Fair Dynamic Routing in Large-Scale Heterogeneous-Server Systems

Mor Armony
Stern School of Business, New York University, New York, New York 10012, marmony@stern.nyu.edu

Amy R. Ward
Marshall School of Business, University of Southern California, Los Angeles, California 90089, amyward@marshall.usc.edu

In a call center, there is a natural trade-off between minimizing customer wait time and fairly dividing the workload among agents of different skill levels. The relevant control is the routing policy, that is, the decision concerning which agent should handle an arriving call when more than one agent is available. We formulate an optimization problem for a call center with heterogeneous agent pools, in which each pool is distinguished by the speed at which agents in that pool handle calls. The objective is to minimize steady-state expected customer wait time subject to a “fairness” constraint on the workload division.

We first solve the optimization problem by formulating it as a Markov decision process (MDP), and solving a related linear program. We note that this approach does not in general lead to an optimal policy that has a simple structure. Fortunately, the optimal policy does appear to have a simple structure as the system size grows large, in the Halfin-Whitt many-server heavy-traffic limit regime. Therefore, we solve the diffusion control problem that arises in this regime and interpret its solution as a policy for the original system. The resulting routing policy is a threshold policy that determines server pool priorities based on the total number of customers in the system. We prove that a continuous modification of our proposed threshold routing policy is asymptotically optimal in the Halfin-Whitt limit regime. We furthermore present simulation results to illustrate that our proposed threshold routing policy outperforms a common routing policy used in call centers (that routes to the agent that has been idle the longest).

Subject classifications: probability; diffusion: stochastic model applications; queues; approximations; diffusion models; limit theorems.

Area of review: Stochastic Models.

History: Received January 2008; revisions received December 2008, June 2009; accepted July 2009. Published online in Articles in Advance February 26, 2010.

1. Introduction

Server heterogeneity is ubiquitous in large-scale service systems. Even when customers are homogeneous in their service requests, different employees have different skill levels and handle customer requests at different speeds. For example, experienced employees, on average, process customers faster than newly hired employees.

A natural question arises in service systems with heterogeneous servers: When a customer arrives and more than one server is available, which server should serve him? Of course, the customer prefers the fastest available server. However, if faster servers always receive priority, then the faster servers will experience a heavier workload than the slower servers. In fact, as the number of servers becomes large and the arrival rate approaches the service capacity, the faster-server-first (FSF) policy asymptotically minimizes expected customer waiting time but also asymptotically only allows slower servers to idle (Armony 2005). Hence prioritizing faster servers does not evenly distribute idle time between servers.

Do service organizations care that the FSF policy is unfair to the faster servers? First, it is generally acknowledged in the organizational behavior and human resource management literature that perceived injustice among employees leads to low employee satisfaction and hampers performance; see, for example, Colquitt et al. (2001) and Cohen-Charash and Spector (2001). Furthermore, high employee satisfaction implies increased employee retention, and Whitt (2006) shows that increased employee retention improves service. Finally, the appeal of additional idle time for relaxation might provide faster servers with an incentive for slowing their service rate, which would increase customer service and waiting times.

Call centers provide a strong motivating example of a service organization that cares about the issue of server fairness. In particular, many call centers follow a longest-idle-server-first (LISF) routing policy; that is, newly arriving calls are routed to the server that has experienced the longest idle time. The LISF policy is “fairer” than the FSF policy in the following asymptotic sense. Consider the inverted-V model shown in Figure 1, with exponential interarrival and service times, \(K\) server types distinguished by their service rates \(\mu_k, k \in \{1, 2, \ldots, K\}\), and \(N_k\) servers...
of each type. For this system, the type \( k \) server idleness proportion, by which we mean the steady-state proportion of idle servers that are of type \( k \), is asymptotically \( N_k \mu_k / (N_1 \mu_1 + \cdots + N_K \mu_K) \) as the number of servers becomes large, and the arrival rate \( \lambda \) approaches the service capacity (Atar 2008). In particular, idleness is shared proportionally among the \( K \) server pools.

The question that then arises concerns the performance of the LISF policy as measured by expected customer waiting time. Specifically, how much longer is expected customer waiting time under the LISF policy as compared to the FSF policy? Furthermore, does another policy exist that achieves the same server idleness proportion as the LISF policy, and also has a lower expected customer waiting time? Intuitively, in the case of two server types, a threshold policy that routes according to a FSF policy when the number of customers in the system is large (i.e., above some threshold level) and routes according to a SSF (slower-server-first) policy when the number of customers in the system is small (i.e., below some threshold level) should have a lower expected customer waiting time than the LISF policy when the threshold level is set to achieve the same server idleness proportions.

Figure 2 presents the results of a simulation study that compares steady-state expected customer waiting time under the threshold and LISF policies. The system simulated has parameters \( N_1 = N_2 = 100 \) and \( \mu_1 = 1 \). The speed at which the faster servers serve \( \mu_2 \) is varied, and the arrival rate \( \lambda \) was adjusted according to Halfin-Whitt limit regime. The performance of the FSF policy is presented for comparison purposes. We record the mean number of customers waiting and the mean slow server idleness proportion for 100 runs, where each run has a 100,000 arrival "warm-up" period (in which statistics are not recorded), and then 500,000 subsequent arrivals (for which statistics are recorded). We then report the average of the number of customers waiting and the mean slow server idleness proportion over the 100 runs.

Notice in Figure 2(a) that the expected waiting time under the threshold policy is consistently lower than under the LISF policy. Specifically, for the higher values of \( \mu_2 \), the expected customer waiting time is approximately 5%-6% higher under the LISF policy as compared to the threshold policy. (In comparison to the FSF policy, the expected customer waiting time under the threshold policy is approximately 20% higher.) Furthermore, as displayed in Figure 2(b) the slow server idleness proportions for the threshold policy and the LISF policy are approximately equal (within 0.02), whereas under the FSF policy the slow servers experience over 90% of the total idle time. In other words, the simulation study supports the intuition that a threshold policy can both have a better performance than the LISF policy and achieve the same server idleness proportions.

Our objective in this paper is to find a policy that minimizes the steady-state customer waiting time subject to any given proportions that specify how idleness is shared between the \( K \) server pools. (There is no reason to restrict ourselves to the idleness proportions attained by the LISF policy.) We first solve this problem by formulating it as
a Markov decision process (MDP) and numerically solving a related linear program. We note that characterizing a simple form for the optimal policy for this problem is hard. Furthermore, most call centers have hundreds or thousands of agents, and the time required to solve the MDP numerically becomes prohibitively long as the number of agents becomes large. Therefore, also noting that many call centers operate in a regime in which the arrival rate and service capacity are close, we consider the many-server heavy-traffic regime in Halfin and Whitt (1981). The diffusion control problem that arises in this regime is analytically tractable, and we solve this explicitly to find that a threshold control specified via \( K - 1 \) threshold levels is optimal for the diffusion control problem. We then propose a threshold policy for the original setting using the threshold parameters obtained from the diffusion control problem, and we prove that a continuous modification of that threshold policy is asymptotically optimal.

The remainder of this paper is organized as follows. We first review relevant literature. In §2, we present our basic model formulation. We formulate our problem as a MDP, and solve it numerically in §3. We construct and solve an approximating diffusion control problem in §4. In §5, we propose a policy for the original system that is based on the solution to the approximating diffusion control problem, and in §6 we show that our proposed policy is asymptotically optimal in the Halfin-Whitt many-server heavy traffic regime. Finally, we make concluding remarks in §7.

The e-companion contains technical details omitted from the main paper body and the proofs of all results stated in the main paper body.

1.1. Literature Review

Fairness in queueing systems has been a topic of interest to researchers and practitioners alike for a while. Especially, the fairness among flows in telecommunication and computer networks has gotten a lot of attention over the years. More recently, researchers have studied fairness in queues from the point of view of individual customers. Two recent overviews of this line of research are Avi-Itzhak et al. (2010) and Wierman (2007). Interestingly, fairness among servers in multiserver queueing systems has gotten relatively little attention in the literature. This is surprising given the strong indication that fairness matters to organizations due to its effect on employee performance and overall satisfaction, as evident by the HRM literature (e.g., Colquitt et al. 2001, Cohen-Charash and Spector 2001) and the practice at many call centers to use fair policies such as LISF. Three papers that do address the server fairness issue are Lu et al. (2007) (in the context of bandwidth allocation in telecommunication networks), Cabral (2007), and Tseytlin (2007). Cabral (2007) examines the question of which servers work more in a heterogeneous server system with equally likely random customer-server assignments among idle servers, and concludes that the effective service rate of the faster servers is higher but that also the faster servers enjoy a larger proportion of the idle time. Tseytlin (2007) discusses fairness in the context of patient allocation to internal wards in a hospital and shows that a policy that balances idleness by routing patients to a server pool with probability that is proportional to the pool’s idleness ratio is fair in the sense that faster servers tend to rest more than slower servers.

