OPERATIONS RESEARCH
Vol. 59, No. 1, January–February 2011, pp. 50–65
ISSN 0030-364X | EISSN 1526-5463 | 11 | 5901 | 0050

informs

DOI 10.1287/opre.1100.0878


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

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

**Avishai Mandelbaum**
Department of Industrial Engineering and Management, Technion Institute of Technology, 32000 Haifa, Israel, avim@ie.technion.ac.il

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 in the heavy-traffic many-server QED, ED, and ED + QED regimes, respectively. For the QED regime, our proposed routing rule is FSF, that assigns customers to the fastest server available first. In the ED and ED + QED regimes, all work-conserving policies perform (asymptotically) equally well. In all these regimes, the form of the asymptotically optimal staffing is consistent with the asymptotically optimal staffing in the same regimes in the single-pool case, respectively. In particular, the total service capacity is (asymptotically) equal to a term that is proportional to the arrival rate plus, possibly, a term that is proportional to the square-root of the arrival rate, with both terms being regime dependent. Our specific proposed approximation for the optimal staffing vector is obtained via a straightforward solution to a deterministic optimization problem subject to a linear feasible region.

*Subject classifications*: queues: applications, balking and reneging, diffusion models, limit theorems.
*Area of review*: Stochastic Models.
*History*: Received June 2008; revisions received June 2009, October 2009, January 2010, March 2010; accepted April 2010. Published online in *Articles in Advance* February 8, 2011.

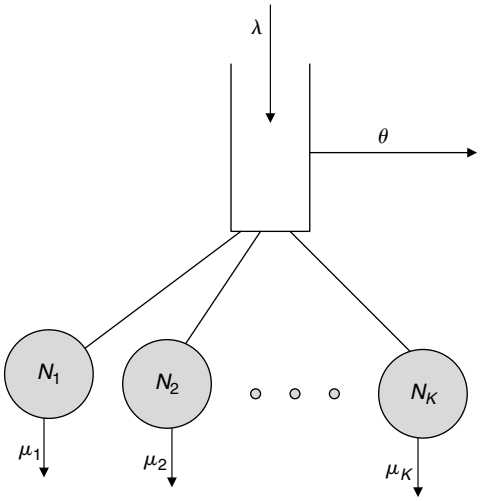
## 1. Introduction

In this paper we consider large-scale service systems, such as customer contact centers (Aksin et al. 2007, Gans et al. 2003) or hospitals (Tseytlin 2007), 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. 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 and in practice. In particular, authors have assumed away the routing question to address the staffing problem (e.g., Borst et al. 2003, Mandelbaum and Zeltyn 2009) or have assumed a particular staffing rule to solve the routing problem (e.g., Armony 2005, Tezcan and Dai 2010, Dai and Tezcan 2008). In contrast, we address this joint problem by rigorously justifying a "divide and conquer" approach; that is, we first find a routing scheme that is (asymptotically) optimal given *any "reasonable"* 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 joint staffing-routing problem is asymptotic. Specifically, we identify rules that are asymptotically optimal as both the arrival rate and number of servers of each pool grow to infinity within the QED, ED, and ED + QED regimes, respectively. The QED regime was first formalized by Halfin and Whitt (1981) and was later adapted to queues with abandonment by Garnett et al. (2002). 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. In the ED regime (Whitt 2004, Mandelbaum and Zeltyn 2009), the fraction of abandonment approaches a limit that is strictly between 0 and 1, while the delay probability converges to 1. Finally, in the ED + QED regime (Baron and Milner 2009, Mandelbaum and Zeltyn 2009), the probability that the waiting time will exceed a prespecified (positive) upper bound also approaches a limit that is strictly between 0 and 1.

**Figure 1.** The inverted-V model.




An analogous model was studied in Armony (2005) in the QED regime, under the assumption that customers are infinitely patient. The routing scheme, FSF, was proposed in Armony (2005). 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 Armony (2005) are that here (a) customers are impatient, (b) the ED and ED + QED regimes are considered in addition to the QED regime, and (c) 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 at least three fronts:

**1. Asymptotic Regimes.** Without abandonment, a natural approach with respect to quality of service is to minimize the delay probability (which is indeed the approach taken in Armony 2005). In that scenario, one is naturally led to work with the QED asymptotic regime. With abandonment, other performance measures become relevant, such as the fraction of abandonment, and the waiting time distribution. While the QED regime is still relevant here, the ED and ED + QED regimes are also realistic and relevant in practice.

**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 because 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 §5.2).

**3. Assumptions.** 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 if customers' patience is stochastically longer than their service time, or otherwise, if one is restricted to working with work-conserving policies only. In contrast, if one wishes to stochastically minimize the cumulative number of abandonment at any point in time, these additional assumptions are not required.

### 1.1. Summary of the Results

The approach we take in the paper is as follows. For each of the three asymptotic regimes, we first assume that the staffing vector is of a form that is consistent with that regime. Under this assumption, we identify a routing rule that asymptotically minimizes the relevant performance measure. We then establish that the assumed form of the staffing vector is necessary and sufficient for asymptotic feasibility of the joint staffing and routing problem. Finally, we propose a staffing vector that minimizes the staffing cost within the asymptotically feasible region and establish the asymptotic optimality of our proposed solution.

Consider a sequence of systems indexed by the arrival rate $\lambda$, where $\lambda \uparrow \infty$.[1] For any fixed value of $\lambda$ (the system indexed by $\lambda$ will be referred to as the $\lambda$-system), let $N_k^\lambda$ represent the number of servers of type $k$, $k = 1, \ldots, K$. Also, let $\vec{N}^\lambda = (N_1^\lambda, N_2^\lambda, \ldots, N_K^\lambda)$ be the staffing vector, and $N^\lambda = N_1^\lambda + N_2^\lambda + \cdots + N_K^\lambda$ be the total number of servers. Suppose that the service rates $\mu_1, \ldots, \mu_K$ and the abandonment rate $\theta$ are fixed independently of $\lambda$. Let $W := W(\infty)$ be the steady-state waiting time, and let $P(ab)$ be the steady-state abandonment probability. The joint staffing and routing problem can be described as minimizing the staffing cost, subject to a quality-of-service constraint. The first row in Table 1 specifies the constraint associated with each regime. To be consistent with the various asymptotic regimes, we first assume that the total service capacity, $\mu_1 N_1^\lambda + \mu_2 N_2^\lambda + \cdots + \mu_K N_K^\lambda$, is of the form described in the second row of Table 1.[2] In words, in the QED regime, the total service capacity is equal to the demand ($\lambda$) plus (or minus) a safety capacity that is in the order of square-root of the demand. In the ED regime, the service capacity is linearly proportional to, and slightly less than, the demand. In particular, the traffic intensity is strictly greater than 1. Finally, in the QED + ED regime, the basic ED capacity is supplemented by a term that is of the order of square-root of the demand. The purpose of this additional term is to fine-tune the target waiting-time tail probability.

**Table 1.** Summary of the results.

| | QED | QED | ED | ED + QED |
|---|---|---|---|---|
| Constraint | $\sqrt{\lambda}P(ab) \leqslant \Delta$ | $P(W > T/\sqrt{\lambda}) \leqslant \alpha$<br>$T \geqslant 0$ | $P(ab) \leqslant \Delta$ | $P(W > T) \leqslant \alpha$<br>$T > 0$ |
| | $\Delta \in (0, \infty)$ | $\alpha \in (0, 1)$ | $\Delta \in (0, 1)$ | $\alpha \in (0, e^{-\theta T})$ |
| $\sum_{k=1}^{K} \mu_k N_k^\lambda =$ | $\lambda + \delta\sqrt{\lambda}$, | $\lambda + \delta_0\sqrt{\lambda}$, | $\lambda(1-\Delta)$ | $\lambda(1-\Delta_1) + \delta_1\sqrt{\lambda}$, |
| | $\delta = \delta(\Delta, \theta, \mu_1)$ | $\delta_0 = \delta_0(\alpha, \theta, \mu_1)$ | | $\Delta_1 = \Delta_1(T)$, |
| | $\lim_{\lambda\to\infty} \frac{N_1^\lambda}{N^\lambda} > 0$ | $\lim_{\lambda\to\infty} \frac{N_1^\lambda}{N^\lambda} > 0$ | | $\delta_1 = \delta_1(T, \alpha)$ |
| Assumptions | | FCFS<br>$\theta \leqslant \mu_1$ or W.C. | | FCFS<br>$\theta \leqslant \mu_1$ or W.C. |
| Routing | FSF | FSF | W.C. | W.C. |
| $C(\vec{N}^{*\lambda}) - C(\widetilde{N}^\lambda) =$ | $o(\lambda^{p-1/2})$ | $o(\lambda^{p-1/2})$ | $o(\lambda^p)$ | $o(\lambda^{p-1/2})$ |
| $\|\vec{N}^{*\lambda} - \widetilde{N}^\lambda\| = o(\lambda)$ | ✓ | ✓ | ✓ | ✓ |
| $\lim_{\lambda\to\infty} \frac{C(\vec{N}^{*\lambda}) - \underline{C}^{\lambda(1-\Delta)}}{C(\widetilde{N}^\lambda) - \underline{C}^{\lambda(1-\Delta)}} = 1$ | ✓ ($\delta \neq 0$) | ✓ ($\delta_0 \neq 0$) | | ✓ ($\delta_1 \neq 0$) |
| | $\Delta = 0$ | $\Delta = 0$ | | $\Delta > 0$ |

*Note.* FCFS = first-come-first-served; W.C. = work conservation.

These proposed staffing forms are later shown to be necessary and sufficient for asymptotic feasibility of the original problem.

Assuming that the staffing vector is indeed consistent with the specified asymptotic form, we identify routing rules that asymptotically minimize the relevant performance measure, under (possibly) some additional assumptions. Specifically, the policy FSF is asymptotically optimal in the QED regime, and all work-conserving policies are asymptotically optimal in the ED and ED + QED regimes.

