

Fair Dynamic Routing in Large-Scale Heterogeneous-Server Systems

Mor Armony¹

Amy R. Ward²

January 10, 2008

Abstract

In a call center, there is a natural trade-off between minimizing customer wait time and fairly dividing the workload amongst 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 two heterogeneous agent pools, one that handles calls at a faster speed than the other, and a single customer class. The objective is to minimize steady-state expected customer wait time subject to a “fairness” constraint on the workload division.

The optimization problem we formulate is difficult to solve exactly. Therefore, we solve the diffusion control problem that arises in the many-server heavy-traffic QED limiting regime. The resulting routing policy is a threshold policy that prioritizes faster agents when the number of customers in the system exceeds some threshold level and otherwise prioritizes slower agents. We prove our proposed threshold routing policy is near-optimal as the number of agents increases, and the system’s load approaches its maximum processing capacity. We further show simulation results that evidence 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) in terms of the steady-state expected customer waiting time for identical desired workload divisions.

Acknowledgement: We thank Rami Atar, Itay Gurvich, Tolga Tezcan and Assaf Zeevi for many valuable discussions.

¹Stern School of Business, New York University, marmony@stern.nyu.edu.

²Marshall School of Business, University of Southern California, amy.ward@marshall.usc.edu

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, as was empirically shown in [30] and [24].

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 [2]. 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 amongst employees leads to low employee satisfaction and hampers performance; see for example [16] and [15]. Furthermore, high employee satisfaction implies increased employee retention, and [46] shows that increased employee retention improves service. Finally, the appeal of additional idle time for relaxation may provide faster servers with an incentive for slowing their service rate, which would increase customer waiting times.

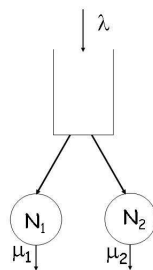
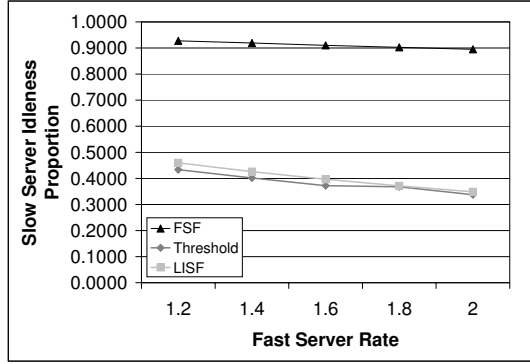
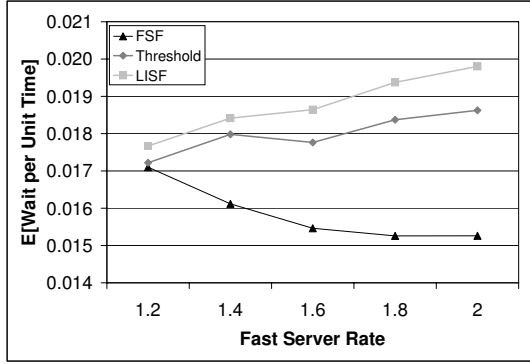


Figure 1.1: The inverted-V model.

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 (first introduced in Armony [2]) shown in Figure 1.1, with exponential inter-arrival and service times, two server types distinguished by their service rates $\mu_k, k = 1, 2$,



(a) The steady-state expected customer waiting time.

(b) The slow server idleness proportions.

Figure 1.2: A comparison of the performance of the threshold, LISF, and FSF policies for $N_1 = N_2 = 100$ and $\mu_1 = 1$.

and N_k servers of each type. For this system, the proportion of idleness experienced by servers of type k is asymptotically $N_k \mu_k / (N_1 \mu_1 + N_2 \mu_2)$ as the number of servers becomes large, and the arrival rate λ approaches the service capacity [5]. In particular, idleness is shared proportionally among the two 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 in order to achieve the same server idleness proportions.

Figure 1.2 (a) and (b) present 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 μ_2 is varied, and the arrival rate λ was adjusted according to the asymptotic regime of [5]. 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 1.2(a) that the expected waiting time under the threshold policy is consistently lower than under the LISF policy. Specifically, for the higher values of μ_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 1.2(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 fairness constraint on the slow server idleness proportion. (There is no reason to restrict ourselves to the idleness proportions attained by the LISF policy.) Such a problem is very difficult to solve exactly. Therefore, noting that call centers generally have a large number of agents and operate in a regime in which the arrival rate and service capacity are close, we consider the many-server heavy-traffic regime first appearing in [20], and formally introduced in Halfin and Whitt [32]. The diffusion control problem that arises in this regime is analytically tractable, and we solve this explicitly to find that a threshold control 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. However, proving the asymptotic optimality of this policy is difficult due to the existence of a discontinuity in the infinitesimal drift. Therefore, we define a notion of ϵ -asymptotic optimality, and prove that a continuous adjustment of the proposed threshold policy is ϵ -asymptotically optimal.

The remainder of this paper is organized as follows. We first review relevant literature. In Section 2, we present our basic model formulation. We construct and solve an approximating diffusion control problem in Section 3. In Section 4, we propose a family of policies based on the solution to the approximating diffusion control problem whose performance is provably near-optimal as λ grows large. Finally, we make concluding remarks in Section 5.

Due to the technical nature of our results, our approach in their presentation is to state them formally and precisely in the body of the paper, but to have the formal proofs appear in a technical appendix [4].

1.1 Literature Review

Inter-server fairness

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 [9] and [47]. 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. [16, 15]) and the practice at many call centers to use fair policies such as the Longest-Idle-Server-First. Two papers that do address this issue are [48] (in the context of bandwidth allocation in telecommunication networks) and [14]. Cabral [14] examines the question of which servers work more in a heterogeneous server system with equally likely random customer-server assignments among idle servers. The author in [14] shows that, in a comparison between any two servers, the faster server is idle for a greater fraction of the time. He also shows that the effective service rate is higher for the faster server.

The slow server problem

Heterogeneity among servers has brought researchers to ask the following two questions: a) When is it optimal to remove the slowest server from a queueing system, to minimize the mean sojourn time in the system (e.g. [41], [13]), and b) Given a set of heterogeneous servers, how to dynamically route customers to servers in order to minimize the mean sojourn time (e.g. [37], [38], [42] and [18]). Both these problems have been coined the slow server problem. For a while, only results for the two-server system have been published (e.g. [41],[38]), but recently results for the general heterogeneous multi-server system have appeared ([13], [18]). Note, though, that the latter problem for the general multiserver case is still open [19].

Inverted-V and asymptotic analysis

The difficulty in identifying optimal controls for the general heterogeneous server problem has prompted researchers to examine this question in various asymptotic regimes. For example, in the conventional heavy traffic regime, for a two server system with two queues, in which routing decisions must be made at the time of each arrival, [21] shows that shortest-expected-delay-first routing is asymptotically optimal, and [43] identify necessary and sufficient conditions for a threshold priority policy to be asymptotically optimal.

More recently, heterogeneous server systems have been studied in the many-server heavy-traffic regime commonly referred to as the Halfin-Whitt regime [32] or the QED regime [23]. In this regime the arrival rate and the number of servers grow to infinity according to a square-root safety staffing rule. This rule has been shown to be asymptotically optimal in various settings ([11], [39], [26]), including the inverted-V model [3].

Several papers have examined the question of dynamic control for the inverted-V system in the QED regime. These include [2], [44], [5], [8] and [3]. With the exception of [5] none of those deals with the inter-server fairness issue. Armony [2] shows that the faster-server first (FSF) policy is asymptotically optimal in the sense that it asymptotically minimizes the expected steady-state waiting time and delay probability. These results have been extended to an inverted-V system with

abandonment in Armony and Mandelbaum [3]. Tezcan [44] examines a similar control question with service times that are hyper-exponential. The author shows that while a priority type policy is still asymptotically optimal the actual priorities depend on other factors beyond the mean service time.

Recently Atar [5] has established that both the FSF and the LISF policies exhibit state-space collapse in the QED regime, even in settings where the service rates are random. Loosely speaking, state-space collapse implies that the system dynamics can be described in the limit by a lower dimensional process than the original process. In [5] the state-space collapse is into a one-dimensional process. Interestingly, according to [5], if the LISF policy is applied in our setting the diffusion limit may be described by a fixed ratio policy in which the fraction of idle servers of each pool is fixed over time. This suggests that LISF is indeed more fair than FSF. Most recently, Atar and Shwartz [8] have shown that in an environment where service rates are heterogeneous and unknown, it is sufficient to take a very small sample of service times to come up with a routing policy that is asymptotically optimal.

Beyond the control problem for the inverted-V problem, there is a growing body of literature that deals with dynamic control of multiclass parallel server systems with heterogeneous servers. This problem is often referred to as skill-based routing. Recently, it has been shown by Gurvich and Whitt [28, 29] that if a general multiskill system has service rates that are server dependent (i.e. they are independent of the customer class) then the system can be reduced to an inverted-V system. Gurvich and Whitt propose a general Fixed-Idleness-Ratio (FIR) policy and prove that the system exhibits state-space collapse under this policy. Moreover, they show that if the ratios are appropriately chosen then FIR is asymptotically optimal in minimizing convex delay costs. The ϵ -threshold policy that we propose in Section 4.2 turns out to be a special case of the FIR policy. We utilize this fact in proving some of our results. Interestingly, our proposed Threshold policy is *not* a special case of FIR due to its discontinuous nature.

2 Model Formulation

Consider a service system with a single customer class and two server types (each type in its own server pool), both capable of fully handling customers' service requirements. Service times are independent and exponential, and the average service time of a customer served by a server from pool k ($k = 1, 2$) is $1/\mu_k$. There are N_k servers in pool k , and $\vec{N} = (N_1, N_2)$ denotes the staffing vector. (Here and elsewhere, \vec{x} is used to denote a vector whose elements are x_1, x_2, \dots). We assume $\mu_1 < \mu_2$, meaning the faster servers reside in pool 2.

Customers arrive to the system according to an independent Poisson process with rate λ . We assume the following necessary condition for stability is satisfied:

$$N_1\mu_1 + N_2\mu_2 > \lambda, \tag{2.1}$$

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 both pools have idle servers must be routed to 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 fraction of idleness experienced by each server pool.

Denote by $\pi := \pi(\lambda, \vec{N})$ a policy that operates in a system with arrival rate λ and staffing vector \vec{N} (at times we will omit the arguments λ and \vec{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 ($k = 1, 2$) 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,

$$X(t; \pi) = Z_1(t; \pi) + Z_2(t; \pi) + Q(t; \pi).$$

Finally, let $W(t; \pi)$ be the virtual waiting time at time t , and let

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

be the number of idle servers in pool k , $k = 1, 2$, with

$$I(t; \pi) = I_1(t; \pi) + I_2(t; \pi)$$

the total number of idle servers. We use $t = \infty$ whenever we refer to the steady-state. Also, we omit the time argument when we refer to the entire process. At times, we will omit π if it is clear from the context which routing policy is used.

Definition: A policy π is called *work conserving* if there are no idle servers whenever there are some delayed customers in the queue. In other words, π is work conserving if $Q(t; \pi) > 0$ implies that $Z_1(t; \pi) + Z_2(t; \pi) = N$, where

$$N := N_1 + N_2$$

is the total number of servers.

Note that in general a 3-dimensional vector is required to specify the state of the system, namely, $(Q(t; \pi), Z_1(t; \pi), Z_2(t; \pi))$. However, for work conserving policies, the state space can be described by the 2-dimensional vector $\vec{X}(t; \pi) = (Z_1(t; \pi) + Q(t; \pi), Z_2(t; \pi))$. In fact, the queue length can be added to the number of busy servers of pool k , for any $k = 1, 2$, because if π is work conserving then $Q(t; \pi) = [Q(t; \pi) + Z_k(t; \pi) - N_k]^+$ (where $[x]^+ := \max\{x, 0\}$) and $Z_k(t; \pi) = \min\{Q(t; \pi) + Z_k(t; \pi), N_k\}$.

Let Π be the set of all non-preemptive, non-anticipating, work-conserving policies. By non-preemptive, we mean that once a call is assigned to a particular server, it cannot be transferred to another server of a different pool. The problem we would like to solve is as follows:

$$\begin{aligned} & \text{minimize}_{\pi \in \Pi} && EW(\infty; \pi) \\ & \text{subject to:} && \frac{EI_1(\infty; \pi)}{EI_1(\infty; \pi) + EI_2(\infty; \pi)} = f_1, \end{aligned} \tag{2.2}$$

where and $0 < f_1 < 1$ is the target steady-state fraction of pool 1 idleness. Specifically, given fixed values of μ_1, μ_2, λ and $\vec{N} = (N_1, N_2)$, one needs to find a policy $\pi = \pi(\lambda, \vec{N}) \in \Pi$ that minimizes the expected steady-state waiting time subject to the constraint in (2.2), which we henceforth refer to as “the fairness constraint”. Note that the fairness constraint in (2.2) is equivalent to the constraint $\frac{EI_2(\infty; \pi)}{EI(\infty; \pi)} = 1 - f_1$.

