Chapman University
Chapman University Digital Commons

Business Faculty Articles and Research

Business

10-1-2016

Optimality of the Fastest Available Server Policy

William P. Millhiser
The City University of New York,

Charu Sinha
Chapman University, csinha@chapman.edu

Matthew J. Sobel
Case Western Reserve University

Recommended Citation

Millhiser, W. P., Sinha, C., & Sobel, M. J. (2016). Optimality of the fastest available server policy. *Queueing Systems, 84*(3–4), 237–263. https://doi.org/10.1007/s11134-016-9502-1

Optimality of the Fastest Available Server Policy

William P. Millhiser
Department of Management, Zicklin School of Business
Baruch College, The City University of New York, New York, NY
william.millhiser@baruch.cuny.edu

Charu Sinha
Argyros School of Business and Economics
Chapman University, Orange, CA
csinha@chapman.edu

Matthew J. Sobel
Department of Operations, Weatherhead School of Management
Case Western Reserve University, Cleveland, OH
matthew.sobel@case.edu

July 28, 2013

Abstract

We give sufficient conditions under which a policy that assigns customers to the **F**astest **A**vailable **S**erver, labelled FAS, is optimal in queueing models with multiple independent Poisson arrival processes and heterogeneous parallel exponential servers. The criterion is to minimize the long-run average cost per unit time. We obtain results for loss models and for queueing systems with a finite-capacity or infinite-capacity buffer under a head-of-the-line priority scheme. The results depend on cost assumptions, so we analyze the robustness of the cost structure and present counter-examples to illustrate when FAS is *not* optimal.

1 Introduction

We consider a model in which costs are incurred as a stream of incoming customers to a service facility are assigned to idle servers. A problem arises of how best to match customers to servers because there are multiple classes of customers, nonidentical servers, and assignment costs that depend on class and service rate. This problem is motivated by urban emergency response systems, such as police, emergency medical, and fire/rescue services, and commercial emergency services such as roadside assistance, where response vehicles can be seen as heterogeneous servers due to different capabilities. Upon an emergency call (i.e., a customer arrival), if multiple vehicles are idle, a dispatcher must decide which vehicle to send considering the costs of matching the needs of the caller with the capabilities of available vehicles. A relatively simple decision rule is to send an idle vehicle that has the lowest assignment cost among all idle vehicles. The tradeoff is that the assigned vehicle will be temporarily unavailable for a possibly lower-cost assignment in the immediate future.

Assignment policies that optimize this tradeoff are not necessarily straightforward, so we seek a simplification: under what conditions would assigning the customer to the fastest available server minimize expected costs? Such decision-making problems arise also in other contexts such as mainframe computers with heterogeneous processors and call-centers with varying degrees of representative skill and experience (for a survey of the latter, see Akşin et al. 2007).

There is a rich literature on the assignment of customers to parallel servers. Policies that assign each customer to the Fastest Available Server (FAS) have been called *full-service* (Sobel 1982, 1990), *myopic* (Katehakis 1985), *fastest-queue* (Koole 1992), and *faster-servers-first* (Armony 2005). In earlier work, Seth (1977) shows that the FAS policy minimizes the long-run loss of customers due to blocking and abandonments when two heterogeneous servers have a finite number of buffers; Derman et al. (1980) give sufficient conditions for the FAS policy to maximize throughput in a loss system of K parallel servers. For subsequent results in loss models, see Shanthikumar and Yao (1987), Cooper and Palakurthi (1989), Sobel (1990), and Koole (1992). In the last-mentioned, Koole generalizes the conditions for which the FAS policy minimizes discounted and long-run average costs when the system owner must first accept or reject customer arrivals prior to the routing decision and each class has a unique blocking cost (but does not have a unique cost of assignment to each server).

The assignment of customers to heterogeneous servers by the FAS policy does not necessarily optimize models with costs that correspond to operating characteristics. Known as the *slow server problem* (Lin and Kumar 1984), a policy that intentionally idles slower servers and queues customers when the queue length falls below a particular threshold can minimize the mean sojourn time or the mean number of customers in the system (see additional references in Armony 2005, de Véricourt and Zhou 2006, and Kim et al. 2011). To complicate matters, policies that minimize the long-run average cost per unit time corresponding to such performance metrics do not necessarily minimize every reasonable cost criterion, as we demonstrate with a numerical example in §5.

That said, FAS assignment is intuitively appealing and easy to implement in practice. Thus, in both loss and waiting room models we present the first set of sufficient conditions on the reward/cost structure under which the FAS assignment of multiple customer classes to non-identical servers is optimal. The result for the loss model remedies a shortcoming in Sobel (1990) which is discussed in Appendix D. Also, §5 presents a counter-example to the optimality of the FAS policy in a generalized loss model with service rates depending *both* on the type of customer being served and the identity of the server.

Our approach exploits the fact that the loss and waiting room models each yield a continuous-time Markov decision chain (abbreviated MDC) which is uniformizable (see Jensen 1953 and a subsequent exposition in Serfozo 1979). We obtain sufficient conditions for the FAS policy to minimize the total expected discounted cost in n transitions (where n may be finite or infinite) in the uniformized model, hence it minimizes the total discounted cost and the long-run average cost per unit time in the MDC.

The sufficient conditions are robust assumptions about the costs of matching customers with idle servers at different levels of system congestion. These conditions include a property we call "faster is cheaper" where the cost of a customer's assignment to server j is less than the cost of assignment to server k if j's service rate is higher than k's, all else the same. This is reasonable in many contexts including those in which the costs include social elements, such as in emergency response, and requires no orderings of costs by customer classes, as a numerical example for the loss model demonstrates in Appendix A. Our cost structure also requires "more congestion is more expensive" which is to say that assignment costs are non-decreasing in the number of busy servers.

We model congestion in two ways: a *loss model* where arrivals are blocked when all servers are busy (e.g., when ambulances are diverted upon discovering all trauma bays occupied at an emergency department), and a single queue of infinite capacity which we call a *delay model*. In the latter, consistent with the literature on FAS policies, we adopt a head-of-the-line queue discipline. An alternative approach would be prioritization, say, via the $c\mu$ rule (Cox and Smith 1961). To our knowledge there are few results on the optimality of a $c\mu$-rule-like policy when many heterogeneous servers experience light traffic; the intuition for this is evident in the counter-example presented in section 5. Nonetheless, Harrison and López (1999), Williams (2000), and Mandelbaum and Stolyar (2004) show that the $c\mu$ rule is optimal in heterogeneous multi-server systems when traffic is sufficiently intense.

We assume that one server is required per customer. This models many phenomena well, but it precludes using a batch of servers for each customer, as is sometimes needed in emergency response (Green 1984), and it does not capture synergies when combinations of servers influence non-cost queueing performance (Andradóttir et al. 2011) and cost (Ahn and Lewis 2011).

This paper is organized as follows. Section 2 formulates the loss model and §3 analyzes it. Then §4 exploits the similarity between the loss and delay models to prove that the FAS policy is optimal in the delay model. Section 5 extends the result to the case of finite-capacity buffers, and gives counter-examples to show why the optimality of the FAS policy cannot be sustained (unless, perhaps, additional restrictions are imposed) when service rates depend on both the server's identity and the customer's type. Section 5 also shows why policies that optimize queueing costs that correspond to operating characteristics do not necessarily minimize other reasonable cost criteria, which we illustrate through a discussion of the differences between the results in Shanthikumar and Yao (1987) and this paper. Section 6 summarizes the paper. Appendix A contains a numerical example of a cost structure without an ordering of customer classes that satisfies our assumptions. Appendix B proves several results stated in the paper, including the main theorem. Appendix C gives sufficient conditions for the functional equation in the delay model to be well-defined. Appendix D explains why the proof of the main result in Sobel (1990) is flawed and presents a counter-example.

2 Formulation of the Loss Model

In this section we formulate the loss model as a Markov decision chain (MDC) which is regrettably difficult to analyze. So we transform it to an equivalent uniformized MDC that is less difficult to analyze, and specify a recursion with a value function that converges to that of the uniformized MDC. We show in §3 that the original MDC inherits the structure of an optimal policy in the recursion. Let $\hat{D}$ denote the original MDC which we now formulate.

A stream of customers arrive one at a time and, unless all servers are busy, are assigned to idle servers. If all servers are busy, then the arrival is blocked from entering the service facility and turned away. A cost is incurred whenever a customer arrives and is assigned or is blocked. The initial criterion is to assign arriving customers to idle servers so as to minimize the expected present value (abbreviated EPV) of the costs over a finite horizon. Subsequent results extend FAS optimality to the criteria of the EPV of the costs over an infinite horizon and, therefore, the long-run average cost per unit time. The remainder of this section presents the detailed formulation of the model and the cost assumptions on which the optimality of FAS is based.

There are J classes of customers who arrive via J independent Poisson arrival processes, and there are K heterogeneous exponential servers with rates that depend on server identity (but not on customer type). Let λ_j be the intensity of the jth arrival process, let $\lambda = \sum_{j=1}^{J} \lambda_j$ denote the cumulative arrival intensity, and let $\mathcal{J} = \{1, 2, ..., J\}$ denote the set of customer classes. We refer to an arrival from the j^{th} process as a "type-j" customer and we assume that a customer's type is known upon arrival. Let μ_k be the rate of server k; we label the servers with decreasing service rates, so that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K > 0$.

At any epoch, the set of K-dimensional *busy vectors* is

$$X = \{(x_1, x_2, \ldots, x_K)\colon\ x_k \in \{0, 1\},\ k = 1, \ldots, K\}$$

in which x_k is 0 or 1 according to whether server k is idle or busy at that epoch. Let $e \in X$ be the vector that indicates that every server is busy (i.e., $x_k = 1$ for all k). Let e_j be the j^{th} unit vector in X with all coordinates 0 except for 1 in the j^{th} coordinate.

