e.g., ≥15 minutes) before entering station B. In the former case, we say that station A precedes a problematic station, while in the latter case we say that it precedes a "potentially long wait." Note that while the set of "problematic stations" is fixed based on the analysis of long-term trends in the process, the set of "stations with potentially long waits" is dynamic, as it is related to the current state of the system. The discussion above leads to the following conditions as being consistent with the use of SI to avoid long waits:

Preceding problematic stations (PPS). OWs at stations preceding problematic stations are longer than OWs at stations preceding nonproblematic stations.

Preceding long waits (PLW). OWs at stations preceding stations with potentially long waits are longer than OWs at other stations.

To summarize, we have identified seven testable conditions that should be indicative of the prevalence of SS activities versus SI as the primary cause of OWs. These conditions, the corresponding null hypotheses, and the results of the statistical tests are summarized in Table 4. Some related station-specific results are presented in Table 5. The results indicate that in all five cases, the statistical tests are consistent with the SI hypothesis. Below, we comment on individual results in this table.

The CP condition is tested with a t-test, which rejects the null hypothesis at a high significance level: the OWs after 11 A.M. are longer than during the less congested pre-11 A.M. period, which is more consistent with the SI hypothesis.

The null hypothesis for the NZI condition is similarly rejected: as can be seen from Table 5 (last column), near-zero idle times are quite common for most stations. Even for X-ray, the station with the minimal

Table 4. Data-Based Evidence for SS vs. SI Origins of OWs

		Agrees with		
Condition	Null hypothesis	SS	SI	Result (comments)
CP	[Mean OWs before 11 A.M.] ≥ [Mean OW after 11 A.M.]	Y	N	H_0 rejected with $p < 0.0001$; mean OWs are 1:50 vs. 3:28
NZIs	Instances with (idle time + OW) = 0 or ≤10 sec. are rare	Y	N	Rejected; such instances are common (see Table 5)
PSs	No difference in mean OWs between problematic and nonproblematic stations	Y	N	H_0 rejected with $p < 0.0001$, mean OWs are 2:13 vs. 2:42
PPS	No difference in mean OWs between stations preceding problematic and all others	Y	N	H_0 rejected with $p < 0.0001$; mean OWs are 2:55 vs. 2:23
PLW	No difference in mean OWs between stations preceding "long waits" and all others	Y	N	H_0 rejected with $p < 0.0001$; mean OWs are 3:38 vs. 2:09

Table 5. Distributions of Service Times, Station Idle Times, and OWs

	Service time		OW				Station idle time		
Station	Mean	CV	Mean	CV	CI(95%)$_{\text{min}}$	CI(95%)$_{\text{max}}$	Mean	CV	$P(\leq 10$ s) (%)
X-ray	0:06:14	0.60	0:04:19	0.90	0.82	0.99	0:17:21	1.08	1.82
Ab ultrasound	0:16:16	0.36	0:01:44	1.80	1.53	2.25	0:07:09	1.74	26.25
Lab work	0:03:23	0.48	0:02:26	1.01	0.93	1.11	0:10:11	1.55	3.17
Treadmill	0:21:50	0.24	0:02:20	1.50	1.32	1.77	0:08:42	1.40	19.24
AV test	0:16:51	0.32	0:03:10	1.35	1.20	1.55	0:09:46	1.66	9.31
Physician exam	0:32:44	0.39	0:01:48	2.42	1.90	3.79	0:08:11	1.61	21.80
Review	0:14:36	0.61	0:02:05	1.80	1.52	2.29	0:10:05	1.74	16.03
Nutrition	0:19:49	0.33	0:02:13	1.45	1.27	1.71	0:09:29	1.54	10.76
Fitness	0:20:28	0.35	0:01:46	2.20	1.77	3.13	0:08:46	3.84	24.14
Total	0:17:09	0.63	0:02:25	1.55	1.36	1.84	0:09:46	1.88	18.06

proportion of near-zero idle times, the standard error is around 0.3%, and thus the observed proportion is more that five standard errors away from zero. It should also be noted that the proportion of near-zero idle times among the three problematic stations is above 10% in all cases. Overall, these results are not supportive of setups (and thus SS activities) as being the main cause of OWs.

