Partial Flexibility in Routing and Scheduling

by

Osman Tansu Akgun

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering-Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Rhonda L. Righter, Chair
Professor Ronald W. Wolff
Professor Lee W. Schruben
Professor Jean Walrand

Fall 2012

Partial Flexibility in Routing and Scheduling

Abstract

Partial Flexibility in Routing and Scheduling

by

Osman Tansu Akgun

Doctor of Philosophy in Engineering - Industrial Engineering and Operations Research

University of California, Berkeley

Professor Rhonda L. Righter, Chair

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 study partial flexibility in multi-server queueing systems. Some (dedicated) arrivals are obliged to use a particular station, while others (flexible) have the ability to use any of the stations.

We first consider the optimal routing policy for the flexible customers under various settings. Flexible customers should be routed by the decision maker to one of queues upon arrival. When the only information available upon arrival is the queue lengths, then"Join the Shortest Queue" (JSQ) has been shown to be optimal in a variety of contexts. We generalize earlier results on the optimality of "Join the Shortest Queue" for flexible arrivals to the following: arbitrary arrivals where only a subset are flexible, multiple-server stations, and abandonments. Our optimality results are very strong; we minimize the queue length process in the weak majorization sense. When the actual workload at each queue is known upon arrival but the required work of the arriving customer is unknown, we show that routing flexible customers to the queue with the shortest workload, known in the literature as the "Join the Shortest Work" (JSW) policy, is optimal for general arrival and service processes.

Secondly, we study the scheduling problem in an alternate design where the flexible customers have a separate queue to wait. We show that the optimal scheduling policy is "Dedicated Customers First" (DCF) policy under various settings. This design is better than the routing design in terms of system performance; however it is not a good design in terms of fairness, because flexible customers face long waiting times.

Finally we consider the marginal impact of customer flexibility. We present our new results showing that the stationary expected waiting time is decreasing and convex in

the proportion of flexible customers. Although convexity in the proportion of flexible customers is intuitive, it does not hold in the strong sense that monotonicity holds, and it is surprisingly difficult to prove. We develop a new approach that combines marginal analysis with coupling to show convexity in the stationary mean waiting time. Our results reinforce the idea that a little flexibility goes a long way.

To Leah,
who has always made me happy...

Contents

List of Tables

List of Figures

Acknowledgments

First and above all, I would like to thank my advisor, Professor Rhonda Righter. Far away from my real family, she was like a mother to me at Berkeley. She never ceased to believe in me. It won't be enough no matter how much I appreciate what she has done for me.

I would also like to thank my co-advisor Professor Ronald Wolff. Working with a legendary name like him was a privilege and honor for me. I would like to thank my dissertation committee members Professor Lee Schruben and Professor Jean Walrand for their valuable comments and feedback on this research.

I was thousands of miles away from my family, but I never had a moment without feeling their support. I would like to thank my parents Gulten Akgun and Ahmet Akgun, to my brother and my biggest supporter Tolga Akgun, to his lovely wife Yurdanur Abla, and of course to the most recent family member and our greatest source of joy, Batu Akgun.

I would like to acknowledge my good old friend, Tevfik Turhan. I would thank him for understanding everything I mean and everything I feel, whenever I needed someone for that. I would also thank another old friend Resat Sen for his mental and physical (mac laptop) support before this journey began. I would like to acknowledge my Berkeley friends Emrehan Baba, Kerem Tutuncu and Serhat Sag. They contributed to my success at Berkeley much much more than I contributed to their Masters studies, and of course I couldn't be a better PES player without them. Last but not the least, I would like to thank Sarp Istanbulluoglu for his valuable support, and above all, I would like to thank him for pulling me out of deep waters when I needed.

Chapter 1

Introduction

1.1 Flexibility in Multi-Server Systems

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 flexibility to these systems through servers or through customers. In the former case, systems have flexible servers that can serve multiple types of customers, whereas in the latter, systems have flexible customers that can be served by different servers.

Server flexibility is a widely studied problem in the literature of multi-server systems. See, e.g., Aksin, Karaesmen and Ormeci [6], Graves and Tomlin [26], Hopp, Tekin and van Oyen [34], Hopp and van Oyen [35], and Jordan and Graves [39]. However, such flexibility is still generally expensive, because having a flexible server costs more than a non-flexible server. 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. Therefore we focus on customer flexibility.

Our following motivating application about the design of call centers is a good example to differentiate between the two types of flexibility. Design of call centers has been a popular research area in queueing systems. Aksin et al. [5] and Gans et al. [24] provide extensive literature surveys. We consider a call center that 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. The first option for the company to tackle this problem is to train or hire bilingual agents (flexible servers). Because of the training expense and high turnover of agents, the company is in favor of the second option, which is to add a "Press 0" option for bilingual customers (flexible 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, because it is taking advantage of flexibility that is already present in the customers.

Another example of customer flexibility is to highway traffic, through the Mobile Millennium project, 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 participating users for personal use in choosing alternate routes. A similar application is to communications and Internet routing, in which some 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.

1.2 Analysis of Flexibility

Flexible multi-server systems are in general designed and controlled by a centralized decision maker to optimize the system performance relative to some objective. This objective can be smaller costs, higher throughput, smaller waiting times, etc. The decision maker's task can be grouped into the following three categories [57].

(i) System Design: This task involves the design of the overall server network, and addresses questions such as how the customers should be queued, which types of customers a server should be able to serve, what the capacity of the system should be. Some of the well-known network topologies related to the multi-server system design are given in Figure 1.1 (from Garnett and Mandelbaum [25]).

For instance the "I" design is the generic single queue model. There is a single type of customers and customers join the queue in the order of arrival.

In "N" designs, there are two types of customers; one (flexible) can be served by either server, whereas the other (dedicated) can be served by only one (dedicated) server. In the N_s design, customers of each type join their own queue.

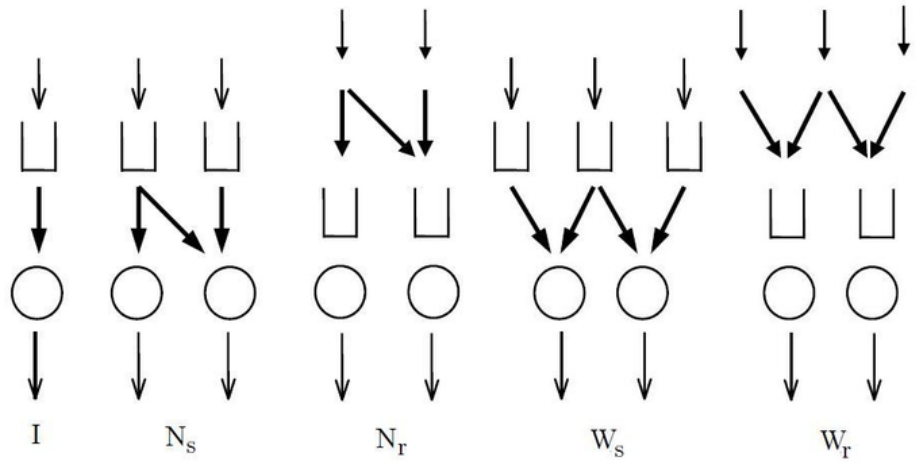

Figure 1.1: Some canonical network designs.

We have a scheduling problem. When a server is free, a decision is made as to whether to serve a customer there, and when there is a choice, which type to serve. In the N_r design, each server has its own queue. We have a routing problem. Flexible customers are assigned to a queue on arrival.

In "W" designs, flexible customers can be served by either server, but now each server has a dedicated stream of customers who can be served only there. So in addition to flexible customers, there are two types of dedicated customers. In the W_s design, customers of each type join their own queue, and we have a scheduling problem. In the W_r design, each server has its own queue, and we have a routing problem.

For instance, in our call center example of Section 1.1, if the company prefers to hire only monolingual agents, then one possible design is to route the bilingual customers to Spanish or English speaking agents upon arrival (the W_r design of Figure 1.1). Another alternative is to route bilingual customers to a separate queue (the W_s design of Figure 1.1).

(ii) Optimal Scheduling and Routing: For a specific network design, the decision maker has to determine a control policy to optimize required objectives. This control policy involves scheduling and/or routing problems depending on the system design. Routing is the assignment of a customer to one of the queues upon arrival. On the other hand the scheduling problem involves determining the type of customer to serve, when a server is idle. As an example let us again consider our bilingual call center model with monolingual agents. If the network design is the W_r design of Figure 1.1, then the decision maker has to

decide which queue bilingual customers should be assigned to (the Spanish or the English queue). If the network design is the W_s design of Figure 1.1, then the decision maker has to decide whether an idle agent, say Spanish speaking, should serve a waiting bilingual customer or should serve a waiting Spanish speaking customer.

(iii) Performance Analysis: This step involves the comparison of different designs and policies. The decision maker has to determine the optimal system design and control policy for specific system parameters and objectives. There are various methods to compare different policies and designs such as stochastic majorization, dynamic programming and simulation.

We consider all of the above three tasks for the multi-server system with flexible customers. We study different system designs, optimal control polices and comparison of different designs and policies in detail.

1.3 Partial Flexibility - A Little Flexibility Goes a Long Way

An important aspect of flexible multi-server systems is the amount of flexibility introduced to the system. In general, more flexibility intuitively leads to better system performance, but it may have higher costs; or the amount of flexibility can be restricted (in the bilingual call center model, flexibility is limited by the proportion of bilingual customers). Therefore determining system performance at different levels of flexibility is a significant problem in flexible multi-server systems.

In the case of flexible servers, the marginal impact of additional flexibility is a widely studied topic. The striking outcome of these studies is the fact that small amounts of flexibility can achieve a performance that is very close to the performance of full flexibility. This idea in the literature is known as "a little flexibility goes a long way". Jordan and Graves [39] show that in a single period newsvendor network, where products with lower than expected demand can be shifted to those with higher than expected demand, level 2 chaining, which allows a shift of at most two different products, has an expected system utilization that is very close to full-level chaining. Bassamboo et. al. [12] consider a parallel server system with N different types of customers and N servers, and show that when each server is able to process only two type of customers and when each customer can be processed by only two servers (level 2 flexibility), the holding cost plus capacity cost is asymptotically minimized. In the

context of server flexibility in a computer processing environment, Tsitsiklis and Xu [64] show that a little flexibility, in terms of the proportion of an available resource that can be deployed in a centralized (flexible) manner rather than being allocated in a decentralized fashion to local servers dedicated to local requests, can yield a large impact in heavy traffic.

As noted earlier, in our multi-server problem with flexible customers, the amount of flexibility in general is limited. Therefore validity of the "a little flexibility goes a long way" idea is very important. For instance, in the bilingual call center model, if the benefit is small when the proportion of bilingual customers is small, then implementing the "Press 0" option for bilingual customers would not be very applicable, because the proportion of bilingual customers is in fact often small.

Figure 1.2 depicts simulation results for two M/M/1 queues with overall traffic intensity $\rho = .9$ and where $\overline{N}$ denotes the long run average number of customers in the system and p denotes the proportion of flexible arrivals among all arrivals. When none of the arrivals are flexible ($p = 0$), this system becomes two separate $M/M/1$ queues. On the other hand when all customers are flexible and are routed to the shortest queue upon arrival (i.e. $p = 1$), the mean number of customers in the system is still (slightly) larger than the corresponding mean for an $M/M/2$ queue with each server having a service rate of μ, and this in turn has more customers on average than a single-server system with twice the service rate ($M/M/1(2\mu)$. We can see from this graph that at $p = 20\%$ we have about 80% of the benefit relative to the total benefit of going from $p = 0$ to $p = 1$. That is, roughly, "a little bit of flexibility goes a long way."

We study the marginal impact of customer flexibility in multi-server systems in detail and develop results supporting the idea that "a little bit of flexibility goes a long way."

1.4 Outline of the Thesis

Chapter 2 introduces majorization which is an analysis tool used widely in the thesis. We give definitions of both deterministic and stochastic majorization as well as providing lemmas and corollaries that are used in the later chapters.

In Chapter 3 we study the W_r design of Figure 1.1 and consider the relevant routing problem. We discuss the optimal control policy under different settings where the available information upon arrival is either the queue length or the exact workload

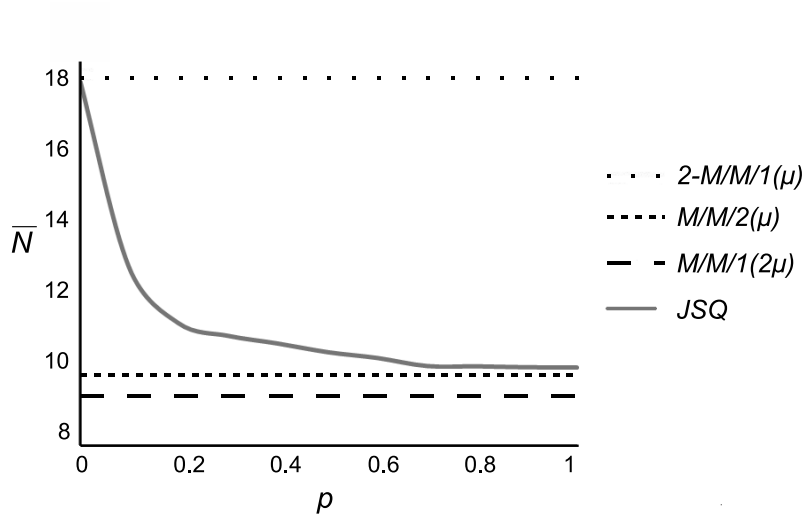

Figure 1.2: Long-run average of the total number of customers.

at each queue.

In Chapter 4, we consider the W_s design of Figure 1.1 and the related scheduling problem. For this design, we show that in many situations the dedicated customers first (DCF) policy is optimal. Under DCF, whenever a server's dedicated queue is non-empty it gives priority to dedicated customers and does not idle.

Finally in Chapter 5, we focus on the marginal impact of customer flexibility. We show that, for the routing problem when the available information is the queue-length vector, the stationary expected waiting time is decreasing and convex in the proportion of flexible customers. We show a stronger convexity result for a variant of our model that includes inventory. Our results reinforce the idea that "a little flexibility goes a long way."

We discuss the relevant literature for each of our topics within the corresponding sections and subsections (for our earlier discussions about the topics, see [7] for Chapter 3, [10] for Chapter 4 and [9] for Chapter 5).

Chapter 2

Majorization

2.1 Background Material on Majorization

In this chapter we study the theory of majorization. Majorization is a mathematical tool which is widely used in the analysis of queueing systems. We also make frequent use of majorization throughout this study. We first present some notations.

$$
\begin{array}{rcl}
\mathbb{R}^n & : & \text{n-dimensional real vector space.} \\
x_{[i]} & : & i\text{th largest component of vector } x \in \mathbb{R}^n,\ i = 1, ..., n. \\
x_{(i)} & : & i\text{th smallest component of vector } x \in \mathbb{R}^n,\ i = 1, ..., n. \\
e_i & : & i\text{th unit vector in } \mathbb{R}^n,\ i = 1, ..., n. \\
x - t & : & (x_1 - t, ..., x_n - t) \text{ for } x \in \mathbb{R}^n \text{ and } t \in \mathbb{R}. \\
x^+ & : & (max\{x_1, 0\}, ..., max\{x_n, 0\}) \text{ for } x \in \mathbb{R}^n. \\
x^- & : & (min\{x_1, 0\}, ..., min\{x_n, 0\}) \text{ for } x \in \mathbb{R}^n. \\
{}^i x & : & \big((x + e_i)_{[1]}, ..., (x + e_i)_{[n]}\big) \text{ for } x \in \mathbb{R}^n,\ i = 1, ..., n. \\
{}_i x & : & \big((x - e_i)_{[1]}, ..., (x - e_i)_{[n]}\big) \text{ for } x \in \mathbb{R}^n,\ i = 1, ..., n.
\end{array}
$$

Majorization is a preordering of $\mathbb{R}^n$ denoted by $\prec$ and is defined as follows. For $x, y \in \mathbb{R}^n$,

$$
x \prec y \text{ if } \begin{cases} \sum\limits_1^k x_{[i]} \le \sum\limits_1^k y_{[i]}, & k = 1, ..., n \\ \sum\limits_1^n x_{[i]} = \sum\limits_1^n y_{[i]}. \end{cases}
$$

Note that

$$x \prec y \Rightarrow \sum_{1}^{k} x_{(i)} \geq \sum_{1}^{k} y_{(i)}, k = 1, ..., n.$$

If the equality constraint used in the definition of majorization is relaxed, the concept of "weak" majorization arises. Weak submajorization, denoted by $\prec_w$, is defined as follows:

$$x \prec_w y \text{ if } \sum_{1}^{k} x_{[i]} \leq \sum_{1}^{k} y_{[i]}, k = 1, ..., n.$$

Similarly weak supermajorization, denoted by $\prec^w$, is defined as follows:

$$x \prec^w y \text{ if } \sum_{1}^{k} x_{(i)} \geq \sum_{1}^{k} y_{(i)}, k = 1, ..., n.$$

In both cases x is said to be weakly majorized by y. We also have that [48, pp. 122-123]

$$\begin{aligned} x \prec_w y &\Leftrightarrow \text{ for some } u, x \leq u \text{ and } u \prec y \\ x \prec^w y &\Leftrightarrow \text{ for some } v, x \prec v \text{ and } v \geq y. \end{aligned} \tag{2.1}$$

Intuitively, if $x \prec_w y$, then x is better balanced and smaller than y, if $x \prec^w y$, then x is better balanced and larger than y, and if $x \prec y$, then x is better balanced than y.

