May 2020

Scheduling Control for Many-Server Queues When Customers Change Class

Mary Elizabeth Saine
Clemson University, marylib1@gmail.com

Scheduling Control for Many-Server Queues When Customers Change Class

---

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Mathematical Sciences

---

by
Mary Lib Saine
May 2020

Accepted by:
Dr. Xin Liu, Committee Chair
Dr. Brian Fralix
Dr. Peter Kiessler

Abstract

We consider a two class, many-server queueing system which allows for customer abandonment and class changes. With the objective to minimize the long-run average holding cost, we formulate a stochastic queueing control problem. Instead of solving this directly, we apply a fluid scaling to obtain a deterministic counterpart to the problem. By considering the equilibrium of the deterministic solution, we can solve the resulting control problem, referred to as the equilibrium control problem (ECP), and use the solution to propose a priority policy for the original stochastic queueing system. We prove that in an overloaded system, under a fluid scaling, our policy is asymptotically optimal as it attains the lower bound formed by the solution of the ECP.

Table of Contents

Chapter 1

Introduction

Each day, the healthcare industry faces an allocation problem - with limited resources and greater demand than supply, how should patients be served? When patients arrive at a hospital, if they cannot be immediately attended to, they are placed in a queue to wait until a healthcare professional becomes available. However, long wait times can cause patients to become impatient and subsequently leave the queue, or their health could deteriorate due to the lack of medical attention, leading to a more severe condition than when they first arrived. A common approach is to prioritize those with a more serious medical condition; however, doing so may result in worsening conditions for less severe patients. Taking these possibilities into consideration, which patients should receive priority when a healthcare provider becomes available?

To address this question, we propose a many-server queue with two customer classes, where class is used to define the customer's need for service. In the healthcare sense, these two classes can be used to distinguish the severity of the patient's condition, i.e. moderate (low-priority) or urgent (high-priority). Our system has n identical servers, and class-i customers, for $i = 1, 2$, arrive according to a time-homogenous Poisson process with rate λ_i. A class-i customer who arrives and cannot be immediately served is placed in an infinite-capacity queue. In queue i, the duration of a class-i customer's service is exponentially distributed with rate μ_i. Customers may become impatient and choose to abandon the queue, and the patience times of class-i customers are exponentially distributed with rate θ_i. As they wait, customers may also change classes, thus moving from one queue to another. A customer switches from class-i to class-j after an exponential amount of time with rate ρ_{ij}. We assume that the arrival processes, service times, patience times, and the times to

change class are independent.

In this model, we consider a heavily loaded system with a high volume of arrivals and a large number of servers. Our goal is to minimize the long-run average holding cost, where the holding cost per unit time for a customer of class-i is given by the constant $c_i > 0$, by efficiently scheduling waiting customers to available servers. We do not intend to solve the control problem directly. Instead, we construct an asymptotic framework in which the arrival rates are assumed to be $\mathcal{O}(n)$ and the rates of service times, patience times, and times to change class are all assumed to be $\mathcal{O}(1)$. Under an appropriate scaling, referred to as the *fluid scaling*, we expect that our state processes will approach deterministic limits. Considering the equilibrium of the deterministic limits gives rise to a simple linear program (LP) which, when solved, leads to a simple priority policy. Using the solution of the LP, we propose a priority policy for the original queueing system. Our main result shows that the proposed policy is asymptotically optimal under fluid scaling.

We now review some of the existing literature relevant to the current work. One closely related work is Atar, Giat, and Shimkin [3]. In [3], a static priority rule, the so-called $c\mu/\theta$ rule, is designed to minimize the long-run holding cost for a multi-class many-server queue, accounting for the possibility that customers may abandon the queue while waiting to be served. This rule is an adaptation of the well known $c\mu$-rule and assigns priority to the queue class according to the order of their indices, $c_i\mu_i/\theta_i$, where c_i represents the class-i holding cost, μ_i is the service rate for customers of class i, and θ_i is the class-i abandonment rate. In other words, when a server becomes available, the queue with the largest $c_i\mu_i/\theta_i$ value is selected to be served. The distinguishing feature of our model, compared to [3], is that we introduce the capability for customers to change classes, in either direction, within the system - to move from, say, the low priority class to the high priority class, or vice versa. The ability for customers to change class in such a way is an important aspect of queueing systems due to its relevance in applications, such as call centers or, particularly, hospital waiting rooms. In the present work, following the main idea of [3], we also develop a static priority policy where the priority of each class depends on an index which depends on the holding cost, service rate, abandonment rate, and furthermore the class change rate in a more complicated way.

Similar to our model, Hu, Chan, and Dong [13] consider a two class many-server queue with abandonment, which allows customers of class 1 to move to class 2. They focus on a healthcare setting with two different classes of patients - moderate (class 1) and urgent (class 2), and a moderate patient who is not given proactive care may change classes to become an urgent patient. They also

consider the simple deterministic LP and derive a similar static priority policy. Their work is not concerned with the asymptotic optimality analysis of the priority policy. Instead, they focus on the equilibrium analysis of the fluid system under the priority policy, and the transient analysis of the system before reaching equilibrium state.

Scheduling plays a central role in many applications including manufacturing, computing, service, and healthcare systems (cf. [18, 2, 6, 11, 12]). There is much literature dedicated to the study of scheduling control of multiclass queues using fluid models. A recent tutorial work [19] considers a general multiclass many-server queue with abandonments - more particularly, a $G/GI/N + GI$ queue - and develops a fluid control problem for long-run average cost functionals. In addition, in Atar, Giat, and Shimkin [4], an ergodic cost function is considered for the same queueing system as [3], and the same $c\mu/\theta$ rule is shown to be asymptotically optimal. Fluid models are also commonly used for time-inhomogeneous systems. In [5, 17, 10, 16], heavily loaded systems are considered, and asymptotically optimal policies are developed, all under fluid scaling. Additionally, queueing models with class changes have been developed for organ transplant systems. Fluid queueing models which incorporate class changes are created for kidney and liver transplant systems to develop efficient allocation policies in [20] and [1]. Recently, [15] models a general transplant system as a stochastic matching queue, and develops an asymptotically optimal allocation policy under the fluid scaling. At last, the paper [9] considers a multiclass single-serve queueing system with class change and formulates the scheduling control problem as a Markov decision process.