The asymptotic optimality of our proposed routing rules, given any regime-consistent staffing vectors, facilitates the solution of the joint staffing and routing problem. Specifically, when focusing on an asymptotically optimal staffing, one can simply assume that those rules are used as the routing rule. This allows us to identify the relevant region to be the asymptotically *feasible* region for the corresponding problem. Moreover, we explicitly provide a one-to-one correspondence between the staffing parameters and the parameters associated with the quality-of-service constraint.

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 a region that approximates the asymptotically feasible region. We explicitly provide the solution under homogeneous, polynomial, and additive cost functions of the form $C(\vec{N}) = c_1 N_1^p + c_2 N_2^p + \cdots + c_K N_K^p$, with $p > 1$ and $c_k > 0$, for all $k$, and establish its asymptotic optimality with respect to the criteria specified in the last three rows of Table 1.

The paper is organized as follows: We conclude the introduction by reviewing the relevant literature. Section 2 gives the model formulation as well as the formulation of the joint staffing and routing problem, focusing on quality-of-service constraints that are associated with the fraction of abandonment. We then proceed to §3, in which the preemptive version of FSF is introduced and its optimality is established. Next, §4 introduces the asymptotic framework associated with the QED and ED regimes and verifies asymptotic optimality of our proposed routing rules. Section 5 then establishes asymptotic feasibility and optimality of our proposed solutions in those regimes. Finally, §6 extends our results to the problem that focuses on the probability that the waiting time exceeds a certain threshold. This is where the ED + QED regime is discussed. Section 7 concludes the paper. An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/. All the proofs are given in the electronic companion to allow for a fluent reading of the paper.

### 1.2. Literature Review

Service systems with heterogeneous servers arise naturally due to training and learning effects (Gans et al. 2010) and also due to heterogeneity in the workforce. In particular, such heterogeneity arises in a co-sourcing environment. (Co-sourcing in call centers is a common arrangement in which the firm outsources part of its call center operations and keeps the rest in-house.) In such an environment, it is likely that the in-house customer service representatives (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 this dynamic routing problem in conjunction with the staffing problem of determining

the number of servers required of each pool. We begin our review by briefly describing the relevant asymptotic staffing literature. Next we detail the relevant literature in the control of the inverted-V system. In passing, we comment on those papers that have combined the staffing and routing components.

**Staffing.** Staffing of large-scale service systems via asymptotic analysis was first formalized by Halfin and Whitt (1981) for the Erlang-C (M/M/N) model. The paper (Halfin and Whitt 1981) showed that a square-root safety-staffing is necessary and sufficient for the delay probability to approach a limit that is strictly between 0 and 1, as the system load grows large. This regime has been coined the Halfin-Whitt (or QED) regime. The same regime was later shown to be cost effective (Borst et al. 2003) and profit maximizing (Maglaras and Zeevi 2003) in various settings (the control component of the latter work is elaborated upon in Maglaras and Zeevi 2004, 2005). For the Erlang-A model (M/M/N + M) Garnett et al. (2002) showed that a similar square-root safety-staffing rule guarantees that the abandonment probability is of the order of $1/\sqrt{N}$. The efficiency-driven (ED) regime was studied extensively by Whitt (2004, 2005, 2006a, b, c) via both fluid and diffusion models. In this regime, the traffic intensity is strictly greater than 1, customers are delayed with probability 1, and the probability of abandonment is strictly between 0 and 1. Recently, Baron and Milner (2009) and Mandelbaum and Zeltyn (2009) have identified the so-called ED + QED regime, in which the basic ED staffing is supplemented by a square-root term. This regime was shown to be necessary and sufficient in guaranteeing that the probability that the waiting time will exceed a prespecified positive threshold is strictly between 0 and 1. In the present paper, we establish asymptotic optimality of these three staffing regimes, under appropriate form and scaling of the desired quality of service, for the inverted-V model.

**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., Rubinovich 1983, Cabral 2005)? (b) Given a set of heterogeneous servers, how should customers be routed dynamically to servers in order to minimize the mean sojourn time (e.g., Larsen and Agrawala 1983, Lin and Kumar 1984, Stockbridge 1991, and de Véricourt and Zhou 2005)? Both of these problems have been coined "the slow server problem." For awhile, only results for the two-server system had been published (e.g., Rubinovich 1983, Lin and Kumar 1984), but recently results for the general heterogeneous multiserver system have appeared (Cabral 2005, de Véricourt and Zhou 2005). Note, though, that the problem (b) for the general multiserver case is still open (de Véricourt and Zhou 2006). 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 Tseytlin (2007), who is applying these concepts to patients flows in 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, Foschini (1977) shows that shortest-expected-delay-first routing is asymptotically optimal, and Teh and Ward (2002) identify necessary and sufficient conditions for a threshold priority policy to be asymptotically optimal.

Several papers have examined the question of dynamic control for the inverted-V system in the QED regime. These include Armony (2005), Tezcan (2007), Atar (2008), Atar and Shwartz (2008), Atar et al. (2011), and Armony and Ward (2010). Armony (2005) shows that FSF 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. Tezcan (2007) 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 (2008) has established that both the FSF and the longest-idle-server-first (LISF) policies exhibit a state-space reduction in the QED regime, even in settings where the service rates are random. The policy that routes to the server pool with the longest cumulative idle time has been studied in Atar et al. (2011), where fairness in idle time is shown to be obtained in the QED regime. With respect to fairness, Armony and Ward (2010) establish that if the system parameters are known, then it is asymptotically optimal to use a threshold-based routing rule, where the faster servers are kept busy when the system is most congested and are idle otherwise. Finally, Atar and Shwartz (2008) have shown that in an environment where service rates are heterogeneous and unknown, it is sufficient to sample a relatively small number of service times to come up with a routing policy that is asymptotically optimal. To our knowledge, our paper is the first to investigate asymptotically optimal routing for the inverted-V model in the ED and ED + QED regimes.

Beyond the control problem for the inverted-V system, 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 (2007b, 2010) 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 Armony (2005) 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. The results of this paper can be used to extend their results to models with customers abandonment, if abandonment rates were appropriately ordered. A more general skill-based routing model is covered in Atar et al. (2009). There, the authors establish that if service rates are pool-dependent, then the control problem is asymptotically reducible to a one-dimensional control problem, which is based on the implicit solution of an Hamilton-Jacobi-Bellman (HJB) equation. Their result is structural in nature, and as such does not reveal special structure like we do in this paper. In Gurvich and Whitt (2007a), the authors establish reduction in dimensionality 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 is a special case of the FIR policy and as such, state-space collapse and weak convergence follow (see Propositions 4.1, and Remark 4.1).

Another line of research represented by Harrison and Zeevi (2005) and Bassamboo et al. (2006a, b) considers joint 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 a service rate that depends on the pool (skill) of the particular server. Specifically, the average service time of a customer who is served by a server of skill $k$ is $1/\mu_k$, $k = 1, 2, \ldots, K$. We assume that the service rates are ordered as follows: $\mu_1 < \mu_2 < \cdots < \mu_K$. Customers arrive to the system according to a Poisson process with rate $\lambda$. Delayed customers wait in a buffer with infinite capacity. Customers are impatient. In particular, if a customer's service does not start within a time that is exponentially distributed with rate $\theta$, 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 independent.

Let $N_k$ be the number of servers in pool $k$. Also, let $\vec{N} = (N_1, N_2, \ldots, N_K)$ be the staffing vector. (Here and elsewhere, $\vec{x}$ is used to denote a vector whose elements are $x_1, x_2, \ldots$.) Let $\Pi := \Pi(\lambda, \vec{N})$ be the set of all nonpreemptive nonanticipating routing policies. Denote by $\pi \in \Pi(\lambda, \vec{N})$, a policy that operates in a system with arrival rate $\lambda$ and staffing vector $\vec{N}$. (At times we will omit the arguments $\lambda$ and $\vec{N}$ when it is clear or immaterial from the context which arguments should be used.) Given a policy $\pi \in \Pi(\lambda, \vec{N})$, let $V(t)$ be the offered (virtual) waiting time of an arbitrary, infinitely patient customer who arrives to the system at time $t$. That is, $V(t)$ is the waiting time that a customer who arrives at time $t$ would experience, if this customer never abandons. Let $V(\infty)$ be the offered waiting time in steady-state. Denote by $W := W(\infty)$ the actual waiting time in steady-state, defined as $W(\infty) := V(\infty) \wedge \tau$, where $\tau \sim \exp(\theta)$ (that is, $\tau$ has the same distribution of the time to abandonment) and is independent of $V^{\lambda}(\infty)$. Accordingly, let $P_\pi(W > T)$ be the steady-state probability that a customer is delayed more than $T$ time units (before starting service or abandoning), and let $P_\pi(ab)$ be the steady-state probability that a customer abandons.[3] Our first goal in this paper is to find a policy in $\Pi$ that minimizes the latter probability.

A more ambitious goal is to jointly identify staffing levels $N_1, \ldots, N_K$ and a routing policy to minimize staffing costs subject to a constraint on system performance (such as the probability of waiting more than $T$ 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 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 relevant* 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 the cost of staffing the system with $N_k$ servers of skill $k$ is $C_k(N_k)$. The total staffing cost is hence $C(N_1, N_2, \ldots, N_K) = C_1(N_1) + C_2(N_2) + \cdots + C_K(N_K)$. 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 first 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

$$\begin{aligned} \text{minimize} \quad & C_1(N_1) + C_2(N_2) + \cdots + C_K(N_K), \\ \text{subject to} \quad & P_\pi(ab) \leqslant \Delta, \quad \text{for some } \pi \in \Pi(\lambda, \vec{N}), \\ & N_1, N_2, \ldots, N_K \in \mathbb{Z}_+. \end{aligned} \tag{1}$$

It is implied from (1) that the decision variables are $N_1, \ldots, N_K$. (We use the convention that the bottom line in an optimization problem corresponds to the decision variables.) However, given the optimal staffing vector, one must also specify the routing policy $\pi$ that would obtain the desired abandonment probability. For the rest of the paper we consider *homogeneous* cost functions of the form $C(\vec{N}) = c_1 N_1^p + c_2 N_2^p + \cdots + c_K N_K^p$, with $p > 1$ and $c_k > 0$ for all $k$.