We assume that an arriving customer must be assigned to a server if at least one is idle, so nontrivial decision-making epochs occur when a customer arrives to find more than one server idle. It follows that the respective sets of idle and busy servers to which a customer can and cannot be assigned, respectively, when x is the busy vector at an arrival epoch, are $I(x) = \{k : x_k = 0, k = 1, \ldots, K\}$ if $x \neq e$ and $B(x) = \{k : x_k = 1, k = 1, \ldots, K\}$ for all x. Let "b" label the "action" when a customer arrives to find every server busy and is, therefore, blocked from entering. It is convenient to define $I(e) = \{b\}$. Let $\#I(\cdot)$ denote the cardinality of $I(\cdot)$.

If x is the busy vector at an epoch, the aggregate service rate is

$$M(x) = \sum_{k=1}^{K} x_k \mu_k = \sum_{k \in B(x)} \mu_k, \quad x \in X,$$

and so $M(e)$ denotes the maximum aggregate service rate with all servers busy.

2.1 Cost Structure

In the original MDC of the loss model, $\hat{D}$, we assume costs are nonnegative, are incurred only at epochs when customers arrive, and are discounted at continuous-time discount rate $\alpha > 0$. The FAS policy is invariant with respect to α, so using the results in this paper does not entail estimating α. Although we formulate the loss model assuming that costs are incurred only at arrival epochs, later in this subsection we observe that this encompasses models in which costs are incurred continuously or at epochs when customers depart. Let $\hat{c}(x, j, k)$ denote the EPV of the cost that is incurred when an arriving type-j customer encounters busy vector x and is assigned to server k.

It is convenient to adopt the convention that $\hat{c}(e, j, k)$ is the same for all $k = 1, ..., K$, so henceforth we simply write $\hat{c}(e, j, b)$ to denote the cost when a type-j arrival encounters busy vector e and, hence, is blocked.

For busy vectors x and x' and $i \leq m$, write $x \preceq x'$ if $x = x' - e_m + e_i$. An example with $K = 5$ is $x = (1, 0, 0, 1, 0) \preceq x' = (1, 0, 0, 0, 1)$. A consequence of the labeling convention $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$ is that $x \preceq x'$ implies that x and x' have the same number of busy servers and $M(x) \geq M(x')$. Observe that X is not partially ordered by $\preceq$ because it lacks transitivity. The usual ordering is applied to $I(x)$ $(x \in X - \{e\})$.

Let $\theta(x) = \min\{k : k \in I(x)\}$ denote the fastest available server. If an arriving customer encounters busy vector x, the *fastest available server* policy (FAS) assigns the customer to server $\theta(x)$.

Cost assumptions. We make the following assumptions for each $j \in \mathcal{J}$ and $x \in X$ with $\#I(x) \geq 1$:

$$\frac{\mu_p}{\mu_q} \leq \frac{\hat{c}[x+e_q, j, \theta(x+e_q)]}{\hat{c}[x+e_p, j, \theta(x+e_p)]} \qquad \text{for all } p \leq q \text{ with } x_p = x_q = 0; \tag{1}$$

$$\frac{\mu_{\theta(x)}}{\mu_k} \leq \frac{\hat{c}[x, j, k]}{\hat{c}[x, j, \theta(x)]} \qquad \text{for all } k \in I(x); \tag{2}$$

$$\hat{c}(x, j, k) \leq \hat{c}(x+e_q, j, k) \qquad \text{for all } k \in I(x+e_q),\ q \in I(x). \tag{3}$$

Assumption (3) posits that expected costs are lower if there are fewer customers present. It is instructive to rewrite (1) and (2) as

$$\frac{\hat{c}[x+e_p, j, \theta(x+e_p)]}{1/\mu_p} \leq \frac{\hat{c}[x+e_q, j, \theta(x+e_q)]}{1/\mu_q} \qquad \text{if } p \leq q \text{ with } p \in I(x) \text{ and } q \in I(x);$$

$$\frac{\hat{c}[x, j, \theta(x)]}{1/\mu_{\theta(x)}} \leq \frac{\hat{c}[x, j, k]}{1/\mu_k} \qquad \text{for all } k \in I(x).$$

First, if the right-side denominator in either (1) or (2) is 0, then interpret it in the rewritten form. Second, the denominators in the *rewritten forms of* (1) and (2) are mean service times, and the left-hand numerators refer to systems with faster aggregate service rates the moment *after* assignment. That is, $x + e_p + e_{\theta(x+_p)} \preceq x + e_q + e_{\theta(x+_q)}$ (see Lemma 1 for a proof) and $x + e_{\theta(x)} \preceq x + e_k$. Thus, cost assumptions (1) and (2) can be interpreted as "faster is cheaper" because the assignment resulting in the overall faster system also has the lower expected cost per unit time. This interpretation is straightforward in (2) where the assigned server denoted in each numerator matches the index of the expected service time in each denominator. The same interpretation can be given to (1) by first noting that if $p \leq q$ and $x_p = x_q = 0$, then

$$\frac{\mu_{\theta(x+e_p)}}{\mu_{\theta(x+e_q)}} \leq \frac{\mu_p}{\mu_q}$$

(see the proof in Lemma 3, Appendix B). Therefore, if $p \leq q$, (1) implies

$$\frac{\hat{c}[x+e_p, j, \theta(x+e_p)]}{1/\mu_{\theta(x+e_p)}} \leq \frac{\hat{c}[x+e_q, j, \theta(x+e_q)]}{1/\mu_{\theta(x+e_q)}}$$

with $p \in I(x)$, $q \in I(x)$, and matching server indices in numerators and denominators.

Notice that (2) implies

$$\hat{c}[x, j, \theta(x)] = min\{\hat{c}(x, j, k) : k \in I(x)\} \text{ because}$$

$$\mu_k \hat{c}[x, j, \theta(x)] \leq \mu_{\theta(x)} \hat{c}[x, j, \theta(x)] \leq \mu_k \hat{c}(x, j, k)$$

with the left and right inequalities implied by $\mu_{\theta(x)} \geq \mu_k$ for all $k \in I(x)$ and (2), respectively.

Robustness of the cost structure. The foregoing specification of costs incurred only at arrival epochs is known to encompass various queueing-cost models in which costs are incurred continually while a customer is present in the system, and at departure epochs. Therefore, there is no loss of generality in our formulation with the costs incurred only upon arrival. For example, suppose that L_{jk} is incurred when a type-j customer departs after having been served by server k, and let τ_k be an exponential random variable with parameter μ_k. If a type-j customer is assigned to server k, then the EPV of L_{jk} *evaluated at the customer's arrival epoch* is

$$L_{jk}E\big(e^{-\alpha\tau_k}\big) = L_{jk}\frac{\mu_k}{\alpha+\mu_k}$$

which depends only on j and k, so it can become an element of $\hat{c}(x,j,k)$. Similarly, if $\mathcal{L}_{jk}$ is a continuously incurred cost rate from the arrival epoch to the departure epoch during a class-j customer's service by server k, then its EPV (evaluated at the customer's arrival epoch) is

$$\mathcal{L}_{jk}E\bigg(\int_0^{\tau_k} e^{-\alpha t}dt\bigg) = \mathcal{L}_{jk}\frac{\alpha}{\alpha+\mu_k}$$

which again depends only on j and k, so it too can become part of $\hat{c}(x,j,k)$. Such continuous cost rates may include the activation of server k during its processing of a type-j job, or alternatively the cost of holding that job in the system. In systems with a queue such as the delay model, $\mathcal{L}_{jk}$ excludes any costs incurred during a type-j job's wait in line. We show in Section 4 that such costs can also be made part of the cost at arrival $\hat{c}(x,j,k)$.

2.2 Markov Decision Chain (MDC) Formulation

In this model, decisions are made only at arrival epochs, so the specification of the state must provide relevant information at such epochs. Thus, the state must specify whether each server is busy or idle, i.e., the busy vector x, and the type of the newly arrived customer j. Therefore, the state space is $X \times \mathcal{J}$ and $I(x)$ is the set of feasible actions in state (x,j).

The initial criterion is to minimize the EPV of the costs over a finite horizon. We begin with a finite-horizon to infer a functional equation satisfied by the value function of the infinite-horizon problem, extend it to the infinite horizon, identify the corresponding equation in the uniformized MDC, use the latter equation in §3 to infer the optimality of FAS for EPV criteria in the uniformized MDC, and, therefore, its optimality for the criterion of the long-run average cost per unit time in the original MDC. Let $\hat{f}_n(x,j)$ denote the minimal EPV of the time stream of costs evaluated at the moment when a type-j customer arrives at a system with busy vector x that will shut down after n transitions (a combined total of n arrivals and service completions), $n = 1, 2, ...$

If $\#I(x) \geq 1$, then the type-j customer is assigned to an idle server k, the expected cost $\hat{c}(x,j,k)$ is incurred, and the new busy vector is $x + e_k$. It is useful to appreciate that the time stream of future costs depends on the new busy vector but not on j, the type of the customer who has just been assigned; all of the cost consequences of j are captured in $\hat{c}(x,j,k)$.

The moment after a transition—i.e., immediately after either a newly arrived customer has been assigned to an idle server (or blocked from entry if $x = e$) or a service completion has occurred—we let x denote the resulting busy vector and $\hat{h}_n(x)$ the minimal EPV of the time stream of costs if the system will shut down after n transitions. As we subsequently explain, $\hat{f}_n(x, j)$ and $\hat{h}_n(x)$ satisfy the following recursion where $\hat{h}_0(\cdot) \equiv 0$:

$$\hat{f}_n(x,j) \quad = \begin{cases} \min\{\hat{c}(x,j,k) + \hat{h}_n(x + e_k) : k \in I(x)\} & \text{if } \#I(x) \geq 1 \\ \hat{c}(e,j,b) + \hat{h}_n(e) & \text{if } x = e \end{cases} \qquad j \in \mathcal{J} \tag{4a}$$

$$\hat{h}_n(x) \quad = \sum_{m \in B(x)} \left(\frac{\mu_m}{\lambda + M(x) + \alpha} \right) \hat{h}_{n-1}(x - e_m) + \sum_{i=1}^{J} \left(\frac{\lambda_i}{\lambda + M(x) + \alpha} \right) \hat{f}_{n-1}(x, i). \tag{4b}$$

The recursion begins with the specification of $\hat{h}_0 \equiv 0$ in (4a) to calculate $\hat{f}_0$; then $\hat{h}_0$ and $\hat{f}_0$ are used in (4b) to calculate $\hat{h}_1$, and so on.

