Adv. Appl. Prob. **43,** 985–1004 (2011)
Printed in Northern Ireland

MULTIPLE-SERVER SYSTEM WITH FLEXIBLE ARRIVALS

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

Abstract

In many service, production, and traffic systems there are multiple types of customers requiring different types of 'servers', i.e. different services, products, or routes. Often, however, a proportion of the customers are flexible, i.e. they are willing to change their type in order to achieve faster service, and even if this proportion is small, it has the potential of achieving large performance gains. We generalize earlier results on the optimality of 'join the shortest queue' (JSQ) for flexible arrivals to the following: arbitrary arrivals where only a subset are flexible, multiple-server stations, and abandonments. Surprisingly, with abandonments, the optimality of JSQ for minimizing the number of customers in the system depends on the relative abandonment and service rates. We extend our model to finite buffers and resequencing. We assume exponential service. Our optimality results are very strong; we minimize the queue length process in the weak majorization sense.

Keywords: JSQ; majorization; abandonment; finite buffer; multiple-server station

2010 Mathematics Subject Classification: Primary 90B22
Secondary 60K25

1. Introduction and literature review

In many service, production, and traffic systems there are multiple types of customers requiring different types of 'servers', i.e. different services, products, or routes. Often, the underlying infrastructure is expensive, and, hence, so are the opportunity costs incurred when servers of one type are idle while others are congested. This cost can be reduced by introducing flexible servers that can serve multiple types of customers, but the cost of providing this flexibility may be very high. On the other hand, in many situations a proportion of the customers may be flexible, i.e. may be willing to change their type in order to achieve faster service, and the infrastructure to take advantage of this customer flexibility is often relatively inexpensive. This latter setting can be thought as an example of the network topology known as 'W' design [9].

This research is in part motivated by a call center which provides service in both English and Spanish. Callers currently have the option of pressing '1' for English and '2' for Spanish, but there are times when many Spanish speakers, for example, are on hold while all the Spanish speaking agents are busy, and yet there are idle English-speaking agents. Because of the training expense and high turnover of agents, the company policy is to train agents to handle calls in only one language. The company would like to know the benefit of adding a 'press 3' option

Received 23 September 2010; revision received 10 August 2011.

* Postal address: Department of Industrial Engineering and Operations Research, 4141 Etcheverry Hall, Berkeley, CA 94720, USA.
** Email address: akguno@ieor.berkeley.edu

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

for bilingual customers willing to have their question answered in either language in exchange for reducing their waiting time. Note that this option has a small incremental infrastructure cost only, because it is taking advantage of flexibility that is already present in the customers.

Another important application of our model and analysis is to highway traffic, through the Mobile Millennium project (see http://www.traffic.berkeley.edu/), in which user-generated content on current highway speed, collected by GPS-enabled cell phones, is contributed to a central system that provides information back to the user for personal use in choosing alternate routes. A similar application is to communications and Internet routeing, in which some but not all users have the ability to query alternate routes and use the shortest. In a make-to-order manufacturing context, some customers may not care, for example, about the color of the product they are ordering. Another application is to national boarder crossings with different queues for different nationalities, and where some customers may have dual citizenship. This will be important when a customer's preferred server turns out to be congested, or in other situations in which individual objectives and social welfare objectives are not aligned.

Note that the flexibility we are studying is customer flexibility, not server flexibility. The latter has received a lot of attention in the operations management community, and in particular for call centers. See, e.g. [3], [10], [14], [15], and [18]. However, such flexibility is still generally expensive, particularly in terms of training costs. Customer flexibility, on the other hand, is often already present, but may not be exploited. Generally, it is inexpensive to take advantage of customer flexibility.

We study the effect of customer flexibility. We consider a queueing system with c parallel multiple-server stations having exponentially distributed service times with the same rate μ. All of these servers follow a nonidling arbitrary discipline (first-come–first-served (FCFS), last-come–first-served, etc.). We assume arrivals to the system form an arbitrary process that is independent of the state of the system. Some (*dedicated*) arrivals are obliged to use a particular station, while others (*flexible*) have the ability to use any of the c stations. Dedicated arrivals are equally likely to require a particular station. Let A be the set of arrival points, and let $F \subseteq A$ denote the time points where a flexible arrival occurs. Note that F is an arbitrary subset of A.

Routeing these flexible customers to the station with the least number of customers (ties are broken arbitrarily) is known in the literature as the 'join the shortest queue' (JSQ) policy. The optimality of JSQ, assuming that all customers are flexible (i.e. $F = A$) and assuming a single server at each station, has been shown in a variety of contexts (see, e.g. [5], [6], [16], [17], [19], [24], [27], [28], [29], [30], and [32]). Whitt [31] gave a counterexample that JSQ is not necessarily optimal (even as an individual optimum rather than a social optimum) when processing times are not exponential. More recently, Gupta *et al.* [12] studied JSQ in processor sharing (PS) server farms for nonexponential service times. When both dedicated and flexible arrivals are present, Menich and Serfozo [22] proved the optimality of JSQ for flexible customers among interchangeable routeing policies when processing times are exponential and arrivals are Poisson. Argon *et al.* [4] considered a model with nonidentical exponential service rates and general delay costs and developed heuristic methods for routeing.

Only a few papers have addressed the marginal impact of customer flexibility. When a subset of customers are flexible and flexible customers follow JSQ, Foley and McDonald [7] studied the large deviations (rare event) behavior of the system, and gave conditions under which the rate at which the total queue length reaches a large level is the same as the rate when all customers are flexible. He and Down [13] showed, building on an earlier analysis of Reiman [25] using diffusion limits, that in heavy traffic the full benefit of having some flexible customers can be achieved with an arbitrarily small proportion of customers being flexible. In a follow-up paper

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

we show that waiting times are convex in the proportion of customers who are flexible, which we shorten to the *proportion flexible* [2]. Here we just consider monotonicity.

In Section 2 we study multiple stations having multiple identical servers. Johri [17] and Sparaggis *et al.* [28] allowed service rates to depend on queue lengths, and showed the optimality of JSQ without dedicated customers and under appropriate conditions on the service rates. Multiple identical servers at a station is a special case. We extend this result to general arrivals of both flexible and dedicated customers using weak majorization and by developing a new approach for coupling potential service completions to prove sample-pathwise optimality. We also show that, when flexible customers follow JSQ, the total number of customers in the system is stochastically decreasing in the proportion flexible, so there is an advantage to having customer flexibility. Note that minimizing the total number in the system is equivalent to minimizing the mean waiting time from Little's law. We also show that the waiting time for dedicated customers is decreasing in the proportion flexible. That is, the monolingual customers, on average, benefit from having bilingual customers.

In the remainder of the paper we consider several practically important extensions. In Section 3 we consider customer abandonments, and show that when customers abandon only from the queue, and the abandonment rate is greater than the service rate, even though JSQ no longer minimizes the number of customers in the system, it still maximizes the service completion process. In Section 4 we consider finite buffers, and in Section 5 we consider several other extensions.

2. Multiple-server stations

Sparragis *et al.* [28] showed that JSQ minimizes queue lengths in the weak majorization sense when the service rate for each server is an identical increasing and concave function of the queue length. For the case of multiple servers at a station (m servers at each of the c stations, all with rate μ), we give a new coupling of the servers across stations to show the optimality of JSQ.

While having multiple servers is a special case of the concave model of Sparragis *et al.* [28], they did not allow dedicated customers, whereas we do. Furthermore, we use the same coupling in our proof here to show new results for impatient customers in the next section.

A detailed discussion about weak submajorization, denoted by '$\prec_{\mathrm{w}}$', is given in Appendix A. The following lemma about the weak submajorization of integer-valued vectors is important and will be referred to throughout. It extends the result of Fulkerson and Ryser [8], who showed (1) below. The proof is similar to that of Marshall and Olkin [21, pp. 135–136], and is thus omitted. Note that the proof is not trivial because adding or subtracting e_i may change the relative ordering of a_i.

Lemma 1. *Let $a_1 \geq \cdots \geq a_c$ and $b_1 \geq \cdots \geq b_c$ be integers. If $a \prec_{\mathrm{w}} b$ then*

$$\begin{aligned} a + e_i &\prec_{\mathrm{w}} b + e_j \quad \textit{for all } i \geq j, \\ a - e_i &\prec_{\mathrm{w}} b - e_j \quad \textit{for all } i \leq j, \end{aligned} \tag{1}$$

where e_k is the kth unit vector.