The question of how to route customers to servers in a heterogeneous server system in order to minimize a given performance measure such as the mean wait time is also well studied. However, it is a difficult problem, and so most exact analysis has been for a two-pool system (e.g., Rubinovich 1983, Larsen and Agrawala 1983, Lin and Kumar 1984, and Stockbridge 1991). Recent results have appeared for the general heterogeneous multiserver system case (e.g., Cabral 2005 and de Vericourt and Zhou 2006), but it appears that the full exact solution to this problem in the more general setting is still open. Hence researchers have been prompted to study the problem in various asymptotic regimes, including both conventional heavy traffic (e.g., Harrison 1998, Bell and Williams 2001, Glazebrook and Nino-Mora 2001, Teh and Ward 2002, Mandelbaum and Stolyar 2004, Bell and Williams 2005, Stolyar 2005) and the Halfin-Whitt many-server limit regime (e.g., Atar et al. 2004; Armony 2005; Atar 2005; Gurvich and Whitt 2009a, b, 2010; Tezcan 2007; Atar 2008; Atar and Shwartz 2008; Armony and Mandelbaum 2010; Atar et al. 2009). Some of these papers look at the more general skill-based routing problem in which there are multiple server pools and multiple customer classes and abandonments. However, with the exception of Atar (2008) and Atar et al. (2009), none of these papers deals with the interserver fairness issue. The main difference between this paper and those two papers is that this paper formulates an optimization problem, whereas those two papers analyze a single policy (the longest idle server first and the longest idle pool first, respectively).

This paper integrates the two aforementioned literature streams by considering the question of how to dynamically route customers to minimize mean wait time, but still be fair to the servers.

2. Model Formulation

Consider a service system with a single customer class and \( K \geq 2 \) server types (each type in its own server pool), all capable of fully handling customers’ service requirements. Let \( \mathcal{K} = \{1, \ldots, K\} \). Service times are independent and exponential, and the average service time of a customer served by a server from pool \( k, k \in \mathcal{K}, \) is \( 1/\mu_k \).

There are \( N_k \) servers in pool \( k \). We let \( N = \sum_{k \in \mathcal{K}} N_k \) be the total number of servers, and \( \hat{x} = (N_1', N_2', \ldots, N_K') \) denote the staffing vector. Here and elsewhere, \( \hat{x} \) is used to denote a vector whose elements are \( x_1, x_2, \ldots \). We assume \( \mu_1 < \mu_2 < \cdots < \mu_K \), meaning that server pools are ordered with respect to their service speed.
Customers arrive to the system according to an independent Poisson process with rate $\lambda$. Let $\tilde{\mu} := \sum_{k \in \mathbb{N}} N_k \mu_k$ be the total service capacity of the system when all servers are busy. We assume the following necessary condition for stability is satisfied: $\rho := \lambda / \tilde{\mu} < 1$; that is, the total service capacity is larger than the arrival rate. Delayed customers wait in an infinite buffer and are served according to a FCFS discipline. Customers that arrive to a system in which all pools have idle servers must be routed to a server of a specific pool. We would like to route customers in a way that minimizes their steady-state expected waiting time subject to a fairness constraint on the steady-state server idleness proportions.

Denote by $\pi := (\lambda, \tilde{N})$ a policy that operates in a system with arrival rate $\lambda$ and staffing vector $\tilde{N}$ (at times we will omit the arguments $\lambda$ and $\tilde{N}$ when it is clear from the context which arguments should be used). Let $t \geq 0$ be an arbitrary time point. We denote by $Z_k(t; \pi)$ the number of busy servers of pool $k$, $k \in \mathbb{N}$, at time $t$, and $Q(t; \pi)$ the queue length at this time. Also, let $X(t; \pi)$ be the total number of customers in the system. That is,

$$X(t; \pi) = Q(t; \pi) + \sum_{k \in \mathbb{N}} Z_k(t; \pi).$$

Let

$$I_k(t; \pi) = N_k - Z_k(t; \pi)$$

be the number of idle servers in pool $k$, $k \in \mathbb{N}$, with

$$I(t; \pi) = \sum_{k \in \mathbb{N}} I_k(t; \pi)$$

the total number of idle servers. Finally, let $W(t; \pi)$ be the waiting time that would be experienced by a customer arriving at time $t$.

We omit the time argument when we refer to the entire process. We use $t = \infty$ whenever we refer to a process in steady-state. Also, we omit $\pi$ from the notation unless it is necessary to avoid confusion between different routing policies.

Let $\Pi$ be the set of all nonanticipating, nonpreemptive, nonidling policies under which a unique steady-state for $X$, $Q$, and $Z_k$, $k \in \mathbb{N}$, exists (which implies a unique steady-state exists for $I_k$, $k \in \mathbb{N}$, and $W$ as well). Nonanticipating (roughly speaking) means that the policy cannot require knowledge of the future. By nonpreemptive, we mean that once a call is assigned to a particular server, it cannot be transferred to another server of a different pool, nor can it be preempted by another call. A policy is nonidling if there are no idle servers whenever there are some delayed customers in the queue; in other words, for any $t \geq 0$, $Q(t) > 0$ implies that $\sum_{k \in \mathbb{N}} Z_k(t) = N$.

The problem we would like to solve is as follows:

$$\text{minimize } \quad EW(\infty; \pi),$$

subject to

$$\frac{EI_k(\infty; \pi)}{EI(\infty; \pi)} = f_k, \quad k \in \mathbb{N},$$

where $\sum_{k=1}^{K} f_k = 1$ and $0 < f_k < 1$ is the target steady-state fraction of pool $k$ idleness. Specifically, given fixed values of $\tilde{\mu}$, $\lambda$, and $\tilde{N}$, one needs to find a policy $\pi = (\pi(\lambda, \tilde{N})) \in \Pi$ that minimizes the expected steady-state waiting time subject to the constraint in (1), which we henceforth refer to as “the fairness constraint.”

It is worthwhile to remark on the nonidling assumption. The assumption is natural in practice, because call centers do not generally keep customers waiting when there are idle servers, and the assumption is common in the literature (see, for example, Atar 2005 and Atar et al. 2004). However, its formal justification is far from obvious. This is because one might intentionally choose to idle servers even when there is work to be done in order to ensure that the fairness constraint is met (nonidling has been shown to be suboptimal in other call-center related settings in the presence of constraints; see Bhulai and Koole 2003, Gans and Zhou 2003, Gurvich et al. 2008, Armony and Gurvich 2009). We do not address this question further in this paper, and we leave it as an interesting direction for future research.

The most obvious question that arises from the problem formulation (1) is how to determine the idleness fraction parameters $f_1, \ldots, f_K$. There are various factors a manager might wish to consider. First, it is intuitive that the expected waiting time should be decreasing in the pool 1 idleness proportion $f_1$, and increasing in the pool $K$ idleness proportion $f_K$. Hence, one would obviously tend to choose higher values of $f_1$ and lower values of $f_K$. But how would this choice affect system fairness? Moreover, how does the choice of $f_k$ for $k \neq 1, K$ affect expected waiting times?

The full answer to the question of how to choose $f_1, \ldots, f_K$ is outside the scope of this paper. However, our analysis does provide guidance on how to choose those fractions, because it allows us to estimate the percentage increase in steady-state wait time over the FSF policy, as a function of $f_1, \ldots, f_K$, under a policy that minimizes the objective function in (1). For a system with two server pools (so that it is only necessary to specify $f_1$ because $f_2 = 1 - f_1$) and large $N_1$ and $N_2$ (in a way that is made precise in Assumptions (A1) and (A2) in the first paragraph of §4), Figure 3 plots this estimated increase. In particular, for any target steady-state fraction of idleness for the slow-server pool, Figure 3 estimates the minimum possible percentage increase in steady-state expected wait time over the FSF policy. The system manager can then decide how he would like to trade off server fairness and expected customer wait time.

3. The Markov Decision Process

In this section, we formulate the fairness problem as a Markov decision process (MDP) and establish that it can be solved using a linear programming (LP) framework. Using numerical experimentation we observe that the optimal policy has a switching curve structure, where, roughly
The predicted minimum percentage increase in steady-state wait time over the FSF policy as a function of $f_1$, when $K = 2$, $\mu_1 = 1$, $\mu_2 = 2$, and $N_1$ and $N_2$ are large and satisfy $\sum_{k=1}^2 \mu_k N_k - \lambda \approx \sqrt{\lambda}$.