Suppose that the routing policy $\pi \in \Pi$ is used, and let $t \geqslant 0$ be an arbitrary time point. We denote by $Z_k(t; \pi)$

the number of busy servers of pool $k$ ($k = 1, 2, \ldots, K$) 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, $Y(t; \pi) = Z_1(t; \pi) + Z_2(t; \pi) + \cdots Z_K(t; \pi) + Q(t; \pi)$. We use $t = \infty$ whenever we refer to steady-state. At times, we omit $\pi$ if it is clear from the context which routing policy is used. Work-conserving policies will play an important role in our paper.

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, $\pi$ is work conserving if $Q(t; \pi) > 0$ implies that $Z_1(t; \pi) + Z_2(t; \pi) + \cdots + Z_K(t; \pi) = N$, where

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

is the total number of servers.

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

$$\begin{aligned} \text{minimize} \quad & P_\pi(ab), \\ & \pi \in \Pi(\lambda, \vec{N}). \end{aligned} \tag{2}$$

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

## 3. Optimal Preemptive Routing

In this section we describe a simple *preemptive* policy, $\text{FSF}_p$ ((preemptive) faster server available first (the subscript $p$ is for preemptive)), which is optimal within the set of all nonanticipating, but possibly preemptive, policies with respect to minimizing the fraction of customers who abandon. Section 4.3 will describe our proposed nonpreemptive policy, FSF, which is also simple but is not necessarily optimal for any fixed size system. However, it is *asymptotically* optimal as the system grows large according to 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 nonanticipating, possibly preemptive policies. What is meant by preemptive 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. We first show that one can restrict attention to first-come-first-served (FCFS) policies.

Proposition 3.1 (FCFS Is Optimal). *Consider the set $\Pi_{all}$ of all nonanticipating, possibly preemptive policies that are not necessarily FCFS. Then to minimize $P_\pi(ab)$ within $\Pi_{all}$, it is sufficient to consider FCFS policies.*

Let $\text{FSF}_p \in \Pi_p$ be the policy in $\Pi_p$ that is FCFS and is characterized by the following two properties: At any time point $t \geqslant 0$: (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 $Z_1(t; \text{FSF}_p) + Z_2(t; \text{FSF}_p) + \cdots + Z_K(t; \text{FSF}_p) < N$, then $Q(t; \text{FSF}_p) = 0$. The next proposition establishes the optimality of $\text{FSF}_p$ within $\Pi_p$.

Proposition 3.2 (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 $t \geqslant 0$, within the family of nonanticipating, possibly preemptive, possibly non-FCFS, and possibly not work-conserving policies. In particular, $FSF_p$ minimizes the abandonment probability $P(ab)$. (By ergodicity, $P(ab)$ is also equal to the long-time fraction of abandonment.)*

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 $E[W(\infty)]$.*

Remark 3.1 (State-Space Collapse Under $\text{FSF}_p$). Note the *state-space collapse* associated with the policy $\text{FSF}_p$. For a work-conserving policy, the state-space is generally $K$-dimensional. However, under $\text{FSF}_p$ 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 $\lambda(y) \equiv \lambda, \ \forall\, y \geqslant 0$, and a piecewise-linear death rate function:

$$\mu(y) = \begin{cases} y\mu_K & \text{if } y \leqslant N_K \\ (y - N_K)\mu_{K-1} + N_K\mu_K & \text{if } N_K < y \leqslant N_{K-1} + N_K \\ \vdots & \vdots \\ (y - (N_2 + \cdots + N_K))\mu_1 + N_2\mu_2 + \cdots + N_K\mu_K & \\ \qquad \text{if } N_2 + \cdots + N_K < y \leqslant N & \\ (y - N)\theta + N_1\mu_1 + N_2\mu_2 + \cdots + N_K\mu_K & \\ \qquad \text{if } y > N. & \end{cases} \tag{3}$$

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 4.3) and express limiting distributions (proof of Proposition 6.2) of some of the relevant scaled processes.

PROPOSITION 3.3. *Consider three systems*:

(A) *System A is the inverted-V system of this paper, with K pools of servers, service rates* $\mu_1 < \mu_2 < \cdots < \mu_K$ *and pool sizes* $N_1, N_2, \ldots, N_K$, *respectively. 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* $\mu_K$, *and* $N_1\mu_1 + N_2\mu_2 + \cdots + N_K\mu_K \geqslant N^B\mu_K$.

(C) *System C is an* $M/M/N^C + M$ *system with* $N^C$ *servers, all working with rate* $\mu_1$ *and* $N_1\mu_1 + N_2\mu_2 + \cdots + N_K\mu_K \leqslant N^C\mu_1$.

*All systems have the same arrival rate* $\lambda$ *and the same individual abandonment rate* $\theta$. *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* $Q^C \leqslant Q^A \leqslant Q^B$ *and* $N^B - Z^B \leqslant (N_1 + N_2 + \cdots + N_K) - Z^A \leqslant N^C - Z^C$, *at all times, almost surely.*

## 4. Asymptotically Optimal Control

The joint staffing and routing problem (1) is difficult to solve in a closed form. Specifically, given fixed values of $\mu_1 < \mu_2 < \cdots < \mu_K, \lambda$ and $\vec{N} = (N_1, N_2, \ldots, N_K)$, one needs to find a policy $\pi \in \Pi(\lambda, \vec{N})$ that minimizes the probability of abandonment in order to determine if the staffing vector $\vec{N}$ is feasible for Problem (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 $\lambda$ and $\vec{N}$). To this end, we consider a sequence of systems indexed by $\lambda$ (to appear as a superscript) with increasing arrival rates $\lambda \uparrow \infty$, and increasing total number of servers $N^\lambda$ but with *fixed* service rates $\mu_1, \mu_2, \ldots, \mu_K$, and a *fixed* abandonment rate $\theta$. We first solve the control problem asymptotically. We then use this solution in identifying asymptotically optimal staffing in §5.

We analyze this problem in both the ED and the QED regimes. As is often the case for this type of analysis, the ED regime is much simpler to study than its QED counterpart. Consistently, the analysis for the QED regime is much more detailed. The analysis of the ED + QED regime is deferred to §6.

### 4.1. ED Regime: Asymptotic Framework and Results

In the ED regime the number of servers of each pool $N_k^\lambda$, $k = 1, 2, \ldots, K$, grows with $\lambda$ with traffic intensity $\rho^\lambda$ that satisfies

$$\rho^\lambda := \frac{\lambda}{\sum_{k=1}^K \mu_k N_k^\lambda} \to \rho > 1, \quad \text{as } \lambda \to \infty. \tag{4}$$

In particular, the system is overloaded. Under these circumstances, flow conservation considerations, as in Whitt (2004), establish that under any work-conserving policy (including the policy $\text{FSF}_P$)

$$P^\lambda(ab) = 1 - \frac{\sum_{k=1}^K \mu_k N_k^\lambda}{\lambda} \to \frac{\rho - 1}{\rho}, \quad \text{as } \lambda \to \infty. \tag{5}$$

Let $\vec{N}^\lambda$ be the staffing vector of the $\lambda$-system. Then, we define the *ED scaled* version of the routing problem (2) to be

$$\begin{aligned} \text{minimize} \quad & \limsup_{\lambda \to \infty} P^\lambda_{\pi^\lambda}(ab), \\ & \pi^\lambda \in \Pi(\lambda, \vec{N}^\lambda). \end{aligned} \tag{6}$$

In light of the optimality of $\text{FSF}_P$ and the fact that $P^\lambda(ab)$ is asymptotically the same for all work-conserving policies (5), we conclude that *all work-conserving policies are asymptotically optimal* with respect to (6) in the ED regime.

### 4.2. QED Regime: Asymptotic Framework

In contrast to the ED regime in the QED regime the traffic intensity is assumed to converge to 1 (see (7) below). In addition, in this regime one needs to carefully define the limiting proportions of the size of the various server pools. Assume that there are $K$ numbers $a_k \geqslant 0$, $k = 1, \ldots, K$, with $a_1 > 0$ and $\sum_{k=1}^K a_k = 1$, such that the number of servers of each pool $N_k^\lambda$, $k = 1, 2, \ldots, K$, grows with $\lambda$ as follows:

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

$$\lim_{\lambda \to \infty} \frac{\mu_k N_k^\lambda}{\lambda} = a_k. \tag{A1}$$

Condition (A1) guarantees that the total traffic intensity,

$$\rho^\lambda := \frac{\lambda}{\sum_{k=1}^K \mu_k N_k^\lambda}, \tag{7}$$

converges to 1, as $\lambda \to \infty$, and hence, for large $\lambda$, 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

$$\mu := \left[ \sum_{k=1}^K a_k/\mu_k \right]^{-1}; \tag{8}$$

then $\lambda/\mu$ can be interpreted as the total offered load for the whole system. Given this definition of $\mu$, (A1) implies that the total number of servers $N^\lambda := \sum_{k=1}^K N_k^\lambda$ satisfies

$$N^\lambda = \frac{\lambda}{\mu} + o(\lambda), \quad \text{as } \lambda \to \infty, \quad \text{or,} \quad \lim_{\lambda \to \infty} \frac{\lambda}{N^\lambda} = \mu. \tag{9}$$

Also,

$$\rho^\lambda \approx \frac{\lambda}{N^\lambda \mu}, \tag{10}$$

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

$$q_k := \lim_{\lambda\to\infty} \frac{N_k^\lambda}{N^\lambda} = \frac{a_k}{\mu_k}\mu \geqslant 0, \quad k = 1, \ldots, K, \tag{11}$$

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 nonnegligible in size. Clearly, $\sum_{k=1}^K q_k = 1$ and $\sum_{k=1}^K q_k\mu_k = \mu$.

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 $N^\lambda$. Therefore, one expects to get finite limits of these quantities when dividing all of them by $N^\lambda$. As it turns out, due to the functional strong law of large numbers (FSLLN), this scaling leads to the fluid dynamics of the system, in the limit as $\lambda\to\infty$. The details are provided in §A of the online companion. In addition to the fluid scaling, we introduce a more refined *diffusion* scaling.