The following corollary provides conditions under which we extend the previous lemma.

Corollary 1. *Let $a_1 \geq \cdots \geq a_c \geq 0$ and $b_1 \geq \cdots \geq b_c \geq 0$ be integers, and fix $i > j$. If $a \prec_{\mathrm{w}} b$ and*

$$\sum_{k=1}^{s} a_k < \sum_{k=1}^{s} b_k \tag{2}$$

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

is true for all $j \leq s < i$, then

$$a - e_i \prec_{\mathrm{w}} b - e_j.$$

Proof. See Appendix B.

Let $N^1(t) = (N_1^1(t), N_2^1(t), \ldots, N_c^1(t))$ denote the number of customers at each station when the JSQ policy is followed for the flexible customers. By shortest queue we mean the station with the fewest number of customers. Let $I^1(t) = (I_{11}^1(t), I_{12}^1(t), \ldots, I_{1m}^1(t), I_{21}^1(t), \ldots, I_{cm}^1(t))$ denote the vector for the number of customers at each server for each station in $N^1(t)$, $I_{ij}^1(t) \in \{0, 1\}$. Furthermore, let $Q^1(t) = ({Q_1}^1(t), {Q_2}^1(t), \ldots, {Q_c}^1(t))$ be the vector of queue lengths (excluding customers at servers) in $N^1(t)$. Hence, $N_i^1(t) = Q_i^1(t) + \sum_{j=(m-1)i+1}^{m*i} I_{ij}^1$. Define $N^2(t)$, $I^2(t)$, and $Q^2(t)$ similarly, assuming that some arbitrary policy is followed.

For convenience, let us label the stations under each policy at time t in decreasing order, so that $N_{[i]}^1(t) = N_i^1(t)$ and $N_{[i]}^2(t) = N_i^2(t)$. Then $Q_{[i]}^1(t) = Q_i^1(t)$ and $Q_{[i]}^2(t) = Q_i^2(t)$. For convenience, we will use a single index for components of $I(t)$, i.e. we let $I_k^1(t) = I_{ij}^1(t)$ and $I_k^2(t) = I_{ij}^2(t)$, where $k = (m-1) * i + j$.

Theorem 1. *The process $\{N^1(t)\}_{t=0}^{\infty}$ is stochastically smaller than $\{N^2(t)\}_{t=0}^{\infty}$ in the sense of weak submajorization:*

$$\{N^1(t)\}_{t=0}^{\infty} \prec_{\mathrm{w}} \{N^2(t)\}_{t=0}^{\infty}.$$

Proof. The proof uses coupling and forward induction. Suppose that $\{\tilde{N}^1(t)\}$ and $\{\tilde{N}^2(t)\}$ are stochastic processes having the same stochastic laws as $\{N^1(t)\}$ and $\{N^2(t)\}$. Define $\{\tilde{I}^1(t)\}$, $\{\tilde{I}^2(t)\}$, $\{\tilde{Q}^1(t)\}$, and $\{\tilde{Q}^2(t)\}$ similarly. We will couple these processes so that

$$\mathrm{P}(\{\tilde{N}^1(t)\}_{t=0}^{\infty} \prec_{\mathrm{w}} \{\tilde{N}^2(t)\}_{t=0}^{\infty}) = 1. \tag{3}$$

To ease the notational burden, we will omit the tildes henceforth on the coupled versions and just use $\{N^1(t)\}$, $\{N^2(t)\}$, $\{I^1(t)\}$, $\{I^2(t)\}$, $\{Q^1(t)\}$, and $\{Q^2(t)\}$.

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, (3) holds for $t = 0$ because $N^1(0) = N^2(0)$. Assume that it is also true for t such that $t_{n-1} \leq t < t_n$. Then, because the state does not change for $t_n \leq t < t_{n+1}$, it is sufficient to show that (3) holds for t_n. We consider two cases separately.

Case 1: arrival. We couple the arrival times so that if a dedicated arrival occurs at the kth largest queue in $N^1(t)$, the same thing happens to $N^2(t)$. For such an arrival, using Lemma 1 with $k = j$ immediately yields $N^1(t_n) \prec_{\mathrm{w}} N^2(t_n)$. Next consider the case where the arrival is flexible and the arbitrary policy chooses to send the customer to the mth largest queue at time t_n. Since $m \leq c$, again by Lemma 1, we obtain $N^1(t_n) \prec_{\mathrm{w}} N^2(t_n)$.

Case 2: departure. It is intuitive to couple potential service completion times (a potential service completion will result in an actual service completion if and only if the queue is not empty) such that, for $N^1(t)$, if a potential service completion occurs at the kth largest station's lth largest server, where the server size within each station is ordered according to $I(t)$, then the same is true in $N^2(t)$. However, this coupling will not work. Consider the counterexample in Table 1. If a potential service completion occurs in the second largest station's smallest server (I_6^j), the majorization will not be valid anymore. Hence, we must define a new way of coupling the potential service completions.

We label the busy servers in order from the largest station to the smallest station, whereas idle servers will be ordered after the busy ones regardless of the station. Then the coupling will

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

Table 1: Two coupled systems at time t with $c = 3$ and $m = 3$.

System 1									System 2								
I_1^1	I_2^1	I_3^1	I_4^1	I_5^1	I_6^1	I_7^1	I_8^1	I_9^1	I_1^2	I_2^2	I_3^2	I_4^2	I_5^2	I_6^2	I_7^2	I_8^2	I_9^2
1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0	0
Q_1^1			Q_2^1			Q_3^1			Q_1^2			Q_2^2			Q_3^2		
1			0			0			1			0			0		

Table 2: Sample server states and related ordering.

I_1^j	I_2^j	I_3^j	I_4^j	I_5^j	I_6^j	I_7^j	I_8^j	I_9^j	I_{10}^j	I_{11}^j	I_{12}^j
1	1	1	1	1	1	1	1	0	1	0	0
Q_1^j			Q_2^j			Q_3^j			Q_4^j		
2			0			0			0		
Server ordering											
1	2	3	4	5	6	7	8	10	9	11	12

be done on this ordering accordingly, i.e. if a potential service completion occurs at the pth server in this ordering in system 1, then a potential service completion occurs at the pth server in system 2. See Table 2 for an ordering example.

There are four cases. If no actual service completion occurs in either system, or if the potential service completion is an actual service completion in system 1 but not in system 2, then majorization continues to hold trivially.

Now, suppose that the potential service completion is an actual service completion in both systems. Suppose that u is such that the actual service completion takes place in the uth largest station in system 1, and define v for system 2 similarly. First, if $u \leq v$ then, by Lemma 1, $N^1(t_n) \prec_{\mathrm{w}} N^2(t_n)$. Now suppose that $u > v$ (see Table 3), so, intuitively, to get the pth nonidle server, more idle servers are skipped over in system 1 than in system 2. Formally, because of our definitions of p, u, and v, stations up to r have more empty servers in system 1 for $v \leq r < u$, that is,

$$\sum_{k=1}^{r*m} I_k^1(t) < p \leq \sum_{k=1}^{v*m} I_k^2(t) \leq \sum_{k=1}^{r*m} I_k^2(t). \tag{4}$$

Now, define q as the smallest indexed station in system 1 without an empty server ($q = 0$ if all stations have at least one empty server, $q = 1$ in the example of Table 3). Then, by (4) and because $I_k^1(t) = 1$ for $k \leq q * m$, we have $\sum_{k=q*m+1}^{r*m} I_k^1(t) < \sum_{k=q*m+1}^{r*m} I_k^2(t)$. Hence,

$$\sum_{k=1}^{r} N_k^1(t) = \sum_{k=1}^{q} N_k^1(t) + \sum_{k=q*m+1}^{r*m} I_k^1(t) < \sum_{k=1}^{q} N_k^2(t) + \sum_{k=q*m+1}^{r*m} I_k^2(t) = \sum_{k=1}^{r} N_k^2(t)$$

for $v \leq r < u$. Therefore, by Corollary 1, we have $N^1(t_n) = N^1(t) - e_u \prec_{\mathrm{w}} N^2(t) - e_v = N^2(t_n)$. Now, suppose that the potential service completion is an actual service completion in system 2, but not in system 1. See Table 4. Define v as the station where the actual service

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

TABLE 3: (a) Server states at time t. A departure takes place from the $(p = 6)$th largest server in both systems, so $u = 3$ and $v = 2$. (b) States after departure. Majorization is still valid.