The first term on the right side of (4b) is associated with the next transition being the departure of the customer who is currently being processed by server m, and the second term is associated with the next transition being the arrival of a type-i customer. The probability of the former event is $\mu_m/[\lambda + M(x)]$ and, if it occurs, the expected value of the discounted elapsed time until the next transition is $[\lambda + M(x)]/[\lambda + M(x) + \alpha]$. The product of these quantities is the coefficient of $\hat{h}_{n-1}(x - e_m)$ in (4b). Similarly in the second term on the right side of (4b), the probability that the next transition is due to the arrival of a type-i customer is $\lambda_i/[\lambda + M(x)]$ and, if it occurs, the expected value of the discounted elapsed time until the next transition is $[\lambda + M(x)]/[\lambda + M(x) + \alpha]$. The product of these quantities is the coefficient of $\hat{f}_{n-1}(x, i)$ in (4b).

The recursion (4a) and (4b) corresponds to a finite-horizon discounted Markov decision process (MDP) with finitely many states and actions. Therefore, as $n \to \infty$ there are unique limit functions $\hat{f}(\cdot, \cdot)$ and $\hat{h}(\cdot)$ that satisfy the following functional equation (and which do not depend on the specification of $\hat{f}_0(\cdot, \cdot) \equiv 0$ and $\hat{h}_0(\cdot) \equiv 0$):

$$\hat{f}(x,j) \quad = \begin{cases} \min\{\hat{c}(x,j,k) + \hat{h}(x + e_k) : k \in I(x)\} & \text{if } \#I(x) \geq 1 \\ \hat{c}(e,j,b) + \hat{h}(e) & \text{if } x = e \end{cases} \qquad j \in \mathcal{J} \tag{5a}$$

$$\hat{h}(x) \quad = \sum_{m \in B(x)} \left(\frac{\mu_m}{\lambda + M(x) + \alpha} \right) \hat{h}(x - e_m) + \sum_{i=1}^{J} \left(\frac{\lambda_i}{\lambda + M(x) + \alpha} \right) \hat{f}(x, i). \tag{5b}$$

2.3 Uniformized MDC Formulation

In (5b), the dependence on x of the coefficients $\mu_m/[\lambda + M(x) + \alpha]$ and $\lambda_i/[\lambda + M(x) + \alpha]$ obstructs the analysis of (5a) and (5b). Replacing the original MDC, $\hat{D}$, with its uniformized counterpart, labelled D, removes this obstacle and considerably simplifies the ensuing analysis.

In $\hat{D}$, when an MDC enters a state, it resides there for a length of time with an exponential distribution having a rate that depends on the state and the action taken. In $\hat{D}$, let $\gamma(x, j, k)$ denote the rate of the exponentially distributed transition time immediately following the assignment of a type-j customer to server k. Since $\hat{D}$ corresponds to an MDC with finitely many states and actions, $\gamma(x, j, k)$ has a uniform upper bound. Specifically, $\gamma(x, j, k) \leq \lambda + M(e) < \infty$, and so MDC $\hat{D}$ corresponds to a *uniformized* MDC D, in which the transition probabilities and costs are modified to retain the marginal distributions of $\hat{D}$, *but the transition rates are the same for every combination of state and action* (Jensen 1953, Serfozo 1979). This property is achieved by introducing fictional transitions back to the state from which the process is departing.

The costs in D are denoted $c(\cdot, \cdot, \cdot)$ and they are given by the formula $c(x, j, k) = \hat{c}(x, j, k)[\alpha + \gamma(x, j, k)]/[\alpha + \lambda + M(e)]$. We choose the unit of time so $\lambda + M(e) = 1$. Therefore,

$$c(x, j, k) = \hat{c}(x, j, k)[\alpha + \lambda + M(x + e_k)]/(\alpha + 1) \tag{6a}$$

$$c(e, j, b) = \hat{c}(e, j, b)[\alpha + \lambda + M(e)]/(\alpha + 1) = \hat{c}(e, j, b). \tag{6b}$$

Let $\beta = 1/(1+\alpha)$ and let f denote the value function for the uniformized MDC. The domain of f, like $\hat{f}$, is $X \times \mathcal{J}$, and (5a) and (5b) imply that it satisfies the following equation:

$$f(x, j) \quad = \begin{cases} \min\{c(x, j, k) + h(x + e_k) : k \in I(x)\} & \text{if } \#I(x) \geq 1 \\ c(e, j, b) + h(e) & \text{if } x = e \end{cases} \qquad j \in \mathcal{J} \tag{7a}$$

$$h(x) \quad = \beta \sum_{m \in B(x)} \mu_m h(x - e_m) + \beta \sum_{i=1}^{J} \lambda_i f(x, i) + \beta[M(e) - M(x)]h(x). \tag{7b}$$

Analogous to (5a), the minimand of (7a) is the sum of the immediate cost $c(x, j, k)$ and the EPV of the subsequent cost. The latter, $h(x + e_k)$, depends only on the busy vector $x + e_k$ resulting from the assignment of the new customer. In (7b), μ_m is the probability that the next event will be the completion of service by server m, λ_i is the probability that the next event will be the arrival of a type-i customer, and

$$1 - \left(\sum_{m \in B(x)} \mu_m + \lambda \right) = M(e) - M(x)$$

is the probability of a fictional return to $h(x)$. So in the uniformized MDC, when the assignment of an arriving customer results in the busy vector x, the right side of (7b) is the EPV of the subsequent costs.

Equations (7a) and (7b) correspond to a discrete-time MDP with the criterion of minimizing the EPV of costs over an infinite horizon. The discount factor is $\beta = 1/(1 + \alpha)$ which, unlike the discount factor in $\hat{D}$, is invariant with respect to x. Therefore, the value functions f_n in the following recursion ($n = 0, 1, 2, \dots$) converge point-wise to f, the unique solution of (7a) and

(7b) *regardless of the choice of* h_0:

$$f_n(x,j) \quad = \begin{cases} \min\{c(x,j,k) + h_n(x+e_k) : k \in I(x)\} & \text{if } \#I(x) \geq 1 \\ c(e,j,b) + h_n(e) & \text{if } x = e \end{cases} \qquad j \in \mathcal{J} \tag{8a}$$

$$h_n(x) \quad = \beta \sum_{m \in B(x)} \mu_m h_{n-1}(x - e_m) + \beta \sum_{i=1}^{J} \lambda_i f_{n-1}(x,i) + \beta[M(e) - M(x)]h_{n-1}(x). \tag{8b}$$

The recursion begins with a specification of h_0 in (8a) to calculate f_0, then h_0 and f_0 are used in (8b) to calculate h_1, and so on.

3 Optimality of the FAS Policy

This section presents the primary properties of the loss model with results that ultimately lead to the optimality of the FAS policy. All proofs are available in Appendix B.

Theorem 1. *Assumptions (1), (2), (3), and* $h_0(\cdot) \equiv 0$ *imply for each* n *that* $k = \theta(x)$ *achieves the minimum in* (8a)*.*

The theorem states that the FAS policy minimizes the EPV of costs in the uniformized MDC for every length of finite horizon. Using standard arguments, it follows that the FAS policy minimizes the EPV of costs in the original MDC for every finite horizon and in the uniformized MDC with an infinite horizon.

Corollary 1. *Assumptions (1), (2), and (3) imply that* $k = \theta(x)$ *achieves the minimum in (4a) and (7a).*

Due to the relation between MDCs with and without uniformization, Corollary 1 implies that the FAS policy is optimal in the infinite-horizon MDC prior to uniformization. This result, stated next, has considerable practical importance because it neither depends on the value of the continuous-time discount rate α nor requires estimating it.

Corollary 2. *Assumptions (1), (2), and (3) imply that* $k = \theta(x)$ *achieves the minimum in (5a).*

The proof of the theorem is intertwined with the proof of the following proposition.

Proposition 1. *For all* $k \in I(x)$*,* $x \in X$ *with* $\#I(x) \geq 1$*, and* $n = 1, 2, ...$*,*

$$h_n(x) \quad \leq \quad h_n(x') \qquad \textit{if } x \preceq x', \quad \textit{and} \tag{9}$$

$$h_n(x) \quad \leq \quad h_n(x + e_k), \quad k \in I(x). \tag{10}$$

That is, for every finite horizon length in the uniformized model, there are two ways in which the minimal EPV of costs is a monotone function of the busy vector. The first is in the sense of the ordering $\preceq$, namely that if x and x' have the same number of busy servers, but the aggregate service rate is higher at x than at x', then the EPV will be lower with x than with x'. The second is in the sense that converting a server from idle to busy leads to higher costs.

The unified proof of Theorem 1 and Proposition 1 uses the following two preliminary results (with proofs in Appendix B).

Lemma 1. *If $x \preceq x'$, then $\theta(x) \geq \theta(x')$ and $x + e_{\theta(x)} \preceq x' + e_{\theta(x')}$.*

Thus, if x and x' have the same number of busy servers, but the aggregate service rate is higher at x than at x', there are two consequences. First, the index of the fastest idle server at x is at least as high as the index of the fastest idle server at x'. Second, if the fastest idle server becomes busy at x and x', then the augmented version of x retains a higher aggregate service rate than the augmented version of x'.