**Diffusion Scaling:** For $\lambda > 0$ and any fixed sequence of *work-conserving* policies $\pi^\lambda \in \Pi(\lambda, N^\lambda)$ (omitted from the notation), define the centered and scaled process $\vec{X}^\lambda(\cdot) = (X_1^\lambda(\cdot), \ldots, X_K^\lambda(\cdot))$ as follows:

$$X_1^\lambda(t) := \frac{Q^\lambda(t) + Z_1^\lambda(t) - N_1^\lambda}{\sqrt{N^\lambda}}, \tag{12}$$

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

$$X_k^\lambda(t) := \frac{Z_k^\lambda(t) - N_k^\lambda}{\sqrt{N^\lambda}}. \tag{13}$$

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

$$X^\lambda(t) := \sum_{k=1}^K X_k^\lambda(t) = \frac{Q^\lambda(t) + \sum_{k=1}^K Z_k^\lambda(t) - N^\lambda}{\sqrt{N^\lambda}} = \frac{Y^\lambda(t) - N^\lambda}{\sqrt{N^\lambda}}. \tag{14}$$

From work conservation it follows that $[X^\lambda(t)]^-$ is the total number of idle servers, and $[X^\lambda(t)]^+ = [X_1^\lambda(t)]^+$ is the queue length, both scaled by $1/\sqrt{N^\lambda}$. Finally, note that if $X_k^\lambda(t) < 0$ for some $k$, then $X_1^\lambda(t) \leqslant 0$.

For all $\lambda > 0$, the scaled *offered* waiting time process is defined as $\hat{V}^\lambda(t) = \sqrt{N^\lambda} V^\lambda(t)$, $t \geqslant 0$, $\lambda > 0$, and its scaled steady-state is $\hat{V}^\lambda(\infty) = \sqrt{N^\lambda} V^\lambda(\infty)$, $\lambda > 0$. Finally, the scaled steady-state *actual* waiting time is $\hat{W}^\lambda(\infty) = \sqrt{N^\lambda} W^\lambda(\infty)$, $\lambda > 0$.

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

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

$$\lim_{\lambda\to\infty} \frac{\sum_{k=1}^K \mu_k N_k^\lambda - \lambda}{\sqrt{\lambda}} = \delta \tag{A2}$$

for some $\delta \in (-\infty, \infty)$.

Condition (A2) is a square-root staffing rule (similar to Halfin and Whitt 1981 and Borst et al. 2003). As shown later (Corollary 4.2), it guarantees that under the appropriate routing, the fraction of abandonment is of the order of $1/\sqrt{\lambda}$. For $k = 1, \ldots, K$, and $\lambda > 0$, let $-\infty < \delta_k^\lambda < \infty$ be defined as: $\delta_k^\lambda := (\mu_k N_k^\lambda - a_k\lambda)/\sqrt{\lambda}$. 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 $\delta_k^\lambda \geqslant 0$ if $a_k = 0$, $\delta_k^\lambda = o(\sqrt{\lambda})$, as $\lambda\to\infty$, $\forall k = 1, \ldots, K$, and $\delta^\lambda := \sum_{k=1}^K \delta_k^\lambda \to \delta$, as $\lambda\to\infty$. Note that we do not require the individual sequences $\{\delta_k^\lambda\}_{\lambda>0}$ to have a limit for any value of $k = 1, \ldots, K$. All that is assumed is that their sum converges to $\delta$, and the following additional condition is also assumed to hold:

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

Let $\vec{N}^\lambda$ be the staffing vector of the $\lambda$-system. Then, we define the QED *scaled* version of the routing problem (2) to be

$$\begin{aligned} \text{minimize} \quad & \limsup_{\lambda\to\infty} \sqrt{\lambda}\, P_{\pi^\lambda}^\lambda(ab), \\ & \pi^\lambda \in \Pi(\lambda, \vec{N}^\lambda). \end{aligned} \tag{15}$$

Corollary 4.2 justifies the above as the sensible scaling of $P^\lambda(ab)$ under Conditions (A1) and (A2).

### 4.3. QED Regime: Faster Server First (FSF) is Asymptotically Optimal

The preemptive policy FSF can be described 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. Judging by the literature on the slow server problem (e.g., Lin and Kumar 1984), this policy is not likely to be optimal. However, as we show in this section, it is *asymptotically* optimal as the arrival rate $\lambda$ grows to $\infty$ 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 within the family of non-preemptive nonanticipating policies.

THEOREM 4.1 (ASYMPTOTIC OPTIMALITY OF FSF). *Consider a sequence of systems indexed by the arrival rate $\lambda$, that satisfy conditions* (A1) *and* (A2). *Then the non-preemptive policy FSF is asymptotically optimal with respect to* (15) *within the family $\Pi_p$ of all possibly preemptive, nonanticipating policies. In particular, it is asymptotically optimal with respect to* (15) *within the family $\Pi$ of non-preemptive, nonanticipating policies.*

COROLLARY 4.1. *Consider a sequence of systems indexed by the arrival rate $\lambda$ that satisfy Conditions* (A1) *and* (A2). *Then the non-preemptive policy FSF is asymptotically optimal with respect to the minimization of* $\limsup_{\lambda\to\infty} E\widehat{Q}^{\lambda}(\infty;\pi^{\lambda})$ *and* $\limsup_{\lambda\to\infty} E\widehat{W}^{\lambda}(\infty;\pi^{\lambda})$.

To prove the asymptotic optimality of FSF as $\lambda\to\infty$, we will show that as $\lambda$ grows, the process $(X_1^{\lambda}(\cdot), X_2^{\lambda}(\cdot), \ldots, X_K^{\lambda}(\cdot))$ (recall the diffusion scaling in §4.2) under FSF becomes close to the same process under the preemptive policy $\text{FSF}_p$; and in the limit, as $\lambda\to\infty$, the two processes coincide. Taking the limits as $t\to\infty$, we will also show that the corresponding steady-state processes become close, and hence the optimality of $\text{FSF}_p$ 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 $(X_1^{\lambda}(\cdot), X_2^{\lambda}(\cdot), \ldots, X_K^{\lambda}(\cdot))$ under FSF, into a one-dimensional process, as $\lambda\to\infty$. Recall that such state-space collapse holds for every $\lambda$ under $\text{FSF}_p$ (by Proposition 3.2 and Remark 3.1). When FSF is used, however, this is no longer true, but the state-space collapse is attained when $\lambda\to\infty$, as will be shown in Proposition 4.1 below.

**4.3.1. State-Space Collapse.** In this section we establish the state-space collapse result with respect to the policy FSF and the process $\vec{X}^{\lambda}(\cdot) = (X_1^{\lambda}(\cdot), \ldots, X_K^{\lambda}(\cdot))$. Because the policy here is fixed, we omit FSF from all notation. Essentially, the state-space collapse result indicates that as $\lambda$ grows, the one-dimensional process $X^{\lambda}(\cdot)$ (see (14)) becomes sufficient in describing the whole $K$-dimensional process $\vec{X}^{\lambda}(\cdot)$. Specifically, we show that as $\lambda\to\infty$, all the faster servers (from pools $k = 2, \ldots, K$) 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 $\lambda$ grows, the processes $X_2^{\lambda}(\cdot), \ldots, X_K^{\lambda}(\cdot)$ become negligible, while the processes $X^{\lambda}(\cdot)$ and $X_1^{\lambda}(\cdot)$ are close. This result is presented in the following proposition.

PROPOSITION 4.1 (STATE-SPACE COLLAPSE). *Suppose that conditions* (A1) *and* (A2) *hold and that the work-conserving non-preemptive policy FSF is used. In addition, suppose that $\vec{X}^{\lambda}(0) \to \vec{X}(0) = \vec{x} = (x_1, \ldots, x_K)$, in probability, as $\lambda\to\infty$. Then, $X_k^{\lambda}(\cdot) \xrightarrow{p} 0$, uniformly on compact intervals, as $\lambda\to\infty$, $\forall\, k \geqslant 2$.*

REMARK 4.1 (STATE-SPACE COLLAPSE FOR $\text{FSF}_p$ IN THE QED REGIME). Proposition 4.1 is also true if the preemptive policy $\text{FSF}_p$ is used. Here the proof is much simpler.

We note that the state-space collapse result of Proposition 4.1 essentially shows that to describe the limiting behavior of the scaled process $\vec{X}^{\lambda}$ it is sufficient to find the diffusion limit of the total customer count (centered and scaled) $X^{\lambda}$. Denoting this limit by $X$, we have that the limit of $X_k^{\lambda}$, for $k \geqslant 2$, is identically zero, and the limit of $X_1^{\lambda}$ is hence equal to $X$.

**4.3.2. Stationary Diffusion Limit.** In this section we prove that the stationary distributions of the process $\vec{X}^{\lambda}$, under both $\text{FSF}_p$ and FSF, converge to the stationary distribution of its diffusion limit $\vec{X}$, as $\lambda\to\infty$. Specifically, we first spell out the stationary distribution of $X = \sum_{k=1}^{K} X_k$. Next we show that the stationary distribution of $X^{\lambda}$ under both $\text{FSF}_p$ and FSF converges to this stationary distribution of $X$. In all processes we use $\infty$ in place of the time argument to denote steady-state.