The above observations imply that we have a constrained continuous-time Markov decision process with a finite state space and finite action space. Because the maximum total rate for exiting any state is uniformly bounded by $\lambda + \mu_k N$, we can use the uniformization technique (see, for example, §11.5 in Puterman 1994). Then we have a constrained discrete-time Markov decision process with a finite state space and a finite action space. It follows from Altman (1999) that there exists an optimal, stationary policy and, noting that the objective function and the fairness constraints in (1) are linear in the steady-state probabilities of the state-action pairs, that this policy can be computed by solving an LP. We show this LP formulation in (EC.1) in the e-companion.

To study the problem numerically, we used Matlab to program the relevant linear program and ran the program over a large set of parameter values. The solution of two examples is depicted in Figure 4. The figure on the left shows the solution for the case $K = 2$, $N_1 = N_2 = 10$, $\mu_1 = 1$, $\mu_2 = 2$, $\lambda = 27$, and $f_1 = 0.5$. The figure on the right shows the solution for the case $K = 2$, $N_1 = 20$, $N_2 = 30$, $\mu_1 = 2$, $\mu_2 = 3$, $\lambda = 110$, and $f_1 = 0.6$. In both cases the optimal solution follows a switching curve form, where FSF is used when the system is handling a relatively high number of calls, and SSF is used when there are fewer calls to handle. There is randomization in one state. We observe this type of switching curve with randomization in one state when $K = 2$ to be typical in the optimal solution across all the examples that we ran.

The switching curves in Figure 4 are not linear but are almost linear. Furthermore, the linear form emerges as the system size grows large. In the case that the linear form is a 45° linear switching curve, the optimal policy is exactly a simple threshold policy for routing customers that determines server pool priorities based on the total number of customers in the system. These observations motivate us to focus our attention on showing that a threshold routing scheme is near optimal as the system size grows large.

Our approach is to solve the routing problem (1) in the Halfin-Whitt heavy traffic asymptotic regime (also referred to as the QED regime). In this regime, the arrival rate $\lambda$ and the service capacity $\bar{\mu}$ become large and are close. More specifically, following the general approach outlined by Harrison (1988), we will solve the diffusion control problem that arises when formally passing to the limit in the control problem having arrival rate $\lambda$ (§4), interpret its solution as a routing policy in the original system (§5), and prove that that routing policy’s performance is asymptotically optimal (§6). The solution to the diffusion control problem provides a lower bound on the original problem (1) in the Halfin-Whitt limit regime, after appropriately scaling the objective function.

The appeal of this approach is that it motivates using a simple threshold policy for routing customers that determines server pool priorities based on the total number of customers in the system. This is true not only when there
are two server pools, but also when there is an arbitrary number of server pools. This is because a threshold policy is optimal for the aforementioned diffusion control problem, regardless of the number of server pools.

Our analysis shows that the threshold levels in the threshold policy we propose should be of order \( \sqrt{N} \). Furthermore, their exact values can be found according to a fast search procedure that scales in the number of server pools. This is important because call centers, in practice, typically have many agents but a small number of server pools. In contrast, by counting the number of variables and constraints in the LP that arises from the MDP (see (EC.1), the electronic companion), we observe that the problem is polynomial in \( N \) and exponential in \( K \). Our experiments show that even when \( K = 2 \) and the system is of a moderate size (number of servers of each type is 20 or more) the program takes 30 minutes or more to run. This makes the MDP approach impractical to use for large systems, especially if the optimal control needs to be updated frequently.

### 4. The Diffusion Control Problem

In this section, we solve the diffusion control problem that arises in the Halfin-Whitt limit regime when formally passing to the limit in the control problem (1). We first specify the asymptotic framework and then present the diffusion control problem that arises in this framework. We then devote §§4.1–4.3 to solving this diffusion control problem.

We consider a sequence of systems indexed by \( \lambda \) with increasing arrival rates \( \lambda \uparrow \infty \), and increasing total number of servers \( N^{A} \) but with fixed service rates \( \mu_{1}, \ldots, \mu_{K} \). Our convention is to superscript any process or quantity associated with the system having arrival rate \( \lambda \) by \( \lambda \). The assumptions we require are as follows.

(A1) There is heavy traffic; specifically, the number of servers in each pool satisfies

\[
\lim_{\lambda \to \infty} \frac{\mu_{k} N_{k}^{A}}{\lambda} = a_{k} \quad \text{for each } k \in \mathcal{K}, \quad \text{where } a_{k} > 0 \quad \text{and} \quad \sum_{k=1}^{K} a_{k} = 1.
\]

Note that (A1) guarantees that \( \rho^{A} = \lambda/\mu^{A} = \sum_{k=1}^{K} a_{k} N_{k}^{A} \to 1 \) as \( \lambda \to \infty \). It is also true that under (A1), \( \lambda N^{A} \to \mu := [\sum_{k=1}^{K} a_{k} \mu_{k}]^{-1} \) as \( \lambda \to \infty \) and the limiting fraction of busy servers in each pool satisfies

\[
N_{k}^{A}/N^{A} \to q_{k} := (a_{k}/\mu_{k})\mu \quad \text{as } \lambda \to \infty \text{ for each } k \in \mathcal{K}.
\]

(A2) There is square-root safety staffing; specifically,

\[
\lim_{\lambda \to \infty} \frac{\sum_{k=1}^{K} \mu_{k} N_{k}^{A} - \lambda}{\sqrt{\lambda}} = \delta \quad \text{for some finite } \delta > 0.
\]

The condition \( \delta > 0 \) guarantees that the system is stable (or can be stable, under reasonable routing) for all \( \lambda \) large enough. Moreover, the fraction of delayed customers is less than 1 (see Armony 2005). Note that (A2) does not specify how the added safety staffing is divided between server pools.

We are now in a position to specify the diffusion control problem that arises under Assumptions (A1) and (A2), when formally passing to the limit in the control problem (1), as the arrival rate \( \lambda \) increases without bound. Let the process \( \hat{X} \) solve the stochastic integral equation

\[
\hat{X}(t) = \hat{X}(0) + \int_{0}^{t} m(\hat{X}(s), u(\hat{X}(s)^{-})) \, ds + \sqrt{2}\mu B(t), \quad (3)
\]
where $B$ is a standard Brownian motion, $E \hat{X}(0)^2 < \infty$, the control
\[
(u_1(x), \ldots, u_K(x)) \in \mathcal{U}
\]
\[
:= \left\{(u_1, \ldots, u_K): 0 \leq u_k \leq 1 \text{ for all } k \in \mathcal{K} \text{ and } \sum_{k=1}^{K} u_k = 1 \right\}
\]
for all $x \in [0, \infty)$, \hspace{1cm} (4)
and
\[
m(x, u) = -\delta \sqrt{\mu} + \sum_{k=1}^{K} \mu_k u_k(x^-) x^-.
\]

Intuitively, the process $\hat{X}$ approximates the centered and scaled number of customers in the system, $(X^\lambda - N^\lambda)/\sqrt{N^\lambda}$, for large $\lambda$. Then, $[\hat{X}(t)]$ approximates the scaled total number of idle servers, so the infinitesimal rate at which the diffusion drifts upward, $\sum_{k=1}^{K} \mu_k u_k(X^-) \hat{X}(t)^-$, depends on how the control divides the idle servers between the different pools.

The key to solving the diffusion control problem in (6) is to formulate a Lagrangian relaxation problem. Because this Lagrangian relaxation problem is a control problem without constraints, we are able to apply the standard methods in, for example, Fleming and Soner (1993) or Kushner and Dupuis (2001), to characterize the optimal control and optimal objective function value through $K + 1$ second-order differential equations. However, solving for the optimal control is tricky, because it will turn out that the optimal control has a discontinuous drift, which means we must smoothly paste together the solutions of these differential equations.