The rest of the paper is organized as follows: In Section 2.1 we formulate the stochastic processes and queueing control problem. In Section 2.2, we introduce the asymptotic framework and implement the fluid scaling. In Section 2.3, we formulate the fluid control problem. We then translate the fluid model into a linear program which, when solved, leads to our proposed priority policy. Finally, Section 2.4 presents our proposed policy and main theoretic results and Section 2.5 contains proofs for those theoretic results.

Chapter 2

Queueing Model

2.1 Problem Formulation

The queueing system considered in this thesis consists of two classes of customers, class 1 and 2, each forming their own queue. These two queues are often interpreted as a high-priority queue and a low-priority queue. There are n identical servers who serve both classes of customers. This system of n servers is referred to as the nth system. Let $X_i^n(t)$ denote the number of class i customers in the system at time t, $Q_i^n(t)$ the number of class i customers in the queue at time t, and $Z_i^n(t)$ the number of class i customers being served at time t, where $i = 1, 2$. Thus, it must be true that for every $t \geq 0$, and $i = 1, 2$,

$$X_i^n(t) - Z_i^n(t) = Q_i^n(t) \geq 0, \tag{2.1}$$

$$Z_1^n(t) + Z_2^n(t) \leq n, \tag{2.2}$$

$$Z_i^n(t) \geq 0. \tag{2.3}$$

For $i = 1, 2$, the external arrival process to the ith queue is assumed to be a Poisson process with rate λ_i^n and we denote it as $\{A_i^n(t); t \geq 0\}$. We assume that the service times and patience times of customers are all independent of each other. For class i, the service times are exponentially distributed with rate μ_i^n, and the patience times are exponentially distributed with rate θ_i^n. Furthermore, if a customer of class i is still waiting in the queue after an exponential amount of time with rate ρ_{ij}^n, he/she will move to class j, and once the customer joins class j,

he/she becomes a customer of class j, where $i \neq j$. For $t \geq 0$, let $D_i^n(t)$ denote the number of class i service completions by time t, $R_i^n(t)$ denote the number of class i customers who abandon the system by time t, and $M_{ij}^n(t)$ denote the number of class i customers who have moved to class j by time t. The processes D_i^n, R_i^n, and M_{ij}^n, $i, j, = 1, 2, i \neq j$, can be formulated as follows. For $t \geq 0$,

$$D_i^n(t) = \tilde{D}_i^n \left(\int_0^t Z_i^n(s)\ ds \right), \tag{2.4}$$

$$R_i^n(t) = \tilde{R}_i^n \left(\int_0^t Q_i^n(s)\ ds \right), \tag{2.5}$$

$$M_{ij}^n(t) = \tilde{M}_{ij}^n \left(\int_0^t Q_i^n(s)\ ds \right), \tag{2.6}$$

where $\tilde{D}_i^n$, $\tilde{R}_i^n$, and $\tilde{M}_{ij}^n$ are independent Poisson processes with rates μ_i^n, θ_i^n, and ρ_{ij}^n, respectively. Finally, the state process can be described as follows. For $t \geq 0$,

$$X_i^n(t) = X_i^n(0) + A_i^n(t) - D_i^n(t) - R_i^n(t) - M_{ij}^n(t) + M_{ji}^n(t), \quad \text{for } i, j = 1, 2 \text{ and } i \neq j. \tag{2.7}$$

Finally, we assume that the initial state $X_i^n(0)$, the external arrival process A_i^n, and the Poisson processes $\tilde{D}_i^n$, $\tilde{R}_i^n$, and $\tilde{M}_{ij}^n$ are independent.

The n identical servers can serve both classes, which gives rise to a natural scheduling problem. Namely, when both queues are nonempty, which one should the next available server select to serve? Denote by π^n a scheduling policy for the nth system. The π^n is characterized by the system processes operated under it. Thus, we let

$$\pi^n = (X^n, Q^n, Z^n, D^n, R^n, M^n),$$

where $X^n = (X_1^n, X_2^n)^T, Q^n = (Q_1^n, Q_2^n)^T, Z^n = (Z_1^n, Z_2^n)^T, D^n = (D_1^n, D_2^n)^T, R^n = (R_1^n, R_2^n)^T$, and $M^n = (M_{12}^n, M_{21}^n)^T$. We are interested in minimizing the long run average holding cost of customers waiting in the queues by choosing scheduling policies. More precisely, let $c_i \geq 0$ be the holding cost per unit time for each class i customer. Then, the holding cost of the system at time t is given by $c_1 Q_1^n(t) + c_2 Q_2^n(t)$. Under policy π^n, the average holding cost function over the time interval $[0, T]$ is given by

$$C_T(\pi^n) = \frac{1}{T} E \left(\int_0^T c_1 Q_1^n(s) + c_2 Q_2^n(s)\ ds \right). \tag{2.8}$$

Let Π^n be the collection of all scheduling policy π^n (note that policies need not satisfy any work conservation condition) such that $(X^n, Q^n, Z^n, D^n, R^n, M^n)$ is right continuous with left limits. Our goal is to minimize $\lim_{T\to\infty} C_T(\pi^n)$ by choosing $\pi^n \in \Pi^n$.

2.2 Asymptotic Framework

We are interested in a heavily loaded system with large customer arrival rates and a large number of servers. The precise heavy traffic assumption is made below.

Assumption 1 (Heavy traffic condition)**.** *For $i = 1, 2$, let $\lambda_i > 0$, $\mu_i > 0$, $\theta_i > 0$, $\rho_{12} \geq 0$, $\rho_{21} \geq 0$ be constants. Then as $n \to \infty$,*

$$\frac{\lambda_i^n}{n} \to \lambda_i,\ \mu_i^n \to \mu_i,\ \theta_i^n \to \theta_i,\ \rho_{12}^n \to \rho_{12},\ \rho_{21}^n \to \rho_{21}.$$

We also make the following initial condition.

