Management Science

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

Optimal Workflow Decisions for Investigators in Systems with Interruptions

Gregory Dobson, Tolga Tezcan, Vera Tilson,

To cite this article:
Gregory Dobson, Tolga Tezcan, Vera Tilson, (2013) Optimal Workflow Decisions for Investigators in Systems with Interruptions. Management Science 59(5):1125-1141. http://dx.doi.org/10.1287/mnsc.1120.1632

MANAGEMENT SCIENCE
Vol. 59, No. 5, May 2013, pp. 1125–1141
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1632

Optimal Workflow Decisions for Investigators in Systems with Interruptions

Gregory Dobson, Tolga Tezcan, Vera Tilson
Simon School of Business, University of Rochester, Rochester, New York 14617
{greg.dobson@simon.rochester.edu, tolga.tezcan@simon.rochester.edu, vera.tilson@simon.rochester.edu}

We model a system that consists of a stream of customers processed through three steps by two resources. The first resource, an investigator, handles the first step, in which she collects information from the customer and decides what work will be done in the second step by the second resource, the back office. In the third step, the investigator returns to the customer armed with the additional information or analysis done by the back office and provides the customer with a conclusion, solution, or diagnosis. The investigator has to prioritize either seeing a new customer or completing the work with a customer already in the system. While serving one customer, the investigator may be interrupted by requests from the other customers in the system. Our main objective is to understand the impact of the investigator's choices on system throughput. In addition, we are interested in the occupancy of the system (and thus the flow time of customers). We create a stylized queueing model to examine the investigator's decisions and show that, when interruptions are not an issue, the investigator should prioritize new customers to maximize throughput, keeping the system as full as possible. If customers who have been in the system for a long time generate interruptions and thus additional work for the investigator, we show that it is asymptotically optimal for the investigator to keep the system occupancy low and prioritize discharging customers. Our conclusions are based on a model of a re-entrant queue with dedicated servers serving multiple stations, with two novel features: a buffer that is shared between stations, and jobs in the system generating additional work for the servers.

Key words: dynamic programming; optimal control; Markov; healthcare; hospitals; probability; Markov processes; stochastic; queues; limit theorems
History: Received September 28, 2010; accepted July 12, 2012, by Assaf Zeevi, stochastic models and simulation. Published online in *Articles in Advance* January 8, 2013.

1. Introduction

We model a system that consists of a stream of customers processed through three steps by two resources. The first resource, an investigator, handles the first step, in which she collects information from the customer and decides what work will be done in the second step by the second resource, the back office. In the third step, the investigator returns to the customer armed with the additional information or analysis done by the back office and provides the customer with a conclusion, solution, or diagnosis. The model is inspired by our observations of an emergency department (ED). We view the physician as the investigator and the lab or radiology department as the back office. The physician performs the initial exam and orders tests in the first step. The tests, typically radiology or blood work, are done in the second step, and the physician combines the results of the tests with her initial exam to provide the customer with a diagnosis (and possible treatment) in the third step. Our model could also represent a lawyer's office, in which the investigator is the lawyer and the back office is one or more legal assistants providing research and writing support, or a bank where the investigator is a loan officer and the back office generates the background work on the customer to make sure he has a good credit score. Thus, the investigator is a professional who provides the interface to the customer. The investigator does the initial data gathering; the back office provides additional data collection, analysis, or technical assistance to support the investigator; and the investigator provides the diagnosis or conclusion at the third step.

We assume that we are dealing with a system that is overloaded, such as an ED. Hence, our main objective is to understand the impact of the investigator's choices on the throughput of the system. We are also interested in the occupancy of the system (and thus the flow time of customers). We consider two additional features of the system and the impact these have on throughput and occupancy: (1) interruptions to the investigator's work, generated by the customers in the system waiting for the diagnosis, and (2) a limit on the number of slots in which customers can wait between the first and the final step. For some of the examples, this limitation might be imposed by the investigator (or her management) to ensure that the system is not overloading the investigator who is

trying to keep track of and thus recall the details for too many customers. When the customer is present throughout, such as a medical clinic or the emergency room, the slots may be the exam rooms or beds available. At any point that our investigator has customers waiting for her two steps, she has a choice of which customer to handle next. If she takes a new customer, then she reduces the time that customer waits for the first service interaction, a common measure related to customer satisfaction, but the same decision increases the total system time for customers already in the system.

As previously noted, our research was initially motivated by our observations of emergency room physicians in two large academic hospitals. Generally, from the point of view of an ED physician, the process of treating a patient consisted of two steps: (1) the initial physical examination including the ordering of tests and treatments and (2) the decision on disposition once the tests and treatments were completed. Although not all patients followed this route within the ED, a large majority did. Physicians had considerable freedom in deciding how to organize their workflow. Trade-offs were made when a physician decided to examine a new patient, rather than complete the work needed to discharge a patient already in the ED. Such prioritization lowered the door-to-doctor time of the new patient but increased the total time in the ED for the patient who could have been discharged or admitted to the hospital. It also increased physical capacity requirements for the ED, as well as the number of patients who had to be simultaneously managed by the physician.

There is empirical evidence that interruptions occur in EDs and that the number of task interruptions is correlated with the number of patients in the ED (Chisholm et al. 2000, Westbrook et al. 2010). In Chisholm et al. (2000), two types of events are considered: "interruptions" and "break-in-tasks" (longer interruptions). Chisholm et al. (2000, p. 1239) observed 30 ED physicians over 36 three-hour periods, and found that "physicians performed a mean of 67.6 tasks per study period. The mean number of interruptions per three hour study period was 30.9 and the mean number of breaks-in-tasks was 20.7." Because interruptions are most often related to the treatment of patients under the physician's care, it is not surprising that Chisholm et al. (2000) found that both the number of interruptions ($r = 0.63$; $p < 0.001$) and the number of breaks in tasks ($r = 0.56$; $p < 0.001$) per observation period were positively correlated with the average number of patients simultaneously managed. In an experimental setting, Speier and Valacich (1999) and Altman (2002) showed that interruptions cause an increase in the total time on task because of the task "switching costs" and "resumption lag."

We create a stylized queueing model to analyze the impact of an investigator's work prioritization decision on system throughput and occupancy. Using the Markov decision process (MDP) analysis, we show that when interruptions are not an issue, throughput is maximized by a simple policy of prioritizing new customers over existing ones and keeping the system as full as possible. On the other hand, when managing a larger number of customers means more interruptions, the throughput optimal policy becomes more complex. Using asymptotic analysis we show that a simple policy of prioritizing existing customers over the new ones results in better throughput measures and average occupancy rates. Our asymptotic analysis is based on considering a sequence of systems as the buffer size available in the system gets large. We scale the occupancy and abandonment costs as well as the interruption rates. We extend the framework in Bramson (1998) to the finite buffer case and prove that in the limit there is a particular state-space collapse; the proportion of filled slots in the system converges to zero. To the best of our knowledge, the application of this framework in the finite buffer space case is new. The methodology developed in this paper can be used to study the control of other finite buffer systems, such as those studied in Andradottir et al. (2001) and Andradottir and Ayhan (2005). A similar approach has been used in the literature for open and closed networks and many server systems; see Bramson and Dai (2001), Kumar (2000), and Dai and Tezcan (2008), among others. However, because of the differences between our asymptotic regime and those that appear in the literature (e.g., Bramson and Dai 2001, Kumar 2000, Dai and Tezcan 2008), the extension of the results in Bramson (1998) requires new analytical techniques.

This paper proceeds as follows. We present our general model in §2 and review the related queueing literature in §3. In §4 we consider a simplified model with exponential service times and no interruptions, describe a throughput optimal policy, and characterize throughput and average occupancy in terms of the relative rates of the investigator, the back-office services, and the buffer capacity. We focus on the model with interruptions in §5. We first show that the exact optimal solution is complicated, and this motivates system analysis in an asymptotic regime, and we focus on that regime in the rest of that section. We summarize the implications for practitioners and further research opportunities in §6. We present the proofs of our results in Online Appendices EC1–EC3.[1] We present additional numerical experiments in Online Appendix EC4, and in Online Appendix EC5 we present approximations

[1] Online appendices available at http://tolgatezcan.simon.rochester.edu/e-companion.pdf.

to determine the minimum buffer size to achieve a desired level of throughput.

2. General Model

We model the workflow as a network of four queues (illustrated in Figure 1). Customers in queue j are served at station j, for $j = 0, 1, 2$. In addition, jobs created by the customers in the system (interruptions) queue up at queue I and are served at station I. As we will explain, some of the queues are physical, whereas others are virtual; all of the stations are virtual. The arrival process is stochastic. Customers arrive at rate λ and wait in queue 0. Customers who are not served promptly renege from queue 0 at rate γ. At station 0 an investigator performs the initial interview and creates the work orders for station 1. After completing service at station 0, customers are assigned a buffer slot. Next, the customers wait in virtual queue 1 for the back-office work to be completed at station 1. After the back-office work is completed, the customers are transferred to virtual queue 2, where they wait for another service by the investigator. We assume that there is a single investigator and one server in the back office. At station 2 the investigator reviews the back-office work, provides a conclusion, solution, or diagnosis, and delivers it to the customer. After this step the customer leaves our system, releasing the slot that was allocated to him. Unlike the typical case in a manufacturing system, customers do not physically move from queue 1 to queue 2. We denote the service rate at station j with μ_j.