Let $\Delta_1, \ldots, \Delta_K \in \mathcal{K}$ be Lagrange multipliers. A Lagrangian relaxation problem is
\[
\text{minimize } E\hat{X}(\infty)^+ + \sum_{k=1}^{K} \Delta_k [E[u_k(\hat{X}(\infty)^-)\hat{X}(\infty)^-] - f_k d] \ni \in \mathcal{U}_p
\]
\[
= -d \sum_{k=1}^{K} f_k \Delta_k + \text{minimize } E\hat{X}(\infty)^+ + \sum_{k=1}^{K} \Delta_k E[u_k(\hat{X}(\infty)^-)\hat{X}(\infty)^-]. \hspace{1cm} (7)
\]

The objective function follows because by Little’s law the expected waiting time is proportional to the number of customers waiting in queue. The constraints follow because $\hat{X}(\infty)^-$ has approximately the same distribution as the scaled steady-state number of idle servers in the discrete event system when $\lambda$ (and therefore $N^\lambda$ also) is large.

**4.1. The Solution Approach and the Lagrangian Relaxation Problem**

It is more convenient for our analysis to look at an equivalent approximating diffusion control problem. The following proposition establishes that the constraints in (5) imply that the value of $E\hat{X}(\infty)^-$ is fixed, which is a simplification.

**PROPOSITION 1.** The diffusion control problem,
\[
\text{minimize } E\hat{X}(\infty)^+,
\]
subject to $E[u_k(\hat{X}(\infty)^-)\hat{X}(\infty)^-] = f_k d,$ \hspace{1cm} (6)
for all $k \in \mathcal{K},$
has the same optimal solution and optimal objective function value as (5), where
\[
d := \frac{\delta \sqrt{\mu}}{\sum_{k=1}^{K} \mu_k f_k}.
\]

The key to solving the diffusion control problem in (6) is to formulate a Lagrangian relaxation problem. Because this Lagrangian relaxation problem is a control problem without constraints, we are able to apply the standard methods in, for example, Fleming and Soner (1993) or Kushner and Dupuis (2001), to characterize the optimal control and optimal objective function value through $K + 1$ second-order differential equations. However, solving for the optimal control is tricky, because it will turn out that the optimal control has a discontinuous drift, which means we must smoothly paste together the solutions of these differential equations.

Let $\Delta_1, \ldots, \Delta_K \in \mathcal{K}$ be Lagrange multipliers. A Lagrangian relaxation problem is

**DEFINITION 1.** A threshold control policy is defined by a $K - 1$ dimensional vector $L := (L_1, \ldots, L_{K-1})$ having $L_0 := 0 < L_1 < L_2 < \cdots < L_{K-1} < L_K := \infty$, and has $u_k(x) = 1 \{L_{k-1} \leq x < L_k\},$ for all $x \in [0, \infty)$ and $k \in \mathcal{K}$. 

Armony and Ward: Fair Dynamic Routing
Note that a threshold control policy is obviously in \( \mathcal{U}_p \). It is also useful to note that the infinitesimal drift associated with threshold control at levels \( \hat{L} \) is

\[
m_k(x) = \begin{cases} 
-\delta \sqrt{\mu} 
& x \geq 0 \\
-\delta \sqrt{\mu} - x \sum_{k=1}^{K} \mu_k 1\{L_{k-1} \leq -x < L_k\} 
& x < 0.
\end{cases} \tag{8}
\]

Fix \( f_1, \ldots, f_K \). Our solution approach is as follows:

1. Show that for any threshold control, there exist Lagrange multipliers \( \triangle_1, \ldots, \triangle_K \) such that that threshold control policy is optimal for the Lagrangian relaxation problem under those Lagrange multipliers (Lemmas 2–4 and inequality (19)).

2. Find the threshold levels \( L_1^* < \cdots < L_{K-1}^* \) that are feasible for the diffusion control problem (4.3) (Lemma 5).

3. Conclude that there exist Lagrange multipliers \( \triangle_1, \ldots, \triangle_K \) under which threshold control at levels \( L_1^*, \ldots, L_{K-1}^* \) is optimal for the Lagrangian relaxation problem, and show that that implies threshold control at the levels \( L_1^*, \ldots, L_{K-1}^* \) is optimal for the diffusion control problem (Theorem 1).

4. When calculating the optimal policy, it is important to note that the Lagrange multipliers \( \triangle_1, \ldots, \triangle_K \) need never actually be computed to solve the diffusion control problem (6). It is only the threshold levels \( L_1^*, \ldots, L_{K-1}^* \) that must be found.

### 4.2. Threshold Control

One reason that threshold policies are attractive is that the steady-state density of \( \check{X} \) under a threshold control policy has a simple form. In particular, it follows from Browne and Whitt (1995) that the steady-state density is

\[
g(x) = \begin{cases} 
b_0 g_0(x) 
& x \geq 0 \\
b_k g_k(x) 
& -L_k \leq x < -L_{k-1}, \quad k \in \mathbb{K}.
\end{cases} \tag{9}
\]

where

\[
g_0(x) = \frac{\delta}{\sqrt{\mu}} \exp\left(-\frac{\delta x}{\sqrt{\mu}}\right) \quad \text{and} \\
g_k(x) = \frac{\sqrt{\mu_k}}{\sqrt{\mu}} \Phi\left(\frac{\delta}{\sqrt{\mu_k}} + \frac{\sqrt{\mu_k}}{\sqrt{\mu}} x\right)
- \Phi\left(\frac{\delta}{\sqrt{\mu_k}} - \frac{\sqrt{\mu_k}}{\sqrt{\mu}} x\right),
\]

and the constants \( b_k \) satisfy

\[
b_k g_k(-L_{k-1}) = b_{k-1} g_{k-1}(-L_{k-1}), \quad k \in \mathbb{K} \quad \text{and} \quad \sum_{k=0}^{K} b_k = 1.
\]

This means that for any fixed threshold levels \( 0 < L_1 < L_2 < \cdots < L_{K-1} \) we can directly calculate the value of the objective function in (7) and then minimize over the threshold levels. However, that strategy would not show that there exists a threshold control policy that minimizes the objective function in (7) over the larger class of admissible policies. For this, we require diffusion control methodology.

The first step is to characterize the value of the objective function in (7) under a threshold control using second-order differential equations.

**Lemma 2.** Suppose there exists a twice-continuously differentiable function \( V : \mathbb{R} \to \mathbb{R} \) and constants \( c \in \mathbb{R}, 0 < \triangle_k, \ k \in \mathbb{K} \), that solve

\[
\mu V''(x) - \delta \sqrt{\mu} V'(x) + x = c, \quad x \geq 0 \\
\mu V''(x) - (\delta \sqrt{\mu} + \mu_k x) V'(x) - \triangle_k x = c, \quad -L_k \leq x \leq -L_{k-1}, \quad k \in \mathbb{K}.
\tag{10}
\]

Also assume there exist \( b_1, b_2 \in \mathbb{R} \) such that \( |V(x)| \leq b_1 x^2 + b_2 \) for all \( x \in \mathbb{R} \). Then, if \( \check{X} \) satisfies (3) under the threshold control at levels \( L_1, \ldots, L_{K-1} \),

\[
E \check{X}(\infty)^+ + \sum_{k=1}^{K} \Delta_k E[\mu_k (\check{X}(\infty)^-) \check{X}(\infty)^-] = c.
\]

The next step is to construct the solution to (10) that satisfies the conditions of Lemma 2. In doing this, we regard the threshold levels as fixed and assume the Lagrange multipliers \( \triangle_k, k \in \mathbb{K} \) can vary.

The general solution to (10) is

\[
V(x) = \int_{-\infty}^{x} V'(y) dy, \quad \text{where}
\]

\[
V'(x) = \begin{cases} 
V'_0(x) 
& x \geq 0 \\
V'_k(x) 
& x \in [-L_k, -L_{k-1}), \quad k \in \mathbb{K},
\end{cases}
\]

and, for \( \alpha_0, \ldots, \alpha_K \in \mathbb{R} \) constants,

\[
V'_0(x) = \alpha_0 \exp\left(\frac{\delta}{\sqrt{\mu}} x + \frac{x}{\delta \sqrt{\mu}} + \frac{1}{\delta^2} - \frac{c}{\delta \sqrt{\mu}}\right),
\]

\[
V'_k(x) = \exp\left(\frac{1}{2} \frac{\mu_k x}{\sqrt{\mu_k}} + \frac{\delta}{\sqrt{\mu_k}}\right) \left(\alpha_k + \frac{\sqrt{2\pi}}{\sqrt{\mu_k}} \frac{c}{\delta \sqrt{\mu_k}} - \frac{\Delta_k}{\mu_k}\right)
\cdot \left(\frac{\delta}{\sqrt{\mu_k}} - \frac{\Delta_k}{\mu_k}\right) - \frac{\Delta_k}{\mu_k}, \quad k \in \mathbb{K}.
\tag{11}
\]

