

Routing and Staffing in Large-Scale Service Systems: The Case of Homogeneous Impatient Customers and Heterogeneous Servers¹

Mor Armony²

Avishai Mandelbaum³

June 25, 2008

Abstract

Motivated by call centers, we study large-scale service systems with homogeneous impatient customers and heterogeneous servers. The servers differ with respect to their speed of service. For this model, we propose staffing and routing rules that are jointly asymptotically optimal as the system approaches the heavy-traffic many-server QED regime. Our routing rule is FSF, which assigns customers to the Fastest Server available First. The staffing rule is defined as a solution to the problem of minimizing the staffing costs subject to a feasible region with a simple linear boundary. This boundary is obtained by adding (or subtracting) a square-root term to the nominal capacity required for stability, in the lack of customer abandonment.

¹The research of both authors was supported by BSF (Binational Science Foundation) grant 2006379. A.M. was supported by the Technion funds for the promotion of research and sponsored research.

²Stern School of Business, New York University, marmony@stern.nyu.edu.

³Industrial Engineering and Management, Technion Institute of Technology, avim@ie.technion.ac.il.

Contents

1	Introduction	1
1.1	Summary of the results	3
1.2	Literature Review	5
2	Model Formulation	7
3	Optimal Preemptive Routing	10
4	Asymptotically optimal control	13
4.1	Asymptotic Framework	13
4.2	Faster Server First (FSF) is Asymptotically Optimal	18
4.2.1	State-Space Collapse	19
4.2.2	Transient Diffusion Limit	19
4.2.3	Stationary Diffusion Limit	21
5	Asymptotically Optimal Staffing	22
5.1	Asymptotic Feasibility	22
5.2	Asymptotically Optimal Staffing	24
5.3	Extensions	28
6	Conclusions	31
A	Proofs	1

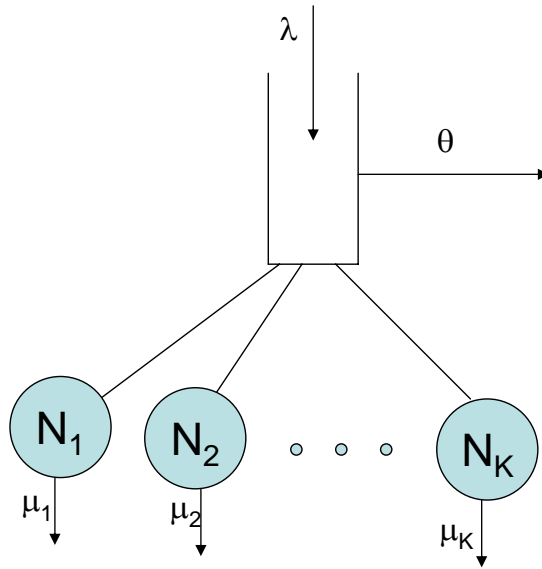


Figure 1.1: The inverted-V model

1 Introduction

In this paper we consider large-scale service systems, such as customer contact centers [1, 16] or hospitals [33], with a homogeneous population of impatient customers and heterogeneous servers that belong to multiple pools. The servers in each pool are statistically identical and all servers are mutually independent. The model is depicted in Figure 1.1. For such systems, our goal is to solve the joint problem of staffing and control. Staffing is concerned with determining the number of servers at each pool, while the control (routing) determines the assignment of customers to those servers. The objective is to minimize staffing costs, subject to an upper bound on the steady-state fraction of customers who abandon.

The joint problem of staffing and control is difficult, hence the two are typically solved separately in the literature. In particular, authors have assumed away the routing question to address

the staffing problem (e.g. [9],[27]), or have assumed a particular staffing rule to solve the routing problem (e.g. [2], [32], [12]). In contrast, we address this joint problem by rigorously justifying a “divide and conquer” approach; that is, we first find a routing scheme which is (asymptotically) optimal given *any* staffing vector. Subsequently, we identify an (asymptotically) optimal staffing rule, assuming that the above-mentioned routing rule is used, thereby solving the *joint* problem.

Our approach in addressing the problem in this paper is asymptotic. Specifically, we identify staffing and routing rules which are asymptotically optimal as both the arrival rate and number of servers of each pool grow to infinity within the QED regime. This regime was first formalized by Halfin and Whitt in [23] and was later adapted to queues with abandonment by Garnett et. al. [18]. In this regime, the delay probability has a limit that is strictly between 0 and 1, and the fraction of abandonment approaches 0, at a rate that is inversely proportional to the square-root of the arrival rate.

An analogous model was studied in [2], under the assumption that customers are infinitely patient. In [2] a routing scheme, FSF, was proposed which assigns customers to the Fastest Server available First. This policy was shown to be asymptotically optimal with respect to minimizing the steady-state delay probability. The main differentiators between the present paper and [2] are that here a) customers are impatient, and b) routing is studied in conjunction with staffing.

One might naturally question the need to dedicate an entire paper to the model with impatient customers. Can it not simply be obtained as a straightforward extension of the model with no abandonment? It turns out that the answer to this question is negative. Customer abandonment introduces some subtle challenges that require and deserve special attention and that we resolve here. To elaborate, the model with abandonment differs from the one without abandonment on three fronts:

1. **Model Formulation:** Without abandonment, a natural approach with respect to quality of service is to minimize the delay probability (which is indeed the approach taken in [2]). With abandonment, this turns out to lead to a behavior that does not make sense service-wise. Specifically, to minimize the delay probability, one would ignore customers who arrive when all servers are busy and simply let them abandon (see Proposition 3.3). Hence, an alternative approach is called for. For example, one could target minimizing the probability of abandonment or the expected waiting time. This is the approach taken in this current paper.
2. **Asymptotic Optimality:** Without abandonment, there is a natural reference point which provides a lower bound on the staffing cost. Specifically, due to stability considerations, the arrival rate is a lower bound on the overall service capacity, which translates into a lower bound on the staffing cost. With abandonment, such a lower bound does not exist, since the system is always stable. This calls for an alternative approach in defining asymptotic optimality. We resolve this issue by introducing an auxiliary optimization problem, which can be thought of as the *fluid-scale* staffing problem. The optimal staffing cost associated

with this fluid problem becomes a centering factor in our definition of asymptotic optimality (see Section 5.2).

3. **Work Conservation Assumption:** Without abandonment, the preemptive version of the FSF policy is optimal with respect to stochastically minimizing the overall number of customers in the system at any time point. With abandonment, the same is true only within the family of work-conserving policies (see Proposition 3.2). If one relaxes the work-conservation assumption then the same policy is optimal with respect to a different criterion (and *not* with respect to the number in the system - see Remark 3.1), namely, the cumulative number of customers who have abandoned the system up to any point in time (see Proposition 3.1).

1.1 Summary of the results

The QED asymptotic regime, considered in this paper, can be characterized as follows: A system with large call volume (demand) and many servers is operating in the QED regime if (a) the delay probability is neither close to 1 nor to 0, or (b) its total service *capacity* is equal to the demand up to a safety capacity which is of the same order of magnitude as the square root of the demand⁴, or (c) the abandonment probability is of the order of 1 over square-root of the demand. It is easily seen that, under (b), the system operates in heavy-traffic, and hence the high server efficiencies. The quality aspect of the QED regime is seen from characterizations (a) and (c). This high performance, which is impossible to achieve for systems in conventional heavy traffic, is obtained here due to the economies of scale associated with the large number of servers. These three characterizations of the QED regime are shown to be equivalent in various settings [18, 27].

The asymptotically optimal routing policy that we propose is the policy Faster Server First (FSF) that simply assigns newly arriving or waiting customers to the fastest server available. FSF is shown to be asymptotically optimal among all the non-anticipating non-preemptive policies (Theorem 4.1). The asymptotic optimality is in terms of the steady-state scaled abandonment probability, as well as the steady-state expected queue length and expected waiting time in the QED regime. More specifically, consider a sequence of systems indexed by the arrival rate λ , where $\lambda \uparrow \infty$.⁵ For any fixed value of λ (the system indexed by λ will be referred to as the λ -system), let N_k^λ represent the number of servers of type k , $k = 1, \dots, K$. Also, let $\vec{N}^\lambda = (N_1^\lambda, N_2^\lambda, \dots, N_K^\lambda)$ be the staffing vector, and $N^\lambda = N_1^\lambda + N_2^\lambda + \dots + N_K^\lambda$ be the total number of servers. Suppose that the service rates μ_1, \dots, μ_K and the abandonment rate θ are fixed independently of λ . To be consistent with the QED regime, assume that the total service capacity, $\mu_1 N_1^\lambda + \mu_2 N_2^\lambda + \dots + \mu_K N_K^\lambda$, is equal

⁴The expert reader may notice that this is a slightly different characterization of the QED regime, in terms of the total service capacity rather than the total number of servers. This characterization is suitable for the \wedge -design studied in our paper.

⁵Since λ is a continuous parameter it is appropriate to refer to a “family” of systems rather than a “sequence”. Instead, we assume a specific increasing subsequence of the family $\{\lambda\}$, and use the superscript λ for simplicity.

to the arrival rate plus (or minus) a square root “safety” capacity. Formally, assume that

$$\sum_{k=1}^K N_k^\lambda \mu_k = \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda}), \quad (1.1)$$

for some number $-\infty < \delta < \infty$.⁶ Let $Q^\lambda(\infty)$ and $W^\lambda(\infty)$ be the steady-state queue length and waiting time, respectively. For asymptotic purposes, let $\hat{Q}^\lambda(\infty) = Q^\lambda(\infty)/\sqrt{N^\lambda}$ and $\hat{W}^\lambda(\infty) = \sqrt{N^\lambda}W^\lambda(\infty)$ be the *scaled* steady-state queue length and waiting time, respectively. Finally, let $P^\lambda(ab)$ be the steady-state abandonment probability. The asymptotic optimality of the FSF policy is with respect to minimization of the limiting expectations of $\hat{Q}^\lambda(\infty)$ and $\hat{W}^\lambda(\infty)$, as $\lambda \rightarrow \infty$, as well as $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda}P^\lambda(ab)$. (see Theorem 4.1 for further details).

Although FSF is a very intuitive policy, its asymptotic optimality is not easy to establish. Particularly, [14] has shown that the characterization of an exact optimal policy for more than two servers is still beyond reach. To establish the asymptotic optimality of FSF we first introduce a related *preemptive* policy, FSF_p . This policy keeps the faster servers busy whenever possible, even at the cost of handing-off customers from slower servers to faster ones. The policy FSF_p is shown (in Proposition 3.1) to stochastically minimize the cumulative number of customers who have abandoned, for any time $t \geq 0$. In particular, FSF_p minimizes $P(ab)$, the steady-state abandonment probability. Consequently, we show that, in the limit as $\lambda \rightarrow \infty$, both policies (FSF and FSF_p) give rise to the same performance measures (Proposition 4.6).

The asymptotic optimality of the FSF routing rule, given any staffing vectors (in the QED regime, and under some mild regularity conditions), facilitates the solution of the joint staffing and routing problem. Specifically, when focusing on an asymptotically optimal staffing, one can simply assume that FSF is used as the routing rule. We then identify the region

$$\left\{ \vec{N}^\lambda : \sum_{k=1}^K N_k^\lambda \mu_k \geq \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda}) \right\}, \quad (1.2)$$

to be the asymptotically *feasible* region for the problem whose constraint is $\limsup_{\lambda \rightarrow \infty} \sqrt{\lambda}P^\lambda(ab) \leq \Delta$, for some $0 < \Delta < \infty$, and we explicitly provide a one-to-one correspondence between δ and Δ (see Proposition 5.1).

Finally, we show that to find an asymptotically optimal staffing rule, it is sufficient to identify the staffing vector that minimizes the staffing cost over the region:

$$\sum_{k=1}^K N_k^\lambda \mu_k \geq \lambda + \delta\sqrt{\lambda}, \quad (1.3)$$

which approximates the asymptotically feasible region, up to terms of order $o(\sqrt{\lambda})$. We explicitly provide the solution under homogeneous and additive cost functions and establish its asymptotic optimality (Proposition 5.3).

⁶Recall that in [2] only positive values of δ were considered. Abandonments render feasible also non-positive values of this parameter.

The paper is organized as follows: We conclude the introduction by reviewing the relevant literature. Section 2 gives the model formulation as well the formulation of the original joint staffing and routing problem. We then proceed to Section 3 in which the preemptive version of FSF is introduced and its optimality is established. Next, Section 4 introduces the asymptotic framework and verifies asymptotic optimality of the FSF routing rule. Finally, Section 5 establishes asymptotic feasibility and optimality of the square-root “safety” capacity rule. Section 6 concludes the paper. All the proofs are given in the Appendix, to allow for a fluent reading of the paper.

1.2 Literature Review

Service systems with heterogeneous servers occur naturally due to training and learning effects [17] and also due to heterogeneity in the workforce. Especially, such heterogeneity arises in a co-sourcing environment (co-sourcing in call centers is a common arrangement in which the firm outsources part of their call center operations and keeps the rest in-house). In such an environment, it is likely that the in-house CSRs are different in their service skills from the outsourcer’s CSRs. Many researchers have addressed the dynamic routing problem of how to assign customers to servers under various assumptions and optimization goals. Only a few papers have tackled the staffing problem in conjunction with this dynamic routing one.

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. [28], [10]), 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. [25], [26], [29] and [13]). 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. [28],[26]), but recently results for the general heterogeneous multi-server system have appeared ([10], [13]). Note, though, that the problem b), for the general multiserver case, is still open [14]. We tackle a problem related to problem b), with the objective of minimizing the abandonment probability. A more detailed summary of the slow server problem is given in [33] who is applying these concepts to hospitals.

Inverted-V and asymptotic analysis

The difficulty in identifying optimal controls for the general heterogenous 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, [15] shows that shortest-expected-delay-first routing is asymptotically optimal, and [30] identifies necessary and sufficient conditions for a threshold priority policy to be asymptotically optimal.

More recently, heterogenous server systems have been studied in the QED many-server heavy-traffic regime. 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 ([9], [27], [19]), including this paper, in which we establish asymptotic optimality of this rule for the inverted-V model.

Several papers have examined the question of dynamic control for the inverted-V system in the QED regime. These include [2], [31], [4], [6] and [3]. 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 are extended in the present paper to the inverted-V system with abandonment, where the minimization is with respect to the steady-state abandonment probability, as well as the expected queue length and waiting time. Tezcan [31] examines a similar routing 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 [4] has established that both the FSF and the Longest-Idle-Server First (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 [4] the state-space collapse is into a one-dimensional process. Most recently, Atar and Shwartz [6] have shown that in an environment where service rates are heterogeneous and unknown, it is sufficient to take a relatively 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 [21, 22] 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. They then rely on the results of [2] to establish asymptotic optimality of their general Fixed-Idleness-Ratio (FIR) policy and of square-root-safety-staffing rule. Both of these papers assume that customers are infinitely patient and therefore do not abandon. They mention that if one were to establish the asymptotic optimality of the FSF policy in the inverted-V model with abandonment (which we do here) then their results could be readily extended to customers' abandonment if abandonment rates were appropriately ordered. The more general case is covered in [5]. In [5] the authors establish that if service rates are pool-dependent, then the control problem is asymptotically reducible to a 1-dimensional control problem, which is based on the implicit solution of an HJB equation. Hence, their result is structural in nature, and as such does not reveal special structure like we do in this paper. In [20], Gurvich and Whitt establish state-space collapse of the FIR policy, as well as weak convergence of the total number in the system, into an appropriate diffusion limit (convergence is for both the transient process and in steady state). Our FSF policy can be considered a special case of the FIR policy and therefore state-space collapse and

weak convergence follow through.

Other papers, which consider control for various systems in the many-server heavy-traffic QED regime, are [32] (N-system) and [12] (more general skill-based routing design). The asymptotic optimization in these papers is with respect to a finite horizon cost of abandonment, and are therefore different from our steady-state setting. Finally, Harrison and Zeevi [24] and Bassamboo et al [7, 8] consider staffing and routing in a general skill-based routing framework, under the assumption that the arrival rate is random and using stochastic fluid models.

2 Model Formulation

Consider a service system with a single customer class and K server skills (each skill in its own server pool), all capable of fully handling customers' service requirements. Service times are assumed to be exponential, with service rate that depends on the pool (skill) of the particular server. Specifically, the average service time of a customer that is served by a server of skill k ($k = 1, 2, \dots, K$) is $1/\mu_k$. We assume that the service rates are ordered as follows: $\mu_1 < \mu_2 < \dots < \mu_K$. Customers arrive to the system according to a Poisson process with rate λ . Delayed customers wait in an infinite buffer, and are served according to a FCFS discipline⁷. Customers are impatient. In particular, if a customer's service does not start within a time that is exponentially distributed with rate θ , this customer abandons (reneges) and does not return to the system. It is assumed that customers do not abandon once their service starts. All interarrival times, service times and time to abandonment are assumed to be statistically independent.

Let N_k be the number of server in pool k . Respectively, let $\vec{N} = (N_1, N_2, \dots, N_K)$ be the staffing vector (here and elsewhere, \vec{x} is used to denote a vector whose elements are x_1, x_2, \dots). Let Π be the set of all non-preemptive non-anticipating routing policies. Denote by $\pi := \pi(\lambda, \vec{N}) \in \Pi$, 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 or immaterial from the context which arguments should be used). Given a policy $\pi \in \Pi$, let $P_\pi(\text{wait} > 0)$ be the steady-state probability that a customer is delayed (before starting service or abandoning), and let $P_\pi(ab)$ be the steady-state probability that a customer abandons.⁸ Our first goal in this paper is to find a policy in Π which minimizes the latter probability.

A more ambitious goal is to jointly identify staffing levels N_1, \dots, N_K and a routing policy to minimize staffing costs subject to a constraint on system performance (such as the probability of waiting and / or the fraction of customers who abandon). Generally, solving the staffing and control problems concurrently has been infeasible. Hence, researchers commonly end up solving

⁷Proposition 3.4 will show that the FCFS assumption can be relaxed.

⁸If the steady-state distribution does not exist, consider $P_\pi(\text{wait} > 0)$ and $P_\pi(ab)$ as the random variables corresponding to the essential limsups of the long term proportions of customers who, respectively, are delayed and abandon.

one while assuming the solution to the other is given. A distinguishing feature of our “divide and conquer” approach is that we identify a control policy which is near-optimal given *any* staffing level, and therefore, we are able to solve the staffing and the control problems concurrently.

The joint staffing and routing problem can be formulated as follows: Suppose that the cost of staffing the system with N_k servers of skill k is $C_k(N_k)$. The total staffing cost is, hence, $C(N_1, N_2, \dots, N_K) = C_1(N_1) + C_2(N_2) + \dots + C_K(N_K)$. We wish to determine the number of servers required of each skill, in order to minimize the staffing cost while maintaining a target service level constraint. The service performance measure that we study is the steady-state probability that a customer abandons the system. Equivalently, we focus on the long-term proportion of customers who abandon. Let $0 < \Delta < 1$ be the target upper bound on the fraction of abandonment. The joint staffing and routing problem is now stated as:

$$\begin{aligned} & \text{minimize} && C_1(N_1) + C_2(N_2) + \dots + C_K(N_K) \\ & \text{subject to} && P_\pi(ab) \leq \Delta, \text{ for some } \pi = \pi(\lambda, \vec{N}) \in \Pi, \\ & && N_1, N_2, \dots, N_K \in \mathbb{Z}_+. \end{aligned} \tag{2.1}$$

Suppose that the routing policy $\pi \in \Pi$ is used, and let $t \geq 0$ be an arbitrary time point. We denote by $Z_k(t; \pi)$ the number of busy servers of pool k ($k = 1, 2, \dots, K$) at time t , and $Q(t; \pi)$ the queue length at this time. Finally, let $Y(t; \pi)$ be the total number of customers in the system (sometimes referred to as the head-count). That is, $Y(t; \pi) = Z_1(t; \pi) + Z_2(t; \pi) + \dots + Z_K(t; \pi) + Q(t; \pi)$. We use $t = \infty$ whenever we refer to steady-state. At times, we omit π if it is clear from the context which routing policy is used.