The routing problem defined in (2.2) is difficult to solve exactly. However, suppose we were to allow for preemption. In other words, suppose that at any point in time a service can be interrupted and the call transferred to another service. Of course, this is not reasonable from a customer service standpoint. However, we expect that the resulting routing problem (2.2) would then be solvable using dynamic programming techniques. This is because when preemption is allowed customers can be distributed between the two server pools in any fashion consistent with equations (2.4)-(2.12). Hence the system state is one-dimensional.

Fortunately, there is a regime in which we expect that the performance of a preemptive policy can be closely modeled using an appropriate non-preemptive policy, meaning that the system state will become one-dimensional. (This intuition follows from Theorem 5 in [6].) This regime is a heavy traffic regime in which the arrival rate λ and the service capacity $N_1\mu_1 + N_2\mu_2$ become large and are close. Therefore, our approach will be to solve the routing problem (2.2) in this heavy traffic asymptotic regime. More specifically, following the general approach outlined by Harrison [31], we will solve the diffusion control problem that arises when formally passing to the limit in the control problem having arrival rate λ , interpret its solution as a routing policy in the original system, and prove that that routing policy’s performance is near optimal.

An attractive feature of the policy obtained by solving the approximating diffusion control problem is that it is a threshold policy. However, the infinitesimal drift of the diffusion associated with the optimal threshold policy is discontinuous. This presents a non-trivial technical difficulty because existing techniques for establishing state-space collapse in parallel server systems (see, for example [17] and [27]) require continuity of this infinitesimal drift. Therefore, we prove that a “continuous adjustment” of a threshold policy for the original system asymptotically obtains performance that can be made arbitrarily close to the optimal solution to the approximating diffusion control problem.

Before setting the stage for our asymptotic analysis, we first discuss our problem formulation (2.2). Then, in Subsection 2.1, we provide the detailed system evolution equations, and in Subsection 2.2 we specify our heavy traffic asymptotic regime.

Discussion of problem formulation (2.2)

The most obvious question that arises from the problem formulation (2.2) is how to determine the idleness fraction parameter f_1 . There are various factors a manager might wish to consider. First, it is fairly intuitive that the expected waiting time should be decreasing in the pool 1 idleness proportion f_1 . Hence, one would obviously tend to choose higher values of f_1 . But how would this

choice affect system fairness?

One way to think about fairness is to consider individual servers and their utilization. Ultimately, one might want to ensure that all servers will have the same utilization. How does this affect the choice of f_1 ? Denote by ρ_k the expected utilization of server of pool k , $k = 1, 2$. We expect all servers of the same pool to have the same utilization. This can be guaranteed by randomly allocating customers to the servers within each pool all with equal weights. One can easily verify that under any policy for which the following steady-state expectations exist, we have that

$$1 - \rho_k = EI_k(\infty)/N_k. \quad (2.3)$$

Therefore, the requirement that $\rho_1 = \rho_2$ translates into $f_1 = \frac{N_1}{N_1+N_2}$.

More generally, any fairness criterion that involves individual server utilizations may be translated into a version of the problem (2.2) with the appropriate choice of f_1 by setting up and solving an equation in f_1 (as long as the solution satisfies that $f_1 \in [0, 1]$). To see this we need the following lemma.

Lemma 2.1 *Suppose that the steady-state of the processes X, I_1 and I_2 are well defined and their expectations are finite. Then, if $EI_1(\infty)/EI(\infty) = f_1$, we have that $EI(\infty) = \frac{\mu_1 N_1 + \mu_2 N_2 - \lambda}{f_1 \mu_1 + \mu_2 (1 - f_1)}$. In particular, $EI_1(\infty) = f_1 \frac{\mu_1 N_1 + \mu_2 N_2 - \lambda}{f_1 \mu_1 + \mu_2 (1 - f_1)}$ and $EI_2(\infty) = (1 - f_1) \frac{\mu_1 N_1 + \mu_2 N_2 - \lambda}{f_1 \mu_1 + \mu_2 (1 - f_1)}$.*

Suppose that the system manager wishes to ensure that the effective rate at which each individual server processes customers is the same over all servers. In particular, this implies the constraint $\rho_1 \mu_1 = \rho_2 \mu_2$. But given the relationship (2.3) and Lemma 2.1 one can easily obtain an equation in f_1 whose solution should be used in the constraint of (2.2).

Another point worth discussing in the problem formulation is the restriction of the family of policies Π to work-conserving policies. While this assumption is fairly intuitive in practice (it appears unnatural to keep customers waiting when there are idle servers) its formal justification is far from obvious. In particular, it is well known that it is sometimes optimal to idle slower servers when there are customers in queue in order not to starve the faster servers (Recall the literature on the slow server problem mentioned in section 1.1). However, if one allows for preemption, idling servers is no longer desired in that context (see for example Proposition 3.1 in [2]).

For the problem (2.2) considered in this paper, it is not clear whether it is optimal to use work-conserving policies even among preemptive policies. 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. So while we are unable to show optimality of work conserving policies among all non-anticipative policies, we assume work-conservation for analytic tractability. We also note that this is a fairly common assumption in the literature (see, for example, [6] and [7]).

2.1 System Evolution Equations

Let $A(t)$ be the total number of arrivals into the system up to time t (that is, $A(t)$, $t \geq 0$ is a Poisson process with rate λ). Also, for $k = 1, 2$ and for a policy $\pi \in \Pi$, let $A_k(t; \pi)$ be the total number of external arrivals joining pool k upon arrival up to time t , and let $R_k(t; \pi)$ be the total number of customers joining server pool k , up to time t , after being delayed in the queue. The number of arrivals into the queue (and not directly to one of the servers) up to time t is denoted by $A_q(t; \pi)$. In addition, let $T_k(t; \pi)$ denote the total time spent serving customers by all N_k servers of pool k up to time t . In particular, $0 \leq T_k(t; \pi) \leq N_k t$. Respectively, let $Y_k(t; \pi)$ be the total idle time experienced by servers of pool k up to time t . Finally, let $D_k(t)$ be a Poisson process with rate μ_k . Then the number of service completions out of server pool k may be written as $D_k(T_k(t; \pi))$.

The above definitions allow us to write the following *flow balance equations*:

$$Q(t; \pi) = Q(0; \pi) + A_q(t; \pi) - \sum_{k=1}^2 R_k(t; \pi), \quad (2.4)$$

$$Z_k(t; \pi) = Z_k(0; \pi) + A_k(t; \pi) + R_k(t; \pi) - D_k(T_k(t; \pi)), \quad k = 1, 2, \quad (2.5)$$

$$T_k(t; \pi) = \int_0^t Z_k(s; \pi) ds, \quad k = 1, 2, \quad (2.6)$$

$$X(t; \pi) = X(0; \pi) + A(t) - \sum_{k=1}^2 D_k(T_k(t; \pi)), \quad (2.7)$$

$$A(t) = A_q(t; \pi) + \sum_{k=1}^2 A_k(t; \pi), \quad (2.8)$$

$$T_k(t; \pi) + Y_k(t; \pi) = N_k t, \quad k = 1, 2, \quad (2.9)$$

Work conserving policies satisfy the following additional equations:

$$Q(t; \pi) \cdot \left(\sum_{k=1}^2 (N_k - Z_k(t; \pi)) \right) = 0, \quad (2.10)$$

$$\int_0^\infty \sum_{k=1}^2 (N_k - Z_k(t; \pi)) dA_q(t; \pi) = 0, \quad (2.11)$$

and

$$\sum_{k=1}^2 \int_0^\infty Q(t; \pi) dY_k(t; \pi) = 0. \quad (2.12)$$

In words, (2.10) means that there are customers in queue only when *all* servers are busy. The verbal interpretation of (2.11) is that new arrivals wait in the queue only when all servers are busy. Finally, (2.12) states that servers can only be idle when the queue is empty.

2.2 Asymptotic Framework

We consider a sequence of systems indexed by λ with increasing arrival rates $\lambda \uparrow \infty$, and increasing total number of servers N^λ but with fixed service rates μ_1, μ_2 . Our convention is to superscript any process or quantity associated with the system having arrival rate λ by λ . Although our focus in obtaining a diffusion control problem that approximates our objective in (2.2) is on the behavior of the system under diffusion scaling, we also require knowledge of the system behavior under fluid scaling. In this section, we present our asymptotic assumptions, and their consequences under both fluid and diffusion scaling.

Fluid Scaling

Assume that there are 2 numbers $a_k > 0$, $k = 1, 2$, with $a_1 + a_2 = 1$, such that the number of servers of each pool N_k^λ , $k = 1, 2$, grows with λ as follows:

$$N_k^\lambda = a_k \frac{\lambda}{\mu_k} + o(\lambda), \text{ as } \lambda \rightarrow \infty, \quad \text{or,} \quad \lim_{\lambda \rightarrow \infty} \frac{\mu_k N_k^\lambda}{\lambda} = a_k. \quad (\text{A1})$$

Condition (A1) guarantees that the total traffic intensity,

$$\rho^\lambda := \frac{\lambda}{\sum_{k=1}^2 \mu_k N_k^\lambda}, \quad (2.13)$$

converges to 1, as $\lambda \rightarrow \infty$, and hence, for large λ , the system is in *heavy traffic*. Also, in view of (A1), the quantity $a_k \lambda / \mu_k$ can be considered as the offered load of server pool k . Let

$$\mu = \left[\sum_{k=1}^2 a_k / \mu_k \right]^{-1}, \quad (2.14)$$

then λ / μ is the total offered load of the whole system. Given this definition of μ , (A1) implies that

$$N^\lambda = \frac{\lambda}{\mu} + o(\lambda), \text{ as } \lambda \rightarrow \infty, \quad \text{or,} \quad \lim_{\lambda \rightarrow \infty} \frac{\lambda}{N^\lambda} = \mu, \quad (2.15)$$

where $N^\lambda = \sum_{k=1}^2 N_k^\lambda$. Also,

$$\rho^\lambda \approx \frac{\lambda}{N^\lambda \mu}, \quad (2.16)$$

in the sense that $\lim_{\lambda \rightarrow \infty} \rho^\lambda / (\lambda / N^\lambda \mu) = 1$. Finally,

$$\lim_{\lambda \rightarrow \infty} \frac{N_k^\lambda}{N^\lambda} = \frac{a_k}{\mu_k} \mu := q_k \geq 0, \quad k = 1, 2, \quad (2.17)$$

where q_k is the limiting fraction of pool k servers out of the total number of servers. The condition $a_k > 0$ guarantees that $q_k > 0$, and hence both server pools are asymptotically non-negligible in size. Clearly, from the definition of μ in (2.14), $q_1 + q_2 = 1$ and $\sum_{k=1}^2 q_k \mu_k = \mu$.

Under assumption (A1), the arrival rate, the offered load, and the size of the different server pools are all of order N^λ . Hence one expects to get finite limits when the system processes in (2.4)-(2.12) are scaled by $1/N^\lambda$. The functional strong law of large numbers shows that this scaling leads to the fluid dynamics of the system, in the limit, as $\lambda \rightarrow \infty$. To see this, for $\lambda \uparrow \infty$, $k = 1, 2$ and a fixed sequence of routing policies $\pi^\lambda := \pi(\lambda, N^\lambda) \in \Pi$ (omitted from the following notation) let $\bar{Q}^\lambda(t) = \frac{Q^\lambda(t)}{N^\lambda}$, and $\bar{Z}_k^\lambda(t) = \frac{Z_k^\lambda(t)}{N^\lambda}$. Similarly, let $\bar{X}^\lambda(t) = \frac{X^\lambda(t)}{N^\lambda}$, $\bar{A}^\lambda(t) = \frac{A^\lambda(t)}{N^\lambda}$, $\bar{A}_k^\lambda(t) = \frac{A_k^\lambda(t)}{N^\lambda}$, $\bar{A}_q^\lambda(t) = \frac{A_q^\lambda(t)}{N^\lambda}$, $\bar{R}_k^\lambda(t) = \frac{R_k^\lambda(t)}{N^\lambda}$, $\bar{T}_k^\lambda(t) = \frac{T_k^\lambda(t)}{N^\lambda}$, and $\bar{Y}_k^\lambda(t) = \frac{Y_k^\lambda(t)}{N^\lambda}$. Finally, let $\bar{D}_k^\lambda(t) = D_k^\lambda(t) = D_k(t)$. That is, $(\bar{Q}^\lambda, \bar{Z}_k^\lambda, \bar{X}^\lambda, \bar{A}^\lambda, \bar{A}_k^\lambda, \bar{A}_q^\lambda, \bar{R}_k^\lambda, \bar{T}_k^\lambda, \bar{Y}_k^\lambda) = (Q^\lambda, Z_k^\lambda, X^\lambda, A^\lambda, A_k^\lambda, A_q^\lambda, R_k^\lambda, T_k^\lambda, Y_k^\lambda)/N^\lambda$, and $\bar{D}_k^\lambda = D_k$. Note that D_k^λ need not be divided by N^λ , due to its definition as a Poisson process with rate μ_k , which is independent of λ .