(a)

System 1									System 2								
I_1^1	I_2^1	I_3^1	I_4^1	I_5^1	I_6^1	I_7^1	I_8^1	I_9^1	I_1^2	I_2^2	I_3^2	I_4^2	I_5^2	I_6^2	I_7^2	I_8^2	I_9^2
1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0	0
	Q_1^1			Q_2^1			Q_3^1			Q_1^2			Q_2^2			Q_3^2	
	2			0			0			2			0			0	
Server ordering									Server ordering								
1	2	3	4	5	8	6	7	9	1	2	3	4	5	6	7	8	9

(b)

System 1									System 2								
I_1^1	I_2^1	I_3^1	I_4^1	I_5^1	I_6^1	I_7^1	I_8^1	I_9^1	I_1^2	I_2^2	I_3^2	I_4^2	I_5^2	I_6^2	I_7^2	I_8^2	I_9^2
1	1	1	1	1	0	0	1	0	1	1	1	1	1	0	1	0	0
	Q_1^1			Q_2^1			Q_3^1			Q_1^2			Q_2^2			Q_3^2	
	2			0			0			2			0			0	

TABLE 4: (a) Server states at time t. A departure takes place from the $(p = 8)$th largest server in both systems, so $v = 3$ is the last nonempty station in system 2. (b) States after departure. Majorization is still valid.

(a)

System 1									System 2								
I_1^1	I_2^1	I_3^1	I_4^1	I_5^1	I_6^1	I_7^1	I_8^1	I_9^1	I_1^2	I_2^2	I_3^2	I_4^2	I_5^2	I_6^2	I_7^2	I_8^2	I_9^2
1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0
	Q_1^1			Q_2^1			Q_3^1			Q_1^2			Q_2^2			Q_3^2	
	2			0			0			2			0			0	
Server ordering									Server ordering								
1	2	3	4	5	8	6	7	9	1	2	3	4	5	6	7	8	9

(b)

System 1									System 2								
I_1^1	I_2^1	I_3^1	I_4^1	I_5^1	I_6^1	I_7^1	I_8^1	I_9^1	I_1^2	I_2^2	I_3^2	I_4^2	I_5^2	I_6^2	I_7^2	I_8^2	I_9^2
1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0	0
	Q_1^1			Q_2^1			Q_3^1			Q_1^2			Q_2^2			Q_3^2	
	2			0			0			2			0			0	

completion in system 2 takes place. We will use the following definition to account for the change in order after departure. Let $v' \geq v$ be such that

$$N_v^2(t) = N_{v+1}^2(t) = \cdots = N_{v'}^2(t) \quad \text{and either} \quad N_{v'}^2(t) > N_{v'+1}^2(t) \quad \text{or} \quad v' = c.$$

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

Then

$$\sum_{k=1}^{i} N_k^2(t_n) = \sum_{k=1}^{i} N_k^2(t) - \mathbf{1}\{i \geq v'\}.$$

Note that the total number of busy servers in system 1 is less than the number of busy servers up to station v in system 2. This follows from

$$\sum_{k=1}^{m*c} I_k^1(t) < p \leq \sum_{i=1}^{m*v} I_k^2(t).$$

Again, define q as the smallest station in system 1 without an empty server. Then,

$$\begin{aligned}\sum_{k=1}^{c} N_k^1(t) &= \sum_{k=1}^{q} N_k^1(t) + \sum_{k=m*q+1}^{m*c} I_k^1(t)\\ &< \sum_{k=1}^{q} N_k^2(t) + \sum_{k=m*q+1}^{m*v} I_k^2(t)\\ &= \sum_{k=1}^{v} N_k^2(t)\\ &\leq \sum_{k=1}^{v'} N_k^2(t).\end{aligned}$$

Hence, $\sum_{k=1}^{i} N_k^1(t_n) \leq \sum_{k=1}^{i} N_k^2(t_n)$, and this concludes the proof.

Remark 1. The weak majorization result also holds for the customers waiting in the queue, that is, $\{Q^1(t)\}_{t=0}^{\infty} \prec_{\mathrm{w}} \{Q^2(t)\}_{t=0}^{\infty}$. We also have the following immediate consequences of our result. The total number of customers in the system, and, hence, the mean waiting time, is stochastically minimized by JSQ for flexible arrivals. That is, letting $\bar{N}^i(t) = \sum_{i=1}^{c} N_i{}^i(t)$, $i = 1, 2$, we have $\{\bar{N}^1(t)\}_{t=0}^{\infty} \prec_{\mathrm{st}} \{\bar{N}^2(t)\}_{t=0}^{\infty}$ and $\mathrm{E}[\bar{N}^1(t)] \leq \mathrm{E}[\bar{N}^2(t)]$, $t \geq 0$. Also, the departure process is stochastically maximized by JSQ for flexible arrivals. That is, letting $D^i(t)$ denote the number of departures by time t in $N^i(t)$, $i = 1, 2$, we have $\{D^1(t)\}_{t=0}^{\infty} \succ_{\mathrm{st}} \{D^2(t)\}_{t=0}^{\infty}$.

Another immediate result is that any increasing Schur convex function of $N(t)$ is stochastically minimized by JSQ for flexible arrivals. An example is when the holding cost is a separable increasing cost function, $\phi(N) = \sum \phi_i(N_i)$, where ϕ_i is increasing convex. Also, note that the cost function could vary over time, or change randomly.

As the subset of the arrivals that are flexible gets larger, which implies that the proportion flexible increases, the number of customers in the system gets smaller in the weak submajorization sense. This result follows from Theorem 1, since we can construct an arbitrary routeing policy where a subset of the flexible arrivals are routed to the shortest queue and the rest are routed arbitrarily. Similarly, the steady-state mean waiting time in the system for all customers and the steady-state mean waiting time for dedicated customers decrease in the usual stochastic order as the subset of the arrivals that are flexible gets larger. That is, not only are all customers better off on average when there are more flexible customers, but the dedicated customers as a group are better off when there are more flexible customers. To see this, note that, by our construction, dedicated arrivals see fewer customers on average as the subset of flexible customers increases, so their mean waiting times will also be smaller by the symmetry of the queues.

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

A special case of our nested arrival process is the case where some proportion of arrivals are flexible. More rigorously, let each arrival be flexible with probability f, and dedicated with probability $1 - f$, where $0 \leq f \leq 1$, independently of the other arrivals. We define N^i as the vector of queue lengths when JSQ is followed for flexible customers and the proportion of flexible customers is f_i, where $f_1 \geq f_2$. Then $\{N^1\}_{t=0}^{\infty} \prec_{\mathrm{w}} \{N^2\}_{t=0}^{\infty}$ and both the steady-state mean waiting time for all customers and steady-state waiting time for dedicated customers are smaller in N^1 for this setting.

3. Impatient customers

We suppose that the customers are impatient and abandon the system after waiting for some exponentially distributed time at rate α. We consider only single-server stations for simplicity. Define $N^1(t) = ({N_1}^1(t), {N_2}^1(t), \ldots, {N_c}^1(t))$ as the vector of queue lengths at time $t \geq 0$ when JSQ is followed for flexible customers, and define $N^2(t)$ as the vector of queue lengths when an arbitrary policy is followed for flexible customers. Also, let $A^1(t)$ denote the total number of abandonments and $D^1(t)$ the total number of service completions up to time t when JSQ is followed for flexible customers. Define $A^2(t)$ and $D^2(t)$ similarly for the arbitrary routeing policy.

Sparaggis *et al.* [28] showed the following results.

- For nondecreasing concave service rates, $\{N^1(t)\}_{t=0}^{\infty} \prec_{\mathrm{w}} \{N^2(t)\}_{t=0}^{\infty}$.
- For nondecreasing convex service rates, $\{N^1(t)\}_{t=0}^{\infty} \prec^{\mathrm{w}} \{N^2(t)\}_{t=0}^{\infty}$.