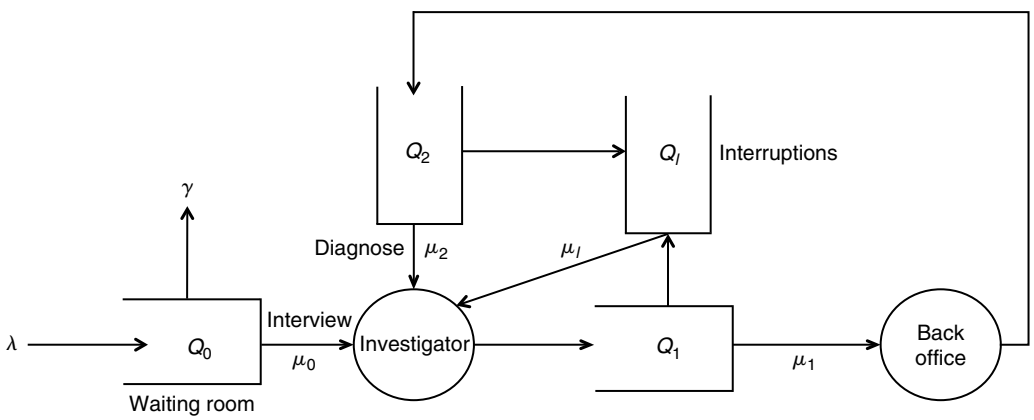
Let $Q_j(t)$ denote the number of customers either waiting or being served at station j at time t. When the customer is in the system, he consumes a resource—a buffer slot. Our main focus is on the case for which there are B buffer slots available and thus at most B customers are in the system, measured from the point when the customer finishes at station 0 to when he finishes at station 2, i.e., $Q_1(t) + Q_2(t) \le B$. When all the slots are occupied, we consider the investigator to be blocked at station 0.

To model the increased workload brought on by interruptions, we assume that a customer who occupies a slot in the system generates additional work for the investigator, i.e., a job or an *interruption*. These interruptions queue for the investigator at queue I, which is given priority and handled by the investigator at rate μ_I. In §2.1 we explain in detail what kind of interruptions arise and how they are handled in an ED. We assume that each customer who does not currently have a job in the interruption queue generates requests at the rate ν, so if there are $Q_1(t) + Q_2(t)$ customers in the system and $Q_I(t)$ jobs either at or in front of the station I at time t, then the instantaneous arrival rate of additional interruptions is $\nu(Q_1(t) + Q_2(t) - Q_I(t))$.

When there are no interruptions to be served, the investigator has three possible actions $\{in, out, idle\}$; Action *in* corresponds to serving a customer at station 0, that is, bringing an additional customer into the system by performing the initial interview. Action *out* corresponds to serving customers at station 2, that is, doing the work to provide a conclusion or decision, which hereafter in the paper we will refer to as a diagnosis, after which a customer may leave the system. Action *idle* means not working at either station. Without interruptions, the investigator is free to take any action in any state, with the exception that there must be a free slot for her to perform action *in* and that $Q_2 > 0$ for her to perform action *out*. When there are interruptions, we assume that the investigator services them immediately, returning to her previous station after the interruption service is completed. Note that although we mainly focus on the case for which the interruptions preempt the investigator's current task, our asymptotic results also hold for the nonpreemptive case.

We are interested primarily in the measures of system throughput and occupancy. We investigate how

Figure 1 Queueing Model



the investigator's decisions affect these measures. In addition to the optimal throughput policy, we investigate two other policies: first buffer first (FBF), in which the investigator gives priority to bringing new customers into the system, i.e., she performs action *in*, whenever there is a slot available to place a customer; and last buffer first (LBF), in which the investigator gives priority to completing her work with a customer by doing the diagnosis step, i.e., she performs action *out*, whenever there is a customer who has completed the back-office step. We investigate under what conditions FBF and LBF are optimal for maximizing throughput, and when neither is optimal, we show whether one of them is close to optimal. In all these cases we examine the impact of these policies on occupancy.

For completeness, we also consider variations of our basic model, relaxing some of the assumptions. We discuss the configurations for which the buffer space is unlimited and for which there are separate limits on the number of customers in queues 1 and 2. For both, we consider cases for which there are interruptions (§5.8) and for which there are no interruptions (§4). In §5.4 we consider the case of customers visiting certain stations more than once. In §5.5 we discuss relaxing the assumption of a single investigator and a single server in the back office.

2.1. ED Operations and Our Model

As previously mentioned, one of the main motivating applications of this paper is an emergency department. In this section we explain the details of how our model reflects ED operations. We note that different EDs have different patient profiles and possibly operate in different ways. We do not claim our model captures the workflow in every ED; however, we believe that similar trade-offs exist in most EDs.

Our focus in this paper is on the decisions of the physician, and so we chose not to model the details of the triage and registration steps. (For a review of triage literature, see §3.) We assume that patients are already triaged and therefore sorted by acuity; thus, we do not explicitly model multiple classes of patients. In a real ED, occasionally there are critically ill patients who need immediate treatment. These cases are handled differently than other cases; these patients are given preemptive priority even if there is no bed available for them. In our observations, such cases were rare; on average, we observed less than one incident per eight-hour shift. Therefore, we do not include such cases in our model.

Thus, we consider the dominant flow in the ED, assume that patients do not differ in the expected service duration, and assume that prior to being picked out of queue 0, patients have already been sorted by waiting costs, so the cost of waiting for the initial service among different service classes is not affected by the physician's decision. Patients who have been seen by the physician and are awaiting test results (the back-office work) are in queue 1, and those who have completed their tests, have their results, and are waiting to see the physician again are in queue 2. In line with our actual observations, even though the patients who have yet to be seen by a physician may generate interruptions, we assume that these interruptions are usually handled by the staff at the front desk, and hence we do not include them within the boundaries of our studied system.

In our observations, ED physicians were practically never idle, and we never observed an ED physician having to go out to the waiting room to examine a patient. Typically, the triage and registration process ensured that patients were in the exam rooms before the physician first saw them. This was mainly because the ED was overloaded (in terms of patient demand) a large portion of the day, and the capacity of staff assigning patients to beds exceeded the total capacity of the physicians. Because of our interest in interruptions of a physician's work, we do not keep track of the number of patients waiting in an ED bed for an initial examination by a physician.

Patients usually do not leave the ED immediately after the physician makes a diagnosis. Rather, they are normally discharged by the nursing staff. Once a decision about the patient has been made and the physician has completed the diagnosis, discharge, or hospital admission instructions, we assume that the responsibility for the patient has been transferred and the patient is unlikely to generate any more interruptions for the physician. Clearly, this is an approximation. Frequently, patients who are to be admitted to the hospital remain in the ED for hours, waiting for an available hospital bed (this practice is called *ED boarding*). However, hospital bed availability is mainly determined by factors other than the actions of ED physicians. Hence, we consider diagnosed patients to be outside our model. If those patients were included in our model, our asymptotic analysis would still be valid, provided that the bottleneck resource in the system is either the investigator or the back office. In addition, the effect of ED boarding on the number of required beds in the ED can be computed independently from our analysis once the throughput from the ED is determined. So in our model, B is the maximum number of patients a physician can manage simultaneously.

We model interruptions generated by patients as jobs arriving to queue I. Interruptions in an ED come in different forms. We focus on the non-life-threatening interruptions that have to be handled by a physician. In our observations, life-threatening interruptions were very rare. Routine interruptions,

on the other hand, happened all the time. Physicians were frequently interrupted while performing indirect patient care, such as filling out orders, dictating medical records, or walking to see a patient. We model these indirect patient care tasks as included in the service times in stations 0 and 2. Although interruptions were less frequent when a physician was examining a patient, they did happen as well—for things as casual as another physician asking whether there were different size examining gloves in the room or as serious as a nurse asking for permission to administer pain medication to another patient. Interruptions were usually short and were normally given preemptive priority by the physicians. For example, if a nurse walked up to a physician to clarify an order for a patient, the physician did not make the nurse wait until he or she finished the current task. A question may arise whether it is throughput optimal for a physician to handle interrupts preemptively. Our computational experiments suggest that there will be a less than 1% improvement in physician throughput if the physician completed the task at hand before servicing an interruption. However, by not allowing preemption, the physician would block another resource, e.g., a nurse, from completing her task. To determine what policy would be optimal in the larger context would require a significantly more complicated queueing model. The model would need to include the tasks performed by providers who are being blocked by the physician's delay in servicing the interruption. The model would be further complicated by the simultaneity requirement of two servers for serving the interruption. We believe that it would be very difficult to get analytical results for such a system. Furthermore, given the very small improvement in the physician throughput, not allowing preemption is likely to be inefficient in the larger picture. Therefore, we do not consider nonpreemptive servicing of interruptions in our paper and instead model the handling of interruptions based on the practice we observed.

Although we assumed a three-step process, patients may have to be rerouted back from the diagnosis stage to complete more tests. Our approximation of the process as a three-step process reflects the majority of cases and offers a parsimonious model to focus on the main trade-offs in the system. We discuss how our results can be extended to the potential rerouting case in §5.4.

3. Literature Review

An extensive healthcare and operations literature addresses the utility of, as well as best practices in, ED triage of newly arrived patients. Oredsson et al. (2011) offer a comprehensive review of medical practitioner papers that address the effects of various triage interventions on patient flow in EDs. In the operations literature, Argon and Ziya (2009) examine the practice of triage under the objective of minimum waiting costs. They consider a system where two customer types differ in waiting costs as well as service times, and where the customer type cannot be perfectly determined in triage. Argon and Ziya (2009) investigate policies that minimize total waiting costs as a function of the strength of the customer's signal type. Dobson and Sainathan (2011) consider a service system with two customer types differentiated by the waiting costs. Dobson and Sainathan (2011) examine the conditions under which sorting of customers results in lower costs. Saghafian et al. (2012) propose a triage system based not only on waiting costs but also on medical complexity. More generally, this research is related to queueing control literature on assigning priority to the newly arrived jobs (e.g., Duenyas et al. 1998, Hu and Benjaafar 2009, Whitt 1999). We consider the initial triage decision to be outside the scope of our model, and we do not differentiate customers based on the cost of the initial wait but only based on their progress in the service process.

