Adv. Appl. Prob. **45,** 673–691 (2013)
Printed in Northern Ireland

PARTIAL FLEXIBILITY IN ROUTEING AND SCHEDULING

OSMAN T. AKGUN,* **
RHONDA RIGHTER * AND
RONALD WOLFF,* *University of California, Berkeley*

Abstract

We consider partial customer flexibility in service systems under two different designs. In the first design, flexible customers have their own queue and each server has its own queue of dedicated customers. Under this model, the problem is a scheduling problem and we show under various settings that the dedicated customers first (DCF) policy is optimal. In the second design, flexible customers are not queued separately and must be routed to one of the server's dedicated queues upon arrival. We extend earlier results about the 'join the smallest work (JSW)' policy to systems with dedicated as well as flexible arrivals. We compare these models to a routeing model in which only the queue length is available in terms of both efficiency and fairness and argue that the overall best approach for call centers is JSW routeing. We also discuss how this can be implemented in call centers even when work is unknown.

Keywords: JSW; scheduling; workload routeing; majorization; sample path analysis

2010 Mathematics Subject Classification: Primary 90B22
Secondary 60K25

1. Introduction

We study a multiserver multiqueue system with partial flexibility. Each server has a queue of dedicated customers that can only be served by that server, and there are also flexible customers that can be served by more than one server. This models both server and customer flexibility. For example, in a call center serving customers in multiple languages, servers (agents), and most customers, could be monolingual (dedicated), with some customers being bilingual (flexible). Alternatively, servers could have bilingual training, with a dedicated language (say Spanish or Mandarin) and a shared (flexible) language (say English), and customers could be monolingual; in this case the English speaking customers would be considered flexible. Given a system with partial flexibility, the question is how best to take advantage of it to minimize average customer waiting times. This depends on whether flexible customers must be immediately routed to a dedicated queue upon arrival (the routeing problem), or whether they have their own queue and the decision is when to assign them to an available server (the scheduling problem). For the routeing problem, the decision depends on the available information. In earlier work [3], we considered the routeing problem when the available information is the queue length, and we gave conditions under which the optimal policy for flexible customers is JSQ (join the shortest queue), in which customers are routed upon arrival to the shortest queue (this is both

Received 2 August 2012; revision received 20 November 2012.
* Postal address: Department of Industrial Engineering and Operations Research, University of California, Berkeley, 4141 Etcheverry Hall, Berkeley, CA 94720, USA.
** Email address: akguno@ieor.berkeley.edu

the socially optimal policy, as well as the individually optimal policy for flexible customers). Here we consider the case when the information available upon arrival of a flexible customer is the workload at each of the servers, and we give conditions under which JSW (join the shortest work) is the optimal routeing policy. We also consider the scheduling problem, and show that, under many conditions DCF (dedicated customers first), is the optimal policy. This design and policy is also the most efficient of the three, in terms of minimizing the overall customer waiting time, but at the expense of the flexible customers, who end up waiting longer on average. This would not be acceptable in call centers, though might be in other queueing systems. We show how JSW can be implemented in call centers, without actually knowing the workloads, and we argue that it is the best design for taking advantage of partial flexibility. It performs better than JSQ and almost as well as DCF in terms of minimizing waiting times, and it is incentive compatible for flexible customers in the sense that it is their individually optimal policy.

Design of call centers has been a popular research area in queueing systems. Aksin *et al.* [4] and Gans *et al.* [12] provided extensive literature surveys. Our flexible scheduling model can be thought of as an example of the network topology known as the 'W' design [12]. Flexibility in service systems has been studied widely under different topologies. A simpler model, the 'N' design, where only one server has dedicated customers, and there is a queue for flexible customers, has been studied widely in the literature. In this case the problem is to determine when the server with the dedicated queue should 'help' the other server, which can only serve flexible customers. Garnett and Mandelbaum [13] showed the difficulty of finding the optimal control policy by considering different examples with different parameters. Harrison [16] constructed a discrete review control policy in which the heavy-traffic limit approaches the bound of a single pooled resource. Bell and Williams [6] showed the optimality of a threshold-type control policy in heavy traffic. Ahn *et al.* [2] considered a clearing system (without arrivals). They showed that even for the clearing system the optimal policy is complex and can either have a monotone switching curve structure or it can be exhaustive for one of the queues. Down and Lewis [8] considered the problem with arrivals and with an upgrading option for low priority customers, and gave conditions under which the $c\mu$-rule is optimal.

The 'W' design is even more complex than the 'N' design and the problem has usually been studied using heavy-traffic approximations or using heuristic control policies. Harrison and Lopez [17] gave conditions for the optimality of a discrete review control policy in heavy traffic for a general model with arbitrary customer and server flexibility, of which the 'W' design is a special case. Mandelbaum and Stolyar [27] also considered a general structure with arbitrary customer and server flexibility and showed the optimality of the generalized $c\mu$-rule in heavy traffic. Gurumurthi and Benjaafar [14] analyzed control policies under arbitrary customer and server flexibility, and they compared the performance of different control policies such as serve the longest queue first (LQF) and strict priority (SP). Saghafian *et al.* [32] studied the 'W' network with Poisson arrivals, exponential service times, Poisson service disruptions, and with preemption permitted. They proposed a control policy called the largest expected workload cost (LEWC), and they compared its performance with the performance of the $c\mu$-rule, the generalized $c\mu$-rule, and the LQF policy. They also gave the conditions under which it is optimal to serve fixed tasks before shared tasks (dedicated before flexible customers) without idling. This result is a special case of Theorem 2 below where we consider a very general arrival process.

For the 'W' design, we show that in many situations the DCF policy is optimal. Under DCF, whenever a server's dedicated queue is nonempty, it gives priority to dedicated customers *and* does not idle.

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

Note that flexible customers tend to be disadvantaged under DCF. We consider the alternative approach of routeing flexible customers upon arrival to a server and then serving both flexible and dedicated customers at each server according to FCFS (first-come–first-served). When the queue lengths are not known upon arrival, Ephremides *et al.* [9] showed that the 'round-robin' policy minimizes the sum of expected completion times when the service times are exponential. Liu and Towsley [25] extended the optimality of the round-robin policy to service times with increasing hazard rate and Liu and Righter [24] showed the optimality of the round-robin policy for general independent and identically distributed (i.i.d.) service time distributions. When the only information available upon arrival is the queue length and all customers are flexible, routeing the flexible customers to the shortest queue is optimal under various settings (see, e.g. [3], [5], [9], [10], [18], [20], [23], [29], [30], [34], [35], [37], [38], and [39]). When the processing times are known upon arrival, Harchol-Balter *et al.* [15] proposed a routeing policy called the 'size interval task assignment with equal load (SITA-E)' where different servers are assigned to jobs with service times falling into particular intervals, and compared the performance of their policy with the performance of the JSW policy under different problem parameters. Hyytia *et al.* [19] discussed the computation of value functions based on different levels of information for the routeing problem.

In this paper we propose a design in which the actual workload at each queue is known upon arrival but the required work of the arriving customer is unknown. Note that in this model, when all customers are flexible, the JSW policy is equivalent to the FCFS policy with a single queue for all customers. Wolff [40], [41] showed that FCFS provides a lower bound for the workload in the system for all policies that are not allowed to depend on the workload, such as round robin. Daley [7], building on Foss's work [11], showed that JSW stochastically minimizes the workload at each time t for general arrival and service processes using weak submajorization arguments. See also [26]. Koole [22] showed the same result using dynamic programming arguments. Stoyan [36] and Koole [22] provided counterexamples showing that the pathwise optimality (jointly across t) of JSW for minimizing the workload is not true. However, Koole [21] showed that departures are sample pathwise maximized by the JSW policy. We extend the earlier results to systems with homogeneous dedicated customers as well as flexible customers. We show that among routeing policies, the JSW policy stochastically maximizes the departure process pathwise, and stochastically minimizes the workload at each time t.