Here '$\prec^{\mathrm{w}}$' denotes weak supermajorization. Again, detailed discussion about supermajorization can be found in Appendix A. Note that customers abandoning both in queue and in service, and customers abandoning only in queue with rate $\alpha \leq \mu$ are special cases of nondecreasing concave service rates, whereas customers abandoning only in queue with rate $\alpha > \mu$ is a special case of nondecreasing convex service rates. However, the above result does not tell us much about the overall system performance, because stochastically minimizing the number in the system does not mean maximizing service completions, as the departures are due to both service completions and abandonments.

Movaghar [24] also studied the JSQ policy with impatient customers. He assumed Poisson arrivals, and gave conditions on the distribution of the patience time such that JSQ minimizes the number of long-run abandonments.

The following result was shown to be true without dedicated arrivals in [27]. We prove it using a new coupling procedure and use it for the proof of Theorem 2.

Corollary 2. *Let each customer in queue (but not in service) abandon the system after waiting an exponentially distributed time at rate α, where $\alpha \leq \mu$. Then,*

$$\{N^1(t)\}_{t=0}^{\infty} \prec_{\mathrm{w}} \{N^2(t)\}_{t=0}^{\infty}, \tag{5}$$

$$\{A^1(t)\}_{t=0}^{\infty} \prec_{\mathrm{st}} \{A^2(t)\}_{t=0}^{\infty}, \tag{6}$$

$$\{D^1(t)\}_{t=0}^{\infty} \succ_{\mathrm{st}} \{D^2(t)\}_{t=0}^{\infty}. \tag{7}$$

Proof. Again, we use induction on event times t_n, so suppose that (5)–(7) hold for $t < t_n$. Now we separate the service completions into two types. We say that a customer finishing service and departing the system will independently be tagged as type 1 with probability α/μ and type 2 with probability $1 - \alpha/\mu$. We couple the arrivals and potential service completions of type 2 as in the proof of Theorem 1. Since we have single-server stations, potential service

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

TABLE 5: Labeling of customers, where a 1 indicates the presence of a customer.

System 1			System 2		
I_1^1 1	I_2^1 1	I_3^1 1	I_1^2 1	I_2^2 1	I_3^2
Q_1^1 1 1	Q_2^1 1 1	Q_3^1	Q_1^2 1 1 1	Q_2^2 1 1 1	Q_3^2
Labeling of customers			Labeling of customers		
5	6	7	5	6	
1 2	3 4		1 2 3	4 7 8	

completion coupling becomes trivial (if a potential service completion occurs at the kth largest queue in $N^1(t)$ then a potential service completion occurs at the kth largest queue in $N^2(t)$). The weak majorization ordering for $N^i(t)$ as well as the stochastic ordering for $A^i(t)$ and $D^i(t)$ will be preserved as before.

Now let us couple potential abandonments and potential service completions of type 1 as follows. In broad terms, we first couple each abandonment in system 1 with an abandonment in system 2, which we can do from Remark 1, because it tells us that the total number in queue will be smaller in system 1. We then couple service completions in both systems as much as possible, and then couple remaining customers. More rigorously, let us label the customers in system 1 as follows. First label the customers in queue, from the largest to the smallest queues. That is, customers $1, 2, \dots, Q_1^1(t)$ are the customers in the first (largest) queue, then customers $Q_1^1(t) + 1, \dots, Q_1^2(t)$ are in the second queue, etc. We then label the customers in service starting from station 1. See the first part of Table 5 for a three-station example, where $I_j^i = \mathbf{1}\{N_j^i(t) > 0\}$ represents the customers in service. For system 2, label the first $Q^1 := \sum Q_i^1(t)$ customers in the queues starting from those in queue 1 as we did for system 1. Then label the customers in service in system 2, and, finally, label any remaining customers in the queues in decreasing order of queue length. Again, see Table 5. Suppose that customer p under the labeling above departs from both systems, $1 \le p \le N^2 := \sum N_i^2(t)$, where, for $p > N^1 := \sum N_i^1(t)$, there is no actual departure in system 1. We have the following cases.

(i) There is an actual abandonment or service completion of type 1 in system 2, but not in system 1 (i.e. $N^1 < p \le N^2$; for example, $p = 8$ in Table 5). Let v be such that this departure takes place in the vth largest queue in system 2, so $\sum_{k=1}^{v} N_k^2(t) \ge p$. Then to account for an order change, let $v' \ge v$ be such that

$$\sum_{k=1}^{i} N_k^2(t_n) = \sum_{k=1}^{i} N_k^2(t) - \mathbf{1}\{i \ge v'\}.$$

Note that

$$N^1 := \sum_{k=1}^{c} N_k^1(t) < p \le \sum_{k=1}^{v} N_k^2(t) \le \sum_{k=1}^{v'} N_k^2(t),$$

so $\sum_{k=1}^{i} N_k^1(t_n) \le \sum_{k=1}^{i} N_k^2(t_n)$ for all i and $N^1(t_n) \prec_{\mathrm{w}} N^2(t_n)$.

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

(ii) There is an actual abandonment in both systems (i.e. $p \leq Q^1$; for example, $p = 2$ in Table 5). Let $u(v)$ be such that the abandonment takes place in the uth largest queue in system 1 and the vth largest queue in system 2. If $u \leq v$ then, by Lemma 1, $N^1(t_n) \prec_{\mathrm{w}} N^2(t_n)$. If $u > v$ then, by the definition of p,

$$\sum_{k=1}^{r} N_k^1(t) \leq r + \sum_{k=1}^{u-1} Q_k^1(t) < r + p \leq r + \sum_{k=1}^{v} Q_k^2(t) \leq \sum_{k=1}^{r} N_k^2(t)$$

for all $v \leq r < u$. Therefore, by Corollary 1, $N^1(t_n) \prec_{\mathrm{w}} N^2(t_n)$.

(iii) There is an actual service completion of type 1 in both systems (i.e. $Q^1 < p \leq Q^1 + \max\{M_1, M_2\}$, where $M_j = \max\{i : N_i^j(t) > 0\}$, the total number of customers in service in system j; for example, $p = 5$ in Table 5). Let $u(v)$ be such that the service completion takes place in the uth largest queue in system 1 and the vth largest queue in system 2. Then, $u = v$ because $Q^1 + u = p = Q^1 + v$. Therefore, by Corollary 1, $N^1(t_n) \prec_w N^2(t_n)$.

(iv) There is an actual abandonment in system 2 and an actual service completion of type 1 in system 1 (i.e. $M_1 > M_2$ and $Q^1 + M_2 < p \leq Q^1 + M_1$; for example, $p = 7$ in Table 5). Let the service completion in system 1 take place in the uth largest queue, and let the abandonment in system 2 take place in the vth largest queue in system 2, so $v \leq M_2 < u$. Therefore, by the definition of p,

$$\sum_{k=1}^{r} N_k^2(t) \geq p - (M_2 - r)^+ > p - (u - r) \geq \sum_{k=1}^{r} N_k^1(t)$$

for all $v \leq r < u$, and, again by Corollary 1, $N^1(t_n) \prec_w N^2(t_n)$.

Note that $A^1(t)$ changes only in case (ii). But in that case $A^2(t_n) = A^2(t)+1 \geq A^1(t)+1 = A^1(t_n)$. Hence, (6) follows by induction. Similar reasoning as in the previous proof shows (7).

We now consider the case where $\mu < \alpha$. As mentioned at the beginning of the section, the total number in the system will not be minimized under JSQ. Intuitively, since abandonments occur only from the queue, and with a higher rate than the service rate, to minimize the number in the system, it is better to have long queues, and, therefore, more abandonments.

We will make use of the following lemma and corollary, which are similar to our earlier submajorization results, so we omit the proofs.

Lemma 2. *Let $a_1 \geq \cdots \geq a_c$ and $b_1 \geq \cdots \geq b_c$ be integers. If $a \prec^{\mathrm{w}} b$ then*

$$a + e_i \prec^{\mathrm{w}} b + e_j \quad \textit{for all } i \geq j,$$
$$a - e_i \prec^{\mathrm{w}} b - e_j \quad \textit{for all } i \leq j,$$

where e_k is the kth unit vector.