So the growth rate of \( V' \) is no more than linear, so the growth of \( V \) is no more than quadratic, we must have \( \alpha_0 = \alpha_k = 0 \). Furthermore, for the continuity of \( V' \) and \( V'' \), we need

\[
V'_k(-L_k) = V'_{k+1}(-L_k), \quad \text{and}
\]

\[
V'_k(-L_k) = V''_{k+1}(-L_k) \quad \text{for all } k \in \{1, \ldots, K-1\},
\]

for which it is sufficient to have

\[
V'_k(0) = \frac{1}{\delta^2} - \frac{c}{\delta \sqrt{\mu}};
\]

\[
V'_k(-L_k) = \frac{\Delta_k - \Delta_{k+1}}{\mu_{k+1} - \mu_k} \quad \text{for all } k \in \{1, \ldots, K-1\},
\tag{12}
\]
and

\[ V'_k(-L_k) = V'_{k+1}(-L_k) \quad \text{for all } k \in \{1, \ldots, K-1\}. \quad (13) \]

Note that it is straightforward to check that conditions (12) and (13) imply \( V'_k(-L_k) = V'_{k+1}(-L_k) \) for all \( k \in \{1, \ldots, K-1\} \) using the fact that

\[ V'_0(x) = \frac{\delta}{\sqrt{\mu}} \alpha_0 \exp\left( \frac{\delta}{\sqrt{\mu}} + \frac{1}{\delta \sqrt{\mu}} \right), \quad \text{and} \]

\[ V'_k(x) = \left( \frac{\mu_k}{\mu} x + \frac{\delta}{\sqrt{\mu}} \right) V'_k(x) + \frac{c_k}{\mu} V'_k(x), \quad k \in \mathcal{K}. \]}

We now use conditions (12) and (13) to determine which particular solution for \( V' \) that satisfies (11) also satisfies the conditions of Lemma 2. Condition (12) implies that, given fixed values of \( \alpha_1, \ldots, \alpha_K \),

\[ V'_k(x) = \frac{\delta \sqrt{2 \pi}}{\sqrt{|\mu_k|}} \left( \frac{c}{\sqrt{|\mu|}} - \frac{\Delta_k}{|\mu|} \right) \exp\left( \frac{1}{2} \left( \frac{\delta}{\sqrt{|\mu|}} + \frac{\mu_k}{\mu} x \right)^2 \right) \]

\[ \cdot \Phi\left( \frac{\mu_k}{\mu} x + \frac{\delta}{\sqrt{|\mu|}} \right) - \Delta_L \frac{\mu_k}{\mu}, \quad (14) \]

and that for \( k \in \{1, \ldots, K-1\} \)

\[ V'_k(x) = \exp\left( \frac{1}{2} \left( \frac{\delta}{\sqrt{|\mu|}} + \frac{\mu_k}{\mu} x \right)^2 \right) \]

\[ \cdot \left( \frac{\Delta_k}{\mu_k} - \frac{\Delta_{k+1}}{\mu_{k+1}} + \frac{\Delta_k}{\mu_k} \right) \frac{\sqrt{2 \pi}}{\sqrt{|\mu_k|}} \phi\left( \frac{\delta}{\sqrt{|\mu|}} - \sqrt{\frac{\mu_k}{\mu}} L_k \right) \]

\[ + \frac{\delta \sqrt{2 \pi}}{\sqrt{|\mu_k|}} \left( \frac{c}{\sqrt{|\mu|}} - \frac{\Delta_k}{\mu_k} \right) \phi\left( \frac{\delta}{\sqrt{|\mu|}} + \sqrt{\frac{\mu_k}{\mu}} L_k \right) \]

\[ - \Phi\left( \frac{\delta}{\sqrt{|\mu|}} - \sqrt{\frac{\mu_k}{\mu}} L_k \right) \right) - \frac{\Delta_k}{\mu_k}, \quad (15) \]

where

\[ \frac{c}{\delta \sqrt{|\mu|}} = \frac{\Delta_1}{\mu_1} + \left( 1 - \exp\left( \frac{1}{2} \frac{\delta^2}{\mu_1} \right) \left( \frac{\Delta_1 - \Delta_2}{\mu_2 - \mu_1} + \frac{\Delta_2}{\mu_1} \right) \right) \]

\[ \cdot \frac{\sqrt{2 \pi}}{\sqrt{|\mu_1|}} \phi\left( \frac{\delta}{\sqrt{|\mu_1|}} - \sqrt{\frac{\mu_1}{\mu_1}} L_1 \right) \]

\[ \cdot \left( 1 + \exp\left( \frac{1}{2} \frac{\delta^2}{\mu_1} \right) \frac{\delta \sqrt{2 \pi}}{\sqrt{|\mu_1|}} \right) \phi\left( \frac{\delta}{\sqrt{|\mu_1|}} - \sqrt{\frac{\mu_1}{\mu_1}} L_1 \right) \]

\[ \cdot \left( \Phi\left( \frac{\delta}{\sqrt{|\mu_1|}} - \sqrt{\frac{\mu_1}{\mu_1}} L_1 \right) - \Phi\left( \frac{\delta}{\sqrt{|\mu_1|}} - \sqrt{\frac{\mu_1}{\mu_1}} L_1 \right) \right)^{-1}. \]

In particular, we have used condition (12) to solve for the constants \( \alpha_0, \ldots, \alpha_K \) that appear in (11) to obtain (14) and (15). The following lemma shows that we can choose the Lagrange multipliers \( \Delta_1, \ldots, \Delta_K \) such that condition (13) is satisfied.

**Lemma 3.** For any \( K-1 \) dimensional vector \( \tilde{L} \) having
\[ 0 < L_1 < L_2 < \cdots < L_{K-1}, \]
there exist Lagrange multipliers \( \Delta_1, \ldots, \Delta_K \) that are positive and satisfy

\[ \frac{\Delta_{K-1} - \Delta_K}{\mu_K - \mu_{K-1}} < \frac{\Delta_{K-2} - \Delta_{K-1}}{\mu_{K-1} - \mu_{K-2}} < \cdots < \frac{\Delta_2 - \Delta_3}{\mu_3 - \mu_2} < \frac{\Delta_1 - \Delta_2}{\mu_2 - \mu_1} < 0, \quad (16) \]

such that \( V'_k \) satisfies (14), \( V'_k \) satisfies (15) for all \( k \in \{1, \ldots, K-1\} \), and condition (13) holds. Furthermore, \( V'_k \) is increasing for each \( k \in \mathcal{K} \).

In summary, for any fixed threshold levels \( 0 < L_1 < L_2 < \cdots < L_{K-1}, \) we have evidenced the existence of a function \( V, \) a constant \( c, \) and Lagrange multipliers \( \Delta_k \) for all \( k \in \mathcal{K} \) such that the conditions of Lemma 2 hold.

### 4.3. The Diffusion Control Problem Solution

We begin with the verification lemma that allows us to argue that for any fixed threshold levels \( 0 < L_1 < L_2 < \cdots < L_{K-1}, \) under Lagrange multipliers \( \Delta_1, \ldots, \Delta_K \) satisfying the conditions of Lemma 3, these threshold levels are optimal for the Lagrangian relaxation problem (7).

**Lemma 4.** Suppose there exists a twice-continuously differentiable function \( V: \mathbb{R} \to \mathbb{R} \) and a constant \( c \in \mathbb{R} \) that solve

\[ \inf_{u \in \mathcal{U}} \mu V''(x) + m(x, u) V'(x) + x^+ + \sum_{k=1}^{K} \Delta_k u_k x^- = c, \]

for all \( x \in \mathcal{X}. \quad (17) \]

Also assume there exist \( b_1, b_2 \in \mathbb{R} \) such that \( |V(x)| \leq b_1 x^2 + b_2 \) for all \( x \in \mathcal{X}. \) Then, if \( \hat{X} \) satisfies (3) under some admissible control \( u \in \mathcal{U}_p, \)

\[ E \hat{X}(\infty)^+ + \sum_{k=1}^{K} \Delta_k E[u_k(\hat{X}(\infty)^- \hat{X}(\infty)^-)] \geq c. \]

To connect the conditions in Lemma 4 to the Lagrange multiplier condition (16) in Lemma 3, it is useful to define the functions

\[ \psi_k(y) := \begin{cases} \mu_k y + \Delta_k, & y \leq 0 \\ 0, & y > 0 \end{cases} \quad \text{for } k \in \mathcal{K}, \quad \text{and} \]

\[ \psi(y) := \min_{k \in \mathcal{K}} \psi_k(y) \quad \text{for all } y \in \mathcal{X}. \]