As in [2], one can show that if $(\bar{Q}^\lambda(0), \bar{Z}_k^\lambda(0), k = 1, 2)$ are bounded, then the process $(\bar{Q}^\lambda, \bar{Z}_k^\lambda, \bar{X}^\lambda, \bar{A}^\lambda, \bar{A}_k^\lambda, \bar{A}_q^\lambda, \bar{R}_k^\lambda, \bar{T}_k^\lambda, \bar{Y}_k^\lambda, \bar{D}_k^\lambda)$ is pre-compact as $\lambda \rightarrow \infty$, and hence any sequence has a converging subsequence. Denote any such *fluid limit* with a ‘‘bar’’ over the appropriate letters but with no superscript (for example, let $\bar{Q}(t)$ be a fluid limit of $\bar{Q}^\lambda(t)$). Note that equations (2.4)-(2.9) imply that the following flow balance equations hold for *any* fluid limit:

$$\bar{Q}(t) = \bar{Q}(0) + \bar{A}_q(t) - \sum_{k=1}^2 \bar{R}_k(t), \quad (2.18)$$

$$\bar{Z}_k(t) = \bar{Z}_k(0) + \bar{A}_k(t) + \bar{R}_k(t) - \mu_k \bar{T}_k(t), \quad k = 1, 2, \quad (2.19)$$

$$\bar{T}_k(t) = \int_0^t \bar{Z}_k(s) ds \quad (2.20)$$

$$\bar{X}(t) = \bar{X}(0) + \mu t - \sum_{k=1}^2 \mu_k \bar{T}_k(t), \quad (2.21)$$

$$\mu t = \bar{A}_q(t) + \sum_{k=1}^2 \bar{A}_k(t), \quad (2.22)$$

$$\bar{T}_k(t) + \bar{Y}_k(t) = q_k t. \quad (2.23)$$

Finally, for work conserving policies, conditions (2.10)-(2.12) imply:

$$\bar{Q}(t) \cdot \left(\sum_{k=1}^2 (q_k - \bar{Z}_k(t)) \right) = 0, \quad (2.24)$$

$$\int_0^\infty \sum_{k=1}^2 (q_k - \bar{Z}_k(t)) d\bar{A}_q(t) = 0, \quad (2.25)$$

and

$$\sum_{k=1}^2 \int_0^\infty \bar{Q}(t) d\bar{T}_k(t) = 0. \quad (2.26)$$

The following proposition is the same as Proposition 2.1 in [2], which shows that for every sequence of work-conserving routing policies and for every fluid limit, the quantities $\bar{Q}(t)$ and $\bar{Z}_k(t)$, $k = 1, 2$, remain constant if starting at time 0 from some appropriate initial conditions. In particular, we assume

$$\bar{Q}^\lambda(0) \rightarrow 0 \text{ and } \bar{Z}^\lambda(0) \rightarrow q_k, \quad k \in \{1, 2\}, \quad (\text{A2})$$

as $\lambda \rightarrow \infty$.

Proposition 2.1 (fluid limits) *For $\lambda > 0$, let $\pi^\lambda := \pi(\lambda, N^\lambda) \in \Pi$ be a sequence of work-conserving policies (omitted from the following notation), and let $(\bar{Q}, \bar{Z}_k, \bar{X}, \bar{A}, \bar{A}_k, \bar{A}_q, \bar{R}_k, \bar{T}_k, \bar{Y}_k, \bar{D}_k)$ be a fluid limit of the processes associated with the system, as $\lambda \rightarrow \infty$. Then, $\bar{Q}(t) = 0$ and $\bar{Z}_k(t) = q_k$, $k = 1, 2$, for all $t \geq 0$.*

Diffusion Scaling

For $\lambda > 0$ and any fixed sequence of work conserving policies $\pi^\lambda = \pi(\lambda, N^\lambda) \in \Pi$ (omitted from the notation), define the centered and scaled process $\hat{X}^\lambda(\cdot) = (\hat{X}_1^\lambda(\cdot), \hat{X}_2^\lambda(\cdot))$ as follows:

$$\hat{X}_1^\lambda(t) := \frac{Q^\lambda(t) + Z_1^\lambda(t) - N_1^\lambda}{\sqrt{N^\lambda}}, \quad (2.27)$$

$$\hat{X}_2^\lambda(t) := \frac{Z_2^\lambda(t) - N_2^\lambda}{\sqrt{N^\lambda}}. \quad (2.28)$$

Note that for $\hat{X}_2^\lambda(t) \leq 0$ for all t , and that for $k = 1, 2$, $\hat{I}_k^\lambda(t) := [\hat{X}_k^\lambda(t)]^- := -\min\{\hat{X}_k^\lambda(t), 0\}$ corresponds to the number of idle servers, scaled by $1/\sqrt{N^\lambda}$. In addition, $\hat{Q}^\lambda(t) := [\hat{X}_1^\lambda(t)]^+$ corresponds to the total queue length, again, scaled by $1/\sqrt{N^\lambda}$. Let

$$\hat{X}^\lambda(t) = \sum_{k=1}^2 \hat{X}_k^\lambda(t) = \frac{Q^\lambda(t) + \sum_{k=1}^2 Z_k^\lambda(t) - N^\lambda}{\sqrt{N^\lambda}} = \frac{X^\lambda(t) - N^\lambda}{\sqrt{N^\lambda}}. \quad (2.29)$$

Note that $\hat{I}^\lambda(t) := [\hat{X}^\lambda(t)]^-$ is the total number of idle servers, and $[\hat{X}^\lambda(t)]^+ = [\hat{X}_1^\lambda(t)]^+$ is the total queue length, both scaled by $1/\sqrt{N^\lambda}$. Finally, note that, from work conservation, if $\hat{X}_2^\lambda(t) < 0$, then $\hat{X}_1^\lambda(t) \leq 0$.

In our heavy traffic asymptotic regime, the queue size becomes large, and waiting times become small. In particular, the scaled waiting time for $\lambda > 0$ is defined as:

$$\hat{W}^\lambda(t) = \sqrt{N^\lambda} W^\lambda(t). \quad (2.30)$$

As will be shown later, in order for the above diffusion-scaled processes to have well defined limits, as $\lambda \rightarrow \infty$, we add the following assumption:

$$\sum_{k=1}^2 \mu_k N_k^\lambda = \lambda + \delta \sqrt{\lambda} + o(\sqrt{\lambda}), \text{ as } \lambda \rightarrow \infty, \quad \text{or,} \quad \lim_{\lambda \rightarrow \infty} \frac{\sum_{k=1}^2 \mu_k N_k^\lambda - \lambda}{\sqrt{\lambda}} = \delta, \quad (\text{A3})$$

for some δ , $0 < \delta < \infty$. Condition (A3) is a square-root safety staffing rule (similar to [32] and [11]). In particular, the condition $\delta > 0$ guarantees that the system is stable (or can be stable, under reasonable routing) for all λ large enough. Moreover, as is shown in [2], it guarantees that under the appropriate routing, the fraction of delayed customers is less than 1. Note that (A3) does not specify how the added safety staffing is divided between the server pools. In particular, it is possible that one server pool will have fewer servers than the nominal allocation of $q_k N^\lambda$, while the other will compensate for this deficit by having more than the nominal staffing. For $k = 1, 2$, and $\lambda > 0$, let $-\infty < \delta_k^\lambda < \infty$ satisfy:

$$\delta_k^\lambda := \frac{\mu_k N_k^\lambda - a_k \lambda}{\sqrt{\lambda}}. \quad (2.31)$$

Then $\delta_k^\lambda \sqrt{\lambda}$ is the safety capacity associated with server pool k , beyond the nominal allocation of $a_k \lambda$. In particular, one can easily verify that

$$\delta_k^\lambda = o(\sqrt{\lambda}), \text{ as } \lambda \rightarrow \infty, \quad \forall k = 1, 2, \quad (2.32)$$

and

$$\delta^\lambda := \sum_{k=1}^2 \delta_k^\lambda \rightarrow \delta, \text{ as } \lambda \rightarrow \infty. \quad (2.33)$$

Note that we do not require the individual sequences $\{\delta_k^\lambda\}_{\lambda > 0}$ to have a limit, for any value of $k = 1, 2$. All that is assumed is that their sum converges to δ .

We also require an assumption on the initial conditions under diffusion scaling:

$$\frac{Q^\lambda(0)}{\sqrt{N^\lambda}} + \sum_{k=1}^2 \frac{Z_k^\lambda(0)}{\sqrt{N^\lambda}} - \sqrt{N^\lambda} \Rightarrow \hat{X}(0), \quad (\text{A4})$$

for some proper random variable $\hat{X}(0)$ having $E\hat{X}(0)^2 < \infty$. The requirement of a finite second moment is used when solving the diffusion control problem that approximates (2.2) in Section 3. In particular, we require it in order to prove what is commonly referred to as a verification Lemma; see Lemma 3.2.

Finally, for any fixed sequence of work conserving policies $\pi^\lambda = \pi(\lambda, N^\lambda) \in \Pi$, we assume that any limiting process arising under diffusion-scaling is Markovian.

$$\text{On any subsequence } \lambda_i \text{ having } \hat{X}^{\lambda_i} \Rightarrow \hat{X} \text{ as } \lambda_i \rightarrow \infty, \text{ the process } \hat{X} \text{ is Markovian.} \quad (\text{A5})$$

We will require assumption (A5) when proving epsilon asymptotic optimality in Section 4. This is because in solving the diffusion control problem that approximates (2.2), we restrict ourselves to

the set of time-homogeneous Markovian policies. Note that although the LISF policy introduced in Section 1 that is commonly used in call centers is itself not Markovian with respect to the state $(Z_1(t) + Q(t), Z_2(t))$, [5] shows that it is in the diffusion limit; i.e., that it satisfies assumption (A5).

3 The Diffusion Control Problem

In this section, we solve the diffusion control problem that arises under assumptions (A1)-(A5) when formally passing to the limit in the control problem (2.2) for the system having arrival rate λ . We begin by providing a heuristic derivation of this diffusion control problem. Recall from the discussion in Section 2 that we expect that the performance of a preemptive policy can be closely modeled by an appropriate non-preemptive policy. Hence to make the heuristic derivation possible, we allow for preemptive policies.

Let $u_1(t)$ and $u_2(t)$ be controls that specify the proportion of idle servers in each pool at time $t \geq 0$, where

$$(u_1(t), u_2(t)) \in \mathcal{U} := \{(u_1, u_2) : 0 \leq u_1 \leq 1, 0 \leq u_2 \leq 1, \text{ and } u_1 + u_2 = 1\}.$$

Then, the system state at time t is fully specified by the total number of customers in the system $X^\lambda(t)$. Furthermore, the infinitesimal drift of the centered and scaled process \hat{X}^λ at time t is

$$\lim_{h \downarrow 0} \frac{1}{h} E \left[\hat{X}^\lambda(t+h) - \hat{X}^\lambda(t) \mid \hat{X}^\lambda(t) = x \right] = \frac{\lambda - \mu_1 N_1^\lambda - \mu_2 N_2^\lambda}{\sqrt{N^\lambda}} + \mu_1 u_1(t)[x]^- + \mu_2 u_2(t)[x]^-,$$

and the infinitesimal variance is

$$\lim_{h \downarrow 0} \frac{1}{h} E \left[\left(\hat{X}^\lambda(t+h) - \hat{X}^\lambda(t) \right)^2 \mid \hat{X}^\lambda(t) = x \right] = \frac{\lambda + \mu_1 N_1^\lambda + \mu_2 N_2^\lambda}{N^\lambda} + \frac{\mu_1 u_1(t)[x]^- + \mu_2 u_2(t)[x]^-}{\sqrt{N^\lambda}},$$

where we have ignored discretization effects that are negligible for large λ . Taking the limit in the above expressions as $\lambda \rightarrow \infty$, and using assumptions (A1), (A3), and the definition of μ in (2.14), suggests convergence to a diffusion process \hat{X} with infinitesimal drift $-\delta\sqrt{\mu} + u_1(t)\mu_1[x]^- + u_2(t)\mu_2[x]^-$ at time t when $\hat{X}(t) = x$ and infinitesimal variance 2μ . In particular, for B a standard Brownian motion and $m : \mathfrak{R} \times \mathcal{U} \rightarrow \mathfrak{R}$ defined as

$$m(x, u) = -\delta\sqrt{\mu} + u_1\mu_1x^- + u_2\mu_2x^-,$$

we expect the limiting process \hat{X} to solve the stochastic integral equation

$$\hat{X}(t) = \hat{X}(0) + \int_0^t m(\hat{X}(s), u(s)) ds + \sqrt{2\mu}B(t). \quad (3.1)$$

Note that the process \hat{X} depends on the control process u , but we suppress this from the notation. We assume the distribution of $\hat{X}(0)^2$ is such that $E\hat{X}(0)^2 < \infty$.

We restrict our attention to the set of admissible controls $\vec{u} = \{(u_1(t), u_2(t)), t \geq 0\}$ which we define as follows:

(C1) \vec{u} is a time-homogeneous Markovian policy with respect to \hat{X} , and

(C2) $(u_1(t), u_2(t)) \in \mathcal{U}$ for all $t \geq 0$.