The null hypotheses for the next three conditions—PS, PPS, and PLW—are all rejected by the t-tests at very high levels of confidence. In all three cases, the data are consistent with SI: the mean OWs are shorter at nonproblematic stations, are longer at stations preceding the problematic stations, and are (substantially) longer at stations preceding potentially long waits. As discussed earlier, this is exactly what we would expect to see if SI was strategically used by process planners to minimize RF incidents.

To summarize, the results of the statistical tests discussed above are quite unambiguous, indicating that OWs are related to SI and are unlikely to be explained by regular service-supporting activities. Our simulation results presented in the following section provide further support for the effectiveness of SI in controlling the number of RF incidents at XYZ. To the best of our knowledge, the empirical results discussed above are the first evidence of the use of SI in practice.

To close this section, we note that OWs may also explain the two sudden changes around 15 and 20 minutes on the histogram of waiting times in Figure 3. Observe that there appears to be an increase in frequency of waits between 15 and 16 minutes and a decrease between 20 and 21 minutes versus the regular pattern observed at other waiting times. A possible explanation is that when a customer is ready to visit an idle server, schedulers may not initiate the service immediately. Instead, if the conditions for SI are present, they may let the customer and the server wait until the customer's waiting time reaches 15 minutes (when a yellow face alarm is generated), but not 20 minutes (which triggers the RF). Indeed, the empirical data show that nearly all (specifically, 96.4%) of 15 to 20 minute waits are associated with OWs, with the average OW being 6:57 (much higher than the mean). On the other hand, only 77.4% of the waiting times of less than 15 minutes are associated with OWs, and the average OW is only 2:05.

5. Evaluation of Dynamic Scheduling Policies and Strategic Idleness

Most real-life service systems, including the one operated by XYZ, are inherently transient: each day starts with an empty system and ends after processing 49 customers on average; it is not clear whether the steady-state regime is ever reached. Since analytical methods run into significant difficulties when analyzing transient behavior, we use simulation models to evaluate the performance of different policies described in Sections 3.2 and 3.3.

We develop two simulation models. In the first one, representing XYZ's system, we keep the available resources (station availabilities and opening times), customers' arrival times, service needs (i.e., the set of stations visited by this customer), and service times at different stations the same as in the empirical data and only change the scheduling policy. The results are described in Sections 5.1–5.3.

The second simulation model is designed to investigate the performance of our DSPs and the effect of using SI through DSP + TBP modification in a broader context. To this end, we model a general open-shop system with randomly generated resource availabilities and customer arrival and service times. We analyze policy performance for both the transient- and steady-state cases. These results are described in Online Appendix A.

5.1. XYZ Simulation: Effect of Overlapped Waits

The objective of the first simulation model is to represent XYZ's system as closely as possible. As noted earlier, we keep stations, the number of servers, opening and closing times as well as customers' arrival times,

Table 6. Performance of Different Scheduling Policies With and Without SI

	System time			Long wait incidents				
Policies\Measures	Mean	Std. dev.	≥4 hrs (%)	#(≥15 mins)	#(≥ 20 mins) = #RFs	Ave. #RFs /RF-Cstmr.	$E[W \mid RF]$ $\pm 2\,Std[W \mid RF]$	#End with doc review
1. Empirical data	4:04:26	1:00:51	52.5	1,645	456	1.09	35:37 ± 1:54	898
2. ED + Nonidle	3:46:37	56:53	38.3	1,846	1,094	1.25	29:48 ± 0:37	898
3. LAW + Nonidle	3:23:20	49:44	20.2	931	700	1.09	35:25 ± 0:59	1,155
4. LS + Nonidle	3:23:01	47:04	19.2	822	615	1.09	39:53 ± 1:36	1,140
5. LMOP + Nonidle	3:23:58	51:49	20.2	762	574	1.09	46:41 ± 2:48	1,127
6. LCW + Nonidle	3:23:26	49:44	20.4	939	706	1.09	35:22 ± 0:58	1,138
7. SERP + Nonidle	3:22:36	50:41	20.1	877	666	1.09	35:39 ± 1:08	1,140
8. LAW + Overtake-free TBP	3:31:41	51:53	26.3	747	500	1.11	33:11 ± 1:06	1,071
9. LS + Overtake-free TBP	3:34:25	49:59	27.4	725	460	1.07	36:59 ± 1:35	1,078
10. LMOP + Overtake-free TBP	3:32:12	52:54	26.2	632	442	1.11	42:06 ± 2:53	1,042
11. LCW + Overtake-free TBP	3:34:45	52:36	29.1	777	467	1.07	31:45 ± 1:03	1,082
12. SERP + Overtake-free TBP	3:31:36	52:18	26.5	729	472	1.11	34:03 ± 1:12	1,067
13. LAW + Overtaking TBP	3:30:45	51:25	37.3	702	478	1.08	32:46 ± 1:05	1,014
14. LS + Overtaking TBP	3:44:26	55:14	37.8	697	476	1.08	34:28 ± 1:24	1,027
15. LMOP + Overtaking TBP	3:44:18	56:53	36.0	654	443	1.13	40:30 ± 2:18	1,046
16. LCW + Overtaking TBP	3:42:20	54:42	35.6	739	497	1.09	32:30 ± 1:04	1,069
17. SERP + Overtaking TBP	3:43:58	56:11	37.5	685	453	1.07	33:17 ± 1:10	1,026