Under certain conditions, we show that in terms of overall efficiency (e.g. minimizing overall sojourn times), having a separate queue for flexible customers and using DCF outperforms the immediate routeing of flexible customers using JSW, and this, in turn, outperforms JSQ routeing. However, as mentioned before, using DCF with a separate flexible queue is unfair to flexible customers; their performance is worse than under the alternatives, and worse than that of dedicated customers. Hence, in practice, flexible customers would want to declare themselves as dedicated customers if they knew that otherwise the policy would discriminate against them (so, for example, instead of pressing '1' for 'bilingual', they would press '2' for 'English'). We therefore conclude that the best policy overall is JSW routeing for flexible customers. Note that this can be implemented, even when the workload is not known (e.g. in call centers), as follows. Upon arrival suppose that multiple (virtual) copies are created for a flexible customer and a copy is sent to each queue. When one of these copies gets the chance to start service, then the real customer is served at that particular server and the other copies are deleted. We show that following the JSW routeing policy for flexible customers stochastically minimizes the sojourn time of *dedicated* customers (as well as of customers overall) among all routeing policies. Also, emprically, the overall performance of JSW is very close to that

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

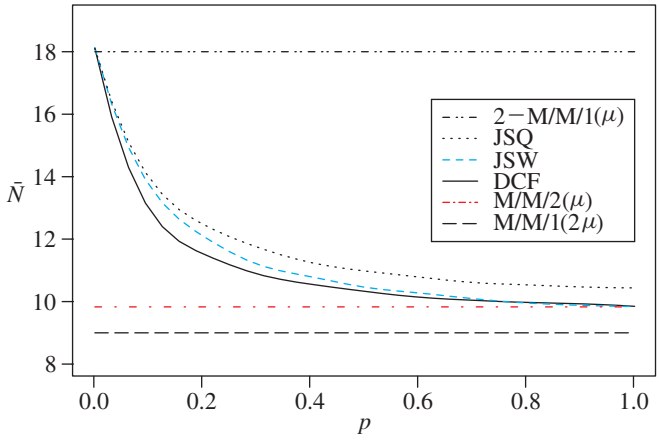

FIGURE 1: Comparison of policies. Total number in the system ($\bar{N}$) versus the proportion of flexible customers (p).

of DCF (see Figure 1). Moreover, JSW is incentive compatible for flexible customers in the sense that, given any fixed policy for all other flexible customers, an arriving flexible customer minimizes its own sojourn time by joining the queue with the shortest work.

The outline of the paper is as follows. In Section 2 we study the 'W' design scheduling problem. In Section 3 we consider the routeing problem. Finally, in Section 4 we compare the performance of different designs.

2. The scheduling problem

In this section we study the problem where flexible customers have their own queue and each server should decide which type of customer to serve next upon a service completion at that particular server. Service time distributions may depend on the server, but (generally) not on the customers. We give a variety of conditions for the arrival and service processes, and restrictions on service disciplines, under which the departure process is stochastically maximized by a DCF policy. Note that although we define the DCF policy as one that does not idle a server when it has dedicated customers, it could idle when there are flexible customers present but no dedicated customers. Also, note that a DCF policy does not specify the service discipline among dedicated customers. It is well known that, for a single-server queue, the FCFS discipline among all nonpreemptive, work conserving, nonanticipating service disciplines, minimizes the stationary sojourn time in the convex ordering sense (see [42]). See Righter and Shantikumar [31] and Aalto *et al.* [1] for optimal preemptive policies within a single class of customers. For example, if service times are DFR (decreasing failure rate), LAST (least attained service time) stochastically minimizes the number of customers in the system among all work conserving nonanticipating policies.

In the following, when we say DCF, we mean a policy that gives priority to dedicated customers *and* is nonidling for dedicated customers. When we say nonidling DCF, we mean that the policy is *also* nonidling for flexible customers. In this case, when preemption is not permitted, a nonidling DCF policy completely specifies the policy for all states.

We also assume, for ease of exposition, that there are two servers, server A and server B. All the results can easily be extended to more than two servers and multiserver stations.

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

TABLE 1: Summary of the results.

Model				Objective	Optimal policy
Preemption permitted	Idling permitted	Arrivals	Services		
Yes	Yes (or no)	General	Exponential (different rates)	Stochastically maximize $\{D(t)\}_{t=0}^{\infty}$	Nonidling DCF
No	Yes	General	General	Stochastically maximize $\{D(t)\}_{t=0}^{\infty}$	DCF
No	Yes	Poisson (homogeneous)	Exponential (same rates)	Minimize mean sojourn time	Nonidling DCF
No	No	General (homogeneous)	Exponential (same rates)	Weakly submajorize queue length process	Nonidling DCF

We consider both preemptive and nonpreemptive policies. When preemption is permitted, a job in service can be removed from that server at any time. The preempted job can resume service at a later time on the same server if it is dedicated, or on either server if it is flexible. We define a very general preempt-resume methodology in the next section. Within both preemptive and nonpreemptive classes of policies, we also consider two subclasses, policies where idling is permitted and those where we force nonidling. We consider different arrival and service processes within each setting. We also present counterexamples showing cases where the results do not hold for more general arrival or service processes. We summarize our results in Table 1 for the different policy classes. Let $\{D_t\}_{t=0}^{\infty}$ denote the departure process, where D_t is the number of departures by time t. We also show that idling will not be optimal when preemption is permitted; hence, the yes or no case for idling in Table 1.

2.1. Preemption and idling are permitted

We first consider permitting both preemption and idling. In this case, flexible customers can be removed from or assigned to any server at any time, whereas dedicated customers can only be reassigned to their dedicated server. We assume work is done in a preempt-resume fashion as described below. In the literature, preempt-resume has only been defined for a single service time distribution. We must extend the notion for flexible customers that can receive service from multiple servers, each with its own service time distribution. Rougly, by preempt-resume, we mean that work that is done on a customer before preemption is not lost. More rigorously, let server A (B) have service time distribution F_A (F_B) which is continuous with a positive density f_A (f_B). Let $\{U_k\}_{k=1}^{\infty}$ be a sequence of uniform(0, 1) random variables denoting the required work of the customers, and let $X_{i,k} = F_i^{-1}(U_k)$ be the service time of customer k if served completely on server i, $i = A, B$. At any time we define the completed work for customer k, U_k^c and the equivalent completed service time if served on server i, $X_{i,k}^c$, such that

$$U_k^c < U_k, \qquad X_{i,k}^c = F_i^{-1}(U_k^c),$$

and where upon arrival of customer k, $U_k^c = 0$. Suppose that customer k with completed work U_k^c begins or returns to service at time $t = 0$ at server i where it is served until service is either interrupted or completed at some $t > 0$, so, either $t < X_{i,k} - X_{i,k}^c$ or $t = X_{i,k} - X_{i,k}^c$ and service is complete. In the former case, we update the completed work using the relation $U_k^c = F_i(X_{i,k}^c + t)$. This is repeated each time customer k resumes service. Note that, in fact,

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

our construction and result for this section does not require the work of the customers ($\{U_k\}_{k=1}^{\infty}$) to be independent. Also, the distributions could be customer dependent.

In this section we assume the arrival process is independent of the state of the system and the policy, but is otherwise arbitrary.

We have the following lemma.

Theorem 1. *When preemption is permitted, idling is not optimal. That is, for each policy that idles, there exists a policy that does not, such that* $\{D_t\}_{t=0}^{\infty}$ *is stochastically larger under the nonidling policy.*

Proof. Consider a policy Π that, without loss of generality, idles server A at $t = 0$ when there is at least one job waiting in queue A or the flexible queue. Let $\tilde{\Pi}$ be the policy that starts serving one of the waiting customers in either queue A or the flexible queue at $t = 0$ (call it customer 1). We couple the future arrival processes for these systems so that each customer has the same arrival epoch in both systems. We also couple the service times (the U_ks) in both systems. We let $\tilde{\Pi}$ follow the same decisions as Π after $t = 0$ (this is possible since preemption is permitted), except for the times where Π serves customer 1 on one of the servers. During these times we will let $\tilde{\Pi}$ serve customer 1 on the same server as Π if it is still in the system, and we let $\tilde{\Pi}$ idle the server that under Π is serving customer 1 if customer 1 has already departed the system under $\tilde{\Pi}$. Then all departure times under $\tilde{\Pi}$ and Π will be the same except for customer 1, which departs earlier under $\tilde{\Pi}$, because the completed work of a customer is strictly increasing in the time it spends in service under the preempt-resume discipline described above, i.e. remaining service time of customer 1 will be less under $\tilde{\Pi}$ than under Π, regardless of the type of server it uses. We can repeat the argument each time our new policy idles until we have a nonidling policy.