Lemma 2. *Assumption (1) implies*

$$c[x, j, \theta(x)] \leq c[x', j, \theta(x')] \quad \textit{for all } x \preceq x'. \tag{11}$$

So "faster is cheaper" in the sense of (1) has the following consequence. Suppose that x and x' have the same number of busy servers, but the aggregate service rate is higher at x than at x'. Then *in the uniformized model,* there is a lower immediate cost of assigning a newly-arrived customer to the fastest idle server in x than in x'.

Both the MDC model (prior to uniformization) and the uniformized MDC model correspond to MDPs with finitely many states and actions. So standard arguments (see Appendix B) have the following consequence of Corollaries 1 and 2 which is the primary result for the loss model.

Corollary 3. *In the loss model (in both the original MDC and the uniformized MDC), assumptions (1), (2), and (3) imply that the FAS policy minimizes the long-run average cost per unit time.*

We conclude that under assumptions (1), (2), (3), the FAS policy minimizes the EPV of costs over an infinite horizon and, most importantly, minimizes the long-run average cost per unit time.

4 Delay Model

This section explains that the FAS policy is optimal in a delay model with an unbounded queue capacity under the same assumptions that sustained optimality in the loss model. As in the loss model, there are J independent Poisson arrival processes and K heterogeneous exponential servers labeled so that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. So an assignment decision must be made whenever a customer arrives at a system with two or more free servers; the decision is trivial if there is one free server. Otherwise, the customer joins the queue and is matched with the next available server using a head-of-the-line priority scheme. We further assume that queued customers are patient and that no servers are idle while there are customers in the queue.

Notation. Due to the unbounded waiting room, we replace the K-dimensional busy vector in the loss model with an infinite-dimensional vector

$$x = (x_1, x_2, \ldots, x_K, x_{K+1}, x_{K+2}, \ldots)$$

where the first K components are the same as in the loss model: if $1 \leq k \leq K$, then $x_k = 0$ or 1 according to whether server k is idle or busy.

Let $N(x)$ be the number of customers in the system when the vector is x; $x_k = 0$ for all $k > max\{N(x), K\}$. If $N(x) > K$, then there are $N(x) - K$ customers in the queue and $x_k \in \mathcal{J}$ for $k = K+1, \ldots, N(x)$. That is, if $N(x) > K$, then $x_{K+1}, \ldots, x_{N(x)}$ are the "types" of the customers in the queue. With this specification, $N(x)$ is the number of non-zero components of x which we continue to call a *busy vector*.

For busy vectors x and x', we extend the notation $x \preceq x'$ to the delay model as follows: $x \preceq x'$ if $x = x' - e_m + e_i$ for $i \leq m$. Notice that this is possible only if $N(x) = N(x') < K$. Thus, Lemmas 1 and 2 remain valid for the delay model (with unaltered proofs). In the sequel, we refer to those lemmas although they were originally stated for the loss model. Let e_j denote the j^{th} unit vector with denumerably many coordinates all of which are 0 except for the j^{th} coordinate which is 1.

Upon a departure, if x is the busy vector the moment before server i completes serving a customer, we let $T(x)$ denote the new busy vector immediately after the customer leaves the system. If $N(x) \leq K$, then $T(x) = x - e_i$. If $N(x) > K$, then $T(x)$ does not depend on i and its components are $T(x)_k = x_k = 1$, $k = 1, ..., K$, and $T(x)_k = x_{k+1}$ for all $k > K$.

The following notation is similar to definitions in the loss model (but not precisely the same because x here has denumerably many components). Let $B(x) = \{k : x_k = 1, 1 \leq k \leq K\}$, $I(x) = \{k : x_k = 0, 1 \leq k \leq K\}$, and $M(x) = \sum_{k \in B(x)} \mu_k$. As in the loss model, choose units so $\lambda + \sum_{m=1}^{K} \mu_m = 1$ and let $\beta = 1/(1+\alpha)$. Therefore, if $N(x) \geq K$, $M(x) = 1 - \lambda$.

Cost structure. We continue to use the label b when a customer arrives to find every server busy; here, such a customer joins the queue, but b in the loss model signified that the customer was blocked from entering the system. If $N(x) < K$, then $\hat{c}(x, j, k)$ denotes the expected cost when a type-j arriving customer encounters busy-vector x and is assigned to server $k \in I(x)$. If $N(x) \geq K$, then $\hat{c}(x, j, b)$ denotes the expected cost when a type-j arriving customer encounters busy-vector x and must join the queue.

Robustness of the cost structure. The foregoing specification of costs encompasses many special cases. First, we give an example in which $\hat{c}(x, j, b)$ depends on the customer-type j. Suppose that customers impose costs on the system that depend on how long they wait in the queue, and their cost rates depend on their types. Let b_j denote the rate for a type-j customer. When a type-j customer joins the queue with the system having a busy vector x, there are $N(x) - K$ customers already in the queue, so the new customer's queueing time is a gamma random variable (r.v.) that is the sum of $N(x) - K + 1$ independent exponential r.v.s, each with the rate $\sum_{k=1}^{K} \mu_k$ which we denote μ. The reason is that the newly arrived customer's service begins only after all of the $N(x) - K$ customers ahead of the new arrival have been assigned to servers, and the first of them begins service only after one of the K customers currently being served leaves the system.

Thus, the new arrival's queueing time is a gamma r.v. with parameters μ and $n = N(x) - K + 1$. Therefore, the EPV of the queueing cost that this customer imposes on the system is

$$b_j \int_0^\infty \mu e^{-\mu y} \frac{(\mu y)^{n-1}}{(n-1)!} \int_0^y e^{-\alpha a} da dy = b_j \frac{1}{\alpha} \left[1 - \left(\frac{\mu}{\mu + \alpha} \right)^n \right]$$

in which the right-hand side results from a standard expansion and rearrangement of the left-hand side.

Second, we give an example in which $\hat{c}(x, j, b)$ depends on the identity of the server who processes the type-j customer. Let L_{jk} be the EPV of cost, billed at the beginning of service, when a type-j customer is processed by server k. Note that *regardless of the numerical value of $N(x)$ or of $N(x) - K$ (the size of the queue), the probability that a customer who joins the queue will be processed by server k is μ_k/μ.* So the EPV of the cost of "matching" the arriving type-j customer with the server who processes that customer is

$$\sum_{k=1}^{K} L_{jk} \frac{\mu_k}{\mu} = \frac{1}{\mu} \sum_{k=1}^{K} \mu_k L_{jk}.$$

Similarly, one may compute the EPV of a cost incurred at the end of service or the EPV of continuous costs incurred during service.

4.1 MDC in the Delay Model

This notation yields the following representation of the finite-horizon MDC (which is analogous to (4a) and (4b)) in which the boundary conditions are $\hat{f}_0(\cdot, \cdot) \equiv 0$ and $\hat{h}_0(\cdot) \equiv 0$:

$$\hat{f}_n(x,j) = \begin{cases} \min\{\hat{c}(x,j,k) + \hat{h}_n(x+e_k) : k \in I(x)\} & \text{if } N(x) < K \\ \hat{c}(x,j,b) + \hat{h}_n(x + je_{N(x)+1}) & \text{if } N(x) \geq K \end{cases} \tag{12a}$$

$$\hat{h}_n(x) = \begin{cases} \sum_{m\in B(x)} \left(\frac{\mu_m}{\lambda+M(x)+\alpha}\right) \hat{h}_{n-1}(x-e_m) & \\ \quad + \sum_{i=1}^{J} \left(\frac{\lambda_i}{\lambda+M(x)+\alpha}\right) \hat{f}_{n-1}(x,i) & \text{if } N(x) < K \\ \beta \sum_{m=1}^{K} \mu_m \hat{h}_{n-1}(x-e_m) & \\ \quad +\beta \sum_{i=1}^{J} \lambda_i \hat{f}_{n-1}(x,i) & \text{if } N(x) = K \\ \beta(1-\lambda)\hat{h}_{n-1}[T(x)] & \\ \quad +\beta \sum_{i=1}^{J} \lambda_i \hat{f}_{n-1}(x,i) & \text{if } N(x) > K. \end{cases} \tag{12b}$$

The coefficient β in (12b) is implied by $\beta = 1/(1+\alpha)$ and $1 = \lambda + \sum_{m=1}^{K} \mu_m$.

$$\sum_{m\in B(x)} \frac{\mu_m}{\lambda + M(x) + \alpha} = \frac{\sum_{m=1}^{K} \mu_m}{\lambda + \sum_{m=1}^{K} \mu_m + \alpha} = \frac{1-\lambda}{1+\alpha} = \beta(1-\lambda).$$

The recursion (12a) and (12b) corresponds to a finite-horizon discounted MDP with finitely many actions and denumerably many states. Therefore, by suitably bounding the rate at which $\hat{c}(x,j,k)$ can grow as $N(x)$ increases, as $n \to \infty$ there are unique limit functions $\hat{f}(\cdot,\cdot)$ and $\hat{h}(\cdot)$. See Appendix C for a suitable bound. The limit functions satisfy the following functional equations (which are analogous to (5a) and (5b) and do not depend on the specification of $\hat{f}_0(\cdot,\cdot) \equiv 0$ and $\hat{h}_0(\cdot) \equiv 0$):

$$\hat{f}(x,j) = \begin{cases} \min\{\hat{c}(x,j,k) + \hat{h}(x+e_k) : k \in I(x)\} & \text{if } N(x) < K \\ \hat{c}(x,j,b) + \hat{h}(x + je_{N(x)+1}) & \text{if } N(x) \geq K \end{cases} \tag{13a}$$

$$\hat{h}(x) = \begin{cases} \sum_{m\in B(x)} \left(\frac{\mu_m}{\lambda+M(x)+\alpha}\right) \hat{h}(x-e_m) & \\ \quad + \sum_{i=1}^{J} \left(\frac{\lambda_i}{\lambda+M(x)+\alpha}\right) \hat{f}(x,i) & \text{if } N(x) < K \\ \beta \sum_{m=1}^{K} \mu_m \hat{h}(x-e_m) & \\ \quad +\beta \sum_{i=1}^{J} \lambda_i \hat{f}(x,i) & \text{if } N(x) = K \\ \beta(1-\lambda)\hat{h}[T(x)] & \\ \quad +\beta \sum_{i=1}^{J} \lambda_i \hat{f}(x,i) & \text{if } N(x) > K. \end{cases} \tag{13b}$$