There is a rich queueing literature on the dynamic allocation of resources (or servers) to different classes of customers; see Gans et al. (2003), Bell and Williams (2005, 2001), Mandelbaum and Stolyar (2004), and Dai and Wuqin (2008). However, in most of these studies, the buffer space is assumed to be infinite. Most relevant to our analysis is the research on tandem lines with limited buffer space in front of each station. An extensive review of this literature can be found in Van Oyen and Hopp (2004) on a cross-trained workforce. Ahn et al. (2002, 1999), Duenyas et al. (1998), Iravani et al. (1997), Kaufman et al. (2005), Rosberg et al. (1982), and Wu et al. (2006, 2008) study tandem queues with flexible servers with the objective of minimizing the holding costs. Bartholdi and Eisenstein (1996), Ahn and Righter (2006), Gel et al. (2002), McClain et al. (1992), and Zavadlav et al. (1996) consider the line-balancing problem in various tandem systems. Andradottir et al. (2001), Arumugam et al. (2009), Andradottir and Ayhan (2005), Andradóttir et al. (2008), and Kirkizlar (2008) consider maximizing throughput in tandem lines with finite buffer space and different characteristics. For example, Andradottir et al. (2001) identify the optimal policies for a system with two servers and two stations and derive heuristics for larger systems, and Andradóttir et al. (2008) consider the effect of service failures.

Asymptotic analysis enables the study of queueing systems that are beyond the realm of exact analysis. In many cases, the asymptotic analysis also helps identify the important characteristics of the underlying system. In Harrison and Wein (1990),

a Brownian control problem with the objective of maximizing the throughput for a two-station system is formulated and solved. Kumar (2000) proves the asymptotic optimality of the policy proposed in Harrison and Wein (1990). Chevalier and Wein (1993) extend the results in Harrison and Wein (1990) to multistation closed queueing networks. Gilland (2001) proposes control policies, which are based on the solution of a particular Brownian control problem that performs better than those proposed in Chevalier and Wein (1993) in certain situations. In the proof of our asymptotic results, we rely heavily on the framework established in Bramson (1998); also see Dai and Tezcan (2011), Gurvich and Whitt (2009), and Besbes and Maglaras (2009) for other applications.

There are two major differences between the models considered in the literature and our model. First, we include interruptions in our model. Second, we consider systems that have a capacity constraint on the total number of "jobs" in the system, not on each buffer. We also consider the case for which there are infinitely many jobs available in the initial buffer. These characteristics of our model mean that the previous results are not directly applicable.

4. The Model Without Interruptions

In this section we consider the queueing model described in §2 in which there are no interruptions. We assume that there is a single investigator, that the back office is a single server, and that the queue of arriving customers is always nonempty so there is a customer available for the investigator whenever she wants to perform an action *in*. Therefore, customer arrivals are irrelevant in this model. We assume that the service times at all three stations are exponentially distributed.

First consider the case in which there is a limit on the total number of jobs in queues 1 and 2, and let the buffer size be denoted by $B > 0$. We can characterize the policy that maximizes throughput. Let

$$m_v = \frac{1}{\mu_0} + \frac{1}{\mu_2} \quad \text{and} \quad m_b = \frac{1}{\mu_1}, \tag{1}$$

where m_v denotes the expected time the investigator spends on each customer, and m_b denotes the expected time the back office spends on each customer. In this section, we assume without loss of generality, that $\mu_0 + \mu_1 + \mu_2 = 1$.

Theorem 1. *Policy FBF is throughput optimal* (*among all preemptive and nonpreemptive policies*) *when there is a limit on the total number of jobs in the system, B. Furthermore, the optimal throughput for a system increases with B and asymptotically approaches* $1/m_v$ *if the investigator is the bottleneck and* $1/m_b$ *if the back office is the bottleneck.*

Details of the proof are presented in Online Appendix EC1. For defining the optimality condition, we rely on results from Andradottir et al. (2001), who showed that maximizing the steady-state throughput of a controlled queueing system is equivalent to maximizing the steady-state departure rate for the embedded Markov chain. Our proof that FBF is optimal uses the policy iteration algorithm for communicating MDPs to show that no other policy dominates FBF. The same proof technique was used in Andradottir and Ayhan (2005), but in our case the state and action space is more complex because a single buffer is shared among multiple stations.

Although we focus on systems with the back office represented by a single server, our numerical experiments suggest that having several slower servers working in parallel does not affect our result. Specifically, we ran several numerical experiments: with $\mu_1 = 0.5$ and two servers, and $\mu_1 = 0.25$ and four servers and within those $\mu_0 = \mu_2 = 1$ and $\mu_0 = \mu_2 = 1/0.6$. We considered $B = 2, \ldots, 20$ for all these parameter settings. In all these experiments we found that FBF maximized throughput.

When $m_b \neq m_v$, the throughput under the FBF policy can be expressed in terms of $\beta(B)$ (see Lemma 2 in the online appendix) as

$$\frac{1 + \beta(B)}{m_b + m_v \beta(B)}, \tag{2}$$

where $\beta(B)$ is defined in (11). The value of $\beta(1)$ is equal to $-m_v/m_b$. When the back office is the bottleneck, $\beta(1) > -1$, the function is increasing and $\beta(B)$ approaches 0 as B increases. When the investigator is the bottleneck, $\beta(1) < -1$, the function is decreasing, and $\beta(B)$ grows in magnitude as B increases. Hence, the function $\beta(B)$ can be considered the term reflecting the effect of finite buffer space on throughput.

Because the expression for $\beta(B)$ is complicated, we discuss an easily computed approximation of the system performance. Under the FBF policy, the number of customers in the system is either $B - 1$ or B. When the number is B and there are customers waiting in queue 2, the investigator discharges a customer; if there are no customers waiting in queue 2, the investigator remains idle until a customer is completed by the back office. Once the investigator discharges a customer, the number of customers in the system falls to $B - 1$, and the investigator then brings a new customer into the system. Therefore, we can model the system as a closed queueing network with two stations and B "jobs." Let the back-office work be performed at station 1, and the investigator's service be performed at station 2. A job is either waiting in queue 2 for the service by an investigator or in queue 1 for the service in the back office. After the investigator processes a job, the job enters the queue

for the back office. After the back office processes the job, the job queues for service by the investigator. When the investigator is not forced to idle, a customer is "transferred" from queue 2 to queue 1 after $expo(\mu_0)+expo(\mu_2)$ elapses, where $expo(\mu)$ denotes an exponential random variable with rate μ.

Unfortunately, this closed queueing network does not have a product-form stationary distribution. To obtain a simple approximation, we approximate the service times in the combined investigator station by an exponential distribution with mean m_v. Let $k = m_v/m_b \geq 1$ (we focus on the case when the investigator is the bottleneck; see §5.7 for a similar approach when the back office is the bottleneck), and let Q_i denote the number of customers in front of and at the ith station of the closed network in steady state. By well-known results on closed queueing networks, we have

$$P\{Q_1 = b_1, Q_2 = B - b_1\} = \begin{cases} \dfrac{k-1}{k-k^{-B}}k^{-b_1} & \text{for } 0 \leq b_1 \leq B, \text{ if } k > 1, \\ 1/B & \text{for } 0 \leq b_1 \leq B, \text{ if } k = 1. \end{cases} \quad (3)$$

Note that the throughput of the system in steady state is given by

$$m_v^{-1}(1 - P\{Q_1 = B, Q_2 = 0\}) \quad (4)$$

or, equivalently, by $m_b^{-1}(1 - P\{Q_1 = 0, Q_2 = B\})$. We test the quality of this approximation in Online Appendix EC5.

4.1. Unlimited Buffer Space

For completeness, we now consider the case with unlimited buffer. We first note that by Theorem 1, the throughput of a system with finite buffer space converges to the theoretical upper bound as B gets large. Therefore, for systems with unlimited buffer, one can use FBF by placing a constraint (B) on the number of customers admitted to the system. For large enough B, the throughput will be arbitrarily close to the theoretical upper bound. In addition, when the investigator is the bottleneck, the LBF policy is also throughput optimal, which is described in detail next. To present the result, let $D(t)$ denote the number of customers who are served in the system by time t.

Theorem 2. *Policy LBF is throughput optimal when there is unlimited buffer space in the sense that*

$$\lim_{t\to\infty} \frac{D(t)}{t} = m_v^{-1} \quad a.s. \quad (5)$$

The proof of this result appears in the online appendix and is based on the analysis of the fluid model of the underlying system.

When there are separate limits on buffer spaces in queues 1 and 2, the throughput optimal policy is not known in general. For some special cases, the optimal policy is established in Kirkizlar et al. (2012).

5. The Model With Interruptions

In this section we introduce interruptions into our model. As described in §2, customers within the system are assumed to generate interruptions as they wait. We start our analysis in §5.1 with numerical experiments that provide motivation for the asymptotic analysis that follows. In §5.2 we describe the asymptotic regime for large systems and present our main results. We provide the sketch of these results in §5.3. We consider systems with more general routing schemes in §5.4. In §5.5 we discuss how our results can be extended to systems with multiple resources at each station. In §5.6 we present numerical results that show that the asymptotic results hold approximately, even when the buffer size is small. We provide asymptotically optimal policies when the back office is the bottleneck in §5.7. Our main focus is on the case with *limited total buffer space*. We comment on the other cases in §5.8.