Corollary 3. *Let $a_1 \geq \cdots \geq a_c \geq 0$ and $b_1 \geq \cdots \geq b_c \geq 0$ be integers, and fix $i > j$. If $a \prec^{\mathrm{w}} b$ and*

$$\sum_{k=s}^{c} a_k > \sum_{k=s}^{c} b_k$$

is true for all $j < s \leq i$, then

$$a - e_i \prec^{\mathrm{w}} b - e_j.$$

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

We also have the following corollary.

Corollary 4. *Let $a_1 \geq \cdots \geq a_c \geq 0$ and $b_1 \geq \cdots \geq b_c \geq 0$ be integers, and let $\sum_{k=1}^{c} a_k > \sum_{k=1}^{c} b_k$. If $a \prec^{\mathrm{w}} b$ then*

$$a - e_1 \prec^{\mathrm{w}} b,$$

where e_k is the kth unit vector.

Proof. See Appendix C.

Theorem 2. *Let each customer in queue abandon the system after waiting an exponentially distributed time at rate α, where $\alpha > \mu$. Then,*

$$\{D^1(t)\}_{t=0}^{\infty} \succ_{\mathrm{st}} \{D^2(t)\}_{t=0}^{\infty}.$$

Proof. We show the stronger result, $\{N^1(t)\}_{t=0}^{\infty} \prec^{\mathrm{w}} \{N^2(t)\}_{t=0}^{\infty}$, as well as $\{D^1(t)\}_{t=0}^{\infty} \succ_{\mathrm{st}} \{D^2(t)\}_{t=0}^{\infty}$ by induction. We will couple the systems in a way that the weak supermajorization is preserved at each $t > 0$. Note that

$$\{N^1(t)\}_{t=0}^{\infty} \prec^{\mathrm{w}} \{N^2(t)\}_{t=0}^{\infty} \quad \Longrightarrow \quad \{\bar{N}^1(t)\}_{t=0}^{\infty} \succ_{\mathrm{st}} \{\bar{N}^2(t)\}_{t=0}^{\infty},$$

where $\bar{N}^j(t)$ denotes the number of customers at time t in system j.

First, we separate the abandonments into two types. We say that a customer abandoning the system will independently be tagged as type 1 with probability μ/α and type 2 with probability $1 - \mu/\alpha$. Next we couple the arrivals as in the proof of Theorem 1. In these cases the weak supermajorization will be preserved by Lemma 2. Next we couple potential service completions and potential abandonments of type 1. Let $d = \sum_{k=1}^{c} N_k^1(t) - \sum_{k=1}^{c} N_k^2(t)$ be the difference between the total number of customers in system 1 and system 2. By induction we have $N^1(t) \prec^{\mathrm{w}} N^2(t)$; hence, $d \geq 0$. We will set these additional d customers apart and couple the other ones together. In other words, departures for these additional d customers will be coupled with dummy (nonactual) departures in system 2. To rigorously define these customers, let V be an integer-valued c-dimensional vector and let e^V be the unit vector such that $e_i^V = 1$ if $V_i = V_{[1]}$ and 0 otherwise. Then we will define a sequence of vectors such that $V^{n+1}(t) = V^n(t) - e^{V^n(t)}$ and $V^0(t) = N^1(t)$. By Corollary 4, $N^1(t) \prec^{\mathrm{w}} V^d(t) \prec^{\mathrm{w}} N^2(t)$ and $\sum_{k=1}^{c} V_k^d(t) = \sum_{k=1}^{c} N_k^2(t)$. Hence, $V^d(t) \prec_{\mathrm{w}} N^2(t)$. If, for the d customers in $N^1(t) - V^d(t)$ there is a service completion or type-1 abandonment in system 1, there is a dummy transition in system 2, so $N^1(t_n) \prec^{\mathrm{w}} N^2(t_n)$. Let this customer be in the kth largest queue in system 1. Then it is immediate that $N^1(t_n) = N^1(t) - e_k \prec^{\mathrm{w}} V^d(t) \prec^{\mathrm{w}} N^2(t) = N^2(t_n)$. Therefore, $N^1(t_n) \prec^{\mathrm{w}} N^2(t_n)$. Also, $D^2(t)$ does not change in this case, still satisfying $D^1(t_n) \geq D^2(t_n)$. For the remaining customers in $V^d(t)$, we couple the service completions and type-1 abandonments with those in $N^2(t)$, as in the proof of Corollary 2, which we can do because $V^d(t) \prec_w N^2(t)$, so we obtain $V^d(t_n) \prec_w N^2(t_n)$ and $D^1(t_n) \geq D^2(t_n)$. Since $\sum_{k=1}^{c} V_k^d(t_n) = \sum_{k=1}^{c} N_k^2(t_n)$ and $N^1(t_n) \geq V^d(t_n)$, we also have $N^1(t_n) \prec^{\mathrm{w}} V^d(t_n) \prec^{\mathrm{w}} N^2(t_n)$.

Finally, we will look at the potential abandonments of type 2. In this case, the number of service completions will not change, satisfying $D^1(t_n) \geq D^2(t_n)$. The coupling procedure is similar to Theorem 1. However, this time we label the customers in queue as we see them from the smallest station to the largest station in both systems. We couple the potential type-2 abandonments such that customer p abandons in both systems (which may correspond to a dummy

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

TABLE 6: Labeling of customers for abandonments of type 2.

System 1			System 2		
I_1^1	I_2^1	I_3^1	I_1^2	I_2^2	I_3^2
1	1	1	1	1	1
Q_1^1	Q_2^1	Q_3^1	Q_1^2	Q_2^2	Q_3^2
1 1	1		1 1		
Labeling of customers			Labeling of customers		
3 2	1		2 1		

abandonment if p exceeds the number of customers). See Table 6. Again, let $I_j^i = \mathbf{1}\{N_j^i(t) > 0\}$ represent the customers in service. It is immediate that $\sum_{k=i}^c I_k^1(t) \geq \sum_{k=i}^c I_k^2(t)$ for all i, as $N^1(t) \prec^{\mathrm{w}} N^2(t)$. The following are the possible cases.

(i) The potential abandonment is not an actual one in system 1, but there is an actual abandonment of type 2 in system 2. Majorization is preserved trivially.

(ii) There is an actual abandonment in both systems. Let $u(v)$ be such that the abandonment takes place in the uth largest queue in system 1 and the vth largest queue in system 2. If $u \leq v$ then, by Lemma 2, $N^1(t_n) \prec^{\mathrm{w}} N^2(t_n)$. If $u > v$ (e.g. $p = 1$ in Table 6, so $u = 2$ and $v = 1$) then, by the definition of p, $\sum_{k=u}^c Q_k^1(t) \geq p$ and $\sum_{k=v+1}^c Q_k^2(t) < p$ because $Q_v^2(t) > 0$. Therefore,

$$\sum_{k=r}^c Q_k^2(t) \leq \sum_{k=v+1}^c Q_k^2(t) < p \leq \sum_{k=u}^c Q_k^1(t) \leq \sum_{k=r}^c Q_k^1(t)$$

for all $v < r \leq u$. Combining this with $\sum_{k=i}^c I_k^1(t) \geq \sum_{k=i}^c I_k^2(t)$ for all i shows that

$$\sum_{k=r}^c N_k^2(t) = \sum_{k=r}^c Q_k^2(t) + \sum_{k=r}^c I_k^2(t) < \sum_{k=r}^c Q_k^1(t) + \sum_{k=r}^c I_k^1(t) = \sum_{k=r}^c N_k^1(t)$$

for all $v < r \leq u$. So, by Corollary 3, $N^1(t_n) \prec^{\mathrm{w}} N^2(t_n)$.