Definition: A control policy $\pi \in \Pi$ 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 $Z_1(t; \pi) + Z_2(t; \pi) + \dots + Z_K(t; \pi) = N$, where

$$N := N_1 + N_2 + \dots + N_K$$

is the total number of servers.

Note that, in general, a $K+1$ dimensional vector is required to specify the state of the system, namely, $Q(t; \pi)$ and $Z_1(t; \pi), \dots, Z_K(t; \pi)$. However, for work conserving policies, the state space can be described by the K -dimensional vector $(Z_1(t; \pi) + Q(t; \pi), Z_2(t; \pi), \dots, Z_K(t; \pi))$. In fact, the queue length can be added to the number of busy servers of pool k , for any k , 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 $Z_k(t; \pi) = \min\{Q(t; \pi) + Z_k(t; \pi), N_k\}$. Work conserving policies also have the appealing property that the waiting probability can be stated in terms of the total number of busy servers. In particular, if $\pi \in \Pi$ is work conserving, and there exist steady-state distributions for their underlying processes, then

$$P_\pi(\text{wait} > 0) = P(Z_1(\infty; \pi) + Z_2(\infty; \pi) + \dots + Z_K(\infty; \pi) = N) = P(Y(\infty; \pi) \geq N), \tag{2.2}$$

where the first equality is due to the PASTA property, and the second follows from work-conservation. Note that if the policy is *not* work conserving then (2.2) need not hold, because one may have customers waiting in queue, even if some of the servers are idle.

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 $k = 1, \dots, K$, 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 $B_k(t; \pi)$ 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 (excluding direct arrivals 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 $I_k(t; \pi)$ be the total idle time experienced by servers of pool k up to time t . Also, let $D_k(t)$, $t \geq 0$, be a Poisson process with rate μ_k . Then the number of service completions out of server pool k may be written as $D_k(T_k(t; \pi))$. In addition, let $E(t; \pi)$ represent the total time spent by customers in the queue up to time t , and let $L(t)$, $t \geq 0$, be a Poisson process with rate θ . Then the number of customers who have abandoned up to time t can be written as $L(E(t; \pi))$. The above definitions allow us to write the following *flow balance equations*:

$$Q(t; \pi) = Q(0; \pi) + A_q(t; \pi) - \sum_{k=1}^K B_k(t; \pi) - L(E(t; \pi)), \quad (2.3)$$

$$E(t; \pi) = \int_0^t Q(s; \pi) ds, \quad (2.4)$$

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

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

$$Y(t; \pi) = Y(0; \pi) + A(t) - \sum_{k=1}^K D_k(T_k(t; \pi)) - L(E(t; \pi)), \quad (2.7)$$

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

$$T_k(t; \pi) + I_k(t; \pi) = N_k t. \quad (2.9)$$

Finally, for any work conserving policy π we have the additional three equations:

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

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

and

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

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.

For a given staffing vector the *routing* (dynamic control) problem is defined as follows:

$$\begin{aligned} \text{minimize } & P_{\pi}(ab), \\ & \pi = \pi(\lambda, \vec{N}) \in \Pi. \end{aligned} \quad (2.13)$$

In the following section we address a simpler version of the above routing problem by considering policies which allow for preemption. Specifically, it is allowed to hand-off, at any time, a customer from one server to another.

3 Optimal Preemptive Routing

In this section we describe a simple *preemptive* policy, FSF_p ((preemptive) Faster Server available First (the subscript p is for preemptive)), which is optimal within the set of all non-anticipating, but possibly preemptive policies, with respect to the fraction of customers who abandon. Section 4.2 will describe our proposed non-preemptive policy, FSF (Faster Server First), which is also simple, but is not necessarily optimal for any fixed size system. However, it is *asymptotically* optimal as the system grows large (that is, as $\lambda \rightarrow \infty$ in the QED regime), in terms of the fraction of abandonment.

Consider a fixed system with fixed arrival rate and staffing vector. Furthermore, consider the more general family of policies $\Pi_p \supseteq \Pi$, which is the family of all non-anticipating policies which are preemptive resume. What is meant by preemptive resume in the context of this paper is that a customer who is served by a particular server may be handed-off to another server, who will resume the service from the point it has been discontinued. In addition, we add the following restriction on each policy belonging to this family: It only performs a *finite* number of *actions* in any finite time interval. An action refers to an assignment of a customer to a certain server, or a hand-off of a customer from one server to another.

Let $\text{FSF}_p \in \Pi_p$ be the policy in Π_p which is characterized by the following two properties: At any time point $t \geq 0$,

1. **Faster servers are used first:** If $Z_k(t; \text{FSF}_p) < N_k$ then $Z_j(t; \text{FSF}_p) = 0$, for all $j < k$.
2. **Work conservation:** If $Z_1(t; \text{FSF}_p) + Z_2(t; \text{FSF}_p) + \dots + Z_K(t; \text{FSF}_p) < N$ then $Q(t; \text{FSF}_p) = 0$.

The next proposition establishes the optimality of FSF_p within Π_p .

Proposition 3.1 (Optimal preemptive routing) *Consider the preemptive routing policy, FSF_p , that keeps the faster servers busy whenever possible. Then it is optimal in the sense that it stochastically minimizes the cumulative number of customers who have abandoned the system by any time $t \geq 0$, within the family of non-anticipating, possibly preemptive and possibly not work-conserving policies. In particular, FSF_p minimizes $P(ab)$, the steady-state fraction of abandonment, which is also, by PASTA, the steady-state probability that an arbitrary customer will abandon.*

Remark 3.1 *In Proposition 3.1 of [2] we show that in the non-abandonment case FSF_p stochastically minimizes the steady-state number of customers in the system within Π_p . This is no longer true when abandonment is introduced into the system. Specifically, to minimize the total number of jobs in the system, it might actually pay to idle some servers and have customers abandon instead of waiting until their service ends. Proposition 3.3 below shows that such intentional idling can also be helpful in minimizing the delay probability.*

Corollary 3.1 *Recall that $Q(\infty)$ and $W(\infty)$ are the steady-state queue length and waiting time, respectively. The preemptive routing policy, FSF_p , that always assigns customers to the faster servers first, is also optimal in the sense that it minimizes the steady-state expected queue length $E[Q(\infty)]$ and the steady-state expected waiting time $E[W(\infty)]$.*

Due to Remark 3.1, Proposition 3.1 is restricted to optimality with respect to minimization of the total number of abandonment and not the total number of customers in the system. However, if we restrict our attention to work conserving policies one can indeed show that FSF_p is also optimal with respect to the head-count process. This is stated in the next proposition.

Proposition 3.2 (Optimal preemptive routing within work conserving policies) *Consider the preemptive routing policy, FSF_p , that keeps the faster servers busy whenever possible. Then it is optimal in the sense that it stochastically minimizes the total number of customers in the system at any time t (including $t = \infty$, i.e. in steady state) within the family of non-anticipating, possibly preemptive policies, which are also work conserving. In other words, for all $\pi \in \Pi_p$ if π is also work-conserving then, assuming both systems start in the same state at time 0, we have $P\{Y(t; \pi) > y\} \geq P\{Y(t; FSF_p) > y\}$, for all $y \geq 0$ and all $0 \leq t \leq \infty$.*

Corollary 3.2 (Optimality of FSF_p with respect to delay probability within work conserving policies) *The policy FSF_p stochastically minimizes both the queue-length and the waiting time in steady-state among all the work conserving policies in Π_p . In particular, FSF_p minimizes the delay probability $P(wait > 0)$ in steady-state within this set of policies.*

It turns out that if we do not restrict our attention to work conserving policies, the steady-state delay probability is *not* minimized by FSF_p if the abandonment rate θ is large enough. The idea is simple; if one is only interested in minimizing $P(\text{wait} > 0)$ then there is no point in serving customers who have waited already. If θ is large enough, those customers abandon the system rapidly, leaving an empty queue for new customers. This is formally stated in the following proposition.

Proposition 3.3 *If the work conservation assumption is relaxed, then FSF_p does not minimize $P(\text{wait} > 0)$ within Π_p . In particular, if θ is large enough, it is possible to reduce the delay probability by idling servers even when customers are waiting in line.*

Remark 3.2 (In the absence of the FCFS assumption) *If one does not require that the service policy is FCFS then it is possible to obtain an even lower waiting probability by assigning an arriving customer to a server immediately upon arrival, if there is an available server at this time, and not to serve the customer at all (that is, let the customer wait until abandonment), otherwise.*

As opposed to minimization with respect to the delay probability, it turns out that, when minimizing the abandonment probability, the FCFS assumption is not necessary. This is stated in the next proposition.

Proposition 3.4 *Consider the set Π_{all} of all non-anticipating policies, which are possibly preemptive, and are not necessarily FCFS. Then to minimize $P_\pi(ab)$ within Π_{all} , it is sufficient to only consider FCFS policies.*

Remark 3.3 (State-space collapse under FSF_p) *Note the state-space collapse associated with the policy FSF_p . For a work conserving policy, the state-space is generally K dimensional. However, under FSF_p it is sufficient to know the total number of customers in the system in order to specify exactly how they are distributed among the server pools and the queue. In particular, the total number of jobs in the system Y may be described as a Birth and Death process with constant birth rates*

$$\lambda(y) \equiv \lambda, \quad \forall y \geq 0,$$

and a piecewise-linear death rate function:

$$\mu(y) = \begin{cases} y\mu_K & \text{if } y \leq N_K \\ (y - N_K)\mu_{K-1} + N_K\mu_K & \text{if } N_K < y \leq N_{K-1} + N_K \\ \cdot & \\ \cdot & \\ (y - (N_2 + \dots + N_K))\mu_1 + N_2\mu_2 + \dots + N_K\mu_K & \text{if } N_2 + \dots + N_K < y \leq N \\ (y - N)\theta + N_1\mu_1 + N_2\mu_2 + \dots + N_K\mu_K & \text{if } y > N. \end{cases} \quad (3.1)$$

The following proposition stipulates that the queue length and the total number of idle servers in the inverted-V system, working under any work-conserving policy, may be bounded from above and from below by the queue length and number of idle servers in two corresponding $M/M/N + M$ systems, respectively. We refer back to it, in our asymptotic analysis, to establish tightness (proof of Proposition 4.6) and uniform integrability (proof of Theorem 4.1) of the relevant scaled processes.

Proposition 3.5 *Consider three systems:*

- A) *System A is the inverted-V system of this paper, with K pools of servers, with respective service rates: $\mu_1 < \mu_2 < \dots < \mu_K$ and respective pool sizes: N_1, N_2, \dots, N_K . Suppose that system A works under an arbitrary work-conserving policy.*
- B) *System B is an $M/M/N_B + M$ system with N_B servers, all working with rate μ_K , and $N_1\mu_1 + N_2\mu_2 + \dots + N_K\mu_K \geq N_B\mu_K$.*
- C) *System C is an $M/M/N_C + M$ system with N_C servers, all working with rate μ_1 and $N_1\mu_1 + N_2\mu_2 + \dots + N_K\mu_K \leq N_C\mu_1$.*

All systems have the same arrival rate λ and the same individual abandonment rate θ . Then there are versions of the queue length processes Q_A, Q_B and Q_C and the total number of busy servers Z_A, Z_B and Z_C associated with systems A, B and C, respectively, such that $Q_C \leq Q_A \leq Q_B$ and $Z_C - N_C \leq Z_A - (N_1 + N_2 + \dots + N_K) \leq Z_B - N_B$, at all times, almost surely.

4 Asymptotically optimal control

4.1 Asymptotic Framework

The joint staffing and routing problem (2.1) is difficult to solve in a closed form. Specifically, given fixed values of $\mu_1 < \mu_2 < \dots < \mu_K, \lambda$ and $\vec{N} = (N_1, N_2, \dots, N_K)$, one needs to find a policy $\pi = \pi(\lambda, \vec{N}) \in \Pi$ that minimizes the probability of abandonment, in order to determine if the staffing vector \vec{N} is feasible for the problem (2.1). This is hard to do. In addition, one would need to develop an efficient search technique to find the optimal staffing vector among all the feasible ones. Instead, we take an asymptotic approach, which leads to *asymptotically* optimal routing rules for systems with many servers and high demand (i.e. large values of λ and \vec{N}). To this end, we consider a sequence of systems indexed by λ (to appear as a superscript) with increasing arrival rates $\lambda \uparrow \infty$, and increasing total number of servers N^λ but with *fixed* service rates $\mu_1, \mu_2, \dots, \mu_K$, and a *fixed* abandonment rate θ . We first solve the control problem asymptotically. We then use this solution in identifying asymptotically optimal staffing in section 5.

Assume that there are K numbers $a_k \geq 0$, $k = 1, \dots, K$, with $a_1 > 0$ and $\sum_{k=1}^K a_k = 1$, such that the number of servers of each pool N_k^λ , $k = 1, 2, \dots, K$, 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,

$$\rho^\lambda := \frac{\lambda}{\sum_{k=1}^K \mu_k N_k^\lambda}, \quad (4.1)$$

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* for server pool k . Now introduce

$$\mu := \left[\sum_{k=1}^K a_k / \mu_k \right]^{-1}; \quad (4.2)$$

then λ / μ can be interpreted as the total offered load for the whole system. Given this definition of μ , (A1) implies that

$$N^\lambda = \frac{\lambda}{\mu} + o(\lambda), \text{ as } \lambda \rightarrow \infty, \quad \text{or,} \quad \lim_{\lambda \rightarrow \infty} \frac{\lambda}{N^\lambda} = \mu, \quad (4.3)$$

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

$$\rho^\lambda \approx \frac{\lambda}{N^\lambda \mu}, \quad (4.4)$$

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

$$q_k := \lim_{\lambda \rightarrow \infty} \frac{N_k^\lambda}{N^\lambda} = \frac{a_k}{\mu_k} \mu \geq 0, \quad k = 1, \dots, K, \quad (4.5)$$

where q_k is the limiting fraction of pool k servers out of the total number of servers. The condition $a_1 > 0$ guarantees that $q_1 > 0$, and hence the slowest server pool 1 is asymptotically non-negligible in size. Clearly, $\sum_{k=1}^K q_k = 1$ and $\sum_{k=1}^K q_k \mu_k = \mu$.

Fluid Scaling: In view of the above discussion, one observes that assumption (A1) implies that quantities involved in the process such as the arrival rate, the offered load, and the size of the different server pools are all of the order of N^λ . Therefore, one expects to get finite limits of these quantities when dividing all of them by N^λ . As it turns out, due the functional strong law of large numbers (FSLLN), this scaling leads to the fluid dynamics of the system, in the limit as $\lambda \rightarrow \infty$. To see this, for $\lambda \uparrow \infty$, $k = 1, \dots, K$, 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 $\bar{Y}^\lambda(t) = \frac{Y^\lambda(t)}{N^\lambda}$, $\bar{A}^\lambda(t) = \frac{A^\lambda(t)}{N^\lambda}$, $\bar{A}_k^\lambda(t) = \frac{A_k^\lambda(t)}{N^\lambda}$, $\bar{A}_q^\lambda(t) = \frac{A_q^\lambda(t)}{N^\lambda}$, $\bar{B}_k^\lambda(t) = \frac{B_k^\lambda(t)}{N^\lambda}$, $\bar{T}_k^\lambda(t) = \frac{T_k^\lambda(t)}{N^\lambda}$, $\bar{I}_k^\lambda(t) = \frac{I_k^\lambda(t)}{N^\lambda}$, and $\bar{E}^\lambda(t) = \frac{E^\lambda(t)}{N^\lambda}$. Finally, let $\bar{D}_k^\lambda(t) = D_k^\lambda(t) = D_k(t)$ and $\bar{L}^\lambda(t) = L^\lambda(t) = L(t)$. That is, as equalities between processes,

$$(\bar{Q}^\lambda, \bar{Z}_k^\lambda, \bar{Y}^\lambda, \bar{A}^\lambda, \bar{A}_k^\lambda, \bar{A}_q^\lambda, \bar{B}_k^\lambda, \bar{T}_k^\lambda, \bar{I}_k^\lambda, \bar{E}^\lambda) = (Q^\lambda, Z_k^\lambda, Y^\lambda, A^\lambda, A_k^\lambda, A_q^\lambda, B_k^\lambda, T_k^\lambda, I_k^\lambda, E^\lambda) / N^\lambda,$$

and $(\bar{D}_k^\lambda, \bar{L}^\lambda) = (D_k, L)$. Note that D_k^λ and L^λ need not be divided by N^λ , due to their definitions as Poisson processes with rates μ_k and θ , respectively, which are independent of λ .

Using standard tools of fluid models (see for example [11], Theorem A.1) one can show that if $(\bar{Q}^\lambda(0), \bar{Z}_k^\lambda(0), k = 1, \dots, K)$ are bounded, then the process $\bar{\mathbb{X}} := (\bar{Q}^\lambda, \bar{Z}_k^\lambda, \bar{Y}^\lambda, \bar{A}^\lambda, \bar{A}_k^\lambda, \bar{A}_q^\lambda, \bar{B}_k^\lambda, \bar{T}_k^\lambda, \bar{I}_k^\lambda, \bar{E}^\lambda, \bar{D}_k^\lambda, \bar{L}^\lambda)$ 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 $\bar{Q}^\lambda(t)$, as $\lambda \rightarrow \infty$). Note that, by Theorem A.1 of [11], equations (2.3)-(2.9) imply that the following flow balance equations hold for *any* fluid limit:

$$\bar{Q}(t) = \bar{Q}(0) + \bar{A}_q(t) - \sum_{k=1}^K \bar{B}_k(t) - \theta \bar{E}(t), \quad (4.6)$$

$$\bar{Q}(t) = \int_0^t \bar{E}(s) ds, \quad (4.7)$$

$$\bar{Z}_k(t) = \bar{Z}_k(0) + \bar{A}_k(t) + \bar{B}_k(t) - \mu_k \bar{T}_k(t), \quad k = 1, \dots, K, \quad (4.8)$$

$$\bar{T}_k(t) = \int_0^t \bar{Z}_k(s) ds \quad (4.9)$$

$$\bar{Y}(t) = \bar{Y}(0) + \mu t - \sum_{k=1}^K \mu_k \bar{T}_k(t) - \theta \bar{E}(t), \quad (4.10)$$

$$\mu t = \bar{A}_q(t) + \sum_{k=1}^K \bar{A}_k(t), \quad (4.11)$$

$$\bar{T}_k(t) + \bar{I}_k(t) = q_k t. \quad (4.12)$$

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

$$\bar{Q}(t) \cdot \left(\sum_{k=1}^K (q_k - \bar{Z}_k(t)) \right) = 0, \quad (4.13)$$

$$\int_0^\infty \sum_{k=1}^K (q_k - \bar{Z}_k(t)) d\bar{A}_q(t) = 0, \quad (4.14)$$

and

$$\sum_{k=1}^K \int_0^\infty \bar{Q}(t) d\bar{I}_k(t) = 0. \quad (4.15)$$

The following proposition shows that for every sequence of work-conserving routing policies and for every fluid limit, the quantities $\bar{Q}(t)$ and $\bar{Z}_k(t)$, $k = 1, \dots, K$, remain constant if starting at time 0 from some appropriate initial conditions.

Proposition 4.1 (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{X} be some fluid limit of the processes associated with the system, as $\lambda \rightarrow \infty$. Recall that $q_k = \lim_{\lambda \rightarrow \infty} \frac{N_k^\lambda}{N^\lambda} = \frac{a_k}{\mu_k} \mu$, $k = 1, \dots, K$, and suppose that $\bar{Q}(0) = 0$ and $\bar{Z}_k(0) = q_k$, $k = 1, \dots, K$. Then, $\bar{Q}(t) \equiv 0$ and $\bar{Z}_k(t) \equiv q_k$, $k = 1, \dots, K$, for all $t \geq 0$.