5.1. Exact Analysis

Assuming that the time until a customer generates an interruption as well as all the service times have exponential distribution, the system can be modeled as a Markov chain. The state space of the Markov chain model used in §4 is augmented with Q_I. Recall that the rate of interruptions at time t is $\nu(Q_1(t) + Q_2(t) - Q_I(t))$. Using standard MDP analysis tools, we can find, at least numerically, a policy that is throughput optimal.

To illustrate, consider the following parameters; $\mu_0 = \mu_1 = \mu_2 = 1$, $B = 5$, $\nu = 0.1$, and $\mu_I = 10$. To give some perspective on the frequency and duration of the interruptions here, if the investigator spends a total of 30 minutes on the customer, then each customer would generate one three-minute interruption every five hours. The optimal policy is presented in Figure 2(a). In this figure, the rows represent the number of customers in queue 2 and the columns represent the number of customers in queue 1. Each cell specifies the optimal action for the investigator when the queue lengths are at these levels. The specified actions are optimal, assuming there are no interruptions waiting for the investigator. When there is at least one interruption, by our assumption, the investigator gives (preemptive) priority to the interruption. Clearly, the optimal policy for the investigator in this system with interruptions is more complicated than the optimal policy in the system without interruptions. If we change ν to 0.05, making interruptions half as frequent, and keep the other parameters fixed, the optimal policy changes to that illustrated in Figure 2(b). Thus, the optimal policy appears to be sensitive to the changes in the system parameters.

We ran several additional experiments for systems with different processing rates and with buffer sizes

Figure 2 Optimal Policy With Interruptions

(a) Parameter set 1

Q_2						
5	Out					
4	Out	Out				
3	Out	Out	Out			
2	In	Out	Out	Out		
1	In	Out	Out	Out	Out	
0	In	In	In	In	In	Idle
	0	1	2	3	4	5

Q_1

(b) Parameter set 2

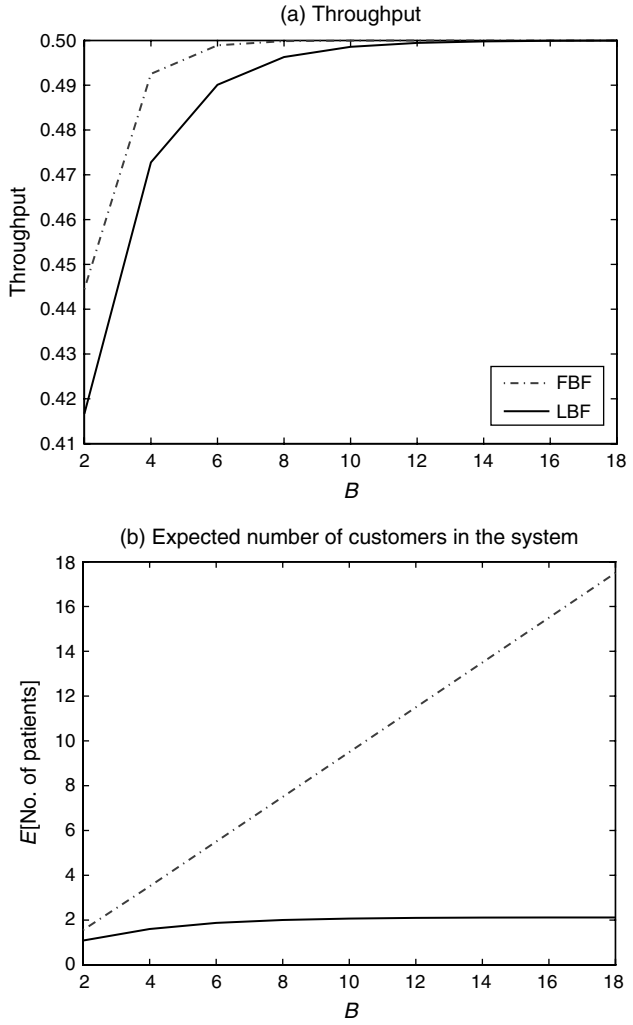
Q_2						
5	Out					
4	Out	Out				
3	In	Out	Out			
2	In	Out	Out	Out		
1	In	In	Out	Out	Out	
0	In	In	In	In	In	Idle
	0	1	2	3	4	5

Q_1

of up to 20 slots. The results were similar to those described above. In light of these experiments, it is unlikely that a simple policy would be optimal in systems with interruptions. Although it may be possible to establish some results about the structure of the optimal policy, we consider it more useful to search for simpler policies that are near optimal, and we seek to establish conditions under which such policies perform well.

A natural question to ask in this setting is whether the FBF policy still performs well. Although we provide a more definitive answer in our asymptotic analysis (and in the numerical results that follow), we can infer from the no-interruption case that FBF does not in fact perform well when there are interruptions. Consider the model without interruptions. Let $\mu_0 = \mu_1 = \mu_2 = 1$. We compare graphically the steady-state throughput and the expected number of customers under the FBF and LBF policies. In Figures 3(a) and 3(b), we plot the throughput and the expected number of customers as a function of buffer size, B. Under FBF the expected number of customers in the system is increasing linearly with B. Under LBF, on the other hand, the expected number of customers in the steady state seems to level off around two for large B. In addition, as B increases, the difference between the throughputs of these two policies becomes very small. The maximum difference is around 4% and falls below 2% for $B \geq 5$. In the presence of interruptions, FBF would imply more interruptions and a significant reduction in throughput. Therefore, it is likely that FBF is no longer optimal in the presence of interruptions, whereas LBF looks like a good candidate policy.

Figure 3 Numerical Results for System Model Without Interruptions



5.2. Asymptotic Analysis

In this section we focus on an asymptotic analysis to identify good policies. The asymptotic analysis not only helps us identify good, easily implementable policies but also enables us to relax some of the restrictive assumptions we have made in previous sections. First, we no longer assume that there are infinitely many customers waiting for the investigator. Instead, we assume that customers arrive to the system with a fixed rate λ according to a renewal process. Practically, this case includes the case with infinite customers waiting, because we can assume that λ is arbitrarily large. The waiting area in front of station 1 is assumed to have infinite capacity. Also, customers waiting to be seen are assumed to have limited patience. Specifically, we assume that each customer's time to renege is exponentially distributed with rate γ. We also allow service times to

have a general distribution that satisfies some general condition that we elaborate on later in the paper; see (22) and (23). In addition, we no longer have to assume that there is only one investigator and one server representing the back office. Although, mostly for notational simplicity, our main focus in this section will be the case with a single investigator and a single server in the back office, our asymptotic analysis can easily be extended to more general cases, as discussed in §5.5.

We begin by assuming that the investigator is the bottleneck, i.e.,

$$m_v > m_b, \tag{6}$$

and that the system is overloaded, i.e.,

$$\lambda > m_v^{-1}. \tag{7}$$

We examine the case in which the back office is the bottleneck in §5.7.

5.2.1. Cost. We consider a more general objective function that consists of throughput and linear occupancy costs. Let $R(t)$ denote the number of customers who reneged from the first queue by time t. Our purpose is to find an optimal policy π to minimize the following objective:

$$C_\pi(T) = aE[R(T)] + E\left[\int_0^T \sum_{i=0}^{2} b_i Q_i(s)\,ds\right], \tag{8}$$

for $a,\ b_i \geq 0,\ i = 0, 1, 2$ and $T > 0$ fixed. This objective includes the objective of throughput optimization; because the number of arrivals on a finite time interval is fixed, setting $b_i = 0$ and minimizing the number of abandonments is equivalent to maximizing throughput.

Under the assumption that the time until the customer reneges has an exponential distribution with rate γ, we have

$$E[R(T)] = E\left[\gamma \int_0^T Q_0(s)\,ds\right]. \tag{9}$$

Hence, our objective function can be written as

$$C_\pi(T) = (a\gamma + b_0)E\left[\int_0^T Q_0(s)\,ds\right] + E\left[\int_0^T \sum_{i=1}^{2} b_i Q_i(s)\,ds\right]. \tag{10}$$

For notational simplicity, we set $b_0 = 0$ and $b_1 = b_2 = c$.

5.2.2. Asymptotic Regime. In our asymptotic analysis, we consider a sequence of systems indexed by r. In this sequence, we assume that the service rates and the arrival rates are fixed and the buffer size is increasing with r; in the rth system the size of the buffer B^r is given by

$$B^r = r^\alpha B, \quad \text{for } 0 < \alpha < 1 \text{ and } B > 0. \tag{11}$$

We denote the rate at which each customer creates a new interruption by ν^r in the rth system and we assume that

$$\nu^r r^\alpha \to \nu > 0, \tag{12}$$

as $r \to \infty$ and

$$B\nu < \mu_I. \tag{13}$$

Assumption (12) implies that customers are generating fewer and fewer interruptions as r increases. This is to balance the fact that B^r is increasing. Assumption (13) is needed to make sure that the system has enough capacity to handle interruptions; otherwise, the investigator may be working only on interruptions.

We assume that in the rth system, the abandonment rate from queue 0 is γ^r with

$$\gamma^r r \to \gamma, \tag{14}$$

as $r \to \infty$ for some $\gamma > 0$. This implies that customers are becoming more patient as r increases. To keep the total abandonment costs of the same order for each r, we assume that

$$a^r / r \to a, \tag{15}$$

as $r \to \infty$ for some $a > 0$, where a^r is the abandonment cost per customer in the rth system.