We will show, through the counterexample below, that when preemption is permitted, DCF will not, in general, be optimal in the sample path sense. An exception is the case of exponential services, which we examine later.

Example 1. Suppose that preemption and idling are both permitted and the required work for all customers is deterministic and equal to 1. Let the service rates for servers A and B respectively be μ_A and μ_B, with $\mu_A < \mu_B$ (so the service times for dedicated customers are $1/\mu_A$ and $1/\mu_B$, respectively). Suppose that at time $t = 0$ there is one job on server B which will finish at ε, one job waiting in the flexible queue, one job waiting in server A's dedicated queue, server A is idle, and that there are no future arrivals. Then the nonidling DCF policy will start serving a dedicated customer on server A at $t = 0$. Let Π start serving the flexible customer on server A at $t = 0$ and then let it move it to server B at time ε and start serving the dedicated customer on A at ε. Then the flexible customer will depart at $\varepsilon + (1 - \mu_A\varepsilon)/\mu_B$ under Π, which is earlier than either departure under DCF (at $1/\mu_A$ and $\varepsilon + 1/\mu_B$).

In the above example DCF is still optimal in the mean sense, i.e. the expected sojourn time of customers is smaller under DCF. The next example shows that even with identical servers, and even in the mean sense, DCF is not necessarily optimal.

Example 2. Let all service times be i.i.d. with distribution S, where

$$S = \begin{cases} 1 & \text{with probability } 0.5, \\ 101 & \text{with probability } 0.5. \end{cases}$$

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

Suppose that at time $t = 0$ there is one dedicated customer at server A, one dedicated customer at server B and one flexible customer in the system, and none of the customers have received any service. Suppose that we do not have any arrivals after $t = 0$. Under DCF, the expected sojourn time of the dedicated customers will be 51 each and the flexible customer's expected sojourn time will be 77 (the expected waiting time is 101 with probability 0.25, and 1 with probability 0.75, and the expected service time is 51). Now consider another policy Π which serves the flexible customer at one of the servers, say A, for 1 unit, and then serves the dedicated customer to completion on that server. Server B serves its customer to completion. The flexible customer (if it does not complete service at time 1) is served to completion by whichever of servers A and B finishes serving its dedicated customer first. Under Π, the expected sojourn time of the dedicated customer at server A will be 52, the expected sojourn time of the dedicated customer at server B will be 51, and the flexible customer's expected sojourn time will be 63.625 (1 with probability 0.5, 201 with probability 0.125, 101 with probability 0.250, and 102 with probability 0.125). Hence, the overall mean sojourn time under Π will be smaller than under the DCF policy.

The next theorem shows that a nonidling DCF policy is optimal in the case of exponential services. Note that because preemption is permitted, under a nonidling DCF policy, servers never idle and flexible customers may be served by both servers. In the case where there is only a single flexible customer, it is served by the faster server.

Theorem 2. *Let the service times at each server be i.i.d. exponential random variables. The servers' rates may differ. If idling and preemption are both permitted, then* $\{D_t\}_{t=0}^{\infty}$ *is stochastically maximized by a nonidling DCF policy. Hence, the same policy is also optimal when idling is not permitted.*

Proof. It is easy to see that, when there is only a single flexible customer, it should be served by the faster server. We show that, for an arbitrary policy Π that does not follow DCF at an arbitrary decision epoch, we can construct a policy that does follow DCF for that decision and that has stochastically earlier departures. Let t be the first time that Π disagrees with DCF and serves a flexible customer when there are dedicated customers for that particular server (call this server A). Let $\tilde{\Pi}$ be a policy that agrees with Π before time t but that serves a dedicated customer from server A's dedicated queue at time t. Consider two systems where Π and $\tilde{\Pi}$ are used respectively as control policies. We couple the future arrival processes for these systems so that each customer has the same arrival epoch in both systems. We also couple the service times at each server in both systems. Let $\tilde{\Pi}$ also agree with Π for service on B while the customer on A is being served.

We first consider the case where Π does not preempt the first service on A before completion. Then, once the first service completes on A after t, at time t' say, Π has one fewer customer in the flexible queue and one more customer in server A's dedicated queue than $\tilde{\Pi}$. We let $\tilde{\Pi}$ agree with Π after t', except the first time Π serves a dedicated customer on server A, $\tilde{\Pi}$ serves a flexible customer on server A. Note that $\tilde{\Pi}$ may then idle server B by choice when Π is forced to idle it (because both the flexible queue and server B's dedicated queue are empty under Π). Then the departures under Π and $\tilde{\Pi}$ will be identical, so $\tilde{\Pi}$ is as good as Π, and a nonidling policy will be better than $\tilde{\Pi}$ by Theorem 1.

Next we consider the case where the flexible customer that starts service on server A at t under Π is preempted before service completion. In this case we let the dedicated customer on server A under $\tilde{\Pi}$ be preempted as well. Then, by the memoryless property of the exponential service, the two systems under Π and $\tilde{\Pi}$ will be stochastically identical, so again $\tilde{\Pi}$ is as good as Π.

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

2.2. Preemption is not permitted

We now consider the case where preemption is not permitted, i.e. once a customer starts service, it has to stay in service until completion. This means that a flexible customer will only be served by one of the two servers, so, when it starts service on a server, it effectively becomes dedicated to that server.

2.2.1. *Idling is permitted.* We first permit idling and show that DCF is optimal for general service times and arrival processes. In Section 2.1 the service process was general and was a customer property, i.e. each customer had an arbitrary amount of required work which was transformed to its service time at a particular server. In the following, since preemption is not permitted, we assume that the service process is a property of the server and that successive service times at each server are an arbitrary sequence of numbers (that do not depend on the customers). That is, we consider an arbitrary realization of service times.

Theorem 3. *Let server j have arbitrary service process $\{S_k^j\}_1^\infty$, $j = A, B$, and let the arrival process be independent of the system state and the policy but otherwise arbitrary. If idling is permitted and preemption is not permitted, then $\{D_t\}_{t=0}^\infty$ is stochastically maximized by a DCF policy.*

Proof. The first part of the proof, that dedicated customers should have priority, is the same as the first (nonpreemptive) case in the proof of Theorem 2. To show that a server with dedicated customers should not idle, consider a policy Π that, without loss of generality, idles server A at $t = 0$ when there is at least one job waiting in queue A. Let $\tilde{\Pi}$ be the policy that starts serving one of the dedicated customers in queue A at $t = 0$ (call it customer 1), and let it agree with Π for server B. Let t_1 denote the departure time of customer 1 under $\tilde{\Pi}$. Since we know dedicated customers should have priority, we can assume without loss of generality that Π will serve a dedicated customer (customer 1) on server A before serving a flexible customer. Suppose that it starts serving customer 1 at t_2 on server A. We couple the service time of customer 1, t_1, under both policies so that customer 1 departs at $t_2 + t_1$ under Π. Let $\tilde{\Pi}$ idle server A from t_1 until $t_1 + t_2$ and follow the same policy as Π after $t_1 + t_2$. Then all departure times under $\tilde{\Pi}$ and Π will be the same except for customer 1 which departs earlier under $\tilde{\Pi}$.

The next question to consider is whether or not a server will idle when there are flexible customers present. Unlike Theorem 3, sample pathwise optimality (stochastic maximization of $\{D_t\}_{t=0}^\infty$) for a nonidling policy for flexible customers will not, in general, hold. Consider the following counterexample.

Example 3. Suppose that preemption is not permitted and that at time $t = 0$ there is a flexible arrival to an empty system. Let Π idle both servers whereas $\tilde{\Pi}$ starts serving the flexible arrival at one of the servers (possibly the faster one). Suppose that the next event is a dedicated arrival to the queue where the initial flexible arrival is still in service. Then the dedicated arrival will wait in queue under $\tilde{\Pi}$ and it will start service under Π. Now if Π also starts serving the initial flexible arrival at the other server then it is not possible to couple the service processes such that departures will be earlier under $\tilde{\Pi}$ than under Π.