A Schur-convex function is defined to be a function that preserves the majorization ordering, i.e., a real valued function ϕ is said to be Schur-convex if $x \prec y \Rightarrow \phi(x) \leq \phi(y)$. Therefore, (2.1) implies that

$$x \prec_w y \Rightarrow \phi(x) \leq \phi(y), \text{ for all increasing Schur-convex functions } \phi,$$

$$x \prec^w y \Rightarrow \phi(x) \leq \phi(y), \text{ for all decreasing Schur-convex functions } \phi.$$

Stochastic weak majorization is also defined analogously to the deterministic case. For two random vectors X, Y, we say that X stochastically weakly submajorizes [supermajorizes] Y, $X \prec_{w.st} [\prec^{w.st}] Y$, if and only if $\phi(X) \prec_{st} \phi(Y)$, for all increasing [decreasing] Schur-convex functions ϕ where $\prec_{st}$ is the usual stochastic order [58, pp. 3-12]. The following definitions are equivalent:

(i) $X \prec_{w.st} [\prec^{w.st}] Y$

(ii) $\phi(X) \prec_{st} \phi(Y)$, $\forall$ increasing[decreasing] Schur-convex functions ϕ

(iii) $E[\phi(X)] \leq E[\phi(Y)]$, $\forall$ increasing[decreasing] Schur-convex functions ϕ

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

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

(b) $\tilde{X} \prec_w [\prec^w] \tilde{Y}$ a.s.

For ease of notation, we use $\prec_w$ and $\prec^w$ for stochastic weak majorization throughout. 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 for any n, $X(t) \prec_w Y(t)$ jointly for all t, with probability 1. A similar definition holds for weak supermajorization.

In the remainder of this chapter we present some majorization lemmas and corollaries that we need in the upcoming chapters.

Lemma 1. *Let $a_1 \geq ... \geq a_n$ and $b_1 \geq ... \geq b_n$ be integers.*

(i) If $a \prec_w b$, then

$$a + e_i \prec_w b + e_j, \forall i \geq j, \tag{2.2}$$

$$a - e_i \prec_w b - e_j, \forall i \leq j. \tag{2.3}$$

(ii) If $a \prec^w b$, then

$$a + e_i \prec^w b + e_j, \forall i \geq j, \tag{2.4}$$

$$a - e_i \prec^w b - e_j, \forall i \leq j. \tag{2.5}$$

Proof. We will just prove (2.2) and (2.4) as the other proofs are similar. First we show (2.2). The vectors $a + e_i$ and $b + e_j$ may not have components in decreasing magnitude, but if $i' \leq i$ is chosen so that

$$a_i = a_{i-1} = ... = a_{i'} \text{ and } \text{ either } a_{i'} < a_{i'-1} \text{ or } i' = 1,$$

then $a + e_{i'}$ has the components of $a + e_i$ reordered decreasingly. Similarly, if $j' \leq j$ satisfies

$$b_j = b_{j-1} = ... = b_{j'} \text{ and } \text{ either } b_{j'} < b_{j'-1} \text{ or } j' = 1,$$

then $b+e_{j'}$ has the components of $b+e_j$ reordered decreasingly. Rather than showing $a+e_i \prec_w b+e_j$, it is more convenient to show the equivalent fact that

$$u \equiv a+e_{i'} \prec_w b+e_{j'} \equiv v.$$

For $r < min(i',j')$ we have

$$\sum_{k=1}^{r} u_k = \sum_{k=1}^{r} a_k \leq \sum_{k=1}^{r} b_k = \sum_{k=1}^{r} v_k.$$

For $r \geq max(i',j')$ we have

$$\sum_{k=1}^{r} u_k = \sum_{k=1}^{r} a_k + 1 \leq \sum_{k=1}^{r} b_k + 1 = \sum_{k=1}^{r} v_k.$$

If $j' \leq i'$, then we have that for $j' \leq r < i'$,

$$\sum_{k=1}^{r} u_k = \sum_{k=1}^{r} a_k < \sum_{k=1}^{r} b_k + 1 = \sum_{k=1}^{r} v_k.$$

It remains to show, for the case that $j' > i'$ and $i' \leq r < j'$, that $\sum_{k=1}^{r} u_k \leq \sum_{k=1}^{r} v_k$. Notice that $\sum_{k=1}^{r} u_k \leq \sum_{k=1}^{r} v_k$ is equivalent to $\sum_{k=1}^{r} a_k < \sum_{k=1}^{r} b_k$. If $a_r < b_r$ then $\sum_{k=1}^{r}(b_k - a_k) > \sum_{k=1}^{r-1}(b_k - a_k) \geq 0$. The remaining case is $a_r \geq b_r$. Because $i' \leq r < j' \leq j \leq i$,

$$a_{j'} = ...a_{r+1} = a_r \geq b_r \geq b_{r+1} \geq ... \geq b_{j'-1} > b_{j'},$$

and

$$0 < -\sum_{k=r+1}^{j'} (b_k - a_k) \leq -\sum_{k=r+1}^{j'} (b_k - a_k) + \sum_{k=1}^{j'} (b_k - a_k) = \sum_{k=1}^{r} (b_k - a_k).$$

Next we will show (2.4). We define $i' \leq i$ and $j' \leq j$ as above, so that $a+e_{i'}$ and $b+e_{j'}$ have their components reordered decreasingly. We will show that

$$u \equiv a+e_{i'} \prec^w b+e_{j'} \equiv v.$$

For $r > max(i',j')$ we have

$$\sum_{k=r}^{n} u_k = \sum_{k=r}^{n} a_k \geq \sum_{k=r}^{n} b_k = \sum_{k=r}^{n} v_k.$$

For $r \leq min(i', j')$ we have

$$\sum_{k=r}^{n} u_k = \sum_{k=r}^{n} a_k + 1 \geq \sum_{k=r}^{n} b_k + 1 = \sum_{k=r}^{n} v_k.$$

If $i' \geq j'$, then we have that for $j' < r \leq i'$,

$$\sum_{k=r}^{n} u_k = \sum_{k=r}^{n} a_k + 1 > \sum_{k=r}^{n} b_k = \sum_{k=r}^{n} v_k.$$

It remains to show, for the case that $i' < j'$ and $i' < r \leq j'$, that $\sum_{k=r}^{n} u_k \geq \sum_{k=r}^{n} v_k$. Notice that it is sufficient to show that $\sum_{k=r}^{n} a_k > \sum_{k=r}^{n} b_k$. If $a_r < b_r$, then $\sum_{k=r}^{n}(a_k - b_r) > \sum_{k=r-1}^{n}(a_k - b_k) \geq 0$. The remaining case is $a_r \geq b_r$. Because $i' < r \leq j' \leq j \leq i$,

$$a_{j'} = a_{j'-1} = ... = a_r \geq b_r \geq b_{r+1} \geq ... \geq b_{j'-1} > b_{j'},$$

and

$$0 < \sum_{k=r}^{j'}(a_k - b_k) \leq \sum_{k=r}^{j'}(a_k - b_k) + \sum_{k=j'+1}^{n}(a_k - b_k) = \sum_{k=r}^{n}(a_k - b_k). \quad \blacksquare$$

The following corollary easily follows from Lemma 1, because the sum of the components of the vectors a and b are still equal after unit vector operations.

Corollary 2. *Let $a_1 \geq ... \geq a_n$ and $b_1 \geq ... \geq b_n$ be integers. If $a \prec b$, then*

$$a + e_i \prec b + e_j, \forall i \geq j,$$
$$a - e_i \prec b - e_j, \forall i \leq j.$$

We also have the following corollary in which the components of the vectors are bounded below.

Corollary 3. *Let $a_1 \geq ... \geq a_n \geq M$ and $b_1 \geq ... \geq b_n \geq M$ be integers. If $a \prec_w b$, then*

$$max\{(a - e_i), M\} \prec_w max\{(b - e_j), M\}, \forall i \leq j.$$

Proof. If $a_i > M$ and $b_j > M$, then the result follows from Lemma 1. If $b_j = M$, then $max\{(a - e_i), M\} \prec_w a \prec_w b = max\{(b - e_j), M\}$. Finally suppose $a_i = M$ and

$b_j > M$. Let $i' \leq i$ be such that $a_{i'} = a_{i'+1} = ... = a_i = M$ and either $a_{i'-1} > M$ or $i' = 1$. Then the vector v defined below is such that $v_{[k]} = v_k$ where:

$$v_r = \begin{cases} a_r & k \neq i' \\ M+1 & k = i' \end{cases}$$

Then for $j < i'$,

$$\sum_{k=1}^{j} v_k = \sum_{k=1}^{j} a_k \leq \sum_{k=1}^{j} b_k$$

And for $j \geq i'$,

$$\sum_{k=1}^{j} v_k = \sum_{k=1}^{j} a_k + 1 = \sum_{k=1}^{i'-1} a_k + 1 \leq \sum_{k=1}^{i'-1} b_k + \sum_{k=i'}^{j} b_k \leq \sum_{k=1}^{j} b_k$$

because $b_{i'} \geq b_{i'+1} \geq ... \geq b_i \geq ... \geq b_j > M$. Hence $v \prec_w b$. Finally $max\{(a - e_i), M\} = (v - e_{i'}) \prec_w (b - e_j) = max\{(b - e_j), M\}$ where the weak majorization follows from Lemma 1. ∎

The following corollary provides conditions under which we can extend Lemma 1.

Corollary 4. *Let $a_1 \geq ... \geq a_n \geq 0$ and $b_1 \geq ... \geq b_n \geq 0$ be integers and fix $i > j$.*

(i) If $a \prec_w b$ and

$$\sum_{k=1}^{s} a_k < \sum_{k=1}^{s} b_k \tag{2.6}$$

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

$$(a - e_i) \prec_w (b - e_j).$$

(ii) If $a \prec^w b$ and

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

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

$$(a - e_i) \prec^w (b - e_j).$$

Proof. We will just prove (i) as the proof of (ii) is similar. Let $i' \geq i$ be such that $a_i = a_{i+1} = ... = a_{i'}$ and either $a_{i'+1} < a_{i'}$ or $i' = c$. Similarly let $j' \geq j$ be such that $b_j = b_{j+1} = ... = b_{j'}$ 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 $i' \leq j'$, $(a - e_i) \prec_w (b - e_j)$ is true trivially. For $i' > j'$, we need to show

$$\sum_{k=1}^{r} a_k < \sum_{k=1}^{r} b_k \tag{2.7}$$

is true for $j' \leq r < i'$. If $j' < i$, then (2.7) is true for $j' \leq r < i$ by (2.6). Therefore it is sufficient to show that (2.7) 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 (2.6) that $\sum_{k=1}^{i-1} a_k < \sum_{k=1}^{i-1} b_k$. This, together with the above, implies $a_i > b_r$. 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 $a \prec_w b$. Thus we have shown that (2.7) is true for $i \leq r < i'$ which concludes the proof. ∎

Corollary 5. *Let $a_1 \geq ... \geq a_n \geq 0$ and $b_1 \geq ... \geq b_n \geq 0$ be integers and $\sum_{k=1}^{n} a_k > \sum_{k=1}^{n} b_k$. Then*

(i) If $a \prec^w b$,

$$(a - e_1) \prec^w b.$$

(ii) If $a \succ_w b$,

$$a \succ_w (b + e_n).$$

Proof. We will just prove (i) as the proof of (ii) is similar. Let $i \geq 1$ be such that $a_i = a_{i-1} = ... = a_1$ and either $a_{i+1} < a_i$ or $i = c$. It is sufficient to show that $\sum_{k=s}^{n} a_k > \sum_{k=s}^{n} b_k$, for $1 \leq s \leq i$. Suppose on the contrary that $\sum_{k=s}^{n} a_k = \sum_{k=s}^{n} b_k$, for some $1 \leq s \leq i$. Since $a \prec^w b$, we have $\sum_{k=s+1}^{n} a_k \geq \sum_{k=s+1}^{n} b_k$. Combining this with the contrary assumption yields $a_s \leq b_s$. Hence:

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

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

Note that even though the components of a vector x are in decreasing order, $x+e_i$ may not be, so we use the notation ${}^{i}x$ to reorder it. For the 2-dimensional vectors we have the following lemma.

Lemma 6. *Let* $a_1 \geq a_2$, $b_1 \geq b_2$, $c_1 \geq c_2$ *and* $d_1 \geq d_2$ *be integers. If* $a \prec b \prec d$, $a \prec c \prec d$ *and* $b + c \prec a + d$, *then*

(i) ${}^{i}b + {}^{j}c \prec {}^{i}a + {}^{j}d, \forall i \geq j = 1, 2,$

(ii) ${}_{i}b + {}_{i}c \prec {}_{i}a + {}_{i}d, i = 1, 2.$

Proof. We just prove (i) as the proof of (ii) is similar. Note that our assumptions are equivalent to $a_1 \leq b_1 \leq d_1$, $a_1 \leq c_1 \leq d_1$, $a_1 + a_2 = b_1 + b_2 = c_1 + c_2 = d_1 + d_2$ and $b_1 + c_1 \leq a_1 + d_1$. For $i = j = 1$, the result is trivial. For $i = j = 2$, we need to take possible order changes into account:

(i) $d_1 = d_2$: Then $a_k = b_k = c_k = d_k$, $k = 1, 2$, and the result follows.

(ii) $d_1 > d_2, b_1 = b_2, c_1 > c_2$: Then $a_k = b_k$, $k = 1, 2$, so $(a + e_2) = (b + e_2)$ and $(b + e_2)_{[1]} + (c + e_2)_{[1]} = b_1 + c_1 + 1 \leq a_1 + d_1 + 1 = (a + e_2)_{[1]} + (d + e_2)_{[1]}$, so the result follows.

(iii) $d_1 > d_2, c_1 = c_2, b_1 > b_2$: Similar to the previous case.

(iv) $d_1 > d_2, c_1 = c_2, b_1 = b_2$: Then $a_1 = a_2 = b_1 = b_2 = c_1 = c_2 < d_1$, so $(a+e_2)_{[1]} = a_1+1 = (b+e_2)_{[1]} = b_1+1 = (c+e_2)_{[1]} = c_1+1 \leq d_1 = (d+e_2)_{[1]}$, and $(b+e_2)_{[1]} + (c+e_2)_{[1]} = b_1+c_1+2 \leq a_1+d_1+1 = (a+e_2)_{[1]} + (d+e_2)_{[1]}$, and the result follows.

(v) $d_1 > d_2, b_1 > b_2, c_1 > c_2$: Then for $a_1 > a_2$, the result is immediate. For $a_1 = a_2$, $(b+e_2)_{[1]} + (c+e_2)_{[1]} = b_1+c_1 \leq a_1+d_1+1 = (a+e_2)_{[1]} + (d+e_2)_{[1]}$, and the result follows.

Finally we look at $j = 1, i = 2$. We again need to take order changes into account. If $b_1 = b_2$, then $a_k = b_k$, $k = 1, 2$, and the result follows. If $b_1 > b_2$ and $a_1 = a_2$, then $(b+e_2)_{[1]} = b_1$, $(a+e_2)_{[1]} = a_1+1$, so $(b+e_2)_{[1]} + (c+e_2)_{[1]} = b_1+c_1+1 \leq a_1+d_1+2 = (a+e_2)_{[1]} + (d+e_2)_{[1]}$, so the result follows. ∎

Note that the above lemma is not true for vectors with larger dimensions (i.e., three or more queues). For a counterexample let $a = (2,1,0)$, $b = (3,1,-1)$, $c = (3,3,-3)$, $d = (4,3,-4)$. If $i = 3$ and $j = 2$, then ${}^i a = (2,1,1)$, ${}^i b = (3,1,0)$, ${}^j c = (4,3,-3)$, ${}^j d = (4,4,-4)$. Hence ${}^i b + {}^j c = (7,4,-3) \not\prec (6,5,-3) = {}^i a + {}^j d$, i.e. the majorization on the sums is violated.

Lemma 7. *Let $a_1 \geq a_2$, $b_1 \geq b_2$, $c_1 \geq c_2$ and $d_1 \geq d_2$ be integers. If $a \prec b \prec d$, $a \prec c \prec d$ and $b + c \prec a + d$, then*

(i) $\sum_{i=1}^{2} \{b_i^+ + c_i^+\} \leq \sum_{i=1}^{2} \{a_i^+ + d_i^+\}$,

(ii) $\sum_{i=1}^{2} \{b_i^- + c_i^-\} \geq \sum_{i=1}^{2} \{a_i^- + d_i^-\}$.

Proof. The proof of (ii) follows from (i) since the majorization assumptions imply $\sum_{i=1}^{2}(b_i + c_i) = \sum_{i=1}^{2}(a_i + d_i)$. For the proof of (i) we will consider the following cases:

(i) $a_1 \leq 0, b_1 \leq 0, c_1 \leq 0$: Then the result follows trivially.

(ii) $a_1 \leq 0, b_1 \leq 0, c_1 > 0$: Since $c \prec d$, the result follows.

(iii) $a_1 \leq 0, b_1 > 0, c_1 \leq 0$: Similar to the previous case.

(iv) $a_1 \leq 0, b_1 > 0, c_1 > 0$: $a_1 \leq 0$ implies $a_2, b_2, c_2, d_2 \leq 0$. And $b_1 + c_1 \leq a_1 + d_1 \leq d_1$, so the result follows.

(v) $a_1 > 0, a_2 \leq 0$: Then $b_2, c_2, d_2 \leq 0$, and $b_1 + c_1 \leq a_1 + d_1$, so the result follows.

(vi) $a_1 > 0, a_2 > 0$: Then all entries are positive and the result follows because $b + c \prec a + d$. ∎

For $x, y \in \mathbb{R}^n$, we say $x \leq y$ 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 8. *Let $a_1 \geq ... \geq a_n \geq 0$ and $b_1 \geq ... \geq b_n \geq 0$ be real numbers. If $a \leq b$, then for any real t,*

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

(ii) $(a + te_i) \leq (b + te_i), \forall i$,

(iii) $(a + te_i) \prec_w (b + te_j), \forall i \geq j$.

Proof. (i) and (ii) follow trivially because $(a_i - t)^+ \leq (b_i - t)^+$ and $a_i + t \leq b_i + t$ for all $i = 1, ..., n$. Now to prove (iii) let $i' \leq i$ be such that either we have $a_{i'-1} \geq a_i + t \geq a_{i'}$ for $i' \geq 2$ or we have $i' = 1$. Let $j' \leq j$ be defined similarly. Let us also define u and v, where we have $u = a + te_i$, $v = b + te_j$ and the components of u and v are in decreasing order.