Assumption 2 (Initial condition)**.** *For a deterministic vector $x(0) \in \mathbb{R}_+^2$, as $n \to \infty$,*

$$\frac{X^n(0)}{n} \to x(0), \quad \textit{in probability.}$$

In the n-th system, we introduce the fluid scaled versions of the aforementioned processes by scaling down the original processes by the factor n.

$$\bar{X}^n = \frac{X^n}{n},\ \bar{Q}^n = \frac{Q^n}{n},\ \bar{Z}^n = \frac{Z^n}{n},\ \bar{A}^n = \frac{A^n}{n},\ \bar{R}^n = \frac{R^n}{n},\ \bar{D}^n = \frac{D^n}{n},\ \bar{M}^n = \frac{M^n}{n}.$$

By (2.1) – (2.3), and (2.7), these scaled processes satisfy: For $i = 1, 2$,

$$\bar{X}_i^n(t) = \bar{X}_i^n(0) + \bar{A}_i^n(t) - \bar{D}_i^n(t) - \bar{R}_i^n(t) - \bar{M}_{ij}^n(t) + \bar{M}_{ji}^n(t),\ j \neq i, \tag{2.9}$$

$$\bar{X}_i^n(t) - \bar{Z}_i^n(t) = \bar{Q}_i^n(t) \geq 0, \tag{2.10}$$

$$\bar{Z}_1^n(t) + \bar{Z}_2^n(t) \leq 1, \tag{2.11}$$

$$\bar{Z}_i^n(t) \geq 0. \tag{2.12}$$

Define the fluid scaled cost function for the n-th system under a policy π^n as follows:

$$\bar{C}_T(\pi^n) = \frac{C_T(\pi^n)}{n} = \frac{1}{T} E\left(\int_0^T c_1 \bar{Q}_1^n(s) + c_2 \bar{Q}_2^n(s) ds \right). \tag{2.13}$$

The fluid scaled control problem is to choose $\pi^n \in \Pi^n$ to minimize $\bar{C}_T(\pi^n)$. It is useful to recall an important result that will be used in this section.

Lemma 1 (Functional Law of Large Numbers (FLLN) for Poisson Processes)**.** *Let $\{N(t); t \geq 0\}$ be a Poisson process with rate λ. For $T \geq 0$,*

$$\sup_{t \in [0,T]} \left| \frac{N(nt)}{n} - \lambda t \right| \to 0, \text{ in probability, as } n \to \infty. \tag{2.14}$$

From Lemma 1, we would expect that

$$\frac{A_i^n(t)}{n} \approx \lambda_i t, \quad \frac{\tilde{D}_i^n(nt)}{n} \approx \mu_i t, \quad \frac{\tilde{R}_i^n(nt)}{n} \approx \theta_i t, \quad \frac{\tilde{M}_{ij}^n(nt)}{n} \approx \rho_{ij} t,$$

and the fluid scaled processes approach the following deterministic limits. For $i, j = 1, 2$ and $i \neq j$,

$$\begin{aligned}
&x_i(t) = x_i(0) + \lambda_i t - (\theta_i + \rho_{ij}) \int_0^t q_i(s)\, ds + \rho_{ji} \int_0^t q_j(s)\, ds - \mu_i \int_0^t z_i(s)\, ds, \\
&x_i(t) - z_i(t) = q_i(t) \geq 0, \\
&z_1(t) + z_2(t) \leq 1, \\
&z_i(t) \geq 0.
\end{aligned} \tag{2.15}$$

The equations in (2.15) will be referred to as the *fluid equations*, and a solution (x, q, z), where $x = \{(x_1(t), x_2(t))^T; t \geq 0\}, q = \{(q_1(t), q_2(t))^T; t \geq 0\}$ and $z = \{(z_1(t), z_2(t))^T; t \geq 0\}$, is called a *fluid limit*. In the next section, we will construct a control problem for the equilibrium of the fluid limit.

2.3 Fluid Model

2.3.1 Equilibrium Control Problem (ECP)

We consider the fluid model defined in (2.15). Corresponding to the fluid scaled control problem to minimize (2.13), we consider a deterministic control problem which is to minimize

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T c_1 q_1(t) + c_2 q_2(t) dt \tag{2.16}$$

by choosing z subject to the constraints of (2.15). We note that if $(q_1(t), q_2(t))$ converges to (q_1^e, q_2^e) as $t \to \infty$, then the objective function (2.16) would converge to $c_1 q_1^e + c_2 q_2^e$ (see Lemma 4 in the Appendix). Thus, we would like to first find q_1^e and q_2^e. This prompts us to find the equilibrium points for the fluid model.

Now consider the derivative of $x_i(t)$ and set it equal to 0. We have for $t \geq 0$,

$$\frac{dx_i(t)}{dt} = \lambda_i - (\theta_i + \rho_{ij})\, q_i(t) + \rho_{ji} q_j(t) - \mu_i z_i(t) = 0.$$

To find the equilibrium points for the fluid equations, let us consider the following system of equations. For $i = 1, 2$,

$$\begin{aligned} \lambda_i - (\theta_i + \rho_{ij})\, q_i^e + \rho_{ji} q_i^e - \mu_i z_i^e = 0,\ j \neq i, \\ x_i^e - z_i^e = q_i^e \geq 0, \\ z_1^e + z_2^e \leq 1, \\ z_i^e \geq 0. \end{aligned}$$

These constraints, together with the objective function $c_1 q_1^e + c_2 q_2^e$, give rise to the equilibrium control problem (ECP):

$$\begin{aligned} \min_{z_1^e, z_2^e} \quad & c_1 q_1^e + c_2 q_2^e \\ \text{s.t.} \quad & \lambda_i = (\theta_i + \rho_{ij}) q_i^e - \rho_{ji} q_j^e + \mu_i z_i^e, \quad i, j = 1, 2, i \neq j, \\ & q_i^e \geq 0, z_i^e \geq 0, \quad i = 1, 2, \\ & z_1^e + z_2^e \leq 1. \end{aligned} \tag{2.17}$$

2.3.2 Solving the ECP

The first constraint in (2.17) can be succinctly written in matrix form as

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} \theta_1 + \rho_{12} & -\rho_{21} \\ -\rho_{21} & \theta_2 + \rho_{21} \end{bmatrix} \begin{bmatrix} q_1^e \\ q_2^e \end{bmatrix} + \begin{bmatrix} \mu_1 z_1^e \\ \mu_2 z_2^e \end{bmatrix}$$