As the above counterexample shows, nonidling DCF will not be optimal in the sample path sense, and we weaken our objective to consider the mean sojourn time (total time in system). When the arrival processes are Poisson and service times are exponentially distributed with the same rate μ, we show that not idling minimizes the mean sojourn time. We assume that arrivals are Poisson and each arrival is flexible with probability p, independent of the state; otherwise, it

is dedicated and is equally likely to go to queue A or queue B, independent of the state, i.e. the arrival processes are independent Poisson processes with rate λp to the flexible queue and with rate $\lambda(1-p)/2$ to each dedicated queue. We assume that the total arrival rate, λ, is such that $\lambda < 2\mu$.

We have the following lemma which directly follows from Theorem 3.

Lemma 1. *For homogeneous exponential servers and symmetric Poisson arrivals, DCF stochastically maximizes* $\{D_t\}_{t=0}^{\infty}$.

It remains to show that the optimal policy is nonidling, i.e. it immediately serves a flexible customer when there are no dedicated customers. Suppose that at time $t = 0$ server A's dedicated queue is empty and that there is at least one job waiting in the flexible queue. Consider two DCF policies Π and $\tilde{\Pi}$, where policy Π chooses to idle server A at time $t = 0$, whereas $\tilde{\Pi}$ starts serving a flexible customer, customer 1, on server A. Without loss of generality, suppose that customer 1 is the first flexible customer Π serves. Note that customer 1 stays in service until departure once service is started under both Π and $\tilde{\Pi}$. Let $\{D_t\}_{t=0}^{\infty}$ and $\{\tilde{D}_t\}_{t=0}^{\infty}$ respectively denote the departure processes under Π and $\tilde{\Pi}$. Now consider a modified nonidling DCF policy $\tilde{\Pi}^p$ where flexible customers have preemptive lower priority than the dedicated customers in the queue of the server where the flexible customer starts service. That is, when a dedicated customer arrives to a server where a flexible customer is taking service, it preempts the flexible customer, and the flexible customer remains in that server's dedicated queue. Let $\{\tilde{D}_t^p\}_{t=0}^{\infty}$, $\{\tilde{D}_{f,t}^p\}_{t=0}^{\infty}$, and $\{\tilde{D}_{d,t}^p\}_{t=0}^{\infty}$ respectively denote the overall departure process, the flexible customers' departure process, and the dedicated customers' departure process under $\tilde{\Pi}^p$. Similarly, we define policy Π^p to agree with Π except flexible customers have lower preemptive priority than the dedicated customers in the queue of the server where they start service, and with departure processes $\{D_t^p\}_{t=0}^{\infty}$, $\{D_{f,t}^p\}_{t=0}^{\infty}$, and $\{D_{d,t}^p\}_{t=0}^{\infty}$, respectively. We have the following lemma, where (i) follows because service times are exponentially distributed, and (ii) follows because dedicated customers are unaffected by the policy for flexible customers in the modified systems.

Lemma 2. (i) $\{D_t\}_{t=0}^{\infty} =_{\text{st}} \{D_t^p\}_{t=0}^{\infty}$ *and* $\{\tilde{D}_t\}_{t=0}^{\infty} =_{\text{st}} \{\tilde{D}_t^p\}_{t=0}^{\infty}$.

(ii) $\{D_{d,t}^p\}_{t=0}^{\infty} =_{\text{st}} \{\tilde{D}_{d,t}^p\}_{t=0}^{\infty}$.

Hence we only need to consider the departure process of flexible customers under Π^p and $\tilde{\Pi}^p$. As far as the flexible customers are concerned, the servers act as independent alternating renewal processes with 'up' times, i.e. times they are available to process flexible customers that are exponentially distributed with rate $\lambda(1-p)/2$, and 'down' times, corresponding to busy periods due to dedicated customers, where they are unavailable. Let S_f be the effective service time of a flexible customer, i.e. the time between its start of service and completion time, including all the down times. Note that S_f also has the same distribution as a dedicated customer busy period.

Consider two systems that are identical except that at time 0 in system 1 some server, A say, is up while in system 2 server A is down.

Lemma 3. *We can couple server A's up and down times in the two systems, by using idling in system 1, such that the first time server A is up in system 2, at time* τ *say, server A will also be up in system 1.*

Proof. We condition on the number of dedicated customers at server A at time 0 in system 2 (by definition, there are 0 dedicated customers at server A in system 1). We couple all dedicated

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

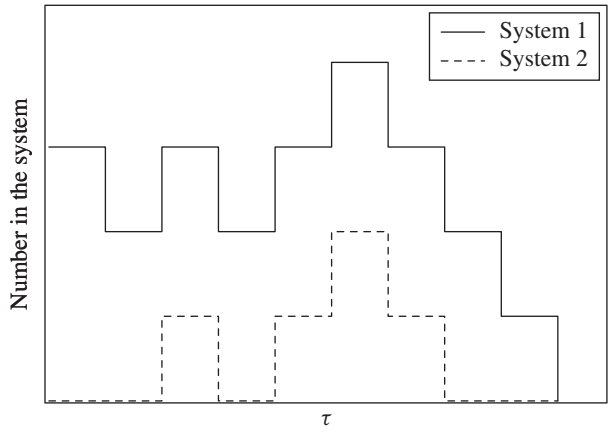

FIGURE 2: Coupling of the two systems.

arrivals and (potential) departures at server A directly. We let system 1 idle server A (not serve flexible customers) during up times before τ. See Figure 2.

We are ready to prove the following result, where '$\prec_{\text{st}}$' denotes the usual stochastic order [33, pp. 3–12].

Lemma 4. $\{D^p_{f,t}\}_{t=0}^{\infty} \prec_{\text{st}} \{\tilde{D}^p_{f,t}\}_{t=0}^{\infty}$.

Proof. We will use the following coupling procedure for arrival and potential service completion times. We generate the effective service time, S_f, for customer 1 under both Π^p and $\tilde{\Pi}^p$, and we couple the dedicated arrivals to queue B and potential service completions on server B under both Π^p and $\tilde{\Pi}^p$. We also generate the dedicated arrivals to queue A and potential service completions on server A under Π^p after $t = 0$, independent of S_f and of the dedicated arrivals and potential service completions for server B (note that server A is idle under Π^p at $t = 0$). We also couple all flexible arrivals directly under Π^p and $\tilde{\Pi}^p$. Let t_1 be the time that Π^p assigns customer 1 to one of the servers. Consider the following possible cases.

Case (i): $S_f \leq t_1$. At time S_f server A is available under $\tilde{\Pi}^p$. Let τ be the first time server A is available after time S_f under Π^p. We couple the process on server A and use idling under $\tilde{\Pi}^p$ so that server A is also available at time τ under $\tilde{\Pi}^p$ by Lemma 3. Since the arrival and service processes of server B are also coupled, both systems are identical at time τ except that customer 1 is still in the system under Π^p. Hence, when customer 1 is being served under Π^p, we can idle that server under $\tilde{\Pi}^p$, and the result follows.

Case (ii): $S_f > t_1$. In this case we consider the following subcases.

Subcase (ii.1): Π^p *serves customer 1 on server A.* In this case, at time S_f, server A is available under $\tilde{\Pi}^p$, while it is busy under Π^p, and server B is in the same state under both policies. Hence, the result follows from Lemma 3 as in the previous case, where we condition on the number of customers (including customer 1) on server A at time S_f, and generate a new remaining busy period (and remaining effective service time for customer 1), with new independent arrivals and services.

Subcase (ii.2): Π^p *serves customer 1 on server B.* In this case we do a cross coupling of servers A and B under two policies as follows. At time t_1 server B is available

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

under $\tilde{\Pi}^p$ and server A may or may not be available under Π^p. We couple the first time server A is available under Π^p so that server B is also available at that time under $\tilde{\Pi}^p$ by Lemma 3. Similarly, at time S_f server A is available under $\tilde{\Pi}^p$ and server B may or may not be available under Π^p. We couple the first time server B is available under Π^p so that server A is also available at that time under $\tilde{\Pi}^p$ by Lemma 3. Again, when customer 1 is being served under Π^p on server B, we can idle server A under $\tilde{\Pi}^p$, and the result follows (where, henceforth, the identities of servers A and B are interchanged).

We have the following theorem that follows easily from Lemmas 1–4, because we can use the same argument each time Π^p or $\tilde{\Pi}^p$ idles to find a better policy that does not idle.