We let \mathcal{U}_P denote the set of all admissible controls; that is, $\vec{u} \in \mathcal{U}_P$ if and only if \vec{u} satisfies conditions (C1)-(C2). The following Lemma establishes some basic properties of the process \hat{X} .

Lemma 3.1 *Conditions (C1)-(C2) imply the following:*

(i) *The process \hat{X} has a strong solution. In particular, \hat{X} has continuous sample paths.*

(ii) *The process \hat{X} has a unique steady-state distribution. Furthermore, if $\hat{X}(\infty)$ is a random variable having the steady-state distribution, $\hat{X}(t) \Rightarrow \hat{X}(\infty)$ as $t \rightarrow \infty$.*

(iii) *The limits*

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{1}{t} E [\hat{X}(t)] &= 0 \\ \lim_{t \rightarrow \infty} \frac{1}{t} E \int_0^t [\hat{X}(s)]^+ ds &= E [\hat{X}(\infty)^+] \\ \lim_{t \rightarrow \infty} \frac{1}{t} E \int_0^t [\hat{X}(s)]^- ds &= E[\hat{X}(\infty)^-] \\ \lim_{t \rightarrow \infty} \frac{1}{t} E \int_0^t u_i(s) [\hat{X}(s)]^- ds &= E[u_i(\infty)\hat{X}(\infty)^-], \quad i = 1, 2 \end{aligned}$$

hold.

The approximating diffusion control problem that parallels the formulation for the discrete-event system in (2.2) is

$$\begin{aligned} &\text{minimize}_{\vec{u} \in \mathcal{U}_P} E\hat{X}(\infty)^+ \\ &\text{subject to: } \frac{E[u_1(\infty)\hat{X}(\infty)^-]}{E\hat{X}(\infty)^-} = f_1. \end{aligned} \tag{3.2}$$

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

3.1 The Solution Approach and the Lagrangian Relaxation Problem

We first simplify the diffusion control problem (3.2) by writing the constraint solely in terms of $E\hat{X}(\infty)^-$. Lemma 3.1 part (iii) implies that

$$\frac{1}{t}E \left[\hat{X}(t) - \hat{X}(0) \right] \rightarrow 0,$$

as $t \rightarrow \infty$. Furthermore, from the stochastic equation (3.1) and Lemma 3.1 part (iii),

$$\begin{aligned} \frac{1}{t}E \left[\hat{X}(t) - \hat{X}(0) \right] &= -\delta\sqrt{\mu} + \frac{1}{t}E \left[\mu_1 \int_0^t u_1(s)[\hat{X}(s)]^- ds + \mu_2 \int_0^t u_2(s)[\hat{X}(s)]^- ds \right] \\ &\rightarrow -\delta\sqrt{\mu} + \mu_1 E[u_1(\infty)\hat{X}(\infty)^-] + \mu_2 E[u_2(\infty)\hat{X}(\infty)^-], \end{aligned}$$

as $t \rightarrow \infty$. It then follows that

$$\delta\sqrt{\mu} = \mu_1 E[u_1(\infty)\hat{X}(\infty)^-] + \mu_2 E[u_2(\infty)\hat{X}(\infty)^-].$$

Since $u_1(t) + u_2(t) = 1$ for all $t \geq 0$,

$$E[u_1(\infty)\hat{X}(\infty)^-] + E[u_2(\infty)\hat{X}(\infty)^-] = E[\hat{X}(\infty)^-].$$

Substitution then shows

$$\delta\sqrt{\mu} = (\mu_1 - \mu_2)E[u_1(\infty)\hat{X}(\infty)^-] + \mu_2 E[\hat{X}(\infty)^-].$$

The constraint $E[u_1(\infty)\hat{X}(\infty)^-] = f_1 E\hat{X}(\infty)^-$ holds if and only if

$$E\hat{X}(\infty)^- = \frac{\delta\sqrt{\mu}}{f_1\mu_1 + (1-f_1)\mu_2}. \quad (3.3)$$

We conclude that an equivalent approximating diffusion control problem to (3.2) is

$$\begin{aligned} &\text{minimize}_{\bar{u} \in \mathcal{U}_P} E\hat{X}(\infty)^+ \\ &\text{subject to:} \quad E\hat{X}(\infty)^- = \frac{\delta\sqrt{\mu}}{f_1\mu_1 + (1-f_1)\mu_2}. \end{aligned} \quad (3.4)$$

Remark 3.1 When $f_1 = 1$, we expect the constraint in (3.4) to be consistent with the limiting expression for $E\hat{X}^\lambda(\infty)^-$, as λ becomes large, when the FSF (fastest server first) routing policy is used in the original system. This is because [2] shows that FSF asymptotically minimizes the steady-state queue-length by allowing only the slow servers to idle. Propositions 4.2 and 4.4 in [2], the continuous mapping theorem, and the uniform integrability established in Proposition A.3 in the technical appendix show that

$$E\hat{X}^\lambda(\infty)^- \rightarrow \frac{\delta\sqrt{\mu}}{\mu_1},$$

as $\lambda \rightarrow \infty$. Hence the constraint in (3.4) is consistent with that known result.

Since the diffusion control problem in (3.4) involves a constraint, to apply the standard methods in, for example [22] or [34], we first formulate the Lagrangian relaxation problem. Let $\Delta \in \mathfrak{R}$ be a penalty parameter, and define

$$d := \frac{\delta\sqrt{\mu}}{f_1\mu_1 + (1 - f_1)\mu_2}.$$

Our solution approach is to solve

$$\text{minimize}_{\vec{u} \in \mathcal{U}_P} E\hat{X}(\infty)^+ + \Delta \left(E\hat{X}(\infty)^- - d \right). \quad (3.5)$$

for a range of penalty parameters. Then, we search for the penalty parameter Δ^* such that the constraint $E\hat{X}(\infty)^- = d$ is satisfied. In this case, the solution to (3.5) is also the solution to (3.4), and so we will have solved the approximating diffusion control problem. This argument is made rigorous in Section 3.3.

3.2 Solving the Lagrangian Relaxation Problem

For a fixed $\Delta \in \mathfrak{R}$, the following verification Lemma is necessary to characterize the form of a control $\vec{u} \in \mathcal{U}_P$ that solves (3.5).

Lemma 3.2 *Suppose there exists a twice-continuously differentiable function $V : \mathfrak{R} \rightarrow \mathfrak{R}$ and a constant $\kappa \in \mathfrak{R}$ that solve*

$$\mu V''(x) + \inf_{u \in \mathcal{U}_P} m(x, u)V'(x) + x^+ + \Delta(x^- - d) = \kappa, \text{ for all } x \in \mathfrak{R}. \quad (3.6)$$

Also assume there exist $b_1, b_2 \in \mathfrak{R}$ such that

$$|V(x)| \leq b_1x^2 + b_2$$

for all $x \in \mathfrak{R}$. Then, if \hat{X} satisfies (3.1) under some admissible control $u \in \mathcal{U}_P$,

$$\liminf_{t \rightarrow \infty} \frac{E \left[\int_0^t \hat{X}(s)^+ + \Delta \left(\hat{X}(s)^- - d \right) ds \right]}{t} \geq \kappa.$$

Observe that for any state $x \leq 0$ such that $V'(x) > 0$,

$$\text{argmin}_{u \in \mathcal{U}} m(x, u)V'(x) = (1, 0),$$

and for any state $x \leq 0$ such that $V'(x) < 0$,

$$\text{argmin}_{u \in \mathcal{U}} m(x, u)V'(x) = (0, 1).$$

Then, assuming that the function V' is increasing, we expect that there exists a control of threshold form that solves (3.6). Specifically, a threshold control at level $L > 0$ has

$$(u_1(t), u_2(t)) = \left(\mathbf{1}\{-L \leq \hat{X}(t) \leq 0\}, \mathbf{1}\{\hat{X}(t) < -L\} \right) \quad (3.7)$$

for all $t \geq 0$. The associated infinitesimal drift is

$$m_L(x) = \begin{cases} -\delta\sqrt{\mu} & x \geq 0 \\ -\delta\sqrt{\mu} - \mu_1 x & -L \leq x < 0 \\ -\delta\sqrt{\mu} - \mu_2 x & x < -L. \end{cases} \quad (3.8)$$

For intuition, a threshold control at level L for the diffusion corresponds to a preemptive policy in the discrete event system that idles only slow servers when $-L \leq \hat{X}^\lambda(t) \leq 0$ and idles only fast servers when $\hat{X}^\lambda(t) < -L$. Note that a threshold control obviously satisfies conditions (C1) and (C2), and so is admissible.

We search for a control that satisfies the conditions of Lemma 3.2 within the class of threshold controls. The following lemma characterizes the cost associated with a threshold control.

Lemma 3.3 *Suppose there exist a twice-continuously differentiable function $V : \Re \rightarrow \Re$ and constants $\kappa \in \Re, L > 0$ that solve*

$$\begin{aligned} \mu V''(x) - \delta\sqrt{\mu}V'(x) + x - \Delta d &= \kappa, & x > 0 \\ \mu V''(x) - (\delta\sqrt{\mu} + \mu_1 x)V'(x) - \Delta(x + d) &= \kappa, & -L \leq x \leq 0 \\ \mu V''(x) - (\delta\sqrt{\mu} + \mu_2 x)V'(x) - \Delta(x + d) &= \kappa, & x < -L. \end{aligned} \quad (3.9)$$

Also assume there exist $b_1, b_2 \in \Re$ such that

$$|V(x)| \leq b_1 x^2 + b_2$$

for all $x \in \Re$. Then, if \hat{X} satisfies (3.1) under the threshold control at level L ,

$$E\hat{X}(\infty)^+ + \Delta \left(E\hat{X}(\infty)^- - d \right) = \kappa.$$

We now present the solution to (3.9), and show it satisfies the conditions of Lemma 3.3. This will allow us to show that there exists a threshold control that satisfies the conditions of Lemma 3.2, and so solves the Lagrangian relaxation problem (3.5). Let ϕ and Φ be respectively the probability density and cumulative distribution functions for a standard normal random variable, and let h be the associated hazard rate function. Define

$$f(x, L) := \sqrt{\frac{2\pi}{\mu_1}} \exp\left(\frac{1}{2} \left(\frac{\delta}{\sqrt{\mu_1}} + \sqrt{\frac{\mu_1}{\mu}} x\right)^2\right) \left[\begin{aligned} &\frac{1}{\sqrt{\mu_1}} \left(\phi\left(\frac{\delta}{\sqrt{\mu_1}} - L\sqrt{\frac{\mu_1}{\mu}}\right) - \phi\left(\frac{\delta}{\sqrt{\mu_1}} + x\sqrt{\frac{\mu_1}{\mu}}\right) \right) \\ &+ \left(\frac{1}{\sqrt{\mu_2}} h\left(L\sqrt{\frac{\mu_2}{\mu}} - \frac{\delta}{\sqrt{\mu_2}}\right) + \frac{\delta}{\mu_2} - \frac{\delta}{\mu_1} \right) \\ &\times \left(\Phi\left(\frac{\delta}{\sqrt{\mu_1}} + \sqrt{\frac{\mu_1}{\mu}} x\right) - \Phi\left(\frac{\delta}{\sqrt{\mu_1}} - L\sqrt{\frac{\mu_1}{\mu}}\right) \right) \end{aligned} \right].$$

Suppose there exists L_Δ^* that solves

$$\frac{1}{\Delta\delta^2} - \frac{1}{\mu_2} = f(0, L) + \frac{1}{\delta\sqrt{\mu_2}} h\left(L\sqrt{\frac{\mu_2}{\mu}} - \frac{\delta}{\sqrt{\mu_2}}\right). \quad (3.10)$$

Then the function defined by

$$V(x) := \begin{cases} \int_0^x V'(y)dy & \text{if } x \geq 0 \\ \int_x^0 V'(y)dy & \text{if } x < 0 \end{cases}$$

for

$$V'(x) := \begin{cases} \frac{x}{\delta\sqrt{\mu}} + \frac{1}{\delta^2} - \frac{\Delta}{\delta\sqrt{\mu_2}} \left(h\left(L_\Delta^*\sqrt{\frac{\mu_2}{\mu}} - \frac{\delta}{\sqrt{\mu_2}}\right) + \frac{\delta}{\sqrt{\mu_2}} \right) & x > 0 \\ \Delta f(x, L_\Delta^*) & -L_\Delta^* \leq x \leq 0 \\ \frac{\Delta}{\mu_2} \sqrt{2\pi} \exp\left(\frac{1}{2}\left(\frac{\delta}{\sqrt{\mu_2}} + \sqrt{\frac{\mu_2}{\mu}}x\right)^2\right) h\left(L_\Delta^*\sqrt{\frac{\mu_2}{\mu}} - \frac{\delta}{\sqrt{\mu_2}}\right) \Phi\left(\frac{\delta}{\sqrt{\mu_2}} + \sqrt{\frac{\mu_2}{\mu}}x\right) - \frac{\Delta}{\mu_2} & x < -L_\Delta^* \end{cases}, \quad (3.11)$$

is a twice-continuously differentiable function. Furthermore, for

$$\kappa := \left[\frac{\Delta}{\sqrt{\mu_2}} \left(h\left(-L_\Delta^*\sqrt{\frac{\mu_2}{\mu}} - \frac{\delta}{\sqrt{\mu_2}}\right) + \frac{\delta}{\sqrt{\mu_2}} \right) - \frac{\Delta d}{\sqrt{\mu}} \right] \sqrt{\mu}, \quad (3.12)$$

the function V solves (3.9). Finally, it is straightforward to verify that $\lim_{x \rightarrow \infty} |V''(x)| < \infty$ and $\lim_{x \rightarrow -\infty} |V''(x)| < \infty$. Then, there exist constants $b_1, b_2 \in \mathfrak{R}$ such that

$$|V(x)| \leq b_1 x^2 + b_2 \text{ for all } x \in \mathfrak{R}.$$

Hence the conditions of Lemma 3.3 will be satisfied.