Solving for $(q_1^e, q_2^e)^T$,

$$\begin{bmatrix} q_1^e \\ q_2^e \end{bmatrix} = A \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} - A \begin{bmatrix} \mu_1 z_1^e \\ \mu_2 z_2^e \end{bmatrix}, \text{ where } A = \frac{1}{\theta_1\theta_2 + \rho_{12}\theta_2 + \rho_{21}\theta_1} \begin{bmatrix} \theta_2 + \rho_{21} & \rho_{21} \\ \rho_{12} & \theta_1 + \rho_{12} \end{bmatrix}.$$

Thus, the ECP is equivalent to maximizing

$$(c_1, c_2)A \begin{bmatrix} \mu_1 z_1^e \\ \mu_2 z_2^e \end{bmatrix} := b_1 z_1^e + b_2 z_2^e \tag{2.18}$$

over $z^e = (z_1^e, z_2^e)^T$, where

$$b_1 = \frac{\mu_1 \left(c_1 \left(\theta_2 + \rho_{21}\right) + c_2\rho_{12}\right)}{\theta_1\theta_2 + \rho_{12}\theta_2 + \rho_{21}\theta_1} > 0, \; b_2 = \frac{\mu_2 \left(c_2 \left(\theta_1 + \rho_{12}\right) + c_1\rho_{21}\right)}{\theta_1\theta_2 + \rho_{12}\theta_2 + \rho_{21}\theta_1} > 0,$$

subject to

$$\begin{aligned} &\begin{bmatrix} q_1^e \\ q_2^e \end{bmatrix} = A \begin{bmatrix} \lambda_1 - \mu_1 z_1^e \\ \lambda_2 - \mu_2 z_2^e \end{bmatrix} \geq 0, \\ &z_i^e \geq 0, \; i = 1, 2, \\ &z_1^e + z_2^e \leq 1. \end{aligned} \tag{2.19}$$

Thus, the ECP is to maximize (2.18) subject to (2.19). The optimal solution is to assign priority to the class with the larger index $b_i, i = 1, 2$. In other words, the optimal solution is to first assign the maximum feasible value to the $z_i^e, i = 1, 2$, corresponding to the larger index $\max(b_1, b_2)$, and then assign the maximum feasible value to the other z_i^e.

Let us consider two special cases of the ECP by limiting the flexibility of our queuing system.

1. Consider a system with no class changes, i.e. $\rho_{12} = \rho_{21} = 0$. This implies

$$b_1 = \frac{c_1\mu_1}{\theta_1} \text{ and } b_2 = \frac{c_2\mu_2}{\theta_2}.$$

We have recovered the $c\mu/\theta$ rule from Atar, Giat, Shimkin [3].

2. Consider a system where only customers in one class can change class. Without loss of generality, we consider a case where class 2 customers can change class to class 1, i.e. $\rho_{12} = 0$ and $\rho_{21} > 0$. In this case, b_1 is recovered as

$$b_1 = \frac{c_1\mu_1}{\theta_1},$$

which is the same as the case without class changes, but b_2 becomes:

$$\begin{aligned} b_2 &= \frac{\mu_2\left(c_2\theta_1 + c_1\rho_{21}\right)}{\theta_1\theta_2 + \rho_{21}\theta_1} \\ &= \frac{\mu_2 c_2}{\theta_2 + \rho_{21}} \frac{\theta_1 + \frac{c_1}{c_2}\rho_{21}}{\theta_1} \\ &= \frac{\mu_2 c_2}{\theta_2 + \rho_{21}} \left(1 + \frac{c_1}{c_2}\frac{\rho_{21}}{\theta_1}\right). \end{aligned}$$

It is important to note the effect that the value of ρ_{21} has on the value of the index b_2. By allowing for a one-directional class change, i.e. $\rho_{21} > 0$ and $\rho_{12} = 0$, the index b_2 increases compared to the case with no class change.

Example 1. *We consider an example with parameters* $c_1 = 4$, $c_2 = 1$, $\theta_1 = 3$, $\theta_2 = 2$, $\mu_1 = 3$, *and* $\mu_2 = 4$. *Thus,* $b_1 = 4$. *Notice that the value of* ρ_{21} *affects whether class 1 or class 2 is of priority. If we let* $\rho_{21} = 1$, *then* $b_2 = \frac{28}{9} < b_1 = 4$, *and so class 1 is of priority. However, if we let* ρ_{21} *take on a larger value, say* $\rho_{21} = 6$, *then* $b_2 = 4.5 > b_1 = 4$, *and so class 2 is of priority. In the healthcare context, this means that if the conditions of the class 2 (moderate) patients worsen very quickly, then it is better to give priority to class 2.*

To simplify our analysis of the ECP, we assume $\rho_{12} = 0$ and $\rho_{21} \geq 0$. In this way, we can solve the ECP explicitly. As mentioned earlier in the discussion of the second special case, letting

$\rho_{12} = 0$ results in the indices b_1, b_2 given as

$$b_1 = \frac{c_1\mu_1}{\theta_1}, \quad b_2 = \frac{\mu_2 c_2}{\theta_2 + \rho_{21}} \left(1 + \frac{c_1}{c_2}\frac{\rho_{21}}{\theta_1}\right). \tag{2.20}$$

Expanding (2.19), we obtain the equivalent set of constraints:

$$z_2^e \leq \frac{\lambda_2}{\mu_2}, \tag{2.21}$$

$$(\theta_2 + \rho_{21})\mu_1 z_1^e + \rho_{21}\mu_2 z_2^e \leq (\theta_2 + \rho_{21})\lambda_1 + \rho_{21}\lambda_2, \tag{2.22}$$

$$z_1^e + z_2^e \leq 1, \tag{2.23}$$

$$z_i^e \geq 0, \ i = 1, 2. \tag{2.24}$$

Thus, the ECP is equivalent to the optimization problem which maximizes (2.18) subject to (2.21) – (2.23). To solve the ECP, we need to consider two cases, when $b_1 < b_2$ and when $b_1 > b_2$.