In addition to the fluid scaling, we introduce a more refined *diffusion* scaling which we proceed to define.

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 $\bar{X}^\lambda(\cdot) = (X_1^\lambda(\cdot), \dots, X_K^\lambda(\cdot))$ as follows:

$$X_1^\lambda(t) := \frac{Q^\lambda(t) + Z_1^\lambda(t) - N_1^\lambda}{\sqrt{N^\lambda}}, \quad (4.16)$$

and, for $k = 2, \dots, K$, let

$$X_k^\lambda(t) := \frac{Z_k^\lambda(t) - N_k^\lambda}{\sqrt{N^\lambda}}. \quad (4.17)$$

Note that for $k = 2, \dots, K$, $X_k^\lambda(t) \leq 0$ for all t , and that for all $k = 1, 2, \dots, K$, $[X_k^\lambda(t)]^- := -\min\{X_k^\lambda(t), 0\}$ corresponds to the number of idle servers, scaled by $1/\sqrt{N^\lambda}$. Similarly, $[X_1^\lambda(t)]^+$ corresponds to the total queue length, again, scaled by $1/\sqrt{N^\lambda}$. Finally, let

$$X^\lambda(t) := \sum_{k=1}^K X_k^\lambda(t) = \frac{Q^\lambda(t) + \sum_{k=1}^K Z_k^\lambda(t) - N^\lambda}{\sqrt{N^\lambda}} = \frac{Y^\lambda(t) - N^\lambda}{\sqrt{N^\lambda}} = \sqrt{N^\lambda} (\bar{Y}^\lambda(t) - 1). \quad (4.18)$$

The process $X^\lambda(\cdot)$ captures the fluctuations of order $1/\sqrt{N^\lambda}$ of $\bar{Y}^\lambda(\cdot)$ about its fluid limit. Also, $[X^\lambda(t)]^-$ is the total number of idle servers, and $[X^\lambda(t)]^+ = [X_1^\lambda(t)]^+$ is the total queue length, both scaled by $1/\sqrt{N^\lambda}$. Finally, note that, from work conservation, if $X_k^\lambda(t) < 0$ for some k , then $X_1^\lambda(t) \leq 0$.

For all $\lambda > 0$, let $V^\lambda(t)$ be the offered (virtual) waiting time of an arbitrary, infinitely patient, customer who arrives to the λ -system at time t , and let $V^\lambda(\infty)$ be the offered waiting time in steady-state. Denote by $W^\lambda(\infty)$ the actual waiting time in steady-state, defined as $W^\lambda(\infty) := V^\lambda(\infty) \wedge \tau$, where τ is the patience of a typical customer, which is independent of $V^\lambda(\infty)$. The scaled offered waiting time process is defined as follows

$$\hat{V}^\lambda(t) = \sqrt{N^\lambda} V^\lambda(t), \quad t \geq 0, \quad \lambda > 0, \quad (4.19)$$

and its scaled steady-state is

$$\hat{V}^\lambda(\infty) = \sqrt{N^\lambda} V^\lambda(\infty), \quad \lambda > 0. \quad (4.20)$$

Finally, the scaled steady-state actual waiting time is

$$\hat{W}^\lambda(\infty) = \sqrt{N^\lambda} W^\lambda(\infty), \quad \lambda > 0. \quad (4.21)$$

As will be shown later, in order for the diffusion scaling to have well defined limits, as $\lambda \rightarrow \infty$, the following assumption must be introduced, in conjunction with (A1):

$$\sum_{k=1}^K \mu_k N_k^\lambda = \lambda + \delta \sqrt{\lambda} + o(\sqrt{\lambda}), \text{ as } \lambda \rightarrow \infty, \quad \text{or,} \quad \lim_{\lambda \rightarrow \infty} \frac{\sum_{k=1}^K \mu_k N_k^\lambda - \lambda}{\sqrt{\lambda}} = \delta, \quad (\text{A2})$$

for some δ , $-\infty < \delta < \infty$.

Condition (A2) is a square-root staffing rule (similar to [23] and [9]). As shown later (Corollary 4.3), it guarantees that under the appropriate routing, the fraction of abandonment is of the order of $1/\sqrt{\lambda}$. Note that (A2) does not specify how the overall staffing level is divided among 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 another will compensate for this deficit by having more than the nominal staffing. For $k = 1, \dots, K$, and $\lambda > 0$, let $-\infty < \delta_k^\lambda < \infty$ be defined as:

$$\delta_k^\lambda := \frac{\mu_k N_k^\lambda - a_k \lambda}{\sqrt{\lambda}}. \quad (4.22)$$

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, clearly, $\delta_k^\lambda \geq 0$ if $a_k = 0$,

$$\delta_k^\lambda = o(\sqrt{\lambda}), \text{ as } \lambda \rightarrow \infty, \quad \forall k = 1, \dots, K, \quad (4.23)$$

and

$$\delta^\lambda := \sum_{k=1}^K \delta_k^\lambda \rightarrow \delta, \text{ as } \lambda \rightarrow \infty. \quad (4.24)$$

Note that we do not require the individual sequences $\{\delta_k^\lambda\}_{\lambda > 0}$ to have a limit, for any value of $k = 1, \dots, K$. All that is assumed is that their sum converges to δ . The one exception to this rule is Proposition 4.4, in which the following additional condition is assumed to hold:

$$\eta := \lim_{\lambda \rightarrow \infty} \sum_{k=1}^K \frac{\delta_k^\lambda}{\mu_k} = \lim_{\lambda \rightarrow \infty} \sqrt{\lambda} \left(\frac{N^\lambda}{\lambda} - \frac{1}{\mu} \right), \text{ exists for some finite number } \eta. \quad (\text{A3})$$

Lastly, let \vec{N}^λ be the staffing vector of the λ -system. Then, we define the *scaled* version of the routing problem (2.13) to be

$$\begin{aligned} \text{minimize} \quad & \limsup_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\pi^\lambda}^\lambda(ab), \\ & \pi^\lambda = \pi(\lambda, \vec{N}^\lambda) \in \Pi. \end{aligned} \quad (4.25)$$

Corollary 4.3 justifies the above as the sensible scaling of $P^\lambda(ab)$ under conditions (A1) and (A2).

4.2 Faster Server First (FSF) is Asymptotically Optimal

In this section we describe a simple non-preemptive policy FSF which is also work-conserving. This policy may be described simply as follows: Upon a customer arrival or a service completion, assign the first customer in the queue (or the one that has just arrived, if the queue is empty) to the fastest available server (which is the server with the largest index k). Judging by the literature on the slow server problem (e.g. [26]), this policy is not likely to be optimal. However, as we show in this section, it is *asymptotically* optimal as the arrival rate λ grows to ∞ and the number of servers per pool grow according to (A1) and (A2); the asymptotic optimality is in terms of the steady-state probability of abandonment, and, equivalently, in terms of the expected queue length and expected waiting time, both in steady-state. The main premise of this section is the asymptotic optimality of FSF within the family of non-preemptive non-anticipating FCFS policies. This is summarized in Theorem 4.1.

Theorem 4.1 (Asymptotic optimality of FSF) *Consider a sequence of systems indexed by the arrival rate λ , that satisfy conditions (A1) and (A2). Then the non-preemptive policy FSF, that assigns customers to the fastest server available whenever a customer arrives, or upon service completion, is asymptotically optimal with respect to (4.25) within the family Π_p of all possibly-preemptive, non-anticipating FCFS policies. In particular, it is asymptotically optimal with respect to (4.25) within the family Π of non-preemptive non-anticipating FCFS policies.*

Corollary 4.1 *Consider a sequence of systems indexed by the arrival rate λ , that satisfy conditions (A1) and (A2). Then the non-preemptive policy FSF, that assigns customers to the fastest server available whenever a customer arrives, or upon service completion, is asymptotically optimal with respect to the minimization of $\limsup_{\lambda \rightarrow \infty} E\hat{Q}^\lambda(\infty; \pi^\lambda)$ and $\limsup_{\lambda \rightarrow \infty} E\hat{W}^\lambda(\infty; \pi^\lambda)$.*

To prove the asymptotic optimality of FSF, as $\lambda \rightarrow \infty$, we will show that, as λ grows, the process $(X_1^\lambda(\cdot), X_2^\lambda(\cdot), \dots, X_K^\lambda(\cdot))$ (recall the diffusion scaling in Section 4.1) under FSF becomes close to the same process under the preemptive policy FSF_p , and in the limit, as $\lambda \rightarrow \infty$, the two processes coincide. Taking the limits as $t \rightarrow \infty$, we will also show that the corresponding steady-state processes become close, and hence, the optimality of FSF_p in steady-state (see Corollary 3.1) will imply the asymptotic optimality of FSF. The key step in the proof of the equivalence between the two processes is the state-space collapse of the process $(X_1^\lambda(\cdot), X_2^\lambda(\cdot), \dots, X_K^\lambda(\cdot))$ under FSF, into a one dimensional process, as $\lambda \rightarrow \infty$. Recall, that such state-space collapse holds for every λ under FSF_p (by Proposition 3.1 and Remark 3.3). When FSF is used, however, this is no longer true, but the state-space collapse is attained when $\lambda \rightarrow \infty$, as will be shown in Proposition 4.2 below.

4.2.1 State-Space Collapse

In this section we establish the state-space collapse result with respect to the policy FSF and the process $\vec{X}^\lambda(\cdot) = (X_1^\lambda(\cdot), \dots, X_K^\lambda(\cdot))$. Since the policy here is fixed we omit FSF from all notation. Essentially, the state-space collapse result indicates that, as λ grows, the one-dimensional process $X^\lambda(\cdot)$ (see (4.18)) becomes sufficient in describing the whole K -dimensional process $\vec{X}^\lambda(\cdot)$. Specifically, we show that, as $\lambda \rightarrow \infty$, all the faster servers (from pools $k = 2, \dots, K$) are constantly busy (or, more accurately, the number of idle servers in these pools is of order $o(\sqrt{N^\lambda})$), and the only possible idleness is within the slowest servers (pool 1). Hence, as λ grows, the processes $X_2^\lambda(\cdot), \dots, X_K^\lambda(\cdot)$ become negligible, while the processes $X^\lambda(\cdot)$ and $X_1^\lambda(\cdot)$ get close. This result is presented in the following proposition:

Proposition 4.2 (State-space collapse) *Suppose that conditions (A1) and (A2) hold, as $\lambda \rightarrow \infty$, and that the work-conserving non-preemptive policy FSF is used. In addition, suppose that $\vec{X}^\lambda(0) \rightarrow \vec{X}(0) = \vec{x} = (x_1, \dots, x_K)$, in probability, as $\lambda \rightarrow \infty$. Then, for all $t > 0$, we have,*

$$X_k^\lambda(t) \xrightarrow{p} 0, \text{ uniformly on compact intervals, as } \lambda \rightarrow \infty, \quad \forall k \geq 2.$$

Remark 4.1 (State-space collapse for FSF_p in the QED regime) Proposition 4.2 is also true if the preemptive policy FSF_p is used. Here the proof is much simpler.

4.2.2 Transient Diffusion Limit

In this section, we establish the form of the diffusion limit of the scaled process \vec{X}^λ . Since our paper is mostly concerned with the optimization of steady-state performance measures, the transient diffusion limit can be regarded as an intermediate step in obtaining the steady-state limit (based on the equivalence between the transient diffusion limits of FSF_p and FSF). However, the form of the diffusion process obtained in the limit is also interesting in its own right. Especially, when compared with the diffusion limit obtained in [18] for the M/M/N+M system (see Corollary 4.2). In a nutshell, this comparison reveals that, for equal levels of overall service capacity, the multi-pool system under the policy FSF (stochastically) outperforms the single type M/M/N+M system.

We note that the state-space collapse result of Proposition 4.2 essentially shows that it is sufficient to find the diffusion limit of the total count of customers (centered and scaled) X^λ . Denoting this limit by X , we have that the limit of X_k^λ , for $k \geq 2$, is identically zero, and the limit of X_1^λ is hence equal to X .

Proposition 4.3 (Transient diffusion limit) *Suppose that $X_k^\lambda(0) \Rightarrow X_k(0)$, as $\lambda \rightarrow \infty$, for $k = 1, \dots, K$, and let $X(0) = \sum_{k=1}^K X_k(0)$. Assume further that (A1) and (A2) hold, and that the policy*

FSF is used. Recall that $\mu_1 < \mu_2 < \dots < \mu_K$, and $\mu = \left[\sum_{k=1}^K a_k / \mu_k \right]^{-1}$. Then, $X^\lambda \Rightarrow X$, as $\lambda \rightarrow \infty$, where X is a diffusion process with the infinitesimal drift

$$m(x) = \begin{cases} -\delta\sqrt{\mu} - \theta x & x \geq 0, \\ -\delta\sqrt{\mu} - \mu_1 x & x < 0, \end{cases} \quad (4.26)$$

and infinitesimal variance

$$\sigma^2(x) = 2\mu. \quad (4.27)$$

Corollary 4.2 (Outperforming the single server skill system) Consider, in comparison to the inverted-V system operating under FSF, a sequence of systems with a single customer class and a single server pool, instead of K pools. Suppose that all these servers have service rate μ (as defined in (4.2)). In addition, suppose that the sequence of arrival rates, $\{\lambda\}$, is identical for both models, and that the number of servers in the single pool model, $N^{S,\lambda}$, satisfies $N^{S,\lambda}\mu = \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda})$, as $\lambda \rightarrow \infty$. That is, in both models the excess capacity is approximately equal to $\delta\sqrt{\lambda}$. For the single-pool model, let $Y^{S,\lambda}(t)$ be the total number of customers in the system at time t , and $X^{S,\lambda}(t) = (Y^{S,\lambda}(t) - N^{S,\lambda})/\sqrt{N^{S,\lambda}}$. Then, by [18], if $X^{S,\lambda}(0) \Rightarrow X^S(0)$, as $\lambda \rightarrow \infty$, then $X^{S,\lambda} \Rightarrow X^S$, as $\lambda \rightarrow \infty$, where X^S is a diffusion process with the infinitesimal drift

$$m^S(x) = \begin{cases} -\delta\sqrt{\mu} - \theta x & x \geq 0, \\ -\delta\sqrt{\mu} - \mu x & x < 0, \end{cases} \quad (4.28)$$

and infinitesimal variance

$$(\sigma^S)^2(x) = 2\mu. \quad (4.29)$$

In particular, the diffusion limits of both processes are of the same form, with the exception that $-\mu x$ in (4.28) replaces $-\mu_1 x$ in (4.26), within the drift component that applies when there are idle servers.

The comparison between the two diffusion processes in Corollary 4.2 indicates that the limiting process associated with the \wedge -design stochastically dominates the process associated with the I -design. This result could be viewed as surprising because one would expect that more variance in service time would lead to worse performance. Moreover, it is an asymptotic result and examples can be identified in which the homogeneous server system actually *outperforms* its heterogenous server counterpart. However, in the QED regime, the FSF policy uses the servers efficiently, and therefore the heterogenous system is indeed better. The managerial implication of this result, on the design and staffing of such large service systems, is that heterogeneity in the server population is an asset and not a liability, if exploited intelligently. An analogous observation was made in [2] for the system with no abandonment.

Remark 4.2 (Transient diffusion limit for FSF_p) Proposition 4.3 and Corollary 4.2 remain true if the preemptive policy FSF_p is used instead. The proofs remain unchanged due to Remark 4.1.

We conclude this section by establishing the transient diffusion limit of the scaled virtual waiting time process, which turns out to be a simple linear function of the corresponding limit of the queue length process. Note that here, in addition to the two assumptions (A1) and (A2), one also needs to assume that (A3) holds.

Proposition 4.4 (Transient diffusion limit of the virtual waiting time process) *Suppose that $X_k^\lambda(0) \Rightarrow X_k(0)$ as $\lambda \rightarrow \infty$, for $k = 1, \dots, K$, and let $X(0) = \sum_{k=1}^K X_k(0)$. Assume further that (A1), (A2) and (A3) hold, and that the policy FSF is used. Then, $\hat{V}^\lambda := \sqrt{N^\lambda} V^\lambda \Rightarrow \hat{V}$, as $\lambda \rightarrow \infty$, where $\hat{V} = [X]^+ / \mu$, and X is the diffusion limit of X^λ as $\lambda \rightarrow \infty$, given in Proposition 4.3.*

4.2.3 Stationary Diffusion Limit

In this section we prove that the stationary distributions of the process \vec{X}^λ , under both FSF_p and FSF, converge to the stationary distribution of \vec{X} , as $\lambda \rightarrow \infty$. In particular, this implies the asymptotic optimality of FSF within Π in terms of the steady-state queue length and waiting time, due to the optimality of FSF_p in Π_p . Specifically, we first spell out the stationary distribution of X , the limiting diffusion process given in Proposition 4.3. Next we show that the stationary distribution of X^λ under both FSF_p and FSF converges to this stationary distribution. In all processes we use ∞ in place of the time argument to denote steady-state.

Proposition 4.5 (Stationary distribution of the diffusion process) *Let $X(\cdot)$ be the diffusion process described in Proposition 4.3, with infinitesimal drift and variance as in (4.26) and (4.27). Then the steady-state distribution of X has the density $f(\cdot)$ given by*

$$f(x) = \begin{cases} \frac{\sqrt{\frac{\theta}{\mu}} \phi\left(\sqrt{\frac{\theta}{\mu}} x + \frac{\delta}{\sqrt{\theta}}\right)}{1 - \Phi\left(\frac{\delta}{\sqrt{\theta}}\right)} \alpha, & \text{if } x \geq 0 \\ \frac{\sqrt{\frac{\mu_1}{\mu}} \phi\left(\sqrt{\frac{\mu_1}{\mu}} x + \frac{\delta}{\sqrt{\mu_1}}\right)}{\Phi\left(\frac{\delta}{\sqrt{\mu_1}}\right)} (1 - \alpha), & \text{if } x < 0, \end{cases} \quad (4.30)$$

where $\alpha := \alpha(\delta, \mu_1, \theta) = \left[1 + \frac{\sqrt{\theta} h(\delta/\sqrt{\theta})}{\sqrt{\mu_1} h(-\delta/\sqrt{\mu_1})}\right]^{-1} = P\{X(\infty) \geq 0\}$, and $h(\cdot) = \frac{\phi(\cdot)}{1 - \Phi(\cdot)}$ is the hazard rate of the standard normal distribution. This steady-state distribution has the following means:

$$EX^+(\infty) = \alpha \left[\frac{-\delta\sqrt{\mu}}{\theta} + \sqrt{\frac{\mu}{\theta}} h\left(\frac{\delta}{\sqrt{\theta}}\right) \right], \quad (4.31)$$

and

$$EX^-(\infty) = (1 - \alpha) \left[\frac{\delta\sqrt{\mu}}{\mu_1} + \sqrt{\frac{\mu}{\mu_1}} h\left(-\frac{\delta}{\sqrt{\mu_1}}\right) \right], \quad (4.32)$$

We now turn to showing that under both FSF_p and FSF, the stationary distribution of $X^\lambda(\cdot)$ weakly converges to the stationary distribution of X , given in (4.30). Note that this convergence

does not immediately follow from Proposition 4.3, Remark 4.2 and Proposition 4.5. In particular, a double limit interchange (as both time and arrival rate go to infinity) must be justified.