If $i' \geq j'$, then we have, $\sum_{k=1}^r u_k = \sum_{k=1}^r a_k \leq \sum_{k=1}^r b_k = \sum_{k=1}^r v_k$, for $r < j'$, $\sum_{k=1}^r u_k = \sum_{k=1}^r a_k + t \leq \sum_{k=1}^r b_k + t = \sum_{k=1}^r v_k$, for $r \geq i'$, and $\sum_{k=1}^r u_k = \sum_{k=1}^r a_k \leq \sum_{k=1}^r b_k < \sum_{k=1}^r v_k$, for $j' \leq r < i'$. Hence (iii) follows.

If $i' \leq j'$, then we again have $\sum_{k=1}^r u_k = \sum_{k=1}^r a_k \leq \sum_{k=1}^r b_k = \sum_{k=1}^r v_k$, for $r < i'$, $\sum_{k=1}^r u_k = \sum_{k=1}^r a_k + t \leq \sum_{k=1}^r b_k + t = \sum_{k=1}^r v_k$, for $r \geq j'$. Finally for $i' \leq r < j'$,

$$\sum_{k=1}^r u_k = \sum_{k=1}^{r-1} a_k + a_i + t \leq \sum_{k=1}^{r-1} b_k + b_j + t \sum_{k=1}^{r-1} b_k + b_{j'-1} \leq \sum_{k=1}^{r-1} b_k + b_r = \sum_{k=1}^r v_k,$$

and the result follows. ∎

Chapter 3

Optimal Routing

In this chapter we study the routing problem in a multi-server multi-queue system with partial flexibility. We consider a queuing system with c parallel servers. Some (*dedicated*) arrivals are obliged to use a particular server, while others (*flexible*) have the ability to use any of the c servers. Flexible customers should be routed by the decision maker to one of the c queues upon arrival and then customers at a particular queue are served according to "First-Come-First-Served (FCFS)" order. This problem can be modeled as the W_r design of Section 1.2.

The optimal routing policy for our model depends on the information available at the decision epochs. We first consider the problem where the only information available at a decision epoch is the history of decisions. In this case, Ephremides et al. [20] show that the "Round Robin" policy minimizes the sum of expected completion times when the service times are exponential with the same rate and when all customers are flexible. Liu and Towsley [45] extend the optimality of the "Round Robin" policy to iid service times with increasing hazard rate, and Liu and Righter [44] show the optimality of the "Round Robin" policy for general iid service time distributions with homogeneous dedicated arrivals. When the service times are not homogeneous, Combe and Boxma [16] discuss the difficulty of the problem relative to the homogeneous case and analyze different solution approaches.

In the next sections we consider the routing problem when additional information is available at decision epochs. First we study the optimal routing when the queue lengths are known upon arrival, and then we discuss the problem when workload information is also available at decision epochs.

3.1 Queue-Length Routing

In this section we consider the optimal routing problem when the decision maker knows the queue lengths at DECing epochs and routes the flexible customer to one of the queues based on the available information. We assume that servers have exponentially distributed service times with the same rate μ, and arrivals to the system form an arbitrary process that is independent of the state of the system. Dedicated arrivals are equally likely to require a particular server. 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.

Routing flexible customers to the queue 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 all customers are flexible (i.e. $F = A$) has been shown in a variety of contexts (e.g., Winston [67], Weber [65], Ephremides, Varaiya and Walrand [20], Hordijk and Koole [36], Koole, Sparaggis, and Towsley [42], Sparaggis and Towsley [59] Sparaggis, Towsley and Cassandras [60], Towsley, Sparaggis and Cassandras [62], Movaghar [53], Johri [38], Bambos and Michailidis [13]). Whitt [66] gives a counterexample showing 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. [28] study JSQ in Processor Sharing (PS) server farms for nonexponential service times. When both dedicated and flexible arrivals are present, Menich and Serfozo [51] prove the optimality of JSQ for flexible customers among interchangeable routing policies when processing times are exponential and arrivals are Poisson. Argon et al. [11] consider a model with nonidentical exponential service rates and general delay costs and develop heuristic methods for routing. When a subset of customers are flexible and flexible customers follow JSQ, Foley and McDonald [21] study the large deviations (rare event) behavior of the system, and give 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.

In Section 3.1.1, we study multiple stations having multiple identical servers. Johri [38] and Sparaggis et al. [60] allow service rates to depend on queue lengths and show 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 sojourn time from Little's law. We also show that the sojourn time for dedicated customers is decreasing in the proportion flexible. That is, the monolingual customers, on average, benefit from having bilingual customers.

We also consider several practically important extensions. In Section 3.1.2 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 3.1.3 we consider finite buffers, and in Section 3.1.4 we consider several other extensions.

3.1.1 Multi-Server Stations

Sparragis et al. [60] 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 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 in [60], they do not allow dedicated customers, whereas we do. Furthermore, we use the same coupling of our proof here to show new results for impatient customers in the next section.

Let $N^1(t) = (N_1^1(t), N_2^1(t), ...N_c^1(t))$ denote the number of customers at each station when the join the shortest queue policy (JSQ) is followed for the flexible customers. By shortest queue, we mean the station with the fewest number of customers. Let $I^1(t) = (I_{11}^1(t), I_{12}^1(t),\ ..., I_{1m}^1(t), I_{21}^1(t)..., I_{cm}^1(t))$ 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 $Q^1(t) = ({Q_1}^1(t), {Q_2}^1(t)..., {Q_c}^1(t))$ 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 $N^2(t)$, $I^2(t)$, $Q^2(t)$ similarly, assuming 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 9. *$\{N^1(t)\}_{t=0}^{\infty}$ is stochastically smaller than $\{N^2(t)\}_{t=0}^{\infty}$ in the sense of weak submajorization:*

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

Proof. The proof uses coupling and forward induction. Suppose $\{\tilde{N}^1(t)\}$ 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

$$P(\{\tilde{N}^1(t)\}_{t=0}^{\infty} \prec_w \{\tilde{N}^2(t)\}_{t=0}^{\infty}) = 1. \tag{3.1}$$

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 $t_1 < t_2 < t_3 < ...$ and $t_0 = 0$. Clearly (3.1) 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 doesn't change for $t_n \leq t < t_{n+1}$, it is sufficient to show that (3.1) holds for t_n. We consider two cases separately.

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 $N^1(t_n) \prec_w N^2(t_n)$. 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 get $N^1(t_n) \prec_w N^2(t_n)$.

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 3.1. 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.

Table 3.1: Two coupled systems at time t with $c = 3, m = 3$.

1st System									2nd System								
$\mathbf{I_1^1}$	$\mathbf{I_2^1}$	$\mathbf{I_3^1}$	$\mathbf{I_4^1}$	$\mathbf{I_5^1}$	$\mathbf{I_6^1}$	$\mathbf{I_7^1}$	$\mathbf{I_8^1}$	$\mathbf{I_9^1}$	$\mathbf{I_1^2}$	$\mathbf{I_2^2}$	$\mathbf{I_3^2}$	$\mathbf{I_4^2}$	$\mathbf{I_5^2}$	$\mathbf{I_6^2}$	$\mathbf{I_7^2}$	$\mathbf{I_8^2}$	$\mathbf{I_9^2}$
1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0	0
$\mathbf{Q_1^1}$			$\mathbf{Q_2^1}$			$\mathbf{Q_3^1}$			$\mathbf{Q_1^2}$			$\mathbf{Q_2^2}$			$\mathbf{Q_3^2}$		
1			0			0			1			0			0		

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 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 3.2 for an ordering example.

There are four cases. If no actual service completion occurs in either system, or if

Table 3.2: Sample server states and related ordering

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

the potential service completion is an actual service completion in system 1 but not in 2, then majorization continues to hold trivially.

Now, suppose the potential service completion is an actual service completion in both systems. Suppose 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, $N^1(t_n) \prec_w N^2(t_n)$. Now suppose $u > v$ (see Table 3.3), 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 definition of p, u and v, stations up to r have more empty servers in system 1 for $v \leq r < u$, that is,

$$\sum_{k=1}^{r*m} I_k^1(t) < p \leq \sum_{k=1}^{v*m} I_k^2(t) \leq \sum_{k=1}^{r*m} I_k^2(t). \tag{3.2}$$

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 3.3). Then by (3.2) 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 4, we have $N^1(t_n) = N^1(t) - e_u \prec_w N^2(t) - e_v = N^2(t_n)$. Now, suppose the potential service completion is an actual service completion

Table 3.3: 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.

t

1st System									2nd System								
$\mathbf{I_1^1}$	$\mathbf{I_2^1}$	$\mathbf{I_3^1}$	$\mathbf{I_4^1}$	$\mathbf{I_5^1}$	$\mathbf{I_6^1}$	$\mathbf{I_7^1}$	$\mathbf{I_8^1}$	$\mathbf{I_9^1}$	$\mathbf{I_1^2}$	$\mathbf{I_2^2}$	$\mathbf{I_3^2}$	$\mathbf{I_4^2}$	$\mathbf{I_5^2}$	$\mathbf{I_6^2}$	$\mathbf{I_7^2}$	$\mathbf{I_8^2}$	$\mathbf{I_9^2}$
1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0	0
	$\mathbf{Q_1^1}$			$\mathbf{Q_2^1}$			$\mathbf{Q_3^1}$			$\mathbf{Q_1^2}$			$\mathbf{Q_2^2}$			$\mathbf{Q_3^2}$	
	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

(**a**)

$\mathbf{t_n}$

1st System									2nd System								
$\mathbf{I_1^1}$	$\mathbf{I_2^1}$	$\mathbf{I_3^1}$	$\mathbf{I_4^1}$	$\mathbf{I_5^1}$	$\mathbf{I_6^1}$	$\mathbf{I_7^1}$	$\mathbf{I_8^1}$	$\mathbf{I_9^1}$	$\mathbf{I_1^2}$	$\mathbf{I_2^2}$	$\mathbf{I_3^2}$	$\mathbf{I_4^2}$	$\mathbf{I_5^2}$	$\mathbf{I_6^2}$	$\mathbf{I_7^2}$	$\mathbf{I_8^2}$	$\mathbf{I_9^2}$
1	1	1	1	1	0	0	1	0	1	1	1	1	1	0	1	0	0
	$\mathbf{Q_1^1}$			$\mathbf{Q_2^1}$			$\mathbf{Q_3^1}$			$\mathbf{Q_1^2}$			$\mathbf{Q_2^2}$			$\mathbf{Q_3^2}$	
	2			0			0			2			0			0	

(**b**)

in system 2, but not in system 1. See Table 3.4. Define v as the station where the actual service 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

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

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,

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

$$= \sum_{k=1}^{v} N_k^2(t) \leq \sum_{k=1}^{v'} N_k^2(t).$$

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. ∎

Table 3.4: 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.

t

1st System									2nd System								
$\mathbf{I_1^1}$	$\mathbf{I_2^1}$	$\mathbf{I_3^1}$	$\mathbf{I_4^1}$	$\mathbf{I_5^1}$	$\mathbf{I_6^1}$	$\mathbf{I_7^1}$	$\mathbf{I_8^1}$	$\mathbf{I_9^1}$	$\mathbf{I_1^2}$	$\mathbf{I_2^2}$	$\mathbf{I_3^2}$	$\mathbf{I_4^2}$	$\mathbf{I_5^2}$	$\mathbf{I_6^2}$	$\mathbf{I_7^2}$	$\mathbf{I_8^2}$	$\mathbf{I_9^2}$
1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0
	$\mathbf{Q_1^1}$			$\mathbf{Q_2^1}$			$\mathbf{Q_3^1}$			$\mathbf{Q_1^2}$			$\mathbf{Q_2^2}$			$\mathbf{Q_3^2}$	
	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

(**a**)

$\mathbf{t_n}$

1st System									2nd System								
$\mathbf{I_1^1}$	$\mathbf{I_2^1}$	$\mathbf{I_3^1}$	$\mathbf{I_4^1}$	$\mathbf{I_5^1}$	$\mathbf{I_6^1}$	$\mathbf{I_7^1}$	$\mathbf{I_8^1}$	$\mathbf{I_9^1}$	$\mathbf{I_1^2}$	$\mathbf{I_2^2}$	$\mathbf{I_3^2}$	$\mathbf{I_4^2}$	$\mathbf{I_5^2}$	$\mathbf{I_6^2}$	$\mathbf{I_7^2}$	$\mathbf{I_8^2}$	$\mathbf{I_9^2}$
1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0	0
	$\mathbf{Q_1^1}$			$\mathbf{Q_2^1}$			$\mathbf{Q_3^1}$			$\mathbf{Q_1^2}$			$\mathbf{Q_2^2}$			$\mathbf{Q_3^2}$	
	2			0			0			2			0			0	

(**b**)

3.1.1.1 Corollaries

The weak majorization result also holds for the customers waiting in the queue, that is, $\{Q^1(t)\}_{t=0}^{\infty} \prec_w \{Q^2(t)\}_{t=0}^{\infty}$. 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 $\overline{N}^i(t) = \sum_{k=1}^{c} {N_k}^i(t),\ i = 1, 2$, we have $\{\overline{N}^1(t)\}_{t=0}^{\infty} \prec_{st} \{\overline{N}^2(t)\}_{t=0}^{\infty}$ and $E[\overline{N}^1(t)] \leq E[\overline{N}^2(t)], t \geq 0$. 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 $\{D^1(t)\}_{t=0}^{\infty} \succ_{st} \{D^2(t)\}_{t=0}^{\infty}$.

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 9, since we can construct an arbitrary routing 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 sojourn time in system for all customers and the steady-state mean sojourn 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.

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 $\{N^1\}_{t=0}^{\infty} \prec_w \{N^2\}_{t=0}^{\infty}$ and both the steady-state mean sojourn time for all customers and steady-state sojourn time for dedicated customers are smaller in N^1 for this setting.

3.1.2 Impatient customers

We suppose 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 $N^1(t) = ({N_1}^1(t), {N_2}^1(t)..., {N_c}^1(t))$ as the vector of queue lengths at time $t \geq 0$ when JSQ is followed for flexible customers and $N^2(t)$ as the vector of queue lengths when an arbitrary policy is followed for flexible customers. Also let $A^1(t)$ denote the total number abandonments and $D^1(t)$ the total number of service completions up to time t when JSQ is followed for flexible customers. Similarly define $A^2(t)$ and $D^2(t)$ for the arbitrary routing policy.

Sparaggis et al. [60] showed the following result.

- For non-decreasing concave service rates $\{N^1(t)\}_{t=0}^{\infty} \prec_w \{N^2(t)\}_{t=0}^{\infty}$, and
- For non-decreasing convex service rates $\{N^1(t)\}_{t=0}^{\infty} \prec^w \{N^2(t)\}_{t=0}^{\infty}$.

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 non-decreasing concave service rates, whereas customers abandoning only in queue with rate $\alpha > \mu$ is a special case of non-decreasing convex service rates. However the above result does not tell much about the overall system performance, because stochastically minimizing the number in system doesn't mean maximizing service completions as the departures are due to both service completions and abandonments.

Movaghar [53] also studies the JSQ policy with impatient customers. He assumes Poisson arrivals, and gives conditions on the distribution of patience time such that JSQ minimizes the number of long-run abandonments.

The following result was shown to be true without dedicated arrivals in [59]. We prove it using a new coupling procedure that we also use for the proof of Theorem 11.

Corollary 10. *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,*

$$\{N^1(t)\}_{t=0}^{\infty} \prec_w \{N^2(t)\}_{t=0}^{\infty} \tag{3.3}$$

$$\{A^1(t)\}_{t=0}^{\infty} \prec_{st} \{A^2(t)\}_{t=0}^{\infty} \tag{3.4}$$

$$\{D^1(t)\}_{t=0}^{\infty} \succ_{st} \{D^2(t)\}_{t=0}^{\infty} \tag{3.5}$$

Proof. Again we use induction on event times t_n so suppose (3.3)-(3.5) 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 $\frac{\alpha}{\mu}$ and type 2 with probability $1 - \frac{\alpha}{\mu}$. We couple the arrivals and potential service completions of type 2 as in the proof of Theorem 9. Since we have single-server stations, potential service 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 together as follows. In broad terms, we first couple each abandonment in system 1 with an abandonment in system 2, which we can do from Section 3.1.1.1, 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 $1, 2, \ldots, Q_1^1(t)$ are the customers in the first (largest) queue, then customers $Q_1^1(t) + 1, \ldots, Q_1^2(t)$ are in the second queue, etc. We then label the customers in service starting from station 1. See the first part of Table 3.5 for a 3-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 3.5. Suppose customer p under the labeling above departs

Table 3.5: Labeling of customers, where a 1 indicates the presence of a customer.

1st System			2nd System		
$\mathbf{I_1^1}$	$\mathbf{I_2^1}$	$\mathbf{I_3^1}$	$\mathbf{I_1^2}$	$\mathbf{I_2^2}$	$\mathbf{I_3^2}$
1	1	1	1	1	
$\mathbf{Q_1^1}$	$\mathbf{Q_2^1}$	$\mathbf{Q_3^1}$	$\mathbf{Q_1^2}$	$\mathbf{Q_2^2}$	$\mathbf{Q_3^2}$
1 1	1 1		1 1 1	1 1 1	
Labeling of Customers			**Labeling of Customers**		
5	6	7	5	6	
1 2	3 4		1 2 3	4 7 8	

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 3.5). Let v be such that this departure takes place in vth largest queue in system 2, so $\sum_{k=1}^{v} N_k^2(t) \ge p$. Then to account for 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 $N^1(t_n) \prec_w N^2(t_n)$.

(ii) There is an actual abandonment in both systems (i.e., $p \le Q^1$; for example, $p = 2$ in Table 3.5). Let $u(v)$ be such that the abandonment takes place in the uth(vth) largest queue in system 1(system 2). If $u \le v$, then by Lemma 1, $N^1(t_n) \prec_w N^2(t_n)$. If $u > v$, then by definition of p,

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

for all $v \le r < u$. Therefore, by Corollary 4, $N^1(t_n) \prec_w N^2(t_n)$.

(iii) There is an actual service completion of type 1 in both systems (i.e., $Q^1 < p \le Q^1 + \max\{M_1, M_2\}$, 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 3.5). Let $u(v)$ be such that the service completion takes place in the uth(vth) largest queue in system 1(system 2). Then, $u = v$ because $Q^1 + u = p = Q^1 + v$. Therefore, by Corollary 4, $N^1(t_n) \prec_w N^2(t_n)$.