(iii) The potential type-2 abandonment is not an actual one in system 2, but there is an actual abandonment of type 2 in system 1 (e.g. $p = 3$, $u = 1$ in Table 6). To account for an order change, let $u' \geq u$ be such that

$$\sum_{k=i}^c N_k^1(t_n) = \sum_{k=i}^c N_k^1(t) - \mathbf{1}\{i \leq u'\}.$$

By the definition of p,

$$\sum_{k=u}^c Q_k^1(t) \geq p \quad \text{and} \quad \sum_{k=1}^c Q_k^2(t) < p.$$

For $r \geq u'$, $\sum_{k=r}^c N^1(t_n) = \sum_{k=r}^c N^1(t) \geq \sum_{k=r}^c N^2(t) = \sum_{k=r}^c N^2(t_n)$. Now suppose, by way of contradiction, for $u' \geq r > u$, that $\sum_{k=r}^c N^1(t_n) < \sum_{k=r}^c N^2(t_n)$,

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

i.e.

$$\sum_{k=r}^{c} N_k^1(t) = \sum_{k=r}^{c} N_k^2(t). \tag{8}$$

By induction we have $\sum_{k=r+1}^{c} N_k^1(t) \ge \sum_{k=r+1}^{c} N_k^2(t)$. Combining this with (8) yields $Q_r^1(t) \le Q_r^2(t)$. On the other hand, $\sum_{k=r}^{c} I_k^1(t) \ge \sum_{k=r}^{c} I_k^2(t)$. This with (8) gives $\sum_{k=r}^{c} Q_k^1(t) \le \sum_{k=r}^{c} Q_k^2(t)$. Therefore,

$$\begin{aligned}\sum_{k=1}^{c} Q_k^2(t) &\ge \sum_{k=u}^{c} Q_k^2(t)\\ &= \sum_{k=u}^{r-1} Q_k^2(t) + \sum_{k=r}^{c} Q_k^2(t)\\ &\ge \sum_{k=u}^{r-1} Q_r^2(t) + \sum_{k=r}^{c} Q_k^2(t)\\ &\ge \sum_{k=u}^{r-1} Q_r^1(t) + \sum_{k=r}^{c} Q_k^1(t)\\ &= \sum_{k=u}^{c} Q_k^1(t)\\ &\ge p,\end{aligned}$$

which is a contradiction. Hence, $\sum_{k=r}^{c} N_k^1(t) > \sum_{k=r}^{c} N_k^2(t)$ for $u' \ge r > u$. Finally, for $r \le u$,

$$\begin{aligned}\sum_{k=r}^{c} N_k^1(t) &= \sum_{k=r}^{c} Q_k^1(t) + \sum_{k=r}^{c} I_k^1(t)\\ &\ge \sum_{k=u}^{c} Q_k^1(t) + \sum_{k=r}^{c} I_k^1(t)\\ &\ge p + \sum_{k=r}^{c} I_k^1(t)\\ &> \sum_{k=1}^{c} Q_k^2(t) + \sum_{k=r}^{c} I_k^1(t)\\ &\ge \sum_{k=1}^{c} Q_k^2(t) + \sum_{k=r}^{c} I_k^2(t)\\ &\ge \sum_{k=r}^{c} Q_k^2(t) + \sum_{k=r}^{c} I_k^2(t)\\ &= \sum_{k=r}^{c} N_k^2(t).\end{aligned}$$

Thus, $\sum_{k=r}^{c} N_k^1(t) > \sum_{k=r}^{c} N_k^2(t)$ for all $r \le u'$, so, $N^1(t_n) \prec^{\mathrm{w}} N^2(t_n)$.

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

4. Finite buffers

In the models studied earlier, each queue had infinite space for waiting. A more realistic extension to this model is the case where queues have finite buffers. This problem is not an immediate extension because the weak majorization will not be preserved upon arrival as in the infinite buffer models. For instance, consider a system with two servers where each queue has a capacity of four customers. Let $N^1(t) = (3, 3)$, and let $N^2(t) = (4, 2)$. Then a dedicated arrival to the longest queue at t_n will violate the weak majorization. Also, with infinite buffers we know that JSQ minimizes the queue length vector process, $\{N(t)\}_{t=0}^{\infty}$, in the weak majorization sense, which implies stochastic maximization of the departure process, $\{D_t\}_{t=0}^{\infty}$ (Remark 1), but here JSQ will only be optimal in the latter sense.

The case where all the customers are flexible and queues might have unequal buffer capacities was studied by Hordijk and Koole [16] and Sparaggis *et al.* [28]. They showed that JSQ (which now means routeing customers to the shortest *nonfull* queue) stochastically maximizes D_t for all $t \geq 0$. Koole *et al.* [19] showed the same result for two queues with equal buffer capacities when all customers are flexible and service times are drawn from an 'increasing likelihood ratio' (ILR) distribution. Here we extend these results by showing that JSQ stochastically maximizes $\{D_t\}_{t=0}^{\infty}$ for two or more queues with exponential service times when there is a mixture of flexible and dedicated arrivals and all buffers have the same capacities. The result does not hold for unequal capacities when there are dedicated customers ($p < 1$), as suggested by the example in the prior paragraph.

First we need the following lemma, which also holds for our earlier models. A sample path argument along the lines of Theorem 3 below proves the lemma; we omit the proof.

Lemma 3. *For any policy that idles a server when customers are present in its queue, we can construct a nonidling policy for which $\{D_t\}_{t=0}^{\infty}$ is stochastically larger.*

Theorem 3. *Let each queue have equal finite buffer capacity. The nonidling join the shortest queue policy, which routes the flexible customers to the shortest of the queues with free capacity, stochastically maximizes $\{D_t\}_{t=0}^{\infty}$.*

Proof. We show that, for an arbitrary policy Π that does not follow JSQ at an arbitrary decision epoch, we can construct a policy that does follow JSQ for that decision and has stochastically earlier departures. Let t be the first time that Π disagrees with JSQ and routes a customer to some queue, A, which is not the shortest nonfull queue. We tag this customer and give it preemptive lower priority compared to other customers, i.e. it always stays at the back of the queue. This is legitimate, since the priority policy within a queue does not affect the departure process, because service times are exponential. Let Π' be a policy that agrees with Π before time t, but routes the arrival at time t to the shortest queue, B. Consider two systems where Π and Π' are used respectively as routeing policies. We couple the arrival process for these systems so that each customer has the same arrival epoch in both systems. We also couple the service times so that each specific customer has the same service time in both systems.

First, assume that A is full before routeing, so that the tagged customer is lost under Π. Now, due to coupling, the tagged customer will either be served under Π' when Π idles server B, so we have one extra departure under Π', or the tagged customer will be lost due to an arrival when queue B is full, so both systems will be in the same state and the departure processes will be the same for both policies.

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

Now assume that A is not full before routeing the tagged customer. Let T be the first time any of the following happens:

(i) queues A and B (excluding the tagged customer) are the same length,

(ii) the tagged customer leaves under Π',

(iii) Π routes to queue A when A is full.

Note that B will not overflow under Π' before one of (i)–(iii) occurs, because it starts with a shorter queue. Now, if (i) occurs first, after interchanging the labels of A and B, both systems are in the same state and the departure processes will be the same for both policies. If (ii) occurs first then, while server B is serving the tagged customer under Π', server B is idle under Π. Let Π' continue to agree with Π until the tagged customer is being served under Π on A, and let Π' idle queue A during that time. Then the systems will agree from the point that the tagged customer leaves under Π, but departures will be stochastically earlier under Π' than under Π because the tagged customer leaves earlier under Π'. If (iii) occurs first, the tagged customer will be lost under Π, as in the case above where Π initially routes the tagged customer to a full queue, so the rest of the argument follows as in that case.

In all cases Π' is as good as Π, and, from Lemma 3, there is a nonidling policy that is as good as Π'. Repeating the argument each time Π deviates from JSQ gives us the result.

When there is a finite *shared* buffer and when dedicated arrivals are present, we do not have the majorization result. For instance, consider a system with two servers and the shared capacity is 6. Let $N^1(t)$ be the vector of queue lengths under the shortest nonfull queue policy, and define $N^2(t)$ similarly for an arbitrary routeing policy. Suppose that $N^1(t) = (3, 2)$, and let $N^2(t) = (3, 3)$. Then a dedicated arrival to the longest queue will violate weak majorization. On the other hand, when all arrivals are flexible, the shortest nonfull queue policy is optimal in the weak majorization sense, that is, $\{N^1(t)\}_{t=0}^{\infty} \prec_{\mathrm{w}} \{N^2(t)\}_{t=0}^{\infty}$. This result can easily be shown again using forward induction and considering the cases where an arrival might be lost due to capacity insufficiency.

5. Other extensions

In this section we present some easy extensions of our results.