Proposition 4.6 (Convergence of the steady-state distributions) *Suppose that conditions (A1) and (A2) hold, and that either FSF or FSF_p is used. Then the stationary distribution of \vec{X}^λ weakly converges, as $\lambda \rightarrow \infty$, to the stationary distribution of $\vec{X} = (X, 0, \dots, 0)$, where the stationary distribution of the first coordinate, X , is given in (4.30).*

Corollary 4.3 *Suppose that conditions (A1) and (A2) hold, and that either FSF or FSF_p is used. Then*

$$\lim_{\lambda \rightarrow \infty} E\hat{Q}^\lambda(\infty) := \lim_{\lambda \rightarrow \infty} E \frac{Q^\lambda(\infty)}{\sqrt{N^\lambda}} = EX^+(\infty), \quad (4.33)$$

$$\lim_{\lambda \rightarrow \infty} E\hat{W}^\lambda(\infty) := \lim_{\lambda \rightarrow \infty} E\sqrt{N^\lambda}\hat{W}(\infty) = EX^+(\infty)/\mu, \quad (4.34)$$

$$\lim_{\lambda \rightarrow \infty} \hat{P}^\lambda(ab) := \lim_{\lambda \rightarrow \infty} \sqrt{N^\lambda}P^\lambda(ab) = \theta EX^+(\infty)/\mu, \quad (4.35)$$

and

$$\lim_{\lambda \rightarrow \infty} \sqrt{\lambda}P^\lambda(ab) = \frac{\theta}{\sqrt{\mu}}EX^+(\infty) = \sqrt{\theta} \alpha \cdot \left[h \left(\frac{\delta}{\sqrt{\theta}} \right) - \frac{\delta}{\sqrt{\theta}} \right]. \quad (4.36)$$

where $EX^+(\infty)$ is given in (4.31).

The proof of the corollary is included in the proof of Theorem 4.1. From Corollary 4.3 it follows that, under conditions (A1) and (A2), $\sqrt{\lambda}P^\lambda(ab)$ converges to a well defined limit, as $\lambda \rightarrow \infty$, under the FSF policy. In particular, $\lim_{\lambda \rightarrow \infty} P^\lambda(ab) = 0$, which implies that the constraint in (2.1) is satisfied trivially in the limit. Therefore, under our asymptotic framework, the sensible optimization problem to focus on is (4.25).

5 Asymptotically Optimal Staffing

5.1 Asymptotic Feasibility

In this section, we wish to characterize the feasible region for the staffing problem (2.1). As was noted before, characterizing this region exactly for fixed $\lambda, \mu_1, \dots, \mu_K, \theta$ and Δ is difficult. Instead, we characterize this region asymptotically, for large values of λ (i.e. as $\lambda \rightarrow \infty$). First we define the asymptotic feasibility problem.

For given values of the parameters $\mu_1 < \mu_2 < \dots < \mu_K$, and θ and a given value of $0 < \Delta < \infty$, a sequence $\{\vec{N}^\lambda\}$ of staffing vectors, for a sequence of systems indexed by their arrival rate λ ,

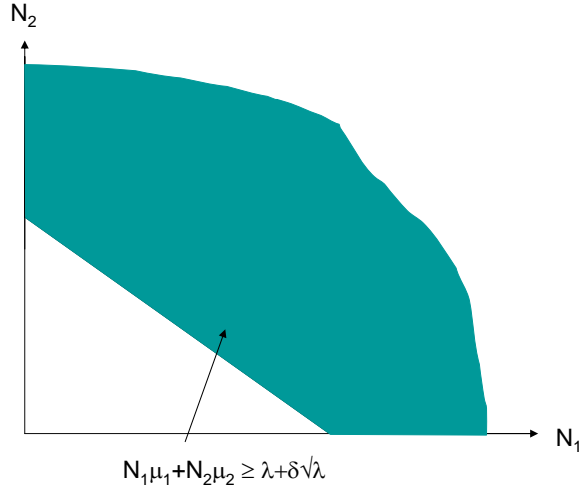


Figure 5.1: The Asymptotically Feasible Region for $K = 2$.

is called *asymptotically feasible* if there exists a sequence of routing policies $\pi^\lambda = \pi(\lambda, \vec{N}^\lambda) \in \Pi$ such that

$$\limsup_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\pi^\lambda}^\lambda(ab) \leq \Delta. \tag{5.1}$$

Proposition 5.1 characterizes the asymptotically feasible region. Although, a-priori, this region can be very complicated, it turns out to have a simple *linear* form. The linearity of the feasible region is not surprising in view of the fact that, under FSF, the limiting scaled abandonment probability depends on the overall service capacity (as long as the slowest server pool is of a non-negligible size - see (5.2) below). In particular, the limiting scaled abandonment probability does not depend on the individual capacities of the different server pools. Note that the overall service capacity is a linear function of the number of servers in each pool. Hence the linearity of the asymptotically feasible region. The asymptotically feasible region is illustrated in Figure 5.1.

Proposition 5.1 (Asymptotically Feasible Region - Square-Root “Safety” Capacity) *Let $0 < \Delta < \infty$, and let $0 < \mu_1 < \mu_2 < \dots < \mu_K$ and θ be fixed. Consider a sequence of systems indexed by the arrival rate $\lambda > 0$, with λ growing to infinity, and N_k^λ servers in pool k ,*

⁹Note that while in (2.1) the value of Δ is restricted to the open interval $(0, 1)$, here we allow for values of Δ in $(0, \infty)$. This is because the probability of abandonment on the left hand side of (5.1) is inflated by $\sqrt{\lambda}$.

$k = 1, \dots, K$. Let $N^\lambda = \sum_{k=1}^K N_k^\lambda$ be the total number of servers in system λ , and assume that

$$\liminf_{\lambda \rightarrow \infty} \frac{N_1^\lambda}{N^\lambda} > 0. \quad (5.2)$$

Then, there exists a sequence $\{\pi^\lambda = \pi^\lambda(\lambda, \vec{N}^\lambda)\}$ of non-preemptive policies, under which

$$\limsup_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\pi^\lambda}(ab) = \Delta \quad (5.3)$$

if and only if

$$\mu_1 N_1^\lambda + \dots + \mu_K N_K^\lambda \geq \lambda + \delta \sqrt{\lambda} + o(\sqrt{\lambda}), \quad \text{as } \lambda \rightarrow \infty, \quad (5.4)$$

where $-\infty < \delta < \infty$ satisfies

$$\Delta := \Delta(\delta, \mu_1, \theta) = \sqrt{\theta} \alpha \cdot \left[h\left(\frac{\delta}{\sqrt{\theta}}\right) - \frac{\delta}{\sqrt{\theta}} \right], \quad (5.5)$$

with $\alpha := \alpha(\delta, \mu_1, \theta) = \left[1 + \frac{\sqrt{\theta} h(\delta/\sqrt{\theta})}{\sqrt{\mu_1} h(-\delta/\sqrt{\mu_1})} \right]^{-1} = P\{X(\infty) \geq 0\}$.

In addition, $\delta = -\infty$ (i.e. (5.4) is violated for all $\delta > -\infty$) if and only if $\Delta = \infty$, and $\delta = \infty$ (i.e. (5.4) holds for any $\delta > 0$) if and only if (5.3) holds for any arbitrary $\infty > \Delta > 0$, with the appropriate choice of π^λ .

Lemma 5.1 *The function $\Delta(\cdot)$ defined in (5.5) is continuous and monotonically decreasing in δ . Moreover, $\lim_{\delta \rightarrow \infty} \Delta(\delta, \mu_1, \theta) = 0$ and $\lim_{\delta \rightarrow -\infty} \Delta(\delta, \mu_1, \theta) = \infty$.*

The proof of Lemma 5.1 follows, in a straightforward manner, from the proof of Theorem 4.1 in [27].

5.2 Asymptotically Optimal Staffing

In this section, we study the staffing problem (2.1). Recall that exact optimality is difficult to obtain, and hence, we present asymptotically optimal solutions. Our previous results already identify, under certain conditions, an asymptotically optimal routing policy (FSF) and the asymptotically feasible region given in (5.4). It is now left to find the asymptotically optimal staffing rule that minimizes the staffing costs among all the sequence of vectors $\vec{N}^\lambda = (N_1^\lambda, \dots, N_K^\lambda)$, which belong to the asymptotically feasible region. For the remainder of this section, we fix a target scaled abandonment probability $0 < \Delta < \infty$.

Consider a cost function $C(\vec{N}) = C_1(N_1) + \dots + C_K(N_K)$, which is increasing and strictly convex in all of its arguments, and such that $C(\vec{N}) \rightarrow \infty$, as $\|\vec{N}\| \rightarrow \infty$. Because of the characterization of the feasible region given in (5.4), it is expected that the staffing cost will be at least

of the order of $C(\lambda \cdot \vec{e})$, where \vec{e} is a vector of 1's of dimension K . In addition, it is expected that differences between staffing costs of two different staffing vectors, which are close to the efficient frontier of the feasible region, will be of the order of $C(\sqrt{\lambda} \cdot \vec{e})$. Hence, in order to establish a meaningful form of asymptotic optimality, one is led to comparing *normalized* staffing costs that measure the difference between the actual staffing costs and a basic cost of order $C(\lambda \cdot \vec{e})$.

To get this basic cost, consider the following related problem:

$$\begin{aligned} & \text{minimize} && C_1(N_1) + C_2(N_2) + \dots + C_K(N_K) \\ & \text{subject to} && \mu_1 N_1 + \mu_2 N_2 + \dots + \mu_K N_K \geq \lambda \\ & && N_1, N_2, \dots, N_K \geq 0, \end{aligned} \tag{5.6}$$

In light of Proposition 5.1, and particularly the relationship (5.4), this problem can be thought of as the *fluid scale* staffing problem, whose solution is, therefore, a natural centering factor for the asymptotic optimality criterion that we present below. This problem, if accompanied by integral constraints, is a special case of the set covering problem. Without the integral constraints, its optimal solution, \vec{N}^* , is uniquely determined by

$$\frac{C'_k(N_k^*)}{\mu_k} = \frac{C'_j(N_j^*)}{\mu_j}, \quad j, k = 1, 2, \dots, K, \tag{5.7}$$

and

$$\mu_1 N_1 + \mu_2 N_2 + \dots + \mu_K N_K = \lambda. \tag{5.8}$$

Let \underline{C}^λ be the optimal cost obtained by solving (5.6). Then, \underline{C}^λ will serve as the normalizing factor.

Definition (Asymptotically Optimal Staffing): Suppose that $\{\vec{N}^{*\lambda}\}_{\lambda>0}$ is a sequence of optimal solutions of the sequence of *scaled* staffing problems:

$$\begin{aligned} & \text{minimize} && C_1(N_1^\lambda) + C_2(N_2^\lambda) + \dots + C_K(N_K^\lambda) \\ & \text{subject to} && P_{\pi^\lambda}^\lambda(ab) \leq \Delta/\sqrt{\lambda}, \quad 0 < \Delta < \infty, \quad \text{for some } \pi^\lambda = \pi(\lambda, \vec{N}^{\lambda}) \in \Pi, \\ & && N_1^\lambda, N_2^\lambda, \dots, N_K^\lambda \in \mathbb{Z}_+, \end{aligned} \tag{5.9}$$

with respect to sequences of arrival rates $\{\lambda\}$ and staffing cost functions $\{C_1(\cdot), \dots, C_K(\cdot)\}$. Let $\{\tilde{N}^\lambda\}_{\lambda>0}$ be another sequence of staffing vectors. Then, $\{\tilde{N}^\lambda\}_{\lambda>0}$ is an *asymptotically optimal staffing sequence* if when used to staff the system,

- a. There exists a sequence of policies $\{\pi^\lambda = \pi^\lambda(\lambda, \tilde{N}^\lambda)\} \subseteq \Pi$ such that $\limsup_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\pi^\lambda}^\lambda(ab) \leq \Delta$, and
- b. $\lim_{\lambda \rightarrow \infty} \left(C(\vec{N}^{*\lambda}) - \underline{C}^\lambda \right) / \left(C(\tilde{N}^\lambda) - \underline{C}^\lambda \right) = 1$.

We now investigate homogeneous cost functions of the form $C^\lambda(\vec{N}) \equiv C(\vec{N}) = c_1 N_1^p + \dots + c_K N_K^p$, where $1 < p < \infty$, and $c_k > 0$ for $k = 1, \dots, K$. Let $-\infty < \delta < \infty$ be such

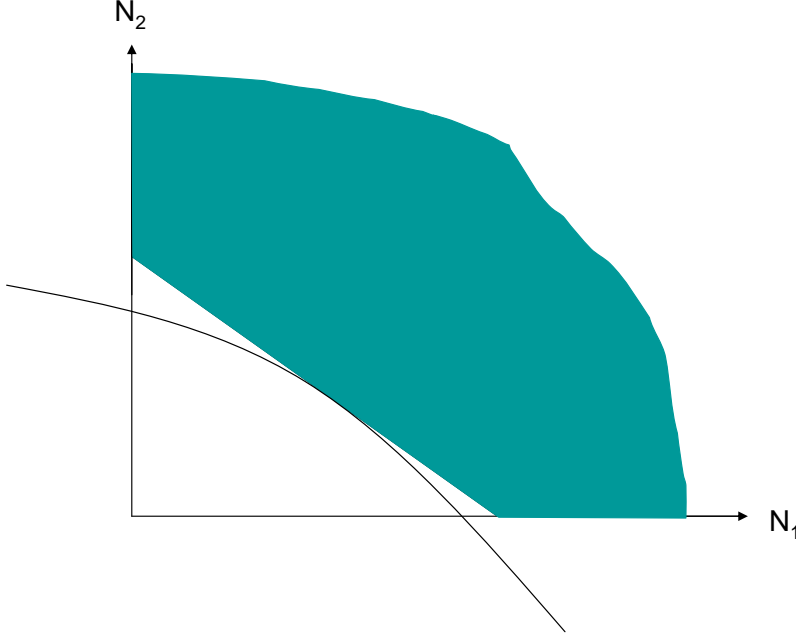


Figure 5.2: Asymptotic Cost Optimization for $K = 2$.

that $\Delta = \Delta(\delta, \mu_1, \theta)$, and let $\vec{M}^{*\lambda}$ be an optimal solution of the problem (5.6) with the right hand side λ replaced by $\lambda + \delta\sqrt{\lambda}$. Note that the vector $\vec{M}^{*\lambda}$ is not necessarily all integers, and let $\tilde{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil := (\lceil M_1^{*\lambda} \rceil, \dots, \lceil M_K^{*\lambda} \rceil)$, that is \tilde{N}^λ is obtained from $\vec{M}^{*\lambda}$ by rounding off its elements to the closest integers from above. We claim that \tilde{N}^λ is an asymptotically optimal staffing vector. Solving for $\vec{M}^{*\lambda}$ is illustrated in Figure 5.2.

Before stating and proving the asymptotic optimality of our proposed solution we state and prove a proposition that shows that the difference in cost between the optimal solution and our proposed solution cannot be too large.

Proposition 5.2 *Consider a fixed target scaled abandonment probability of $\Delta \in (0, \infty)$. Suppose that $C(\vec{N}) = c_1 N_1^p + \dots + c_K N_K^p$, and for $\lambda > 0$ consider the staffing vector $\tilde{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil$, where $\vec{M}^{*\lambda} = (M_1^{*\lambda}, \dots, M_K^{*\lambda})$ is an optimal solution to (5.6), with the right hand side λ replaced by $\lambda + \delta\sqrt{\lambda}$. Here δ satisfies $\Delta = \Delta(\delta, \mu_1, \theta)$ (see (5.5)), and*

$$\vec{M}^{*\lambda} = (\lambda + \delta\sqrt{\lambda}) \frac{((\mu_1/c_1)^{1/(p-1)}, (\mu_2/c_2)^{1/(p-1)}, \dots, (\mu_K/c_K)^{1/(p-1)})}{\sum_{k=1}^K (\mu_k^p/c_k)^{1/(p-1)}}, \quad \lambda > 0. \quad (5.10)$$

Let $\vec{N}^{*\lambda}$ be a sequence of optimal solutions to (5.9). Then

$$\lim_{\lambda \rightarrow \infty} \frac{|C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda})|}{\lambda^{p-1/2}} = 0, \quad (5.11)$$

which also implies that

$$\lim_{\lambda \rightarrow \infty} \frac{|C(\vec{N}^{*\lambda}) - C(\vec{N}^\lambda)|}{\lambda^{p-1/2}} = 0, \quad (5.12)$$

Proposition 5.3 (Asymptotically optimal staffing) Consider a fixed target scaled abandonment probability of $\Delta \in (0, \infty)$. Suppose that $C(\vec{N}) = c_1 N_1^p + \dots + c_K N_K^p$, and for $\lambda > 0$ consider the staffing vector $\vec{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil$, where $\vec{M}^{*\lambda} = (M_1^{*\lambda}, \dots, M_K^{*\lambda})$ is an optimal solution to (5.6), with the right hand side λ replaced by $\lambda + \delta\sqrt{\lambda}$. Here δ satisfies $\Delta = \Delta(\delta, \mu_1, \theta)$ (see (5.5)), and $\vec{M}^{*\lambda}$ is as given in (5.10). Then, if $\delta \neq 0$, $\{\vec{N}^\lambda\}_{\lambda > 0}$ is an asymptotically optimal staffing sequence with respect to (5.9).

Remark 5.1 Proposition 5.3 excludes the case $\delta = 0$. When $\delta = 0$ the difference between $C(\vec{N}^{*\lambda})$ and C^λ might be too small, so asymptotic optimality according to the original definition might not hold. Alternatively, if one defines asymptotic optimality in terms of Proposition 5.2, then asymptotic optimality extends to the case $\delta = 0$.

Example 5.1 Quadratic Cost Functions: Consider the staffing problem:

$$\begin{aligned} & \text{minimize} && c_1 N_1^2 + c_2 N_2^2 + \dots + c_K N_K^2 \\ & \text{subject to} && P_\pi(ab) \leq \Delta/\lambda, \text{ for some } \pi = \pi(\lambda, \vec{N}) \in \Pi, \\ & && N_1, N_2, \dots, N_K \in \mathbb{Z}_+, \end{aligned} \quad (5.13)$$

with a fixed $0 < \Delta < \infty$. To emphasize the dependence of the staffing level on the arrival rate λ , we denote our proposed solution by \vec{N}^λ . To determine the total capacity needed to satisfy the abandonment probability constraint, Proposition 5.1 suggests that

$$\mu_1 N_1^\lambda + \mu_2 N_2^\lambda + \dots + \mu_K N_K^\lambda \geq \lambda + \delta\sqrt{\lambda} + o(\lambda), \text{ as } \lambda \rightarrow \infty,$$

where δ satisfies (5.5). That is, the total capacity required to achieve the target abandonment probability depends asymptotically on the service rates through the service rate μ_1 of the slow servers only. To determine the actual staffing level, one needs to take into account the actual individual service rates. By Proposition 5.3, the proposed staffing vector \vec{N}^λ which satisfies

$$\frac{N_k^\lambda}{N_j^\lambda} = \frac{c_j/\mu_j}{c_k/\mu_k}, \quad k, j = 1, 2, \dots, K, \quad (5.14)$$

and

$$\mu_1 N_1^\lambda + \mu_2 N_2^\lambda + \dots + \mu_K N_K^\lambda = \lambda + \delta\sqrt{\lambda},$$

is asymptotically optimal among all asymptotically feasible vectors. The verbal interpretation of (5.14) is that when the staffing cost is quadratic, then staffing levels for individual server pools are inversely proportional to the ratio c_k/μ_k . This rule is intuitive as it implies that when the cost per unit of service rate is high, the staffing level should be low. Note that the ratio c/μ is not to be confused with the quantity $c\mu$ often used (in different contexts) to determine routing rules when holding costs is associated with waiting customers.

5.3 Extensions

Homogeneous Cost Functions that are Arrival Rate Dependent: Suppose that, instead of the fixed staffing cost function considered in Proposition 5.3, the cost function is dependent on the arrival rate. We capture this dependence through the superscript λ . Particularly, consider, for $\lambda > 0$, the staffing cost function is $C^\lambda(\vec{N}) = c_1^\lambda N_1^{p^\lambda} + c_2^\lambda N_2^{p^\lambda} + \dots + c_K^\lambda N_K^{p^\lambda}$. For $\lambda > 0$ and $k = 1, \dots, K$, assume that $c_k^\lambda > 0$, $\liminf_{\lambda \rightarrow \infty} c_k^\lambda > 0$, $p^\lambda > 1$, $\liminf_{\lambda \rightarrow \infty} p^\lambda > 1$, and $\limsup_{\lambda \rightarrow \infty} p^\lambda < \infty$.