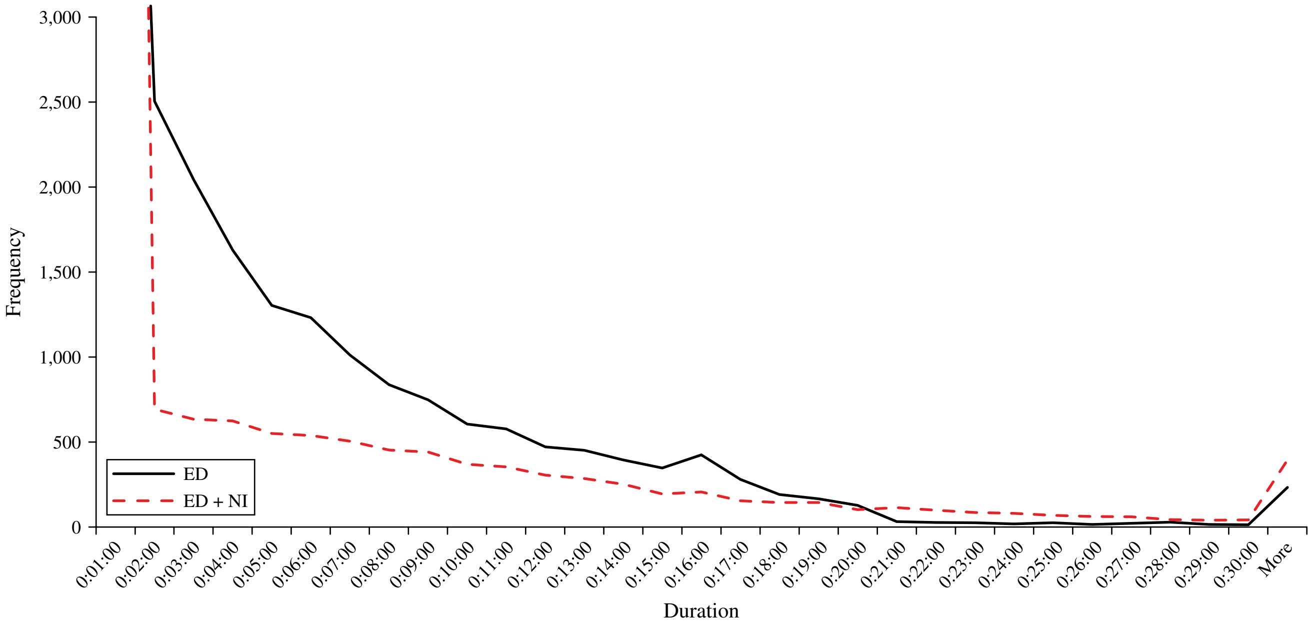
and service needs (i.e., the set of stations visited by this customer) as per empirical data. We also use the actual values for each customer's service times at different stations (recall that customers' service times do not include OWs). Thus, if the schedule (sequence of visits to various stations by each customer) and the OWs before services are also kept the same as in the empirical data, we simply recreate the values of our three SLMs in the empirical data. We refer to this schedule as the empirical data (ED) policy. The relevant values are provided in row 1 of Table 6.