5.1. Slotted service

Suppose that all service times are geometrically distributed, and slotted so that services start and end at integer time points. A special case of this model will be deterministic service times. JSQ will again minimize the queue length vector process in the weak majorization sense. To prove this, we couple an arbitrary routeing policy with JSQ so that, if there is a potential service completion in the kth largest queue in one system at time n, $n \in \{0, 1, 2, \dots\}$, the same is true for the other system. Note that we might have more than one departure at a single epoch. However, treating these potential service completions one by one (starting from the smallest queue so that ordering changes will not affect other couplings), we can conclude that the weak majorization will be preserved at departure.

5.2. Random service rate

Our results will also hold when the instantaneous service rate (the failure rate of the service times), which is common for all the servers, $\mu(t)$, varies according to an arbitrary stochastic process, as long as the process is independent of the queue lengths and routeing policy.

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

For example, servers could go online and offline according to a random process. Because the servers are still identical at each point in time, we can still couple service completions so that our majorization and stochastic orderings are preserved.

5.3. Random yield

We can also handle models in which service may not be successful, as long as the probability that a service is a success is independent of the state and policy, and of the number of times the service has been repeated. If the customer returns to the same queue upon an unsuccessful completion, and the success probability is constant, our results continue to hold trivially because a geometric sum of exponential service times is exponential. For a varying success probability, or if an unsuccessfully completed customer is treated as a new arrival and all arrivals are flexible, the coupling to preserve our results is easy.

5.4. Power of two choices

Suppose a flexible customer does not have full information about all the queue lengths upon arrival. Instead, two (or more) queues are randomly chosen and the customer learns their queue lengths and joins one of the those queues. For example, suppose that the facility is multilingual, but customers are at most bilingual, and all combinations of bilingualism are equally likely. Mitzenmacher [23] showed that JSQ among two queues yields almost as much advantage as JSQ among all of the queues.

Again, all of our results will hold, where now JSQ means join the shortest of the subset of queues that the customer has available. The proof is a trivial extension of the full information case.

5.5. Resequencing

Suppose that a customer cannot depart from the system until all the customers that arrived before it finish service, i.e. customers are forced to depart in order of arrival. Out-of-order customers wait in a resequencing buffer until earlier arrivals complete service. For this model, we assume that service within a queue is FCFS (and it is easy to see that this is the optimal service order to minimize departure times under resequencing). These kinds of systems are common in telecommunications where jobs (e.g. packets of a video) arrive as a stream, and they should leave in the same order as they arrived. For prior research on the topic, we refer the reader to [1], [11], and [20].

Let D_t be the number of service completions by time t, without considering resequencing, and let E_t be the number of departures actually exiting from the resequencing buffer by time t. From Remark 1 we know that $\{D_t\}_{t=0}^{\infty}$ is stochastically maximized by JSQ for an arbitrary arrival process, but the same will not be true for $\{E_t\}_{t=0}^{\infty}$. However, we can show the weaker result that $\hat{E}_i$ is stochastically minimized by JSQ for all i, where $\hat{E}_i$ is the time at which the ith customer exits the resequencing buffer. Note that JSQ stochastically minimizes $\{\hat{D}_i\}_{i=0}^{\infty}$ from Remark 1, where $\hat{D}_i$ is the ith service completion time, and where, unlike in the case for $\hat{E}_i$, the ith service completion time may not correspond to the completion time of the ith customer to arrive. Let $\hat{C}_i$ be the completion time of the ith customer to arrive, so $\hat{E}_i = \max_{j=1,\dots,i} \hat{C}_j$. Let us fix i, and consider a new arrival process that is identical to our original arrival process for the first i arrivals, but in which no customers arrive after the ith arrival. Let $\tilde{D}_j$ be the time of the jth completion, and let $\tilde{C}_j$ be the service completion time for the jth arrival, $j = 1, \dots, i$, for the system with the new arrival process. Because of our FCFS assumption within queues, the completion time of the jth arrival is unaffected by any arrivals after it, i.e. $\tilde{C}_j = \hat{C}_j$ for $j = 1, \dots, i$. From Remark 1, $\tilde{D}_i$ is stochastically minimized by JSQ, but

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

$\tilde{D}_i = \max_{j=1,\dots,i} \tilde{C}_j = \max_{j=1,\dots,i} \hat{C}_j = \hat{E}_i$, so $\hat{E}_i$ is also stochastically minimized by JSQ for any i.

Appendix A. Majorization

Weak submajorization is a preordering of Re^c denoted by '$\prec_{\mathrm{w}}$' and is defined as follows. For $x, y \in \mathrm{Re}^c$,

$$x \prec_{\mathrm{w}} y \quad \text{if} \quad \sum_1^k x_{[i]} \le \sum_1^k y_{[i]}, \qquad k = 1, \dots, c,$$

where $x_{[i]}$ denotes the components of x in decreasing order. Similarly, weak supermajorization, denoted by '$\prec^{\mathrm{w}}$', is defined as

$$x \prec^{\mathrm{w}} y \quad \text{if} \quad \sum_1^k x_{(i)} \ge \sum_1^k y_{(i)}, \qquad k = 1, \dots, c,$$

where $x_{(i)}$ denotes the components of x in increasing order. In both cases x is said to be weakly majorized by y. When $\sum_1^c x_i = \sum_1^c y_i$, $x \prec^{\mathrm{w}} y$ and $x \prec_{\mathrm{w}} y$ are equivalent, and we say that x is majorized by y, denoted by $x \prec y$.

Intuitively, if $x \prec_{\mathrm{w}} y$ then x is better balanced and smaller than y; if $x \prec^{\mathrm{w}} y$ then x is better balanced and larger than y; and if $x \prec y$ then x is better balanced than y.

Stochastic weak majorization is also defined analogously to the deterministic case. For two random vectors X and Y, we say that X stochastically weakly submajorizes Y, $X \prec_{\mathrm{w.st}} Y$, or supermajorizes Y, $X \prec^{\mathrm{w.st}} Y$, if and only if $\phi(X) \prec_{\mathrm{st}} \phi(Y)$ for all increasing, or, respectively, decreasing Schur-convex functions ϕ, where '$\prec_{\mathrm{st}}$' is the usual stochastic order [26, pp. 3–12] and a Schur-convex function is defined to be a function that preserves the majorization ordering [21]. The following definitions are equivalent:

(i) $X \prec_{\mathrm{w.st}} Y$ (respectively $X \prec^{\mathrm{w.st}} Y$),

(ii) $\phi(X) \prec_{\mathrm{st}} \phi(Y)$ for all increasing (respectively decreasing) Schur-convex functions ϕ,

(iii) $\mathrm{E}[\phi(X)] \le \mathrm{E}[\phi(Y)]$ for all increasing (respectively decreasing) Schur-convex functions ϕ,

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

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

(b) $\tilde{X} \prec_{\mathrm{w}} \tilde{Y}$ (respectively $\tilde{X} \prec^{\mathrm{w}} \tilde{Y}$) almost surely.

For ease of notation, throughout the paper, we use '$\prec_{\mathrm{w}}$' and '$\prec^{\mathrm{w}}$' for stochastic weak majorization. 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

$$\{X(t)\}_{t=0}^{\infty} \prec_{w} \{Y(t)\}_{t=0}^{\infty},$$

if we can couple the processes on the same probability space such that, for any sample path realization and any n, $X(t_i) \prec_{\mathrm{w}} Y(t_i)$ jointly for all t_i, $i = 1, \dots, n$, with probability 1. A similar definition holds for weak supermajorization.