(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 $Q^1 + M_2 < p \le Q^1 + M_1$; for example, $p = 7$ in Table 3.5). 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 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 4, $N^1(t_n) \prec_w N^2(t_n)$.

Note that $A^1(t)$ only changes 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 (3.4) follows by induction. And similar reasoning as in the previous proof shows (3.5). ∎

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 higher rate than the service rate, to minimize the number in the system it is better to have long queues, and therefore more abandonments.

Theorem 11. *Let each customer in queue abandon the system after waiting an exponentially-distributed time at rate α, where $\alpha > \mu$. Then,*

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

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

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

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 $\frac{\mu}{\alpha}$ and type 2 with probability $1 - \frac{\mu}{\alpha}$. Next we couple the arrivals as in the proof of Theorem 9. In these cases the weak supermajorization will be preserved by Lemma 1. Next we couple potential service completions and potential abandonments of type 1 together. 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 $N^1(t) \prec^w N^2(t)$, 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 (non-actual) departures in system 2. To rigorously define these customers, let V be an integer valued c dimensional vector and 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 5, $N^1(t) \prec^w V^d(t) \prec^w N^2(t)$ and $\sum_{k=1}^c V_k^d(t) = \sum_{k=1}^c N_k^2(t)$. Hence, $V^d(t) \prec_w N^2(t)$. 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 $N^1(t_n) \prec^w N^2(t_n)$. Let this customer be in the kth largest queue in system 1. Then it is immediate that $N^1(t_n) = N^1(t) - e_k \prec^w V^d(t) \prec^w N^2(t) = N^2(t_n)$. Therefore $N^1(t_n) \prec^w N^2(t_n)$. 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 10 , which we can do, because $V^d(t) \prec_w N^2(t)$, so we get $V^d(t_n) \prec_w N^2(t_n)$ 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 $N^1(t_n) \prec^w V^d(t_n) \prec^w N^2(t_n)$.

Finally we will look at the potential abandonments of type 2. In this case, number of service completions will not change, satisfying $D^1(t_n) \geq D^2(t_n)$. The coupling procedure is similar to Theorem 9. 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 abandonment if p exceeds the number of customers). See Table 3.6. Again let $I_j^i = \mathbf{1}\{N_j^i(t) > 0\}$ represent the customers

Table 3.6: Labeling of customers for abandonments of type 2.

1st System				2nd System			
	$\mathbf{I_1^1}$	$\mathbf{I_2^1}$	$\mathbf{I_3^1}$		$\mathbf{I_1^2}$	$\mathbf{I_2^2}$	$\mathbf{I_3^2}$
	1	1	1		1	1	1
	$\mathbf{Q_1^1}$	$\mathbf{Q_2^1}$	$\mathbf{Q_3^1}$		$\mathbf{Q_1^2}$	$\mathbf{Q_2^2}$	$\mathbf{Q_3^2}$
1	1	1		1	1		
Labeling of Customers				**Labeling of Customers**			
3	2	1		2	1		

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 $N^1(t) \prec^w N^2(t)$. 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(vth) largest queue in system 1 (system 2). If $u \leq v$, then by Lemma 1, $N^1(t_n) \prec^w N^2(t_n)$. If $u > v$, (e.g., $p = 1$ in Table 3.6 so $u = 2$, $v = 1$), then by 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 4, $N^1(t_n) \prec^w N^2(t_n)$.

(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 3.6). To account for 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 definition of p, $\sum_{k=u}^{c} Q_k^1(t) \geq p$ and $\sum_{k=1}^{c} Q_k^2(t) < p$. 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 $\sum_{k=r}^{c} N^1(t_n) < \sum_{k=r}^{c} N^2(t_n)$, i.e.,

$$\sum_{k=r}^{c} N_k^1(t) = \sum_{k=r}^{c} N_k^2(t). \tag{3.6}$$

By induction we have $\sum_{k=r+1}^{c} N_k^1(t) \geq \sum_{k=r+1}^{c} N_k^2(t)$. Combining this with (3.6) yields $Q_r^1(t) \leq Q_r^2(t)$. On the other hand, $\sum_{k=r}^{c} I_k^1(t) \geq \sum_{k=r}^{c} I_k^2(t)$. This

with (3.6) gives $\sum_{k=r}^{c} Q_k^1(t) \leq \sum_{k=r}^{c} Q_k^2(t)$. Therefore,

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

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

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

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

3.1.3 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 4 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, $\{D(t)\}_{t=0}^{\infty}$ (Section 3.1.1.1), 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 [36] and Sparaggis et al. [60]. They showed that JSQ (which now means routing customers to the shortest *nonfull* queue) stochastically maximizes $D(t)$ for all $t \geq 0$. Koole, Sparaggis, and Towsley

[42] 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 13 below proves the lemma; we omit the proof.

Lemma 12. *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 13. *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 routing 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 routing, 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.

Now assume that A is not full before routing 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 on, 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 12, 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 don't 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 similarly define $N^2(t)$ for an arbitrary routing policy. Suppose $N^1(t) = (3, 2)$ 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, $\{N^1(t)\}_{t=0}^{\infty} \prec_w \{N^2(t)\}_{t=0}^{\infty}$. This result can easily be shown again using forward induction and considering the cases where an arrival might be lost due to capacity insufficiency.

3.1.4 Other Extensions

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

Slotted Service: Suppose 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 routing policy with JSQ so that, if there is a potential service completion in the kth largest queue in one system at time n, $n \in \{0, 1, 2, ...\}$, the same is true for the other system. Notice 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.

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 routing policy. For example, servers could go on- and off-line 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.

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.

Power of 2 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 the facility is multi-lingual, but customers are at most bilingual, and all combinations of bilingualism are equally likely. Mitzenmacher [52] showed that JSQ among 2 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 (see [8] for a detailed discussion).

Resequencing: Suppose 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 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 reader to [3], [27] and [43].

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 Section 3.1.1.1 we have 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 3.1.1.1, 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 $\hat{E}_i = \max_{j=1,\ldots,i} \hat{C}_j$. 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, $j = 1, \ldots, i$ 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 $j = 1, \ldots, i$. From Remark 3.1.1.1, $\tilde{D}_i$ is stochastically minimized by JSQ, but $\tilde{D}_i = \max_{j=1,\ldots,i} \tilde{C}_j = \max_{j=1,\ldots,i} \hat{C}_j = \hat{E}_i$, so $\hat{E}_i$ is also stochastically minimized by JSQ for any i.

3.2 Workload Routing

In this section we study the routing problem when the actual workload at each queue is known upon arrival but the required work of the arriving customer is unknown. We assume that the arrivals to the system form an arbitrary process that is independent of the state of the system, and some arbitrary subset of these arrivals are flexible. Dedicated arrivals are equally likely to require a particular server. The service times are independent and identically distributed with a distribution function $G(\cdot)$ and are independent of the inter-arrival times.

Routing flexible customers to the queue with the shortest workload is known in the literature as the "Join the Shortest Work" (JSW) policy. Note that when all customers are flexible, the "Join the Shortest Work (JSW)" policy is equivalent to the "First-Come-First-Served (FCFS)" policy with a single queue for all customers. Wolff [68, 69] shows 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 [17], building on Foss's work [23], shows that JSW stochastically minimizes the workload at each time t for general arrival and service processes using weak submajorization arguments. See also Liu et al. [46]. Koole [41] shows the same result using dynamic programming arguments. Stoyan [61] and Koole [41] provide counterexamples showing that the pathwise optimality (jointly across t) of JSW for minimizing the workload is not true. However Koole [40] shows 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 routing policies, the JSW policy stochastically maximizes the departure process pathwise, and stochastically minimizes the workload at each time t.

When the required work of the arriving customer is also known upon arrival the JSW policy is not necessarily optimal. Harchol-Balter et al. [30] propose a routing policy called "Size Interval Task Assignment with Equal-Load (SITA-E)" where different servers are assigned to jobs with service times falling into particular intervals, and compare the performance of their policy with the performance of the "Join the Shortest Work (JSW)" policy under different problem parameters. SITA-E outperforms JSW under high task size variance. Similar policies such as EquiLoad and AdaptLoad are suggested in the work of Ciardo, Riska, and Smirni [15] and Riska, Sun, Smirni and Ciardo [56]. Hyytia et al. [37] discuss the computation of value functions based on different levels of information.

The JSW policy can easily be implemented, even when workload is not known (e.g. in our call center example of Section 1.1). Upon arrival suppose 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 now show that 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 (JSW). We also show that the departures are pathwise maximized by the JSW policy. These results are extensions of Daley [17] and Koole [40] to systems with both dedicated and flexible arrivals.

Consider two parallel server systems where V_n^i denotes the workload vector at the instant of the nth arrival in the ith system for $i = 1, 2$ (V_0^i denotes the initial workload at time $t = 0$), and an arbitrary routing 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 nth and $n-1$st customers, and s_n is the service time of nth customer. Let $\{D^i(t)\}$ denote the departure process in the ith system for $i = 1, 2$. Also, suppose the initial customers leave earlier in system 1. Then we have the following corollary.

Corollary 14. *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) $\{D^1(t)\}_{t=0}^{\infty} \geq \{D^2(t)\}_{t=0}^{\infty}$ *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 the kth largest workload in system 2, then system 1 routes it to the queue with the kth largest workload as well. A dedicated arrival to the queue with the kth largest workload in system 1 is also a dedicated arrival to the queue with the kth largest workload in system 2 with the same service time in both systems. Then (i) directly follows from Lemma 8. 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 some arbitrary queue u. We will construct a policy $\tilde{\Pi}$ that routes customer 1 to the queue with the shortest workload, queue v, and that has stochastically smaller workload and pathwise earlier departures than policy Π. Let $V_n = (V_{n1}, V_{n2}, ..., V_{nc})$ and $\tilde{V}_n = (\tilde{V}_{n1}, \tilde{V}_{n2}, ..., \tilde{V}_{nc})$ denote the workload vector at the instant of 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 15. *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 iid. Then*

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

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

Proof. We will first show (i). For $n = 1$, we couple the service time of customer 1 under both policies and the result follows by Lemma 8. Consider $n = 2$. If the second arrival, customer 2, is a dedicated arrival to one of the queues that is different than u or v, say queue r, under Π, we let customer 2 to be a dedicated arrival to the queue r under $\tilde{\Pi}$ as well and we couple service time of customer 1 (customer 2) under Π with the service time of customer 1 (customer 2) and it can easily be seen that $\tilde{V}_2 \leq V_2$.

For other cases for customer 2, we do a sample-path-dependent coupling as follows. If customer 2 is a flexible arrival routed to queue u (queue v) under Π, then we let $\tilde{\Pi}$ route customer 2 to queue v (queue u). If customer 2 is a dedicated arrival to queue u (queue v) under Π, we let customer 2 to be a dedicated arrival to queue v (queue u) under $\tilde{\Pi}$. Let τ be the interarrival time between the first and second arrival. If $V_{0v} > \tau$ then we do a direct coupling of customer service times, i.e., we couple the service time of customer 1 (customer 2) under Π with the service time of customer 1 (customer 2) under $\tilde{\Pi}$. On the other hand if $V_{0v} < \tau$, then we do a cross coupling for service times, i.e., we couple the service time of customer 1 (customer 2) under Π with the service time of customer 2 (customer 1) under $\tilde{\Pi}$. In both cases it can easily be seen that $\tilde{V}_2 \leq V_2$.

Finally for $n > 2$, the result follows by Corollary 14.

Now to prove (ii) we use the same coupling described as above for $n \geq 2$. Because the systems are identical at $t = 0$, the initial customers leave at the same time under $\tilde{\Pi}$ as under Π. Also, customers 1 and 2 leaves earlier under $\tilde{\Pi}$ in all of the above cases. And for the arrivals after second arrival, $n > 2$, the result follows by Corollary 14. ∎

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 $n \geq 2$). Therefore the coupling depends

on time, and the result is stochastic optimality for each t, but not jointly for all t. See Koole [40] 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 time 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

$$W_n^d =_{st} \frac{\sum_{k=1}^{c} (V_{m-1k} - \tau_m)^+}{c} + S_n,$$

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. Hence, we have the following corollary.

Corollary 16.

$$\tilde{W}_n^d \prec_{st} W_n^d, \text{ for all } n \geq 1$$

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 $\min_k\{(V_{m-1k} - \tau_m)^+\}$ 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. That is, JSW is individually optimal for the flexible customers.

Chapter 4

Optimal Scheduling

In this chapter we study the scheduling problem in a multi-server multi-queue system with partial flexibility. Each server has a queue of dedicated customers that can only be served by that server, and there is a separate queue for flexible customers that can be served by more than one server. The decision maker should assign a job from the queues (if there are any jobs waiting) to an idle server upon arrival or service completion. This problem can be modeled as the W_s design of Section 1.2.

A simpler model, the N_s design, where only one server has dedicated customers, and there is a queue for flexible customers, has been studied widely on 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 [25] show the difficulty of finding the optimal control policy by considering different examples with different parameters. Harrison [31] constructs a discrete review control policy in which the heavy traffic limit approaches the bound of a single pooled resource. Bell and Williams [14] show the optimality of a threshold-type control policy in heavy traffic. Ahn et al. [4] consider a clearing system (without arrivals). They show 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 [19] consider the problem with arrivals and with an upgrading option for low priority customers, and give conditions under which the $c\mu$ rule is optimal.

The W_s design is even more complex than the N_s design and the problem has usually been studied using heavy traffic approximations or using heuristic control policies. Harrison and Lopez [32] give 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_s" design is a special case. Mandelbaum and Stolyar [47] also consider a general structure with arbitrary customer and server flexibility and show the optimality of the generalized $c\mu$ rule in heavy traffic. Gurumurthi and Benjaafar [29] analyze control policies under arbitrary customer and server flexibility, and they compare the performance of different control polices such as serve the Longest Queue First (LQF) and Strict Priority (SP). Saghafian et al. [57] study the "W_s" network with Poisson arrivals, exponential service times, Poisson service disruptions and with preemption permitted. They propose a control policy called "Largest Expected Workload Cost (LEWC)," and they compare its performance with the performance of the $c\mu$ rule, the generalized $c\mu$ rule and the LQF policy. They also give 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 20 below where we consider a very general arrival process.

We show that in many situations the dedicated customers first (DCF) policy is optimal. Under DCF, whenever a server's dedicated queue is non-empty it gives priority to dedicated customers *and* does not idle. Note that flexible customers tend to be disadvantaged under DCF. DCF is also more efficient than the two routing designs discussed in Chapter 3, JSQ and JSW, in terms of minimizing the overall customer sojourn 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, because 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 "0" for "bilingual", they would press "1" for "English"). As we discussed earlier, 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 sojourn times. Also, empirically, the overall performance of JSW is very close to that of DCF (see Figure 4.1). Moreover, JSW is incentive compatible for flexible customers in the sense that it is their individually optimal policy, i.e., 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 chapter is as follows. In Section 4.1 we study the "W_s" design scheduling problem. In Section 4.2, we compare the performance of scheduling versus routing designs.

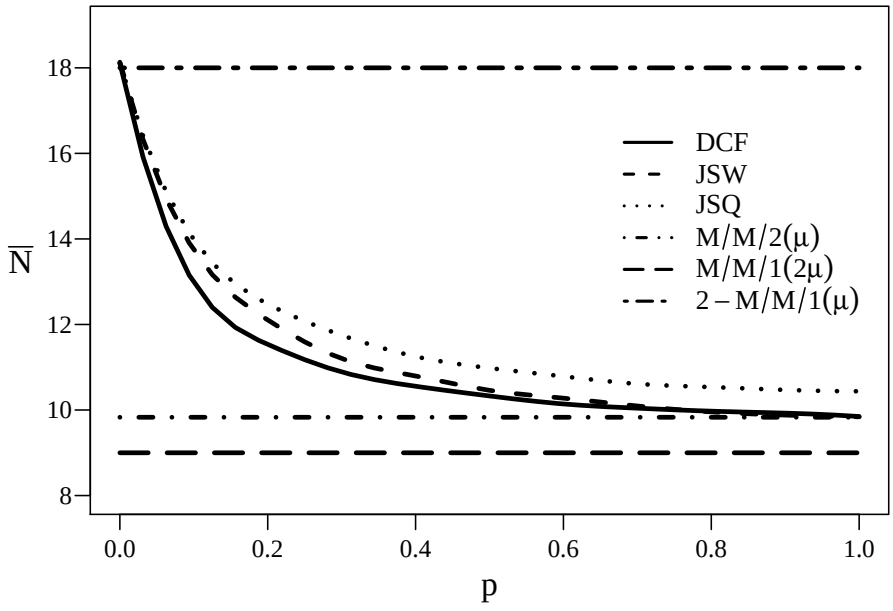

Figure 4.1: Comparison of policies. Total number in system ($\bar{N}$) vs proportion of flexible customers (p).

4.1 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 non-preemptive, work conserving, non-anticipating service disciplines, minimizes the stationary sojourn time in the convex ordering sense (Wolff [70]). See Righter and Shantikumar [55] and Aalto, Ayesta and Righter [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 non-anticipating 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 multi-server stations.

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. Table 4.1 summarizes our results 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 4.1.

Table 4.1: Summary of results.

Model				Objective	Optimal policy
Preemption Permitted	Idling Permitted	Arrivals	Services		
Yes	Yes (or No)	general	exponential (diff. 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

4.1.1 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. Roughly, 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). 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

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

and where upon arrival of customer k, $U_k^c = 0$. Suppose 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, 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 17. *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 process for these systems so that each customer has the same arrival epoch in both systems. We also couple the service times (the U_k's) 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 18. *Suppose preemption and idling are both permitted and 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), with $\mu_A < \mu_B$ (so the service times for dedicated customers are $\frac{1}{\mu_A}$ and $\frac{1}{\mu_B}$ respectively). Suppose at time $t = 0$ there is one job on server B which will finish at $t + \epsilon$, one job waiting in the flexible queue, one job waiting in server A's dedicated queue, server A is idle, and 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 $\epsilon + \frac{1-\mu_A\epsilon}{\mu_B}$ under Π, which is earlier than either departure under DCF (at $\frac{1}{\mu_A}$ and $\epsilon + \frac{1}{\mu_B}$).*

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 19. *Let all service times be iid with the distribution of S where*

$$S = \begin{cases} 1 & w.p.\ 0.5 \\ 101 & w.p.\ 0.5 \end{cases}$$

Suppose 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 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 sojourn 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 doesn't 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 20. *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 dedicated customers first (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 process 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 17.