To investigate the impact of OWs, we define the "nonidle" version of the ED policy by eliminating the OWs while keeping the service schedule completely the same as under the ED policy. For example, suppose customer i finishes service at station A. From the data, we observe that the next station visited by customer i is station B. If station B is idle at the moment and customer i is its next customer (based on the data), it immediately starts serving customer i (eliminating any OW that may have been present in the data). Otherwise, if station B is still busy (e.g., because of the removal of OW when customer i visits station A, customer i finishes service earlier than in the data), customer i waits until station B becomes idle, despite that there might be other idle stations customer i has not visited. Note that in the absence of OWs, the service schedule can potentially be different: customer i can visit other stations when station B is not available. However, we intentionally enforce the service schedule to highlight the effect of OWs on different service-level measures. We follow a similar rule when choosing the next customer for station A. We call this the "ED + Nonidle" policy because it exactly replicates the ED policy except for the removal of OWs. The corresponding results are presented in in row 2 of Table 6.

We observe that had the clinic been operating under the ED + Nonidle policy during the two month in our data, the average total system time would have dropped by 18 minutes to 3:46:37, and the proportion of customers with system time of over four hours would have decreased by 14.3% to 38.3%. Thus, both of the macrolevel SLMs would have improved substantially (though we will soon see that these improvements fall far short of those observed under different DSPs). However, the total number of RFs would increase to 1,094 (140% increase). Note that eliminating OWs can both decrease the number of RFs (by making stations available earlier) and increase their number (by forcing customers to join longer queues earlier). Indeed, out of the 456 red faces observed in the data under the ED policy, 217 disappear under the ED + Nonidle policy, while 855 new ones emerge. Shorter waiting times at previous stations, i.e., elimination of SI in the form of OWs, causes 705 of these new red faces.

The mean and standard deviation of the waiting time from the simulation result for the ED + Nonidle policy are 3.5 minutes and 7.5 minutes, respectively. Figure 5 displays the histograms of waiting time for the ED + Nonidle and ED policies. We see that the two frequency curves cross at around 20 minutes, providing yet another indication that strategic idleness (manifesting itself in the form of OWs in the data) is likely used to target the RF threshold of 20 minutes. It is also interesting to note that the histogram of waiting time for ED + Nonidle policy is much smoother; the odd jumps at $t = 16$ or 21 minutes observed for ED disappear,

Figure 5. (Color online) Histogram of Customers' Waiting Times Between Two Sequential Tests Under Different Scheduling Policies



indicating that interventions in the form of OWs may have caused these jumps. These observations further support the conclusions from our statistical hypothesis tests that XYZ's schedulers are indeed using SI.

5.2. Performance of Work-Conserving Dynamic Scheduling Policies

To analyze the performance of the various dynamic scheduling policies described in Section 3.2, we first simulate their work-conserving versions, i.e., without using any strategic idleness. As before, the station availabilities, customer requirements, and service times are taken from the empirical data. However, the order of service is changed according to each DSP. The results are presented in rows 3–7 of Table 6.

The macrolevel SLMs are reported in the second, third, and fourth columns. We observe that the differences among our six DSPs with respect to these measures are not very large. The average system times are all around 3 hours and 23 minutes. Given the standard deviation of around 50 minutes and that there are 2,012 customers in our data, the standard error is around 1.2 minutes, indicating that none of the differences with respect to system time are statistically significant.

Similarly, the proportion of customers with a system time of over four hours varies in a relatively narrow range, from a low of 19.2% for the LS policy to a high of 20.4% for the LCW policy. For this SLM, the standard error can be estimated as $0.2(1-0.2)/sqrt(2{,}012) = 0.35\%$; thus, differences on the order of 1% are statistically significant (though they may be less significant in practice).

The difference in macrolevel SLMs between our DSPs and the actual schedule (i.e., the ED policy) are quite large: DSPs reduce the average system time by about 41 minutes (approximately 17% reduction) and the proportion of customers over four hours by up to 33.3% (a reduction of nearly 63%). These are more than twice the improvements observed earlier for the ED + Nonidle policy.