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

Appendix B. Proof of Corollary 1

Let $i' \geq i$ be such that $a_i = a_{i+1} = \cdots = a_{i'}$ and either $a_{i'+1} < a_{i'}$ or $i' = c$. Similarly, let $j' \geq j$ be such that $b_j = b_{j+1} = \cdots = b_{j'}$ and either $b_{j'+1} < b_{j'}$ or $j' = c$. Then

$$\sum_{k=1}^{r}(a - e_i)_{[k]} = \sum_{k=1}^{r}(a - e_{i'})_{[k]} = \sum_{k=1}^{r}(a - e_{i'})_k = \sum_{k=1}^{r} a_k - \mathbf{1}\{r \geq i'\}$$

for all $1 \leq r \leq n$. Similarly, $\sum_{k=1}^{r}(b - e_j)_{[k]} = \sum_{k=1}^{r} b_k - \mathbf{1}\{r \geq j'\}$ for all $1 \leq r \leq n$. Therefore, if $i' \leq j'$, $a - e_i \prec_{\mathrm{w}} b - e_j$ is true trivially. For $i' > j'$, we need to show that

$$\sum_{k=1}^{r} a_k < \sum_{k=1}^{r} b_k \tag{9}$$

is true for $j' \leq r < i'$. If $j' < i$ then (9) is true for $j' \leq r < i$ by (2). Therefore, it is sufficient to show that (9) holds for all $i \leq r < i'$. Now suppose on the contrary that $\sum_{k=1}^{r} a_k = \sum_{k=1}^{r} b_k$ for some $i \leq r < i'$. Then, because of our definition of i',

$$\begin{aligned}\sum_{k=1}^{i-1} a_k + (r - i + 1)a_i &= \sum_{k=1}^{i-1} a_k + \sum_{k=i}^{r} a_k \\ &= \sum_{k=1}^{r} a_k \\ &= \sum_{k=1}^{r} b_k \\ &\geq \sum_{k=1}^{i-1} b_k + \sum_{k=i}^{r} b_k \\ &\geq \sum_{k=1}^{i-1} b_k + (r - i + 1)b_r.\end{aligned}$$

We have, by (2), $\sum_{k=1}^{i-1} a_k < \sum_{k=1}^{i-1} b_k$. This, together with the above, implies that $a_i > b_r$. Therefore,

$$\sum_{k=i}^{r+1} a_k = \sum_{k=i}^{r} a_k + a_i > \sum_{k=1}^{r} b_k + b_r \geq \sum_{k=1}^{r} b_k + b_{r+1} = \sum_{k=1}^{r+1} b_k,$$

which contradicts $a \prec_{\mathrm{w}} b$. Thus, we have shown that (9) is true for $i \leq r < i'$, which concludes the proof.

Appendix C. Proof of Corollary 4

Let $i \geq 1$ be such that $a_i = a_{i-1} = \cdots = a_1$ and either $a_{i+1} < a_i$ or $i = c$. It is sufficient to show that $\sum_{k=s}^{c} a_k > \sum_{k=s}^{c} b_k$ for $1 \leq s \leq i$. Suppose on the contrary that $\sum_{k=s}^{c} a_k = \sum_{k=s}^{c} b_k$ for some $1 \leq s \leq i$. Since $a \prec^{\mathrm{w}} b$, we have $\sum_{k=s+1}^{c} a_k \geq \sum_{k=s+1}^{c} b_k$.

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

Combining this with the contrary assumption, yields $a_s \leq b_s$. Hence,

$$\sum_{k=1}^{c} a_k = \sum_{k=s}^{c} a_k + (s-1)a_s = \sum_{k=s}^{c} b_k + (s-1)a_s \leq \sum_{k=s}^{c} b_k + (s-1)b_s \leq \sum_{k=1}^{c} b_k.$$

This contradicts the second assumption in the corollary. Therefore, $\sum_{k=s}^{c} a_k > \sum_{k=s}^{c} b_k$ for $1 \leq s \leq i$ and the result follows.

Acknowledgement

We would like to thank the anonymous referee for his/her useful review and comments.

References

[1] Agrawal, S. and Ramaswamy, R. (1987). Analysis of the resequencing delay for M/M/m systems. In *ACM Sigmetrics Performance Evaluation Review*, Association for Computing Machinery, New York, pp. 27–35.

[2] Akgun, O. T., Righter, R. and Wolff, R. (2011). Understanding the marginal impact of customer flexibility. Submitted.

[3] Aksin, O. Z., Karaesmen, F. and Ormeci, E. L. (2007). Workforce cross training in call centers from an operations management perspective. In *Workforce Cross Training Handbook*, ed. D. Nembhard, CRC Press, Boca Raton, FL, pp. 211–240.

[4] Argon, N. T., Ding, L., Glazebrook, K. D. and Ziya, S. (2009). Dynamic routing of customers with general delay costs in a multiserver queuing system. *Prob. Eng. Inf. Sci.* **23,** 175–203.

[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] Ephremides, A., Varaiya, P. and Walrand, J. (1980). A simple dynamic routing problem. *IEEE Trans. Automatic Control* **25,** 690–693.

[7] Foley, R. D. and McDonald, D. R. (2001). Join the shortest queue: stability and exact asymptotics. *Ann. Appl. Prob.* **11,** 569–607.

[8] Fulkerson, D. R. and Ryser, H. J. (1962). Multiplicities and minimal widths for (0,1)-matrices. *Canad. J. Math.* **14,** 498–508.

[9] Gans, N., Koole, G. and Mandelbaum, A. (2003). Telephone call centers: tutorial, review, and research prospects. *Manufact. Service Operat. Manag.* **5,** 79–141.

[10] Graves, S. C. and Tomlin, B. T. (2003). Process flexibility in supply chains. *Manag. Sci.* **49,** 907–919.

[11] Gogate, N. R. and Panwar, S. S. (1999). Assigning customers to two parallel servers with resequencing. *IEEE Commun. Lett.* **3,** 119–121.

[12] Gupta, V., Balter, M. H., Sigman, K. and Whitt, W. (2007). Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation* **64,** 1062–1081.

[13] He, Y.-T. and Down, D. G. (2009). On accommodating customer flexibility in service systems. *INFOR.* **47,** 289–295.

[14] Hopp, W. J. and Van Oyen, M. P. (2004). Agile workforce evaluation: A framework for crosstraining and coordination. *IIE Trans.* **36,** 919–940.

[15] Hopp, W. J., Tekin, E. and Van Oyen, M. P. (2004). Benefits of skill chaining in serial production lines with cross-trained workers. *Manag. Sci.* **50,** 83–98.

[16] Hordijk, A. and Koole, G. (1990). On the optimality of the generalized shortest queue policy. *Prob. Eng. Inf. Sci.* **4,** 477–487.

[17] Johri, P. K. (1989). Optimality of the shortest line discipline with state dependent service times. *Europ. J. Operat. Res.* **41,** 157–161.

[18] Jordan, W. C. and Graves, S. C. (1995). Principles on the benefits of manufacturing process flexibility. *Manag. Sci.* **41,** 577–594.

[19] Koole, G., Sparaggis, P. D. and Towsley, D. (1999). Minimizing response times and queue lengths in systems of parallel queues. *J. Appl. Prob.* **36,** 1185–1193.

[20] Kuri, J. and Kumar, A. (1994). On the optimal allocation of customers that must depart in sequence. *Operat. Res. Lett.* **15,** 41–46.

[21] Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications.* Academic Press, New York.

[22] Menich, R. and Serfozo, R. F. (1991). Optimality of routing and servicing in dependent parallel processing stations. *Queueing Systems* **9,** 403–418.

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press

[23] Mitzenmacher, M. (2001). The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distributed Systems* **12,** 1094–1104.

[24] Movaghar, A. (2005). Optimal control of parallel queues with impatient customers. *Performance Evaluation* **60,** 327–343.

[25] Reiman, M. I. (1984). Some diffusion approximations with state space collapse. In *Modelling and Performance Evaluation Methodology* (Paris, 1983; Lecture Notes Control Inf. Sci. **60**), Springer, Berlin, pp. 209–240.

[26] Shaked, M. and Shanthikumar, J. G. (1994). *Stochastic Orders and Their Applications.* Academic Press, Boston, MA.

[27] Sparaggis, P. D. and Towsley, D. (1994). Optimal routing and scheduling of customers with deadlines. *Prob. Eng. Inf. Sci.* **8,** 33–49.

[28] 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.

[29] 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. on Decision and Control*, pp. 658–663.

[30] Weber, R. R. (1978). On the optimal assignment of customers to parallel servers. *J. Appl. Prob.* **15,** 406–413.

[31] Whitt, W. (1986). Deciding which queue to join: some counterexamples. *Operat. Res.* **34,** 55–62.

[32] Winston, W. (1977). Optimality of the shortest line discipline. *J. Appl. Prob.* **14,** 181–189.

https://doi.org/10.1239/aap/1324045695 Published online by Cambridge University Press