We append r to all the quantities in the rth system; e.g., $Q_i^r(t)$ is the number of customers (or interruptions) waiting in queue i in the rth system at time t. We define

$$\begin{aligned} \bar{Q}_0^r(t) &= \frac{Q_0^r(rt)}{r} \quad \text{and} \\ \bar{Q}_i^r(t) &= \frac{Q_i^r(rt)}{r^\alpha}, \quad \text{for } i = 1,\ 2,\ I. \end{aligned} \tag{16}$$

Note that we scale the queue length for the queue in front of the system differently from those within the system. This is to take into account that the buffer size scales according to (11). We reflect the fact that the number of customers in front of the system is usually significantly more than the buffer capacity within the system.

To assess the asymptotic performance of a control policy, we consider the following objective function:

$$C_{\pi^r}^r(T) = \frac{1}{T} a^r \gamma^r E\left[\int_0^T \bar{Q}_0^r(s)\,ds\right] + \frac{1}{T} cE\left[\int_0^T (\bar{Q}_1^r(s) + \bar{Q}_2^r(s))\,ds\right], \tag{17}$$

where π^r is the scheduling policy in the rth system. This cost function is very similar to (10) except for the scaling factors r and r^α. We restrict ourselves to the scheduling policies that make control decisions based only on the current state (current server occupancy and queue content). Fix a sequence of policies $\pi = \{\pi^r\}$. We define

$$C_\pi^* = \lim_{T\to\infty} \liminf_{r\to\infty} C_{\pi^r}^r(T) \tag{18}$$

and set

$$C^* = \inf_\pi C_\pi^*. \tag{19}$$

For simplicity, we assume that the initial number of customers in each queue is deterministic and satisfies

$$\bar{Q}_i^r(0) \to \bar{Q}_i(0), \tag{20}$$

as $r \to \infty$ with

$$\bar{Q}_0(0) > 0. \tag{21}$$

Although these two assumptions are not necessary for our results to hold, they simplify the technical details of our analysis. Throughout the rest of this paper, we assume that interarrival times are independent and identically distributed (i.i.d.) with finite mean and make the following technical assumption for the service times. Let $u_i = \{u_i(n)\colon\ n \geq 2\}$ denote an i.i.d. sequence of random variables, where $u_i(n)$ denotes the service times of the nth customer in station $i = 0, 1, 2, I$. We use $u_i(1)$ to denote the service time of the first customer. We assume, as in Bramson (1998), that service times satisfy the following regularity conditions:

$$u_i(1)/r^\alpha \to 0 \quad \text{as } r \to \infty \tag{22}$$

and

$$E[u_i(2)^{1/\alpha} \mid u_i(2) > r] < \varphi(r), \tag{23}$$

for some function φ with $\varphi(r) \to 0$ as $r \to \infty$, for $i = 0, 1, 2, I$.

We first establish a lower bound on the total cost.

Theorem 3. *Assume that* (7), (14), (15), (20), *and* (21) *hold. We have*

$$C^* \geq a(\lambda - m_v^{-1}). \tag{24}$$

The proofs of the results in this section are presented in Online Appendix EC2. To understand the intuition behind this result, assume that it would be possible to keep a few customers in the system while keeping the investigator busy at all times. Then the reneging rate in the long run from the first queue would be approximately $\lambda - m_v^{-1}$ (recall that the investigator is assumed to be the bottleneck). Note that this is a lower bound on the reneging rate because the investigator can see at most m_v^{-1} customers per unit time in the long run. Therefore, (24) consists only of the reneging cost, and the occupancy cost in the system is zero. Although this bound might seem crude, we show below that there exists a policy that achieves this lower bound. A similar result also holds when the back office is the bottleneck (see §5.7).

Next we consider the LBF policy and show that this rule is asymptotically optimal.

Theorem 4. *Assume that* (6), (7), (11)–(15), (20), *and* (21) *hold. Under the LBF rule,*

$$\lim_{T\to\infty} \limsup_{r\to\infty} C_{\text{LBF}}^r(T) = C^*. \tag{25}$$

In the process of proving Theorem 4, we show that both the number of customers in the system and the number of interruptions are asymptotically negligible under LBF. Hence, LBF policy can be said to keep the system "empty." In ED applications, when B is interpreted as the maximum number of patients a physician can manage simultaneously (see §2.1), our result shows that the actual number of patients managed simultaneously is small under LBF. In addition, using this result, we can show that properly scaled flow times are also minimized. Let $W^r(t)$ denote the time a customer enters the system at time t. Because the system is "empty," it follows that $W^r(t)/r^\alpha$ also converges to zero. In this sense, LBF also minimizes the flow times of the customers in the system. In addition, one can deduce from (24) that the throughput of the system, which is equal to m_v^{-1}, is maximized under LBF.

Next, we analyze the asymptotic performance of the FBF policy.

Theorem 5. *Assume that* (6), (7), (11), (12)–(15), (20), *and* (21) *hold. Under the FBF rule,*

$$C_{\text{FBF}}^* = a\left(\lambda - m_v^{-1}\left(1 - \frac{B\nu}{\mu_I}\right)\right) + cB. \tag{26}$$

Remark 1. Clearly, the FBF policy is not asymptotically optimal because $B\nu/\mu_I < 1$ by assumption (12). Another interesting point is that, under FBF, if the number of slots in the system *increases*, the throughput *decreases*. This result is mainly because of the interruptions. As buffer capacity is increased, the FBF rule would keep more customers in the system, in turn increasing the number of interruptions the investigator must handle, which results in lower throughput.

Remark 2 (On the Asymptotic Framework). Our purpose in the paper is to analyze an overloaded system, which is formulated in (7), that has a large

buffer. We focus on the case when the buffer capacity is significantly lower than the average number of customers waiting to be admitted to the system. Therefore, we want Q_0^r to scale faster than $Q_1^r + Q_2^r$ in r. Because the buffer size in the rth system is of order r^α, $0 < \alpha < 1$, by (11), we want to scale the number of customers waiting to be admitted by r. Our analysis will still go through if we assume that it scales in any order larger than r^α.

Because customers in the first queue renege with rate γ, from the conservation of flow in queue 0, the expected number of customers in this queue should be at least $(\lambda - m_v^{-1})/\gamma$. Hence, (14) is needed to make sure Q_0^r is of the order r. However, this implies that the rate at which the system incurs the abandonment cost is $\gamma^r Q_0^r$. To obtain meaningful limits, we also enforce (15). These assumptions imply that customers become more and more patient but abandonment costs increase with r. Given that Q_0^r is of the order of r and $Q_1^r + Q_2^r$ is of the order of r^α, the scalings we use in (16) are natural.

Assumptions (12) and (13) are also needed to have a meaningful limit. They guarantee that the investigator has enough capacity to handle interruptions even when all the slots are full. This assumption is reasonable because, in designing a system, the capacity of the investigator will also be considered. Also, if ν^r is assumed to decrease in a slower order in r, no matter how we control the system, the investigator will be working only on the interruptions. If it decreases faster, interruptions play no role asymptotically.

Remark 3. Another interesting and more traditional asymptotic regime is to assume that the system reaches heavy traffic as $r \to \infty$, e.g., letting $1 - \lambda^r m_v^{-1} \approx r^{-1/2}$ and $B^r = r^{1/2}B$. This analysis, which we leave for future research, should be more appropriate for systems with $m_v \approx m_b$.

5.3. Sketch of the Proof of Theorems 3–5

The proofs of Theorems 3–5 are related, and the details are provided in Online Appendix EC2. Here, we give a brief outline and highlight the main points of their proofs. In proving these results, we first analyze the asymptotic behaviors of $\bar{Q}_1^r$, $\bar{Q}_2^r$, and $\bar{Q}_I^r$. We first show that the interruptions queue, $\bar{Q}_I^r$, is asymptotically "empty" under any admissible scheduling rule. This result is intuitive, because the interruptions queue has priority over other queues. Then we show that the total number of customers in the system under the LBF rule is asymptotically negligible. This conclusion in turn implies there are very few interruptions in the system under LBF. On the other hand, we show that the system is asymptotically full under FBF at all times.

Theorem 3 is established by using the fact that there are asymptotically no interruptions waiting in the system and by assuming that the system is asymptotically empty at all times. Proof of Theorem 4 then follows from the fact that the system is asymptotically empty under LBF. Finally, using the fact that the system is asymptotically full under FBF at all times, we prove Theorem 5.

The main technical challenge in our proofs is to establish the asymptotic results for $\bar{Q}_1^r$, $\bar{Q}_2^r$, and $\bar{Q}_I^r$. Our proof idea is similar to that in Bramson (1998). We first establish similar results for a solution of a set of deterministic equations, commonly referred to as the fluid model in the literature, and then we establish that they imply the desired result in the scaling we use. The scaling we use is different from the diffusion scaling used in Bramson (1998), so we provide a proof for the connection between the fluid model we introduce and our scaling.

5.4. Probabilistic Routing

In this section we relax the assumption on deterministic routing and assume that after receiving service in station i, each customer may be routed to station j with probability p_{ij} or leave the system with probability $1 - \sum_j p_{ij}$. For example, this may be the case when additional tests may be required based on the results obtained from an initial test. We assume that customers cannot be routed to the initial diagnosis stage, so $p_{i0} = 0$, for $i = 0, 1, 2$.