4.2 Uniformized MDC in the Delay Model

The logic that leads from the MDC functional equations (13a) and (13b) to their uniformized counterparts is similar to the logic for the loss model at the beginning of §2.3. There is an analogous parallelism for the uniformized MDCs in the loss and delay models. So here we explain only the cost function formula in the uniformized model. In contrast to (6a) and (6b), here

$$\begin{aligned} c(x,j,k) &= \hat{c}(x,j,k)[\alpha+\lambda+M(x+e_k)]/(\alpha+1) && \text{if } N(x) < K && \text{(14a)}\\ c(x,j,b) &= \hat{c}(x,j,b) && \text{if } N(x) \geq K. && \text{(14b)} \end{aligned}$$

If $N(x) < K$, (14a) corresponds exactly to (6a). If $N(x) \geq K$, (14b) results from

$$\begin{aligned} c(x,j,b) &= \hat{c}(x,j,b)\big[\alpha+\lambda+\sum_{k=1}^{K}\mu_k\big]/(\alpha+1)\\ &= \hat{c}(x,j,b)\big[\alpha+\lambda+1-\lambda\big]/(\alpha+1)\\ &= \hat{c}(x,j,b)(\alpha+1)/(\alpha+1) = \hat{c}(x,j,b). \end{aligned}$$

The following functional equations for the uniformized MDC use $1-\lambda = \sum_{m=1}^{K}\mu_m$ and they correspond to (13a) and (13b) in the same way that (7a) and (7b) correspond to (5a) and (5b):

$$f(x,j) = \begin{cases} \min\{c(x,j,k)+h(x+e_k) : k \in I(x)\} & \text{if } N(x) < K\\ c(x,j,b)+h(x+je_{N(x)+1}) & \text{if } N(x) \geq K \end{cases} \tag{15a}$$

$$h(x) = \begin{cases} \beta\sum_{m\in B(x)}\mu_m h(x-e_m) + \beta\sum_{i=1}^{J}\lambda_i f(x,i) & \\ \quad +\beta[1-\lambda-M(x)]h(x) & \text{if } N(x) < K\\ \beta\sum_{m=1}^{K}\mu_m h(x-e_m) + \beta\sum_{i=1}^{J}\lambda_i f(x,i) & \text{if } N(x) = K\\ \beta(1-\lambda)h[T(x)] + \beta\sum_{i=1}^{J}\lambda_i f(x,i) & \text{if } N(x) > K. \end{cases} \tag{15b}$$

In the case $N(x) = K$ there are two right-hand terms instead of three because the third would be $\beta(1-\lambda-\mu)h(x)$ which is zero because $\lambda+\mu = 1$.

Equations (15a) and (15b) correspond to an MDP with the criterion of minimizing the EPV of costs over an infinite horizon and a discount factor $\beta = 1/(1+\alpha)$. Therefore, the value functions f_n and h_n in the following recursion ($n = 0,1,2,\dots$) converge point-wise to f, the unique solution of (15a) and (15a) *regardless of the choice of* h_0 *and* f_0. This statement is valid if the rate at which $\hat{c}(x,j,k)$ can increase as $N(x)$ increases is suitably bounded. A suitable bound here is essentially the same as for (12a) and (12b). See Appendix C for such a bound. The recursion is

$$f_n(x,j) = \begin{cases} \min\{c(x,j,k)+h_n(x+e_k) : k \in I(x)\} & \text{if } N(x) < K\\ c(x,j,b)+h_n(x+je_{N(x)+1}) & \text{if } N(x) \geq K \end{cases} \tag{16a}$$

$$h_n(x) = \begin{cases} \beta\sum_{m\in B(x)}\mu_m h_{n-1}(x-e_m) + \beta\sum_{i=1}^{J}\lambda_i f_{n-1}(x,i) & \\ \quad +\beta[1-\lambda-M(x)]h_{n-1}(x) & \text{if } N(x) < K\\ \beta\sum_{m=1}^{K}\mu_m h_{n-1}(x-e_m) + \beta\sum_{i=1}^{J}\lambda_i f_{n-1}(x,i) & \text{if } N(x) = K\\ \beta(1-\lambda)h_{n-1}[T(x)] + \beta\sum_{i=1}^{J}\lambda_i f_{n-1}(x,i) & \text{if } N(x) > K. \end{cases} \tag{16b}$$

The key to the proof of the optimality of FAS in the delay model is that the minimization equation in (16a) is nearly the same as the minimization equation in (8a).

4.3 Optimality of FAS in the Delay Model

The same cost assumptions that yield optimality of FAS for the loss model imply that it is optimal in the delay model. The reason is that the proof of Theorem 1 and Proposition 1 remains valid here with no significant change.

Corollary 4. *Assumptions (1), (2), and (3) imply that $k = \theta(x)$ achieves the minimum in* (12a)*,* (13a)*,* (15a)*, and* (16a) *and minimizes the long-run average cost per unit time.*

5 Extensions and Counter-Examples

This section extends the loss and delay models to systems with finite-capacity buffers, addresses the distinction between the minimization of costs corresponding to operating characteristics and the minimization of more general costs, and presents a counterexample and intuition as to why the FAS policy may not minimize costs when service rates depend on the customer types in addition to the server's identity.

5.1 Finite-Capacity Queues

Consider a model similar to the loss and delay models except that it has a finite waiting room of size L. Thus, if there are fewer than $K + L$ customers in the system, an arriving customer is accepted into the system and immediately assigned to a server if one is available and otherwise added to the queue. If the number of customers in the queue grows to L, then further arrivals are rejected until the number in the queue drops below L. For example, when the server assignment is facilitated via a call center, there is a finite number of trunk lines, and, therefore, limited waiting space. The results of §3 can be extended to this model. So in a model with a finite waiting room, the FAS policy is optimal with respect to the expected discounted cost in finite and infinite horizons, and the long-run average cost per unit time under the same assumptions as in the loss model.

5.2 Costs that Correspond to Queueing System Performance

Assignment policies that minimize costs corresponding to an operating characteristic of the queueing system may not minimize every reasonable cost structure. To illustrate, consider the finite buffer capacity model in Shanthikumar and Yao (1987) that differs from our finite waiting room model only in that the inter-arrival times are not necessarily exponentially distributed. The objectives are to stochastically minimize the number of customers in the system and stochastically maximize the number of jobs processed by the system. These criteria induce a partial ordering of the set of all policies while our criterion induces a complete ordering of policies. However, their optimal policy coincides with ours, and, therefore, one might ask whether a policy that stochastically minimizes the number of customers in the system will necessarily minimize every reasonable cost criterion.

The following counterexample shows that their powerful results do not imply the optimality of the FAS policy. Let the cost of assignment be $\hat{c}(x, j, k) = [h_j + v(\mu_k)]/\mu_k$ where h_j is the class-j holding cost per unit time and $v(\mu_k)$ is the service cost per unit time for server k. Let $L = 0$ (i.e., loss model with no buffer), $J = 1$, $K = 3$, $\lambda = 1$, $h = 1$, $\hat{c}(e, j, b) = 5$, $\mu_1 = 6$, $\mu_2 = 5$, $\mu_3 = 1$ and $v(\mu_k) = v_k$ with $v_1 = 20$, $v_2 = 5$, and $v_3 = 1$. First, this cost structure is precluded by our assumptions because it violates (1) and (2). Second, let π denote a policy that assigns incoming customers to the available servers in the following order: server 2, server 1 and server 3. Using their criterion and results, the FAS policy outperforms policy π. However, if the performance of these policies is evaluated by comparing the average cost per unit time in each case, then the average cost per unit time for policy FAS is 1.59, while the average cost per unit time for policy π is 0.78. So the FAS policy is sub-optimal according to our criterion.

Thus the optimality of the FAS policy depends not only on the arrival process and service distribution, but also on the particular structure of the costs.

5.3 Service Rates

In the three models in which we have proved that the FAS policy is optimal, the service rates depend only on the identity of the server. Instead, augment the loss model in §2 so that the service rate μ_{kj} depends both on the the type of server k and the type of customer j being served. The following counter-example refutes a conjecture that the FAS policy is optimal with this generalization without further restrictions.

Consider a loss model with $J = K = 2$, $\lambda_1 = 0.9$, $\lambda_2 = 0.1$, $\mu_{11} = 30$, $\mu_{21} = \mu_{12} = 3$, and $\mu_{22} = 2$. Both servers are faster with a type-1 customer than a type-2 customer, server 1 is much faster than server 2 on a type-1 customer, and type-1 customers arrive more frequently than type-2 customers. Under the FAS policy, both types of customers who arrive to find the system empty are assigned to server 1.

If a customer arrives to find the system empty, another policy labeled ALT assigns customer type j to server j, $j = 1, 2$. This policy reserves server 1 for type-1 customers who arrive more often and are processed more rapidly than type-2 customers. We show that ALT dominates FAS in the sense that the blocking probability is lower.

For each policy, it is a straightforward task to write and solve the balance equations of the induced continuous-time Markov chain. The blocking probability is the stationary probability of having two customers in the system, and corresponds to the cost structure $\hat{c}(e, \cdot, b) \equiv 1$ and

$\hat{c}(x, j, k) = 0$ if $x \neq e$ which satisfies assumptions (1), (2), and (3). These probabilities under policies FAS and ALT, respectively, are 0.00591 and 0.00339. Since blocking probability is a special case of long-run average cost per unit time, we conclude that the FAS policy need not be optimal for models in which service rates depend both on server type and customer type.

