MANAGEMENT SCIENCE
Vol. 54, No. 2, February 2008, pp. 279–294
ISSN 0025-1909 | EISSN 1526-5501 | 08 | 5402 | 0279

DOI 10.1287/mnsc.1070.0825

Service-Level Differentiation in Call Centers with Fully Flexible Servers

Itay Gurvich
Graduate School of Business, Columbia University, New York, New York 10027, ig2126@columbia.edu

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

Avishai Mandelbaum
The William Davidson Faculty of Industrial Engineering and Management, Technion Institute of Technology, Haifa 32000, Israel, avim@ie.technion.ac.il

We study large-scale service systems with multiple customer classes and many statistically identical servers. The following question is addressed: How many servers are required (staffing) and how does one match them with customers (control) to minimize staffing cost, subject to class-level quality-of-service constraints? We tackle this question by characterizing scheduling and staffing schemes that are asymptotically optimal in the limit, as system load grows to infinity. The asymptotic regimes considered are consistent with the efficiency-driven (ED), quality-driven (QD), and quality-and-efficiency-driven (QED) regimes, first introduced in the context of a single-class service system.

Our main findings are as follows: (a) Decoupling of staffing and control, namely, (i) staffing disregards the multiclass nature of the system and is analogous to the staffing of a single-class system with the same aggregate demand and a single global quality-of-service constraint, and (ii) class-level service differentiation is obtained by using a simple idle-server-based threshold-priority (ITP) control (with state-independent thresholds); and (b) robustness of the staffing and control rules: our proposed single-class staffing (SCS) rule and ITP control are approximately optimal under various problem formulations and model assumptions. Particularly, although our solution is shown to be asymptotically optimal for large systems, we numerically demonstrate that it performs well also for relatively small systems.

Key words: call centers; heavy traffic; QED regime; square-root safety staffing; skill-based routing; Halfin-Whitt regime; outsourcing; service-level constraints

History: Accepted by Ger Koole, special issue editor; received November 9, 2004. This paper was with the authors one week for 2 revisions.

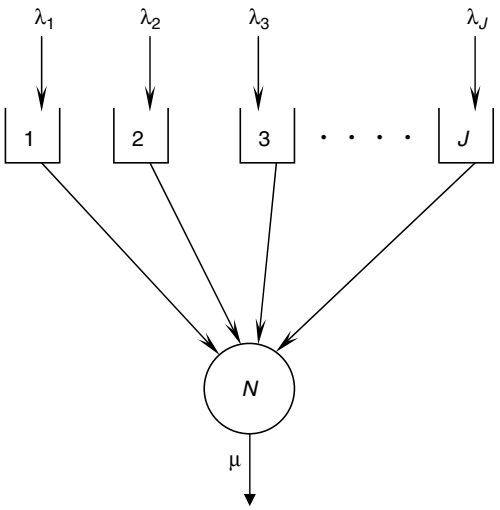
1. Introduction

Modern service systems strive to provide customers with personalized service, which is customized to the customers needs. Recent trends include self-selecting market segmentation, multilingual customer support, and customized cross-sales offerings. With this growing level of service customization, the variety of services provided by any given organization is increasingly high. This variety requires service personnel to possess a large skill set. It has been long recognized that to avoid overstaffing it is important to cross-train customer-service representatives and maintain server flexibility. However, to take full advantage of this high flexibility level, one needs to make efficient customer-server assignments and sensible staffing and cross-training decisions. These staffing and control problems are now receiving increasing attention, and is where this work's contribution lies.

Our work is largely motivated by modern call centers which often consist of dozens, hundreds, or even thousands of agents, and who strive to meet a large variety of customers needs. Examples include direct banking, multilingual services, and help desks. In such centers, a customer class may be characterized by its members special-service needs, their relative importance to the organization, or their quality-of-service expectations or guarantees. We model such systems by a multiclass multiserver queue with many servers, which we call the V-model. This model is depicted in Figure 1.