Theorems 3 and 4 still hold in this case, with the following modifications. Let $P = (p_{ij})$ denote the 3×3 routing matrix and assume that $(I - P')$ is invertible. For $\lambda^{in} \geq 0$, define $\lambda(\lambda^{in}) = (I - P')^{-1}\tilde{\lambda}^{in}$, where $\tilde{\lambda}^{in}$ is a three-dimensional column vector whose first element is λ^{in} and all its other elements are zero. For $\lambda(\lambda^{in}) = (\lambda_0(\lambda^{in}), \lambda_1(\lambda^{in}), \lambda_2(\lambda^{in}))$, the load on the investigator is $\rho_v(\lambda^{in}) = \lambda_0(\lambda^{in})/\mu_0 + \lambda_2(\lambda^{in})/\mu_2$ and on the back office $\rho_b(\lambda^{in}) = \lambda_1(\lambda^{in})/\mu_1$. Let

$$\lambda^* = \sup\{\lambda^{in}\colon \rho_v(\lambda^{in}) \leq 1, \rho_b(\lambda^{in}) \leq 1\}. \tag{27}$$

The rate λ^* is the maximum throughput for system. Assume that

$$\rho_v(\lambda^*) > \rho_b(\lambda^*); \tag{28}$$

i.e., the investigator is the bottleneck and the system is overloaded—

$$\lambda > \lambda^*. \tag{29}$$

Then Theorem 3 holds with (29) replacing assumption (7) and λ^* replacing m_v^{-1} in (24). Also, Theorem 4 holds with the same change in the assumptions and with also assuming (28) instead of (6). The proofs in this case are similar to those of Theorems 3 and 4. We provide the details in Online Appendix EC3.3. However, Theorem 5 may not hold in general in this case.

5.5. Multiple Resources

In some systems there may be multiple resources at each station. In this section we discuss how our results can be extended to systems with multiple resources. Assuming that there are $n \geq 2$ investigators working in the system, we define

$$m_v = n^{-1}\left(\frac{1}{\mu_0} + \frac{1}{\mu_2}\right)$$

and assume that all the conditions hold with this new m_v. In this setting, if each investigator follows the LBF policy and the investigators are pooled, Theorem 4 still holds. A similar analysis can be carried out when there are multiple servers in the back office.

An ED is a typical service system with multiple investigators. (In this case, B models the total number of patients that can be managed simultaneously by the collective group of physicians.) However, in EDs, physicians are not pooled; a patient is normally treated by the same physician in both stations 0 and 2. Analyzing the case when each customer has an assigned investigator can be done following the same approach as with a single investigator, but it would require keeping track of a larger state space. Rather than pursuing analytical results, in §5.6 we use simulations to verify that LBF still performs well when investigators are not pooled.

5.6. Numerical Results

In this section we use numerical experiments to illustrate our main results. We focus on the case when the arrival rate is large compared to the system capacity; hence, we assume there is always a customer waiting to be admitted. We examine the convergence rate of the throughput under LBF to the optimal level under different parameter values.

In the first experiment, we set $\mu_0 = \mu_1 = \mu_2 = 1$, $\mu_I = 10$, and $\nu = 0.1$ (the results of additional experiments are presented in Online Appendix EC4). Figure 4(a) shows throughput as a function of buffer size under three policies: an optimal policy (found numerically), FBF, and LBF. FBF is optimal or close to optimal when B is small. As B increases, the performance of LBF surpasses FBF and approaches the performance of the optimal policy.

In Figure 4(b), we plot the steady-state expected number of customers in the system as a function of buffer size under the three different policies. Under the FBF policy, the expected number of customers increases linearly with B, as it did in the model with no interruptions. Under the optimal policy, the expected number of customers appears to stabilize around two. Hence, increasing the buffer size after a certain point has marginal impact on throughput.

It is clear from these numerical experiments and our theoretical results that the throughput under LBF surpasses that under FBF as the number of slots gets larger. We denote with B^* the minimum number of slots needed for LBF to outperform FBF, and we investigate the effect of different parameters on B^*. We use as the numeraire, m_v, the total expected time spent by a patient at stations 0 and 2. We set m_v equal to 1 and let $\mu_v = 1/m_v$. We vary several factors:

- back-office rates as $\mu_1 \in \{1.1\mu_v, 1.2\mu_v, 1.5\mu_v\}$, investigating the cases for which the back office has 10%, 20%, and 50% more capacity than the investigator;
- interruption service rate $\mu_I \in \{5\mu_v, 10\mu_v\}$, investigating the cases for which the time it takes the investigator to service the interrupt is 1/5th and 1/10th of the time it takes her to provide the main service to the customer;
- interruption arrival rate $\nu \in \{5\%\mu_v, 20\%\mu_v\}$, investigating cases for which the interrupts are very infrequent and for which they are more frequent;

Figure 4 Comparison of Three Policies

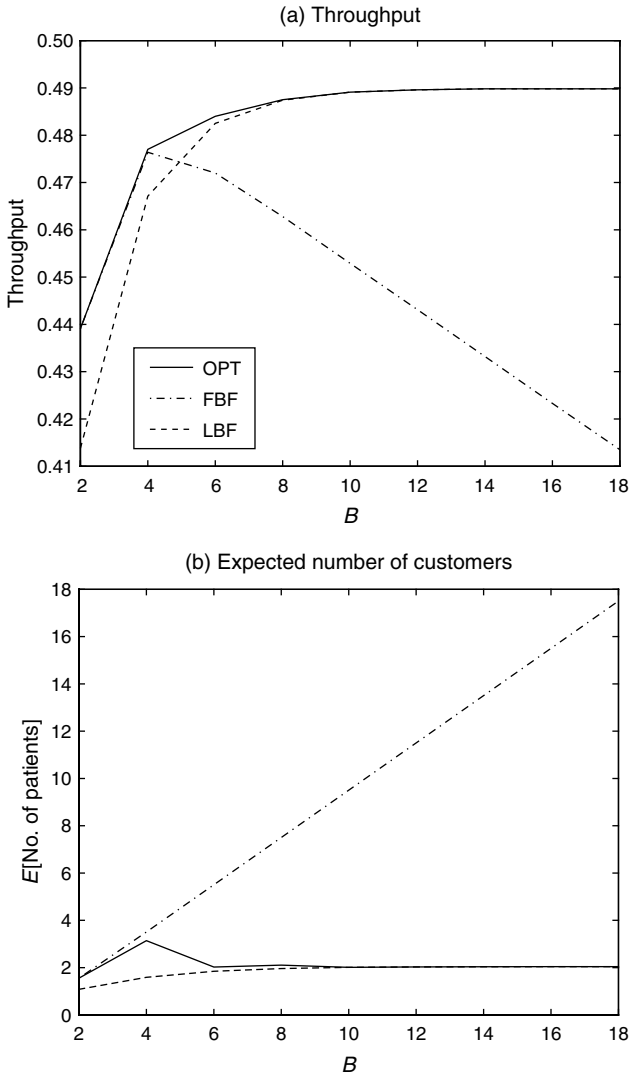


Table 1 B^* **Under Different Parameters**

		Interrupt arrival rate			
		$5\%\mu_v$		$20\%\mu_v$	
		Interrupt service rate		Interrupt service rate	
Back-office service rate	μ_0/μ_2	$5\mu_v$	$10\mu_v$	$5\mu_v$	$10\mu_v$
$1.1\mu_v$	3/7	8	12	4	6
	1	10	14	6	8
	7/3	14	18	6	10
$1.2\mu_v$	3/7	8	10	4	6
	1	10	12	6	8
	7/3	12	16	6	8
$1.5\mu_v$	3/7	6	8	4	6
	1	8	8	4	6
	7/3	10	12	6	8

- the ratio $\mu_0/\mu_2 \in \{3/7, 1, 7/3\}$ (keeping $m_v = 1$), investigating how the split in service time between stations affects the system.

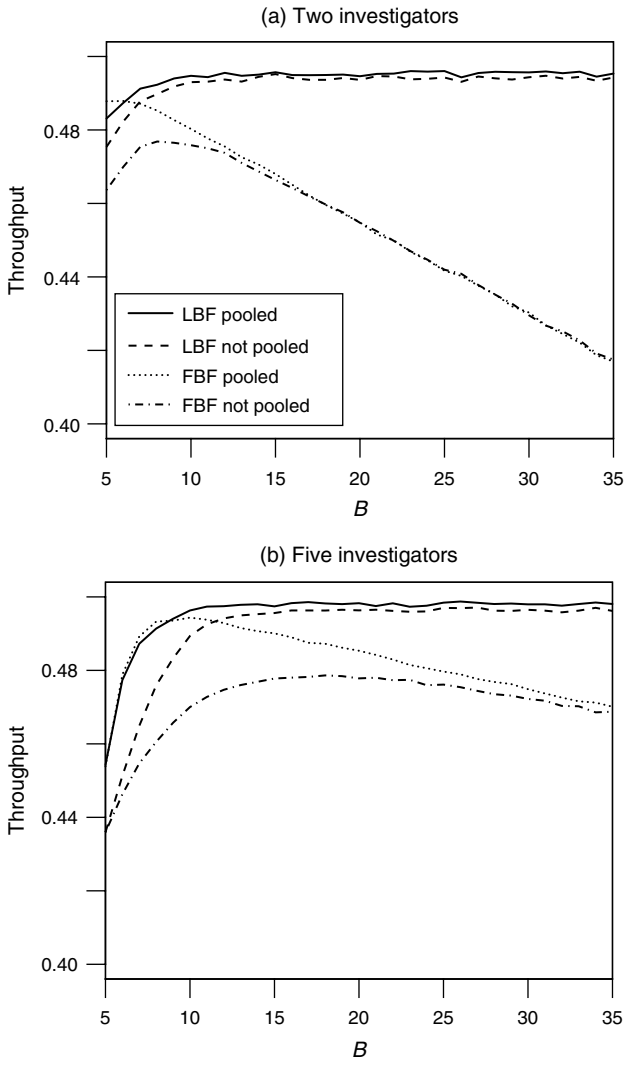
For each of the 36 set of parameters, we calculate the performance of the LBF, FBF, and optimal policies, varying the number of slots, B, in increments of two. The values of B^* are presented in Table 1. Some of these values of B^* may be as high as 18. A single physician is not likely to manage 18 patients at a time in an ED. However, if we consider an ED with multiple physicians, then there may very well be more than 18 slots available for the patients. Online Appendix EC4 includes additional results on the average number of occupied slots and on the differences between the throughputs of the LBF and optimal policies (found numerically) when the buffer capacity is B^*.