Then, an equivalent representation of (17) is

\[ \mu V''(x) - \delta \sqrt{\mu} V'(x) + x^+ + \psi(V'(x)) x^- = c \]

for all \( x \in \mathcal{X}. \quad (18) \]
Note that the intersection point of the functions $\psi_k$ and $\psi_j$ is $(\Delta_j - \Delta_i)/(\mu_k - \mu_j)$. Hence

$$
\psi(y) = \begin{cases} 
\psi_k(y) & y < \frac{\Delta_{k-1} - \Delta_k}{\mu_k - \mu_{k-1}} \\
\psi_{k-1}(y) & \frac{\Delta_{k-1} - \Delta_k}{\mu_k - \mu_{k-1}} \leq y < \frac{\Delta_{k-2} - \Delta_{k-1}}{\mu_{k-1} - \mu_{k-2}} \\
\vdots & \\
\psi_1(y) & \frac{\Delta_1 - \Delta_2}{\mu_2 - \mu_1} \leq y < 0.
\end{cases}
$$

The function $V'$ has been constructed so that

$$V'(-L_k) = \frac{\Delta_k - \Delta_{k+1}}{\mu_{k+1} - \mu_k} \quad \text{for all } k \in \{1, \ldots, K-1\},$$

(see (12)) and, by Lemma 3, $V'$ is increasing and continuous. Hence

$$
\psi(V'(x)) = \begin{cases} 
\psi_k(V'_k(x)) & x < -L_{k-1} \\
\psi_{k-1}(V'_{k-1}(x)) & -L_{k-1} < x < -L_{k-2} \\
\vdots & \\
\psi_1(V'_1(x)) & -L_1 \leq x < 0,
\end{cases}
$$

which implies the ordinary differential equation (ode) in (18) is exactly the ode in (10) in Lemma 2. Therefore, we can conclude that if $\tilde{X}$ satisfies (3) under any admissible control $u \in U_p$, the system remains stable.

$$E[\tilde{X}(\infty)] + \sum_{k=1}^{K} \Delta_k [E[u_k(\tilde{X}(\infty)) - \tilde{X}(\infty)] - f_k d] \geq E[\tilde{X}^*(\infty)] + \sum_{k=1}^{K} \Delta_k E[\tilde{X}^*(\infty) - 1(L_{k-1} \leq -\tilde{X}(\infty) \leq L_k)] - f_k d]. \quad (19)$$

It follows from Inequality (19) that the threshold control policy that is optimal is exactly the one that satisfies the constraints in the diffusion control problem (6). Specifically, let $L_1^*, \ldots, L_{K-1}^*$ be the threshold levels that satisfy $E[\tilde{X}(\infty) - 1(L_{k-1} \leq -\tilde{X}(\infty) \leq L_k)] = f_k d$ for all $k \in \mathcal{K}$. Then, (19) implies that under any control having $E[u_k(\tilde{X}(\infty)) \tilde{X}(\infty)] = f_k d$ for all $k \in \mathcal{K}$, it is also true that $E[\tilde{X}(\infty)] \geq E[\tilde{X}^*(\infty)]$. Hence we have proved the following theorem.

**Theorem 1.** Let $\tilde{X}^*$ satisfy (3) under the threshold control having levels $L_1^*, \ldots, L_{K-1}^*$. Let $\Delta_1^*, \ldots, \Delta_{K-1}^*$ be the associated penalty parameters. Let $\epsilon > 0$ and $\Delta^* = \max_{k \in \mathcal{K}} \Delta_k$. Then, for any $\tilde{X}$ that satisfies (3) under control $u \in U_p$ and has $E[u_k(\tilde{X}(\infty)) \tilde{X}(\infty)] = f_k d$ for all $k \in \mathcal{K}$, it is also true that $E[\tilde{X}(\infty)] \geq E[\tilde{X}^*(\infty)]$.

The following lemma shows that the threshold levels $L_1^*, \ldots, L_{K-1}^*$ exist, are unique, and can be found numerically with relative ease.

**Lemma 5.** For any positive $f_1, \ldots, f_K$ having $\sum_{k=1}^{K} f_k = 1$, there exist unique $0 < L_1 < L_2 < \cdots < L_{K-1} < \infty$ such that

$$E[\tilde{X}(\infty) - 1(L_{k-1} \leq -\tilde{X}(\infty) < L_k)] = f_k d \quad \text{for all } k \in \mathcal{K}.$$}

Furthermore, such threshold levels can be found according to a search procedure that is a sequence of one-dimensional searches.

Finally, to prove asymptotic optimality, it is useful to observe that if the constraints in (6) are violated by no more than a small amount, then the objective function value cannot decrease by more than a small amount.

**Corollary 1.** Let $\tilde{X}^*$ satisfy (3) under the threshold control having levels $L_1^*, \ldots, L_{K-1}^*$. Let $\epsilon > 0$ and $\Delta^* = \max_{k \in \mathcal{K}} \Delta_k$. Then, for any $\tilde{X}$ that satisfies (3) under control $u \in U_p$ and has

$$\frac{\mu \epsilon}{\Delta^*} \tilde{X}(\infty) - f_k d \leq \frac{\mu \epsilon}{\Delta^*} \tilde{X}(\infty) - f_k d \leq \frac{\mu \epsilon}{\Delta^*} \tilde{X}(\infty) - f_k d \quad \text{for all } k \in \mathcal{K}, \quad (20)$$

it is also true that $E[\tilde{X}(\infty)] + \mu \epsilon \geq E[\tilde{X}^*(\infty)]$.

### 5. The Proposed Threshold Policy

There is a natural translation of the optimal threshold policy for the approximating diffusion control in Theorem 1 to a routing policy for the original system. We specify this translation in terms of the FSF-excluding-pool-$k$ routing policy. This policy assigns a newly arriving customer to an idle server according to the priority rule: pool $K$, pool $K-1$, ..., pool $k+1$, pool $k-1$, ..., pool 1, pool $k$, so that the newly arriving customer is routed to pool $k$ only if all the servers from all the other routing pools are busy. In the Halfin-Whitt limit regime, this policy ensures all idle servers are from pool $k$. Note that the FSF excluding pool 1 routing policy is exactly the FSF routing policy.

**The Threshold Policy.** Let $L_1^*, \ldots, L_{K-1}^*$ be the threshold levels and $\tilde{X}^*$ be the associated diffusion, from Theorem 1, so that $E[\tilde{X}^*(\infty) - 1(L_{k-1} \leq -\tilde{X}(\infty) < L_k)] = f_k d$ for all $k \in \mathcal{K}$, and the constraints in the diffusion control problem (6) are satisfied. Fix $\lambda$ and $N^\lambda$, and let $L_0^\lambda := N^\lambda$, $L_k^\lambda = 0$, and

$$L_k^\lambda := N^\lambda - L_{k-1}^\lambda \sqrt{N^\lambda}, \quad k \in \{1, \ldots, K-1\},$$

so that $L_0^\lambda < L_{K-1}^\lambda < \cdots < L_2^\lambda < L_1^\lambda < L_0^\lambda$. Suppose the number of customers in the system, $x$, satisfies $L_k^\lambda \leq x < L_{k-1}^\lambda$. Then, the threshold policy routes in accordance with the FSF excluding pool $k$ policy. Otherwise, if no servers in any pool $k$ are free at the time the new customer arrives, the customer queues.

Note that when $L_k^\lambda \leq x < L_0^\lambda = N^\lambda$, so that there are almost as many customers in the system as servers, the idle servers are the slow servers, from pool 1, and when $L_k^\lambda =$
0 ≤ x < L_{k-1}^k$, so that there are few customers in the system compared with the number of servers, the idle servers are the fast servers, from pool $K$.

We expect our proposed threshold policy to perform well for large values of $\lambda$ because the diffusion control problem we solve in §4 arises when formally passing to the limit in (1) as $\lambda$ becomes large. Recall that this intuition is validated in Figure 3 in §2, when we compare the performance of our proposed threshold policy to the LISF policy that call centers commonly use. However, the LISF policy can achieve only one value of the vector $(f_1, \ldots, f_K)$, and our proposed threshold policy can achieve any value of $(f_1, \ldots, f_K)$ that satisfies $f_k \in [0, 1]$ for all $k \in \mathcal{K}$ and $f_1 + \cdots + f_K = 1$. Therefore, for further evidence that our policy performs well, in the e-companion, we also compare the performance of our policy to a variant of the LISF policy that can achieve the same set of values $f_1, \ldots, f_K$ by weighting the different server pools and routing calls to the server that has experienced the longest weighted idle times.