Naturally, call center managers strive to provide a high quality of service in terms of operational performance measures and other less tangible quality-of-service criteria. From an operational point of view, quality of service is expressed in terms of various performance measures. Those include the average speed of answer (ASA), the fraction of abandoning calls (Abn %), and the service level (SL). The latter measures the fraction of calls that are answered within a prespecified "service-level" target. For example, a call center may wish to have at least 80% of its calls

Figure 1 **The V-Model: Multiple Customer Classes and a Single-Server Type**



answered within 20 seconds. SL targets may be specified internally by the call center or, alternatively, they may be based on contractual agreements between the call center and its clients.

To determine staffing levels that will provide callers with the desired quality of service, many call centers today use the Erlang-C model, which is based on a single-class $M/M/N$ queueing system and its corresponding steady-state performance. To generalize this approach, researchers have proposed using the Erlang-A model, which includes also customer abandonment (see, for example, Garnett et al. 2002; Mandelbaum and Zeltyn 2006; Whitt 2004, 2006). This model has been increasingly adopted by developers of workforce management tools and is consequently becoming more prevalent in call centers. But what if the call center manager wishes to provide a differentiated service level to different customer classes? The V-model studied in this paper is a natural extension of the single-class Erlang-C and Erlang-A models, which allows one to handle situations in which service-level differentiation is desired.

In multiclass call centers, customers have already learned to expect a differentiated service level. Some organizations have special class designations (such as Platinum, Gold, Sliver, Economy, etc. in banks and airlines) where customers receive a differentiated quality of service depending on their class designation. Also, some contact centers provide service via multiple channels. Here too customers will experience different quality of service depending on the channel they have selected. For example, many contact centers answer phone calls within seconds or minutes but will answer e-mail inquiries after a few hours.

To capture the service-level differentiation element in our model, we assign an SL constraint to each of the customer classes, which are relatively important to the organization. In addition, we impose a global ASA constraint, which is much less restrictive than the SL constraints. The classes who do not have an individual SL constraint are referred to as the *best-effort* classes. This formulation is natural in an outsourcing environment, where the outsourcer is likely to have a global quality-of-service constraint in addition to customer-specific contracts.

With respect to this V-model, we ask the following question: How many servers are required (staffing) and how does one match them with customers (control) to minimize the staffing costs, subject to the SL and ASA constraints?

The staffing and control decisions are generally made at different time scales. While the control decisions are made online in real time, the staffing decisions are often made on a weekly basis, or even less frequently. Consistently with this difference in time scale, we show that to make the staffing decision it is sufficient to know only aggregate call volume, while the specific class-level arrival rates are only used later for the purpose of control. Specifically, the staffing rule is robust with respect to class-level arrival rates, as long as their sum can be forecasted accurately. Even though the staffing and control decisions involve different time scales, it is important to consider these problems together in a common framework to avoid suboptimal solutions. Nevertheless, due to the relative complexity of the joint staffing-control problem, they have generally been considered separately in the literature. Recent exceptions include Armony and Maglaras (2004a, 2004b), Armony (2005), Armony and Mandelbaum (2007), Bassamboo et al. (2006a, 2006b), Harrison and Zeevi (2005), Wallace and Whitt (2005), as well as this paper.

Our approach in addressing the staffing and control question is an asymptotic one; specifically, we characterize scheduling and staffing schemes that are asymptotically optimal as the aggregate arrival rate increases to infinity. The analysis following this approach is technically deep, but the final results are simple enough to be stated in a very accessible manner, enough for managers to apply directly (for example, the square-root safety staffing)—hence, we expect this paper to be useful in applications; consequently, the paper is structured such that the technicalities are discussed in the end. The main asymptotic framework considered in this work is the many-server heavy-traffic regime, first introduced by Halfin and Whitt (1981). Within the general framework of the many-server heavy-traffic regime, we focus on the following three more-specific regimes: quality-and-efficiency-driven (QED), quality-driven (QD), and efficiency-driven (ED).

1.1. Main Results

The main results of this paper are as follows:

(1) The joint problem of staffing and control is decoupled into two separate problems, where

(a) The staffing level is the same as in a single-class system with a common total arrival rate, and the global ASA constraint. We call this rule the *single-class staffing* (SCS) rule.

(b) The online control provides quality-of-service differentiation between the various customer classes via an *idle-server-based threshold priority* (ITP)[1] scheduling rule, where the threshold is on the minimal number of idle servers before customers of a particular class may be assigned to servers. The thresholds associated with this rule are state-independent and their values are easily determined as a function of the system parameters.