PROPOSITION 4.2 (STATIONARY DISTRIBUTION OF THE DIFFUSION PROCESS). *Let $X(\cdot)$ be a diffusion process, with infinitesimal drift*

$$m(x) = \begin{cases} -\delta\sqrt{\mu} - \theta x & x \geqslant 0, \\ -\delta\sqrt{\mu} - \mu_1 x & x < 0, \end{cases} \tag{16}$$

*and infinitesimal variance*

$$\sigma^2(x) = 2\mu. \tag{17}$$

*Then the steady-state distribution of $X$ has the density $f(\cdot)$ given by*

$$f(x) = \begin{cases} \dfrac{\sqrt{\theta/\mu}\,\phi(\sqrt{\theta/\mu}\,x + \delta/\sqrt{\theta})}{1 - \Phi(\delta/\sqrt{\theta})}\,\alpha, & \text{if } x \geqslant 0, \\ \dfrac{\sqrt{\mu_1/\mu}\,\phi(\sqrt{\mu_1/\mu}\,x + \delta/\sqrt{\mu_1})}{\Phi(\delta/\sqrt{\mu_1})}\,(1-\alpha), & \text{if } x < 0, \end{cases} \tag{18}$$

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

$$EX^{+}(\infty) = \alpha\left[\frac{-\delta\sqrt{\mu}}{\theta} + \sqrt{\frac{\mu}{\theta}}\,h\left(\frac{\delta}{\sqrt{\theta}}\right)\right], \tag{19}$$

*and*

$$EX^{-}(\infty) = (1-\alpha)\left[\frac{\delta\sqrt{\mu}}{\mu_1} + \sqrt{\frac{\mu}{\mu_1}}\,h\left(-\frac{\delta}{\sqrt{\mu_1}}\right)\right]. \tag{20}$$

Proposition 4.3 (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 $\vec{X}^\lambda$ weakly converges, as $\lambda\to\infty$, to the stationary distribution of $\vec{X}=(X,0,\ldots,0)$, where the stationary distribution of the first coordinate, $X$, is given in* (18). *Moreover, the stationary distribution of the scaled virtual waiting time, $\hat{V}^\lambda$, weakly converges to the stationary distribution of $[X]^+/\mu$.*

Corollary 4.2. *Suppose that Conditions* (A1) *and* (A2) *hold and that either FSF or FSF$_p$ is used. Then*

$$\lim_{\lambda\to\infty} E\hat{Q}^\lambda(\infty) := \lim_{\lambda\to\infty} E\frac{Q^\lambda(\infty)}{\sqrt{N^\lambda}} = EX^+(\infty), \tag{21}$$

$$\lim_{\lambda\to\infty} E\widehat{W}^\lambda(\infty) := \lim_{\lambda\to\infty} E\sqrt{N^\lambda}\widehat{W}(\infty) = EX^+(\infty)/\mu, \tag{22}$$

$$\lim_{\lambda\to\infty} \hat{P}^\lambda(ab) := \lim_{\lambda\to\infty} \sqrt{N^\lambda}P^\lambda(ab) = \theta EX^+(\infty)/\mu, \tag{23}$$

*and*

$$\lim_{\lambda\to\infty} \sqrt{\lambda}P^\lambda(ab) = \frac{\theta}{\sqrt{\mu}}EX^+(\infty) = \sqrt{\theta}\alpha\cdot\left[h\left(\frac{\delta}{\sqrt{\theta}}\right)-\frac{\delta}{\sqrt{\theta}}\right], \tag{24}$$

*where $EX^+(\infty)$ is given in* (19).

From Corollary 4.2 it follows that under Conditions (A1) and (A2), $\sqrt{\lambda}P^\lambda(ab)$ converges to a well-defined limit, as $\lambda\to\infty$, under the FSF policy. In particular, $\lim_{\lambda\to\infty} P^\lambda(ab)=0$, which implies that the constraint in (1) is satisfied trivially in the limit. Therefore, under the QED asymptotic framework, the sensible optimization problem to focus on is (15).

## 5. Asymptotically Optimal Staffing

### 5.1. Asymptotic Feasibility

In this section, we wish to characterize the feasible region for the staffing problem (1). As noted before, characterizing this region exactly for fixed $\lambda, \mu_1, \ldots, \mu_K, \theta$, and $\Delta$ is difficult. Instead, we characterize this region asymptotically for large values of $\lambda$ (i.e., as $\lambda\to\infty$).

It is interesting to note that in both the ED and the QED regimes, the asymptotically feasible region is characterized by a simple *linear* asymptotic inequality that is a function of the staffing vectors through the corresponding total service capacity. 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 individual capacities of the various servers pools through the overall service capacity. Moreover, the latter is a linear function of the number of servers in each pool.

**5.1.1. ED Regime: Asymptotic Feasibility.** For given values of the parameters $\mu_1<\mu_2<\cdots<\mu_K$, and $\theta$ and a given value of $0<\Delta<1$, a sequence $\{\vec{N}^\lambda\}$ of staffing vectors, for a sequence of systems indexed by their arrival rate $\lambda$, is called *asymptotically feasible* if there exists a sequence of routing policies $\pi^\lambda\in\Pi(\lambda,\vec{N}^\lambda)$ such that

$$\limsup_{\lambda\to\infty} P^\lambda_{\pi^\lambda}(ab)\leqslant\Delta. \tag{25}$$

Proposition 5.1 characterizes the asymptotically feasible region for the ED regime.

Proposition 5.1 (Asymptotically Feasible Region: the ED Regime). *Let $0\leqslant\Delta\leqslant 1$, and let $0<\mu_1<\mu_2<\cdots<\mu_K$ and $\theta$ be fixed. Consider a sequence of systems indexed by the arrival rate $\lambda>0$, with $\lambda$ growing to infinity and $N_k^\lambda$ servers in pool $k$, $k=1,\ldots,K$. Let $N^\lambda=\sum_{k=1}^K N_k^\lambda$ be the total number of servers in system $\lambda$. Then there exists a sequence $\{\pi^\lambda\}$, with $\pi^\lambda\in\Pi(\lambda,\vec{N}^\lambda)\}$, of non-preemptive policies, under which*

$$\limsup_{\lambda\to\infty} P^\lambda_{\pi^\lambda}(ab)=\Delta \tag{26}$$

*if and only if*

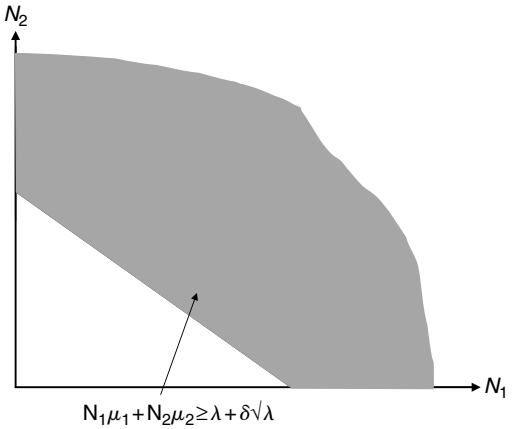
$$\mu_1 N_1^\lambda+\cdots+\mu_K N_K^\lambda\geqslant\lambda(1-\Delta)+o(\lambda),\quad \text{as } \lambda\to\infty. \tag{27}$$

**5.1.2. QED Regime: Asymptotic Feasibility.** For given values of the parameters $\mu_1<\mu_2<\cdots<\mu_K$, and $\theta$ 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 $\lambda$, is called *asymptotically feasible* if there exists a sequence of routing policies $\pi^\lambda\in\Pi(\lambda,\vec{N}^\lambda)$ such that

$$\limsup_{\lambda\to\infty} \sqrt{\lambda}P^\lambda_{\pi^\lambda}(ab)\leqslant\Delta.^4 \tag{28}$$

Proposition 5.2 characterizes the asymptotically feasible region, which is illustrated in Figure 2.[5]

**Figure 2.** The asymptotically feasible region for $K=2$ in the QED regime.



PROPOSITION 5.2 (ASYMPTOTICALLY FEASIBLE REGION-SQUARE-ROOT "SAFETY" CAPACITY). *Let $0 < \Delta < \infty$, and let $0 < \mu_1 < \mu_2 < \cdots < \mu_K$ and $\theta$ be fixed. Consider a sequence of systems indexed by the arrival rate $\lambda > 0$, with $\lambda$ growing to infinity and $N_k^\lambda$ servers in pool $k$, $k = 1, \ldots, K$. Let $N^\lambda = \sum_{k=1}^K N_k^\lambda$ be the total number of servers in system $\lambda$, and assume that*

$$\liminf_{\lambda\to\infty} \frac{N_1^\lambda}{N^\lambda} > 0. \tag{29}$$

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

$$\limsup_{\lambda\to\infty} \sqrt{\lambda} P_{\pi^\lambda}(ab) = \Delta \tag{30}$$

*if and only if*

$$\mu_1 N_1^\lambda + \cdots + \mu_K N_K^\lambda \geqslant \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda}), \quad \text{as } \lambda\to\infty, \tag{31}$$

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

$$\Delta := \Delta(\delta, \mu_1, \theta) = \sqrt{\theta}\ \alpha \cdot \left[ h\left(\frac{\delta}{\sqrt{\theta}}\right) - \frac{\delta}{\sqrt{\theta}} \right], \tag{32}$$

*with $\alpha := \alpha(\delta, \mu_1, \theta) = [1 + \sqrt{\theta} h(\delta/\sqrt{\theta})/(\sqrt{\mu_1} h(-\delta/\sqrt{\mu_1}))]^{-1} = P\{X(\infty) \geqslant 0\}$.*
*In addition, $\delta = -\infty$ (i.e., (31) is violated for all $\delta > -\infty$) if and only if $\Delta = \infty$, and $\delta = \infty$ (i.e., (31) holds for any $\delta > 0$) if and only if (30) holds for any arbitrary $\infty > \Delta > 0$, with the appropriate choice of $\pi^\lambda$.*

## 5.2. Asymptotically Optimal Staffing

In this section, we study the staffing problem (1). Recall that exact optimality is difficult to obtain, and hence we present asymptotically optimal solutions. Our results so far have identified an asymptotically optimal routing policy and the asymptotically feasible regions given in (27) and (31) for the ED and QED regimes, respectively. We now turn to finding an asymptotically optimal staffing rule within the asymptotically feasible region.

**5.2.1. Asymptotic Optimality Definition.** Before stating our proposed solution for the asymptotic staffing problem, the notion of asymptotic optimality needs to be defined. Consider a homogeneous staffing cost function of the form $C(\vec{N}) = C_1 N_1^p + \cdots + C_K N_K^p$, where $p > 1$. In view of the characterization of the feasible region given in (27) and (31), it is expected that the optimal staffing cost will be of the order of $\lambda^p$. In particular, defining asymptotically optimality in terms of the cost ratio between two asymptotically feasible staffing vectors is too crude a criterion, in the sense that many plausible staffing vectors would satisfy it. Hence, 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 $\lambda^p$. One may refer to this criterion as *second-order* asymptotic optimality.