6 Summary

Diverse applications have streams of arriving customers who must be matched with idle heterogeneous servers. Even in simple multi-channel models with heterogeneity, some performance criteria lead to complicated optimal matching policies. So we identify sufficient conditions that yield a simply structured optimal policy that can be implemented easily. That policy, the *fastest available server* (FAS) policy, assigns each customer to the fastest idle server.

In our models, multiple classes of customers arrive via independent Poisson processes and service times are exponentially distributed with server-specific rates. Costs are incurred at moments of time when customers enter the system but this encompasses many models in which costs are incurred continuously or at the end of service. The expected cost at such a moment depends on the "busy vector," namely which servers are free and which are idle, the class of the customer who is being assigned to a server, and the identity of the server to which the customer is assigned.

We consider loss, delay, and finite waiting room queueing models. In the loss model, customers are blocked from entering the service system if all the servers are busy. In the delay model, a customer joins a single queue if all the servers are busy at the moment of arrival. Customers are matched with servers with a head-of-the-line discipline as servers become available. The finite waiting room model is the same as the delay model except that a customer who arrives when the waiting room is full is blocked from entry. We give detailed treatments of the loss and delay models, and in §5 we link the results for the delay model to the results for the finite waiting room model.

In the cost structure of the loss model, we assume that at arrival epochs, more customers in the system imply higher expected costs. We also make assumptions which correspond to "faster is cheaper." None of our assumptions concern the comparative intensities of the arrival processes for different classes of customers. These assumptions and uniformizing the resulting continuous-time Markov decision chain lead to preliminary results. They in turn imply that the FAS policy minimizes the expected discounted cost during all horizons, finite and infinite regardless of the discount factor. Hence, the FAS policy minimizes the long-run average cost per unit time.

In the delay model, the uniformized continuous-time Markov decision chain bears similarities to its counterpart in the loss model because non-trivial assignment decisions are made in systems with two or more free servers. So essentially the same cost structure assumptions lead to the same conclusion: the FAS policy minimizes the expected discounted cost during all finite horizons and the infinite horizon, and, therefore, it minimizes the long-run average cost per unit time.

Our results consist of sufficient, but not necessary, conditions on the cost structure for FAS to be optimal. In Appendix A we give an example of a cost structure that satisfies the sufficient conditions. In §5 we comment on some features of cost structures which would *not* necessarily satisfy the sufficient conditions for optimality of FAS and for which the FAS policy may be sub-optimal. These include customer class-specific holding cost rates and server-specific costs per unit time that depend on whether a server is active or idle. Therefore, we believe that necessary *and* sufficient conditions for optimality of FAS, whatever they may be, are not much weaker than our sufficient conditions.

Appendix A. Numerical Example that Satisfies the Sufficient Conditions

Example 1. The following cost structure in the loss model with $J = 3$ customer classes and $K = 2$ servers with $\mu_1 = 4$ and $\mu_2 = 2$ satisfies assumptions (1), (2), and (3):

$$\begin{array}{lllll}
\hat{c}[(0,0),1,1] = 0.3, & \hat{c}[(0,0),1,2] = 3, & \hat{c}[(1,0),1,2] = 4, & \hat{c}[(0,1),1,1] = 1, & \hat{c}[(1,1),1,b] = 2, \\
\hat{c}[(0,0),2,1] = 0.6, & \hat{c}[(0,0),2,2] = 2.5, & \hat{c}[(1,0),2,2] = 3, & \hat{c}[(0,1),2,1] = 3, & \hat{c}[(1,1),2,b] = 1, \\
\hat{c}[(0,0),3,1] = 1.0, & \hat{c}[(0,0),3,2] = 2, & \hat{c}[(1,0),3,2] = 2, & \hat{c}[(0,1),3,1] = 2, & \hat{c}[(1,1),3,b] = 3.
\end{array}$$

Note that for each x and $k \in I(x)$, there is no particular ordering among $\hat{c}[x,1,k]$, $\hat{c}(x,2,k)$, and $\hat{c}(x,3,k)$.

Appendix B. Proofs of Main Results

Lemma 1. *If $x \preceq x'$ then $\theta(x) \geq \theta(x')$ and $x + e_{\theta(x)} \preceq x' + e_{\theta(x')}$.*

Proof. If $x \preceq x'$, then by definition $x - e_i = x' - e_m$ for some $i \leq m$. If $i = m$, the result is trivial. If $i < m$ and at least one server $k < i$ is idle in x (and therefore in x'), then $\theta(x) = \theta(x')$ and it follows that $x + e_{\theta(x)} \preceq x' + e_{\theta(x')}$. The remainder of the proof is for the case $i < m$ with no idle server $k < i$. This implies either $i = 1$ or all servers $k < i$ are busy in both x and x', and so i is the idle server with the smallest index in x', and the index of the fastest idle server in x exceeds i. Therefore $\theta(x) > \theta(x')$. If $\theta(x') = i < \theta(x) = m$, then $x + e_{\theta(x)} = x' + e_{\theta(x')}$. Otherwise $i < \theta(x) < m$ and so $x + e_{\theta(x)}$ and $x' + e_{\theta(x')}$ coincide in all but positions $\theta(x)$ and m with server $\theta(x)$ busy in $x + e_{\theta(x)}$ but idle in $x' + e_{\theta(x')}$ and server m busy in $x' + e_{\theta(x')}$ but idle in $x + e_{\theta(x)}$. Thus by definition, $x + e_{\theta(x)} \preceq x' + e_{\theta(x')}$. □

Lemma 2. *Assumption (1) implies inequality (11).*

Proof. Using definition (6a),

$$\begin{aligned}
& (\alpha + 1)\left(c[x', j, \theta(x')] - c[x, j, \theta(x)]\right) \\
= \; & \hat{c}[x', j, \theta(x')][\alpha + \lambda + M(x' + e_{\theta(x')})] - \hat{c}[x, j, \theta(x)][\alpha + \lambda + M(x + e_{\theta(x)})] \\
= \; & (\alpha + \lambda)\Big(\hat{c}[x', j, \theta(x')] - \hat{c}[x, j, \theta(x)]\Big) \qquad (17) \\
& + \hat{c}[x', j, \theta(x')]M(x' + e_{\theta(x')}) - \hat{c}[x, j, \theta(x)]M(x + e_{\theta(x)}]. \qquad (18)
\end{aligned}$$

Now $x \preceq x'$ corresponds to $x = x' - e_m + e_i$ with $i \leq m$ so $\mu_i \geq \mu_m$. Therefore, (1) implies $\mu_m \hat{c}[x, j, \theta(x)] \leq \mu_i \hat{c}[x, j, \theta(x)] \leq \mu_m \hat{c}[x', j, \theta(x')]$, so

$$\hat{c}[x, j, \theta(x)] \leq \hat{c}[x', j, \theta(x')]. \tag{19}$$

Therefore, the quantity in (17) is nonnegative and

$$\begin{aligned}
& (\alpha+1)\left(c[x',j,\theta(x')] - c[x,j,\theta(x)]\right) \\
\geq\ & \hat{c}[x',j,\theta(x')]M(x'+e_{\theta(x')}) - \hat{c}[x,j,\theta(x)]M(x+e_{\theta(x)}] \\
=\ & \hat{c}[x',j,\theta(x')]\Big(M(x')+\mu_{\theta(x')}\Big) - \hat{c}[x,j,\theta(x)]\Big(M(x)+\mu_{\theta(x)}\Big) && (20) \\
=\ & \hat{c}[x',j,\theta(x')]M(x') - \hat{c}[x,j,\theta(x)]M(x) && (21) \\
& +\hat{c}[x',j,\theta(x')]\mu_{\theta(x')} - \hat{c}[x,j,\theta(x)]\mu_{\theta(x)}. && (22)
\end{aligned}$$

Lemma 1 implies $\mu_{\theta(x')} \geq \mu_{\theta(x)}$, and so quantity (22) is nonnegative due to (19). Therefore,

$$(\alpha+1)\left(\hat{c}[x',j,\theta(x')] - \hat{c}[x,j,\theta(x)]\right) \geq \hat{c}[x',j,\theta(x')]M(x') - \hat{c}[x,j,\theta(x)]M(x). \qquad (23)$$

Let $x_0 = x - e_i = x' - e_m$ which is the vector of servers that are busy in both x and x'. So $M(x) = M(x_0) + \mu_i$ and $M(x') = M(x_0) + \mu_m$. Substituting in (23),

$$\begin{aligned}
& (\alpha+1)\left(\hat{c}[x',j,\theta(x')] - \hat{c}[x,j,\theta(x)]\right) \\
\geq\ & \Big(\hat{c}[x',j,\theta(x')] - \hat{c}[x,j,\theta(x)]\Big)M(x_0) && (24) \\
& +\mu_m\hat{c}[x',j,\theta(x')] - \mu_i\hat{c}[x,j,\theta(x)] \\
\geq\ & \mu_m\hat{c}[x',j,\theta(x')] - \mu_i\hat{c}[x,j,\theta(x)] \geq 0 && (25)
\end{aligned}$$

with the second inequality implied due to (19), and the third due to (1). □

Theorem 1 & Proposition 1. *Assumptions (1), (2), (3), and $h_0(\cdot) \equiv 0$ imply for each n that $k = \theta(x)$ achieves the minimum in* (8a)*; and for all n and $k \in I(x)$ and $x \in X$ with $\#I(x) \geq 1$, inequalities (9) and (10) are true.*

Proof. The unified inductive proof of the theorem and proposition begins by establishing at $n = 0$ that $k = \theta(x)$ achieves the minimum in (8a). From (8a) and $h_0(\cdot) \equiv 0$,

$$f_0(x,j) = min\{c(x,j,k):\ k \in I(x)\}. \qquad (26)$$

From (6a),

$$\begin{aligned}
& (\alpha+1)\Big(c(x,j,k) - c[x,j,\theta(x)]\Big) \\
=\ & \hat{c}(x,j,k)[\alpha+\lambda+M(x+e_k)] - \hat{c}[x,j,\theta(x)][\alpha+\lambda+M(x+e_{\theta(x)})] \\
=\ & \mu_k\hat{c}(x,j,k) - \mu_{\theta(x)}\hat{c}[x,j,\theta(x)] && (27) \\
& +[\alpha+\lambda+M(x)]\{\hat{c}(x,j,k) - \hat{c}[x,j,\theta(x)]\} && (28)
\end{aligned}$$

Since $\theta(x) \leq k \in I(x)$ implies $\mu_{\theta(x)} \geq \mu_k$, assumption (2) implies that (27) and (28) are nonnegative. Therefore, $k = \theta(x)$ achieves the minimum in (8a) at $n = 0$. Inequalities (9) and (10) are valid at $n = 0$ due to $h_0(\cdot) \equiv 0$.