It is also interesting to note that the DSPs outperform the actual schedule with respect to another measure: the number of customers who end their visits with the doctor review station. The relevant figures can be found in the last column of Table 6. While this is not an official SLM, it is clearly an important indicator: one of the key selling points of XYZ's service is that the doctor has a 360° view of the customer's current state of health during the review; this is clearly compromised if some of the stations have to be scheduled after the doctor review. (The desirable process flow is also reflected in Figure 1.) The ED policy achieves this desired outcome for only 898 out of 2,012 (45%) customers. The scheduling rules for our DSPs also allow some stations to be scheduled after the doctor review (when the latter is busy), but this flexibility is clearly invoked less frequently than under the ED policy: under the LS policy, 1,140 out of 2,012 customers had the doctor review as their last station, representing a 27% improvement. (The results for other DSPs are similar.)

However, when compared with the ED policy, all DSPs have a substantially higher incidence of RFs (sixth column of Table 6). Even the best-performing LMOP policy experiences an increase in this measure of 27% (574 versus 456 for the ED policy)—a clearly undesirable outcome. The DSPs also exhibit a wide range of performance on this measure: the number of RFs ranges from 574 to 706. For the measure of a customer's expected number of RFs given she has at least one RF (seventh column of Table 6), all DSPs perform the same as the ED policy; i.e., under these DSPs, a customer will not get more RFs because she already has one. Furthermore, in the eighth column of Table 6, we focus on those waits that lead to RFs and report the 95% confidence interval of such waits, i.e., $E[W \mid RF] \pm 2\,Std[W \mid RF]$. The result shows that the LAW, LCW, and SERP policies perform better than ED policy (shorter or similar mean and standard deviation), while the LS and LMOP policies lead to higher means for waits more than 20 minutes. Note that the LS and LMOP policies do not take into account customers' past waiting times, so customers who already wait long are not prioritized. It is interesting to note that all DSPs perform better than the ED policy with respect to the incidence of shorter 15 minute waits (see the fifth column of Table 6)—the number of such incidents is about two times higher under the ED policy. This further confirms that the current policy concentrates on reducing the incidents of specific long waits (20 minutes or longer).

We conclude that our work-conserving DSPs outperform the ED policy with respect to all macrolevel SLMs, but not with respect to the microlevel SLM. As discussed in the previous sections, it appears that the ED policy controls the latter measure using SI. In the next section we discuss the performance of our DSPs employing SI (via the TBP modification).

5.3. Performance of DSPs with Strategic Idleness Modification in XYZ's Service Network

To analyze the impact of adding SI to our DSPs, we developed a TBP modification of each DSP using the MW-TBP, as described in Section 3.3. Note that the TBP modification with different thresholds leads to different performances. We therefore use the threshold that minimizes the number of RFs in each policy. The results are reported in rows 8–12 of Table 6 for the overtake-free TBP, and rows 13–17 for the TBP with overtaking.

Comparing rows 3–7 to rows 8–17 reveals that for all five policies, the (overtaking or overtake-free) TBP modification produces the expected result: trading off somewhat worse performance on the macrolevel SLMs for an improved performance on the microlevel SLM. Indeed, we observe that the TBP modification results in an increase in average system time by 7–12 minutes and an increase in the proportion of customers with a ≥4 hour system time by 6%–9%. On the other hand, the TBP modification reduces the number of RF incidents by 132–239 and shortens the $E[W \mid RF]$ by 1.5–6 minutes. Moreover, when combined with LS and LMOP policies, the TBP modification reduces $2\,Std[W \mid RF]$ by 12–30 seconds, which dominates the increases in $2\,Std[W \mid RF]$ when combined with LAW, LCW, and SERP policies (≤ 7 seconds). Thus, the TBP modification does not discriminate customers who already wait more than 20 minutes to improve the microlevel SLM. This is in contrast to the practice described in Milner and Olsen (2008) (see also Soh and Gurvich 2016), where such customers are purposely moved to the back of the line.