It is not surprising that B^* is increasing in the interruption service rate μ_I and decreasing in the interruption arrival rate ν. The faster the service of interruptions and the fewer the number of interruptions, the more the system appears as one without the interruptions, where the FBF policy is throughput optimal. The number of slots B^* also increases as the back-office service rate μ_1 decreases. As μ_1 decreases, the rate of the back office approaches the rate of the investigator. If the investigator prioritizes the existing jobs over the new ones (LBF), then the back-office server is more likely to be starved and throughput will be lower. Similarly, B^* is increasing in the ratio μ_0/μ_2, and the slower the diagnosis service, the more likely that the back-office server is starved.

Another fact we observe in our experiments is that for reasonable buffer size, say more than five slots, even when LBF is not optimal, the optimal policy is only slightly different from LBF. The only differences we observe are in the states in which there are no customers in queue 1 waiting to be served by the back office, but a few are in queue 2 (usually less than three) waiting for a diagnosis. In this case, especially when the service speed of the investigator and back office are close, it may be optimal to see a new customer instead of diagnosing an existing one. This fact is also evident in Figure 2.

Next we present the simulation results for multiple investigators and compare the cases in which the investigators are pooled and not pooled (see §5.5). First we consider systems with two and five investigators. We use the same parameters as the first experiment above ($\mu_0 = \mu_1 = \mu_2 = 1$, $\mu_I = 10$, and $\nu = 0.1$), dividing the service rate at stations 0 and 2 by the number of investigators, so that the nominal capacity of the system is fixed in each experiment. In Figures 5(a) and 5(b) we plot the throughput as a function of the number of slots in systems with two and five investigators, respectively. The trend is very similar to the case with a single investigator. Notably, in these experiments the number of slots per investigator at which LBF surpasses FBF is decreasing in

Figure 5 **Throughput in Systems With Multiple Investigators**



the number of investigators, making LBF more suitable for the ED applications. In addition, the difference between the throughputs when the investigators are pooled and not pooled appears to be very small, especially when the number of slots is not small.

5.7. When the Back Office Is the Bottleneck

So far we have focused on the case when the investigator is the bottleneck, i.e., (6) holds, which is the typical case in most systems. In this section, for completeness, we consider the case when the back office is the bottleneck. Specifically, we assume that

$$m_v < m_b. \tag{30}$$

When the back office is the bottleneck, the LBF policy is no longer optimal. In fact, because the back office is the bottleneck, having the investigator see a new customer whenever it is possible leads to overcrowding the system (recall that LBF would see a new customer if there are no customers in the last station). This policy leads to a performance similar to that of FBF, when the investigator is the bottleneck. In this section we present an ϵ-optimal (to be defined later) policy for the investigator.

Consider the following threshold policy for the investigator: define a constant $\kappa < 1$, then serve customers at station 0 if and only if $Q_1/B < \kappa$; otherwise, serve customers in queue 2, and if there are no customers in queue 2, then remain idle. The purpose of such a policy is to make sure that the back office almost never idles without overcrowding the system. In this setting Theorem 3 still holds with

$$C^* \geq a(\lambda - \mu_1). \tag{31}$$

The proof is similar and hence is skipped.

Now, for the asymptotic analysis, we apply the above threshold policy, with the number of beds in the rth system equal to B^r. Let $\pi^r(\kappa)$ denote the threshold policy with the threshold equal to κ and

$$C^*_{\text{threshold}} = \lim_{T\to\infty} \limsup_{r\to\infty} C^r_{\pi^r(\kappa)}(T).$$

We have the following result.

THEOREM 6. *Assume that* (11)–(15), (20), (21), *and* (30) *hold and that* $\lambda > m_b^{-1}$. *For any* $\epsilon > 0$, *there exists* $\kappa(\epsilon) > 0$ *such that under the threshold policy rule with* $\kappa(\epsilon)$,

$$C^*_{\text{threshold}} \leq C^* + \epsilon. \tag{32}$$

The proof of this result is similar to that of Theorem 4; hence, we provide only a sketch. Fix $\kappa > 0$. Essential to the proof is the following observation: under the conditions of Theorem 6,

$$\sup_{0<t<T} |\bar{Q}^r_0(t) - \kappa| \to 0 \quad \text{and} \quad \sup_{0<t<T} |\bar{Q}^r_1(t)| \to 0 \tag{33}$$

in probability as $r \to \infty$, under the threshold policy. Observe that (33) implies that the back office almost never idles (asymptotically) and that there are only κr^α customers in the system. By selecting κ small enough, one can prove (32).

When the back office is the bottleneck, it is also possible to derive an approximation for the throughput. Consider the threshold policy described above. We assume that service times are exponential and the service is preemptive; the investigator will start serving a customer in the first queue if the number of customers in the second station is below the threshold value $N = \kappa$. Then the number of customers in the second station can be modeled as a $M/M/1/N$ system, and the probability that the back office is idle in the steady state is given by

$$P_0 = \frac{\rho_1 - 1}{\rho_1^N - 1}, \tag{34}$$

where $\rho_1 = \mu_0/\mu_1$. Because we assume that the back office is the bottleneck, service rates at stations 0 and 2 must be faster than the service rate at station 1; therefore, we necessarily have $\mu_1 < \mu_0$. Hence, the throughput of the system can be approximated by $\mu_1(1 - P_0)$. In our simulation experiments, the approximation (34) seems to work especially well when μ_0 is not close to μ_1 and the interruption arrival rate and service times are relatively small, e.g., similar in the range to those tested in Table 1.

5.8. Other Cases

Now we discuss why the LBF policy and its variants are asymptotically optimal in the cases of separate limits on buffer space and no limits on buffer space. In the case of separate buffer space limitations, the analysis is very similar. The only difference is that instead of (11), we assume that

$$B^r_i = r^\alpha B_i, \text{ for } 0 < \alpha < 1 \quad \text{and} \quad B_i > 0, \ i = 1, 2. \tag{35}$$

Let $B = B_1 + B_2$. Then, under assumptions (6), (7), (12)–(15), (35), (20), and (21), Theorems 3 and 4 still hold.

If there is no limit on buffer space, showing the optimality of LBF is a little bit more involved. First we note that Theorem 3 still holds under (7), (14), (15), (20), and (21), because it does not require any of the buffer space assumptions. We also know that LBF attains this lower bound in the case of limited buffer space. Instead of applying LBF directly, we modify it as follows. Let $B > 0$ denote a constant. The investigator follows LBF as long as the total number of admitted customers is strictly less than B. If it is equal to B, the investigator never admits a new customer to the system and idles if there are no customers waiting to be discharged. Refer to this policy as LBF-B and

assume that (12), (14), and (15) hold. Now choose B^r in the rth system such that (11) and (13) hold. Then under LBF-B^r, Theorem 4 obviously holds with the following additional assumption on the initial state:

$$\bar{Q}_1(0) + \bar{Q}_2(0) < \mu_I/\nu. \tag{36}$$

This condition is similar to (13), but because there is no restriction on the number of customers within the system, it is possible at time 0 that there is an overwhelming number of customers in the system, in which case the investigator may have to handle interruptions forever. Thus, (36) is a reasonable assumption.

5.9. Implications of Our Asymptotic Results for ED Operations

The details of the relationship between the ED operations and our model are explained in §2.1. In this section we highlight the implications of our asymptotic results (mainly) on ED operations, using this relationship. We consider the case with a single physician handling at most B patients (in applications a physician usually handles fewer than 10 patients). The following implications are derived from our theoretical results as well as from our numerical experiments above.

(i) Our numerical results in Figure 3(b) (and in Online Appendix EC4) imply that, for a set of fixed parameters μ_0, μ_1, μ_2, μ_I, and B, there exists a threshold value ν^* for the interruption rate ν such that if $\nu \geq \nu^*$, then LBF performs better than FBF and otherwise FBF performs better than LBF.

(ii) Our theoretical results Theorems 3 and 4 imply that as B gets larger, ν^* (as defined in the previous item) becomes smaller (assuming all the other parameters are fixed), making LBF more and more attractive, and eventually ν^* converges to 0 as $B \to \infty$, implying asymptotic optimality of LBF. This can also be observed in Figure 3(b), because FBF keeps more and more customers in the system as B gets larger.

(iii) Our numerical results in §5.6 (and those that follow in Online Appendix EC5) indicate that the quality of our asymptotic results in smaller systems deteriorates when k gets closer to 1, i.e., when the load of the physician and the back office are balanced. The exact level of k depends on the other parameters.

(iv) As illustrated in Figure 3(b) (and in Online Appendix EC4), the expected number of patients in $Q_1 + Q_2$ under LBF is smaller than it is under FBF. Although this comes at a certain cost in terms of throughput in our model, especially when B is small (see Figure 3(a)), under LBF, a physician has to take care of fewer patients than he or she has to under FBF. Therefore, LBF may be even more attractive in ED applications because decreased multitasking is likely to decrease the service times of the physician (this feature is not included in our model).