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. ∎

4.1.2 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.

4.1.2.1 Idling is permitted

We first permit idling and show that DCF is optimal for general service times and arrival processes. In Section 4.1.1 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 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 21. *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 dedicated customers first (DCF) policy.*

Proof. The first part of the proof, that dedicated customers should have priority, is the same as the first (non-preemptive) case in the proof of Theorem 20. To show that a server with dedicated customers shouldn't 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 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 21, 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 22. *Suppose preemption is not permitted and 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 the total arrival rate, λ, is such that $\lambda < 2\mu$.

We have the following lemma which directly follows from Theorem 21.

Lemma 23. *For homogeneous exponential servers and 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 at time $t = 0$ server A's dedicated queue is empty and 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 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 $\{D(t)\}_{t=0}^{\infty}$, $\{\tilde{D}(t)\}_{t=0}^{\infty}$ 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 $\{\tilde{D}^p(t)\}_{t=0}^{\infty}$, $\{\tilde{D}_f^p(t)\}_{t=0}^{\infty}$, $\{\tilde{D}_d^p(t)\}_{t=0}^{\infty}$ 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 $\{D^p(t)\}_{t=0}^{\infty}$, $\{D_f^p(t)\}_{t=0}^{\infty}$, $\{D_d^p(t)\}_{t=0}^{\infty}$ 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 24. *(i)* $\{D(t)\}_{t=0}^{\infty} =_{st} \{D^p(t)\}_{t=0}^{\infty}$, $\{\tilde{D}(t)\}_{t=0}^{\infty} =_{st} \{\tilde{D}^p(t)\}_{t=0}^{\infty}$

(ii) $\{D_d^p(t)\}_{t=0}^{\infty} =_{st} \{\tilde{D}_d^p(t)\}_{t=0}^{\infty}$

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 25. *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 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 4.2. ∎

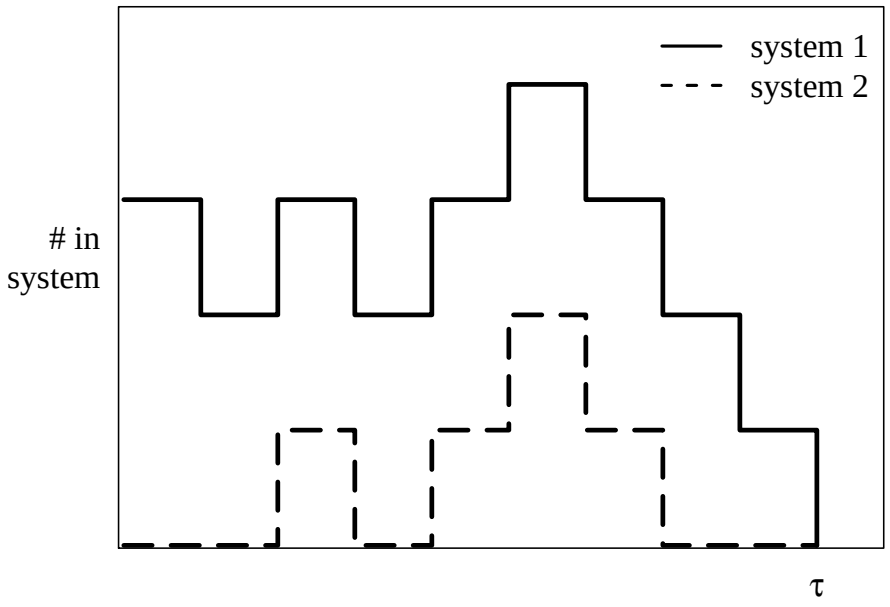

Figure 4.2: Coupling of the two systems.

We are ready to prove the following result.

Lemma 26. $\{D_f^p(t)\}_{t=0}^{\infty} \prec_{st} \{\tilde{D}_f^p(t)\}_{t=0}^{\infty}$

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.

(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 25. 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.

(ii) $S_f > t_1$: In this case we consider following sub cases.

(a) 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 25 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.

(b) 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 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 25. 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 25. 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 23-26, 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 27. *For homogeneous exponential servers and Poisson arrivals, a nonidling dedicated customers first (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 24 and 26 do not imply the sample-path result,

$$\{D^p(t)\}_{t=0}^{\infty} \prec_{st} \{\tilde{D}^p(t)\}_{t=0}^{\infty},$$

because we don't have the joint sample-path result,

$$\{D_d^p(t), D_f^p(t)\}_{t=0}^{\infty} \prec_{st} \{\tilde{D}_d^p(t), \tilde{D}_f^p(t)\}_{t=0}^{\infty},$$

which indeed is not true by Example 22.

4.1.2.2 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 28. *Suppose neither idling nor preemption are permitted and consider the setting of Example 18. 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* $\frac{1}{\mu_A}$. *Hence if* $\frac{2}{\mu_A} < t + \epsilon + \frac{1}{\mu_B}$, *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 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 27. Here we will extend the result to sample-pathwise optimality using weak submajorization.

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 the 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 $\bar{N}(t) = N_A(t) + N_f(t) + N_B(t)$ 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 $(N_A(0), N_f(0), N_B(0)) = \big(N'_A(0), N'_f(0), N'_B(0)\big)$. 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 29. *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) $\{N'(t)\}_{t=0}^{\infty} \prec_w \{N(t)\}_{t=0}^{\infty}$

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

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

$$P(\{\tilde{N}'(t)\}_{t=0}^{\infty} \prec_w \{\tilde{N}(t)\}_{t=0}^{\infty}) = 1 \tag{4.1}$$

$$P(\{\tilde{\bar{N}}'(t)\}_{t=0}^{\infty} \leq \{\tilde{\bar{N}}(t)\}_{t=0}^{\infty}) = 1 \tag{4.2}$$

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 $t_1 < t_2 < t_3 < ...$, and $t_0 = 0$. Clearly (4.1) and (4.2) hold for $t = 0$ because $N(0) = N'(0)$. Assume that they also hold for t such that $t_{n-1} \leq t < t_n$. Then, because the state doesn't change for $t_n \leq t < t_{n+1}$, it is sufficient to show that (4.1) and (4.2) hold for t_n.

If t_n is an arrival time then (4.2) is trivially true. If the arrival is a dedicated arrival, (4.1) 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 $N'(t_n) = N'(t) \leq N(t) \leq N(t_n)$.

(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 (4.1) follows.

(ii) $N'_{[1]}(t) > 0$, $N'_{[2]}(t) = 0$: Then the flexible arrival is routed to the idle server, $N'_{[1]}(t_n) = N'_{[1]}(t)$ and $N'_{[2]}(t_n) = 1$. Either $1 \leq N_{[2]}(t) = N_{[2]}(t_n)$ (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 (4.1) follows.

Next, suppose t_n is a potential service completion. We first show (4.1). Note that $N'(t_n) \leq N'(t)$ and $N(t_n) \leq N(t)$ but we may have equality even if the potential service completion is an actual service completion. Consider the following cases:

(i) $N'_f(t) = N_f(t) = 0$: Then the result follows from Corollary 3 with $M = 0$, i.e., the potential service completion is not an actual completion if the server is idle.

(ii) $N'_f(t) > 0$, $N_f(t) > 0$: 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 (4.1) follows. On the other hand, if the arbitrary policy agrees with DCF at t_n, then the result will follow from Corollary 3 with $M = 1$, i.e., the queue length where the service completion occurred will decrease by one 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, (4.1) follows.

(iii) $N'_f(t) > 0$, $N_f(t) = 0$: First suppose 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 Corollary 3 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 (4.2) and the fact that $N_f(t) = 0$, $N_{[1]}(t) \geq 2$ and the result follows. Secondly, suppose the potential service completion is from the shortest queue in both systems. If $N'_{[2]}(t) > 1$, then again the result follows from Corollary 3 with $M = 0$. If $N'_{[2]}(t) = 1$, then $N'_{[2]}(t_n) = 1, N'_f(t_n) = N'_f(t) - 1$ 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.

(iv) $N'_f(t) = 0$, $N_f(t) > 0$: 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, so (4.1) 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

$$N'(t_n) = (N'(t) - e_i)^+ \prec_w (N(t) - e_i)^+ \leq max\{(N(t) - e_i), 1\} = N(t_n)$$

where the majorization inequality follows from Corollary 3 with $M = 0$. Hence (4.1) follows.

Now to show (4.2), 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 $N'_{[1]}(t) = 0$. 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.

4.2 Comparison of policies

Let $\{D^{JSQ}(t)\}_{t=0}^{\infty}$ ($\{D^{JSW}(t)\}_{t=0}^{\infty}$) 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 $\{D^{DCF}(t)\}_{t=0}^{\infty}$ denote the departure process when there is a separate queue for flexible customers to wait, DCF is followed (might be idling), and preemption 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 routing, which is better than JSQ. The following theorem follows directly from Theorem 15 of Section 3, since JSQ becomes the arbitrary policy described in the proof.

Theorem 30. *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 independent and identically distributed with a distribution function $G(\cdot)$ and be independent of the inter-arrival times. Then*

$$\{D^{JSW}(t)\}_{t=0}^{\infty} \prec_{st} \{D^{JSQ}(t)\}_{t=0}^{\infty}.$$

The following theorem follows from Theorem 21, 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 31. *Let the arrival process be arbitrary, and let server j have arbitrary service process $\{S_k^j\}_1^\infty$, $j = A, B$. Suppose the actual workload at each queue is known upon arrival and the service time of the arriving customer is not known. We have*

- $\{D^{DCF}(t)\}_{t=0}^\infty \prec_{st} \{D^{JSQ}(t)\}_{t=0}^\infty$
- $\{D^{DCF}(t)\}_{t=0}^\infty \prec_{st} \{D^{JSW}(t)\}_{t=0}^\infty$

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.

4.3 Extensions

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

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.

Multi-server Stations: Suppose we have multiple service stations with their own queues, but each station has more than one servers. Then the results will still be true, however the proof of Theorem 27 requires a significant modification because the "up" times depend on the overall server state of the stations. Also, the proof of Theorem 29 requires a careful coupling for the potential service completions (see Chapter 3 for the details).

Impatience: Suppose 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 chapter will still be true.

Chapter 5

Marginal Impact of Customer Flexibility

In Chapter 3 we showed for the W_r design that, when the only information available upon arrival is the queue length, routing flexible customers to the shortest queue, JSQ, is optimal in a very strong sense; it minimizes the queue-length vector process in a sample-path weak majorization sense. A consequence of this result is that the stationary waiting time is stochastically decreasing in the proportion of flexible customers, p. Now we are interested in the marginal impact of customer flexibility, so we consider the convexity of waiting time in p. Convexity means that the marginal advantage of flexibility is largest at small proportions. That is, roughly, "a little bit of flexibility goes a long way." Unfortunately, it is not possible to obtain convexity in the strong sense for which monotonicity holds. Instead, we show a weaker form of convexity, that the stationary mean waiting time is convex in the proportion of flexible customers.

Although convexity in p is intuitive, it does not hold in the strong sense that monotonicity holds, and it is surprisingly difficult to prove. We develop a new approach that combines marginal analysis with coupling to show convexity in the stationary mean waiting time. We consider a tagged customer in steady-state that has lowest preemptive priority relative to the other customers so that the other customers are unaffected by the tagged customer. We show that the derivative of the stationary waiting time with respect to p (the marginal value of customer flexibility) can be expressed in terms of the difference in expected waiting time between going to the long and the short queue for the tagged customer. We then show, using another coupling argument, that this difference is decreasing in p.

We also consider a high-production Inventory Model, where we obtain a sample-

pathwise convexity result using majorization of the queue lengths. In this model, there are two servers and they never idle but instead build up inventory when no customers are waiting. This is reasonable in production environments where demand is high and where it is expensive to idle machines, e.g., due to high backorder costs or server shutdown costs. The non-idling model is also appropriate when a new product is being launched and the company wants to avoid stockouts as much as possible. For this Inventory Model, we do have convexity across sample paths over any finite time interval. Thus, the result holds over all non-idling periods, even in systems where there may be other periods where servers are permitted to idle (e.g. after sufficient inventory ramp up). After showing sample-pathwise convexity for this model, we show why the approach that works for the Inventory Model fails for the Service Model.

The only work we are aware of that addresses convexity is the paper by He and Down [33]. They show that in heavy traffic with Poisson arrivals and multiple parallel servers the full benefit of having some flexible customers can be achieved with an arbitrarily small proportion of customers being flexible. Another type of convexity result is based on the number of queues that flexible customers can access. Here also "a little flexibility goes a long way" with most of the advantage of routing choice being attained when there are only two choices, the so-called "power-of-two choices" (e.g. having bilingual customers in a multi-lingual facility) [52, 63].

In this chapter we consider two models called the *Service Model* and a related *Inventory Model*. The Service Model corresponds to the W_r design and is a multi-server queueing system with i.i.d. exponentially distributed service times for all servers, with an arrival process that is independent of the state of the system, but is otherwise arbitrary. The servers follow a nonidling but otherwise arbitrary service discipline (FCFS, LCFS, etc.). Some proportion of arrivals (dedicated arrivals) are obliged to use a particular dedicated server, while others (flexible arrivals) have the ability to use any server and they join the shortest queue upon arrival. Each arrival is flexible with probability p, independent of the other arrivals, and dedicated arrivals are equally likely to require each particular server. In the Service Model, of course, a server is busy only when there is a customer to serve. The Inventory Model, on the other hand, differs only in this respect: The servers are always busy, where a service completion at a server builds up inventory when that server does not have a customer to serve. For the Service Model, we allow any number of servers, but we require additional conditions on the arrival process so that the stationary expected waiting time is well defined. For the Inventory Model, our analysis is restricted to two servers.

In Section 5.1 we show convexity in a strong sample-path sense for the Inventory Model, and we show that convexity will not hold in this sense for the Service Model.

In Section 5.3 we show convexity for the stationary waiting time for the Service Model, and in Section 5.4 we describe extensions of our results.

5.1 The Inventory Model and Sample-Pathwise Convexity

We now consider an Inventory Model with two servers that are always busy, so, when no customers are waiting, each server continues working and builds up inventory. When a customer arrives to a queue that has inventory, it does not wait for service but leaves immediately while decreasing the inventory by one. For this model a negative value for the number of customers waiting at a server corresponds to the amount of inventory held. In this setting, we prove that the number of (actual) customers waiting in the system, and the inventory held, are both decreasing and convex in the proportion of flexible customers in a strong sample-path sense, as defined below.

For the Inventory Model we show a very strong form of convexity in which we have four coupled systems with different proportions of flexible customers, indexed by superscript j, where the proportion of flexible customers is $p^1 = p$, $p^2 = p^3 = p+\delta$, $p^4 = p + 2\delta$, respectively, with $\delta > 0$ and $0 \leq p \leq 1 - 2\delta$. For system $j = 1, 2, 3, 4$, let $N^j(t)$ be the vector of queue lengths at time t, where the components are ordered in decreasing order and $\bar{N}^j(t) = \sum_{i=1}^2 N_i^j(t)$. For system j at time t, define $Q^j(t) = \sum_{i=1}^2 (N_i^j(t) \vee 0)$ as the total number of actual customers and $I^j(t) = -\sum_{i=1}^2 (N_i^j(t) \wedge 0)$ as the total number of items in inventory in system j at time t. Finally let $W^j(t)$ denote the total time that customers spend in system j by time t, i.e., $W^j(t) = \sum_{k=1}^n Q^j(t_{k-1}^j)(t_k^j - t_{k-1}^j) + Q^j(t_n^j)(t - t_n^j)$, where $0 = t_0^j < t_1^j < ... < t_n^j < t$ denote the ordered arrival and potential service completion times in system j up to time t.

The following lemma gives monotonicity in p.

Lemma 32. *Under JSQ for flexible arrivals,*

(i) $\{N^4(t)\}_{t=0}^\infty \prec \{N^2(t)\}_{t=0}^\infty \prec \{N^1(t)\}_{t=0}^\infty$,
$\{N^4(t)\}_{t=0}^\infty \prec \{N^3(t)\}_{t=0}^\infty \prec \{N^1(t)\}_{t=0}^\infty$,

(ii) $\{\bar{N}^4(t)\}_{t=0}^\infty \leq_{st} \{\bar{N}^2(t)\}_{t=0}^\infty \leq_{st} \{\bar{N}^1(t)\}_{t=0}^\infty$,
$\{\bar{N}^4(t)\}_{t=0}^\infty \leq_{st} \{\bar{N}^3(t)\}_{t=0}^\infty \leq_{st} \{\bar{N}^1(t)\}_{t=0}^\infty$,

(iii) $\{Q^4(t)\}_{t=0}^{\infty} \leq_{st} \{Q^2(t)\}_{t=0}^{\infty} \leq_{st} \{Q^1(t)\}_{t=0}^{\infty}$,
$\{Q^4(t)\}_{t=0}^{\infty} \leq_{st} \{Q^3(t)\}_{t=0}^{\infty} \leq_{st} \{Q^1(t)\}_{t=0}^{\infty}$,

(iv) $\{I^4(t)\}_{t=0}^{\infty} \leq_{st} \{I^2(t)\}_{t=0}^{\infty} \leq_{st} \{I^1(t)\}_{t=0}^{\infty}$,
$\{I^4(t)\}_{t=0}^{\infty} \leq_{st} \{I^3(t)\}_{t=0}^{\infty} \leq_{st} \{I^1(t)\}_{t=0}^{\infty}$,

(v) $\{W^4(t)\}_{t=0}^{\infty} \leq_{st} \{W^2(t)\}_{t=0}^{\infty} \leq_{st} \{W^1(t)\}_{t=0}^{\infty}$,
$\{W^4(t)\}_{t=0}^{\infty} \leq_{st} \{W^3(t)\}_{t=0}^{\infty} \leq_{st} \{W^1(t)\}_{t=0}^{\infty}$.