To obtain this basic cost, consider the following auxiliary staffing problem. For a positive constant $x$, let

$$\begin{aligned} \text{minimize} \quad & C_1(N_1) + C_2(N_2) + \cdots + C_K(N_K), \\ \text{subject to} \quad & \mu_1 N_1 + \mu_2 N_2 + \cdots + \mu_K N_K \geqslant x \\ & N_1, N_2, \ldots, N_K \geqslant 0. \end{aligned} \tag{SP}$$

We will use the notation SP($x$) to underline the dependence of the problem (SP) on the parameter $x$.

In the QED regime, in light of Proposition 5.2, and particularly the relationship (31), the problem SP($\lambda$) 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

$$\frac{C_k'(N_k^*)}{\mu_k} = \frac{C_j'(N_j^*)}{\mu_j}, \quad j, k = 1, 2, \ldots, K, \tag{33}$$

and

$$\mu_1 N_1 + \mu_2 N_2 + \cdots + \mu_K N_K = \lambda. \tag{34}$$

Let $\underline{C}^\lambda$ be the optimal cost obtained by solving SP($\lambda$). Then, $\underline{C}^\lambda$ will serve as the normalizing factor in this regime.

DEFINITION (ASYMPTOTICALLY OPTIMAL STAFFING IN THE QED REGIME). Suppose that $\{\vec{N}^{*\lambda}\}_{\lambda>0}$ is a sequence of optimal solutions of the sequence of *scaled* staffing problems:

$$\begin{aligned} \text{minimize} \quad & C_1(N_1^\lambda) + C_2(N_2^\lambda) + \cdots + C_K(N_K^\lambda), \\ \text{subject to} \quad & P_{\pi^\lambda}^\lambda(ab) \leqslant \Delta/\sqrt{\lambda}, \quad 0 < \Delta < \infty \\ & \qquad \text{for some } \pi^\lambda \in \Pi(\lambda, \vec{N}^\lambda), \\ & N_1^\lambda, N_2^\lambda, \ldots, N_K^\lambda \in \mathbb{Z}_+, \end{aligned} \tag{35}$$

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

(a) there exists a sequence of policies $\{\pi^\lambda\}$, with $\pi^\lambda \in \Pi(\lambda, \widetilde{N}^\lambda)$ such that $\limsup_{\lambda\to\infty} \sqrt{\lambda} P_{\pi^\lambda}(ab) \leqslant \Delta$, and
(b) $\lim_{\lambda\to\infty} (C(\vec{N}^{*\lambda}) - \underline{C}^\lambda)/(C(\widetilde{N}^\lambda) - \underline{C}^\lambda) = 1$.

Condition (b) requires that the proposed staffing does not only match the optimal solution at the fluid scale but also at the more refined normalized scale defined by $|C(\vec{N}^{*\lambda}) - \underline{C}^\lambda|$.

For the ED regime, the fluid scale problem is given by SP($\lambda(1-\Delta)$). The solution of this problem coincides with

our proposed solution in this regime. Hence, normalizing the cost by the optimal cost of the fluid scale problem is not meaningful in this case. Instead, for the ED regime we establish a form of asymptotic optimality that is expressed in terms of the *distance* of the proposed solution from the true optimal solution.

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

$$\begin{aligned} \text{minimize} \quad & C_1(N_1^\lambda) + C_2(N_2^\lambda) + \cdots + C_K(N_K^\lambda), \\ \text{subject to} \quad & P^\lambda_{\pi^\lambda}(ab) \leqslant \Delta, \quad 0 < \Delta < 1, \\ & \qquad \text{for some } \pi^\lambda \in \Pi(\lambda, \vec{N}^\lambda), \\ & N_1^\lambda, N_2^\lambda, \cdots, N_K^\lambda \in \mathbb{Z}_+, \end{aligned} \tag{36}$$

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

(a) there exists a sequence of policies $\{\pi^\lambda\}$, with $\pi^\lambda \in \Pi(\lambda, \widetilde{N}^\lambda)$ such that $\limsup_{\lambda\to\infty} P_{\pi^\lambda}(\text{ab}) \leqslant \Delta$, and

(b) $\|\vec{N}^{*\lambda} - \widetilde{N}^\lambda\| = o(\lambda)$.

Note that this definition is consistent with the asymptotically optimal definition used for this regime in Mandelbaum and Zeltyn (2009).

Remark 5.1. One would expect that for the QED regime a similar distance-based definition for asymptotically optimal staffing would work with (b) above replaced by $\|\vec{N}^{*\lambda} - \widetilde{N}^\lambda\| = o(\sqrt{\lambda})$. However, we were unable to identify a simple staffing rule that would satisfy this criterion in that regime. Nevertheless, the criterion $\|\vec{N}^{*\lambda} - \widetilde{N}^\lambda\| = o(\lambda)$ can be shown to be satisfied in the QED regime, as well as the ED + QED regime studied in §6.

**5.2.2. ED Regime: Asymptotically Optimal Staffing.** Fix $p > 1$, $c_k > 0$, $k = 1, \ldots, K$ and $0 < \Delta < 1$ and consider the problem (36) with respect to the cost function $C(\vec{N}^\lambda) = c_1(N_1^\lambda)^p + \cdots + c_K(N_K^\lambda)^p$. Let $\vec{M}^{*\lambda}$ be an optimal solution of the problem $\text{SP}(\lambda(1-\Delta))$. Note that the vector $\vec{M}^{*\lambda}$ is not necessarily all integers, and let $\widetilde{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil := (\lceil M_1^{*\lambda} \rceil, \ldots, \lceil M_K^{*\lambda} \rceil)$; that is, $\widetilde{N}^\lambda$ is obtained from $\vec{M}^{*\lambda}$ by rounding its elements to the closest integers from above. We claim that $\widetilde{N}^\lambda$ is an asymptotically optimal staffing vector in the ED regime. Solving for $\vec{M}^{*\lambda}$ is illustrated in Figure 3.

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

Proposition 5.3. *Consider a fixed target scaled abandonment probability of* $\Delta \in (0,1)$. *Suppose that* $C(\vec{N}) = c_1 N_1^p + \cdots + c_K N_K^p$, $p > 0$ *and for* $\lambda > 0$ *consider the staffing vector* $\widetilde{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil$, *where* $\vec{M}^{*\lambda} = (M_1^{*\lambda}, \ldots, M_K^{*\lambda})$ *is an optimal solution to* $SP(\lambda(1-\Delta))$. *Then*

$$\vec{M}^{*\lambda} = \begin{cases} (\lambda(1-\Delta)) \dfrac{((\mu_1/c_1)^{1/(p-1)}, (\mu_2/c_2)^{1/(p-1)}, \ldots, (\mu_K/c_K)^{1/(p-1)})}{\sum_{k=1}^K (\mu_k^p/c_k)^{1/(p-1)}}, \\ \qquad\qquad p > 1, \\ \dfrac{\lambda(1-\Delta)}{\mu_k} e_k, \quad k = \min\{\text{argmax}_{k=1,\ldots,K}\{c_k/(\mu_k)^p\}\}, \\ \qquad\qquad 0 < p \leqslant 1, \end{cases} \tag{37}$$

*where* $e_k$ *is a* $K$*-dimensional vector with the number* 1 *is the* $k$*th position and* 0 *otherwise. Let* $\vec{N}^{*\lambda}$ *be a sequence of optimal solutions to* (35). *Then*

$$\lim_{\lambda\to\infty} \frac{|C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda})|}{\lambda^p} = 0, \tag{38}$$

*which also implies that*

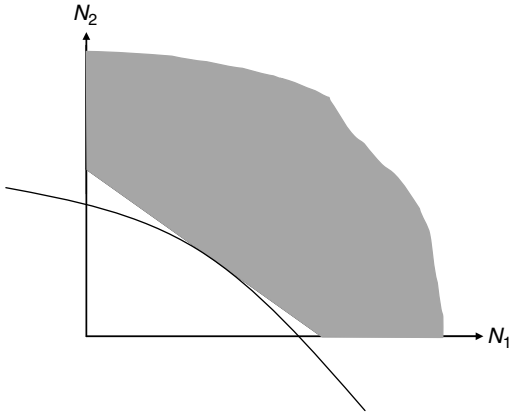
$$\lim_{\lambda\to\infty} \frac{|C(\vec{N}^{*\lambda}) - C(\widetilde{N}^\lambda)|}{\lambda^p} = 0. \tag{39}$$

Note that in the above proposition we consider all values of $p > 0$. For the asymptotic optimality proposition below, we need to restrict our attention to strictly convex cost functions, i.e., $p > 1$.

Proposition 5.4 (Asymptotically Optimal Staffing in the ED Regime). *Consider a fixed target abandonment probability of* $\Delta \in (0,1)$. *Suppose that* $C(\vec{N}) = c_1 N_1^p + \cdots + c_K N_K^p$, $p > 1$, *and for* $\lambda > 0$ *consider the staffing vector* $\widetilde{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil$, *where* $\vec{M}^{*\lambda} = (M_1^{*\lambda}, \ldots, M_K^{*\lambda})$ *is an optimal solution to* $SP(\lambda(1-\Delta))$ *given in* (37). *Then* $\{\widetilde{N}^\lambda\}_{\lambda>0}$ *is an asymptotically optimal staffing sequence in the ED regime with respect to* (36).

**5.2.3. QED Regime: Asymptotically Optimal Staffing.** Fix $p > 1$, $c_k > 0$, $k = 1, \ldots, K$ and $0 < \Delta < \infty$ and consider the problem (35) with respect to the cost functions $C_k(N_k^\lambda) = c_k(N_k^\lambda)^p$. Let $-\infty < \delta < \infty$ be such

**Figure 3.** 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 SP($\lambda + \delta\sqrt{\lambda}$). Let $\widetilde{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil := (\lceil M_1^{*\lambda} \rceil, \ldots, \lceil M_K^{*\lambda} \rceil)$. We show that $\widetilde{N}^\lambda$ is asymptotically optimal in the QED regime.

The following proposition is analogous to Proposition 5.3 with respect to the QED regime. It shows that the difference in cost between the optimal solution and our proposed solution cannot be too large. In fact, the distance between these two vectors is even smaller than the distance between the corresponding vectors in the ED regime. This is to be expected because of the more refined scaling in this regime.