The following Lemma shows the condition under which L_Δ^* exists.

Lemma 3.4 *Suppose*

$$0 < \Delta < \left[\frac{\delta^2}{\mu_2} + \frac{\delta}{\sqrt{\mu_2}} h\left(\frac{-\delta}{\sqrt{\mu_2}}\right) \right]^{-1}. \quad (3.13)$$

Then, there exists L_Δ^ that solves (3.10).*

Figure 3.1 graphs the function V' for a problem with parameters $\mu_1 = 1$, $\mu_2 = 2$, $\delta = 1$, $\mu = 1.5$, and penalty parameter $\Delta = 0.3$ (We choose $\Delta = 0.3$ because we show in Section 3.3 that $\Delta = 0.3$ is the correct choice in order to have a solution to the original approximating diffusion control problem (3.4)). In this case, it follows from equation (3.10) that $L_\Delta^* = 1.5343$. Observe that $V'(x) > 0$ for all $x \in (-L_{0.3}^*, 0]$ and $V'(x) < 0$ for all $x < -L_{0.3}^*$, which implies

$$\begin{aligned} \operatorname{argmin}_{u \in \mathcal{U}} m(x, u) V'(x) &= (1, 0) \text{ if } -L_{0.3}^* \leq x \leq 0 \\ \operatorname{argmin}_{u \in \mathcal{U}} m(x, u) V'(x) &= (0, 1) \text{ if } x \leq -L_{0.3}^*. \end{aligned}$$

Hence our numerics suggest that a solution to (3.9) satisfies the conditions of the verification Lemma 3.2. The following proposition makes this intuition rigorous.

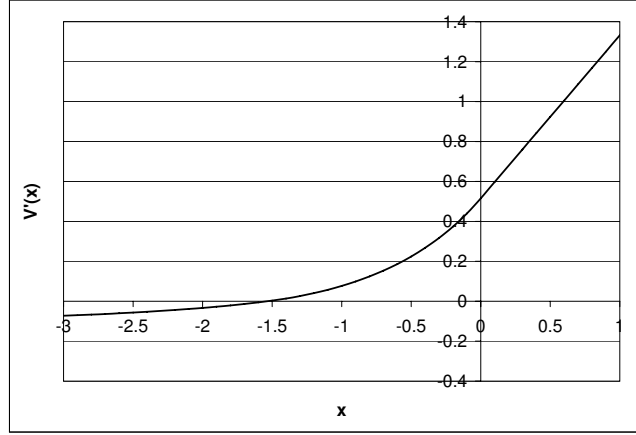


Figure 3.1: The function V' for a problem having $\mu_1 = 1$, $\mu_2 = 2$, $\delta = 1$, $\mu = 1.5$, and penalty parameter $\Delta = 0.3$. Then, $L_{0.3}^* = 1.5343$.

Proposition 3.1 *Assume Δ satisfies (3.13). Let \hat{X}_{Δ}^* satisfy (3.1) under the threshold control at level L_{Δ}^* , and let \hat{X} satisfy (3.1) under any other control $u \in \mathcal{U}_P$. Then,*

$$E\hat{X}_{\Delta}^*(\infty)^+ + \Delta \left(E\hat{X}_{\Delta}^*(\infty)^- - d \right) \leq E\hat{X}(\infty)^+ + \Delta \left(E\hat{X}(\infty)^- - d \right).$$

3.3 Solving the Diffusion Control Problem

Let \hat{X}_{Δ}^* satisfy (3.1) under the threshold control at level L^* when the penalty parameter is Δ . To solve the approximating diffusion control problem (3.4), it suffices to find Δ^* that satisfies condition (3.13) under which

$$E\hat{X}_{\Delta^*}^*(\infty)^- = d.$$

To see this, observe that it follows from the choice of Δ^* and from Proposition 3.1 that for any other \hat{X} satisfying the stochastic integral equation (3.1) under control $u \in \mathcal{U}_P$,

$$E\hat{X}_{\Delta^*}^*(\infty)^+ = E\hat{X}_{\Delta^*}^*(\infty)^+ + \Delta^* \left(E\hat{X}_{\Delta^*}^*(\infty)^- - d \right) \leq E\hat{X}(\infty)^+ + \Delta^* \left(E\hat{X}(\infty)^- - d \right).$$

In particular, the above inequality must hold under any control $u \in \mathcal{U}_P$ having $E\hat{X}(\infty)^- = d$, and so

$$E\hat{X}_{\Delta^*}^*(\infty) \leq E\hat{X}(\infty)^+.$$

We conclude that the above inequality holds for any $u \in \mathcal{U}_P$ satisfying the constraint in the approximating diffusion control problem (3.4), and so threshold control at the level L^* associated with the penalty parameter Δ^* solves (3.4).

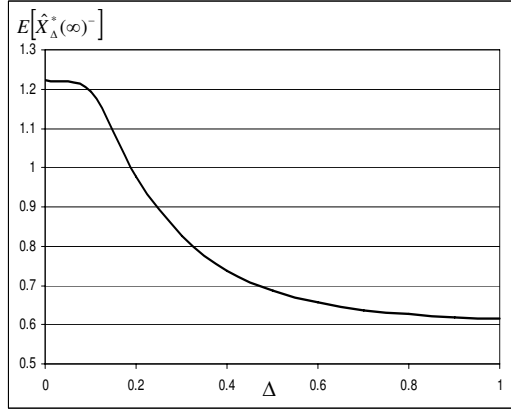


Figure 3.2: The value of $E[\hat{X}_{\Delta}^*(\infty)^-]$ as a function of the penalty parameter Δ for a problem having $\mu_1 = 1$, $\mu_2 = 2$, $\delta = 1$, $\mu = 1.5$, and $f_1 = 0.5$. Then, from (3.4), $d = 0.816497$.

Figure 3.2 graphs $E\hat{X}_{\Delta}^*(\infty)^-$ as a function of the penalty parameter Δ for a problem with parameters $\mu_1 = 1$, $\mu_2 = 2$, $\delta = 1$, $\mu = 1.5$, and $f_1 = 0.5$, which implies from (3.4) that $d = 0.816497$. In this case, we see numerically that $\Delta^* = 0.3$, and it is straightforward to verify that condition (3.13) is satisfied. (Recall from Section 3.1 that the associated optimum threshold level is $L_{0.3}^* = 1.5343$, as demonstrated in Figure 3.1.) The following lemma ensures the existence of a Δ^* that satisfies condition (3.13) for any problem parameters.

Lemma 3.5 *There exists Δ^* that satisfies (3.13) such that $E\hat{X}_{\Delta^*}^*(\infty)^- = d$.*

Finally, we state our main theorem, that provides a solution to the approximating diffusion control problem (3.4).

Theorem 3.1 *Let Δ^* be such that $E\hat{X}_{\Delta^*}^*(\infty)^- = d$. Then, for any \hat{X} that satisfies (3.1) under control $u \in \mathcal{U}_P$ and has $E\hat{X}(\infty)^- = d$,*

$$E\hat{X}_{\Delta^*}^*(\infty)^+ \leq E\hat{X}(\infty)^+.$$

Note that while Δ^* and d were used in this section to help solve the diffusion control problem (3.4), now that it has been established that the optimal solution is of threshold type, it is no longer necessary to find the values of these parameters in solving this problem. Instead, it is sufficient to find the threshold level L^* that would ensure that the constraint of (3.4) with its particular choice of f_1 is met. To elaborate, we next spell out the value of f_1 as a function of the threshold level L . A simple search can then allow one to find L as a function of f_1 .

Let \hat{X}_L satisfy (3.1) under the threshold control at level L . In order to satisfy the constraint

in the original approximating diffusion control problem (3.2), we have

$$f_1 = \frac{E[u_1(\infty)\hat{X}_L(\infty)^-]}{E[\hat{X}_L(\infty)^-]} \quad (3.14)$$

$$= \frac{E[\hat{X}_L(\infty)^- | -L \leq \hat{X}_L(\infty) < 0] P(-L \leq \hat{X}_L(\infty) < 0)}{E[\hat{X}_L(\infty)^- | \hat{X}_L(\infty) < -L] P(\hat{X}_L(\infty) < -L) + E[\hat{X}_L(\infty)^- | -L \leq \hat{X}_L(\infty) < 0] P(-L \leq \hat{X}_L(\infty) < 0)}$$

By Proposition 18.3 in Browne and Whitt [12]

$$E[\hat{X}_L(\infty)^- | \hat{X}_L(\infty) < -L] = \frac{\delta\sqrt{\mu}}{\mu_2} + \sqrt{\frac{\mu}{\mu_2}} h \left(L\sqrt{\frac{\mu_2}{\mu}} - \frac{\delta}{\sqrt{\mu_2}} \right), \quad (3.15)$$

$$E[\hat{X}_L(\infty)^- | -L \leq \hat{X}_L(\infty) < 0] = \frac{\delta\sqrt{\mu}}{\mu_1} - \sqrt{\frac{\mu}{\mu_1}} \frac{\phi\left(\frac{\delta}{\sqrt{\mu_1}} - L\sqrt{\frac{\mu_1}{\mu}}\right) - \phi\left(\frac{\delta}{\sqrt{\mu_1}}\right)}{\Phi\left(\frac{\delta}{\sqrt{\mu_1}}\right) - \Phi\left(\frac{\delta}{\sqrt{\mu_1}} - L\sqrt{\frac{\mu_1}{\mu}}\right)}, \quad (3.16)$$

$$P(\hat{X}_L(\infty) < -L) = 1/D(L), \quad (3.17)$$

and

$$P(-L \leq \hat{X}_L(\infty) < 0) = \sqrt{\frac{\mu_2}{\mu_1}} h \left(L\sqrt{\frac{\mu_2}{\mu}} - \frac{\delta}{\sqrt{\mu_2}} \right) \frac{\Phi\left(\frac{\delta}{\sqrt{\mu_1}}\right) - \Phi\left(\frac{\delta}{\sqrt{\mu_1}} - L\sqrt{\frac{\mu_1}{\mu}}\right)}{\phi\left(\frac{\delta}{\sqrt{\mu_1}} - L\sqrt{\frac{\mu_1}{\mu}}\right)} \frac{1}{D(L)}, \quad (3.18)$$

where

$$D(L) = 1 + \sqrt{\frac{\mu_2}{\mu_1}} h \left(L\sqrt{\frac{\mu_2}{\mu}} - \frac{\delta}{\sqrt{\mu_2}} \right) \frac{\Phi\left(\frac{\delta}{\sqrt{\mu_1}}\right) - \Phi\left(\frac{\delta}{\sqrt{\mu_1}} - L\sqrt{\frac{\mu_1}{\mu}}\right)}{\phi\left(\frac{\delta}{\sqrt{\mu_1}} - L\sqrt{\frac{\mu_1}{\mu}}\right)} \quad (3.19)$$

$$+ \frac{\sqrt{\mu_2}}{\delta} h \left(L\sqrt{\frac{\mu_2}{\mu}} - \frac{\delta}{\sqrt{\mu_2}} \right) \frac{\phi\left(\frac{\delta}{\sqrt{\mu_1}}\right)}{\phi\left(\frac{\delta}{\sqrt{\mu_1}} - L\sqrt{\frac{\mu_1}{\mu}}\right)}. \quad (3.20)$$

Plugging (3.15)-(3.19) back into (3.14) yields the desired relationship between L and f_1 .

To be able to solve for the threshold L given any value of the fraction f_1 , one still needs to show that (3.14) has a solution in L for any value of $f_1 \in (0, 1)$. This can be easily established by noting that according to the proof of Lemma 3.5 we have that for any value of d such that

$$\frac{\delta\sqrt{\mu}}{\mu_2} < d < \frac{\delta\sqrt{\mu}}{\mu_1}, \quad (3.21)$$

there exists an $L > 0$ such that $E\hat{X}_L(\infty)^- = d$. But by (3.3) the constraint $f_1 = \frac{E[u_1(\infty)\hat{X}_L(\infty)^-]}{E[\hat{X}_L(\infty)^-]}$ holds if and only if the constraint $E\hat{X}_L(\infty)^- = d$ holds with $d = \frac{\delta\sqrt{\mu}}{f_1\mu_1 + (1-f_1)\mu_2}$. Finally, noting that for $0 < f_1 < 1$ the above value of d always satisfies (3.21) the argument is complete.