At $n = t$, if $k = \theta(x)$ is optimal in (8a) and inequality (9) is valid, then using (8b) and letting $p \leq q$,

$$\begin{aligned} h_{t+1}(x+e_q) \quad - \quad & h_{t+1}(x+e_p) = \sum_{m \in B(x+e_q)} \mu_m h_t(x+e_q-e_m) - \sum_{m \in B(x+e_p)} \mu_m h_t(x+e_p-e_m) \\ & + \sum_j \lambda_j [f_t(x+e_q,j) - f_t(x+e_p,j)] \\ & + [M(e) - M(x+e_q)] h_t(x+e_q) - [M(e) - M(x+e_p)] h_t(x+e_p) \\ = \quad & \sum_{m \in B(x)} \mu_m [h_t(x+e_q-e_m) - h_t(x+e_p-e_m)] \\ & + (\mu_q - \mu_p) h_t(x) + \sum_j \lambda_j [f_t(x+e_q,j) - f_t(x+e_p,j)] \\ & + \sum_{m \in I(x+e_q)} \mu_m h_t(x+e_q) - \sum_{m \in I(x+e_p)} \mu_m h_t(x+e_p) \end{aligned}$$

Observe that $-h_t(x+e_p)\sum_{m \in I(x+e_p)} \mu_m = -h_t(x+e_p)\sum_{m \in I(x+e_q)} \mu_m + (\mu_p - \mu_q) h_t(x+e_p)$. So

$$\begin{aligned} h_{t+1}(x+e_q) \quad - \quad & h_{t+1}(x+e_p) \\ = \quad & \sum_{m \in B(x)} \mu_m [h_t(x+e_q-e_m) - h_t(x+e_p-e_m)] && (29) \\ & + [h_t(x+e_q) - h_t(x+e_p)] \sum_{m \in I(x+e_q)} \mu_m && (30) \\ & + \sum_j \lambda_j [f_t(x+e_q,j) - f_t(x+e_p,j)] && (31) \\ & + (\mu_p - \mu_q)[h_t(x+e_p) - h_t(x)]. && (32) \end{aligned}$$

Inequality (9) at $n = t$ and $p \leq q$ cause $x+e_p-e_m \preceq x+e_q-e_m$ which implies that (29) and (30) are nonnegative. The quantity in (32) is non-negative due to (10) in the induction hypothesis. Each term in (31) is nonnegative via the following expansion which uses the assumed optimality of $k = \theta(x)$ at $n = t$ in (8a):

$$\begin{aligned} & f_t(x+e_q,j) - f_t(x+e_p,j) \\ = \quad & c[x+e_q, j, \theta(x+e_q)] - c[x+e_p, j, \theta(x+e_p)] \\ & + \beta[h_t(x+e_q+e_{\theta(x+e_q)}) - h_t(x+e_p+e_{\theta(x+e_p)})] \\ \geq \quad & c[x+e_q, j, \theta(x+e_q)] - c[x+e_p, j, \theta(x+e_p)] \geq 0. \end{aligned}$$

Lemma 1, the induction hypothesis, and the assumed validity of inequality (9) at $n = t$ imply the first inequality; Lemma 2 implies the second inequality. This completes the proof of inequality (9) at $n = t + 1$.

The remainder of the theorem's proof shows that assumption (2) and (9) imply that $k = \theta(x)$ minimizes each term of the minimand in (8a) with $n = t + 1$ and, therefore, is optimal in (8a) with $n = t+1$. The term $h_{t+1}(x+e_k)$ is minimized with respect to $k \in I(x)$ by $k = \theta(x)$ due to (9) and $\theta(x) \leq k$ for all $k \in I(x)$. The next paragraph concerns the minimand's term $c(x, j, k)$.

Using definition (6a) and $\mu_{\theta(x)} \geq \mu_k$ for all $k \in I(x)$, assumption (2) implies

$$\begin{aligned} c(x,j,k) - c[x,j,\theta(x)] &= \hat{c}(x,j,k)[\alpha + \lambda + M(x+e_k] - \hat{c}[x,j,\theta(x)][\alpha+\lambda+M(x+e_{\theta(x)})] \\ &= [\alpha+\lambda+M(x)]\Big(\hat{c}(x,j,k) - \hat{c}[x,j,\theta(x)]\Big) \qquad (33) \\ &\quad +\mu_k \hat{c}(x,j,k) - \mu_{\theta(x)}\hat{c}[x,j,\theta(x)] \qquad (34) \end{aligned}$$

The expression in (33) is nonnegative because assumption (2) and $\mu_{\theta(x)} \geq \mu_k$ imply $\hat{c}(x,j,k) \geq \hat{c}[x,j,\theta(x)]$. Assumption (2) implies that the expression in (34) is nonnegative.

Finally, we show that (10) is true when $n = t+1$. Using (8b),

$$\begin{aligned} h_{t+1}(x+e_k) \quad - \quad h_{t+1}(x) &= \sum_{m\in B(x+e_k)} \mu_m h_t(x+e_k-e_m) - \sum_{m\in B(x)} \mu_m h_t(x-e_m) \\ &\quad +\sum_j \lambda_j [f_t(x+e_k,j) - f_t(x,j)] \\ &\quad +[M(e) - M(x+e_k)]h_t(x+e_k) - [M(e)-M(x)]h_t(x) \\ &= \sum_{m\in B(x)} \mu_m [h_t(x+e_k-e_m) - h_t(x-e_m)] \\ &\quad +\sum_j \lambda_j [f_t(x+e_k,j) - f_t(x,j)] \\ &\quad + \sum_{m\in I(x+e_k)} \mu_m h_t(x+e_k) - \sum_{m\in I(x)} \mu_m h_t(x) + \mu_k h_t(x) \\ &= \sum_{m\in B(x)} \mu_m [h_t(x+e_k-e_m) - h_t(x-e_m)] \qquad (35) \\ &\quad + \sum_{m\in I(x+e_k)} \mu_m [h_t(x+e_k) - h_t(x)] \qquad (36) \\ &\quad +\sum_j \lambda_j [f_t(x+e_k,j) - f_t(x,j)] \qquad (37) \end{aligned}$$

The quantities in (35) and (36) are non-negative by (10) in the induction hypothesis, and (37) is non-negative by the following two cases of $\#I(x)$.

- Case 1 [$\#I(x) \geq 2$]: Using assumption (3), (10) in the induction hypothesis, and set inclusion in (8a),

$$\begin{aligned} f_n(x,j) &= min\{c(x,j,i) + \beta h_n(x+e_i) : i \in I(x)\} \\ &\leq min\{c(x,j,i) + \beta h_n(x+e_i) : i \in I(x+e_k)\} \\ &\leq min\{c(x+e_k,j,i) + \beta h_n(x+e_i) : i \in I(x+e_k)\} \\ &\leq min\{c(x+e_k,j,i) + \beta h_n(x+e_k+e_i) : i \in I(x+e_k)\} \\ &= f_n(x+e_k,j) \end{aligned}$$

- Case 2 [$\#I(x) = 1$]: Let i be the idle server. Equation (8a) implies

$$\begin{aligned} f_n(x+e_k, j) - f_n(x, j) &= c(e, j, b) - c(x, j, i) \\ &+ h_n(e) - \beta h_n(x+e_i) \\ &\geq c(e, j, b) - c(x, j, i) \geq 0 \end{aligned}$$

where the first inequality is due to the definition of β and $h_n(e) = h_n(x+e_i)$. The second inequality follows from assumption (3) because the right-hand side of assumption (3) is $\hat{c}(e, j, b)$ in this case.

This completes the proof of the theorem and the proposition. □

Corollary 3. *In the loss model, under the assumptions of Theorem 1, the FAS policy minimizes the long-run average cost per unit time.*

Proof. The proof depends on three observations. First, there are only finitely many states and actions so there is no loss of optimality in confining attention to stationary policies. Second, the long-run average cost per unit time of a stationary policy is the same in $\hat{D}$ and D. Third, the continuous-time discount factor α plays a role in D only via the discount factor $\beta = 1/(1+\alpha)$ in the MDP corresponding to D.

In this MDP, let X_t be the reward in period t. There are only finitely many states and actions so, regardless of the policy, $|E(X_t)| \leq u$ for all t for some $u < \infty$. Let δ denote the FAS policy and let π be any other stationary policy. The proof of the theorem based on the recursion (8a) and (8b) establishes

$$E_\delta \sum_{t=1}^{n} \beta^{t-1} X_t \leq E_\pi \sum_{t=1}^{n} \beta^{t-1} X_t$$

for all $\beta \leq 1$, i.e., for all $\alpha \geq 0$, and all $n = 1, 2, ...$ Let $a_t = E_\delta(X_t)$ and $b_t = E_\pi(X_t)$. Therefore,

$$\frac{1}{n} \sum_{t=1}^{n} a_t \leq \frac{1}{n} \sum_{t=1}^{n} b_t$$

for every n. So

$$\lim \inf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} a_t \leq \lim \inf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} b_t.$$

□

Lemma 3. *For all $x \in X$ with $\#I(x) \geq 1$, if $p \leq q$ with $x_p = x_q = 0$, then $\mu_{\theta(x+e_p)}/\mu_{\theta(x+e_q)} \leq \mu_p/\mu_q$.*

Proof. When $p = q$, the result is trivial. So consider the case of $p < q$ which implies $\theta(x+e_q) < q$ and, therefore, $\mu_{\theta(x+e_q)} \geq \mu_q$. If there is no available server faster than server p, (i.e.