Proposition 5.5. *Consider a fixed target scaled abandonment probability of $\Delta \in (0, \infty)$. Suppose that $C(\vec{N}) = c_1 N_1^p + \cdots + c_K N_K^p$, and for $\lambda > 0$ consider the staffing vector $\widetilde{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil$, where $\vec{M}^{*\lambda} = (M_1^{*\lambda}, \ldots, M_K^{*\lambda})$ is an optimal solution to SP($\lambda + \delta\sqrt{\lambda}$). Here $\delta$ satisfies $\Delta = \Delta(\delta, \mu_1, \theta)$ (see* (32)*), and*

$$\vec{M}^{*\lambda} = (\lambda + \delta\sqrt{\lambda}) \frac{((\mu_1/c_1)^{1/(p-1)}, (\mu_2/c_2)^{1/(p-1)}, \ldots, (\mu_K/c_K)^{1/(p-1)})}{\sum_{k=1}^K (\mu_k^p/c_k)^{1/(p-1)}} \quad \lambda > 0. \tag{40}$$

*Let $\vec{N}^{*\lambda}$ be a sequence of optimal solutions to* (35)*. Then*

$$\lim_{\lambda \to \infty} \frac{|C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda})|}{\lambda^{p-1/2}} = 0, \tag{41}$$

*which also implies that*

$$\lim_{\lambda \to \infty} \frac{|C(\vec{N}^{*\lambda}) - C(\widetilde{N}^\lambda)|}{\lambda^{p-1/2}} = 0. \tag{42}$$

Proposition 5.6 (Asymptotically Optimal Staffing in the QED Regime). *Consider a fixed target scaled abandonment probability of $\Delta \in (0, \infty)$. Suppose that $C(\vec{N}) = c_1 N_1^p + \cdots + c_K N_K^p$, and for $\lambda > 0$ consider the staffing vector $\widetilde{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil$, where $\vec{M}^{*\lambda} = (M_1^{*\lambda}, \ldots, M_K^{*\lambda})$ is an optimal solution to SP($\lambda + \delta\sqrt{\lambda}$). Here $\delta$ satisfies $\Delta = \Delta(\delta, \mu_1, \theta)$ (see* (32)*), and $\vec{M}^{*\lambda}$ is as given in* (40)*. Then, if $\delta \neq 0$, $\{\widetilde{N}^\lambda\}_{\lambda > 0}$ is an asymptotically optimal staffing sequence in the QED regime with respect to* (35)*.*

Remark 5.2. Proposition 5.6 excludes the case $\delta = 0$. When $\delta = 0$ the difference between $C(\vec{N}^{*\lambda})$ and $\underline{C}^\lambda$ 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 5.5, then asymptotic optimality extends to the case $\delta = 0$.

## 6. Constraints on the Waiting Time Distribution

Our discussion thus far was focused on solving the problem of minimizing staffing costs subject to an upper bound on the fraction of abandonment. Another commonly used constraint is an upper bound on the tail probability of the waiting time distribution. Specifically, call centers are often interested in the following problem version:

$$\begin{aligned} \text{minimize} \quad & C_1(N_1) + C_2(N_2) + \cdots + C_K(N_K), \\ \text{subject to} \quad & P_\pi(W > T) \leqslant \alpha, \\ & \qquad \text{for some } \pi \in \Pi(\lambda, \vec{N}), \\ & N_1, N_2, \ldots, N_K \in \mathbb{Z}_+, \end{aligned} \tag{43}$$

for some fixed $T \geqslant 0$ and $\alpha \in (0, 1)$. In this section, we solve (43) and explore its similarities and differences with the model formulation (1).

**Preemptive Routing.** In §3 we established the optimality of the policy FSF$_P$ with respect to the steady-state fraction of abandonment among all nonanticipating, possibly preemeptive routing policies. In particular, while FSF$_P$ is FCFS and work conserving, no such restrictions were imposed on the set of admissible policies. In contrast, if one wishes to establish that FSF$_P$ stochastically minimizes the steady-state waiting time distribution, some restrictions are needed. Otherwise, one might be led to favor policies that do not make sense service-wise.

To elaborate, the FCFS assumption is necessary for the optimality of FSF$_P$. If this assumption is removed, it is optimal to give priority to customers who have already waited $T$ time units or more. Similarly, if the abandonment rate is large, work-conservation becomes necessary for the optimality of FSF$_P$. In its absence, it makes sense to, at times, idle servers while customers are in queue so that those customers would abandon quickly, instead of being served slowly. It turns out that if $\theta \leqslant \min\{\mu_k,\ k = 1, \ldots, K\} = \mu_1$, then such intentional idling is not beneficial. In words, the assumption $\theta \leqslant \min\{\mu_k,\ k = 1, \ldots, K\} = \mu_1$ implies that average patience exceeds all average service times, which is indeed what has been observed in practice; see, for example, Tables 36 and 49 in Mandelbaum et al. (2000).

To conclude this discussion, we state the following proposition, which is an analog of Proposition 3.2.

Proposition 6.1 (Optimal Preemptive Routing). *Consider the preemptive routing policy, FSF$_p$. 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 nonanticipating, possibly preemptive policies, which are* also FCFS and work conserving. *In other words, for all $\pi \in \Pi_p$, if $\pi$ is also FCFS and work conserving, then assuming both systems start in the same state at time* 0*, we have $P\{Y(t; \pi) > y\} \geqslant P\{Y(t; FSF_p) > y\}$ for all $y \geqslant 0$ and all $0 \leqslant t \leqslant \infty$. Alternatively, if*

$$\theta \leqslant \min\{\mu_k, k = 1, \ldots, K\}, \tag{44}$$

*then the work conservation assumption can be removed.*

COROLLARY 6.1 (OPTIMALITY OF $\mathrm{FSF}_p$ WITH RESPECT TO THE TAIL PROBABILITY). *The policy $\mathrm{FSF}_p$ stochastically minimizes both the queue length and the waiting time in steady-state among all FCFS work conserving policies in $\Pi_p$. In particular, $\mathrm{FSF}_p$ minimizes the tail probability $P(W > T)$ in steady-state within this set of policies. Alternatively, if* (44) *holds, then the work-conservation assumption can be removed.*

**ED + QED Regime: Asymptotically Optimal Routing and Staffing.** For a constraint of the type $P(W > T) \leqslant \alpha$, $T > 0$, $0 < \alpha < 1$, the papers of Baron and Milner (2009) and Mandelbaum and Zeltyn (2009) showed that the appropriate asymptotic regime is the so-called ED + QED regime. In this regime, the ED staffing is refined by an additional square-root capacity. In the context of the inverted-V model, this translates into a staffing vector that satisfies

$$\sum_{k=1}^{K} \mu_k N_K = \lambda(1-\Delta_1) + \delta_1\sqrt{\lambda} + o(\sqrt{\lambda}), \tag{45}$$

where $\Delta_1 = G(T)$, $G(T) = 1 - \exp(-\theta T)$, and $-\infty < \delta_1 < \infty$.

PROPOSITION 6.2. *Let $T > 0$, $0 < \alpha < 1 - G(T)$, and let $0 < \mu_1 < \mu_2 < \cdots < \mu_K$ and $\theta$ be fixed. Suppose that either $\theta \leqslant \mu_1$ or that only work-conserving policies are allowed. Consider a sequence of systems indexed by the arrival rate $\lambda > 0$, with $\lambda$ growing to infinity, and $N_k^\lambda$ servers in pool $k$, $k = 1, \ldots, K$. Let $N^\lambda = \sum_{k=1}^{K} N_k^\lambda$ be the total number of servers in system $\lambda$. Then there exists a sequence $\{\pi^\lambda\}$, with $\pi^\lambda \in \Pi(\lambda, \vec{N}^\lambda)$, of FCFS non-preemptive policies, under which*

$$\limsup_{\lambda\to\infty} P_{\pi^\lambda}^\lambda(W > T) = \alpha \tag{46}$$

*if and only if*

$$\mu_1 N_1^\lambda + \cdots + \mu_K N_K^\lambda \geqslant \lambda(1-\Delta_1) + \delta_1\sqrt{\lambda} + o(\sqrt{\lambda}), \quad \text{as } \lambda\to\infty, \tag{47}$$

*where $\Delta_1 = G(T)$, $\alpha = ((1 - G(T)) \cdot \bar{\Phi}(\delta_1/\sqrt{g(T)}))$, $g(T) = \theta\exp(-\theta T)$, and $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$ is the survival function of the standard normal distribution. Moreover, under* (45)*, all work-conserving policies satisfy* (46)*.*

Notice the upper bound $\alpha < 1 - G(T)$. This follows from the fact that if we end up serving no customers—and as a consequence, all customers abandon—we have $P(W > T) = 1 - G(T)$.

In light of Proposition 6.2, our proposed solution for the problem (43) is $\widetilde{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil$, where $\vec{M}^{*\lambda}$ is an optimal solution to $\mathrm{SP}(\lambda(1-\Delta_1) + \delta_1\sqrt{\lambda})$. (Recall the definition of $\mathrm{SP}(\cdot)$ given in §5.2.1.) Suppose that $\vec{N}^{*\lambda}$ is an optimal solution for the problem (43) with arrival rate $\lambda$. Then, similarly to Proposition 5.5, we can establish that $\lim_{\lambda\to\infty} |C(\vec{N}^{*\lambda}) - C(\widetilde{N}^\lambda)|/\lambda^{p-1/2} = 0$. In turn, analogously to Proposition 5.6, if $\delta_1 \neq 0$,

$$\lim_{\lambda\to\infty} (C(\vec{N}^{*\lambda}) - \underline{C}^{\lambda(1-\Delta_1)})/t(C(\widetilde{N}^\lambda) - \underline{C}^{\lambda(1-\Delta_1)}) = 1,$$

where $\underline{C}^{\lambda(1-\Delta_1)}$ is the staffing cost associated with the optimal solution of $\mathrm{SP}(\lambda(1-\Delta_1))$.