### 6. Asymptotic Analysis

We validate the performance of our proposed threshold policy in §5 analytically by establishing that a continuous modification of it is asymptotically optimal in the Halfin-Whitt limit regime. We start with the definitions required to make precise what we mean by asymptotically optimal.

First, note that in the Halfin-Whitt limit regime, waiting times become small so that the scaled waiting time is defined as: $\hat{W}_\lambda(t) = \sqrt{N^\lambda} W_\lambda(t)$.

**Definition (Asymptotic Feasibility).** Consider a sequence of systems, indexed by $\lambda$, that operates under a sequence of admissible policies $\pi = \pi(\lambda, N^\lambda) \in \Pi$ for every value of $\lambda$. Then, $\pi$ is **asymptotically feasible** with respect to (1) if $\lim_{\lambda \to \infty} (E I^k_\lambda(\infty; \pi) / E I^k_\lambda(\infty; \pi_0)) = f_k$.

**Definition (Asymptotic Optimality).** Consider a sequence of systems, indexed by $\lambda$, that operates under an asymptotically feasible sequence of policies $\pi = \pi(\lambda, N^\lambda)$. Then, $\pi$ is **asymptotically optimal** with respect to (1) if for any other sequence of asymptotically feasible policies $\pi'$ $\limsup_{\lambda \to \infty} E \hat{W}_\lambda(\infty; \pi) \leq \liminf_{\lambda \to \infty} E \hat{W}_\lambda(\infty; \pi')$.

Intuitively, as $\lambda \to \infty$, under the proposed threshold policy, the number of idle servers in any pool besides $k$, $k \in \mathcal{K}$, will be negligibly small when the number of customers in the system is between the threshold levels $L_k^k$ and $L_{k-1}^k$. The total number of idle servers in pool $k$ will be exactly the total number of idle servers when the total number of customers in the system is between $L_k^k$ and $L_{k-1}^k$. This suggests a noncontinuous form of state-space collapse at the threshold levels. Existing techniques do not allow for the establishment of such a form of state-space collapse (for example, both Dai and Tezcan 2005 and Gurvich and Whitt 2009a assume a continuous form of state-space collapse). Therefore, we first show that a “continuous adjustment” of the threshold policy proposed in §5 into what we call the $\epsilon$-threshold policy is $\epsilon$-asymptotically feasible

$$\lim_{\lambda \to \infty} E I^k_\lambda(\infty; \pi) / E I^k_\lambda(\infty; \pi_0) - f_k < \epsilon$$

for all $k \in \mathcal{K}$, and $\epsilon$-asymptotically optimal

$$\limsup_{\lambda \to \infty} E \hat{W}_\lambda(\infty; \pi) \leq \liminf_{\lambda \to \infty} E \hat{W}_\lambda(\infty; \pi') + \epsilon,$$

for $\epsilon > 0$ arbitrarily small. We then allow $\epsilon$ to depend on $\lambda$ in such a way that $\epsilon(\lambda)$ goes to 0 as $\lambda$ becomes large, so that we can establish asymptotic optimality of the $\epsilon$-threshold policy.

Our proof of asymptotic optimality requires an additional assumption on the behavior of an admissible policy in the Halfin-Whitt limit regime that is in essence a requirement that there is a state-space collapse. This assumption is needed because the scaled number of idle servers might converge to a limit process with discontinuous drift under an arbitrary admissible policy, and as noted previously, existing techniques do not allow the establishment of such a form of state-space collapse.

To state the aforementioned assumption, we require the centered and scaled processes $\hat{X}_\lambda(t) = (\hat{X}_k^\lambda(t), \ldots, \hat{X}_k^\lambda(t))$ defined as follows:

$$\hat{X}^\lambda_i(t) := Q^\lambda_i(t) + Z^\lambda_i(t) - N^\lambda_i$$

$$\hat{X}^\lambda_k(t) := Z^\lambda_k(t) - N^\lambda_k, \quad k \in \{2, \ldots, K\}.$$

Note that $\hat{X}^\lambda_i(t) \leq 0$ for all $t$, and that for $k \in \mathcal{K}$, $\hat{I}^\lambda_i(t) := [\hat{X}^\lambda_i(t)]^+$ corresponds to the number of idle servers, scaled by $1/\sqrt{N^\lambda}$. Because we consider only nonidling policies, if $\hat{X}^\lambda_i(t) < 0$ for any $k \in \mathcal{K}$, then $\hat{X}^\lambda_i(t) \leq 0$. Let

$$\hat{X}^\lambda(t) := \sum_{k=1}^K \hat{X}^\lambda_k(t) = \frac{Q^\lambda(t) + \sum_{k=1}^K Z^\lambda_k(t) - N^\lambda}{\sqrt{N^\lambda}}$$

$$\hat{X}^\lambda(t) = \frac{X^\lambda(t) - N^\lambda}{\sqrt{N^\lambda}}.$$

Then, $\hat{I}^\lambda(t) := [\hat{X}^\lambda(t)]^+$ is the total number of idle servers, and $[\hat{X}^\lambda(t)]^+ = [\hat{X}^\lambda(t)]^+$ is the total queue length, both scaled by $1/\sqrt{N^\lambda}$. (A3) There is state-space collapse; that is, on any subsequence $\lambda_i$ having $X^\lambda_i \Rightarrow \hat{X}$ as $\lambda_i \to \infty$, there exist functions $\hat{u}_k : [0, \infty) \to [0, 1]$ having $\sum_{k \in \mathcal{K}} \hat{u}_k(x) = 1$ for any $x \in [0, \infty)$, and only a finite number of discontinuities, such that

$$\hat{I}^\lambda_k \Rightarrow \hat{I}_k \quad \text{for all } k \in \mathcal{K} \text{ as } \lambda_i \to \infty,$$

where $\hat{I}_k(t) := \hat{u}_k(\hat{X}(t^-)) \hat{X}(t^-)$. 

---

*Armony and Ward: Fair Dynamic Routing*  
Assumption (A3) is true for any QIR policy proposed in Gurvich and Whitt (2009a, b, 2010), which has a continuous form of state-space collapse. We conjecture that it holds for the proposed threshold policy in §5.

There are three key steps in our asymptotic optimality argument.

1. In §6.1, we propose the $\epsilon$-threshold policy $\hat{TH}_\epsilon$ for the diffusion $\hat{X}$ in (3) and show in Proposition 2 that it is $\epsilon$-feasible and $\epsilon$-optimal (terms that are defined precisely in Proposition 2) for the diffusion control problem (6). The $\epsilon$-threshold policy modifies the threshold control at levels $L^*_1 < L^*_2 < \cdots < L^*_K$, into a control that has a continuous infinitesimal drift.

2. In §6.2, we propose the $\epsilon$-threshold policy $TH_\epsilon$ for the original queueing problem (1) and show that its performance asymptotically approaches the performance of $\hat{TH}_\epsilon$.

3. Finally, in §6.3, we establish asymptotic optimality.

### 6.1. The Diffusion $\epsilon$-Threshold Policy

Let $\hat{X}^*$ denote the diffusion in Theorem 1 that satisfies (3) under the optimal threshold control at levels $L^*_1, \ldots, L^*_K$. Then, $\hat{X}^*$ has a piecewise linear drift $m_{\hat{X}}(x)$ as in (8) and an infinitesimal variance $2\mu$. We would like to replace the process $\hat{X}^*$ by another diffusion process whose steady-state performance is close to that of $\hat{X}^*$, but whose infinitesimal drift term is continuous. To do this, we propose the diffusion process with continuous infinitesimal drift $\hat{X}_\eta$, and show that

$$E[\hat{X}_\eta(\infty)1\{-L^*_1 \leq \hat{X}_\eta(\infty) \leq L^*_K\}] = f_k d, \quad \text{for all } k \in \mathcal{H},$$

and $E\hat{X}_\eta(\infty)^- \rightarrow E\hat{X}^*(\infty)^+$, as $\eta \downarrow 0$.