Comparing rows 8–12 with rows 13–17 shows that the performances of the overtaking and overtake-free TBPs are similar. Although, when combined with LS, LMOP, and LCW policies, the overtake-free TBP dominates the overtaking TBP in all measures, and when combined with LAW policy, the overtaking TBP dominates the overtake-free TBP. For the SERP policy, the overtaking TBP has fewer RFs (453 versus 472) but longer average system time than the overtake-free TBP (3:43:58 versus 3:31:36). Regarding the 95% confidence interval of waits more than 20 minutes, the overtaking TBP does better than the overtake-free TBP when combined with most DSPs, except the LCW policy.

According to XYZ's three main SLMs, the most attractive TBP-modified policies appear to be the LMOP + Overtake-free TBP (lowest number of RFs, lowest proportion with a ≥ 4 hour system time, and third best average system time) and LAW + Overtaking TBP (lowest average system time, with reasonable performance on the other two measures).

The performance of the LMOP + Overtake-free TBP policy exceeds the ED (actual scheduling rule) with respect to all three SLMs: it reduces the system time from 4:04:26 to 3:32:12, it reduces the proportion of customers with a ≥4 hour system time from 52.2% to 26.2%, and it reduces the number of RF incidents from 456 to 442. In addition, it improves the number of customers who end their visits with the doctor review from 898 to 1,042. Thus, the main goal of our analysis of the XYZ system—construction of an automatic scheduling policy that would outperform the current system with respect to all measures—has been achieved.

However, it should be noted that the cost of improving the microlevel SLM via the TBP modification is not minor. For example, the nonidle LMOP policy has only 20.2% of customers exceeding the four hours system time, while the number of RFs is 574. Thus, the overtake-free TBP modification of this policy decreases the number of RFs by 132, but increases the number of customers taking over four hours in the system by 121 (as well as decreasing the number of customers ending their visits with the doctor review by 85). It is up to XYZ management to decide whether this trade-off is worthwhile and whether a TBP-modified or the original nonidling version of the policy should be implemented.

In Online Appendix A, we derive the performance of SI modification, when combined with DSPs, in randomly generated open-shop systems in both transient and steady states and show similar trade-offs between macro- and microlevel SLMs.

To summarize, our main finding from this section is that the usage of SI in conjunction with the DSPs, such as the LMOP and LAW policies, provides a trade-off between macro- and microlevel measures in a systematic fashion. Moreover, because of the simple and transparent structure of these policies, the cost of implementing them should be low. In addition to supporting the main theme of this paper on the joint usage of DSPs and SI, this finding is also encouraging for the use of decision support systems for managing complex service networks in practice.

6. Summary and Open Questions

The main feature of this paper is the need to balance the more traditional macrolevel service-level measures, such as the total system time, with the customer-focused microlevel measure related to incidents of excessive waits at individual stations within the system. Such incidents can be managed by introducing strategic idleness, where intentional (small) waits are introduced in upstream stations to prevent (longer) waits at busy downstream stations. Our work was motivated by the observations from a real-life medical clinic operated by XYZ. Through process analysis and statistical hypothesis tests, we demonstrate that system schedulers appear to use SI to manage the microlevel measure.

We introduced a number of completely reactive (online) dynamic scheduling policies by defining simple scoring rules. We also showed how a given policy can be modified to include an SI component allowing it to account for both micro- and macrolevel measures. By developing simulation models based on the empirical data from XYZ, we showed that our automated scheduling policies with SI modification appear to be quite promising: achieving substantial improvements on the macrolevel measures while exceeding the performance of actual policies on the microlevel measure. We further examined the effect of SI modification in a randomly generated open-shop network in both transient and steady states and showed that the same intuition holds.

Because of the complexity of the underlying system, our results are mostly computational. Analytical substantiation of some of our conclusions would be quite interesting. A step in this direction was taken by Baron et al. (2014), who investigated policies with SI analytically for a tandem queue. An extension of their results to an open-shop or more complex stochastic network requires further work. In addition, other questions remain open: In what structure of systems is the trade-off between nonidle and SI policies (using TBP or other implementations) the most significant? Would the trade-off be stronger in a tandem system or in an open-shop system? Would the load and scale of the system affect the magnitude of the trade-off? Would the offered load analysis (see, e.g., Whitt 2013) improve the performance of DSPs? How could we jointly determine DSP and SI policies instead of adding SI to existing DSPs? More generally, it would be beneficial to understand the influence of precedence constraints on the potential advantage of SI. Finally, a broader research question is, what is the optimal SI policy in a tandem queue or open-shop service network?