Theorem 4. *For homogeneous exponential servers and symmetric Poisson arrivals, a nonidling DCF policy minimizes the mean sojourn time when preemption is not permitted. Therefore, a nonidling DCF policy also minimizes the mean sojourn time when neither idling nor preemption is permitted.*

Note that Lemmas 2 and 4 do not imply the sample path result, $\{D_t^p\}_{t=0}^{\infty} \prec_{\mathrm{st}} \{\tilde{D}_t^p\}_{t=0}^{\infty}$, because we do not have the joint sample path result, $\{D_{d,t}^p, D_{f,t}^p\}_{t=0}^{\infty} \prec_{\mathrm{st}} \{\tilde{D}_{d,t}^p, \tilde{D}_{f,t}^p\}_{t=0}^{\infty}$, which indeed is not true by Example 3.

2.2.2. *Idling is not permitted.* Finally, we consider the case where neither idling nor preemption is permitted. Our next example shows that, in general, DCF will not be optimal in the sample path sense.

Example 4. Suppose that neither idling nor preemption are permitted and that the required work for all customers is deterministic and equal to 1. Let the service rate for server A (B) be μ_A (μ_B). Suppose that at time $t = 0$ there is one job on server B which will finish at ε, one job waiting in the flexible queue, one job waiting in server A's dedicated queue, server A is idle, and that there are no future arrivals. Then the DCF policy will start A's dedicated job on A at time 0 and the flexible job on B at time ε. Let Π start the flexible customer on A at time 0 and start the dedicated customer on A at time $1/\mu_A$. Hence, if $2/\mu_A < \varepsilon + 1/\mu_B$, the second departure will be earlier under Π, so departures are not sample pathwise earlier under DCF.

Next we will show that, when preemption and idling are not permitted, and the service times are exponential with the same rate, the number of customers in the system is minimized by the DCF policy in the sample path sense, so the departure process is stochastically maximized. We assume that the overall arrival process is one at a time (batch arrivals are not allowed) and independent of the system state and policy but otherwise arbitrary, that a subset of customers are flexible, and that dedicated customers are equally likely to go to queue A or queue B. Note that a weaker result, showing the optimality of the DCF policy in the mean sense when the overall arrival process is Poisson, follows from Theorem 4. Here we will extend the result to sample pathwise optimality using weak submajorization. Weak submajorization is a preordering of $\mathbb{R}^c$, denoted by '$\prec_{\mathrm{w}}$' and defined as follows. For $x, y \in \mathbb{R}^c$,

$$x \prec_{\mathrm{w}} y \quad \text{if} \quad \sum_{1}^{k} x_{[i]} \le \sum_{1}^{k} y_{[i]}, \qquad k = 1, \ldots, c,$$

where $x_{[i]}$ denotes the components of x in decreasing order. Intuitively, x is both smaller and better balanced than y. For two random vectors X and Y, we say that X stochastically

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

weakly submajorizes Y, written $X \prec_{\text{w.st}} Y$, if $\phi(X) \prec_{\text{st}} \phi(Y)$ for all increasing Schur-convex functions ϕ, where a Schur-convex function is defined to be a function that preserves the majorization ordering [28]. The following definitions are equivalent.

(i) $X \prec_{\text{w.st}} Y$.

(ii) $\phi(X) \prec_{\text{st}} \phi(Y)$ for all increasing Schur-convex functions ϕ.

(iii) $\mathbb{E}[\phi(X)] \le \mathbb{E}[\phi(Y)]$ for all increasing Schur-convex functions ϕ.

(iv) There exist random variables $\tilde{X}$ and $\tilde{Y}$ such that

  (a) $X =_{\text{st}} \tilde{X}$ and $Y =_{\text{st}} \tilde{Y}$,

  (b) $\tilde{X} \prec_{\text{w}} \tilde{Y}$ almost surely (a.s.).

For ease of notation, throughout the paper we use '$\prec_{\text{w}}$' for stochastic weak submajorization. Next, let $\{X(t)\}_{t=0}^{\infty}$ and $\{Y(t)\}_{t=0}^{\infty}$ be stochastic processes. We say that $\{X(t)\}_{t=0}^{\infty}$ is stochastically less than $\{Y(t)\}_{t=0}^{\infty}$ in the sense of weak submajorization, denoted by $\{X(t)\}_{t=0}^{\infty} \prec_{\text{w}} \{Y(t)\}_{t=0}^{\infty}$, if we can couple the processes on the same probability space such that, for any sample path realization and any n, $X(t_i) \prec_{\text{w}} Y(t_i)$ jointly for all t_i, $i = 1, \ldots, n$, with probability 1.

We have the following lemma.

Lemma 5. *Let $a_1 \ge \cdots \ge a_c \ge M$ and $b_1 \ge \cdots \ge b_c \ge M$ be integers. If $a \prec_{\text{w}} b$ then*

$$\max\{(a - e_i), M\} \prec_{\text{w}} \max\{(b - e_j), M\} \quad \textit{for all } i \le j,$$

where e_k is the kth unit vector.

For an arbitrary policy at time t, let $N_A(t)$ ($N_B(t)$) denote the number of customers in queue A (B) including the customer in service (regardless of the type of customer in service), let $N(t) = (N_A(t), N_B(t))$, let $N_f(t)$ denote the number of flexible customers waiting in the flexible queue, and let $\bar{N}(t) = N_A(t) + N_f(t) + N_B(t)$ be the total number of customers. We similarly define $N'_A(t)$, $N'_B(t)$, $N'_f(t)$, $N'(t)$, and $\bar{N}'(t)$ for the nonidling DCF. Also, let $(N_A(0), N_f(0), N_B(0)) = (N'_A(0), N'_f(0), N'_B(0))$. Note that a flexible customer effectively becomes a dedicated customer once it starts service because it cannot be put back in the flexible queue.

Theorem 5. *Let the overall arrival process be arbitrary, with an arbitrary subset of customers being flexible, and where the dedicated customers are equally likely to go to queue A or queue B. If the service times are exponentially distributed with the same rate, and neither preemption nor idling are permitted, then*

(i) $\{N'(t)\}_{t=0}^{\infty} \prec_{\text{w}} \{N(t)\}_{t=0}^{\infty}$,

(ii) $\{\bar{N}'(t)\}_{t=0}^{\infty} \le_{\text{st}} \{\bar{N}(t)\}_{t=0}^{\infty}$.

Proof. The proof uses coupling and forward induction. Suppose that $\{\tilde{N}'(t)\}$ $\{\tilde{N}(t)\}$ are stochastic processes having the same stochastic laws as $\{N'(t)\}$ and $\{N(t)\}$. We will couple these processes so that

$$\mathbb{P}(\{\tilde{N}'(t)\}_{t=0}^{\infty} \prec_{\text{w}} \{\tilde{N}(t)\}_{t=0}^{\infty}) = 1, \tag{1}$$

$$\mathbb{P}(\{\tilde{\bar{N}}'(t)\}_{t=0}^{\infty} \le \{\tilde{\bar{N}}(t)\}_{t=0}^{\infty}) = 1. \tag{2}$$

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

To ease the notational burden, we will omit the tildes henceforth on the coupled versions and just use $\{N'(t)\}$ and $\{N(t)\}$.

We couple the arrivals and potential service completion times so that a flexible arrival in $N'(t)$ is also a flexible arrival in $N(t)$, a dedicated arrival to the longest (shortest) dedicated queue in $N'(t)$ is also a dedicated arrival to the longest (shortest) dedicated queue in $N'(t)$. If a potential service completion occurs at the longest (shortest) queue in $N'(t)$ then a potential service completion occurs at the longest (shortest) queue in $N(t)$. A potential service completion will result in an actual service completion if and only if the server is idle (both the flexible and its dedicated queue are empty).

We use induction on t_n, where $\{t_n\}$ denotes the ordered arrival and potential service completion times such that $t_1 < t_2 < t_3 < \cdots$, and $t_0 = 0$. Clearly, (1) and (2) hold for $t = 0$ because $N(0) = N'(0)$. Assume that they also hold for t such that $t_{n-1} \le t < t_n$. Then, because the state does not change for $t_n \le t < t_{n+1}$, it is sufficient to show that (1) and (2) hold for t_n.