Proof. Parts *(ii)*, *(iii)* and *(iv)* follow from *(i)*, because $\bar{N}^j(t)$, $Q^j(t)$ and $I^j(t)$ are all Schur-convex functions of $N^j(t)$ for all $j = 1, 2, 3, 4$. Also *(v)* easily follows from *(iii)* by definition. The proof of *(i)* uses coupling and forward induction similarly to the proof of Theorem 9 of Section 3. We will just prove $\{N^4(t)\}_{t=0}^{\infty} \prec \{N^2(t)\}_{t=0}^{\infty}$ as the other proofs are the same. Suppose $\{\tilde{N}^4(t)\}$ and $\{\tilde{N}^2(t)\}$ are stochastic processes having the same stochastic laws as $\{N^4(t)\}$ and $\{N^2(t)\}$. We will couple these processes so that

$$P(\{\tilde{N}^1(t)\}_{t=0}^{\infty} \prec \{\tilde{N}^2(t)\}_{t=0}^{\infty}) = 1. \tag{5.1}$$

Again, to ease the notational burden, we will omit the tildes henceforth on the coupled versions and just use $\{N^j(t)\}$, and we assume components are ordered in decreasing order.

We use induction on t_n, where t_n denotes the ordered arrival and potential service completion times described as above. Clearly (5.1) holds for $t = 0$ because $N^2(0) = N^4(0)$. 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 (5.1) holds for t_n.

We couple the arrivals so that a flexible arrival in $N^2(t)$ is also a flexible arrival in $N^4(t)$, and a dedicated arrival to the kth largest queue in $N^4(t)$ is to the kth largest queue in $N^2(t)$ if it is also dedicated. We couple the service completions so that whenever there is a service completion at the kth largest queue in $N^4(t)$, then a service completion occurs at the kth largest queue in $N^2(t)$. Then (5.1) holds for t_n by Corollary 2 and the proof is complete. ∎

Now we show sample-pathwise convexity.

Theorem 33. *Under JSQ for flexible arrivals,*

(i) $\{N^2(t) + N^3(t)\}_{t=0}^{\infty} \prec \{N^1(t) + N^4(t)\}_{t=0}^{\infty}$,

(ii) $\{Q^2(t) + Q^3(t)\}_{t=0}^{\infty} \leq_{st} \{Q^1(t) + Q^4(t)\}_{t=0}^{\infty}$,

(iii) $\{I^2(t) + I^3(t)\}_{t=0}^{\infty} \leq_{st} \{I^1(t) + I^4(t)\}_{t=0}^{\infty}$,

(iv) $\{W^2(t) + W^3(t)\}_{t=0}^{\infty} \leq_{st} \{W^1(t) + W^4(t)\}_{t=0}^{\infty}$.

Proof. Parts (ii) and (iii) follow from (i) by Lemma 7 and (iv) follows from (ii) by definition. Our proof of (i) uses coupling and forward induction similarly to the above proof of Lemma 32. Suppose $\{\tilde{N}^j(t)\}, j = 1, 2, 3, 4$, are stochastic processes having the same stochastic laws as $\{N^j(t)\}$. We will couple these processes so that

$$P(\{\tilde{N}^2(t) + \tilde{N}^3(t)\}_{t=0}^{\infty} \prec \{\tilde{N}^1(t) + \tilde{N}^4(t)\}_{t=0}^{\infty}) = 1. \tag{5.2}$$

Once again, to ease the notational burden, we will omit the tildes henceforth on the coupled versions and just use $\{N^j(t)\}$, and we assume components are ordered in decreasing order.

We couple the arrivals as follows. Arrivals occur at the same time in all systems. Let $\{U_n; n = 0, 1, 2...\}$ be a sequence of independent and identically distributed uniform random variables on the interval $[0, 1]$. We define the random variables $\mathbf{1}^j$ as:

$$\mathbf{1}^1 = 1\{U_n \in [0, p]\}$$

$$\mathbf{1}^2 = 1\{U_n \in [0, p + \delta]\}$$

$$\mathbf{1}^3 = 1\{U_n \in [0, p]\} + 1\{U_n \in [p + \delta, p + 2\delta]\}$$

$$\mathbf{1}^4 = 1\{U_n \in [0, p + 2\delta]\}$$

If $\mathbf{1}^j = 1$, then the corresponding arrival is flexible in $N^j(t)$, $j = 1, ..., 4$. If a dedicated arrival is to the kth largest queue in one system, then it is to the kth largest queue in all the systems where it is also dedicated, $k = 1, 2$. We couple the service completions so that whenever there is a service completion at the kth largest queue in $N^1(t)$, then a service completion occurs at the kth largest queue in $N^j(t)$, $j = 2, 3, 4$.

We again need to show that (5.2) holds for t_n. First suppose t_n is an arrival time. Due to our coupling, the following are the possible cases:

(i) The arrival is flexible in all systems. Then, from the first part of Lemma 6, with $i = j = 2$, $N^2(t_n) + N^3(t_n) \prec N^1(t_n) + N^4(t_n)$.

(ii) The arrival is flexible in $N^4(t)$ and $N^3(t)$, but dedicated in $N^1(t)$ and $N^2(t)$. If the dedicated arrival is to the longer queue, then from the first part of Lemma 6, with $i = 2, j = 1$, we have $N^2(t_n) + N^3(t_n) \prec N^1(t_n) + N^4(t_n)$. If the dedicated arrival is to the shorter queue, then from the first part of Lemma 6, with $i = j = 2$, $N^2(t_n) + N^3(t_n) \prec N^1(t_n) + N^4(t_n)$.

(iii) The arrival is flexible in $N^4(t)$ and $N^2(t)$, but dedicated in $N^1(t)$ and $N^3(t)$. Similar to previous case.

(iv) The arrival is dedicated in all systems. Then from the first part of Lemma 6, with $i = j$, $N^2(t_n) + N^3(t_n) \prec N^1(t_n) + N^4(t_n)$.

Second suppose t_n is a service completion time in the kth largest queue. Then using the second part of Lemma 6, with $k = i$ immediately yields $N^2(t_n) + N^3(t_n) \prec N^1(t_n) + N^4(t_n)$. ∎

Let $Q_t(p)$, $I_t(p)$ denote the number of (actual) customers and number of inventory in the system respectively at time $t > 0$, when the proportion of flexible customers is p. Also let $\overline{Q}_T(p) = \frac{1}{T}\int_0^T Q_s(p)ds$ and $\overline{I}_T(p) = \frac{1}{T}\int_0^T I_s(p)ds$ be the averages for a finite horizon, T. Finally let $\overline{W}_T(p)$ be the average time spent in the system by the (actual) customers until T, when the proportion of flexible customers is p. We then have the following corollary.

Corollary 34. *$\overline{Q}_T(p)$, $\overline{I}_T(p)$ and $\overline{W}_T(p)$ are all decreasing and convex in p for all $T > 0$.*

5.2 Sample-pathwise convexity for the Service Model

Note that if one tries to generalize the approach of Section 5.1 to the Service Model (the W_r design) with two servers, in which there is no inventory and servers idle when no customers are present, and using weak submajorization instead of majorization (because the total number of customers in the four systems will not be equal), it will not work. For an example consider the following case where the weak submajorization version of Lemma 6 fails. Let $a = (2, 1)$, $b = (3, 3)$, $c = (4, 2)$, $d = (5, 4)$. Then $a \prec_w b \prec_w d$, $a \prec_w c \prec_w d$ and $b + c \prec_w a + d$. However if $i = j = 2$, one can see that ${}^i b + {}^j c = (8, 6) \not\prec_w {}^i a + {}^j d = (7, 7)$. The following theorem generalizes this result.

Theorem 35. *For the Service Model, the total number in the system, $\sum_{i=1}^{2} N_i(t,p)$, is not decreasing and convex in p in the sample-path sense, where $N(t,p)$ is the queue-length vector at time $t > 0$ when the proportion of flexible customers is p.*

Proof. Consider four systems under JSQ for flexible arrivals, with proportion of flexible customers $p^1 = p$, $p^2 = p^3 = p + \delta$, $p^4 = p + 2\delta$ respectively, and where $\delta > 0$. For $j = 1, 2, 3, 4$, let $N^j(t)$ be the vector of queue lengths at time t. By definition, to show stochastic convexity in the sample-path sense [58], one needs to construct vectors $\tilde{N}^j(t)$ having the same stochastic laws as $N^j(t)$, such that:

(i) $\sum_{i=1}^{2} \tilde{N}_i^1(t) \geq max\{\sum_{i=1}^{2} \tilde{N}_i^2(t), \sum_{i=1}^{2} \tilde{N}_i^3(t)\}$ a.s.,

(ii) $\sum_{i=1}^{2} \tilde{N}_i^2(t) + \sum_{i=1}^{2} \tilde{N}_i^3(t) \leq \sum_{i=1}^{2} \tilde{N}_i^1(t) + \sum_{i=1}^{2} \tilde{N}_i^4(t)$ a.s..

To do this, we need to couple the processes such that whenever there is an arrival in one system, there is an arrival in the other systems as well. Furthermore if it is a flexible arrival in $\tilde{N}^4(t)$, then it has to be a flexible arrival in at least one of $\tilde{N}^2(t)$, $\tilde{N}^3(t)$. Otherwise, suppose all systems are empty initially and consider two arrivals which are both flexible in $\tilde{N}^4(t)$, and dedicated to the same queue in $\tilde{N}^j(t)$, $j = 1, 2, 3$. Then $\tilde{N}^1(t) = \tilde{N}^2(t) = \tilde{N}^3(t) = (2, 0)$, whereas $\tilde{N}^4(t) = (1, 1)$. So if the next time epoch is a service completion, no matter how you couple the potential service completions, because potential service completions from both queues have non-zero probability, there must be some realizations such that (ii) will be violated. Similarly if an arrival is a flexible one in $\tilde{N}^1(t)$, then it has to be a flexible arrival in both $\tilde{N}^2(t)$, $\tilde{N}^3(t)$. Otherwise as above, one can come up with $\tilde{N}^2(t) = \tilde{N}^3(t) = (2, 0)$, whereas $\tilde{N}^1(t) = (1, 1)$, and, if the next epoch is a service completion, again (ii) will be violated. So the only possible way of coupling the arrival processes is as we did in the proof of Theorem 33. For potential service completions, the only choice that will preserve (i) and (ii) is to couple the potential service completion in the kth largest queue together in all systems, $k = 1, 2$.

Under the above coupling (i) will be satisfied. Furthermore we have $\tilde{N}^4(t) \prec_w \tilde{N}^2(t) \prec_w \tilde{N}^1(t)$ and $\tilde{N}^4(t) \prec_w \tilde{N}^3(t) \prec_w \tilde{N}^1(t)$. However, these two will not be sufficient, since we do not have weak majorization on the sums as discussed above. In particular, one can come up with $\tilde{N}^1(t) = (2, 1)$, $\tilde{N}^2(t) = (2, 0)$, $\tilde{N}^3(t) = (2, 1)$, $\tilde{N}^4(t) = (1, 1)$, starting from all systems empty, after a couple of epochs, see Table

Table 5.1: Evolution of 4 coupled systems, where D_k denotes a dedicated arrival to the kth largest queue, S_k denotes a potential service completion in the kth largest queue, F denotes a flexible arrival and where $N^j = (0,0), \forall j$ at t_0.

j	$N^j(t_1)$			$N^j(t_2)$			$N^j(t_3)$			$N^j(t_4)$			$N^j(t_5)$			$N^j(t_6)$			$N^j(t_7)$		
1	D_1	1	0	D_1	2	0	S_2	2	0	F	2	1	S_1	1	1	D_1	2	1	S_2	2	0
2	D_1	1	0	D_1	2	0	S_2	2	0	F	2	1	S_1	1	1	F	2	1	S_2	2	0
3	D_1	1	0	F	1	1	S_2	1	0	F	1	1	S_1	1	0	D_1	2	0	S_2	2	0
4	D_1	1	0	F	1	1	S_2	1	0	F	2	1	S_1	1	0	F	1	1	S_2	1	0

5.1. Then, if the next epoch is a potential service completion in the shorter queue, (ii) will be violated. ∎

5.3 The Service Model and Stationary Mean Convexity

For the Service Model we must consider a weaker form of convexity, in particular for the stationary mean waiting time when the proportion of flexible customers is p, $W(p)$. For simplicity we first assume we have only two servers; later we discuss the extension to more than two servers. To guarantee that $W(p)$ is well defined and finite, we require that the interarrival sequence, $\{T_n, n = 0, \pm 1, ...\}$ be strongly mixing and stationary ergodic with $\frac{1}{2\mu} < ET$, where T is a random variable with the marginal distribution of any T_n. See [18] for a rigorous definition of strongly mixing. Intuitively, in a strongly mixing sequence, the dependencies between the intervals disappear as the time between the intervals goes to infinity. Under additional regularity conditions [18, Lemma 7], $W(0)$ exists and is finite. We showed in Theorem 9 of Section 3 that $W(p) \leq W(0), 0 \leq p \leq 1$, so the stationary waiting times are finite for all p. We start with a sample path for the model with fixed $\lambda = \frac{1}{ET}$, $\mu > \frac{\lambda}{2}$, and p (call this the p system) and construct a coupled sample path when the proportion of flexible customers becomes $p + \varepsilon$ (the $p + \varepsilon$ system). All arrivals and (potential) service completions occur at the same times in both systems, and if a service completion is from the shorter (longer) queue in the p system then it is also from the shorter (longer) queue in the $p + \varepsilon$ system. Let $\{U_j; j = 0, 1, 2...\}$ be a sequence of independent and identically distributed uniform random variables on the interval $[0, 1]$ and, for the jth arrival, let $A_j = 1\{U_j \in [0, p]\} + 1\{U_j \in [0, p + \varepsilon]\}$, so A_j represents the number of

flexible customers for the jth arrival for both systems. Thus, $A_j = 2$ if the arrival is flexible in both systems; $A_j = 0$ if it is dedicated in both (we call both of these "regular" customers), and $A_j = 1$ if the arrival is an "extra" arrival, i.e., it is flexible for the $p+\varepsilon$ system but not for the p system. When $A_j = 0$, if the dedicated customer goes to the shorter (longer) queue in the p system, then it does the same in the $p+\varepsilon$ system.

Rather than strict FCFS, we consider the following alternative service discipline that will make only extra customers experience the difference in waiting times for the two systems, i.e., it will maintain the same waiting times for regular customers in both systems. The extra customers in both systems have lowest preemptive priority, i.e., they are always at the back of whichever queue they join, and, among extra customers, the priority is LCFS-PR. Among regular customers the discipline is FCFS. We also have the following switching rule. Suppose a flexible regular customer ($A_j = 2$) joins the shorter queue (just observing the total queue lengths) and realizes that it would have been better off if it had joined the longer queue as it could have preempted the extra customers in the longer queue (i.e., the longer queue has fewer regular customers). Then we switch this customer with the first extra customer in the other queue. This makes the routing of regular flexible customers in the p and $p+\varepsilon$ systems the same without changing the queue lengths. Our alternative service discipline has the same overall mean waiting time as standard FCFS for all customers, because service times are i.i.d. and exponential, independent of the type of the customer. Also, because of our LCFS-PR and switching assumptions, the waiting times of regular customers, both flexible and dedicated, are the same for both systems. Hence the average waiting times, W_f and W_d of customers in these groups respectively, are also the same for both systems. Let $W_e(p)$ ($W_e(p+\varepsilon)$) be the stationary expected waiting time for the extra customers in the p ($p+\varepsilon$) system. We have

$$\begin{aligned} W(p) &= pW_f + (1-p-\varepsilon)W_d + \varepsilon W_e(p) \\ W(p+\varepsilon) &= pW_f + (1-p-\varepsilon)W_d + \varepsilon W_e(p+\varepsilon) \end{aligned}$$

so

$$\frac{dW(p)}{dp} = \lim_{\varepsilon\to 0} \frac{W(p+\varepsilon) - W(p)}{\varepsilon} = \lim_{\varepsilon\to 0}(W_e(p+\varepsilon) - W_e(p)).$$

Let us tag a single extra customer and let W_L (W_S) denote the waiting time of the tagged customer in the p ($p+\varepsilon$) system. We will condition on the number of other extra customers that arrive while our tagged customer is present, so first we show that this random variable has finite expectation. Under our assumptions on

the arrival process and because service times are i.i.d. and exponential, the mean waiting time for all customers is finite, even if customers are randomly assigned to types upon arrival and different types face different policies, as long as the policies are non-idling, keep the number in the two queues the same, and are stochastically equivalent to JSQ. Because the overall mean is a weighted average of type means, the type means must also be finite, i.e., $EW_L < \infty$, $EW_S < \infty$. Let K_L (K_S) denote the set of arrivals to the tagged customer's queue until it is gone from the p ($p+\varepsilon$) system. Because these customers have higher preemptive priority than the tagged customer we have

$$W_L \geq \sum_{i \in K_L} S_i \ , \ W_S \geq \sum_{i \in K_S} S_i,$$

where S_i denotes the service time of the ith customer. Taking the expectations yields

$$\infty > EW_L \geq \frac{E|K_L|}{\mu} \ , \ \infty > EW_S \geq \frac{E|K_S|}{\mu},$$

hence $E|K_L| < \infty$, $E|K_S| < \infty$. We now let K denote the set of arrivals that could affect the tagged customer, that is, $K = K_S \cup K_L$. Then we have $E|K| \leq E[|K_S| + |K_L|] = E|K_S| + E|K_L| < \infty$.