In this case, one can verify that the sequence of staffing vectors $\lceil \vec{M}^{*\lambda} \rceil$ proposed in (5.10) - with superscripts λ accompanying c_k and p , is an asymptotically optimal staffing. Specifically, let

$$\vec{M}^{*\lambda} = (\lambda + \delta\sqrt{\lambda}) \frac{\left((\mu_1/c_1^\lambda)^{1/(p^\lambda-1)}, (\mu_2/c_2^\lambda)^{1/(p^\lambda-1)}, \dots, (\mu_K/c_K^\lambda)^{1/(p^\lambda-1)} \right)}{\sum_{k=1}^K \left(\mu_k^{p^\lambda}/c_k^\lambda \right)^{1/(p^\lambda-1)}}, \quad \lambda \geq 1, \quad (5.15)$$

then one can show that $\lceil \vec{M}^{*\lambda} \rceil$ is asymptotically optimal.

Linear Cost Functions with Constraints: In many practical situations, one is interested in determining staffing levels to minimize linear staffing costs. This is the case where the staffing costs are associated, for example, with salaries of the servers. However, the linear cost case has not been included in our discussion so far. To illustrate why this case is problematic within our framework, consider the following example: Suppose that one is interested in solving the staffing problem

$$\begin{aligned} & \text{minimize} && c_1 N_1 + c_2 N_2 + \dots + c_K N_K \\ & \text{subject to} && P_\pi(ab) \leq \Delta/\lambda, \text{ for some } \pi = \pi(\lambda, \vec{N}) \in \Pi, \\ & && N_1, N_2, \dots, N_K \in \mathbb{Z}_+, \end{aligned} \quad (5.16)$$

for some fixed value of $0 < \Delta < \infty$. If one, instead, solved the deterministic problem:

$$\begin{aligned} & \text{minimize} && c_1 N_1 + c_2 N_2 + \dots + c_K N_K \\ & \text{subject to} && \mu_1 N_1 + \mu_2 N_2 + \dots + \mu_K N_K \geq \lambda + \delta\sqrt{\lambda}, \\ & && N_1, N_2, \dots, N_K \geq 0, \end{aligned} \quad (5.17)$$

then any optimal solution \vec{N}^λ would satisfy:

$$N_k^\lambda > 0 \text{ only if } \frac{c_k}{\mu_k} = \min_{j=1, \dots, K} \left\{ \frac{c_j}{\mu_j} \right\}, \quad k = 1, 2, \dots, K.$$

In particular, if $\frac{c_1}{\mu_1} \neq \min_j \left\{ \frac{c_j}{\mu_j} \right\}$, then $N_1^\lambda = 0$ for all λ . The problem in this case is that one is no longer guaranteed that the proposed staffing vector is asymptotically *feasible* because condition (5.2) of Proposition 5.1 is not satisfied. In fact, it can be shown that when $N_1^\lambda = 0$ a higher overall capacity level is needed in order to obtain the same limiting abandonment probability.

Note that if $\frac{c_1}{\mu_1} = \min_j \left\{ \frac{c_j}{\mu_j} \right\}$, then one can choose N_1^λ to be non-negligible relatively to the other server pools, and then the proposed solution is indeed asymptotically optimal (the proof follows through similarly to the proof of Proposition 5.3). However, even in the case that $\frac{c_1}{\mu_1} \neq \min_j \left\{ \frac{c_j}{\mu_j} \right\}$, there are scenarios where our theory can provide useful solutions. Consider the following staffing problem with linear staffing costs and additional linear constraints:

$$\begin{aligned}
& \text{minimize} && c_1 N_1 + c_2 N_2 + \dots + c_K N_K \\
& \text{subject to} && P_\pi(ab) \leq \Delta/\lambda, \text{ for some } \pi = \pi(\lambda, \vec{N}) \in \Pi, \\
& && A\vec{N} \geq b \\
& , && N_1, N_2, \dots, N_K \in \mathbb{Z}_+,
\end{aligned} \tag{5.18}$$

where A is a $i \times K$ matrix, and b is an i -dimensional matrix for some $i \geq 1$. Examples for such additional constraints can include $N_1/(N_1 + \dots + N_K) \geq p$ for some $0 < p < 1$, or $l_k \leq N_k/(N_1 + \dots + N_K) \leq u_k$ for some $0 \leq l_k \leq u_k \leq 1$. The first example can result from a case where servers of pool 1 are trainees, and one wants to make sure that they get the experience they need. The second set of constraints can be the result of given proportions of servers skills in the particular server population.

For the problem (5.18), we claim that if the set of problems

$$\begin{aligned}
& \text{minimize} && c_1 N_1 + c_2 N_2 + \dots + c_K N_K \\
& \text{subject to} && \mu_1 N_1 + \mu_2 N_2 + \dots + \mu_K N_K \geq \lambda + \delta\sqrt{\lambda}, \\
& && A\vec{N} \geq b \\
& , && N_1, N_2, \dots, N_K \in \mathbb{Z}_+,
\end{aligned} \tag{5.19}$$

has a sequence of solutions \vec{N}^λ which satisfy $\liminf_{\lambda \rightarrow \infty} N_1^\lambda/(N_1^\lambda + \dots + N_K^\lambda) > 0$, then the proposed sequence is an asymptotically optimal staffing. The proof follows similarly to the proof of Proposition 5.3.

The Model without Abandonment: If customers are infinitely patient, [2] shows that FSF is asymptotically optimal in minimizing the steady-state delay probability, $P(\text{wait})$. A natural analogous staffing problem to (2.1) in that case is:

$$\begin{aligned}
& \text{minimize} && C_1(N_1) + C_2(N_2) + \dots + C_K(N_K) \\
& \text{subject to} && P_\pi(\text{wait}) \leq \alpha, \text{ for some } \pi = \pi(\lambda, \vec{N}) \in \Pi, \\
& && N_1, N_2, \dots, N_K \in \mathbb{Z}_+,
\end{aligned} \tag{5.20}$$

where $0 < \alpha < 1$. For the problem (5.20) we can establish an analogous asymptotically feasible region, as given in the following proposition:

Proposition 5.4 (Asymptotically Feasible Region: The Case without Abandonment) Let $0 < \alpha < 1$ and $0 < \mu_1 < \mu_2 < \dots < \mu_K$ be fixed. Consider a sequence of systems indexed by the arrival rate $\lambda > 0$, with λ growing to infinity, and N_k^λ servers in pool k , $k = 1, \dots, K$. Let $N^\lambda = \sum_{k=1}^K N_k^\lambda$ be the total number of servers in system λ , and suppose that

$$\liminf_{\lambda \rightarrow \infty} \frac{N_1^\lambda}{N^\lambda} > 0. \quad (5.21)$$

Then, there exists a sequence $\{\pi^\lambda = \pi^\lambda(\lambda, \vec{N}^\lambda)\}$ of non-preemptive policies, under which

$$\limsup_{\lambda \rightarrow \infty} P_{\pi^\lambda}^\lambda(\text{wait}) = \alpha \quad (5.22)$$

if and only if

$$\mu_1 N_1^\lambda + \dots + \mu_K N_K^\lambda \geq \lambda + \delta \sqrt{\lambda} + o(\sqrt{\lambda}), \quad \text{as } \lambda \rightarrow \infty, \quad (5.23)$$

$$\alpha \triangleq \alpha(\delta/\sqrt{\mu_1}) = \left[1 + \frac{(\delta/\sqrt{\mu_1}) \Phi(\delta/\sqrt{\mu_1})}{\phi(\delta/\sqrt{\mu_1})} \right]^{-1}. \quad (5.24)$$

In addition, $\delta = 0$ (i.e. (5.23) is violated for all $\delta > 0$) if and only if $\alpha = 1$, and $\delta = \infty$ (i.e. (5.23) holds for all $\delta < \infty$) if and only if (5.22) holds for any arbitrary $1 > \alpha > 0$, with the appropriate choice of π^λ .

Proposition 5.4 identifies the asymptotically feasible region with respect to (5.20). Similarly, one could define asymptotic optimal staffing in this case and establish that the square-root safety-staffing rule is indeed asymptotically optimal. The definition here is more straightforward than in the case with abandonment, because the stability region, as given in the constraint of (5.6), provides a natural lower bound on the staffing cost. Accordingly, asymptotic optimal staffing in this case may be defined as follows:

Definition (Asymptotically Optimal Staffing: The Case without Abandonment) Suppose that $\{\vec{N}^{*\lambda}\}_{\lambda > 0}$ is a sequence of optimal solutions of the sequence of staffing problems:

$$\begin{aligned} & \text{minimize} && C_1(N_1^\lambda) + C_2(N_2^\lambda) + \dots + C_K(N_K^\lambda) \\ & \text{subject to} && P_{\pi^\lambda}(\text{wait}) \leq \alpha, \quad 0 < \alpha < 1, \quad \text{for some } \pi^\lambda = \pi(\lambda, \vec{N}^\lambda) \in \Pi, \\ & && N_1^\lambda, N_2^\lambda, \dots, N_K^\lambda \in \mathbb{Z}_+, \end{aligned} \quad (5.25)$$

with respect to sequences of arrival rates $\{\lambda\}$ and staffing cost functions $\{C_1(\cdot), \dots, C_K(\cdot)\}$. Let $\{\tilde{N}^\lambda\}_{\lambda > 0}$ be another sequence of staffing vectors. Then, $\{\tilde{N}^\lambda\}_{\lambda > 0}$ is an *asymptotically optimal staffing sequence* if when used to staff the system,

- a. There exists a sequence of policies $\{\pi^\lambda = \pi^\lambda(\lambda, \tilde{N}^\lambda)\} \subseteq \Pi$ such that $\limsup_{\lambda \rightarrow \infty} P_{\pi^\lambda}(\text{wait}) \leq \alpha$, and

- b. $\lim_{\lambda \rightarrow \infty} \left(C(\vec{N}^{*\lambda}) - \underline{C}^\lambda \right) / \left(C(\tilde{N}^\lambda) - \underline{C}^\lambda \right) = 1$, where \underline{C}^λ is the optimal value of the objective function of (5.6).

With this definition in mind, one could prove asymptotic optimality of the square-root safety-staffing rule, analogously to Proposition 5.3.

Proposition 5.5 (Asymptotically Optimal Staffing: The Case without Abandonment) *Consider a fixed target delay probability of $\alpha \in (0, 1)$. Suppose that $C(\vec{N}) = c_1 N_1^p + \dots + c_K N_K^p$, and for $\lambda > 0$ consider the staffing vector $\tilde{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil$, where $\vec{M}^{*\lambda} = (M_1^{*\lambda}, \dots, M_K^{*\lambda})$ is an optimal solution to (5.6), with the right hand side λ replaced by $\lambda + \delta\sqrt{\lambda}$. Here $\delta > 0$ satisfies $\alpha = \alpha(\delta/\mu_1)$ (see (5.24)), and $\vec{M}^{*\lambda}$ is as given in (5.10). Then, $\left\{ \tilde{N}^\lambda \right\}_{\lambda > 0}$ is an asymptotically optimal staffing sequence with respect to (5.25).*

6 Conclusions

We have studied the joint staffing and routing problem with respect to the inverted-V system with abandonment. Recognizing that abandonment introduces new challenges, we come up with a robust problem formulation that seeks to minimize staffing costs subject to an upper bound on the abandonment probability. With respect to this problem formulation, we show that it is asymptotically optimal to use work-conserving FCFS policies. Specifically, we show that the Faster-Server-First (FSF) policy is asymptotically optimal in the sense that it asymptotically minimized the steady-state fraction of customer abandonment in the QED regime. We then proceed to define the asymptotically feasible region which consists of staffing vectors under which a target abandonment probability can be obtained. This region is shown to have a simple linear boundary, that is defined by the total service capacity exceeding the arrival rate plus a square-root term. Finally, we show that minimizing the staffing costs over this asymptotically feasible region results in an asymptotically optimal staffing rule.

This research can be extended in several ways. First, one might consider models with more general service time, interarrival time, or time-to-abandon distributions, and / or with more general staffing costs. Second, it is relevant to explore the same model in the Efficiency Driven (ED) regime, where the abandonment probability does not converge to 0. Finally, Armony and Ward in [3] consider the issue of fairness among the servers in the inverted-V model without abandonment. This paper's results may be used to study fairness in the model with impatient customers.

References

- [1] Aksin, Z., Armony, M., Mehrotra, V. (2007) The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research, *Production and Operations Management, Special issue on Service Operations in honor of John Buzacott* (ed. G. Shanthikumar and D. Yao), 16 (6), 665-688. [1](#)
- [2] Armony, M. (2005), Dynamic routing in large-scale service systems with heterogeneous servers, *Queueing Systems*, **51**(3-4), pp. 287-329. [1](#), [1](#), [6](#), [1.2](#), [3.1](#), [4.2.2](#), [5.3](#)
- [3] Armony, M. and Ward, A. (2008), Fair dynamic routing in large-scale heterogeneous-server systems. Working paper. [1.2](#), [6](#)
- [4] Atar, R. (2007), Central limit theorem for a many-server queue with random service rates, *Annals of Applied Probability*, forthcoming. [1.2](#)
- [5] Atar, R., Mandelbaum, A. and Shaikhet, G. (2008), Simplified control problems for multi-class many-server queueing systems. Working paper. [1.2](#)
- [6] Atar, R. and Shwartz, A. (2007), Efficient routing in heavy traffic under partial sampling of service times. Working paper. [1.2](#)
- [7] Bassamboo, A. J.M. Harrison and A. Zeevi (2006), Design and control of a large call center: Asymptotic analysis of an LP-based method *Operations Research* **54**, pp. 419-435. [1.2](#)
- [8] Bassamboo, A. J.M. Harrison and A. Zeevi (2006), Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems: Theory and Applications*, **51**, pp. 249-285. [1.2](#)
- [9] Borst, S., Mandelbaum, A. and Reiman, M. (2003), Dimensioning large call centers, *Operations Research*, **52**(1), pp. 17-34. [1](#), [1.2](#), [4.1](#)
- [10] Cabral F.B. (2005), The slow server problem for uninformed customers, *Queueing systems*, **50**(4), pp. 353-370. [1.2](#)
- [11] J. G. Dai and T. Tezcan (2005), State space collapse in many-server diffusion limits of parallel server systems. Working paper. [4.1](#)
- [12] J. G. Dai and T. Tezcan (2007) Optimal Control of Parallel Server Systems with Many Servers in Heavy Traffic. Working paper. [1](#), [1.2](#)
- [13] de Véricourt, F. and Zhou, Y.-P. (2005), A routing problem for call centers with customer callbacks after service failure, *Operations Research*, **53**(6). [1.2](#)
- [14] de Véricourt, F. and Zhou, Y.-P. (2006), On the Incomplete Results for the Heterogeneous Server Problem, *Queueing Systems* **52**(3). [1.1](#), [1.2](#)
- [15] 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. [1.2](#)

- [16] Gans, N., Koole, G., Mandelbaum, A. 2003. Telephone Call Centers: Tutorial, Review and Research Prospects. Invited review paper by *Manufacturing and Service Operations Management* **5**(2), pp. 79-141. [1](#)
- [17] Gans, N. and Shen, H. (2007), Service time heterogeneity in an inbound call center, In Preparation. [1.2](#)
- [18] Garnett, O., Mandelbaum, A. and Reiman, M. (2002), Designing a call center with impatient customers, *Manufacturing & Service Operations Management*, **4**(3), pp. 208-227. [1](#), [1.1](#), [4.2.2](#), [4.2](#)
- [19] Gurvich I., Armony M. and Mandelbaum A. (2006), Service level differentiation in call centers with fully flexible servers. *Management Science, Special Issue: on Call Center Management* (ed. G. Koole), **54**(2), 279-294. [1.2](#)
- [20] Gurvich, I. and Whitt, W. (2007) Queue-and-Idleness-Ratio Controls in Many-Server Service Systems. Working paper. [1.2](#)
- [21] Gurvich, I. and Whitt, W. (2007), Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing. Working paper. [1.2](#)
- [22] Gurvich, I. and Whitt, W. (2007), Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing and Service Operations Management*, forthcoming. [1.2](#)
- [23] Halfin, S. and Whitt, W. (1981), Heavy-traffic limits for queues with many exponential servers, *Operations Research*, **29**(3), pp. 567–588. [1](#), [4.1](#)
- [24] J.M. Harrison and A. Zeevi (2005), A method for staffing large call centers using stochastic fluid models. *Manufacturing & Service Operations Management*, **7**, pp. 20–36. [1.2](#)
- [25] 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. [1.2](#)
- [26] 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. [1.2](#), [4.2](#)
- [27] Mandelbaum, A. and Zeltyn, S. (2007), Staffing many-server queues with impatient customers: constraint satisfaction in call centers. Working paper. [1](#), [1.1](#), [1.2](#), [5.1](#)
- [28] Rubinovich M. (1983), The slow server problem, *Journal of Applied Probability* **22**, pp. 205-213. [1.2](#)
- [29] Stockbridge R.H. (1991), A Martingale approach to the slow server problem, *Journal of Applied Probability* **28**, pp 480-486. [1.2](#)
- [30] Teh, Y., Ward, A. R. (2002), Critical thresholds for dynamic routing in queueing networks. *Queueing Systems*, **42**, pp. 297–315. [1.2](#)
- [31] Tezcan, T. (2007), Asymptotically optimal control of many-server heterogeneous service systems with hyper-exponential service times. Working paper. [1.2](#)

- [32] T. Tezcan and J.G. Dai (2006) Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. Working paper. [1](#), [1.2](#)
- [33] Tseytlin, Y. (2007), Queueing Systems with Heterogeneous Servers: Improving Patients' Flow in Hospitals. Technion M.Sc. Research Proposal. [1](#), [1.2](#)

Technical Appendix for:

Routing and Staffing in Large-Scale Service Systems: The Case of Homogeneous Impatient Customers and Heterogeneous Servers

Mor Armony¹

Avishai Mandelbaum²

June 25, 2008

A Proofs

Proof of Proposition 3.1: We prove the Proposition using sample-path coupling arguments. Consider two coupled systems both with the same initial conditions, and the same sequence of arrivals. System 1 operates under an arbitrary policy $\pi \in \Pi_p$ while System 2 operates under FSF_p . For all $t \geq 0$ and $i = 1, 2$, let $Q^i(t)$, $Y^i(t)$, and $Ab^i(t)$ be the queue length at time t , the head-count at this time, and the total number of abandonment up to this time in System i , respectively. We claim that the two systems can be coupled such that the following three properties hold for all $t \geq 0$.

$$Ab^1(t) \geq Ab^2(t), \tag{A.1}$$

$$Q^2(t) - Q^1(t) \leq Ab^1(t) - Ab^2(t), \tag{A.2}$$

and

$$Y^2(t) - Y^1(t) \leq Ab^1(t) - Ab^2(t). \tag{A.3}$$

Establishing property (A.1) will complete the proof of the proposition. Let $0 = t_0 < t_1 < t_2 \dots$ be the union of the sets of transition and action time points in both systems. We prove (A.1)-(A.3) by induction on t_n , $n = 0, 1, 2, \dots$. At time $t_0 = 0$ both systems are assumed to have the same state and therefore properties (A.1)-(A.3) are trivially satisfied. Suppose that these properties are satisfied for all $t \leq t_n$. We need to establish they are also satisfied at $t_j < t \leq t_{n+1}$. Since nothing happens at either system until $t = t_{n+1}$ the only issues is to show that all properties still hold immediately after either a transition or an action at $t = t_{n+1}$. The validation of these properties is due to the specific system coupling which is done as follows:

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

²Industrial Engineering and Management, Technion Institute of Technology, avim@ie.technion.ac.il.