Thus, the staffing rule has the desirable property that it only requires partial demand information. Particularly, no class-level arrival-rate information is needed. When these arrival rates become known in real time, the control decisions make full use of this new information.

(2) Robustness of staffing and control: the SCS rule together with the ITP control are shown to be asymptotically optimal (under all three asymptotic regimes) for a variety of problem formulations and model assumptions, including our original constraint satisfaction problem, but also cost-minimization and profit-maximization problems, with or without customer abandonment.

The simplicity of the suggested staffing rule is of great importance. A priori, staffing decisions that need to take into consideration the service requirement of multiple customer classes can potentially be very complex. Our result, that only total arrival rate and the global ASA constraint are needed, simplifies the staffing decision tremendously. Moreover, the form of this SCS rule as a function of these two arguments is also very simple. For example, a special case of the SCS rule is the familiar square-root safety staffing rule (see Borst et al. 2004).

The dynamic control we propose of matching servers to customers is based on priorities and thresholds. In a nutshell, according to the ITP control, customer classes are prioritized with respect to their SL targets, with lowest priority to the best-effort customers. A customer of a certain priority can enter service only if there are no higher-priority customers waiting, and the number of idle servers exceeds a class-dependent threshold. A similar threshold policy has also been proposed in a call blending environment (i.e., call centers that handle both inbound and outbound calls) (Bhulai and Koole 2003, Gans and Zhou 2003). The role of the thresholds is to ensure that enough servers are available to serve future arrivals of higher priorities. This resembles the principle of capacity reservation in telecommunication networks (Puhalskii and Reiman 1998) and also of stock rationing in make-to-stock systems (Deshpande et al. 2003, de Véricourt et al. 2004). The thresholds in ITP can be easily adjusted to provide the right level of service.

The rest of this paper is organized as follows: The introduction is concluded with a brief literature review. We formally introduce the joint staffing and control problem in §2. The threshold-priority (ITP) rule and the corresponding queueing model (denoted by $M/M/N/\{K_i\}$), as well as the single-class staffing (SCS) rule are also introduced in §2. Section 3 then presents a numerical example to illustrate the applicability of our solution to both large and moderate-size systems. Section 4 introduces the asymptotic framework used in this paper. Section 5 establishes the asymptotic feasibility of our proposed joint staffing and control policies. Section 6 then shows the asymptotic optimality of SCS and ITP. In §7, staffing and control are discussed for an extension of the original model that includes customer abandonment. To conclude, §8 discusses the results and suggests directions for further research.

Due to the technical nature of our results, our approach in their presentation is to state them formally and precisely in the body of the paper, but the formal proofs appear in the online technical appendix (provided in the e-companion).[2]

1.2. Literature Review

There is extensive literature dealing with the V-model both in terms of performance analysis and in terms of performance optimization and control. However, little work has been done on the staffing problem and especially on the combined solution of staffing and control. Next, we mention only the papers most closely related to our work.

In the context of performance analysis, exact steady-state performance analysis of the V-model, under the threshold-priority scheme that we use in this paper, is given in Schaack and Larson (1986).

In the context of control, to differ from much of the literature on the V-model, our model formulation is characterized by imposing quality-of-service constraints rather then by assigning costs to the quality of

[1] We use the acronym ITP to describe this rule instead of simply TP (for threshold priority) to differentiate from the queue-length-based threshold priority rule (QTP), which has been suggested in other contexts (e.g., Bell and Williams 2001).

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

service and aiming at the overall cost minimization. There is vast literature on control of the V-model under cost-minimization objectives, and due to space limitations we do not give a comprehensive list here; instead, we refer the interested reader to Gurvich (2004) for a detailed survey of relevant papers both in the context of cost minimization as well as asymptotic performance analysis of the V-model under given policies. An example for the use of the same solution approach used here, to solve a different constraint satisfaction problem, can be found in the papers by Armony and Maglaras (2004a, 2004b), who were the first to consider dynamic control in the QED regime. Maglaras and Zeevi (2005) consider profit maximization for a loss system two-class V-model with pricing, sizing, and admission control. To distinguish this paper from our setting, the server-allocation scheme (Maglaras and Zeevi 2005) is fixed rather than a decision variable.