Although our results have practical implications for emergency departments, a number of issues remain to be addressed. For example, it is worthwhile to pursue approximations for ν^*. In addition, the behavior of these systems seems to be more complicated when $k \approx 1$. More analysis needs to be done using the heavy-traffic framework similar to that in Harrison and Wein (1990) and Kumar (2000). We leave these directions for future research.

6. Conclusion

We considered a stylized model for a workflow that consists of an investigator, a back office, and a limited number of buffer slots available for customers. Customers receive service in three stations: in the first station, the investigator collects initial data, performs an initial interview, and generates the orders for the back office; in the second station, the back office performs the work ordered by the investigator; and in the last station, the investigator delivers the result or diagnosis based on the results from the initial exam and the work done by the back office. Customers who spend too much time in the system generate interruptions—additional work that must be handled by the investigator. In this setting, the investigator has to determine whether to serve a customer in the first station or a customer in the last station. By focusing on overloaded systems, we analyze the throughput optimal workflow decisions for the investigator. Our main findings are as follows.

For systems with no interruptions, the FBF policy, which always gives priority to seeing new customers, maximizes the throughput when there is a limit on the total buffer space in the system. If there is no limit, LBF is throughput optimal. When there are separate limits on the size of the buffer in front of each station, the throughput optimal policy is not known.

For systems with interruptions, our findings are as follows:

- The exact optimal policy is complicated, with the investigator switching priorities between seeing new customers and handling customers already in the system.
- The throughput under FBF decreases (asymptotically) in the number of available buffer slots.
- The LBF policy, which always gives priority to serving customers already in the system, is asymptotically optimal in maximizing the throughput when there is a large limit on the total buffer size or separate limits on each buffer. If the buffer size is unlimited, then a slight variation of LBF policy is asymptotically optimal.
- The LBF policy is nearly optimal even in relatively small systems.
- For systems with interruptions, we recommend that investigators in the system give priority to station 2 except when the queue in front of the back

office is almost empty and there are just a few customers waiting before station 2.

The last recommendation has implications for the design of information systems that support service facilities. For instance, we observed information systems in a number of hospital EDs. In these EDs, separate systems were used for processing imaging studies, laboratory tests, nursing notes, etc. It was, therefore, difficult to create a single signal that all the physician's orders were complete. As new systems are put in place or the legacy systems integrated, adding such functionality could help make the EDs more efficient, enabling the LBF policy.

In this paper, we focus only on overloaded systems. For systems that are critically loaded, the optimal policy for the investigator may be different. A more traditional heavy-traffic analysis may be more suitable for those situations to identify the optimal policy.

Acknowledgments

The research of Tolga Tezcan was supported by the National Science Foundation [Grants CMMI-0954126, CMMI-1130346].

References

Ahn H, Righter R (2006) Dynamic load balancing with flexible workers. *Adv. Appl. Probab.* 38(3):621–642.

Ahn H, Duenyas I, Lewis ME (2002) Optimal control of a two-stage tandem queuing system with flexible servers. *Probab. Engrg. Informational Sci.* 16(4):453–469.

Ahn H, Duenyas I, Zhang RQ (1999) Optimal stochastic scheduling of a two-stage tandem queue with parallel servers. *Adv. Appl. Probab.* 31(4):1095–1117.

Altman E (2002) Applications of Markov decision processes in communication networks: A survey. Feinberg EA, Shwartz A, eds. *Handbook of Markov Decision Processes* (Kluwer, Dordrecht, The Netherlands).

Andradottir S, Ayhan H (2005) Throughput maximization for tandem lines with two stations and flexible servers. *Oper. Res.* 53(3):516–531.

Andradottir S, Ayhan H, Down DG (2001) Server assignment policies for maximizing the steady-state throughput of finite queueing systems. *Management Sci.* 47(10):1421–1439.

Andradóttir S, Ayhan H, Down DG (2008) Maximizing the throughput of tandem lines with flexible failure-prone servers and finite buffers. *Probab. Engrg. Informational Sci.* 22(2):191–211.

Argon NT, Ziya S (2009) Priority assignment under imperfect information on customer type identities. *Manufacturing Service Oper. Management* 11(4):674–693.

Arumugam R, Mayorga M, Taaffe K (2009) Inventory based allocation policies for flexible servers in serial systems. *Ann. Oper. Res.* 172(1):1–23.

Bartholdi JJ III, Eisenstein DD (1996) A production line that balances itself. *Oper. Res.* 44(1):21–34.

Bell SL, Williams RJ (2001) Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Ann. Appl. Probab.* 11(3):608–649.

Bell SL, Williams RJ (2005) Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy. *Electron. J. Probab.* 10(33):1044–1115.

Besbes O, Maglaras C (2009) Revenue optimization for a make-to-order queue in an uncertain market environment. *Oper. Res.* 57(6):1438–1450.

Bramson M (1998) State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems: Theory Appl.* 30(1–2):89–148.

Bramson M, Dai JG (2001) Heavy traffic limits for some queueing networks. *Ann. Appl. Probab.* 11(1):49–90.

Chevalier PB, Wein LM (1993) Scheduling networks of queues: Heavy traffic analysis of a multistation closed network. *Oper. Res.* 41(4):743–758.

Chisholm CD, Collison EK, Nelson DR, Cordell WH (2000) Emergency department workplace interruptions: Are emergency physicians "interrupt-driven" and "multitasking"? *Acad. Emergency Medicine* 7(11):1239–1243.

Dai JG, Tezcan T (2008) Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems* 59(2):95–134.

Dai JG, Tezcan T (2011) State space collapse in many server diffusion limits of parallel server systems. *Math. Oper. Res.* 36(2):271–320.

Dai JG, Wuqin L (2008) Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Ann. Appl. Probab.* 18(6):2239–2299.

Dobson G, Sainathan A (2011) On the impact of analyzing customer information and prioritizing in a service system. *Decision Support Systems* 51(4):875–883.

Duenyas I, Gupta D, Olsen TL (1998) Control of a single-server tandem queueing system with setups. *Oper. Res.* 46(2):218–230.

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

Gel ES, Hopp WJ, Van Oyen MP (2002) Factors affecting opportunity of worksharing as a dynamic line balancing mechanism. *IIE Trans.* 34(10):847–863.

Gilland WG (2001) Effective sequencing rules for closed manufacturing networks. *Oper. Res.* 49(5):759–770.

Gurvich I, Whitt W (2009) Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* 11(2):237–253.

Harrison JM, Wein LM (1990) Scheduling networks of queues: Heavy traffic analysis of a two-station closed network. *Oper. Res.* 38(6):1052–1064.

Hu B, Benjaafar S (2009) Partitioning of servers in queueing systems during rush hour. *Manufacturing Service Oper. Management* 11(3):416–428.

Iravani SMR, Posner MJM, Buzacott JA (1997) A two-stage tandem queue attended by a moving server with holding and switching costs. *Queueing Systems: Theory Appl.* 26(3–4):203–228.

Kaufman D, Ahn H, Lewis M (2005) On the introduction of an agile, temporary workforce into a tandem queueing system. *Queueing Systems: Theory Appl.* 51(1):135–171.

Kirkizlar HE (2008) Performance improvements through flexible workforce. Unpublished doctoral dissertation, Georgia Institute of Technology, Atlanta.

Kirkizlar HE, Andradottir S, Ayhan H (2012) Flexible servers in understaffed tandem lines. *Production Oper. Management* 21(4):761–777.

Kumar S (2000) Two-server closed networks in heavy traffic: Diffusion limits and asymptotic optimality. *Ann. Appl. Probab.* 10(3): 930–961.

Mandelbaum A, Stolyar AL (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* 52(6):836–855.

McClain JO, Thomas LJ, Sox C (1992) "On-the-fly" line balancing with very little WIP. *Internat. J. Production Econom.* 27(3):283–289.

Oredsson S, Jonsson H, Rognes J, Lind L, Göransson KE, Ehrenberg A, Asplund K, Castrén M, Farrohknia N (2011) A systematic review of triage-related interventions to improve patient flow in emergency departments. *Scandinavian J. Trauma, Resuscitation and Emergency Medicine* 19(43), http://www.sjtrem.com/content/19/1/43.

Rosberg Z, Varaiya P, Walrand J (1982) Optimal control of service in tandem queues. *IEEE Trans. Automatic Control* 27(3):600–609.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Complexity-based triage: A tool for improving patient safety and operational efficiency. Ross School of Business Paper 1161, University of Michigan, Ann Arbor.

Speier C, Valacich J (1999) The influence of task interruption on individual decision making: An information overload perspective. *Decision Sci.* 30(2):337–360.

Van Oyen MP, Hopp WJ (2004) Agile workforce evaluation: A framework for cross-training and coordination. *IIE Trans.* 36(10):919–940.

Westbrook, JI, Woods A, Rob MI, Dunsmuir WTM, Day RO (2010) Association of interruptions with an increased risk and severity of medication administration errors. *Arch. Internal Medicine* 170(8):683–690.

Whitt W (1999) Partitioning customers into service groups. *Management Sci.* 45(11):1579–1592.

Wu C, Down DG, Lewis ME (2008) Heuristics for allocation of reconfigurable resources in a serial line with reliability considerations. *IIE Trans.* 40(6):595–611.

Wu C, Lewis ME, Veatch M (2006) Dynamic allocation of reconfigurable resources in a two-stage tandem queueing system with reliability considerations. *IEEE Trans. Automatic Control* 51(2):309–314.

Zavadlav E, McClain JO, Thomas LJ (1996) Self-buffering, self-balancing, self-flushing production lines. *Management Sci.* 42(8):1151–1164.