If t_n is an arrival time then (2) is trivially true. If the arrival is a dedicated arrival, (1) is also obvious. If it is a flexible arrival, it may go to one of the dedicated queues if the server is idle. It is sufficient to consider the following cases because, for all other cases, the flexible arrival will stay in the flexible queue in the N' system and, therefore, we would have $N'(t_n) = N'(t) \prec_{\mathrm{w}} N(t) \le N(t_n)$.

Case (i): $N'(t) = 0$. Then the flexible arrival is routed to one of the idle servers and $N'_{[1]}(t_n) = 1$. Either $0 < N_{[1]}(t) = N_{[1]}(t_n)$ or $N(t) = 0$, and (1) follows.

Case (ii): $N'_{[1]}(t) > 0$ *and* $N'_{[2]}(t) = 0$. Then the flexible arrival is routed to the idle server, $N'_{[1]}(t_n) = N'_{[1]}(t)$ and $N'_{[2]}(t_n) = 1$. Either $1 \le N_{[2]}(t) = N_{[2]}(t_n)$ (the flexible arrival stays in the flexible queue in the N system) or $0 = N_{[2]}(t)$ (the flexible arrival is routed to the idle server) and $N_{[2]}(t_n) = 1$. Hence, (1) follows.

Next, suppose that t_n is a potential service completion. We first show (1). Note that $N'(t_n) \le N'(t)$ and $N(t_n) \le N(t)$, but we may have equality even if the potential service completion is an actual service completion. Consider the following cases.

Case (i): $N'_f(t) = N_f(t) = 0$. Then the result follows from Lemma 5 with $M = 0$, i.e. the potential service completion is not an actual completion if the server is idle.

Case (ii): $N'_f(t) > 0$ *and* $N_f(t) > 0$. Then the potential service completion is an actual completion in both systems. If the arbitrary policy does not follow the DCF policy then $N(t_n) = N(t)$, since a flexible customer will start being served on the just idled server (and $N_f(t_n) = N_f(t) - 1$), so (1) follows. On the other hand, if the arbitrary policy agrees with DCF at t_n then the result will follow from Lemma 5 with $M = 1$, i.e. the queue length where the service completion occurred will decrease by 1 if there are dedicated customers waiting in the queue, and otherwise a flexible customer will start being served and $N(t_n) = N(t)$ with $N_{[2]}(t_n) = 1$ and, again, (1) follows.

Case (iii): $N'_f(t) > 0$ *and* $N_f(t) = 0$. First suppose that the potential service completion is from the largest queue in both systems. If $N'_{[1]}(t) > 1$ then $N_{[1]}(t) > 1$ and the result follows from Lemma 5 with $M = 0$. If $N'_{[1]}(t) = 1$ then we have $N'_{[2]}(t) = 1$ since $N'_f(t) > 0$. Therefore, from the inductive hypothesis for (2) and the fact that $N_f(t) = 0$, $N_{[1]}(t) \ge 2$ and the result follows. Secondly, suppose that the potential service completion is from the shortest queue in both systems. If $N'_{[2]}(t) > 1$ then again the result follows from Lemma 5 with $M = 0$. If $N'_{[2]}(t) = 1$ then $N'_{[2]}(t_n) = 1$,

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

$N'_f(t_n) = N'_f(t) - 1$, and we have two possibilities for $N_{[2]}(t)$. If $N_{[2]}(t) = 0$ then the potential service completion is not an actual service completion in the N system, i.e. $N(t_n) = N(t)$ and $N'(t_n) = N'(t)$. If $N_{[2]}(t) > 0$ then $N_{[1]}(t) + N_{[2]}(t) = \bar{N}(t) \geq \bar{N}'(t) > N'_{[1]}(t) + N'_{[2]}(t)$, because we have $N'_f(t) > 0$. Hence, the result follows.

Case (iv): $N'_f(t) = 0$ and $N_f(t) > 0$. If the arbitrary policy does not follow the DCF policy then $\bar{N}(t_n) = N(t)$, since a flexible customer will start being served on the just idled server, so (1) follows. If the arbitrary policy follows the DCF policy and if the potential service completion occurs in the ith largest queue, $i = 1, 2$, then

$$\begin{aligned} N'(t_n) &= \max\{(N'(t) - e_i), 0\} \prec_{\mathrm{w}} \max\{(N(t) - e_i), 0\} \\ &\leq \max\{(N(t) - e_i), 1\} \\ &= N(t_n), \end{aligned}$$

where the majorization inequality follows from Lemma 5 with $M = 0$. Hence, (1) follows.

Now to show (2), it is sufficient to look at the case $\bar{N}(t) = \bar{N}'(t)$. In this case, whenever there is an actual service completion in $N(t)$, there has to be an actual service completion in $N'(t)$. To see this, first suppose that $N'_{[1]}(t) = 0$. Then $\bar{N}(t) = \bar{N}'(t) = 0$, so there cannot be an actual service completion in $N_{[1]}(t)$. If $N'_{[2]}(t) = 0$ then $N_{[2]}(t) \leq \bar{N}(t) - N_{[1]}(t) = \bar{N}'(t) - N_{[1]}(t) \leq \bar{N}'(t) - N'_{[1]}(t) = N'_{[2]}(t) = 0$, and again there cannot be an actual service completion in $N_{[2]}(t)$. Hence, $\bar{N}'(t_n) \leq \bar{N}(t_n)$.

So far we have shown that (nonidling) DCF minimizes the number of customers in three different senses depending on the model; the mean number of customers is minimized, the number of customers is sample pathwise stochastically minimized (which follows when departures are stochastically earlier), or the number of customers is sample pathwise stochastically weakly submajorized. If the (nonidling) DCF policy is implemented we have the following monotonicity result. As the subset of the arrivals that are flexible gets larger, which implies that the proportion that are flexible increases, the number of customers in the system gets smaller in the respective sense. This result follows because we can construct an arbitrary policy where a subset of the flexible arrivals are served before dedicated arrivals, and this system becomes stochastically identical to the system where the (nonidling) DCF policy is implemented throughout, but where the same subset of arrivals are dedicated.

3. Optimal routeing of flexible customers

In this section we consider the following setting. Now each server has its own queue but flexible customers do not have their own queue and must be routed to one of the dedicated queues upon arrival. Within a queue services are nonpreemptive FCFS. When the only information available is the queue length, routeing the flexible customers to the shortest queue is optimal under various settings, as noted in the introduction. In this section we consider the model where the exact workload in each queue is known upon arrival, but the required work of the arrival is not yet known, and the arrival is routed to the queue with the shortest work.

Let the overall arrival process be arbitrary, with an arbitrary subset of customers being flexible, and with the dedicated customers equally likely to go to either queue. Let the service times be i.i.d. with distribution function $G(\cdot)$ and be independent of the interarrival times.

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

We now show that, when the workload is known, the workload in the system at each $t > 0$ is stochastically minimized by sending the flexible customers to the queue with the shortest workload upon arrival (the JSW policy). We also show that the departures are pathwise maximized by the JSW policy. These results are extensions of Daley [7] and Koole [21] to systems with both dedicated and flexible arrivals.

For $x, y \in \mathbb{R}^c$, we say that $x \leq y$ if $x_{[j]} \leq y_{[j]}$ for all j, i.e. the jth largest component in x is smaller than the jth largest component in y for all j. We have the following lemma.

Lemma 6. *Let $a_1 \geq \cdots \geq a_c \geq 0$ and $b_1 \geq \cdots \geq b_c \geq 0$ be real numbers.*

(i) *If $a \leq b$ then, for any real t,*

$$(a - t)^+ \leq (b - t)^+,$$

where $(x - t)_j^+ = \max\{x_j - t, 0\}$ for all j.

(ii) *If $a \leq b$ then, for any real u,*

$$(a + ue_i) \leq (b + ue_i) \quad \text{for all } i,$$

where e_k is the kth unit vector.

Consider two parallel server systems where V_n^i, $i = 1, 2$, denote the workload vector at the instant of the nth arrival in the ith system (V_0^i denotes the initial workload at time $t = 0$), and an arbitrary routeing policy is followed in system 2. Then we have

$$V_n^i = (V_{n-1}^i - \tau_n)^+ + s_n e_j$$

if the nth arrival in system i is either a dedicated arrival to queue j or it is a flexible arrival routed to queue j, and where τ_n is the interarrival time between the nth and $(n-1)$th customers, and s_n is the service time of the nth customer. Let $\{D_t^i\}$ denote the departure process in the ith system for $i = 1, 2$. Also, suppose that the initial customers leave earlier in system 1. Then we have the following corollary.