4 The Proposed Policy

The threshold policy specified in Theorem 3.1 motivates our proposed routing policy. A natural translation of the optimal threshold policy for the approximating diffusion control to a routing policy for the original system is as follows.

The Threshold Policy: Let $L(f_1)$ be such that

$$E\hat{X}_L(\infty)^- = \frac{\delta\sqrt{\mu}}{f_1\mu_1 + (1-f_1)\mu_2},$$

so that the constraint in the equivalent approximating diffusion control problem (3.4) is satisfied, as explained at the end of Section 3.3. Fix λ and $N^{\vec{\lambda}}$, and let $L^\lambda = N^\lambda - L(f_1)\sqrt{N^\lambda}$. Then the threshold policy assigns newly arriving customers to the faster (pool 2) servers first when the total number of customers in the system is greater than or equal to L^λ , and it assigns such customers to the slower (pool 1) servers first, when the number in the system is below L^λ .

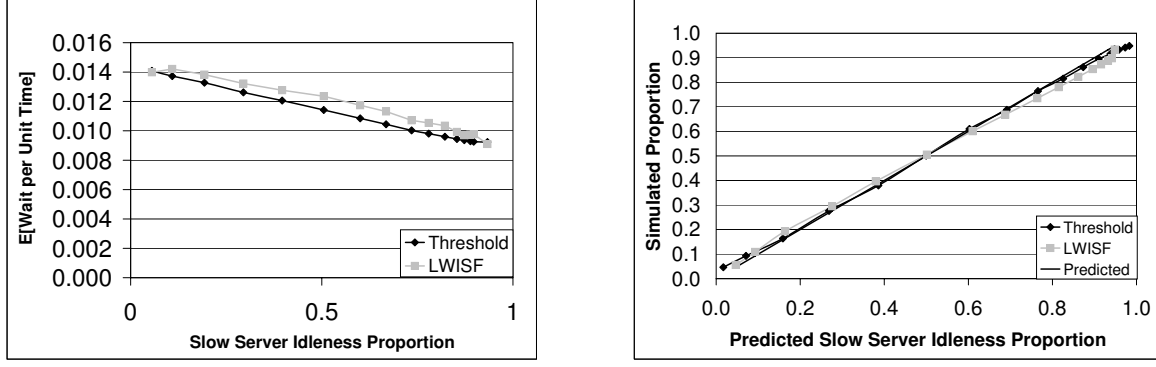
We expect our proposed threshold policy to perform well for large values of λ because the diffusion control problem we solve in Section 3 arises when formally passing to the limit in (2.2) as λ becomes large. Recall from Figure 1.2 in Section 1 that, for the same slow server idleness proportion f_1 , our proposed threshold policy has a lower expected waiting time than the LISF policy that call centers often use to route calls. However, the LISF policy can achieve only one f_1 , and our proposed threshold policy can achieve any $f_1 \in [0, 1]$. Therefore, for comparison purposes, we extend the LISF policy to a variant of the LISF policy that can achieve any $f_1 \in [0, 1]$ by weighting the different server pools and routing calls to the server that has experienced the longest weighted idle times.

The Longest-Weighted-Idle-Server-First (LWISF) Policy: Let $w_1 \in [0, 1]$ and $w_2 \in [0, 1]$ be such that $w_1 + w_2 = 1$. Suppose that at the time a new customer arrives there is at least one server in each pool free, and let i_1 represent the idle time of the server in pool 1 that has been idle the longest and i_2 represent the idle time of the server in pool 2 that has been idle the longest. Then, the LWISF policy routes the customer to the pool 1 server that has been idle the longest if $w_1 i_1 > w_2 i_2$, and otherwise routes the customer to the pool 2 server that has been idle the longest. If servers in only one pool are free when a new customer arrives, the LWISF policy routes the customer to the server that has been idle the longest in that pool. Finally, if no servers in any pool are free at the time a new customer arrives, the customer queues.

Note that for the LWISF policy, we expect that

$$f_1 \approx \frac{\mu_1 N_1 / w_1}{\mu_1 N_1 / w_1 + \mu_2 N_2 / w_2}. \quad (4.1)$$

To see (4.1), let U_k be the average idleness period duration for a server in pool k . Then, the slow



(a) The steady-state expected customer waiting time.

(b) The slow server idleness proportions.

Figure 4.1: A comparison of the performance of the threshold and LWISF policies for $N_1 = 300$, $N_2 = 200$, $\mu_1 = 1$, and $\mu_2 = 2$.

server idleness proportion is

$$f_1 = \frac{N_1 \left(\frac{U_1}{U_1+1/\mu_1} \right)}{N_1 \left(\frac{U_1}{U_1+1/\mu_1} \right) + N_2 \left(\frac{U_2}{U_2+1/\mu_2} \right)}.$$

Similar to the LISF policy in [5], the LWISF policy tends to equalize $w_1 U_1$ and $w_2 U_2$. Hence, substituting into the above expression, and noting that idle periods are much shorter than service periods ($U_k = o(1)$ in our asymptotic regime), shows

$$f_1 = \frac{N_1 \left(\frac{U_1}{U_1+1/\mu_1} \right)}{N_1 \left(\frac{U_1}{U_1+1/\mu_1} \right) + N_2 \left(\frac{w_1 U_1 / w_2}{w_1 U_1 / w_2 + 1/\mu_2} \right)} \approx \frac{N_1}{N_1 + \frac{\mu_2 w_1}{\mu_1 w_2} N_2},$$

which is equivalent to the expression (4.1).

The expression for f_1 in (4.1) shows that the LWISF policy can attain any value of f_1 between 0 and 1 by varying the weights w_1 and w_2 .

Figure 4.1 shows that our proposed threshold policy outperforms the LWISF policy in terms of the steady-state expected waiting time for a range of targeted slow server idleness proportion values f_1 . We simulate a system having parameters $N_1 = 300$, $N_2 = 200$, $\mu_1 = 1$, $\mu_2 = 2$, and $\lambda = 674$, and vary the threshold level L and the weights w_1 and w_2 so as to achieve the desired f_1 for both the threshold and LWISF policies. We report the average of the mean number of customers waiting and the mean slow server idleness proportion over 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 (in which statistics are recorded). For values of f_1 between approximately 0.4 and 0.8, the percentage decrease using the LWISF policy as compared to the threshold policy is

between 6% and 8%. It is to be expected that there is no performance difference between the two policies when either $f_1 = 1$ or $f_1 = 0$ because then both policies are exactly the same, either faster-server-first ($f_1 = 1$) or slower-server-first ($f_1 = 0$).

Unfortunately, establishing asymptotic optimality is technically very challenging. The difficulty is in establishing state-space collapse. Specifically, one would expect that as $\lambda \rightarrow \infty$ the number of fast idle servers will be negligibly small every time the number in the system exceeds the threshold, and, similarly, the number of slow idle servers will be negligibly small when the number in the system is below the threshold. This suggests a non-continuous form of state-space collapse at the threshold level $L(f_1)$. Existing techniques do not allow for such a form of state-space collapse (for example, both [17] and [27] assume a continuous form of state-space collapse). Therefore, we show that a “continuous adjustment” of the non-preemptive threshold policy into what we call the ϵ -threshold policy is feasible (that is, satisfies the fairness constraint), and asymptotically, as $\lambda \rightarrow \infty$ has an expected waiting time that is ϵ away from the optimal solution, where ϵ is arbitrarily small.

We start with some relevant definitions.

Definition (Asymptotic Feasibility): Consider a sequence of systems indexed by λ , with N_k^λ servers in pool k ($k = 1, 2$). Suppose that the system operates under a sequence of policies $\pi = \pi(\lambda, \vec{N}^\lambda) \in \Pi$, and that the process $\vec{X}^\lambda(\cdot; \pi)$ has a steady-state for every value of λ . Then, π is *asymptotically feasible* with respect to (2.2) if

$$\lim_{\lambda \rightarrow \infty} \frac{EI_1^\lambda(\infty; \pi)}{EI^\lambda(\infty; \pi)} = f_1.$$

Definition (ϵ -Asymptotic Optimality): Consider a sequence of systems indexed by λ , with N_k^λ servers in pool k ($k = 1, 2$). Suppose that the system operates under an *asymptotically feasible* sequence of policies $\pi = \pi(\lambda, \vec{N}^\lambda)$. Let $\epsilon > 0$. Then π is *ϵ -asymptotically optimal* with respect to (2.2) if for any other sequence of asymptotically feasible policies $\pi' = \pi'(\lambda, \vec{N}^\lambda) \in \Pi$ we have

$$\limsup_{\lambda \rightarrow \infty} E\hat{W}^\lambda(\infty; \pi) \leq \liminf_{\lambda \rightarrow \infty} E\hat{W}^\lambda(\infty; \pi') + \epsilon.$$

The derivation of our proposed ϵ -threshold policy is done in three steps.

1. In Subsection 4.1, we propose the ϵ -threshold policy \hat{TH}_ϵ for the diffusion control problem (3.4), and establish its ϵ -optimality. Note that the ϵ -threshold policy modifies the threshold control at a given level L into a control that has a continuous infinitesimal drift.
2. In Subsection 4.2, we propose the ϵ -threshold policy TH_ϵ for the original queueing problem (2.2), establish its asymptotic feasibility, and show that its performance asymptotically approaches the performance of \hat{TH}_ϵ .

3. Finally, in Subsection 4.3, we establish ϵ - asymptotic optimality by showing that the optimal objective value in (3.4) is a lower bound on the scaled objective value in (2.2) under any asymptotically feasible policy.

4.1 The diffusion ϵ - threshold policy

Let \hat{X}^* denote the diffusion in Theorem 3.1 that satisfies (3.1) under the optimal threshold control at level L^* when the penalty parameter is Δ^* . (For notational simplicity, we have dropped the subscript Δ^* .) Then, \hat{X}^* has a piecewise linear drift $m(x)$ as in (3.8) and an infinitesimal variance 2μ . 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 that, we first propose upper bound and lower bound diffusion processes: \hat{X}_η^u and \hat{X}_η^l , respectively, such that:

$$E \left[\hat{X}_\eta^l(\infty) \right]^+ \leq E \left[\hat{X}^*(\infty) \right]^+ \leq E \left[\hat{X}_\eta^u(\infty) \right]^+, \quad (4.2)$$

and

$$E \left[\hat{X}_\eta^u(\infty) \right]^- \leq E \left[\hat{X}^*(\infty) \right]^- \leq E \left[\hat{X}_\eta^l(\infty) \right]^-, \quad (4.3)$$

where $\eta > 0$ is some parameter whose role will become clear later.

As the next step, we define a process $\hat{X}_{\eta,\gamma}$ whose infinitesimal drift is a convex combination of the drift terms of \hat{X}_η^u and \hat{X}_η^l , with weights γ and $1 - \gamma$ respectively ($0 \leq \gamma \leq 1$). We then prove the existence of $\gamma(\eta)$ such that

$$E \left[\hat{X}_{\eta,\gamma(\eta)}(\infty) \right]^- = E \left[\hat{X}^*(\infty) \right]^-. \quad (4.4)$$

In particular, if $T\hat{H}_\epsilon(\eta, \gamma)$ is a control policy that corresponds to the diffusion process $\hat{X}_{\eta,\gamma}$, then $T\hat{H}_\epsilon(\eta, \gamma(\eta))$ is feasible for the problem (3.4), for all values of $\eta > 0$.

Finally, to prove the ϵ - optimality of $T\hat{H}_\epsilon(\eta, \gamma(\eta))$ we show that

$$\lim_{\eta \rightarrow 0} E \left[\hat{X}_{\eta,\gamma(\eta)}(\infty) \right]^+ = E \left[\hat{X}^*(\infty) \right]^+. \quad (4.5)$$

The upper and lower bound diffusion processes

Fix a threshold value L and let the infinitesimal drift associated with the resulting diffusion process be

$$m(x) = \begin{cases} -\delta\sqrt{\mu} & x \geq 0 \\ -\delta\sqrt{\mu} - \mu_1 x & -L \leq x < 0 \\ -\delta\sqrt{\mu} - \mu_2 x & x < -L. \end{cases} \quad (4.6)$$

Let $0 < \eta < L$. We define \hat{X}_η^u and \hat{X}_η^l to be two diffusion processes with infinitesimal variance 2μ and infinitesimal drift:

$$m_\eta^u(x) = \begin{cases} -\delta\sqrt{\mu} & x \geq 0 \\ -\delta\sqrt{\mu} - \mu_1 x & -L + \eta \leq x < 0 \\ -\delta\sqrt{\mu} - \mu_1 x + \frac{L}{\eta}(\mu_2 - \mu_1)(\eta - L - x) & -L \leq x < -L + \eta \\ -\delta\sqrt{\mu} - \mu_2 x & x < -L. \end{cases} \quad (4.7)$$

and

$$m_\eta^l(x) = \begin{cases} -\delta\sqrt{\mu} & x \geq 0 \\ -\delta\sqrt{\mu} - \mu_1 x & -L \leq x < 0 \\ -\delta\sqrt{\mu} - \mu_2 x + \frac{L}{\eta}(\mu_2 - \mu_1)(-x - L - \eta) & -L - \eta \leq x < -L \\ -\delta\sqrt{\mu} - \mu_2 x & x < -L - \eta. \end{cases} \quad (4.8)$$

respectively. We make the following three observations:

1. The functions $m_\eta^u(\cdot)$ and $m_\eta^l(\cdot)$ are piecewise linear. Therefore, by Section 18.4 of [12] the steady-state distributions of \hat{X}_η^u and \hat{X}_η^l exist.
2. The functions $m_\eta^u(\cdot)$ and $m_\eta^l(\cdot)$ are continuous.
3. The relationship $m_\eta^l(x) \leq m(x) \leq m_\eta^u(x)$ holds for all x . Therefore, by Proposition 18.5 in [12], we have

$$\hat{X}_\eta^l(\infty) \stackrel{st}{\leq} \hat{X}^*(\infty) \stackrel{st}{\leq} \hat{X}_\eta^u(\infty). \quad (4.9)$$

In particular, the inequalities (4.2) and (4.3) both hold.