Choosing the quality-of-service constraints to use is not a trivial task because the obvious formulation leads to some undesirable performance characteristics. This issue is addressed by both Koole (2003), which is dedicated to the discussion of quality-of-service performance measures for call centers, and Milner and Olsen (2008), which deals with this issue in the context of contracts in the call center industry.

Finally, the literature on staffing of single-class systems is extremely relevant to this paper due to the structure of our suggested staffing rule, which is based on single-class considerations. The most relevant references in this context are the papers by Borst et al. (2004) and Mandelbaum and Zeltyn (2006), which cover in great generality staffing problems for the single-class $M/M/N$ and $M/M/N+G$ systems.

2. Model Formulation

Consider a large service system modelled as a multiclass queueing system with J customer classes and N statistically identical servers. Customers of class i arrive according to a Poisson process with rate λ_i, independently of other classes. We define $\lambda = \sum_{i=1}^{J} \lambda_i$ to be the aggregate arrival rate. Service times are assumed to be exponential with rate μ for all customer classes. Delayed customers of class i wait in an infinite buffer queue i.

We start by assuming that customers do not abandon (the model which includes abandonment is described in §7). We consider the minimization of the number of servers (staffing level) subject to quality-of-service constraints. These constraints are expressed in terms of the fraction of class i customers who wait more than T_i units of time before starting service. We refer to T_i as the SL target and to the set of constraints as the SL constrains. Customer classes who do not have an SL constraint associated with them are referred to as the *best-effort classes*. Because we do not differentiate between the different best-effort classes in terms of quality-of-service constraints, we may assume, without loss of generality (w.l.o.g), that there is a single best-effort class which is class J. In addition, we impose a constraint on the average speed of answer (ASA) of the entire customer population. This constraint is referred to as the *global ASA constraint*. Let W and W_i be, respectively, the steady-state waiting time of the entire customer population and the steady-state waiting time of class i customers. Let α_i be class i target SL probability. The problem is formally given as follows:

$$\begin{aligned} \text{minimize} \quad & N \\ \text{subject to} \quad & E[W] \le T, \\ & P\{W_i > T_i\} \le \alpha_i, \quad i = 1, \ldots, J-1, \\ & N \in \mathbb{Z}_+, \quad \pi \in \Pi. \end{aligned} \tag{1}$$

We refer to (1) as the combined best-effort/SL constraints formulation, or the best-effort formulation, for short. Here T and the SL targets T_i, $i = 1, \ldots, J-1$, are strictly positive constants and $0 < \alpha_i < 1$. We assume w.l.o.g that classes are ordered in increasing order of T_i, that is, $T_1 \le T_2 \le \cdots \le T_{J-1} < T$, and $\alpha_i < \alpha_{i+1}$ if $T_i = T_{i+1}$. For a given staffing level N, a control policy, π, is a set of rules that determine how to match calls with servers at any given time. The set of admissible policies, Π, is defined as follows:

Definition 2.1 (Admissible Policies). We say that π is an admissible scheduling policy if it is nonpreemptive nonanticipating and it satisfies the following two conditions:

(1) *Class FCFS*: Customers are served first-come first-served (FCFS) within each class.

(2) *All Customers are Served*: We assume that it is not allowed to block customers or send them elsewhere.[3]

Here, informally, nonanticipation means that scheduling decisions at time t can be based only on information that is available up to time t. For the sake of coherence, we postpone the discussion of the model formulation and the restriction on admissible policies to the end of §2.1.

2.1. The Proposed Solution

We provide here an informal description of our solution. This description is sufficient for practical purposes and does not require the use of asymptotic

[3] Formally, we assume that $Q(t) = A(t) - D(t) - Z(t)\ \forall t \ge 0$, where $A(t)$ and $D(t)$ are, respectively, the cummulative number of arrivals and service completions up to time t, and $Z(t)$ and $Q(t)$ are, respectively, the number of busy agents and the overall number of customers in all the J queues at time t.

framework or terminology. A more formal description is given in §5. Clearly, any solution to the staffing problem should specify both the staffing rule and the control to be used in real time. We will end the section with a complete description of the solution. We start, however, by introducing the control rule that we use and some of its characteristics that are essential to the understanding of our complete solution and its good performance. The control we use is called the *idle-server-based threshold priority* (*ITP*) rule and is defined as follows:

Upon a customer arrival or a service completion, assign the head-of-the-line class i customer to an idle server if and only if (1) queue j is empty for all classes j, such that $j < i$, and (2) the number of idle servers exceeds a threshold K_i, where, $0 = K_1 \le K_2 \le \cdots \le K_J$. We denote the queueing model associated with this policy as $M/M/N/\{K_i\}$.