Corollary 1. *If $V_0^1 \leq V_0^2$ then we can couple the arrival and service processes in both systems such that*

(i) $\{V_n^1\}_{n=1}^\infty \leq \{V_n^2\}_{n=1}^\infty$ *a.s.,*

(ii) $\{D_t^1\}_{t=0}^\infty \geq \{D_t^2\}_{t=0}^\infty$ *a.s.*

Proof. We couple the arrivals so that a flexible arrival in system 1 is also a flexible arrival in system 2 with the same service time in both systems and, if it is routed to the queue with longest (shortest) workload in system 2, then system 1 routes it to the queue with longest (shortest) workload as well. A dedicated arrival to the queue with longest (shortest) workload in system 1 is also a dedicated arrival to the queue with longest (shortest) workload in system 2 with the same service time in both systems. Then (i) directly follows from Lemma 6. Now consider the nth arrival after $t = 0$ in both systems. It will join the same queue in both systems and will depart earlier in system 1 than system 2 by (i); hence, (ii) follows.

Consider an arbitrary policy Π that, without loss of generality, routes the first flexible arrival after $t = 0$ (customer 1) to queue A, which has larger workload than queue B. We will construct a policy $\tilde{\Pi}$ that routes customer 1 to queue B and that has a stochastically smaller workload and

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

pathwise earlier departures than policy Π. Let $V_n = (V_{nA}, V_{nB})$ and $\tilde{V}_n = (\tilde{V}_{nA}, \tilde{V}_{nB})$ denote the workload vectors at the instant of the nth arrival under Π and $\tilde{\Pi}$, respectively, and let $V(t)$ ($\tilde{V}(t)$) be the workload vector at time t under Π ($\tilde{\Pi}$). Also, let $\{D_t\}$ ($\{\tilde{D}_t\}$) denote the departure process under Π ($\tilde{\Pi}$). Then we have the following theorem.

Theorem 6. *Let the overall arrival process be arbitrary, with an arbitrary subset of customers being flexible, and with the dedicated customers equally likely to go to either queue. Let the service time of each customer be i.i.d. Then*

(i) $\tilde{V}_n \prec_{\mathrm{w}} V_n$ *for all* $n \geq 1$ *and* $\tilde{V}(t) \prec_{\mathrm{w}} V(t)$ *for all* t,

(ii) $\{D_t\}_{t=0}^{\infty} \leq_{\mathrm{st}} \{\tilde{D}_t\}_{t=0}^{\infty}$.

Proof. The proof follows similarly to the proof of Theorem 3.2 of [21], by using Corollary 1, so we present only a brief sketch of the proof here. First consider (i). Let m be such that the mth arrival (customer m) is the first arrival to queue B after $t = 0$ under policy Π (it is either a dedicated arrival to queue B or a flexible arrival routed to queue B). We cross couple the arrival processes so that, for all $k \leq m$, if the kth arrival is a dedicated arrival to queue A (B) under Π, then it is a dedicated arrival to queue B (A) under $\tilde{\Pi}$. Similarly, if the kth arrival is a flexible arrival routed to queue A (B) under Π, then it is a flexible arrival routed to queue B (A) under $\tilde{\Pi}$. To couple the service times, we first consider $1 \leq n < m$. In this case we couple the service time of customer k under Π with the service time of customer k under $\tilde{\Pi}$ for $1 \leq k \leq n$. Then it can easily be seen that $\tilde{V}_n \prec_{\mathrm{w}} V_n$. Next we consider $n = m$. In this case we couple the service time of customer k under Π with the service time of customer k under $\tilde{\Pi}$ for $1 < k < m$. Let τ be the time of the mth arrival. If $V_{0B} < \tau$ then we couple the service time of customer 1 (customer m) under Π with the service time of customer 1 (customer m) under $\tilde{\Pi}$. However, if $V_{0B} > \tau$ then we cross couple the service time of customer 1 (customer m) under Π with the service time of customer m (customer 1) under $\tilde{\Pi}$. With this coupling procedure, it can be seen that $\tilde{V}_m \prec_{\mathrm{w}} V_m$ and $\tilde{V}_m \leq V_m$. Hence, (i) follows by Corollary 1 for $n > m$.

The proof of (ii) is similar and can be omitted.

The workload result, unlike the departure result, is not sample pathwise optimal, because the coupling procedure required for the proof depends on n (specifically, for $n = 1$, the procedure is different than for $n \geq m$). Therefore, the coupling depends on time, and the result is stochastic optimality for each t, but not jointly for all t. See [21] for a concrete counterexample to pathwise optimality for the workload process when all customers are flexible.

Let W_n^d and $\tilde{W}_n^d$ denote the sojourn times of the nth dedicated arrival under Π and $\tilde{\Pi}$, respectively. Let the nth dedicated arrival be the mth overall arrival for some $m \geq n$. Then we have

$$W_n^d =_{\mathrm{st}} \frac{(V_{m-1A} - \tau_m)^+ + (V_{m-1B} - \tau_m)^+}{2} + S_n,$$

where S_n is the service time of the nth dedicated arrival ($\tilde{W}_n^d$ can be defined similarly), which is an increasing Schur-convex function of the workload. A Schur-convex function is defined to be a function that preserves the majorization ordering. Therefore, if a function $f: \mathcal{R}^c \to \mathcal{R}$ is increasing and Schur convex, then $f(x) \leq f(y)$ whenever $x \prec_{\mathrm{w}} y$. Because $f(x_1, x_2) = c_1[(x_1 - a)^+ + (x_2 - a)^+] + c_2$ is an increasing Schur-convex function, we have the following corollary.

Corollary 2. *It holds that*

$$\tilde{W}_n^d \leq_{\mathrm{st}} W_n^d \quad \textit{for all } n \geq 1.$$

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

Therefore, the dedicated customers on average are better off under the JSW policy. Note that the same result does not follow for flexible customers because $\min\{(V_{m-1A} - \tau_m)^+, (V_{m-1B} - \tau_m)^+\}$ is not an increasing Schur-convex function of the workload vector. That is, flexible customers as a group are not necessarily better off under JSW. However, if we consider a single tagged flexible customer and if the policy is fixed for all other flexible customers, then the tagged flexible customer will be better off by joining the queue with the shortest work. Therefore, JSW is individually optimal for the flexible customers.

4. Comparison of policies

Let $\{D_t^{\mathrm{JSQ}}\}_{t=0}^{\infty}$ ($\{D_t^{\mathrm{JSW}}\}_{t=0}^{\infty}$) denote the departure process when each server has its own queue and flexible customers are routed to the queue with the shortest number of customers (shortest workload). Let $\{D_t^{\mathrm{DCF}}\}_{t=0}^{\infty}$ denote the departure process when there is a separate queue for flexible customers to wait, DCF is followed (might be idling), and premption is not permitted (nonpreemption is more realistic and is more applicable to service systems).

A consequence of our earlier results is that, in general, the system with a separate flexible queue following DCF is better than JSW routeing, which is better than JSQ. The following theorem follows directly from Theorem 6, since JSQ becomes the arbitrary policy described in the proof.

Theorem 7. *Let the overall arrival process be arbitrary, with an arbitrary subset of customers being flexible, and with the dedicated customers equally likely to go to either queue. Let the service times be i.i.d. with distribution function $G(\cdot)$ and be independent of the interarrival times. Then*

$$\{D_t^{\mathrm{JSW}}\}_{t=0}^{\infty} \geq_{\mathrm{st}} \{D_t^{\mathrm{JSQ}}\}_{t=0}^{\infty}.$$

The following theorem follows from Theorem 3, because we can mimic the JSQ or JSW policy in a system with a separate flexible queue and with a control policy which serves the customers at the servers in the same order as they are served under JSQ or JSW. This policy becomes an arbitrary control policy and, therefore, is worse than the DCF policy.