Case 1: $b_1 < b_2$

When $b_1 < b_2$, class 2 is of priority. Therefore, in order to maximize (2.18), the optimal solution is to first assign the maximum feasible value to z_2^e, and then assign the maximum feasible value to z_1^e. In other words, the optimal solution (z_1^*, z_2^*) is given by

$$z_1^* = \min\left\{1 - \min\left\{1, \frac{\lambda_2}{\mu_2}\right\}, \frac{\lambda_1}{\mu_1}\right\}, \ z_2^* = \min\left\{1, \frac{\lambda_2}{\mu_2}\right\}. \tag{2.25}$$

Case 2: $b_1 > b_2$

When $b_1 > b_2$, class 1 is of priority. Consider (2.22). Letting $z_2^e = 0$, (2.22) yields

$$z_1^e \leq \frac{(\theta_2 + \rho_{21})\lambda_1 + \rho_{21}\lambda_2}{(\theta_2 + \rho_{21})\mu_1} = \frac{\lambda_1}{\mu_1} + \frac{\rho_{21}\lambda_2}{(\theta_2 + \rho_{21})\mu_1} := A$$

and letting $z_1^e = 0$, (2.22) yields

$$z_2^e \leq \frac{(\theta_2 + \rho_{21})\lambda_1 + \rho_{21}\lambda_2}{\rho_{21}\mu_2} = \frac{\lambda_2}{\mu_2} + \frac{(\theta_2 + \rho_{21})\lambda_1}{\rho_{21}\mu_2} := B$$

Also notice that because $b_1 > b_2$, we must have

$$\mu_1 > \frac{\mu_2\rho_{21}}{\theta_2 + \rho_{21}},$$

which implies

$$A = \frac{\lambda_1}{\mu_1} + \frac{\rho_{21}\lambda_2}{(\theta_2+\rho_{21})\mu_1} = \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}\frac{\mu_2\frac{\rho_{21}}{\theta_2+\rho_{21}}}{\mu_1} < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2},$$
$$B = \frac{\lambda_2}{\mu_2} + \frac{(\theta_2+\rho_{21})\lambda_1}{\rho_{21}\mu_2} = \frac{\lambda_2}{\mu_2} + \frac{\lambda_1}{\mu_1}\frac{\mu_1}{\mu_2\frac{\rho_{21}}{\theta_2+\rho_{21}}} > \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}.$$

This establishes

$$A < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < B. \tag{2.26}$$

The linear program, i.e. the ECP, can be explicitly solved for (z_1^*, z_2^*) by noting that the optimal solution must be on one of the vertices of the feasible region. The constraints given by (2.21) - (2.24) are simply lines bounding our feasible region. In the first quadrant, consider the three lines below:

$$z_1^e + z_2^e = 1, \tag{2.27}$$
$$z_2^e = \frac{\lambda_2}{\mu_2}, \tag{2.28}$$
$$(\theta_2+\rho_{21})\mu_1 z_1^e + \rho_{21}\mu_2 z_2^e = (\theta_2+\rho_{21})\lambda_1 + \rho_{21}\lambda_2. \tag{2.29}$$

To determine the optimal solution to the ECP, we must analyze the intersections of these three lines.

1. The intersection of (2.27) and (2.28) is the point $\left(1-\frac{\lambda_2}{\mu_2}, \frac{\lambda_2}{\mu_2}\right)$, denoted I_1.
2. The intersection of (2.27) and (2.29) is the point $\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$, denoted I_2.
3. The intersection of (2.28) and (2.29) is the point $(\alpha_0, 1-\alpha_0)$, denoted I_3,where

$$\alpha_0 = \frac{\theta_2\lambda_1 + \rho_{21}\left(\lambda_1+\lambda_2-\mu_2\right)}{\mu_1(\theta_2+\rho_{21}) - \rho_{21}\mu_2}.$$

The optimal solution would be one of the points I_1, I_2, I_3, or $(1,0)$. We consider three parameter regimes: $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2} \leq 1$, $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2} > 1, A \leq 1$, and $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2} > 1, A > 1$. The table below lists the optimal solution (z_1^*, z_2^*) for the ECP and the corresponding (q_1^*, q_2^*) under each case.

The overloaded regime is the only case that we are interested in since it is the only one where $q^* \neq 0$. When $q^* = 0$, we already have optimality, as queues are empty in any optimal solution.

$b_1 > b_2$		
	(z_1^*, z_2^*)	(q_1^*, q_2^*)
$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq 1$	$\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$	$(0, 0)$
$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} > 1$, $A < 1$	$(\alpha_0, 1-\alpha_0)$	$\left(0, \frac{\mu_1\mu_2\left(\frac{\lambda_2}{\mu_2}+\frac{\lambda_1}{\mu_1}-1\right)}{\mu_1(\theta_2+\rho_{21})-\rho_{21}\mu_2}\right)$
$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} > 1$, $A \geq 1$	$(1, 0)$	$\left(\frac{\lambda_1+\frac{\rho_{21}\lambda_2}{\theta_2+\rho_{21}}-\mu_1}{\theta_1}, \frac{\lambda_2}{\theta_2+\rho_{21}}\right)$

Table 2.1: Optimal solution of the ECP when $b_1 > b_2$.

2.4 Proposed Policy and Main Results

Throughout this section, we assume $\lambda_1/\mu_1 + \lambda_2/\mu_2 > 1$; thus, we are working with an overloaded system. Furthermore, we require $\rho_{12}^n = 0$ for all n. Based on the optimal solution of the ECP, we propose the following preemptive priority scheduling policy for the n-th system. Define

$$b_1^n = \frac{\mu_1^n\left(c_1\left(\theta_2^n + \rho_{21}^n\right) + c_2\rho_{12}^n\right)}{\theta_1^n\theta_2^n + \rho_{12}^n\theta_2^n + \rho_{21}^n\theta_1^n}, \quad \text{and} \quad b_2^n = \frac{\mu_2^n\left(c_2\left(\theta_1^n + \rho_{12}^n\right) + c_1\rho_{21}^n\right)}{\theta_1^n\theta_2^n + \rho_{12}^n\theta_2^n + \rho_{21}^n\theta_1^n}.$$

Without loss of generality, we assume $b_1^n > b_2^n$ and $b_1 > b_2$. Our policy assigns priority to customers of class 1. Denote by $\pi^{n,*}$ the proposed policy. Our main theoretic results are provided below.