References

Afèche P (2013) Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing Service Oper. Management* 15(3):423–443.

Alcaide D, Rodriguez-Gonzalez A, Sicilia J (2006) A heuristic approach to minimize expected makespan in open shops subject to stochastic processing times and failures. *Internat. J Flexible Manufacturing Systems* 17(3):201–226.

Alcaide D, Sicilia J, Vigo D (1997) A tabu search algorithm for the open shop problem. *Trabajos de Investigación Operativa* 5(2): 283–296.

Atar R, Mandelbaum A, Zviran A (2012) Control of fork-join networks in heavy traffic. 2012 50*th Annual Allerton Conf., Comm., Control, Comput.* 823–830.

Baron O, Milner J (2009) Staffing to maximize profit for call centers with alternate service level agreements. *Oper. Res.* 57(3): 685–700.

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.

Bouch A, Kuchinsky A, Bhatti N (2000) Quality is in the eye of the beholder: Meeting users' requirements for internet quality of service. *Proc. CHI*2000 *Human Factors in Comput. Systems Conf.* (ACM Press, New York), 297–394.

Chen H, Yao D (2001) *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization* (Springer-Verlag, New York).

Chhaochhria P, Graves S (2013) A Forecast-driven tactical planning model for a serial manufacturing system. *Internat. J. Production Res.* 51(23–24):6860–6879.

de-Véricourt F, Jennings O (2011) Review on nurse staffing in medical units: A queueing perspective. *Oper. Res.* 59(6):1320–1331.

de-Véricourt F, Zhou Y-P (2005) A routing problem for call centers with customer callbacks after service failure. *Oper. Res.* 53(6):968–981.

Friedman HH, Friedman LW (1997) Reducing the "wait" in waiting-line systems: Waiting line segmentation. *Bus. Horizons* 40(4): 54–58.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Graves S (1986) A tactical planning model for a job shop. *Oper. Res.* 34(4):522–533.

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.

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

Masin M, Herer Y, Dar-el EM (2005) Design of self-regulating production control systems by tradeoffs programming. *IIE Trans.* 37(3):217–232.

Mehrotra V, Ross K, Ryder G, Zhou YP (2012) Routing to manage resolution and waiting time in call centers with heterogeneous servers. *Manufacturing Service Oper. Management* 14(1):66–81.

Milner J, Olsen TL (2008) Service level agreements in call centers: Perils and prescriptions. *Management Sci.* 54(2):238–252.

Ouelhadj D, Petrovic S (2009) A survey of dynamic scheduling in manufacturing systems. *J. Scheduling* 12(4):417–431.

Pinedo M (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 M (2012) *Scheduling: Theory, Algorithms, and Systems* (Springer, New York).

Pinedo M, Ross SM (1982) Minimizing expected makespan in stochastic open shops. *Adv. Appl. Probab.* 14(4):898–911.

Roemer TA (2006) A note on the complexity of the concurrent open shop problem. *J. Scheduling* 9(4):389–396.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Oper. Res.* 60(5):1080–1097.

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

Simchi-Levi D (2014) OM forum—OM research: From problem-driven to data-driven research. *Manufacturing Service Oper. Management* 16(1):2–10.

Soh SB, Gurvich I (2016) Call center staffing: Service-level constraints and index priorities. *Oper. Res.*, ePub ahead of print October 28, http://dx.doi.org/10.1287/opre.2016.1532.

Soman D, Shi M (2003) Virtual progress: The effect of path characteristics on perceptions of progress and choice. *Management Sci.* 49(9):1229–1250.

Stidham S (2002) Analysis, design, and control of queueing systems. *Oper. Res.* 50(1):197–216.

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

Teo CC, Bhatnagar R, Graves S (2012) An application of master schedule smoothing and planned lead time control. *Production Oper. Management* 21(2):211–223.

Whitt W (2013) Offered load analysis for staffing. *Manufacturing Service Oper. Management* 15(2):166–169.