Theorem 8. *Let the arrival process be arbitrary, and let server j have arbitrary service process $\{S_k^j\}_1^{\infty}$, $j = A, B$. Suppose that the actual workload at each queue is known upon arrival and that the service time of the arriving customer is not known. We have*

(i) $\{D_t^{\mathrm{DCF}}\}_{t=0}^{\infty} \geq_{\mathrm{st}} \{D_t^{\mathrm{JSQ}}\}_{t=0}^{\infty}$,

(ii) $\{D_t^{\mathrm{DCF}}\}_{t=0}^{\infty} \geq_{\mathrm{st}} \{D_t^{\mathrm{JSW}}\}_{t=0}^{\infty}$.

As mentioned earlier, the DCF policy is unfair to flexible customers. Therefore, it can be argued that the best system, in terms of both fairness and efficiency, is JSW, which can easily be implemented in call centers.

5. Extensions

In this section we present some easy extensions for our models studied earlier in the paper.

- *More than two servers.* As mentioned earlier in the text, for ease of exposition, we assumed that there are two servers, server A and server B. All the results can easily be extended to more than two servers.

https://doi.org/10.1239/aap/1377868534 Published online by Cambridge University Press

- *Multiple stations.* Suppose that we have multiple service stations with their own queues, but each station has more than one server. Then the results will still be true; however, the proof of Theorem 4 requires a significant modification because the 'up' times depend on the overall server state of the stations. Also, the proof of Theorem 5 requires a careful coupling for the potential service completions (see [3] for the details).
- *Impatience.* Suppose that the customers are impatient and abandon the system after waiting for some exponentially distributed time at rate α (either in queue or in service). Then we can couple the abandonments as we coupled the (potential) service completions in each proof, and all the results discussed in the paper will still be true.

References

[1] Aalto, S., Ayesta, U. and Righter, R. (2009). On the Gittins index in the M/G/1 queue. *Queueing Systems* **63,** 437–458.
[2] Ahn, H.-S., Duenyas, I. and Zhang, R. Q. (2004). Optimal control of a flexible server. *Adv. Appl. Prob.* **36,** 139–170.
[3] Akgun, O. T., Righter, R. and Wolff, R. (2011). Multiple-server system with flexible arrivals. *Adv. Appl. Prob.* **43,** 985–1004.
[4] Aksin, Z., Armony, M. and Mehrotra, V. (2007). The modern call-center: A multi-disciplinary perspective on operations management research. *Prod. Operat. Manag.* **16,** 665–688.
[5] Bambos, N. and Michailidis, G. (2002). On parallel queueing with random server connectivity and routing constraints. *Prob. Eng. Inf. Sci.* **16,** 185–203.
[6] Bell, S. L. and Williams, R. J. (2001). Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *Ann. Appl. Prob.* **11,** 608–649.
[7] Daley, D. J. (1987). Certain optimality properties of the first-come first-served discipline for $G/G/s$ queues. *Stoch. Process. Appl.* **25,** 301–308.
[8] Down, D. G. and Lewis, M. E. (2010). The N-network model with upgrades. *Prob. Eng. Inf. Sci.* **24,** 171–200.
[9] Ephremides, A., Varaiya, P. and Walrand, J. (1980). A simple dynamic routing problem. *IEEE Trans. Automatic Control* **25,** 690–693.
[10] Foley, R. D. and McDonald, D. R. (2001). Join the shortest queue: stability and exact asymptotics. *Ann. Appl. Prob.* **11,** 569–607.
[11] Foss, S. G. (1980). Approximation of multichannel queueing systems. *Siberian Math. J.* **21,** 851–857.
[12] Gans, N., Koole, G. and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review and research prospects. *Manufact. Service Operat. Manag.* **5,** 79–141.
[13] Garnett, O. and Mandelbaum, A. (2000). An introduction to skills-based routing and its operational complexities. Teaching note, Technion, Haifa, Israel.
[14] Gurumurthi, S. and Benjaafar, S. (2004). Modeling and analysis of flexible queuing systems. *Naval Res. Logistics* **51,** 755–782.
[15] Harchol-Balter, M., Crovella, M. E. and Murta, C. D. (1999). On choosing a task assignment policy for a distributed server system. *J. Paral. Distr. Comput.* **59,** 204–228.
[16] Harrison, J. M. (1998). Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Ann. Appl. Prob.* **8,** 822–848.
[17] Harrison, J. M. and López, M. J. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing Systems* **33,** 339–368.
[18] Hordijk, A. and Koole, G. (1990). On the optimality of the generalized shortest queue policy. *Prob. Eng. Inf. Sci.* **4,** 477–487.
[19] Hytia, E., Aaalto, S., Penttinen, A. and Virtamo, J. (2012). On the value function of the $M/G/1$ FIFO and LIFO queues. Submitted.
[20] Johri, P. K. (1989). Optimality of the shortest line discipline with state-dependent service times. *Europ. J. Operat. Res.* **41,** 157–161.
[21] Koole, G. (1992). On the optimality of FCFS for networks of multi-server queues. Tech. Rep. BS-R923, CWI, Amsterdam.
[22] Koole, G. (1992). Stochastic scheduling and dynamic programming. Doctoral Thesis, Leiden University.
[23] Koole, G., Sparaggis, P. D. and Towsley, D. (1999). Minimising response times and queue lengths in systems of parallel queues. *J. Appl. Prob.* **36,** 1185–1193.
[24] Liu, Z. and Righter, R. (1998). Optimal load balancing on distributed homogeneous unreliable processors. *Operat. Res.* **46,** 563–573.

[25] Liu, Z. and Towsley, D. (1994). Optimality of the round-robin routing policy. *J. Appl. Prob.* **31,** 466–475.
[26] Liu, Z., Nain, P. and Towsley, D. (1995). Sample path methods in the control of queues. *Queueing Systems* **21,** 293–335.
[27] Mandelbaum, A. and Stolyar, A. L. (2004). Scheduling flexible servers with convex delay costs: heavy-traffic optimality of the generalized $c\mu$-rule. *Operat. Res.* **52,** 836–855.
[28] Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. Academic Press, New York.
[29] Menich, R. and Serfozo, R. F. (1991). Optimality of routing and servicing in dependent parallel processing stations. *Queueing Systems* **9,** 403–418.
[30] Movaghar, A. (2005). Optimal control of parallel queues with impatient customers. *Performance Evaluation* **60,** 327–343.
[31] Righter, R. and Shanthikumar, J. G. (1989). Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Prob. Eng. Inf. Sci.* **3,** 323–333.
[32] Saghafian, S., Van Oyen, M. P. and Kolfal, B. (2010). The "W" network and the dynamic control of unreliable flexible servers. *IIE Trans.* **43,** 893–907.
[33] Shaked, M. and Shanthikumar, G. J. (1994). *Stochastic Orders and Their Applications*. Academic Press, Boston, MA.
[34] Sparaggis, P. D. and Towsley, D. (1994). Optimal routing and scheduling of customers with deadlines. *Prob. Eng. Inf. Sci.* **8,** 33–49.
[35] Sparaggis, P. D., Towsley, D. and Cassandras, C. G. (1993). Extremal properties of the shortest/longest nonfull queue policies in finite-capacity systems with state-dependent service rates. *J. Appl. Prob.* **30,** 223–236.
[36] Stoyan, D. (1976). A critical remark on a system approximation in queueing theory. *Math. Operationsforsch. Statist.* **7,** 953–956.
[37] Towsley, D., Sparaggis, P. D. and Cassandras, C. G. (1990). Stochastic ordering properties and optimal routing control for a class of finite capacity queueing systems. In *Proc. 29th IEEE Conf. Decision and Control*, pp. 658–663.
[38] Weber, R. R. (1978). On the optimal assignment of customers to parallel servers. *J. Appl. Prob.* **15,** 406–413.
[39] Winston, W. (1977). Optimality of the shortest line discipline. *J. Appl. Prob.* **14,** 181–189.
[40] Wolff, R. W. (1977). An upper bound for multichannel queues. *J. Appl. Prob.* **14,** 884–888.
[41] Wolff, R. W. (1987). Upper bounds on work in system for multi-channel queues. *J. Appl. Prob.* **24,** 547–551.
[42] Wolff, R. W. (1989). *Stochastic Modeling and Theory of Queues*. Prentice Hall, Englewood Cliffs, NJ.