Under the proposed policy $\pi^{n,*}$, we have for $t \geq 0$ and $i = 1, 2$,

$$\begin{aligned} \bar{Z}_i^n(t) &= \bar{X}_i^n(t) \wedge \left[1 - \sum_{k=1}^{i-1} \bar{X}_k^n(t)\right]^+, \\ \bar{Q}_i^n(t) &= \bar{X}_i^n(t) \wedge \left[\sum_{k=1}^{i} \bar{X}_k^n(t) - 1\right]^+. \end{aligned} \tag{2.30}$$

The corresponding fluid equations become

$$\begin{aligned} x_i(t) &= x_i(0) + \lambda_i t - \mu_i \int_0^t x_i(s) \wedge \left[1 - \sum_{k=1}^{i-1} x_k(s)\right]^+ ds \\ &\quad - (\theta_i + \rho_{ij}) \int_0^t x_i(s) \wedge \left[\sum_{k=1}^{i} x_k(s) - 1\right]^+ ds \\ &\quad + \rho_{ji} \int_0^t x_j(s) \wedge \left[\sum_{k=1}^{j} x_k(s) - 1\right]^+ ds, \qquad (2.31) \\ z_i(t) &= x_i(t) \wedge \left[1 - \sum_{k=1}^{i-1} x_k(t)\right]^+, \\ q_i(t) &= x_i(t) \wedge \left[\sum_{k=1}^{i} x_k(t) - 1\right]^+. \end{aligned}$$

Theorem 1. *Under the proposed policy $\pi^{n,*}$, we have for any $T \geq 0$, as $n \to \infty$,*

$$\sup_{t\in[0,T]} \left\|(\bar{X}^n(t), \bar{Q}^n(t), \bar{Z}^n(t)) - (x(t), q(t), z(t))\right\| \to 0, \quad \textit{in probability}, \qquad (2.32)$$

where $(\bar{X}^n, \bar{Q}^n, \bar{Z}^n)$ and (x, q, z) are as in (2.30) and (2.31), and as $t \to \infty$,

$$(x(t), q(t), z(t)) \to (x^*, q^*, z^*), \qquad (2.33)$$

where z^* is the optimal solution of the ECP, and (x^*, q^*) is the corresponding state process.

Let $V^* = (c_1, c_2)^T(q_1^*, q_2^*)$ be the optimal solution of the ECP.

Theorem 2 (Asymptotic optimality)**.** *The proposed policy $\pi^{n,*}$ is asymptotically optimal, i.e., for an arbitrary scheduling policy π^n,*

$$\limsup_{T\to\infty} \limsup_{n\to\infty} \bar{C}_T(\pi^{n,*}) = V^* \leq \liminf_{T\to\infty} \liminf_{n\to\infty} \bar{C}_T(\pi^n). \qquad (2.34)$$

2.5 Proofs of Theorems

We first present the C-tightness of the fluid scaled process in the following lemma. Its proof can be adapted from the proof of Proposition 7.1 in [15], and thus will be omitted in this thesis.

Lemma 2. *Under any scheduling policy π^n, the fluid scaled process $(\bar{X}^n, \bar{Q}^n, \bar{Z}^n)$ is C-tight.*

Proof of Theorem 1. We consider the proposed policy $\pi^{n,*}$. The state process $(\bar{X}^n, \bar{Q}^n, \bar{Z}^n)$ satisfies (2.30). From Lemma 2, $(\bar{X}^n, \bar{Q}^n, \bar{Z}^n)$ is C-tight. Let $(\bar{X}, \bar{Q}, \bar{Z})$ be a weak limit. Then, $(\bar{X}, \bar{Q}, \bar{Z})$ satisfies (2.31). From [8], there exists a unique solution to (2.31). Thus, $(\bar{X}^n, \bar{Q}^n, \bar{Z}^n)$ converges to the unique solution of the fluid equations (2.31) weakly, which establishes (2.32). The convergence in (2.33) follows from Theorem 2 and Theorem 3 in [13]. □

Proof of Theorem 2. We first show the inequality in (2.34), and consider an arbitrary policy π^n. From Lemma 2, $(\bar{X}^n, \bar{Q}^n, \bar{Z}^n)$ is C-tight. Let $(\bar{X}, \bar{Q}, \bar{Z})$ be a weak limit of $(\bar{X}^n, \bar{Q}^n, \bar{Z}^n)$ along a subsequence $\{n_k\}_{k=1}^{\infty}$. By the Skorohod representation theorem, without loss of generality, we can assume

$$(\bar{X}^{n_k}, \bar{Q}^{n_k}, \bar{Z}^{n_k}) \to (\bar{X}, \bar{Q}, \bar{Z})$$

almost surely and uniformly over $[0, T]$ for $T > 0$. From Lemma 1, $(\bar{X}, \bar{Q}, \bar{Z})$ satisfies the fluid equations (2.15). Since $x(0)$ is deterministic, the limit $(\bar{X}, \bar{Q}, \bar{Z})$ is deterministic. Using Fatou's Lemma,

$$\begin{aligned}
\liminf_{n\to\infty} \bar{C}_T(\pi^n) &= \liminf_{n\to\infty} \frac{1}{T} \int_0^T E\left(c_1 \bar{Q}_1^n(s) + c_2 \bar{Q}_2^n(s)\right) ds \\
&\geq \frac{1}{T} \int_0^T \liminf_{n\to\infty} E\left(c_1 \bar{Q}_1^n(s) + c_2 \bar{Q}_2^n(s)\right) ds \\
&= \frac{1}{T} \int_0^T c_1 E\left(\liminf_{n\to\infty} \bar{Q}_1^n(s)\right) + c_2 E\left(\liminf_{n\to\infty} \bar{Q}_2^n(s)\right) ds \\
&= \frac{1}{T} \int_0^T c_1 \bar{Q}_1(s) + c_2 \bar{Q}_2(s) ds,
\end{aligned}$$

where $(\bar{X}, \bar{Q}, \bar{Z})$ is a solution of the fluid equations (2.15). Let $\bar{q} = \frac{1}{T}\int_0^T \bar{Q}(s)ds$ and $\bar{z} = \frac{1}{T}\int_0^T \bar{Z}(s)ds$. Then,

$$\frac{\bar{X}_i(T)}{T} = \frac{x_i(0)}{T} + \lambda_i - (\theta_i + \rho_{ij})\,\bar{q}_i - \mu_i \bar{z}_i + \rho_{ji}\bar{q}_j.$$