Let K_e denote the number of extra customers among K, so conditioned on the total number of customers $|K| = k$, we have $[K_e||K| = k] \sim Binomial(\varepsilon, k)$ for all k. Therefore,

$$P(K_e = 0) = \sum_{k=0}^{\infty} (1-\varepsilon)^k P(|K| = k) = G(1-\varepsilon)$$

where G is the generating function of $|K|$. We have $\lim_{\varepsilon \to 0} P(K_e = 0) = G(1) = 1$, so $P(K_e = 0) = 1 - O(\varepsilon)$. Similarly,

$$\frac{P(K_e = 1)}{\varepsilon} = \sum_{k=1}^{\infty} k(1-\varepsilon)^{k-1} P(|K| = k),$$

so $\lim_{\varepsilon \to 0} \frac{P(K_e=1)}{\varepsilon} = E|K|$. Because $E|K| < \infty$, we have $P(K_e = 1) = O(\varepsilon)$. Note that $E(K_e||K| = k) = \varepsilon k$, so $EK_e = \varepsilon E|K| < \infty$. We also have $EK_e \geq P(K_e = 1) + 2P(K_e \geq 2)$. Therefore

$$0 \leq \frac{P(K_e \geq 2)}{\varepsilon} \leq \frac{1}{2} \left(E|K| - \frac{P(K_e = 1)}{\varepsilon} \right)$$

implying $\lim_{\varepsilon\to 0}\frac{P(K_e\geq 2)}{\varepsilon}=0$, so $P(K_e\geq 2)=o(\varepsilon)$. We summarize these results as

$$P(K_e=0)=1-O(\epsilon), P(K_e=1)=O(\epsilon), P(K_e\geq 2)=o(\epsilon).$$

Let $W_e^i(p)$ ($W_e^i(p+\varepsilon)$) be the mean waiting time incurred by the tagged customer in the p ($p+\varepsilon$) system, given that $K_e=i$, so $W_e(p)=W_e^0(p)+O(\varepsilon)$, and similarly, $W_e(p+\varepsilon)=W_e^0(p+\varepsilon)+O(\varepsilon)$. (Under LCFS-PR, earlier arriving extra customers will have no effect on the tagged customer.) If the tagged customer happens to join the shortest queue in the p system (this will happen with probability 1/2), then there is no difference in waiting times for that customer in the two systems. Suppose the tagged customer goes to the longer queue in the p system (also with probability 1/2). Let $Y(p)$ be the *additional* expected waiting time for an extra customer that goes to the longer queue (in the p system) rather than the shorter queue (in the $p+\varepsilon$ system) given that no other extra customers arrive while it is in the system. We have

$$\begin{aligned} W'(p) &= \frac{dW(p)}{dp}=\lim_{\varepsilon\to 0}(W_e(p+\varepsilon)-W_e(p)) \\ &= \lim_{\varepsilon\to 0}(W_e^0(p+\varepsilon)-W_e^0(p)+O(\varepsilon)) \\ &= \lim_{\varepsilon\to 0}(-Y(p)/2+O(\varepsilon))=-Y(p)/2. \end{aligned}$$

An important note here is that we may assume, without loss of generality, that the tagged customer will not be switched if it is the only extra customer in the system at a given time, by assuming that if a flexible customer sees equal queues, it will go to the queue with the tagged customer. This is the only case where a switch after choosing the shortest queue would be beneficial to the regular customer. Then, for a particular sample path in both systems, the tagged customer will remain at the same queue until either it leaves in the $p+\varepsilon$ system, or the two queue lengths, not counting the tagged customer, become the same in both systems. In the former case the waiting time is smaller in the $p+\varepsilon$ system, i.e., $Y(p)\geq 0$ and $W(p)$ is decreasing in p. In the latter case the waiting time for the tagged customer in the two systems will be the same, so, for both cases $Y(p)\geq 0$ and $W(p)$ is decreasing in p. Note that although the overall arrival process is general, because dedicated arrivals are equally likely to join either of the two queues and service times are exponential, once the two queue lengths are equal, from that point on the two queues will be stochastically identical.

Summarizing the above we have the following theorem

Theorem 36. *$Y(p) \geq 0$ so $W'(p) = -\frac{Y(p)}{2} \leq 0$.*

Note that $W'(p) \leq 0$ follows from our earlier much stronger result, Theorem 9 of Section 3.

Now let us consider $Y(p)$, the additional stationary expected waiting time a random arrival (the tagged customer) must spend if it goes to the long queue rather than the short queue when the proportion of flexible customers is p, it has lowest preemptive priority, all other customers are regular customers and regular *flexible* customers will join the queue with the tagged customer when the queue lengths are equal. Let $Y(L, S, p)$ be the same thing conditional on the tagged customer finding L in the long queue and $S \leq L$ in the short queue, so $Y(L, L, p) = 0$. Throughout the rest of the paper when we use or make a statement about the term "queue length," we count the customers in service as well. Let $\hat{W}(A, B, p)$ be the additional expected waiting time for our tagged customer given it is at the back of a queue with A regular customers while the other queue has B regular customers. Here we use A and B instead of L and S because A can be shorter or longer than B. We will need the following lemma to derive further properties of $Y(L, S, p)$, which says that the waiting time of the tagged customer is increasing in both queue lengths.

Lemma 37.

$$\hat{W}(A+1, B, p) \geq \hat{W}(A, B, p), \ \forall A, B \tag{5.3}$$

$$\hat{W}(A, B+1, p) \geq \hat{W}(A, B, p), \ \forall A, B \tag{5.4}$$

Thus $\hat{W}(A, B, p)$ is increasing in A and B.

Proof. We prove (5.3) and (5.4) together. Let $N^1(t)$ ($N^2(t)$) be the vector of queue lengths, including the tagged customer, at time $t \geq 0$, where at $t = 0$, the two systems have the same queue lengths, except that system 1 has one more customer in one of the two queues (so it corresponds to the left hand side of (5.3) and (5.4), while system 2 corresponds to the right hand side). That is,

$$N^2_{[1]}(0) \leq N^1_{[1]}(0) \leq N^2_{[1]}(0) + 1 \text{ and } \sum_{i=1}^{2} N^1_i(0) = \sum_{i=1}^{2} N^2_i(0) + 1.$$

We couple the arrivals and potential service completion times so that a flexible arrival in $N^1(t)$ is also a flexible arrival in $N^2(t)$, and if a dedicated arrival occurs at the kth largest queue in $N^1(t)$ then a dedicated arrival occurs at the kth largest queue in

$N^2(t)$, and similarly for a potential service completion. A potential service completion will result in an actual service completion if and only if the queue is not empty. We will show, by induction, that

$$N^2_{[1]}(t) \leq N^1_{[1]}(t) \leq N^2_{[1]}(t) + 1 \text{ and}$$
$$\sum_{i=1}^{2} N_i^2(t) \leq \sum_{i=1}^{2} N_i^1(t) \leq \sum_{i=1}^{2} N_i^2(t) + 1, \forall t \geq 0, \tag{5.5}$$

so that either $N^1(t) = N^2(t)$ or $N^1(t)$ has one extra customer than $N^2(t)$ in one of the two queues. This is true at $t = 0$, so, suppose it is also true for some $t > 0$. If $\sum_{i=1}^{2} N_i^1(t) = \sum_{i=1}^{2} N_i^2(t)$, then $N^1_{[i]} = N^2_{[i]}$, $i = 1, 2$, so both systems are in the same state and we can couple them so that they always remain in the same state. Suppose $\sum_{i=1}^{2} N_i^1(t) = \sum_{i=1}^{2} N_i^2(t) + 1$. Let t_0 be the next arrival or potential service completion epoch after t. First, consider the case that t_0 is an arrival. It is sufficient to look at the following cases where the queue order changes in at least one system. This occurs when the queue lengths are equal in one of the two systems.

(i) $N^2_{[2]}(t) = N^2_{[1]}(t) = N^1_{[2]}(t) = N^1_{[1]}(t) - 1$. If the arrival is a dedicated arrival to the shorter queue or a flexible arrival, then $N^1_{[1]}(t_0) = N^1_{[2]}(t_0) = N^2_{[1]}(t_0) = N^2_{[2]}(t_0) + 1$, satisfying (5.5). If the arrival is a dedicated arrival to the longer queue, then $N^1_{[1]}(t_0) - 2 = N^1_{[2]}(t_0) = N^2_{[2]}(t_0) = N^2_{[1]}(t_0) - 1$, again satisfying (5.5).

(ii) $N^1_{[1]}(t) = N^1_{[2]}(t) = N^2_{[1]}(t) = N^2_{[2]}(t) + 1$. If the arrival is a dedicated arrival to the shorter queue or a flexible arrival, then $N^1_{[1]}(t_0) - 1 = N^1_{[2]}(t_0) = N^2_{[1]}(t_0) = N^2_{[2]}(t_0)$, satisfying (5.5). If the arrival is a dedicated arrival to the longer queue, then $N^1_{[2]}(t_0) + 1 = N^1_{[1]}(t_0) = N^2_{[1]}(t_0) = N^2_{[2]}(t_0) + 2$, satisfying (5.5).

Secondly, suppose t_0 is a potential service completion. If t_0 is an actual completion in both systems then the proof is very similar to the arrival case and (5.5) is satisfied at t_0. On the other hand if t_0 is an actual service completion in $N^1(t)$ but not in $N^2(t)$, then $N^1_{[2]}(t_0) = N^2_{[2]}(t_0) = 0$ and $N^1_{[1]}(t_0) = N^2_{[2]}(t_0)$, satisfying $N^1(t) = N^2(t)$.

Now we can prove (5.3) and (5.4). We will prove that the tagged customer in system 2 will be closer to the server at any given time $t \geq 0$ than the tagged customer in system 1. When the two queues are equal we label the queue with the

tagged customer as the short queue. We again use induction. The statement is true for $t = 0$. So suppose for some $t > 0$, the tagged customer in system 2 is closer to the server and that (5.5) holds. Let t_0 be the next arrival or potential service completion epoch. We will show that at t_0, the tagged customer in system 2 is still closer to the server. It is sufficient to consider the case that $N^1(t) \neq N^2(t)$ and the tagged customer is equally far from the server in $N^1(t)$ and $N^2(t)$ at t (so the queue without the tagged customer has one more customer in system 1). Consider the following four cases.

(i) The tagged customer is in the shorter queue in both systems (so $N^1_{[2]}(t) = N^2_{[2]}(t)$). If there is a flexible arrival or dedicated arrival to the shorter queue, the distance of the tagged customer will increase by one in both systems. If there is an actual service completion from the shorter queue their distance will both decrease by one or they will both depart.

(ii) The tagged customer is in the longer queue in both systems (so $N^1_{[1]}(t) = N^2_{[1]}(t)$). Similar to the previous case.

(iii) The tagged customer is in the shorter queue in $N^2(t)$ but it is in the longer queue in $N^1(t)$ (so $N^1_{[1]}(t) = N^2_{[2]}(t)$). This case cannot happen because it would violate (5.5).

(iv) The tagged customer is in the shorter queue in $N^1(t)$ but it is in the longer queue in $N^2(t)$ (so $N^1_{[2]}(t) = N^2_{[1]}(t)$). Then $N^1_{[1]}(t) = N^1_{[2]}(t) = N^2_{[1]}(t) > N^2_{[2]}(t)$. Now we couple dedicated arrivals and service completions so that a dedicated arrival or a service completion occurs at the queue with the tagged customer in both systems or it occurs at the queue without the tagged customer in both systems. Note that this coupling is consistent with the coupling of arrivals and service completions at the kth largest queue that we used to show (5.5), because the queue lengths are equal in system 1. If we have a flexible arrival it will go to the queue with the tagged customer in system 1 and to the other queue in system 2, and we still have the tagged customer closer to the server in system 2.

We have proved (5.3) and (5.4). ∎

Let L (S) be the number of customers in the longer (shorter) queue when the tagged customer arrived. The following theorem tells us that the advantage of being

in the shorter queue for the tagged customer decreases in the length of the short queue and increases in the length of the longer queue. Hence, the advantage is smaller when the queues are better balanced.

Theorem 38.

$$Y(L, S+1, p) \leq Y(L, S, p),\ \forall S < L \tag{5.6}$$

$$Y(L, S, p) \leq Y(L+1, S, p),\ \forall S \leq L \tag{5.7}$$

Proof. Suppose, without loss of generality, that the tagged customer arrives at $t = 0$. We first prove (5.6). Let $M^1(t)$ and $M^2(t)$ be the vector of queue lengths at time $t \geq 0$ excluding the tagged customer, so that $M^1_{[1]}(0) = M^2_{[1]}(0) = L$, $M^1_{[2]}(0) = S+1$ and $M^2_{[2]}(0) = S$. We couple the arrivals and the potential service completions as in the previous proof until $min(\tau, t_1)$, where t_i denotes the time that the tagged customer could depart, assuming it joins the shorter queue in the ith system, and τ denotes the time until the queue lengths excluding the tagged customer first become equal in $M^1(t)$. If $\tau \leq t_1$, then $Y(L, S+1, p) = 0 \leq Y(L, S, p)$ from Theorem 36, so suppose $\tau > t_1$. Because of our coupling, $t_2 < t_1$, so that, for $\tau > t_1$, $M^1(t_1) = M^2(t_1)$ and $T_1 = T_2$ where T_i denotes the time that the tagged customer could depart, assuming it joins the longer queue in the ith system. Therefore $Y(L, S+1, p) = T_1 - t_1 < T_2 - t_2 = Y(L, S, p)$, so (5.6) is true.

Next we prove (5.7). Now we have $M^1_{[2]}(0) = M^2_{[2]}(0) = S$, $M^1_{[1]}(0) = L$, $M^2_{[2]}(0) = L+1$. With τ, t_i, T_i defined as above, if $\tau \leq t_1$, then $Y(L, S, p) = 0 \leq Y(L+1, S, p)$. If $\tau > t_1$, then $t_1 = t_2$, $M^1_{[2]}(t_1) = M^2_{[2]}(t_1) = 0$, and $M^1_{[1]}(t_1) = M^2_{[1]}(t_1)+1$. So $Y(L, S, p)$ and $Y(L+1, S, p)$ will be the additional waiting time of the tagged customer from t_1 on, when it is at the back of the longer queue and the shorter queue is empty at t_1. We have by Lemma 37 that $T_2 \leq T_1$, so $Y(L, S, p) \leq Y(L+1, S, p)$ in this case. Hence, (5.7) is true and we are done. ∎

We are now ready to prove convexity.

Theorem 39. *$Y(p)$ is decreasing in p, so $W(p)$ is convex in p.*

Proof. We again consider two coupled systems, with respective proportions of flexible customers p and $p+\delta$, and a proportion δ of customers go to the shortest queue in the $p+\delta$ system whereas they are equally likely to go to either queue in the p system. Call these customers "δ customers." Again our tagged customer has

lowest priority in both systems. Arguing as we did for Theorem 39, we have that $Y(p) = Y_0(p)(1-O(\delta))+Y_1(p)(O(\delta))+o(\delta)\sum_{i=2}^{n} Y_i(p)$, where $Y_i(p)$ is the value of $Y(p)$ when there are i δ customers that arrive while the tagged customer is in the system. We have a similar expression for $Y(p+\delta)$. Note that $Y_0(p) = Y_0(p+\delta)$ because the tagged customer will not experience any effect from δ customers in either system. We now show that $Y_1(p+\delta) \leq Y_1(p)$. Let t_δ denote the arrival time of the δ customer, let t_d denote the time that the tagged customer could depart, assuming it joins the shorter queue, and let τ denote the time when the two queue lengths excluding our tagged customer become equal. If $\tau < min\{t_d, t_\delta\}$, then the two systems are stochastically identical from time τ on, so, $Y_1(p+\delta) = Y_1(p) = 0$. If $t_\delta < min\{t_d, \tau\}$, then the δ customer will join the shorter queue in $Y_1(p+\delta)$, whereas it is equally likely to go to either queue in $Y_1(p)$. If it goes to the shorter queue in $Y_1(p)$, then, $Y_1(p+\delta) = Y_1(p)$. Suppose it goes to the longer queue in $Y_1(p)$. Let L_{t_δ} and S_{t_δ} be the lengths of the long and short queues not counting the tagged customer at the time the δ customer arrives. Then $Y_1(p+\delta) = Y(L_{t_\delta}, S_{t_\delta}+1, p) \leq Y(L_{t_\delta}, S_{t_\delta}, p) \leq Y(L_{t_\delta}+1, S_{t_\delta}, p) = Y_1(p)$ by Theorem 38. If $t_d < min\{\tau, t_\delta\}$, then, arguing as in the previous case, either $Y_1(p+\delta) = Y_1(p)$ or $Y_1(p) - Y_1(p+\delta) = \hat{W}(L_{t_\delta}+1, S_{t_\delta}, p) - \hat{W}(L_{t_\delta}, S_{t_\delta}+1, p) \geq 0$ by Lemma 37. Finally

$$\begin{aligned} Y'(p) &= \frac{dY(p)}{dp} = \lim_{\delta\to 0} \frac{Y(p+\delta) - Y(p)}{\delta} \\ &= \lim_{\delta\to 0} \frac{(Y_1(p+\delta) - Y_1(p))O(\delta) + o(\delta)}{\delta} \leq 0, \end{aligned}$$

where we use the fact that the $O(\delta)$ term is non-negative because it is the probability of having one δ customer. Thus, $Y(p)$ is decreasing in p, and $W(p)$ is decreasing and convex in p. ∎

Finally, we calculate a closed form expression for the derivative when $p = 0$ and arrivals are Poisson. Let W_S, W_L denote the time that the tagged customer will spend in the short and long queues respectively. Since it stays at the back of the queue, by PASTA we have:

$$W_S = \frac{(S_\infty + 1)E[B]}{\mu}$$

$$W_L = \frac{(L_\infty + 1)E[B]}{\mu}$$

where $E[B]$ denotes expected busy cycle length and S_∞ (L_∞) denotes the stationary average length of the shorter (longer) queue. Since when $p = 0$, the two queues

are independently and geometrically distributed with common parameter $1-\rho$, their minimum is geometrically distributed with parameter $1-\rho^2$, so, $S_\infty = \frac{\rho^2}{1-\rho^2}$. Using $x \wedge y + x \vee y = x + y, \forall x, y \in \mathcal{R}$, we get $L_\infty = \frac{(2+\rho)\rho}{1-\rho^2}$. Finally, for $M/M/1$ queues, we have $E[B] = \frac{1}{\mu(1-\rho)}$. Hence,

$$W'(0) = -\frac{Y(0)}{2} = \frac{W_S - W_L}{2} = \frac{-\rho}{\mu(1-\rho^2)(1-\rho)} \leq W'(p), 0 \leq p \leq 1.$$

We also have that $W'(0)/W(0) = \frac{-\rho}{1-\rho^2}$.