Coupling:

- **Arrivals:** The systems are coupled such that customers arrive into both systems at the same time.

The only property that might be violated due to this kind of transition is (A.2), if $Q^2(t_n) - Q^1(t_n) \leq Ab^1(t_n) - Ab^2(t_n)$, and if the arrival into system 1 joins the service, but the arrival into system 2 has to wait in line. But this would mean that all the servers are busy in system 2 and therefore $Y^2(t_n) - Y^1(t_n) \geq Q^2(t_n) - N - (Q^1(t_n) - N) = Q^2(t_n) - Q^1(t_n) = Ab^1(t_n) - Ab^2(t_n)$ which violates (A.3).

- **Service completions:** A service completion in system i ($i = 1, 2$) implies a service completion in system j ($j = 1, 2, j \neq i$) with probability $\min \left\{ 1, \frac{\sum_{k=1}^K Z_k^j(t_n) \mu_k}{\sum_{k=1}^K Z_k^i(t_n) \mu_k} \right\}$, where $Z_k^i(t_n), Z_k^j(t_n)$ are the number of busy servers of pool k ($k = 1, \dots, K$) in systems i and j , respectively, at time t_n .

This kind of transition might violate properties (A.2) and / or (A.3). The latter might be violated if $Y^2(t_n) - Y^1(t_n) = Ab^1(t_n) - Ab^2(t_n)$ and the service completion occurs in system 1 only. But, due to (A.1), $Y^2(t_n) \geq Y^1(t_n)$. In particular, due to the non work conserving nature of FSF_p , there are more busy servers in system 2 than in system 1. Now, due to the fast server first property of FSF_p this also implies that $\sum_{k=1}^K Z_k^2(t_n) \mu_k \geq \sum_{k=1}^K Z_k^1(t_n) \mu_k$ and therefore this service completion in system 1 also implies a service completion in system 2. To show that (A.2) is also not violated by such a transition, suppose that $Q^2(t_n) - Q^1(t_n) = Ab^1(t_n) - Ab^2(t_n)$ and that the service completion reduces Q^1 only. This can happen only if (a) $Q^2(t_n) = 0$ or (b) $Q^2(t_n) > 0$ and the number of busy servers in system 1 is larger than in system 2. But in case (a) if $Q^2(t_n) = 0$, then since $Q^2(t_n) - Q^1(t_n) = Ab^1(t_n) - Ab^2(t_n) \geq 0$, $Q^1(t_n) = 0$. Also, in case (b), if $Q^2(t_n) > 0$ then due to work-conservation all server are busy and, hence, a service completion in system 1 will imply a service completion in system 2.

- **Abandonment:** An abandonment from system i ($i = 1, 2$) implies an abandonment in system j ($j = 1, 2, j \neq i$) with probability $\min \left\{ 1, \frac{Q^j(t_n)}{Q^i(t_n)} \right\}$.

This transition might violate any of the three properties if there is an abandonment out of system 2 which is not coupled with an abandonment in system 1. But this may only happen if $Q^2(t_n) > Q^1(t_n)$. In particular, this implies that $Ab^1(t_n) > Ab^2(t_n)$ (due to (A.2)). Therefore (A.1) will not be violated. Similarly, such a transition will result in a unit decrease

of both the left-hand-side and the right-hand-side of both (A.2) and (A.3) and, therefore, these two properties still hold.

- **Policy Actions:** Policy actions include the assignment of a customer to a particular server, and a hand-off of a customer from one server to another. If these actions occur in one system they do not affect the other system. It is easily verified that these actions will not cause the violation of any of the three properties. ■

Proof of Corollary 3.1: Notice that, in steady-state, the following balance equation holds for any policy $\pi \in \Pi_p$ (see also equation (2) in [4]):

$$\theta \cdot E[Q(\infty; \pi)] = \lambda \cdot P_\pi(ab). \quad (\text{A.4})$$

The left-hand-side corresponds to the rate of abandonment from the system, and the right-hand-side describes the rate of arrival of customers who will eventually abandon. From Little's law and (A.4) we also obtain a relationship between the expected waiting time and probability of abandonment in steady-state:

$$\theta \cdot E[W(\infty; \pi)] = P_\pi(ab). \quad (\text{A.5})$$

Proposition 3.1 together with the relationships (A.4) and (A.5) completes the proof. ■

Proof of Proposition 3.2: We prove this proposition using sample-path coupling arguments to show that at all times $t \geq 0$ we have:

$$Y(t; \pi) \geq Y(t; \text{FSF}_p). \quad (\text{A.6})$$

The coupling is done the same as the coupling in the proof of Proposition 3.1, observing that, due to work-conservation, if $Q(t; \pi) > Q(t; \text{FSF}_p)$, then necessarily (A.6) also holds. ■

Proof of Corollary 3.2: The proof follows from the relationships $Q(t; \pi) = [Y(t; \pi) - N]^+$ and (2.2) which hold for all $\pi \in \Pi_p$ that are also work conserving. ■

Proof of Proposition 3.3: Suppose that $\theta > \sum_{k=1}^K N_k \mu_k$ and consider the following two systems: System 2 operates under FSF_p and System 1 operates under the policy $\pi \in \Pi_p$, described as follows:

1. π devotes the fastest servers available to the jobs which are in service (in this sense it mimics FSF_p). Formally, this means that if $Z_l(t; \pi) > 0$ for some $l \in \{1, \dots, K\}$, then $Z_k(t; \pi) = N_k$ for all $k > l$.

2. If a job arrives when either all servers are busy or there are other jobs waiting in the queue, it is left in the queue until it abandons the system. In particular, π is *not* work conserving because it may idle servers even when there are jobs in the queue.

We claim that $P_\pi(\text{wait} > 0) < P_{\text{FSF}_p}(\text{wait} > 0)$. We use sample-path coupling arguments to prove that the two systems can be coupled such that the following properties hold, for all $t \geq 0$ (The superscripts 1 and 2 here correspond to systems 1 and 2, respectively):

$$Y^2(t) \geq Y^1(t), \tag{A.7}$$

$$Z_k^2(t) \geq Z_k^1(t), \quad k = 1, \dots, K, \tag{A.8}$$

$$Q^2(t) \geq Q^1(t) - 1, \tag{A.9}$$

and

$$\text{if } Z_1^2(t) < N_1, \text{ then } Q^1(t) = 0. \tag{A.10}$$

Properties (A.8) and (A.10) together imply that $P_\pi(\text{wait} > 0) \leq P_{\text{FSF}_p}(\text{wait} > 0)$. The strict inequality

$$P_\pi(\text{wait} > 0) < P_{\text{FSF}_p}(\text{wait} > 0) \tag{A.11}$$

will be argued in the course of the proof.

Let $0 = t_0 < t_1 < t_2 \dots$ be the sequence of transition times in either system (defined inductively). We will prove that (A.7)-(A.10) hold for all $0 \leq t \leq t_n$ by induction on n , supposing that those properties hold at time $t_0 = 0$. Suppose that they hold for all $t \leq t_n$. We show that this implies that the four properties also hold at $t_n < t \leq t_{n+1}$. Since no transitions occur at $t_n < t < t_{n+1}$ the only time point in question is the transition point t_{n+1} . To verify the induction we need to describe the coupling between these two systems.

Coupling:

- **Arrivals:** Arrivals into both system occur at the same time. The only potential problem with such a transition is that the arrival joins the queue in system 1 but joins a server in system 2. But this is only possible if there are available servers in system 2 (in particular, $Z_1^2(t_n) < N_1$), and $Q^1(t_n) > 0$. But this violates property (A.10) and therefore is not possible.

Case	System 1 Departure	System 2 Departure	Transition rate
1) $Q^1 = 0$	a service completion	a service completion	$\sum_{k=1}^K Z_k^1 \mu_k$
	–	a service completion	$\sum_{k=1}^K (Z_k^2 - Z_k^1) \mu_k$
	–	an abandonment	θQ^2
2) $Z_1^1 = N_1$	a service completion	a service completion	$\sum_{k=1}^K N_k \mu_k$
	an abandonment	an abandonment	θQ^1
	–	an abandonment	$\theta(Q^2 - Q^1)$
3) $Q^2 = Q^1 - 1$ and $Z_1^1 < N_1$	1 job abandonment	a service completion	$\sum_{k=1}^K N_k \mu_k$
	1 abandonment out of $Q^1 - 1$ jobs	1 abandonment out of Q^2 jobs	θQ^2
	a service completion	–	$\sum_{k=1}^K Z_k^1 \mu_k$
	1 job abandonment	–	$\theta - \sum_{k=1}^K N_k \mu_k$
4) $Q^2 \geq Q^1 > 0$ and $Z_1^1 < N_1$	1 job abandonment	a service completion	$\sum_{k=1}^K N_k \mu_k$
	1 abandonment out of $Q^1 - 1$ jobs	1 abandonment out of $Q^1 - 1$ jobs	$\theta(Q^1 - 1)$
	a service completion	1 job abandonment	$\sum_{k=1}^K Z_k^1 \mu_k$
	1 job abandonment	1 job abandonment	$\theta - \sum_{k=1}^K N_k \mu_k$
	–	1 abandonment out of $Q^2 - (Q^1 - 1)$ jobs	$\theta(Q^2 - Q^1)$ $+ (\sum_{k=1}^K N_k \mu_k - \sum_{k=1}^K Z_k^1 \mu_k)$

Table 1: Coupling of departures from both systems (see proof of Proposition 3.3)

- **Departures:** This case requires a more careful consideration. Specifically, we identify four different sub-cases, each requires a different type of coupling between the various transitions. Table A summarizes those transition rates and coupling in this case³.

It is left to verify that properties (A.7)-(A.10) are satisfied after a transition which involves a job departure. To verify (A.7) note that the only cases where there is a departure from system 2 which is not accompanied by a departure out of system 1 is when $Y^2(t_n) > Y^1(t_n)$. To verify (A.8) we note that whenever there is a service completion from system 2 alone, we have $\sum_{k=1}^K Z_k^2(t_n) > \sum_{k=1}^K Z_k^1(t_n)$. Since both policies use the fastest servers first, this property remains true at time t_{n+1} . For property (A.9) one needs to verify two facts: a) that an abandonment from system 2 only occurs only when $Q^2(t_n) > Q^1(t_n) - 1$, and b) that when a service completion from system 2 is not accompanied by an abandonment from

³All relevant quantities in the table refer to time $t = t_n$, which is omitted for brevity.

system 1 then either $Q^2(t^n) > Q^1(t_n) - 1$ or $Q^2(t_n) = 0$. Both facts are indeed true. Finally, to confirm that property (A.10) also holds at time t_{n+1} one can verify that whenever there is a service completion in system 2 and $Q^2(t_n) = 0$ then $Q^1(t_n) = 0$ and therefore, after a (possible) departure, it is still true that $Q^1(t_{n+1}) = 0$. ■

Proof of Proposition 3.4: Due to (A.4) which relates between the abandonment probability and the expected queue length, minimizing $P_\pi(ab)$ is equivalent to minimizing $EQ(\infty; \pi)$. We show that $EQ(\infty; \pi)$ under any policy π which is not necessarily FCFS is equal to $EQ(\infty; \pi')$, where π' is a corresponding FCFS policy. We prove this using a construction of the policy π' and a sample path coupling. Consider the system under a particular sample path ω and the policy π . Construct a policy π' with a sample path ω' as follows: The arrival times under both ω and ω' are the same. Every time the policy π serves a tagged customer which is *not* at the head of the line, the policy π' leaves this customer in line, and instead serves the head-of-line (HOL) customer. The service time of this HOL customer under ω' is set equal to the service time of the tagged customer under ω . Similarly, the time to abandon from that moment on of the tagged customer under ω' is set equal to the time to abandon of the HOL customer under ω . Since the time to abandon distribution is exponential one can couple those two systems and get the same steady-state expected queue length. Also, by construction, π' is a FCFS policy. ■

Proof of Proposition 3.5: The proof is shown for $K = 2$. The general case follows similarly. We first show that $Q_A \leq Q_B$. We will show, using uniformization, coupling and induction, that there are versions of these processes such that

1. $Q_A(n) \leq Q_B(n)$, and
2. $Z_A(n) - Z_B(n) \leq N_1 + N_2 - N_B$,

where n is the discrete time index, and Z_A and Z_B are the total number of busy servers in systems A and B, respectively. Suppose 1. and 2. hold at time n . We show that they will also hold at time $n + 1$. 1. or 2. might be violated at time $n + 1$ if and only if:

- a. $Q_A(n) = Q_B(n) > 0$ and there is a departure in system B but not in system A , or
- b. $Z_A(n) - Z_B(n) = N_1 + N_2 - N_B$, and there is a departure in system B but not in system A .

a. is easy to deal with, because if $Q_A(n) = Q_B(n) > 0$, then due to work conservation, all servers are busy in both systems, and therefore, the total service rates satisfy: $N_1\mu_1 + N_2\mu_2 \geq N_B\mu_2$. Hence, every departure in system B can be coupled with a departure in system A. For b., note that if $Z_B = N_B - m$ then $Z_A = N_1 + N_2 - m$, where $m < N_B$, then the total service rate in system A is minimal when the idle servers are fast. In other words, the total service rate in system A is greater than or equal to

$$\begin{cases} N_1\mu_1 + (N_2 - m)\mu_2, & \text{if } m \leq N_2 \\ (N_1 - (m - N_2))\mu_1, & \text{if } m > N_2 \end{cases} \geq (N_B - m)\mu_2.$$

In particular, the total service rate in system A is greater than or equal to the total service rate in system B.

We next show that $Q_C \leq Q_A$. Similarly to the above we show that for every n the following two properties are satisfied:

1. $Q_C(n) \leq Q_A(n)$, and
2. $Z_C(n) - Z_A(n) \leq N_C - (N_1 + N_2)$.

Again, two scenarios might lead to a violation of 1. or 2. at time $n + 1$ if they are satisfied at time n . These are:

- a. $Q_A(n) = Q_C(n) > 0$ and there is a departure in system A but not in system C, or
- b. $Z_C(n) - Z_A(n) = N_C - (N_1 + N_2)$, and there is a departure in system A but not in system C.

a. is easy to rule out, using work conservation and the fact that $\mu_1 N_1 + \mu_2 N_2 \leq N_C \mu_1$. To show b. notice that if this equality holds at time n and $Z_C = N_C - m$, for some $m \leq N_C$, then we have $Z_A = N_1 + N_2 - m$. In this case, the service rate is maximized if the idle servers are the slow ones. In other words, the total service rate in system A is less than or equal to:

$$\begin{cases} (N_1 - m)\mu_1 + N_2\mu_2, & \text{if } m \leq N_1 \\ (N_2 - (m - N_1))\mu_2, & \text{if } m > N_1 \end{cases} \leq (N_C - m)\mu_1.$$

The proof is then complete, by realizing that the right hand side corresponds to the service rate of system C. ■

Proof of Proposition 4.1: Let $f(t) = |\bar{Y}(t) - 1| = \left| \sum_{k=1}^K (\bar{Z}_k(t) - q_k) + \bar{Q}(t) \right|$, then $f(t) \geq 0$ and $f(t) = 0$ if and only if $\bar{Q}(t) = 0$ and $\bar{Z}_k(t) = q_k$ for all $k = 1, \dots, K$. By Lemma C.1 of [6], and from the fact that $f(\cdot)$ is absolutely continuous, it is sufficient to show that whenever $t \geq 0$ is such that f is differentiable at t , we have $\dot{f}(t) \leq 0$. Suppose that t is such that $\bar{Y}(t) \geq 1$. Then, by (4.13) $\bar{Z}_k(t) = q_k$, for all k . In particular, if f is differentiable at t , then

$$\dot{f}(t) = \dot{\bar{Y}}(t) = \mu - \sum_{k=1}^K \mu_k \bar{Z}_k(t) - \theta \bar{Q}(t) \leq \mu - \sum_{k=1}^K \mu_k q_k = 0.$$

If t is such that $\bar{Y}(t) < 1$, then $\bar{Z}_k(t) < q_k$ for at least one k , and hence, by (4.13), $\bar{Q}(t) = 0$. If f is differentiable at t then,

$$\dot{f}(t) = -\dot{\bar{Y}}(t) = \sum_{k=1}^K \mu_k \bar{Z}_k(t) + \theta \bar{Q}(t) - \mu < \sum_{k=1}^K \mu_k q_k - \mu = 0.$$

■

Proof of Proposition 4.2: The proof follows directly from Theorem 3.1 of [5] (our model satisfies both assumptions C-1 and C-2 of that theorem). ■

Proof of Remark 4.1: The proof follows directly from Theorem 3.1 of [5] (our model satisfies both assumptions C-1 and C-2 of that theorem). ■

Proof of Proposition 4.3: The proof follows directly from Theorem 5.1 of [5]. ■

Proof of Proposition 4.4: The proof follows directly from Theorem 5.1 of [5]. ■

Proof of Proposition 4.5: The proof follows from [2]. Note that the process $X(\cdot)$, restricted to $[0, \infty)$, is an O-U process with infinitesimal drift $-\delta\sqrt{\mu} - \theta x$ and variance 2μ . Hence, according to [2, (18.33)], its steady-state density, conditional on $X(\infty) \geq 0$, is normal with mean $-\delta/\sqrt{\mu}/\theta$ and variance μ/θ , conditioned on having non-negative values only (see [2, (18.28)]). Similarly, the process $X(\cdot)$ restricted to the negative half-line is an O-U process with infinitesimal drift $-\delta\sqrt{\mu} - \mu_1 x$ and variance 2μ . Therefore, its stationary density, conditional on $X(\infty) < 0$, is the density of a normal random variable with mean $-\delta\sqrt{\mu}/\mu_1$, and variance μ/μ_1 , conditioned on having negative values only. Putting these two densities together, establishes that $f(x)$ is indeed the steady-state density of X , with $\alpha = P(X(\infty) \geq 0)$. To find the value of α , note that $f(\cdot)$ is continuous because the infinitesimal variance is continuous on the whole real line (see [2, p. 471]). Hence, α may be solved for by a smooth fit, namely, by equating the limits of $f(\cdot)$ at 0 from both left and right. ■

Proof of Proposition 4.6: The proof is based on Ethier and Kurtz [3, Theorem 9.10 and Remark 9.11, p. 244]. According to [3] and based on our Propositions 4.2 and 4.3, it suffices to show that:

1. There exists a stationary distribution of $\vec{X}^\lambda(\cdot)$ for all λ .
2. The sequence of stationary distributions of $\vec{X}^\lambda(\cdot)$ is tight.

We establish 1. and 2. for $K = 2$. The general case follows similarly.

1. Fix $\lambda > 0$. First note that under FSF_p the total number in the system Y^λ is a Birth and Death process with birth rates $\lambda(y) = \lambda$ and death rates $\mu^\lambda(y)$ as given in (3.1). Due to abandonment, the system is stable for all λ , and the stationary distribution is given by $p_n^\lambda := P(Y^\lambda(\infty) = n) = p_0^\lambda \pi_n^\lambda$, $n = 0, 1, \dots$, where $\pi_n^\lambda = \frac{\lambda^n}{\prod_{i=1}^n \mu^\lambda(i)}$, $n = 0, 1, \dots$, and $p_0^\lambda = [\sum_{n=0}^{\infty} \pi_n^\lambda]^{-1}$. Clearly, the stationary distribution of $X^\lambda = \frac{Y^\lambda - N^\lambda}{\sqrt{N^\lambda}}$, can be easily obtained from the stationary distribution of Y^λ . Finally, since \vec{X}^λ is easily obtained as a one-to-one function of its sum X^λ , the existence of a steady-state distribution for \vec{X}^λ has been established.

To show the existence of a stationary distribution of \vec{X}^λ under the non-preemptive policy FSF one can use the stationarity of the process with respect to FSF_p and the dominance of FSF_p over FSF which was established in Proposition 3.2, noting that FSF is work-conserving. The details are omitted as the proof is identical to the proof of part 1. of Proposition 4.6 in [1].