Exact analysis of the $M/M/N/\{K_i\}$ was conducted in Schaack and Larson (1986). Theoretically, then, for fixed-system parameters and staffing levels and under a given set of threshold levels $\{K_i\}$, one could use the results of Schaack and Larson (1986) to calculate $P\{W_j > T_j\}$ for each j. We claim that staffing and routing using the $M/M/N/\{K_i\}$ model is approximately optimal for problem (1). That is, to solve (1), one could use the following recipe: *Assuming ITP is used, find the least staffing level N for which there exists a set of thresholds $\{K_i\}$ so that $E[W] \le T$ and $P\{W_j > T_j\} \le \alpha_j$ $\forall\, j \le J-1$.*

This, however, is not a particularly practical solution because calculating the correct optimal parameters N and $\{K_i\}$ requires an extensive search. It turns out, however, that one can approximate the performance measures under $M/M/N/\{K_i\}$ in a way that simplifies the solution tremendously. Specifically, we show in subsequent sections that a good approximation for the tail probabilities, $P\{W_j > T_j\}$, under the ITP rule is given by the following simple recursion:

$$P\{W_j > T_j\} \approx P\{W_{j+1} > 0\}\sigma_j^{K_{j+1}-K_j}\bar{F}(N\cdot T_j;\ \sigma_j, \sigma_{j-1}) \qquad \forall\, j \le J-1, \quad (2)$$

where $\sigma_j = \sum_{k=1}^{j}\rho_k$. Also, for given values $0 < y < x < 1$, $F(\cdot; x, y)$ is a distribution function ($\bar{F}(\cdot; x, y)$ is its complement) with Laplace transform $\psi(s/N;\ \sigma_i, \sigma_{i-1})/s$, where

$$\psi(s;\sigma_i,\sigma_{i-1}) = \begin{cases} \dfrac{\mu(1-\sigma_1)}{s(s+\mu(1-\sigma_1))}, & i=1, \\ \dfrac{\mu(1-\sigma_i)(1-\tilde{\gamma}_i(s))}{s(s-\hat{\lambda}_i+\hat{\lambda}_i\tilde{\gamma}_i(s))}, & i=2,\ldots,J-1, \end{cases} \quad (3)$$

and

$$\tilde{\gamma}_i(s) = \frac{s+\mu}{2\sigma_{i-1}\mu} + \frac{1}{2} - \sqrt{\left(\frac{s+\mu}{2\sigma_{i-1}\mu}+\frac{1}{2}\right)^2 - \frac{1}{\sigma_{i-1}}}. \quad (4)$$

By setting $T_j = 0$ in (2), we have that

$$P\{W_j > 0\} \approx P\{W_{j+1} > 0\}\sigma_j^{K_{j+1}-K_j}. \quad (5)$$

The important thing to note here is that the approximate waiting-time distribution of class j does not have any evident dependence on the thresholds of class $i = 1, \ldots, j-1$. This is not entirely true because some dependence exists through the value of $P\{W_J > 0\}$ which is required to initialize the recursion, and $P\{W_J > 0\}$ is indeed dependent on the threshold values. It turns out, however, that this dependence can be approximately removed. Specifically, we show that using the ITP rule with appropriately chosen thresholds, one has that the probability of delay of class J, $P\{W_J > 0\}$, can be approximated by the probability of delay in a simple $M/M/N$ FCFS queue. That is,

$$P\{W_J > 0\} \approx P\{W_{\lambda,\mu}^{\mathrm{FCFS}} > 0\}, \quad (6)$$

where $W_{\lambda,\mu}^{\mathrm{FCFS}}$ is the steady-state waiting time in an $M/M/N$ FCFS queue with arrival rate λ, service rate μ, and N agents. Considering (2) again, one can now see that the waiting-time distribution of class j is easily approximated using only the performance of class $j+1$.