From Lemma 3, $\frac{\bar{X}_i(T)}{T} - \frac{x_i(0)}{T} \to 0$ as $T \to \infty$. Let $\tilde{\lambda}_i = \lambda_i - \frac{\bar{X}_i(T) - x_i(0)}{T}$. We now have $(\bar{q}, \bar{z})$ satisfies

$$\tilde{\lambda}_i - (\theta_i + \rho_{ij})\,\bar{q}_i - \mu_i \bar{z}_i + \rho_{ji}\bar{q}_j = 0.$$

From Lemma 5 in the Appendix, we have

$$\begin{aligned}\liminf_{T\to\infty}\liminf_{n\to\infty}\bar{C}_T(\pi^n) &\geq \liminf_{T\to\infty}\frac{1}{T}\int_0^T c_1\bar{Q}_1(s)+c_2\bar{Q}_2(s)ds\\ &= \liminf_{T\to\infty}(c_1\bar{q}_1+c_2\bar{q}_2)\\ &\geq V^*.\end{aligned}$$

We next consider the proposed policy $\pi^{n,*}$, and note that the equality in (2.34) follows from (2.32) and (2.33). □

Appendices

Appendix A Some Elementary Lemmas

Lemma 3. *For any solution (x, q, z) of the fluid equations (2.15),*

$$\left(\frac{x(t)}{t}, \frac{q(t)}{t}, \frac{z(t)}{t}\right) \to (0,0,0), \quad \text{as } t \to \infty.$$

Proof. Define $u(t) = x_1(t) + x_2(t), t \geq 0$. Let $\alpha = \min\{\theta_1, \theta_2, \mu_1, \mu_2\}$. For $t \geq 0$, we have

$$\begin{aligned}
u(t) &= u(0) + (\lambda_1 + \lambda_2)t - \theta_1 \int_0^t q_1(s)ds - \theta_2 \int_0^t q_2(s)ds - \mu_1 \int_0^t z_1(s)ds - \mu_2 \int_0^t z_2(s)ds \\
&= u(0) + (\lambda_1 + \lambda_2)t - \alpha \int_0^t q_1(s) + q_2(t) + z_1(s) + z_2(s)ds - (\theta_1 - \alpha) \int_0^t q_1(s)ds \\
&\quad - (\theta_2 - \alpha) \int_0^t q_2(s)ds - (\mu_1 - \alpha) \int_0^t z_1(s)ds - (\mu_2 - \alpha) \int_0^t z_2(s)ds \\
&= u(0) + (\lambda_1 + \lambda_2)t - \alpha \int_0^t u(s)ds - \Delta(t),
\end{aligned}$$

where $\Delta(t) = (\theta_1 - \alpha) \int_0^t q_1(s)ds + (\theta_2 - \alpha) \int_0^t q_2(s)ds + (\mu_1 - \alpha) \int_0^t z_1(s)ds + (\mu_2 - \alpha) \int_0^t z_2(s)ds$. We next define

$$v(t) = u(0) + (\lambda_1 + \lambda_2)t - \alpha \int_0^t v(s)ds, \quad t \geq 0.$$

In the following, we show that $v(t) \geq u(t)$ for each $t \geq 0$ and $v(t)/t \to 0$ as $t \to \infty$.

Let $w(t) = v(t) - u(t)$. Then,

$$w(t) = -\alpha \int_0^t w(s)ds + \Delta(t).$$

Let $t_0 = \inf\{t \geq 0 : w(t) < 0\}$. Thus, $w(t_0) = 0$ and $\dot{w}(t_0) < 0$. We have

$$\dot{w}(t_0) = -\alpha w(t_0) + \dot{\Delta}(t_0) = \dot{\Delta}(t_0).$$

However, $\dot{\Delta}(t_0) \geq 0$ since $q_i(t), z_i(t) \geq 0$ for all t, $i = 1, 2$. Therefore, $w(t) = v(t) - u(t) \geq 0$ for all t.

Now, consider $\dot{v}(t) = (\lambda_1 + \lambda_2) - \alpha v(t) = -\alpha \left[v(t) - \frac{\lambda_1 + \lambda_2}{\alpha}\right]$. We define a homogenous differential equation $\dot{\tilde{v}}(t) = -\alpha \tilde{v}(t)$. This has a simple solution given by $\tilde{v}(t) = \tilde{v}(0)e^{-\alpha t}$. Setting

$\dot{v}(t) = \dot{\tilde{v}}(t)$, we have $\tilde{v}(t) = v(t) - \frac{\lambda_1+\lambda_2}{\alpha}$. So,

$$\begin{aligned} v(t) &= \tilde{v}(t) + \frac{\lambda_1 + \lambda_2}{\alpha} \\ &= \tilde{v}(0)e^{-\alpha t} + \frac{\lambda_1 + \lambda_2}{\alpha} \\ &= \left[v(0) - \frac{\lambda_1 + \lambda_2}{\alpha}\right] e^{-\alpha t} + \frac{\lambda_1 + \lambda_2}{\alpha}, \end{aligned}$$

which says $v(t)/t \to 0$ as $t \to \infty$. □

Lemma 4. *For a continuous, integrable function $y(t)$, if $\lim_{t\to\infty} y(t) = y^e$, then*

$$\frac{1}{t}\int_0^t y(s)ds \to y^e \text{ as } t \to \infty.$$

Proof. For all $\epsilon > 0$, there exists $T > 0$ such that for $t \geq T$,

$$|y(t) - y_0| < \epsilon.$$

Then,

$$\begin{aligned} \left|\frac{1}{t}\int_0^t y(s)ds - y_0\right| &= \left|\frac{1}{t}\int_0^T y(s)ds + \frac{1}{t}\int_T^t y(s)ds - y_0\right| \\ &\leq \left|\frac{1}{t}\int_0^T y(s)ds\right| + \left|\frac{1}{t}\int_T^t (y(s) - y_0)\, ds + \frac{t-T}{t}y_0 - y_0\right| \\ &= \left|\frac{1}{t}\int_0^T y(s)ds\right| + \frac{t-T}{t}\epsilon + \frac{T}{t}y_0 \\ &\to \epsilon \quad \text{as } t \to \infty. \end{aligned}$$