2. Tightness of $\vec{X}^\lambda(\infty)$, $0 < \lambda < \infty$, is established in two stages. First, we show that $\vec{X}^\lambda(\infty)$ is tight under FSF_p . We then conclude that this sequence is also tight under FSF .

Tightness under FSF_p : Suppose that the policy FSF_p is used (to be omitted from the notation for brevity). We start by establishing the tightness of $X^\lambda(\infty) = \sum_{k=1}^K X_k^\lambda(\infty)$. Assume, without loss of generality, that $K = 2$. Along the lines of Proposition 3.5 define two related sequences of systems. One is sequence B which is a sequence of $M/M/N_B^\lambda + M$ systems with N_B^λ servers all working with rate μ_2 , where $N_B^\lambda = \left\lfloor \frac{N_1^\lambda \mu_1 + N_2^\lambda \mu_2}{\mu_2} \right\rfloor$. Similarly, define the sequence C to be a sequence of $M/M/N_C^\lambda + M$ systems with N_C^λ servers all working with rate μ_1 , where $N_C^\lambda = \left\lfloor \frac{N_1^\lambda \mu_1 + N_2^\lambda \mu_2}{\mu_1} \right\rfloor$. The sequences B and C both have the same sequence of arrival rates λ as the original system, and the same abandonment rate of θ . Then, according to Proposition 3.5, for every fixed λ , $X^\lambda(\infty)$ is stochastically dominated from above by $Q_B^\lambda(\infty)/\sqrt{N^\lambda}$ and is stochastically dominated from below by

$\frac{Z_C^\lambda(\infty) - N_C^\lambda}{\sqrt{N^\lambda}}$. Tightness of $X^\lambda(\infty)$ now follows from [4, Theorem 2], and the facts that $N_B^\lambda = \lambda/\mu_2 + \frac{\delta}{\sqrt{\mu_2}}\sqrt{\lambda/\mu_2} + o(\sqrt{\lambda})$, $N_C^\lambda = \lambda/\mu_1 + \frac{\delta}{\sqrt{\mu_1}}\sqrt{\lambda/\mu_1} + o(\sqrt{\lambda})$, and that both $\sqrt{N_B^\lambda/N^\lambda}$ and $\sqrt{N_C^\lambda/N^\lambda}$ have finite limits.

Now that we have established that $X^\lambda(\infty)$ is tight, we proceed by showing that $\vec{X}^\lambda(\infty)$ is tight (again for $K = 2$, without loss of generality). Note that under FSF_p , $Q^\lambda + Z_1^\lambda = [Y^\lambda - N_2^\lambda]^+$ and $Z_2^\lambda = \min\{Y^\lambda, N_2^\lambda\}$. Therefore, as long as $Y^\lambda(\infty) \geq N_2^\lambda$, $X_1^\lambda(\infty) = X^\lambda(\infty)$ and $X_2^\lambda(\infty) = 0$. But $Y^\lambda(\infty) \geq N_2^\lambda$ is equivalent to $X^\lambda(\infty) \geq -\frac{N_1^\lambda}{\sqrt{N^\lambda}}$, whose probability goes to 1 as $\lambda \rightarrow \infty$ by tightness of $X^\lambda(\infty)$. Therefore, the vector $\vec{X}^\lambda(\infty)$ is tight.

Tightness under FSF: To establish the tightness of $\vec{X}^\lambda(\infty)$ under FSF, we can use a proof which is essentially identical the proof of part 2. in [1, Proposition 4.6]. All that is missing is to establish that the steady-state probability that all the servers are busy under FSF_p goes to a non-zero limit as $\lambda \rightarrow \infty$. Since we have already established that under FSF_p , $X^\lambda(\infty)$ weakly converges to $X(\infty)$, it is left to show that the probability that $X(\infty)$ is non-negative is non-zero. But this probability is equal to α (in the statement of Proposition 4.5) which is clearly positive. ■

We are now finally in a position to prove the asymptotic optimality of FSF as stated in Theorem 4.1.

Proof of Theorem 4.1: Let $\{\pi^\lambda\}_{\lambda>0} \subseteq \Pi$ be a sequence of policies, and suppose that the steady-state distributions of $Q^\lambda(\cdot; \pi^\lambda)$, $V^\lambda(\cdot; \pi^\lambda)$ and $P_{\pi^\lambda}^\lambda(ab, \cdot)$ exist for all $\lambda > 0$ (here $P_{\pi^\lambda}^\lambda(ab, t)$ is defined as the probability of abandonment for a virtual customer who arrives at time t .) In addition, for $\lambda > 0$, define $\hat{Q}^\lambda(\infty; \pi^\lambda) := Q^\lambda(\infty; \pi^\lambda)/\sqrt{N^\lambda}$, $\hat{W}^\lambda(\infty; \pi^\lambda) := \sqrt{N^\lambda}W^\lambda(\infty; \pi^\lambda)$, and $\hat{P}_{\{\pi^\lambda\}}^\lambda(ab) := \sqrt{N^\lambda}P_{\pi^\lambda}^\lambda(ab)$.

We prove the theorem in three steps:

1. First we show asymptotic optimality of FSF_p in terms of minimizing $\limsup_{\lambda \rightarrow \infty} E\hat{Q}^\lambda(\infty)$, as $\lambda \rightarrow \infty$.
2. The asymptotic optimality of FSF in terms of minimizing $\limsup_{\lambda \rightarrow \infty} E\hat{Q}^\lambda(\infty)$ as $\lambda \rightarrow \infty$ is shown next.

3. We conclude by showing the asymptotic optimality of FSF with respect to both $E\hat{W}^\lambda(\infty)$, $\hat{P}^\lambda(ab)$, and $\sqrt{\lambda}P^\lambda(ab)$ as $\lambda \rightarrow \infty$.

Step 1. In Corollary 3.1 we have shown that FSF_p minimizes $E[Q^\lambda(\infty)]$ for every fixed λ . Therefore, we can conclude that

$$\limsup_{\lambda \rightarrow \infty} E\hat{Q}^\lambda(\infty; \text{FSF}_p) \leq \liminf_{\lambda \rightarrow \infty} E\hat{Q}^\lambda(\infty; \pi^\lambda) \quad (\text{A.12})$$

Step 2. In light of 1. it is sufficient to show that $\lim_{\lambda \rightarrow \infty} E\hat{Q}^\lambda(\infty; \text{FSF}) = \lim_{\lambda \rightarrow \infty} E\hat{Q}^\lambda(\infty; \text{FSF}_p)$, in order to establish the asymptotic optimality of FSF with respect to $E\hat{Q}^\lambda(\infty)$ as $\lambda \rightarrow \infty$. From Proposition 4.6 and the continuous mapping theorem it follows that $\hat{Q}^\lambda(\infty)$ converges weakly to $[X(\infty)]^+$ under both FSF and FSF_p . In order to show that $\lim_{\lambda \rightarrow \infty} E\hat{Q}^\lambda(\infty) = E[X(\infty)]^+$ under both these policies it is sufficient to establish uniform integrability (UI) of $\{\hat{Q}^\lambda(\infty)\}_{\lambda \in (0, \infty)}$ under both policies.

We establish UI of FSF and FSF_p by showing that $\{\hat{Q}^\lambda(\infty)\}_{\lambda \in (0, \infty)}$ is bounded stochastically from above by another UI sequence. Since both sequences are positive it follows that $\{\hat{Q}^\lambda(\infty)\}_{\lambda \in (0, \infty)}$ is UI. Suppose, without loss of generality that $K = 2$. The candidate sequence to be used as an upper bound is $\{Q_B^\lambda(\infty)/\sqrt{N^\lambda}\}_{\lambda \in (0, \infty)}$, where $Q_B^\lambda(\infty)$ is the steady-state queue length in an $M/M/N_B^\lambda + M$ system with arrival rate λ , service rate μ_2 , abandonment rate θ and number of servers $N_B^\lambda = \lfloor \frac{N_1^\lambda \mu_1 + N_2^\lambda \mu_2}{\mu_2} \rfloor$. It has already been established in Proposition 3.5 that for every fixed λ , $Q_B^\lambda(\infty)/\sqrt{N^\lambda}$ is a stochastic upper bound on $\hat{Q}^\lambda(\infty)$ under any work-conserving policy, including FSF and FSF_p . The fact that $\{Q_B^\lambda(\infty)/\sqrt{N^\lambda}\}_{\lambda \in (0, \infty)}$ is UI follows by establishing that $\lim_{\lambda \rightarrow \infty} E Q_B^\lambda(\infty)/\sqrt{N^\lambda} = E \lim_{\lambda \rightarrow \infty} Q_B^\lambda(\infty)/\sqrt{N^\lambda}$. The latter follows from [4, Theorem 4], the relationship (A.4), and [4, Theorem 3].

Step 3. The asymptotic optimality of FSF with respect to $E\hat{W}^\lambda(\infty)$ and $\hat{P}^\lambda(ab)$ follows from Little's law and the relationship (A.4). Finally, the asymptotic optimality of FSF with respect to $\sqrt{\lambda}P^\lambda(ab)$ follows from (4.3).

■

Proof of Corollary 4.1: The proof of this corollary is included in the proof of Theorem 4.1. ■

Proof of Proposition 5.1: We prove the proposition for $K = 2$. The general case follows similarly. Fix $-\infty < \delta < \infty$, and suppose that (5.4) holds. Let $a_k = \liminf_{\lambda \rightarrow \infty} \frac{\mu_k N_k^\lambda}{\lambda}$, $k = 1, 2$. Clearly, $a_1 + a_2 \geq 1$ and $a_1 > 0$. Suppose first that $a_1 + a_2 > 1$. In this case, we can obtain (5.3) with $\Delta := \Delta(\delta, \mu_1, \theta)$ by choosing to use only a subset of each server pool of size $\tilde{N}_k^\lambda = \frac{(a_k/(a_1+a_2))\lambda + (\delta/2)\sqrt{\lambda}}{\mu_k}$, $k = 1, 2$, and apply the policy FSF. Corollary 4.3 then confirms that (5.3) is satisfied. Now, suppose that $a_1 + a_2 = 1$, and without loss of generality, let $a_k = \lim_{\lambda \rightarrow \infty} \frac{\mu_k N_k^\lambda}{\lambda}$. Let $\tilde{\delta} = \liminf_{\lambda \rightarrow \infty} \frac{\mu_1 N_1^\lambda + \mu_2 N_2^\lambda - \lambda}{\sqrt{\lambda}}$ (again, without loss of generality, assume that $\tilde{\delta} = \lim_{\lambda \rightarrow \infty} \frac{\mu_1 N_1^\lambda + \mu_2 N_2^\lambda - \lambda}{\sqrt{\lambda}}$). Clearly, $\tilde{\delta} \geq \delta$, and possibly, $\tilde{\delta} = \infty$. If $\tilde{\delta} > \delta$, then one is able to obtain (5.3) by using FSF with respect to a subset of each server pool of size $\tilde{N}_k^\lambda = \frac{\mu_k N_k^\lambda - (d^\lambda/2)\sqrt{\lambda}}{\mu_k}$, where $d^\lambda := \frac{\mu_1 N_1^\lambda + \mu_2 N_2^\lambda - \lambda}{\sqrt{\lambda}} - \delta$. Finally, if $\tilde{\delta} = \delta$, then (5.3) holds if FSF is used by Corollary 4.3.

Now suppose (5.3) holds for some $0 < \Delta_0 < \infty$, and let $-\infty < \delta_0 < \infty$ be such that $\Delta(\delta_0, \mu_1, \theta) = \Delta_0$ (such δ exists due to Lemma 5.1). Assume by contradiction that (5.4) is violated with respect to $\delta = \delta_0$. Then if (5.4) holds with respect $\delta = \delta_1$ for some $-\infty < \delta_1 < \delta_0$, then by the monotonicity of $\Delta(\delta)$ and Corollary 4.3, FSF will satisfy $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} P^\lambda(ab) = \Delta_1$ where $\Delta_1 > \Delta_0$, which contradicts the asymptotic optimality of FSF (Theorem 4.1). Finally, if (5.4) is violated with respect to any $\delta > -\infty$ then the case $\delta = -\infty$ applies. This case is dealt with next.

To complete the proof, we need to examine the cases where $\delta = -\infty$ and $\delta = \infty$. Suppose first that $\delta = -\infty$, and assume, by contradiction, that there exists a sequence of policies $\{\pi^\lambda = \pi^\lambda(\lambda, \vec{N}^\lambda)\}$, such that $\limsup_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\pi^\lambda}^\lambda(ab) = \Delta < \infty$. Let $-\infty < \delta_0 < \infty$ be such that $\Delta(\delta, \mu_1, \theta) = \Delta$ (δ_0 exists due to Lemma 5.1). Consider another sequence of systems with server pools of size $\tilde{N}_k^\lambda = N_k^\lambda + (\delta_0/4)\sqrt{\lambda}/\mu_k$, $k = 1, 2$. Clearly, (5.4) holds for the new sequence, with $\delta = \delta_0/2$. Now, according to Corollary 4.3, if FSF is used with the new sequence of systems, then $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\text{FSF}}^\lambda(ab) = \Delta(\delta_0/2, \mu_1, \theta) > \Delta$. However, $\{\pi^\lambda\}$ is assumed to obtain a scaled abandonment probability of Δ (asymptotically, over a subsequence) by using only a subset of the servers $(\tilde{N}_1^\lambda, \tilde{N}_2^\lambda)$. This is a contradiction to the asymptotic optimality of FSF (Theorem 4.1). Finally, if $\mu_1 N_1^\lambda + \mu_2 N_2^\lambda \geq \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda})$ for all $\delta < \infty$, then by using FSF with a subset of the servers, one can obtain that $\limsup_{\lambda \rightarrow \infty} \sqrt{\lambda} P^\lambda(ab) = \Delta$, for all $0 < \Delta < \infty$. ■

Proof of Proposition 5.4: The proof is analogous to the proof of Proposition 5.1. The details are omitted. ■

Proof of Proposition 5.2: We prove the proposition for the case $K = 2$. The general case follows similarly. Let $\vec{M}^{*\lambda}$ be the non-negative vector on the half-plane $\mu_1 M_1 + \mu_2 M_2 \geq \lambda + \delta\sqrt{\lambda}$ that

minimizes the staffing cost $C(M)$, $\lambda > 0$. Clearly, $\mu_1 M_1^{*\lambda} + \mu_2 M_2^{*\lambda} = \lambda + \delta\sqrt{\lambda}$. Let $\tilde{N}_k^\lambda = \lceil M_k^{*\lambda} \rceil$, $k = 1, 2$. We prove (5.11), which also implies the validity of (5.12). The outline of the proof is as follows:

1. We solve for $\vec{M}^{*\lambda}$, and $C(\vec{M}^{*\lambda})$ for all $\lambda > 0$, and show that $\lceil \vec{M}^{*\lambda} \rceil$ satisfies the conditions of Proposition 5.1. Now assume by contradiction that $\limsup_{\lambda \rightarrow \infty} \frac{|C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda})|}{\lambda^{p-1/2}} > 0$, and without loss of generality assume that

$$\lim_{\lambda \rightarrow \infty} \frac{|C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda})|}{\lambda^{p-1/2}} = \epsilon > 0. \quad (\text{A.13})$$

2. Assuming first that $C(\vec{M}^{*\lambda_n}) \leq C(\vec{N}^{*\lambda_n})$ on a subsequence $\{\lambda_n\}$, we show that (A.13) implies that for all n large enough there exists a staffing vector \vec{L}^{λ_n} which is feasible for the problem (5.9), but $C(\vec{L}^{\lambda_n}) < C(\vec{N}^{*\lambda_n})$, which is a contradiction to the optimality of $\vec{N}^{*\lambda_n}$.
3. Assuming now that $C(\vec{N}^{*\lambda_n}) \leq C(\vec{M}^{*\lambda_n})$ on a subsequence $\{\lambda_n\}$, we show that (A.13) implies that for all n large enough there exists a staffing vector \vec{L}^{λ_n} which is feasible for the problem (5.6) with the λ on the right-hand-side replaced by $\lambda + \delta\sqrt{\lambda}$, but $C(\vec{L}^{\lambda_n}) < C(\vec{M}^{*\lambda_n})$, which is a contradiction to the optimality of $\vec{M}^{*\lambda_n}$.

We now proceed with the details of steps 1-3.

1. To find $\vec{M}^{*\lambda}$ and $C(\vec{M}^{*\lambda})$ one needs to solve the problem:

$$\begin{aligned} & \text{minimize} && c_1 M_1^p + c_2 M_2^p \\ & \text{subject to} && \mu_1 M_1 + \mu_2 M_2 \geq \lambda + \delta\sqrt{\lambda} \\ & && M_1, M_2 \geq 0. \end{aligned} \quad (\text{A.14})$$

The solution to (A.14) is given in (5.10), and for $K = 2$ it satisfies

$$(M_1^{*\lambda}, M_2^{*\lambda}) = (\lambda + \delta\sqrt{\lambda}) \cdot \frac{((\mu_1 c_2)^{1/(p-1)}, (\mu_2 c_1)^{1/(p-1)})}{(\mu_1^p c_2)^{1/(p-1)} + (\mu_2^p c_1)^{1/(p-1)}},$$

and

$$C(\vec{M}^{*\lambda}) = \frac{(\lambda + \delta\sqrt{\lambda})^p c_1 c_2}{((\mu_1^p c_2)^{1/(p-1)} + (\mu_2^p c_1)^{1/(p-1)})^{p-1}} \triangleq (\lambda + \delta\sqrt{\lambda})^p \xi. \quad (\text{A.15})$$

In particular, $\lceil \vec{M}^{*\lambda} \rceil$ satisfies condition (5.2) of Proposition 5.1, because

$$\frac{M_1^{*\lambda}}{M_1^{*\lambda} + M_2^{*\lambda}} \equiv \frac{(\mu_1 c_2)^{1/(p-1)}}{(\mu_1 c_2)^{1/(p-1)} + (\mu_2 c_1)^{1/(p-1)}} > 0.$$

2. Assume that (A.13) holds and that, without loss of generality, $C(\vec{M}^{*\lambda}) \leq C(\vec{N}^{*\lambda})$ for all λ . By (A.13), we have that for all λ large enough

$$C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda}) \geq \lambda^{p-\frac{1}{2}} \frac{\epsilon}{2}.$$

Let \vec{M}^λ be the solution of (5.6) with λ on the right-hand side replaced by $\lambda + (\delta + \eta)\sqrt{\lambda}$, $\eta = \frac{\epsilon}{8\xi p}$. Then,

$$\begin{aligned} \frac{C(\vec{M}^\lambda) - C(\vec{M}^{*\lambda})}{\lambda^{p-\frac{1}{2}}} &= \xi \lambda^p \frac{\left(1 + \frac{\delta+\eta}{\sqrt{\lambda}}\right)^p - \left(1 + \frac{\delta}{\sqrt{\lambda}}\right)^p}{\lambda^{p-\frac{1}{2}}} \\ &= \xi \lambda^p \frac{1 + p(\delta + \eta)/\sqrt{\lambda} - 1 - p\delta/\sqrt{\lambda} + o(1/\sqrt{\lambda})}{\lambda^{p-\frac{1}{2}}} = \xi[p\eta + o(1)] < \frac{\epsilon}{4}, \end{aligned}$$

for all λ large enough.