Note that due to (4.3) we expect the policies associated with the upper and lower bound processes to not necessarily be feasible for the problem (3.4). Hence, we propose the process $\hat{X}_{\eta,\gamma}$ whose drift is a convex combination of $m^u(x)$ and $m^l(x)$ which, with the appropriate choice of the parameter γ is feasible.

The feasible diffusion process $\hat{X}_{\eta,\gamma}$

Fix the parameter $0 < \eta < L$ and let γ be an arbitrary number in the interval $[0, 1]$. We define the diffusion process $\hat{X}_{\eta,\gamma}$ to have an infinitesimal variance 2μ and an infinitesimal drift that satisfies $m_{\eta,\gamma}(x) = \gamma m_\eta^l(x) + (1 - \gamma)m_\eta^u(x)$, or:

$$m_{\eta,\gamma}(x) = \begin{cases} -\delta\sqrt{\mu} & x \geq 0 \\ -\delta\sqrt{\mu} - \mu_1 x & -L + \eta \leq x < 0 \\ -\delta\sqrt{\mu} - \mu_1 x + (1 - \gamma)\frac{L}{\eta}(\mu_2 - \mu_1)(\eta - L - x) & -L \leq x < -L + \eta \\ -\delta\sqrt{\mu} - \mu_2 x + \gamma\frac{L}{\eta}(\mu_2 - \mu_1)(-x - L - \eta) & -L - \eta \leq x < -L \\ -\delta\sqrt{\mu} - \mu_2 x & x < -L - \eta. \end{cases} \quad (4.10)$$

The following observations are true with respect to the process $\hat{X}_{\eta,\gamma}$:

1. The function $m_{\eta,\gamma}(\cdot)$ is piecewise linear. Therefore, by section 18.4 of [12] the steady-state distribution of $\hat{X}_{\eta,\gamma}$ exists.
2. The function $m_{\eta,\gamma}(\cdot)$ is continuous.
3. The relationship $m_{\eta}^l(x) \leq m_{\eta,\gamma}(x) \leq m_{\eta}^u(x)$ holds for all x . Therefore, by Proposition 18.5 in [12], we have

$$\hat{X}_{\eta}^l(\infty) \stackrel{st}{\leq} \hat{X}_{\eta,\gamma}(\infty) \stackrel{st}{\leq} \hat{X}_{\eta}^u(\infty). \quad (4.11)$$

In particular, the following two inequalities hold:

$$E \left[\hat{X}_{\eta}^l(\infty) \right]^+ \leq E \left[\hat{X}_{\eta,\gamma}(\infty) \right]^+ \leq E \left[\hat{X}_{\eta}^u(\infty) \right]^+, \quad (4.12)$$

and

$$E \left[\hat{X}_{\eta}^u(\infty) \right]^- \leq E \left[\hat{X}_{\eta,\gamma}(\infty) \right]^- \leq E \left[\hat{X}_{\eta}^l(\infty) \right]^- . \quad (4.13)$$

We next define the ϵ - threshold policy $\hat{T}H_{\epsilon}(\eta, \gamma)$ that corresponds to the process $\hat{X}_{\eta,\gamma}$. Let

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

and

$$u_{2,\eta,\gamma}(x) = \begin{cases} 0 & -L + \eta \leq x < 0 \\ -(1 - \gamma) \frac{L}{\eta} \left(\frac{\eta - L}{x} - 1 \right) & -L \leq x < -L + \eta \\ 1 + \gamma \frac{L}{\eta} \left(1 + \frac{L + \eta}{x} \right) & -L - \eta \leq x < -L \\ 1 & x < -L - \eta. \end{cases}$$

Then one can easily verify that $u_{\eta,\gamma} = (u_{1,\eta,\gamma}, u_{2,\eta,\gamma})$ is the policy associated with the diffusion process $\hat{X}_{\eta,\gamma}$ by confirming that:

- a. $0 \leq u_{i,\eta,\gamma} \leq 1, i = 1, 2,$
- b. $u_{1,\eta,\gamma} + u_{2,\eta,\gamma} = 1,$ and
- c. $m_{\eta,\gamma}(x) = -\delta\sqrt{\mu} + u_{1,\eta,\gamma}(x)\mu_1x^+ + u_{2,\eta,\gamma}(x)\mu_2x^+,$ for all $x < 0.$

We next show that there exists $\gamma(\eta) \in [0, 1]$ such that (4.4) holds. This will establish feasibility of the policy $\hat{T}H_{\epsilon}(\eta, \gamma(\eta)).$

Proposition 4.1 *Fix $0 < \eta < L$ and define the processes $\hat{X}_{\eta}^u, \hat{X}_{\eta}^l$ and $\hat{X}_{\eta,\gamma}$ as above. Then there exists $\gamma(\eta) \in [0, 1]$ such that $E[\hat{X}_{\eta,\gamma(\eta)}(\infty)]^- = E[\hat{X}^*(\infty)]^-.$*

Establishing ϵ - optimality of $\hat{T}H_\epsilon(\eta, \gamma(\eta))$

Now that we have established the existence of $\gamma(\eta)$ such that the policy $\hat{T}H_\epsilon(\eta, \gamma(\eta))$ is feasible with respect to the problem (3.4), it is left to show that this policy is also ϵ - optimal for the same problem. We show this property by establishing that $E[\hat{X}_{\eta, \gamma(\eta)}(\infty)]^+$ converges to $E[\hat{X}^*(\infty)]^+$ as $\eta \downarrow 0$.

Proposition 4.2 *The policy $\hat{T}H_\epsilon(\eta, \gamma(\eta))$ is ϵ - optimal with respect to the problem (3.4). That is:*

- i. $\hat{T}H_\epsilon(\eta, \gamma(\eta))$ is feasible for all values of $0 < \eta < L$, and
- ii. For all $\epsilon > 0$, there exists $\eta_\epsilon > 0$ such that $E[\hat{X}_{\eta_\epsilon, \gamma(\eta_\epsilon)}(\infty)]^+ \leq E[\hat{X}^*(\infty)]^+ + \epsilon$.

4.2 The ϵ - threshold policy

In this section we define the ϵ - threshold policy $TH_\epsilon(\eta, \gamma)$ for the original queueing system (fixed λ). We then establish that, in the limit as $\lambda \rightarrow \infty$ the scaled number of idle servers of each pool and the scaled queue length converge to the appropriate quantities associated with the diffusion process $\hat{X}_{\eta, \gamma}$.

Fix the values of $L > 0$, $0 < \eta < L$ and $0 < \gamma < 1$. Consider the following two functions $v_1, v_2 : \mathcal{R}_+ \rightarrow [0, 1]$:

$$v_1(x) = \begin{cases} 1 & 0 < x \leq L - \eta \\ 1 + (1 - \gamma) \frac{L}{\eta} \left(\frac{L - \eta}{x} - 1 \right) & L - \eta < x \leq L \\ -\gamma \frac{L}{\eta} \left(1 - \frac{L + \eta}{x} \right) & L < x \leq L + \eta \\ 0 & L + \eta < x, \end{cases}$$

and

$$v_2(x) = \begin{cases} 0 & 0 < x \leq L - \eta \\ -(1 - \gamma) \frac{L}{\eta} \left(\frac{L - \eta}{x} - 1 \right) & L - \eta < x \leq L \\ 1 + \gamma \frac{L}{\eta} \left(1 - \frac{L + \eta}{x} \right) & L < x \leq L + \eta \\ 1 & L + \eta < x. \end{cases}$$

Note that $v_i(x) = u_{i, \eta, \gamma}(-x)$ for all $x > 0$ and $i = 1, 2$, where $u_{1, \eta, \gamma}, u_{2, \eta, \gamma}$ is the control associated with the diffusion ϵ - threshold policy: $\hat{T}H_\epsilon(\eta, \gamma)$. Roughly speaking, according to $\hat{T}H_\epsilon(\eta, \gamma)$, if at time t the state of the diffusion system is $x < 0$, then a fraction $u_{i, \eta, \gamma}$ of the total number of idle servers x^- receives service from pool i . Our purpose in $TH_\epsilon(\eta, \gamma)$ is to imitate this policy, keeping in mind that for the original system the policy must be non-preemptive.

The ϵ - threshold policy: For $\lambda > 0$, let $\hat{I}_{k, \eta, \gamma}^\lambda$ be the scaled number of idle servers of pool k , and let $\hat{X}_{\eta, \gamma}^\lambda$ be the centered and scaled total number of customers in the system. For brevity, omit η

and γ from the notation. Then, the policy $TH_\epsilon(\eta, \gamma)$ assigns a new arrival at time t to a server of pool $k^*(t)$ where

$$k^*(t) \in \operatorname{argmax} \left\{ \hat{I}_k^\lambda(t) - [\hat{X}^\lambda(t)]^- \cdot v_k([\hat{X}^\lambda(t)]^-) \right\}.$$

Ties are resolved arbitrarily.

Note that our policy is a special case of the Queue-and-Idleness-Ratio (QIR) control introduced in [27].

Our goal now is to establish that as $\lambda \rightarrow \infty$ the appropriate (scaled) performance under $TH_\epsilon(\eta, \gamma)$ converges to the performance obtained under $\hat{T}H_\epsilon(\eta, \gamma)$. A step towards establishing this ultimate goal is proving a state-space-collapse (SSC) result. Roughly speaking, SSC implies that at any time point t the fraction of idle servers of pool i is approximately equal to v_i . We state this formally in the following proposition:

Proposition 4.3 *Suppose that $(\hat{X}^\lambda(0), \hat{Z}^\lambda(0)) \Rightarrow (\hat{X}(0), \hat{Z}(0))$ and that $\hat{I}_k^\lambda(0) - [\hat{X}^\lambda(0)]^- \cdot v_k([\hat{X}^\lambda(0)]^-) \Rightarrow 0$, as $\lambda \rightarrow \infty$. Also suppose that the system works under the ϵ -threshold policy $TH_\epsilon(\eta, \gamma)$. Then we have the state-space collapse:*

$$\hat{I}_k^\lambda(t) - [\hat{X}^\lambda(t)]^- \cdot v_k([\hat{X}^\lambda(t)]^-) \Rightarrow 0 \text{ in } D \text{ as } \lambda \rightarrow \infty, k = 1, 2,$$

where D is the space of all RCLL functions with values in \mathcal{R} , equipped with the Skorohod J_1 metric.

We next establish weak convergence of the scaled process $(\hat{X}^\lambda, \hat{Q}^\lambda, \hat{I}_1^\lambda, \hat{I}_2^\lambda)$.

Proposition 4.4 *Consider a sequence of systems operating under the ϵ -threshold policy $TH_\epsilon(\eta, \gamma)$. Suppose that $\vec{\hat{X}}^\lambda(0) \Rightarrow \vec{\hat{X}}(0)$, as $\lambda \rightarrow \infty$, with $\hat{X}_1(0) + \hat{X}_2(0) = x$ and $\hat{I}_k(0) = \hat{X}_k^-(0) = x^- \cdot v_k(x^-)$, $k = 1, 2$. Then $\hat{X}^\lambda(\cdot) \Rightarrow \hat{X}(\cdot)$, where $\hat{X}(\cdot)$ is the diffusion process associated with the policy $\hat{T}H_\epsilon(\eta, \gamma)$ whose infinitesimal variance is 2μ and infinitesimal drift is given in (4.10). In addition, $\hat{Q}^\lambda \Rightarrow \hat{X}^+$ and $\hat{I}_k^\lambda \Rightarrow \hat{X}^- \cdot v_k(\hat{X}^-)$, $k = 1, 2$.*

Finally, we are ready to establish that the ϵ -threshold policy $TH_\epsilon(\eta, \gamma)$ asymptotically performs as well as its diffusion counterpart $\hat{T}H_\epsilon(\eta, \gamma)$

Theorem 4.1 *Fix the values of η and γ (omitted from the notation), and consider a sequence of systems operating under the ϵ -threshold policy $TH_\epsilon(\eta, \gamma)$. Then,*

$$\lim_{\lambda \rightarrow \infty} E[\hat{X}^\lambda(\infty)]^+ = E[\hat{X}(\infty)]^+, \quad (4.14)$$

and

$$\lim_{\lambda \rightarrow \infty} E[\hat{I}_k^\lambda(\infty)] = E[\hat{X}(\infty)^- \cdot v_k(\hat{X}(\infty)^-)], \quad k = 1, 2, \quad (4.15)$$

where \hat{X} is the diffusion process associated with the policy $\hat{T}H_\epsilon(\eta, \gamma)$.