$\theta(x+e_p) > p$), then $\mu_{\theta(x+e_p)} \leq \mu_p$ and so $\mu_p/\mu_q \geq \mu_{\theta(x+e_p)}/\mu_{\theta(x+e_q)}$. Alternatively, if there is an available server faster than p (i.e., $\theta(x+e_p) < p$), then $\theta(x+e_p) = \theta(x+e_q)$. Thus $\mu_p/\mu_q \geq \mu_{\theta(x+e_p)}/\mu_{\theta(x+e_q)}$ because $\mu_p/\mu_q \geq 1$. □

Appendix C. Bounding the Recursion in the Delay Model

The basic idea is to preclude $\hat{c}$ in the delay model from growing too quickly as the number of waiting customers gets larger. We illustrate the idea with a system that is empty at time zero. Let T_n be the time at which the n^{th} customer arrives, and let x^n and j_n be the x-vector at that time (including customer n) and customer n's type. If $N(x^n) \geq K$, let $k_n = b$; otherwise, let k_n label the server to which customer n is assigned. Also, we assume that the costs are bounded above by a geometrically increasing sequence, i.e., $0 \leq \hat{c}(x,j,k) \leq A(1+\eta)^{N(x)}$ (where $0 < A$ and $0 \leq \eta$). This bound would be trivially satisfied by the most likely applications of the model where $\hat{c}(x,j,k)$ would be the difference between nonnegative revenue and cost terms, viz., $\hat{c}(x,j,k) = A_0(j,k) - B(x,j,k)$ with A_0 and B nonnegative at all arguments. Let $A = max\{A_0(j,k) : j = 1,...,J, k = 1,...,K\}$, so $\hat{c}(x,j,k) \leq A(1+\eta)^{N(x)}$ with $\eta = 0$.

With this notation in the delay model,

$$\begin{aligned} \hat{f}_n(x,j) &= E\Big(\sum_{i=1}^{n} e^{-\alpha T_i}\hat{c}(x^i,j_i,k_i)\Big) \leq E\Big(\sum_{i=1}^{\infty} e^{-\alpha T_i}\hat{c}(x^i,j_i,k_i)\Big) \\ &\leq E\Big(\sum_{i=1}^{\infty} e^{-\alpha T_i}A(1+\eta)^i\Big) = A(1+\eta)\sum_{i=1}^{\infty}\left(\frac{\lambda(1+\eta)}{\lambda+\alpha}\right)^{i-1} \end{aligned}$$

which is finite if (and only if) $\lambda < \alpha/\eta$.

This bound is calculated with $N(x^i) \leq i$ and, thus, it ignores the likelihood that customers will be served and leave the system. Less restrictive conditions (but more complicated expressions) for a bound can be obtained by assuming that all customers are served by the slowest server, i.e., $k_i = K$ for all i.

Appendix D. Counter-Example for Sobel (1990)

Our key preliminary results are Lemma 1, which does not depend on the cost assumptions (1)-(3), and Lemma 2, which depends on assumption (1). Sobel (1990), which analyzes only the loss model, neither assumes (1) nor asserts Lemmas 1 and 2. Also, it does not assume (2) or (3). That would raise the possibility that the FAS policy is optimal under a very different and possibly weaker set of assumptions than (1)-(3). However, the following logical gap and counter-example make it unlikely.

It seems that the final inequality in the proof of Theorem 1 in Sobel (1990) cannot be sustained without Lemmas 1 and 2 and we see no way to prove the latter without assuming

(1). Although a gap in a result's proof does not imply that the result is false, the following counter-example confirms that this gap cannot be filled.

The counter-example using the loss model has $J = 2$ customer classes and $K = 2$ servers, so the set of busy-vectors is $X = \{(0,0),(0,1),(1,0),(1,1)\}$. Let $\lambda_1 = 0.4$, $\lambda_2 = 0.1$, $\mu_1 = 0.4$, $\mu_2 = 0.1$, $\hat{c}(\cdot, 1, 2) \equiv 1$, and $\hat{c}(x, j, k) \equiv 0$ if $(j, k) \neq (1, 2)$. This structure violates (1) because

$$\begin{aligned} &\mu_2 \hat{c}[(0,1), 1, \theta((0,1)] - \mu_1 \hat{c}[(1,0), 1, \theta(1,0)] \\ &= \mu_2 \hat{c}[(0,1), 1, 1] - \mu_1 \hat{c}[(1,0), 1, 2] = -0.4. \end{aligned}$$

Since server 1 is faster than server 2, FAS assigns a customer who arrives when the busy-vector is $(0,0)$ to server 1 regardless of the customer's type. Therefore, occasionally a type-2 customer will be assigned to server 1 and soon thereafter a type-1 customer will have to be assigned to server 2. The following calculation shows that the initial assignment of the type-2 customer is sub-optimal because positive costs are incurred only when type-1 customers are assigned to server 2.

An alternative policy, labelled ALT, assigns a type-1 customer who arrives when the busy-vector is $(0,0)$ to server 1, and to server 2 if the customer is type-2. We confirm that the long-run average cost per unit time is lower with ALT than with FAS. This is contrary to the main assertion in Sobel (1990) because this example satisfies the isotonicity assumption made there.

Under either FAS or ALT, let p denote the stationary probability of the busy-vector being $(0,1)$. Since a positive cost is incurred only if a type-1 customer arrives when the busy vector is $(0,1)$ and thus must be assigned to server 2, the long-run average cost per unit time is $\lambda_1 p = 0.4p$. Solving for the stationary probabilities of the continuous-time Markov chains induced by each policy yields $p = 0.139$ with FAS and $p = 0.115$ with ALT, so the long-run average cost per unit time is lower with ALT than with FAS.

References

Ahn, H.-S., M.E. Lewis. 2011. Flexible Server Allocation and Customer Routing Policies for Two Parallel Queues when Service Rates Are Not Additive. Working paper, University of Michigan.

Akşin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6), 665-688.

Andradóttir, S., H. Ayhan, D.G. Down. 2011. Queueing systems with synergistic servers. *Operations Research*, articles in advance.

Armony, M. 2005. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems*, 51(3-4), 287-329.

Blackwell, D. 1967. Positive Dynamic Programming. In L.M. LeCam and J. Neyman, eds., *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, vol. 1, 415-418.

Buyukkoc, C., P. Varaiya, J. Walrand. 1985. The $c\mu$ Rule Revisited. *Advances in Applied Probability*, 17(1), 237-238.

Cooper, R.B. and S. Palakurthi. 1989. Heterogeneous-server loss systems with ordered entry: an anomaly. *Operations Research Letters*, 8(6), 347-349.

Cox, D.R. and W.L. Smith, 1961. *Queues*, Chapman and Hall, London.

de Véricourt, F. and Y.P. Zhou. 2006. On the incomplete results for the heterogeneous server problem. *Queueing Systems*, 52(3), 189-191.

Derman, C. 1962. On Sequential Decisions and Markov Chains. *Management Science*, 9(1), 16-24.

Derman, C., G.J. Lieberman, S.M. Ross. 1980. On the Optimal Assignment of Servers and a Repairman. *Journal of Applied Probability*, 17(2), 577-581.

Green, L. 1984. A Multiple Dispatch Queueing Model of Police Patrol Operations. *Management Science*, 30(6), 653-664.

Harrison, J.M., M.J. López. 1999. Heavy Traffic Resource Pooling in Parallel-Server Systems. *Queueing Systems*, 33(4), 339-368.

Jensen, A. 1953. Markov chains as an aid in the study of Markov processes. *Skandinavisk Aktuarietidskrift*, 36, 87-91.

Katehakis, M.N. 1985. A note on the hypercube model. *Operations Research Letters*, 3(6), 319-322.

Kim, J.H., H.-S. Ahn, R. Righter. 2011. Managing Queues with Heterogeneous Severs. *Journal of Applied Probability*, 48(2), 435-452.

Koole, G. 1992. Stochastic Scheduling and Dynamic Programming. Ph.D. dissertation, Leiden University, Leiden, Netherlands.

Lin, W., P. Kumar. 1984. Optimal Control of a Queueing System with Two Heterogeneous Servers. *IEEE Trans. on Automatic Control*, 29(8), 696-703.

Mandelbaum, A., A.L. Stolyar. 2004. Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized $c\mu$-Rule. *Operations Research*, 52(6), 836-855.

Shanthikumar, J.G. and D.D. Yao. 1987. Comparing ordered-entry queues with heterogeneous servers. *Queueing Systems*, 2(3), 235-244.

Serfozo, R. 1979. An Equivalence Between Continuous and Discrete Time Markov Decision Processes. *Operations Research*, 27(3), 616-620.

Seth, K. 1977. Optimal Service Policies, Just after Idle Periods, in Two-Server Heterogeneous Queuing Systems. *Operations Research*, 25(2), 356-360.

Sobel, M.J. 1982. The Optimality of Full Service Policies. *Operations Research*, 30(4), 636-649.

Sobel, M.J. 1990. Throughput Maximization in a Loss Queueing System with Heterogeneous Servers. *Journal of Applied Probability*, 27(3), 693-700.

van Mieghem, J.A. 1995. Dynamic Scheduling with Convex Delay Costs: The Generalized $c\mu$ Rule. *The Annals of Applied Probability*, 5(3), 809-833.

Williams, R.J. 2000. On Dynamic Scheduling of a Parallel Server System with Complete Resource Pooling. In *Analysis of Communication Networks: Call Centres, Traffic and Performance*, D.R. McDonald and S.R.E. Turner (eds.), American Mathematical Society, Providence, RI, 49-71.