We define $\hat{X}_\eta$ to be the diffusion process with infinitesimal variance $2\mu$ and continuous infinitesimal drift:

$$m_{\hat{X}}(x) = \begin{cases} 
-\delta \sqrt{\mu} & x \geq 0 \\
-\delta \sqrt{\mu} - \mu_k x & -L^*_1 + \eta \leq x < -L^*_{k-1} \\
-\delta \sqrt{\mu} - \mu_k x + \frac{L^*_k}{\eta} (\mu_{k+1} - \mu_k) \cdot (\eta - L^*_k - x) & -L^*_k \leq x < -L^*_1 + \eta \\
-\delta \sqrt{\mu} - \mu_k x & x < -L^*_k, 
\end{cases}$$

for $0 < \eta < \min_{k \in \mathcal{K}}(L^*_k - L^*_{k-1})$ and $k \in \{1, \ldots, K - 1\}$. The control policy that corresponds to the process $\hat{X}_\eta$ is

$$u_{\eta, k}(x) = \begin{cases} 
1 & -L^*_1 + \eta \leq -x < 0 \\
1 + \frac{L^*_k}{\eta} \left( \frac{\eta - L^*_1}{-x} - 1 \right) & -L^*_1 \leq -x < -L^*_1 + \eta, 
\end{cases}$$

for $k \in \{2, \ldots, K - 1\}$,

This is straightforward to verify by confirming that

$$0 \leq u_{\eta, k} \leq 1 \quad \text{for all } k \in \mathcal{H}, \quad \sum_{k=1}^{\infty} u_{\eta, k} = 1, \quad \text{and}$$

$$m_{\hat{X}}(x) = -\delta \sqrt{\mu} + \sum_{k=1}^{\infty} u_{\eta, k}(x) \mu_k x^- \quad \text{for all } x < 0.$$

For any $\epsilon > 0$, the $\epsilon$-threshold policy $\hat{TH}_\epsilon$ is defined by choosing $\eta_\epsilon$ so that

$$|E\hat{X}_{\eta_\epsilon}(\infty)^- u_{\eta_\epsilon, k}(\hat{X}(\infty)^-) - f_k d| < \frac{\mu \epsilon}{\Delta^* K}, \quad \text{and}$$

$$|E\hat{X}_{\eta_\epsilon}(\infty)^+ - E\hat{X}^*(\infty)^+| < \mu \epsilon$$

are satisfied. Conditions (24) and (25) require that $\hat{TH}_\epsilon$ be $\epsilon$-feasible and $\epsilon$-optimal, respectively. Note that condition (24) is exactly the $\epsilon$-feasibility condition (20) in Corollary 1.

The following proposition establishes that such an $\eta_\epsilon$ exists.

**Proposition 2.** For any $\epsilon > 0$, there exists $\eta_\epsilon$ such that (24) and (25) hold.

### 6.2. The $\epsilon$-Threshold Policy

In this section, we define the $\epsilon$-threshold policy $TH_\epsilon$ for the original queueing system (fixed $\lambda$). We then establish that, in the limit, under appropriate (diffusion) scaling, as $\lambda \to \infty$, the $\epsilon$-threshold policy $TH_\epsilon$ asymptotically performs as well as its diffusion counterpart $\hat{TH}_\epsilon$.

Roughly speaking, according to $\hat{TH}_\epsilon$, if at time $t$ the state of the diffusion is $x < 0$, then a fraction $u_{\eta_\epsilon, k}(x^-)$ of the total number of idle servers $x^-$ receives service from pool $k$. Our purpose in $TH_\epsilon$ is to imitate this policy, keeping in mind that for the original system the policy must be nonpreemptive.

**The $\epsilon$-Threshold Policy.** The policy $TH_\epsilon$ assigns a new arrival at time $t$ to a server of pool $k^*(t)$ where
Assume the initial system conditions satisfy Theorem 3. We are able to use Proposition 3 to establish asymptotic the same conditions as in Theorem 2. The sequence of polynomials $X^k = \arg\max_{\lambda} [\hat{I}^k(i) - (\hat{X}^k(i))^* \cdot u_{n_k,i}((\hat{X}^k(i))^*)]$. Ties are resolved arbitrarily.

The $\epsilon$-threshold policy is a special case of the queue-and-idleness-ratio (QIR) control introduced in Gurvich and Whitt (2009a).

**Theorem 2.** Consider a sequence of systems operating under the $\epsilon$-threshold policy $TH_\epsilon$, and (A3) holds. Suppose the initial system conditions are such that

$$\frac{Q^\lambda(0)}{N^\lambda} \to 0 \quad \text{and} \quad \frac{Z^\lambda(0)}{N^\lambda_k} \to q_k, \quad k \in \mathcal{K}, \quad \text{and}$$

$$\hat{X}^\lambda(0) \Rightarrow \hat{X}(0), \quad \text{as} \quad \lambda \to \infty,$$

with $\sum_{k=1}^K \hat{X}_k(0) = x$ and $\hat{I}_k(0) = \hat{X}_k^*(0) = x^- \cdot u_{n_k,i}(x^-)$, for all $k \in \mathcal{K}$. Then,

$$\lim_{\lambda \to \infty} E[\hat{X}(\infty)] = E[\hat{X}(\infty)], \quad \text{and}$$

$$\lim_{\lambda \to \infty} E[I,(\infty)] = E[\hat{X}(\infty)^* \cdot u_{n_k,i}(\hat{X}(\infty)^-)], \quad k \in \mathcal{K}, \quad (26)$$

where $\hat{X}$ is the diffusion process associated with the policy $TH_\epsilon$.

### 6.3. Asymptotic Optimality

We first observe that the $\epsilon$-threshold policy $TH_\epsilon$ is $\epsilon$-asymptotically feasible and $\epsilon$-asymptotically optimal.

**Proposition 3.** Fix $\epsilon > 0$, and consider a sequence of systems, indexed by $\lambda$, that operates under the $\epsilon$-threshold policy $TH_\epsilon$. Suppose the initial system conditions satisfy the same conditions as in Theorem 2.

(i) There is $\epsilon$-asymptotic feasibility; that is,

$$\lim_{\lambda \to \infty} \frac{|E[I,(\infty); TH_\epsilon]|}{E[I,(\infty); TH_\epsilon]} - f_k < \frac{\mu \epsilon}{\Delta^* Kd} \quad \text{for all} \quad k \in \mathcal{K};$$

(ii) There is $\epsilon$-asymptotic optimality; that is, if $\pi = \pi(\lambda, N^\lambda_k)$ is any other sequence of admissible policies that satisfies (i) and (A3),

$$\lim_{\lambda \to \infty} E[\hat{W}(\infty); TH_\epsilon] \leq \liminf_{\lambda \to \infty} E[\hat{W}(\infty); \pi] + 2\epsilon.$$

By letting $\epsilon$ depend on $\lambda$ so that $\epsilon(\lambda) \downarrow 0$ as $\lambda \to \infty$, we are able to use Proposition 3 to establish asymptotic optimality.

**Theorem 3.** Assume the initial system conditions satisfy the same conditions as in Theorem 2. The sequence of policies $TH_\epsilon^k$ is asymptotically optimal.

### 7. Conclusions and Future Research

Service systems with heterogeneous servers are concerned about two conflicting goals: minimizing expected customers waiting times and maintaining fairness among their servers. We formulate this problem as a dynamic control problem with waiting time performance as the objective function and a fairness criterion as the constraint. For this problem we propose a simple threshold policy that is based on the total number of customers in the system, and controls which server pool has idle servers. Our proposed threshold policy has the intuitively desirable property that the faster servers idle only when there are few customers in the system, leaving the slower servers to idle (and so the faster servers to work) when the number of customers in the system is close to the number of servers. This policy is numerically shown to be fair and to improve on the expected waiting time in comparison with the longest idle server first (LISF) policy commonly used in call centers. Formally, we show that a continuous version of the threshold policy is asymptotically optimal in the Halfin-Whitt limit regime.

An interesting direction for future research is to further study how to best define the notion of server fairness (rather than assuming it is defined through the constraint in (1)). This raises the question of also exploring the incentive issues associated with various control schemes. For example, it is fairly obvious that the faster server first policy gives an incentive to faster servers to slow down in order to get a break. However, an appropriate compensation mechanism (e.g., pay-per-call) could motivate all agents to work as fast as they can. The downside there is that quality might suffer.

Finally, understanding how to adapt the problem formulation and extend the results to more general multiskill networks, with heterogeneous servers and customer abandonment, without the non-idling assumption is important.

### 8. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

### Acknowledgments

The authors thank Rami Atar, Ashish Goel, Iltur Gurvich, John Hasenbein, Tolga Tezcan, and Assaf Zeevi for many valuable discussions. They are also grateful to the associate editor and two anonymous referees for a careful review of the paper and providing the authors with insightful suggestions.

### References


Tezcan, T. 2007. Asymptotically optimal control of many-server heterogeneous service systems with hyperexponential service times. Working paper, University of Illinois at Urbana–Champaign, Urbana, IL.