In particular, $C(\vec{M}^\lambda) < \lambda^{P-\frac{1}{2}} \frac{\epsilon}{4} + C(\vec{M}^{*\lambda}) \leq C(\vec{N}^{*\lambda}) - \lambda^{P-\frac{1}{2}} \frac{\epsilon}{4}$ for all λ . Let \vec{L}^λ be such that $L_k^\lambda = \lceil M_k^\lambda \rceil$, $k = 1, 2$, for all λ . Then, for all λ large enough, we also have that

$$C(\vec{L}^\lambda) \leq C(\vec{N}^{*\lambda}) - \lambda^{P-\frac{1}{2}} \frac{\epsilon}{8} < C(\vec{N}^{*\lambda}). \quad (\text{A.16})$$

Now, note that by Corollary 4.3, we have that, when staffing the λ system with \vec{L}^λ ,

$$\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\text{FSF}}^\lambda(ab) = \Delta(\delta + \eta, \mu, \theta) < \Delta(\delta, \mu, \theta).$$

In particular, $\sqrt{\lambda} P_{\text{FSF}}^\lambda(ab) \leq \Delta(\delta, \mu, \theta)$ for all λ large enough, which implies that \vec{L}^λ is a feasible solution of (5.9), which by (A.16) is a contradiction to the optimality of $\vec{N}^{*\lambda}$.

Before we turn to step 3 of the proof, we state and prove two lemmas.

Lemma A.1 *Suppose that for all $\lambda > 0$, $\vec{N}^{*\lambda}$ is an optimal solution of (2.1) and*

$$\liminf_{\lambda \rightarrow \infty} \frac{N_1^{*\lambda}}{N_1^{*\lambda} + N_2^{*\lambda}} > 0. \text{ Then, } \mu_1 N_1^{*\lambda} + \mu_2 N_2^{*\lambda} = \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda}).$$

Proof: By contradiction, assume that either there exists a subsequence $\{\lambda_j\}$ for which $\mu_1 N_1^{*\lambda_j} + \mu_2 N_2^{*\lambda_j} < \lambda_j + \delta\sqrt{\lambda_j} + o(\sqrt{\lambda_j})$, or there exists $\tilde{\epsilon} > 0$ such that $\mu_1 N_1^{*\lambda_j} + \mu_2 N_2^{*\lambda_j} \geq \lambda_j + (\delta + \tilde{\epsilon})\sqrt{\lambda_j} + o(\sqrt{\lambda_j})$. In the first case, by Proposition 5.1, $\limsup_{j \rightarrow \infty} \sqrt{\lambda_j} P_{\pi^{\lambda_j}}^{\lambda_j}(ab) > \Delta$, for all $\pi^{\lambda_j} \in \Pi$, which is a contradiction to the feasibility of $\vec{N}^{*\lambda_j}$, for some large values of j . In the second case, let $\vec{N}^{\lambda_j} = \vec{N}^{*\lambda_j} - \vec{e}$ (where \vec{e} is a vector of 1's). Then $C(\vec{N}^{\lambda_j}) < C(\vec{N}^{*\lambda_j})$, and by Proposition 5.1, there exists a sequence of policies $\{\pi^{\lambda_j} = \pi^{\lambda_j}(\lambda_j, \vec{N}^{\lambda_j})\} \subseteq \Pi$ under which $\limsup_{j \rightarrow \infty} \sqrt{\lambda} P_{\pi^{\lambda_j}}^{\lambda_j}(ab) < \Delta$. This is a contradiction to the optimality of $\vec{N}^{*\lambda_j}$ for all large j . ■

Lemma A.2 Let \vec{N}^λ be a sequence of staffing vectors satisfying $\mu_1 N_1^\lambda + \mu_2 N_2^\lambda < \lambda + \delta\sqrt{\lambda}$ for some $-\infty < \delta < \infty$, and $\lim_{\lambda \rightarrow \infty} \frac{N_1^\lambda}{N_1^\lambda + N_2^\lambda} = 0$. Suppose that there exists a sequence of policies $\{\pi^\lambda = \pi^\lambda(\lambda, \vec{N}^\lambda)\} \subseteq \Pi$ such that $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\pi^\lambda}^\lambda(ab) = \Delta$, where $\Delta = \Delta(\delta, \mu_1, \theta)$. Then, \vec{N}^λ satisfies

$$\mu_1 N_1^\lambda + \mu_2 N_2^\lambda \geq \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda}). \quad (\text{A.17})$$

Proof: Let $\delta_1 = \lim_{n \rightarrow \infty} \frac{\mu_1 N_1^{\lambda_n}}{\sqrt{\lambda_n}}$, $0 \leq \delta_1 \leq \infty$ and let $\delta_2 = \lim_{n \rightarrow \infty} \frac{\mu_2 N_2^{\lambda_n} - \lambda_n}{\sqrt{\lambda_n}}$, $-\infty \leq \delta_2 \leq \infty$, where $\{\lambda_n\}$ is a subsequence along which these limits are well defined. Without loss of generality, assume that $\{\lambda_n\} \equiv \{\lambda\}$. We show that if the policy FSF_p is used to process the system (which by Proposition 3.1 implies that $\limsup_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\text{FSF}_p}^\lambda(ab) \leq \Delta$), then (A.17) is satisfied.

We consider three different cases with respect to the value of δ_1 : (a) $0 < \delta_1 < \infty$, (b) $\delta_1 = 0$ and (c) $\delta_1 = \infty$, and note that since $\delta_1 + \delta_2 < \delta$, we have $\delta_2 < \infty$. Denote $\tilde{\delta} := \delta_1 + \delta_2$.

Case (a) ($0 < \delta_1 < \infty$): First suppose that $\delta_2 > -\infty$. In this case, we can show, using Stone's criterion and the birth-and-death representation of the total number in the system given in (3.1), that the scaled process $X^\lambda(t)$ weakly converges to a diffusion process $X(t)$ with infinitesimal drift

$$m(x) = \begin{cases} -\tilde{\delta}\sqrt{\mu_2} - \theta x & x \geq 0 \\ -\tilde{\delta}\sqrt{\mu_2} - \mu_1 x & -\frac{\delta_1}{\mu_1}\sqrt{\mu_2} \leq x < 0 \\ -\mu_2 \left(\frac{\delta_1}{\mu_1} + \frac{\delta_2}{\mu_2} \right) \sqrt{\mu_2} - \mu_2 x & x < -\frac{\delta_1}{\mu_1}\sqrt{\mu_2} \end{cases}$$

and infinitesimal variance $\sigma^2(x) = 2\mu_2$.

Let \underline{X} be another diffusion process with the same infinitesimal variance and with infinitesimal drift

$$\underline{m}(x) = \begin{cases} -\tilde{\delta}\sqrt{\mu_2} - \theta x & x \geq 0 \\ -\tilde{\delta}\sqrt{\mu_2} - \mu_1 x & x < 0. \end{cases}$$

Then, clearly $m(x) \geq \underline{m}(x)$ for all x , and therefore, by Proposition 18.5 of [2] we have that $X(\infty) \stackrel{st}{\geq} \underline{X}(\infty)$. In particular, $\frac{\theta}{\sqrt{\mu_2}} EX^+(\infty) \geq \frac{\theta}{\sqrt{\mu_2}} E\underline{X}^+(\infty) = \Delta(\tilde{\delta}, \mu_1, \theta)$. Therefore, by establishing that $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\text{FSF}_p}^\lambda(ab) = \frac{\theta}{\sqrt{\mu_2}} EX^+(\infty)$, the proof is complete for this case (recalling that Δ is a decreasing function of δ (see Lemma 5.1). Note that this latter limit holds due to the tightness and uniform integrability results established in the proofs of Proposition 4.6 (step 2) and of Theorem 4.1 (step 2).

Next, consider the case $\delta_2 = -\infty$. We claim that in this case, $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\text{FSF}_p}^\lambda(ab) = \infty$. To see this, we first note that if $-\infty < \delta_2 < \infty$, then by [2] along with the tightness and uniform integrability results, we have that

$$\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\text{FSF}_p}^\lambda(ab) = \tilde{\Delta}(\delta_1, \delta_2, \mu_1, \mu_2, \theta) := \sqrt{\theta} \alpha_1 \left[h \left(\frac{\delta_1 + \delta_2}{\sqrt{\theta}} \right) - \frac{(\delta_1 + \delta_2)}{\sqrt{\theta}} \right],$$

where $\alpha_1 := \alpha_1(\delta_1, \delta_2, \mu_1, \mu_2, \theta) = \left[1 + \sqrt{\frac{\theta}{\mu_1}} h \left(\frac{\tilde{\delta}}{\sqrt{\theta}} \right) \frac{\Phi(\tilde{\delta}/\sqrt{\mu_1}) - \Phi(\delta_2/\sqrt{\mu_1})}{\phi(\tilde{\delta}/\sqrt{\mu_1})} + \sqrt{\frac{\theta}{\mu_2}} \frac{h(\tilde{\delta}/\theta)}{h(-\delta_2/\sqrt{\mu_2})} \right]^{-1}$, with $\tilde{\delta} := \delta_1 + \delta_2$. Note that $\tilde{\Delta}$ is continuous in δ_2 and that $\lim_{\delta_2 \rightarrow \infty} \tilde{\Delta} = 0$ and $\lim_{\delta_2 \rightarrow -\infty} \tilde{\Delta} = \infty$ (these limits may be obtained by a successive application of L'Hôpital's rule). Therefore, by an argument analogous to the proof of Proposition 5.1 for the case $\delta = -\infty$, one can show that indeed if $\delta_2 = -\infty$ then $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\text{FSF}_p}^\lambda(ab) = \infty$. This finally leads to a contradiction due to the optimality of FSF_p .

Case (b) ($\delta_1 = 0$): In this case, if $\delta_2 > -\infty$, then one can show that the scaled process $X^\lambda(t)$ weakly converges to a diffusion process $X(t)$ with infinitesimal drift

$$m(x) = \begin{cases} -\delta_2 \sqrt{\mu_2} - \theta x & x \geq 0 \\ -\delta_2 \sqrt{\mu_2} - \mu_2 x & x < 0 \end{cases}$$

and infinitesimal variance $\sigma^2(x) = 2\mu_2$. Consider another diffusion process \underline{X} with the same infinitesimal variance and with infinitesimal drift equal to

$$\underline{m}(x) = \begin{cases} -\delta_2 \sqrt{\mu_2} - \theta x & x \geq 0 \\ -\delta_2 \sqrt{\mu_2} - \mu_1 x & x < 0. \end{cases}$$

Then, clearly, $m(x) \geq \underline{m}(x)$ for all x , and hence, by analogous arguments to the ones used in case (a), we have that

$$\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\text{FSF}_p}^\lambda(ab) \geq \Delta(\delta_2, \mu_1, \theta),$$

which implies by Lemma 5.1 that $\delta_2 \geq \delta$, and in turn, that $\mu_1 N_1^\lambda + \mu_2 N_2^\lambda \geq \lambda + \delta \sqrt{\lambda} + o(\sqrt{\lambda})$. The case $\delta_2 = -\infty$ may be analyzed analogously to $\delta_2 = -\infty$ in case (a) to show that if $\delta_2 = -\infty$ then $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\text{FSF}_p}^\lambda(ab) = \infty$, which leads to a contradiction.

Case (c) ($\delta_1 = \infty$): In this case, since $\delta_1 + \delta_2 := \lim_{\lambda \rightarrow \infty} \frac{\mu_1 N_1 + \mu_2 N_2 - \lambda}{\sqrt{\lambda}} \leq \delta$, we have that necessarily, $\delta_2 = -\infty$. Assume first that $\delta_1 + \delta_2 = \tilde{\delta} > -\infty$. Then, in this case, the scaled process $X^\lambda(t)$ weakly converges to a diffusion process $X(t)$ with infinitesimal drift

$$m(x) = \begin{cases} -\tilde{\delta} \sqrt{\mu_2} - \theta x & x \geq 0 \\ -\tilde{\delta} \sqrt{\mu_2} - \mu_1 x & x < 0 \end{cases}$$

and infinitesimal variance $\sigma^2(x) = 2\mu_2$. This process has the same law as the diffusion process defined through (4.26) and (4.27) with $\delta = \tilde{\delta}$ and $\mu = \mu_2$. In particular, one can show that $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} P_{\text{FSF}_p}^\lambda(ab) = \Delta(\tilde{\delta}, \mu_1, \theta)$, which, in turn, implies that $\tilde{\delta} \geq \delta$, so that $\mu_1 N_1 + \mu_2 N_2 \geq \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda})$. Finally, if $\delta_1 + \delta_2 = -\infty$, one can obtain a contradiction in the same way that was done for cases (a) and (b). ■

We now return to step 3. in the proof of Proposition 5.2.

3. Assume that (A.13) holds and that, without loss of generality, $C(\vec{N}^{*\lambda}) \leq C(\vec{M}^{*\lambda})$ for all λ . By (A.13), we have that for all λ large enough

$$C(\vec{M}^{*\lambda}) - C(\vec{N}^{*\lambda}) \geq \lambda^{p-\frac{1}{2}} \frac{\epsilon}{2}.$$

By the optimality of $\vec{M}^{*\lambda}$, we have that $\mu_1 N_1^{*\lambda} + \mu_2 N_2^{*\lambda} < \lambda + \delta\sqrt{\lambda}$ for all λ large enough. Therefore, by Lemmas A.1 and A.2, we have that $\mu_1 N_1^{*\lambda} + \mu_2 N_2^{*\lambda} \geq \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda})$.

Let $f^\lambda := \lambda + \delta\sqrt{\lambda} - (\mu_1 N_1^* + \mu_2 N_2^*)$. Then, $f^\lambda > 0$ and $f^\lambda = o(\sqrt{\lambda})$. Let $\vec{L}^\lambda = b^\lambda \cdot \vec{N}^{*\lambda}$, where $b^\lambda := \frac{\lambda + \delta\sqrt{\lambda}}{\lambda + \delta\sqrt{\lambda} - f^\lambda} = \left(1 - \frac{f^\lambda}{\lambda + \delta\sqrt{\lambda}}\right)^{-1}$. Then, one can verify that $\mu_1 L_1^\lambda + \mu_2 L_2^\lambda = \lambda + \delta\sqrt{\lambda}$. In particular,

$$C(\vec{L}^\lambda) \geq C(\vec{M}^{*\lambda}). \quad (\text{A.18})$$

Note that $(b^\lambda)^p = \frac{1}{\left(1 - \frac{f^\lambda}{\lambda + \delta\sqrt{\lambda}}\right)^p} = \frac{1}{1 - p \frac{f^\lambda}{\lambda + \delta\sqrt{\lambda}} + o(1/\sqrt{\lambda})} = 1 + p \frac{f^\lambda}{\lambda + \delta\sqrt{\lambda}} + o(1/\sqrt{\lambda})$. We now have that

$$\begin{aligned} \frac{C(\vec{L}^{*\lambda}) - C(\vec{N}^{*\lambda})}{\lambda^{p-\frac{1}{2}}} &= \frac{C(\vec{N}^{*\lambda})((b^*)^p - 1)}{\lambda^{p-\frac{1}{2}}} = \frac{C(\vec{N}^{*\lambda}) \left(p \frac{f^\lambda}{\lambda + \delta\sqrt{\lambda}} + o\left(1/\sqrt{\lambda}\right)\right)}{\lambda^{p-\frac{1}{2}}} \\ &\leq \frac{C(\vec{M}^{*\lambda}) \left(p \frac{f^\lambda}{\lambda + \delta\sqrt{\lambda}} + o\left(1/\sqrt{\lambda}\right)\right)}{\lambda^{p-\frac{1}{2}}} = \frac{\xi \left(\lambda + \delta\sqrt{\lambda}\right)^p \left(p \frac{1}{\sqrt{\lambda}} \frac{f^\lambda}{\lambda + \delta} + o\left(1/\sqrt{\lambda}\right)\right)}{\lambda^{p-\frac{1}{2}}} \\ &= \xi \left(p \left(1 + \frac{\delta}{\sqrt{\lambda}}\right)^p \frac{f^\lambda}{\sqrt{\lambda} + \delta} + o\left(1/\sqrt{\lambda}\right)\right) \rightarrow 0 \text{ as } \lambda \rightarrow \infty. \end{aligned}$$

In particular,

$$C(\vec{L}^{*\lambda}) \leq \lambda^{p-\frac{1}{2}} \frac{\epsilon}{4} + C(\vec{N}^{*\lambda}) \leq C(\vec{M}^{*\lambda}) - \lambda^{p-\frac{1}{2}} \frac{\epsilon}{4} < C(\vec{M}^{*\lambda}),$$

for all λ large enough. This is in contradiction to (A.18). ■

Proof of Proposition 5.3: We prove the proposition for the case $K = 2$. The general case follows similarly. Let $\vec{M}^{*\lambda}$ be the non-negative vector on the half-plane $\mu_1 M_1 + \mu_2 M_2 \geq \lambda + \delta\sqrt{\lambda}$ that

minimizes the staffing cost $C(M)$, $\lambda > 0$. Clearly, $\mu_1 M_1^{*\lambda} + \mu_2 M_2^{*\lambda} = \lambda + \delta\sqrt{\lambda}$. Let $\tilde{N}_k^\lambda = \lceil M_k^{*\lambda} \rceil$, $k = 1, 2$. We prove that $\lim_{\lambda \rightarrow \infty} \frac{C(\tilde{N}^{*\lambda}) - \underline{C}^\lambda}{C(\vec{M}^{*\lambda}) - \underline{C}^\lambda} = 1$. The asymptotic optimality of \tilde{N}^λ then easily follows. The outline of the proof is as follows:

1. We solve for \underline{C}^λ , and explicitly express $C(\vec{M}^{*\lambda}) - \underline{C}^\lambda$.
2. The asymptotic optimality then follows easily from Proposition 5.2.

1. To find \underline{C}^λ , one needs to solve the problem (5.6). Simple constrained optimization obtains:

$$\underline{C}^\lambda = \frac{\lambda^p c_1 c_2}{((\mu_1^p c_2)^{1/(p-1)} + (\mu_2^p c_1)^{1/(p-1)})^{p-1}} = \lambda^p \xi. \quad (\text{A.19})$$

It follows that

$$C(\vec{M}^{*\lambda}) - \underline{C}^\lambda = \xi p \delta \lambda^{p-1/2} + o(\lambda^{p-1/2}).$$

2. By proposition 5.2, we have that $C(\tilde{N}^{*\lambda}) = C(\vec{M}^{*\lambda}) + f^\lambda$, where $f^\lambda = o(\lambda^{p-1/2})$. Therefore,

$$\frac{C(\tilde{N}^{*\lambda}) - \underline{C}^\lambda}{C(\vec{M}^{*\lambda}) - \underline{C}^\lambda} = \frac{\xi p \delta \lambda^{p-1/2} + o(\lambda^{p-1/2}) + f^\lambda}{\xi p \delta \lambda^{p-1/2} + o(\lambda^{p-1/2})} = 1 + \frac{f^\lambda}{\xi p \delta \lambda^{p-1/2} + o(\lambda^{p-1/2})} \rightarrow 1,$$

as $\lambda \rightarrow \infty$, provided that $\delta \neq 0$.

■

Proof of Proposition 5.5: The proof follows similarly to the proof of Proposition 5.3. The details are omitted.

■

References

- [1] Armony, M. (2005), Dynamic routing in large-scale service systems with heterogeneous servers, *Queueing Systems*, **51**(3–4), pp. 287–329. [A](#)
- [2] Browne, S. and Whitt, W. (1995), Piecewise-linear diffusion processes, *Advances in Queueing. Theory, Methods, and Open Problems*, Dshalalow, J.H. (editor), CRC Press, Chapter 18, pp 463–480. [A](#), [A](#)
- [3] Ethier, S.N. and Kurtz, T.G. (1985), *Markov processes, characterization and convergence*, John Wiley & Sons. [A](#)
- [4] Garnett, O., Mandelbaum, A. and Reiman, M. (2002), Designing a call center with impatient customers, *Manufacturing & Service Operations Management*, **4**(3), pp. 208-227. [A](#), [A](#), [A](#)
- [5] Gurvich, I. and Whitt, W. (2007) Queue-and-Idleness-Ratio Controls in Many-Server Service Systems. Working paper. [A](#)
- [6] T. Tezcan and J.G. Dai (2006) Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. Working paper. [A](#)