4.3 Establishing a lower bound

The following theorem will establish that the optimal objective value to the diffusion control problem in (3.2) provides a lower bound on $E[\hat{X}^\lambda(\infty; \pi')^+]$ under any asymptotically feasible policy π' .

Theorem 4.2 *Consider a sequence of systems indexed by λ , with N_k^λ servers in pool k ($k = 1, 2$). Suppose that the system operates under an asymptotically feasible sequence of policies $\pi' = \pi'(\lambda, N_k^\lambda)$. Then,*

$$\liminf_{\lambda \rightarrow \infty} E \left[\hat{X}^\lambda(\infty; \pi') \right]^+ \geq E \left[\hat{X}^*(\infty) \right]^+.$$

To establish ϵ -asymptotic optimality, it is left to show that the asymptotic diffusion-scaled steady-state waiting time under the ϵ -threshold policy TH_ϵ does not exceed the asymptotic diffusion-scaled steady-state waiting time under any other asymptotically feasible policy by any more than ϵ .

Let $\epsilon > 0$, and let \hat{X} be the feasible diffusion process that corresponds to the policy $\hat{\pi} = TH_{\mu\epsilon}(\eta, \gamma(\eta))$. Consider a sequence of systems operating under the $\epsilon\mu$ -threshold policy $\pi = TH_{\mu\epsilon}(\eta, \gamma(\eta))$. It follows from Little's law that

$$\lim_{\lambda \rightarrow \infty} E \left[\hat{W}^\lambda(\infty; \pi) \right] = \lim_{\lambda \rightarrow \infty} \frac{N^\lambda}{\lambda} E \left[\hat{X}^\lambda(\infty; \pi) \right]^+.$$

Theorem 4.1 and the limit in (2.15) show

$$\lim_{\lambda \rightarrow \infty} \frac{N^\lambda}{\lambda} E \left[\hat{X}^\lambda(\infty; \pi) \right]^+ = \frac{E \left[\hat{X}(\infty; \hat{\pi}) \right]^+}{\mu},$$

and, by Proposition 4.2,

$$\frac{E \left[\hat{X}(\infty; \hat{\pi}) \right]^+}{\mu} \leq \frac{E \left[\hat{X}^*(\infty) \right]^+}{\mu} + \epsilon.$$

Theorem 4.2 implies that

$$\frac{E \left[\hat{X}^*(\infty) \right]^+}{\mu} \leq \frac{1}{\mu} \liminf_{\lambda \rightarrow \infty} E \left[\hat{X}^\lambda(\infty; \pi') \right]^+.$$

Since for any other asymptotically feasible sequence of policies π'

$$\frac{1}{\mu} \liminf_{\lambda \rightarrow \infty} E \left[\hat{X}^\lambda(\infty; \pi') \right]^+ = \liminf_{\lambda \rightarrow \infty} \frac{N^\lambda}{\lambda} E \left[\hat{X}^\lambda(\infty; \pi') \right]^+,$$

Little's law shows that

$$\frac{1}{\mu} \liminf_{\lambda \rightarrow \infty} E \left[\hat{X}^\lambda(\infty; \pi') \right]^+ = \liminf_{\lambda \rightarrow \infty} E \left[\hat{W}^\lambda(\infty; \pi') \right]^+.$$

We conclude that

$$\lim_{\lambda \rightarrow \infty} E \left[\hat{W}^\lambda(\infty; \pi) \right] \leq \liminf_{\lambda \rightarrow \infty} E \left[\hat{W}^\lambda(\infty; \pi') \right]^+ + \epsilon,$$

which establishes that the policy π is ϵ -asymptotically optimal.

5 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 utilizes the faster servers first when the number in the system exceeds the threshold and utilizes the slower servers first when that number is below the threshold. 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, and its natural extension, the longest weighted idle server first (LWISF) policy. Formally, we show that a continuous version of the threshold policy is ϵ -asymptotically optimal in the many-server heavy-traffic QED limiting regime.

There are various directions for further research in this domain. First, it might be possible to extend this paper's results to an inverted-V model with an arbitrary number of server pools. One might also be able to incorporate customer abandonment into the model, as well as other forms for the interarrival and service time distributions. Another important extension is to study the notion of server fairness in more general multiskill networks. Here, the problem formulation needs to change due to the multiplicity of both customer classes and server skills.

Another potentially fruitful direction for further research is to explore 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. This raises the question of what kind of control would give each agent an incentive to work as fast as s/he can. It is possible that such a control mechanism would have to be supplemented with an appropriate compensation mechanism (e.g. pay-per call) in order to obtain this goal.

References

- [1] Abate, J. and Whitt, W. (1987), Transient behavior of regulated Brownian motion I: Starting at the origin, *Advances in Applied Probability*, **19**(3), 560-598.
- [2] Armony, M.(2005), Dynamic routing in large-scale service systems with heterogeneous servers, *Queueing Systems*, **51**(3-4), 287-329.
- [3] Armony, M. and Mandelbaum, A. (2007), Routing and staffing in large-scale service systems with heterogeneous servers and impatient customers, Preprint.
- [4] Armony, M. and Ward, A. (2008), Fair dynamic routing in large-scale heterogeneous-server systems: Technical appendix.
- [5] Atar, R. (2007), Central limit theorem for a many-server queue with random service rates, Preprint.
- [6] Atar, R. (2005), Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic, *The Annals of Applied Probability*, **15**(4), pp. 2606-2650.
- [7] Atar R., Mandelbaum A. and Reiman M. (2004), Scheduling a multi-class queue with many i.i.d. servers: Asymptotic optimality in heavy-traffic, *The Annals of Applied Probability*, **14**(3), pp. 1084-1134.
- [8] Atar, R. and Shwartz, A. (2007), Efficient routing in heavy traffic under partial sampling of service times. Working paper.
- [9] B. Avi-Itzhak, H. Levy and D. Raz (2004), Quantifying fairness in queueing systems: Principles and applications. Working paper.
- [10] Billingsley, P. (1999), Convergence of probability measures. Second Edition. New York: John Wiley & Sons.
- [11] Borst, S., Mandelbaum, A. and Reiman, M. (2003), Dimensioning large call centers, *Operations Research*, **52**(1), pp. 17-34.
- [12] Browne, S. and Whitt, W. (1995), Piecewise-linear diffusion processes. *Advances in queueing: Theory, methods, and open problems* (Dshalalow, ed.), Boca Raton, FL: CRC Press, pp. 463-480.
- [13] Cabral F.B. (2005), The slow server problem for uninformed customers, *Queueing systems*, **50**(4), pp. 353-370.
- [14] Cabral F. B. (2007), Queues with heterogeneous servers and uninformed customers: who works the most? Working paper.
- [15] Cohen-Charash, Yochi and Spector, P. E. (2001), The role of justice in organizations: A meta-analysis, *Organizational Behavior and Human Decision Processes*, **86**(2), pp. 278-321.
- [16] Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., and Ng, K. Y. (2001), Justice at the millennium: A meta-analytic review of 25 Years of organizational justice research, *Journal of Applied Psychology*, **86**(3), pp. 425-445.

- [17] J. G. Dai and T. Tezcan, (2005), State space collapse in many-server diffusion limits of parallel server systems, working paper.
- [18] de Véricourt, F. and Zhou, Y.-P. (2005), A routing problem for call centers with customer callbacks after service failure, *Operations Research*, **53**(6).
- [19] de Véricourt, F. and Zhou, Y.-P. (2006), On the incomplete results for the heterogeneous server problem, *Queueing Systems* **52**(3).
- [20] Erlang, A.K. (1948), On the rational determination of the number of circuits, in: *The Life and Works of A. K. Erlang* (E. Brockmeyer, H. L. Halstrom, and A. Jensen, eds.) Copenhagen: The Copenhagen Telephone Company.
- [21] Foschini, G. J. (1977), On heavy traffic diffusion analysis and dynamic routing in packet switched networks. *Computer Performance Measurements, Modeling, and Evaluation* (Reiser, M. and Chandy, K., eds.), Amsterdam: North-Holland, pp. 499-514.
- [22] Fleming, W. H., Soner, H. M. (1993), Controlled Markov processes and viscosity solutions. New York: Springer.
- [23] Gans, N., Koole, G., Mandelbaum, A. (2003), Telephone call centers: Tutorial, review and research prospects. *Manufacturing and Service Operations Management* **5**(2), pp. 79-141.
- [24] Gans, N. and Shen, H. (2007), Service time heterogeneity in an inbound call center, In Preparation.
- [25] Gautschi, W. (1972), Error function and Fresnel integrals. *Handbook of Mathematical Functions* (Abramowitz, M. and Stegun, I., eds.), New York: Dover, pp. 295-329.
- [26] Gurvich I., Armony M. and Mandelbaum A. (2006), Service level differentiation in call centers with fully flexible servers. *Management Science*, forthcoming.
- [27] Gurvich, I. and Whitt, W. (2007) Queue-and-idleness-ratio controls in many-server service systems. Working paper.
- [28] Gurvich, I. and Whitt, W. (2007), Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing. Working paper.
- [29] Gurvich, I. and Whitt, W. (2007), Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing and Service Operations Management*, forthcoming.
- [30] Gustafson, H. W. (1982), Force-loss cost analysis. *Employee Turnover: Causes, Consequences, and Control* (W. H. Mobley, ed.), Reading, MA: Addison-Wesley.
- [31] Harrison, J. M. (1985), Brownian models of queueing networks with heterogeneous customer populations. *Stochastic Differential Systems, Stochastic Control Theory and Applications* (Fleming, W. and P. L. Lions, eds.), IMA Volumes in Mathematics and its Applications, vol. 10, New York: Springer-Verlag, pp. 147-186.
- [32] Halfin, S. and Whitt, W. (1981), Heavy-traffic limits for queues with many exponential servers, *Operations Research*, **29**(3), pp. 567-588.

- [33] Harrison, J. M. and Zeevi, A. (2004), Dynamic scheduling of a multiclass queue in the Halfin and Whitt heavy traffic regims, *Operations Research*, **52**, pp. 243-257.
- [34] Kushner, H. J., Dupuis, P. (2001), Numerical methods for stochastic control problems in continuous time. New York: Springer.
- [35] Karatzas, I. and Shreve, S. (1991), Brownian motion and stochastic calculus. Second Edition. New York: Springer.
- [36] Kurtz, T.G. and Protter, P. (1991), Weak limit theorems for stochastic integrals and stochastic differential equations, *Ann. Probab.*, **19**, pp. 1035-1070.
- [37] Larsen, R. L. and Agrawala, A. K. (1983), Control of a heterogeneous two-server exponential queueing system, *IEEE Transactions on Software Engineering*, July, pp 522-526.
- [38] Lin, W. and P.R. Kumar (1984), Optimal control of a queueing system with two heterogeneous servers, *IEEE Trans. Automat. Control* **29**, pp 696–703.
- [39] Mandelbaum A. and Zeltyn S. (2007), Staffing many-server queues with impatient customers: constraint satisfaction in call centers. Working paper.
- [40] Rogers, L. C. G. and Williams, D. (2000), Diffusions, Markov processes, and martingales. Volume 2, Ito Calculus. Cambridge: Cambridge University Press.
- [41] Rubinovich M. (1983), The slow server problem, *Journal of Applied Probability* **22**, pp. 205-213.
- [42] Stockbridge R.H. (1991), A Martingale approach to the slow server problem, *Journal of Applied Probability* **28**, pp 480-486.
- [43] Teh, Y., Ward, A. R. (2002), Critical thresholds for dynamic routing in queueing networks. *Queueing Systems*, **42**, pp. 297–315.
- [44] Tezcan, T. (2007), Asymptotically optimal control of many-server heterogeneous service systems with hyper-exponential service times. Working paper.
- [45] Whitt, W (1980), Some useful functions for functional limit theorems, *Mathematics of Operations Research*, **5** (1), pp. 67-85.
- [46] Whitt, W (2006), The impact of increased employee retention upon performance in a customer contact center, *Manufacturing and Service Operations Management*, **8** (3), pp. 221-234.
- [47] Wierman, A. (2007) Fairness and classification, *Performance Evaluation Review*, **34**(4), pp. 4–12.
- [48] Wu, H-M., Wu C-C. and Lin, W. (2007), Improving inter-server fairness in active queue management, *IEEE communications letters*, **11**(11), pp. 910-912.
- [49] Whitt, W. (2002), Stochastic-process limits. New York: Springer.
- [50] Zeltyn, S. and Mandelbaum, A. (2005), Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue, *Queueing Systems*, **51** (3/4), pp. 361-402.