Corollary 40. *For Poisson arrivals,* $W'(0) = \frac{-\rho}{\mu(1-\rho^2)(1-\rho)} \to -\infty$ *as* $\rho \to 1$.

Note that the result in the corollary above is consistent with the result of He and Down [33].

Example 41. *Suppose* $\rho = \frac{\lambda}{2\mu} = 0.9$ *(as in Figure 1.1) and* $\mu = 1$. *Because when* $p = 0$ *the two queues are independent* $M/M/1$ *queues,*

$$W(0) = \frac{1}{\mu(1-\rho)} = 10.$$

We also have by Corollary 40 *that* $W'(0) \approx -47$. *Therefore, for example, we could estimate the expected waiting time when* $p = 1\%$ *as* $10 - 0.01(47) \approx 9.5$. *Note that for* $p = 1$, *the expected waiting time is close to (and greater than) the expected waiting time in an* $M/M/2$ *queue, which is given by*

$$\frac{1}{\mu} + \frac{\rho^2}{\mu(1-\rho)(1+\rho^2)},$$

so, $W(1) \approx 5.48$. *Therefore, the maximal benefit of customer flexibility is approximately* $10 - 5.5 = 4.5$, *so* 1% *customer flexibility gives us about* $11\% \approx 0.5/4.5$ *of the maximal benefit due to full customer flexibility.*

Remark 42. *Note that the argument above can easily be extended to more than two servers. We need to consider the additional expected time of the tagged customer that goes to the shortest queue rather than to one of the other queues, say queue* K, *and use the same argument with "queue* K*" replacing "the longer queue."*

5.4 Extensions

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

Multi-Server Stations: Suppose we have two stations with two queues, but each station has $c \geq 1$ servers. Then the results for the Inventory Model will still be true because now each station can be modeled as a single server station with service rate being equal to the sum of the service rates of each server at that station.

For the Service Model, the convexity result will also hold. The proofs of Theorems 38 and 39 will be the same, because the systems are identical except for the tagged customer and the δ customer, so that they can be coupled in the obvious way. Similarly, in the proof of Lemma 37, either the two systems are identical or the first system has one additional customer. Therefore both systems can be coupled in the obvious way, so that at each event (arrival or service completion) either the additional customer remains in system 1, or it departs and both systems become identical, and Lemma 37 will still hold.

Impatience: Suppose the customers are impatient and abandon the system after waiting for some exponentially distributed time at rate α (either in queue or in service). For the Inventory Model, this extension is not proper, because abandonment of an inventory is not well defined. For the Service Model we define the waiting time as the time spent in the system regardless of the departure type (abandonment or service completion). As in the multi-server case, the results will still hold, since the systems are identical except for the tagged customer, the δ customer or the additional customer, and they can be coupled in the obvious way.

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 routing policy. For example, servers could go on- and off-line according to a random process.

Chapter 6

Conclusions

In this dissertation we studied partial flexibility in multi-server systems, where the arrival stream consists of two different types of customers. The first type is the group of customers who are dedicated to use a particular server and the second is the group of flexible customers that can be served by more than one server. We mainly focused on two types of problems. The first one was the routing problem where we sought to find the optimal routing policy for the arriving flexible customers. The second one was the scheduling problem and we studied the optimal policies for assigning customers to idle servers. Below we present a summary of our results.

6.1 Summary of Results

6.1.1 Routing Problem

In Chapter 3 we studied the optimal routing policy for the flexible customers. The optimal policy depends on the available information upon arrival. When the queue lengths are not known upon arrival it has been shown in literature under various settings that the "Round Robin" policy is optimal. When the only information available upon arrival is the queue length and all customers are flexible, routing the flexible customers to the shortest queue, has been shown to be optimal for a variety of problems as well. Routing flexible customers to the queue with the least number of customers is known in the literature as the "Join the Shortest Queue" (JSQ) policy. We extended JSQ policy to the following cases. We studied multiple stations having multiple identical servers. For this problem we showed the optimality of JSQ for 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 showed that when flexible customers followed 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. We also showed that the sojourn time for dedicated customers is decreasing in the proportion flexible. Then we considered customer abandonments, and showed that when customers abandon only from the queue, and the abandonment rate is greater than the service rate, even though JSQ no longer minimized the number of customers in the system, it still maximizes the service completion process. In a more realistic setting we studied queues with finite buffers, and we extended earlier known results by showing that JSQ stochastically maximizes the departure process pathwise when there is a mixture of flexible and dedicated arrivals and all buffers have the same capacities. Finally we considered several practical extensions such as slotted service times and resequencing requirements, and we showed that JSQ is optimal for these problems.

We then studied the problem where 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 "Join the Shortest Work (JSW)" policy is equivalent to the "First-Come-First-Served (FCFS)" policy with a single queue for all customers, and this policy is shown to be optimal for different objectives. It is known that JSW stochastically minimizes the workload at each time t for general arrival and service processes. It has also been proven that departures are sample-pathwise maximized by the JSW policy. We extended the earlier results to systems with homogeneous dedicated customers as well as flexible customers. We showed that among routing policies, the JSW policy stochastically maximized the departure process pathwise, and stochastically minimizes the workload at each time t.

6.1.2 Scheduling Problem

In Chapter 4 we considered the scheduling problem in a multi-server multi-queue system with partial flexibility. We studied the optimal assignment of waiting customers to the idle servers under various settings and objectives. We showed that in many situations the dedicated customers first (DCF) policy is optimal. Under DCF, whenever a server's dedicated queue is non-empty it gives priority to dedicated customers *and* does not idle.

We considered both permitting and not permitting preemption. 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. Within both preemptive and nonpreemptive classes

of policies, we also considered two subclasses, policies where idling is permitted and those where we forced nonidling. We considered different arrival and service processes within each setting. We also presented counterexamples showing cases where the results do not hold for more general arrival or service processes.

The results we proved are as follows.

- When preemption and idling are both allowed, nonidling DCF maximizes departure process pathwise for exponential service times and general arrival processes.
- When neither preemption nor idling is allowed, nonidling DCF weakly submajorizes the queue-length vector process for exponential service times and general arrival processes.
- When idling is allowed but preemption is not allowed, DCF (not necessarily nonidling) maximizes departure process pathwise for general service times and arrival processes.
- When idling is allowed but preemption is not allowed, nonidling DCF minimizes mean sojourn time of customers for exponential service times and Poisson arrivals.

Finally in Chapter 4 we compared different designs and policies. DCF is more efficient than the two routing designs discussed in Chapter 3, JSQ and JSW, in terms of minimizing the overall customer sojourn time, but at the expense of the flexible customers, who end up waiting longer on average. JSW performs better than JSQ and almost as well as DCF in terms of minimizing sojourn times. Also, empirically the overall performance of JSW is very close to that of DCF. Moreover, JSW is incentive compatible for flexible customers in the sense that it is their individually optimal policy.

6.1.3 Marginal Impact of Customer Flexibility

In Chapter 5 we considered the marginal impact of customer flexibility on system performance for the routing problem when queue lengths are observed. In particular we studied the change of overall waiting time with respect to the proportion of flexible customers, p. We showed the convexity of waiting time in p, which means that the marginal advantage of flexibility is largest at small proportions. It was not possible to obtain convexity in the strong sense for which monotonicity held in Chapter 3.

Instead, we showed a weaker form of convexity, that the stationary mean waiting time was convex in the proportion of flexible customers.

Although convexity in p is intuitive, it is surprisingly difficult to prove. To show this difficulty we first studied a high-production Inventory Model. Unlike the original model (Service Model, or W_r design), the servers are always busy, where a service completion at a server builds up inventory when that server does not have a customer to serve. For the Inventory Model, we obtained a sample-pathwise convexity result using majorization of the queue lengths. We also showed why the approach that worked for the Inventory Model failed for the Service Model.

We developed a new approach that combined marginal analysis with coupling to show convexity in the stationary mean waiting time. We considered a tagged customer, which had no effect other customers (lowest preemptive priority relative to the other customers). We showed that the derivative of the stationary waiting time with respect to p (the marginal value of customer flexibility) was the difference in expected waiting time between going to the long and the short queue for the tagged customer. We then showed, using another coupling argument, that this difference was decreasing in p.

6.2 Future Areas of Research

This section discusses future directions for this work.

Extension of the Optimality of JSQ: As discussed in Chapter 3, JSQ is the optimal routing policy under various settings such as service times distributions with increasing likelihood ratio. Weber [65] claimed to prove the result for distributions with increasing hazard rate (IHR), however this results was proven to be wrong by Koole et al. [42]. They also provided a counterexample showing that the JSQ policy did not stochastically maximize the departure process for all $t > 0$.

Note that the above counterexample also suggests that the strong monotonicity result with respect to the proportion of flexible customers is not true. However it is still very intuitive to think that the weaker result should be true, i.e., the stationary mean waiting time of flexible customers should be decreasing in the proportion of flexible customers who follow the JSQ policy when service times are IHR. We would like to use the marginal analysis idea of Chapter 5 to prove this result. When the service times were exponential, the nonnegative difference in expected waiting time

between going to the long and the short queue for the tagged customer, gave us the monotonicity result. For the increasing hazard rate service-time distributions we would like to use a similar idea, but this time we plan to study the difference in expected waiting time in terms of the workload in system rather than the queue lengths.

Convexity for the N_s design: When we studied the scheduling problem of Chapter 4, we also wanted to show the convexity result, i.e., the stationary mean waiting time is decreasing and convex in the proportion of flexible customers. For the W_s design the marginal analysis of Chapter 5 is not applicable because the δ customer of Theorem 39 might be dedicated to the queue that the tagged customer did not consider, and in this case $Y_1(p+\delta) \geq Y_1(p)$. However in a special case, where there is only one dedicated arrival stream, i.e., the N_s design, we conjecture that the above problem would not occur and we would be able to show the convexity result.

Pricing Mechanisms with Partial Flexibility: Pricing in multi-server multi-queue systems has been a recent popular area of study. Researchers have tried to address the optimal incentive compatible pricing and admission control policies for different types of customers in the system (see Mendelson [49], Mendelson and Wang [50], Afeche [2]). We also would like to study pricing strategies for our models, and would like to consider the changes in the optimal strategies with respect to various system parameters such as arrival rate and the proportion of flexible customers.

Bibliography

[1] Afeche P. (2012). Incentive-compatible revenue management in queueing systems: Optimal strategic delay and other delay tactics. Submitted.

[2] Aalto, S., Ayesta, U. and Righter R. (2009). On the Gittins index in the M/G/1 queue. *Queueing Systems.* **63**, 437–458.

[3] Agrawal, S. and Ramaswamy, R. (1987). Analysis of the resequencing delay for M/M/m systems. *ACM Sigmetrics Performance Evaluation Review.* **15,** 27–35.

[4] Ahn, H. S., Duenyas, I. and Zhang, R. Q. (2004). Optimal control of a flexible server. *Adv. Appl. Prob.* **36**, 139-170.

[5] Aksin, Z., Armony, M. and Mehrotra, V. (2007). The modern call-center: A multi-disciplinary perspective on operations management research. *Prod. Operat. Management*, **16**, 665–688.

[6] Aksin, O. Z., Karaesmen, F. and Ormeci, E. L. (2007). A review of workforce cross-training in call centers from an operations management perspective. In *Workforce Cross Training Handbook*, ed. D. Nembhard. CRC Press.

[7] Akgun, O. T., Righter, R. and Wolff, R. (2011). Multiple-server system with flexible arrivals. *Adv. Appl. Prob.* **43**, 985–1004.

[8] Akgun, O. T., Righter R. and Wolff, R. (2011) The power of partial power of two choices. *Performance Evaluation Review.* **39**, 46-48.

[9] Akgun, O. T., Righter, R. and Wolff, R. (2012). Understanding the marginal impact of customer flexibility. *Queueing Systems.* **71**, 5–23.

[10] Akgun, O. T., Righter, R. and Wolff, R. (2012). Partial flexibility in routing and scheduling. Submitted.

[11] Argon, N. T., Ding, L., Glazebrook, K. D. and Ziya, S. (2009). Dynamic routing of customers with general delay costs in a multiserver queueing system. *Prob. Eng. Inf. Sci.* **23,** 175–204.

[12] Bassamboo, A., Randhawa, R. S., and Van Mieghem, J. A. (2009). A little flexibility is all you need: Asymptotic optimality of tailored chaining and pairing in queuing systems. Submitted.

[13] Bambos, N. and Michailidis, G. (2002). On parallel queueing with random server connectivity and routing constraints. *Prob. Eng. Inf. Sci.* **16**, 185–204.

[14] 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.

[15] Ciardo, G., Riska, A. and Smirni, E. (2001). Equiload: A load balancing policy for clustered web servers. *Performance Evaluation.* **46**, 101-124.

[16] Combe, M. B. and Boxma, O. J. (1994). Optimization of static traffic allocation policies. *Theoret. Comput. Sci.* **125**, 17-43.

[17] Daley, D. J. (1987). Certain optimality properties of the first-come-first-served discipline for $G/G/s$ queues. *Stoch. Proc. Appl.* **25**, 301–308.

[18] Daley, D. J. and Rolski, T. (1992). Finiteness of waiting-time moments in general stationary single server queues. *Ann. Appl. Prob.* **2**, 987–1008.

[19] Down, D. G. and Lewis, M. E. (2010). The N-network model with upgrades. *Prob. Eng. Inf. Sci.* **24**, 171–200 (2010)

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

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

[22] Foschini, G. J. and Salz, J. (1978). A basic dynamic routing problem and diffusion. *IEEE Trans. Commun.* **26**, 320–327.

[23] Foss, S. G. (1980). Approximation of multichannel queueing systems. *Siberian Math. J.* **21**, 851–857.

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

[25] Garnett O. and Mandelbaum, A. (2000). An introduction to skills-based routing and its operational complexities. Teaching note, Technion, Haifa, Israel.

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

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

[28] 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.

[29] Gurumurthi S. and Benjaafar, S. (2004). Modeling and Analysis of Flexible Queuing Systems. *Naval Res. Logist.* **51**, 755–782.

[30] Harchol-Balter, M., Crovella, M. and Murta, C. (1999). On choosing a task assignment policy for a distributed server system. *J. Paral. Distr. Comput.* **59**, 204-228.

[31] Harrison, J. M. (1998). Heavy traffic analysis of a system with parallel servers: Asymptotic analysis of discrete-review policies. *Ann. Appl. Prob.* **8**, 822-848.

[32] Harrison, J. M. and Lopez, M. J. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing Systems.* **33**, 339-368.

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

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

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

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

[37] Hytia, E., Aaalto S., Penttinen A. and Virtamo, J. (2012). On the value function of the $M/G/1$ FIFO and LIFO queues. Submitted.

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

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

[40] Koole, G. (1992). On the optimality of FCFS for networks of multi-server queues. Technical report BS-R923, CWI, Amsterdam.

[41] Koole, G. (1992). Stochastic scheduling and dynamic programming, PhD Thesis, Leiden University.

[42] 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.

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

[44] Liu, Z. and Righter, R. (1998). Optimal load balancing on distributed homogeneous unreliable processors. *J. Operat. Res.* **46**, 563-573.

[45] Liu, Z. and Towsley, D. (1994). Optimality of the round-robin routing policy. *J. Appl. Prob.* **31**, 466-475.

[46] Liu, Z., Nain, P. and Towsley, D. (1995). Sample path methods in the control of queues. *Queueing Systems.* **21**, 293-335.

[47] Mandelbaum, A. and Stolyar, A. L. (2004). Scheduling flexible servers with convex delay costs: Heavy-taffic optimality of the generalaized $c\mu$-rule. *Operat. Res.* **52**, 836-855.

[48] Marshall, A. and Olkin, I. (1979). *Inequalities:Theory of Majorization and Its Applications.* Academic Press, Orlando.

[49] Mendelson, H. (1985). Pricing computer services: Queueing effects. *Comm. ACM.* **28**, 312–321.

[50] Mendelson, H. and Whang, S. (1990). Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Operat. Res.* **38**, 870–883.

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

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

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

[54] Reiman, M. I. (1984). Some diffusion approximations with state space collapse. In *Lecture Notes in Control and Information Sciences.* Springer, Berlin-New York, **60,** 209–240.

[55] 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**, 323–334.

[56] Riska, A., Sun W., Smirni, E. and Ciardo, G., (2002). AdaptLoad: Effective balancing in clustered web servers under transient load conditions. n *Proceedings of the 22nd International Conference on Distributed Computing Systems.* 104-111.

[57] 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.

[58] Shaked, M. and Shanthikumar, G. (1994). *Stochastic Orders and Their Applications.* Academic Press, San Diego.

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

[60] Sparaggis, P. D., Towsley, D. and Cassandras, C. G. (1993). Extremal properties of the shortest/longest non-full queue policies in finite-capacity systems with state-dependent service rates. *J. Appl. Prob.* **30,** 223–236.

[61] Stoyan D. (1976). A critical remark on a system approximation in queueing theory. *Math. Operationsforsch. Statist.* **7**, 953-956.

[62] 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 *Proceedings of the 29th IEEE Conference on Decision and Control.* **2,** 658–663.

[63] Turner, S. R. E. (1998). The effect of increasing routing choice on resource pooling. *Prob. Eng. Inf. Sci.* **12**, 109–124.

[64] Tsitsiklis, J.N., and Xu, K. (2011) On the power of (even a little) centralization in distributed processing. *ACM SIGMETRICS Performance Evaluation Review.* **39**, 121132.

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

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

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

[68] Wolff, R. (1977). An upper bound for multi-channel queues. *J. Appl. Prob.* **14**, 884–888.

[69] Wolff, R. (1987). Upper bounds on work in system for multi-channel queues. *J. Appl. Prob.* **24**, 547–551.

[70] Wolff, R. (1989). *Stochastic Modeling and Theory of Queues.* Prentice Hall, Englewood Cliffs, NJ.