□

Lemma 5. *The solution of the ECP is continuous in the parameters $(\lambda, \mu, \theta, \rho)$.*

Proof. This result follows from the exact solutions derived in (2.25) and Table 2.1. □

Appendix B Tightness of stochastic processes

We adapt notation and definitions from [7]. Let S be a metric space, and $\mathcal{S}$ be the class of Borel subsets of S. A probability measure μ on $(S, \mathcal{S})$ is said to be *tight* if for each $\epsilon > 0$ there exists a compact set K such that $\mu(K) > 1 - \epsilon$. A family Π of probability measures on $(S, \mathcal{S})$ is said to be *tight* if for each $\epsilon > 0$ there exists a compact set K such that $\mu(K) > 1 - \epsilon$ for all $\mu \in \Pi$. A sequence $\{U_n(t); t \geq 0\}_{n=1}^{\infty}$ of stochastic processes is tight if the family of probability measures induced by $\{U_n(t); t \geq 0\}_{n=1}^{\infty}$ is tight. Furthermore, a sequence $\{U_n(t); t \geq 0\}_{n=1}^{\infty}$ of stochastic processes is C-tight if the sequence is tight and all weak limits are almost surely continuous.

The following theorem can be used to prove the C-tightness of $(\bar{X}^n, \bar{Q}^n, \bar{Z}^n)$.

Theorem 3 (Theorem VI.3.26 in [14]). *The sequence of stochastic processes $\{X^n(t); t \geq 0\}$ in $D([0, \infty), \mathbb{R}^K)$ is C-tight if and only if the following two conditions hold:*

(i) *For any $T \geq 0$,*

$$\lim_{a \uparrow \infty} P\left(\sup_{0 \leq t \leq T} |X^n(t)| > a \right) = 0, \quad n \geq 1.$$

(ii) *For any $\epsilon > 0$ and $0 \leq t_1 \leq t_2 < \infty$,*

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P\left(\sup_{0 \leq t_1 \leq t_2 \leq t_1 + \delta} |X^n(t_2) - X^n(t_1)| > \epsilon \right) = 0.$$

Theorem 4 (Skorohod representation theorem). *Let $\{\mu_n\}$ be a sequence of probability measures on a metric space S such that μ_n converges weakly to a probability measure μ on S as $n \to \infty$. Suppose also that μ is separable. Then there exists random variables X_n and X defined on a probability space such that X_n has the same law as μ_n for all n, and X has the same law as μ, and X_n converges to X almost surely.*

Bibliography

[1] Mustafa Akan, Oguzhan Alagoz, Baris Ata, Fatih Safa Erenay, and Adnan Said. A broader view of designing the liver allocation system. *Operations research*, 60(4):757–770, 2012.

[2] Zeynep Aksin, Mor Armony, and Vijay Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management*, 16(6):665–688, 2007.

[3] Rami Atar, Chanit Giat, and Nahum Shimkin. The cμ/θ rule for many-server queues with abandonment. *Operations Research*, 58(5):1427–1439, 2010.

[4] Rami Atar, Chanit Giat, and Nahum Shimkin. On the asymptotic optimality of the cμ/θ rule under ergodic cost. *Queueing Systems*, 67(2):127–144, 2011.

[5] Achal Bassamboo, J Michael Harrison, and Assaf Zeevi. Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems*, 51(3-4):249–285, 2005.

[6] Achal Bassamboo and Ramandeep Singh Randhawa. Scheduling homogeneous impatient customers. *Management Science*, 62(7):2129–2147, 2016.

[7] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[8] G Birkhoff, G Rota, and John Wiley. *Ordinary Differential Equations*. John Wiley and Sons, New York, 1989.

[9] Ping Cao and Jingui Xie. Optimal control of a multiclass queueing system when customers can change types. *Queueing Systems*, 82(3-4):285–313, 2016.

[10] Milica Čudina and Kavita Ramanan. Asymptotically optimal controls for time-inhomogeneous networks. *SIAM journal on control and optimization*, 49(2):611–645, 2011.

[11] Noah Gans, Ger Koole, and Avishai Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

[12] Wallace J Hopp and William S Lovejoy. *Hospital operations: Principles of high efficiency health care*. FT Press, 2012.

[13] Yue Hu, Carri W Chan, and Jing Dong. Optimal scheduling of proactive care with patient deterioration. Technical report, Columbia University Working Paper. http://www. columbia. edu/ jd2736 . . . , 2019.

[14] Jean Jacod and Albert Shiryaev. *Limit theorems for stochastic processes*, volume 288. Springer Science & Business Media, 2013.

[15] AMIN Khademi and X Liu. Asymptotically optimal allocation policies for transplant queueing systems. Technical report, Working Paper, 2018.

[16] Chihoon Lee, Xin Liu, Yunan Liu, and Ling Zhang. Optimal control of a time-varying double-ended production queueing model. *Available at SSRN 3367263*, 2019.

[17] Erhun Özkan and Amy R Ward. Dynamic matching for real-time ride sharing. *Stochastic Systems*, 2017.

[18] Michael Pinedo. *Scheduling: theory, algorithms, and systems.*, volume 5. Springer Science & Business Media Business Media, 2012.

[19] Amber L Puha and Amy R Ward. Scheduling an overloaded multiclass many-server queue with impatient customers. In *Operations Research & Management Science in the Age of Analytics*, pages 189–217. INFORMS, 2019.

[20] Stefanos A Zenios, Glenn M Chertow, and Lawrence M Wein. Dynamic allocation of kidneys to candidates on the transplant waiting list. *Operations Research*, 48(4):549–569, 2000.