**QED Regime: Asymptotically Optimal Routing and Staffing.** For systems where short delays are desired or prevalent, it might be more appropriate to consider a constraint of the form $P(W > T/\sqrt{\lambda}) \leqslant \alpha$, where $T \geqslant 0$ and $0 < \alpha < 1$. This constraint is consistent with the QED regime. Working with FCFS policies only and under the assumption that either $\theta \leqslant \mu_1$ or otherwise only work-conserving policies may be considered, we end up with results that are analogous to the QED regime analysis in §§4 and 5 and hence will not be elaborated upon here. The one detail that is noteworthy is how to determine the coefficient of the square-root safety staffing as a function of $T$ and $\alpha$.

Using observations from Propositions 4.2 and 4.3, we establish that if $\lim_{\lambda\to\infty} N_1^\lambda/N^\lambda > 0$, then under the FSF policy

$$\lim_{\lambda\to\infty} P(W > T/\sqrt{\lambda}) = \alpha$$

if and only if

$$\sum_{k=1}^{K} \mu_k N_k = \lambda + \delta_0\sqrt{\lambda} + o(\sqrt{\lambda}),$$

where $-\infty < \delta_0 < \infty$ satisfies

$$\alpha = \frac{1 - \Phi(\sqrt{\theta}T + \delta_0/\sqrt{\theta})}{1 - \Phi(\delta_0/\sqrt{\theta})} \cdot \left(1 + \frac{\sqrt{\theta}h(\delta_0/\sqrt{\theta})}{\sqrt{\mu_1}h(-\delta_0/\sqrt{\mu_1})}\right)^{-1}. \tag{48}$$

In particular, like $\delta$, $\delta_0$ is a function of the service rates through $\mu_1$ only.

## 7. Conclusions

We study the joint staffing and routing problem with respect to the inverted-V system with abandonment. Recognizing that abandonment introduces new challenges, we propose with a robust problem formulation that seeks to minimize staffing costs subject to an upper bound on the abandonment probability or an upper bound on the tail probability of the waiting time. With respect to the former type of constraint problem, we show that it is asymptotically optimal to use work-conserving FCFS policies. With respect to the second constraint, we *assume* that only FCFS policies may be used. Additionally, supported by empirical evidence, we assume that the customer patience is stochastically longer than their service time. In the absence of this assumption, the set of admissible policies needs to be further restricted to work-conserving FCFS policies.

For both types of abandonment and waiting time constraints, we show that in the ED and the ED + QED regimes any FCFS work-conserving policy is an asymptotically optimal control. In the QED regime, we show that the FSF policy is an asymptotically optimal control. We then proceed to define the asymptotically feasible region,

which consists of staffing vectors under which the relevant constraint can be asymptotically satisfied. This region is shown to have a simple linear lower boundary, which is characterized by the total service capacity exceeding a regime-dependent function of the arrival rate. 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, due to the cruder optimization criteria involved with the ED regime, the analysis in that regime is much simpler than that of the QED regime. This suggests that solutions to other relevant skill-based routing problems may be achievable. Indeed, recently, Atar et al. (2010) have shown that a $c\mu/\theta$ index rule for the V-model is asymptotically optimal in the ED regime. Their analysis relies on the relevant fluid model. Surprisingly, while the optimization criteria in the ED + QED regime are just as stringent as those used in the QED regime, the control-related asymptotic analysis in the ED and ED + QED regimes is, nevertheless, very similar.

Third, while we study three separate asymptotic regimes, in reality, one may be dealing with one particular real-life system with a finite arrival rate. In that case, it is of interest to know which asymptotic regime provides the best approximation. In a numerical study, Mandelbaum and Zeltyn (2009, Table 1) find that for the single-pool case, the QED regime consistently provides good approximations except, possibly, for cases in which the constraints are loose and the system is large. It would be interesting to see to what extent their insights extend into the multiserver-pool case.

Finally, Armony and Ward (2010) consider the issue of fairness among the servers in the inverted-V model without abandonment. The present paper can be helpful for studying fairness in the model with impatient customers.

## 8. Electronic Companion

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

## Endnotes

1. Because $\lambda$ is a continuous parameter, it is appropriate to refer to a "family" of systems rather than a "sequence." Instead, we fix a specific increasing subsequence of the family $\{\lambda\}$ and use the superscript $\lambda$ for simplicity.
2. The expert reader might notice that this is a slightly different characterization of the various regimes 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.
3. If the steady-state distribution does not exist, consider $P_\pi(W > T)$ and $P_\pi(ab)$ as the random variables corresponding to the essential limsups of the long-term relevant proportions.
4. Note that while in (1) the value of $\Delta$ is restricted to the open interval $(0, 1)$, here we allow for values of $\Delta$ in $(0, \infty)$. This is because the probability of abandonment on the left-hand side of (28) is inflated by $\sqrt{\lambda}$.
5. The feasible region in the ED regime has a similar linear form as in Figure 2 with a lower bound specified by (27).

## Acknowledgments

The research of both authors was supported by BSF (Binational Science Foundation) grant 2006379. The second author was supported by the ISF (Israeli Science Foundation) grant 1357/08 and by the Technion funds for the promotion of research and sponsored research.

### References

Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call-center: A multi-disciplinary perspective on operations management research. G. Shanthikumar, D. Yao, eds. *Production and Operations Management. Special Issue on Service Operations in Honor of John Buzacott* **16**(6) 665–688.

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

Armony, M., A. Ward. 2010. Fair dynamic routing in large-scale heterogeneous-server systems. *Oper. Res.* **58**(3) 624–637.

Atar, R. 2008. Central limit theorem for a many-server queue with random service rates. *Ann. Appl. Probab.* **18**(4) 1548–1568.

Atar, R., A. Shwartz. 2008. Efficient routing in heavy traffic under partial sampling of service times. *Math. Oper. Res.* **33**(4) 899–909.

Atar, R., C. Giat, N. Shimkin. 2010. The $c\mu/\theta$ rule for many-server queues with abandonment. *Oper. Res.* **58**(5) 1427–1439.

Atar, R., A. Mandelbaum, G. Shaikhet. 2009. Simplified control problems for multiclass many-server queueing systems. *Math. Oper. Res.* **34**(4) 795–812.

Atar, R., Y. Y. Shaki, A. Shwartz. 2011. A blind policy for equalizing cumulative idleness. *Queueing Systems.* Forthcoming.

Baron, O., J. Milner. 2009. Staffing to maximize profit for call centers with alternate service level agreements. *Oper. Res.* **57**(3) 685–700.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2006a. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.* **54**(3) 419–435.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2006b. Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems: Theory Appl.* **51**(3–4) 249–285.

Borst, S., A. Mandelbaum, M. Reiman. 2003. Dimensioning large call centers. *Oper. Res.* **52**(1) 17–34.

Cabral, F. B. 2005. The slow server problem for uninformed customers. *Queueing Systems* **50**(4) 353–370.

Dai, J. G., T. Tezcan. 2008. Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems* **59**(2) 95–134.

de Véricourt, F., Y.-P. Zhou. 2005. Managing response time in a call-routing problem with service failure. *Oper. Res.* **53**(6) 968–981.

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

Foschini, G. J. 1977. On heavy traffic diffusion analysis and dynamic routing in packet switched networks. M. Reiser, K. Chandy, eds. *Computer Performance Measurements, Modeling, and Evaluation*. North-Holland, Amsterdam, 499–514.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.

Gans, N., N. Liu, A. Mandelbaum, H. Shen, H. Ye. 2010. Service times in call centers: Agent heterogeneity and learning with some operational consequences. J. O. Berger, T. T. Cai, I. M. Johnstone, eds. *A Festschrift for Lawrence D. Brown*. IMS Collections, Vol. 6. IMS, Beachwood, OH, 99–123.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4**(3) 208–227.

Gurvich, I., W. Whitt. 2007a. Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.* **34**(2) 363–396.

Gurvich, I., W. Whitt. 2007b. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* **11**(2) 237–253.

Gurvich, I., W. Whitt. 2010. Service-level differentiation in many-server service systems via queue-ratio routing. *Oper. Res.* **58**(2) 316–328.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–588.

Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing Service Oper. Management* **7**(1) 20–36.

Larsen, R. L., A. K. Agrawala. 1983. Control of a heterogeneous two-server exponential queueing system. *IEEE Trans. Software Engrg.* (July) 522–526.

Lin, W., P. R. Kumar. 1984. Optimal control of a queueing system with two heterogeneous servers. *IEEE Trans. Automat. Control* **29** 696–703.

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Sci.* **49**(8) 1018–1038.

Maglaras, C., A. Zeevi. 2004. Diffusion approximations for a multiclass Markovian service system with "guaranteed" and "best-effort" service levels. *Math. Oper. Res.* **29**(4) 786–813.

Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* **53**(2) 242–262.

Mandelbaum, A., S. Zeltyn. 2009. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.* **57**(5) 1189–1205.

Mandelbaum, A., A. Sakov, S. Zeltyn. 2000. Empirical analysis of a call center. Technical report. Accessed January 14, 2011, http://iew3.technion.ac.il/serveng/References/references.html.

Rubinovich, M. 1983. The slow server problem. *J. Appl. Probab.* **22** 205–213.

Stockbridge, R. H. 1991. A martingale approach to the slow server problem. *J. Appl. Probab.* **28** 480–486.

Teh, Y., A. R. Ward. 2002. Critical thresholds for dynamic routing in queueing networks. *Queueing Systems* **42**(3) 297–316.

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

Tezcan, T., J. G. Dai. 2010. Dynamic control of $N$-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Oper. Res.* **58**(1) 94–110.

Tseytlin, Y. 2007. Queueing systems with heterogeneous servers: Improving patients' flow in hospitals. Technion M.Sc. research proposal, Technion, Haifa, Israel.

Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* **50**(10) 1449–1461.

Whitt, W. 2005. Two fluid approximations for multi-server queues with abandonments. *Oper. Res. Lett.* **33**(4) 363–372.

Whitt, W. 2006a. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.

Whitt, W. 2006b. A multi-class fluid model for a contact center with skill-based routing. *Internat. J. Electronics Comm.* (*AEU*) **60**(2) 95–102.

Whitt, W. 2006c. Staffing a call center with uncertain arrival rate and absenteeism. *Production Oper. Management* **15**(1) 88–102.