Recall that our problem formulation requires also the calculation of the global average waiting time. This calculation is rather involved under the $M/M/N/\{K_i\}$ model, but turns out to have also a very simple approximation that uses the $M/M/N$ model. Specifically, when the thresholds are appropriately chosen, we show that

$$E[W] \approx E[W_{\lambda,\mu}^{\mathrm{FCFS}}]. \quad (7)$$

Having these approximations in mind, a naive solution procedure would first find the number of agents required to satisfy the global ASA constraint $E[W] \le T$. By Equation (7), we can find this number of agents approximately using a simple $M/M/N$ model. To this end, we redefine $W_{\lambda,\mu}^{\mathrm{FCFS}}(N)$ to be the steady-state waiting time in an $M/M/N$ system as a function of the number of agents, N. Once we find the optimal $M/M/N$ staffing level, we can determine the thresholds using our recursive expression (2) to determine the thresholds.

It turns out that this naive procedure works extremely well, and is indeed the solution procedure we propose. Specifically, we propose using an SCS rule and a proper use of the ITP rule: under the assumption that $T_i \ll T$ for all $i = 1, \ldots, J-1$ (i.e., that T is orders of magnitude greater than the T_is, e.g., minutes versus seconds, respectively), the following staffing and control procedure is approximately optimal:

- **Staffing:** Find the staffing level through the single class $M/M/N$ (or Erlang-C) model with arrival

rate λ, service rate μ, and FCFS service. Specifically, let

$$N^* = \text{Min}\{N \in \mathbb{Z}_+ \colon E[W^{\text{FCFS}}_{\lambda,\mu}(N)] \le T\}. \tag{8}$$

• **Control:** Use the ITP rule with the differences $\{K_{j+1} - K_j\}_{j \le J-1}$ chosen recursively for $j = J-1, \ldots, 1$ in the following manner:

Compute

$$K_{j+1} - K_j = \left\lceil \frac{\ln(\alpha_j / [P\{W_{j+1} > 0\} \bar{F}(N^* \cdot T_j; \sigma_j, \sigma_{j-1})])}{\ln(\sigma_j)} \right\rceil \vee 0, \quad j = J-1, \ldots, 1. \tag{9}$$

Set

$$P\{W_j > 0\} = P\{W_{j+1} > 0\} \sigma_j^{K_{j+1} - K_j}. \tag{10}$$

In the above, we set $P\{W_J > 0\} := P\{W^{\text{FCFS}}_{\lambda,\mu}(N^*) > 0\}$, and for two real numbers x and y, $x \vee y =: \max\{x, y\}$. The actual threshold values are then determined by setting $K_1 = 0$.

It is important to note that, as mentioned in the introduction, the staffing step in the above procedure requires only the knowledge of the aggregate arrival rate, while the individual class arrival rates are needed only to determine the threshold values in the control step. This way, the joint solution of staffing and control is decoupled into two independent decisions. This property is highly desirable for practical purposes because the information available to the manager when making staffing decisions is limited, but more is revealed when control decisions are made in real time.

Remark 2.1. An alternative equation for the threshold, that does not use the distribution function $\bar{F}$ and hence does not require a Laplace transform inversion, is given by

$$K_{j+1} - K_j = \left\lceil \frac{\ln(\alpha_j T_j / [P\{W_{j+1} > 0\} \hat{w}(N^*, \sigma_j, \sigma_{j-1})])}{\ln(\sigma_j)} \right\rceil \vee 0, \quad j = 1, \ldots, J-1, \tag{11}$$

where $\hat{w}(N, \sigma_j, \sigma_{j-1}) = [N\mu(1-\sigma_j)(1-\sigma_{j-1})]^{-1}$.

However, the simpler formula comes at the cost of a less-precise outcome. Specifically, thresholds calculated using (11) are expected to be significantly less-precise (with respect to the true optimal policy) than those obtained through Equation (9). Formula (11) is obtained using Markov's inequality, so that the inaccuracy of the threshold is influenced by the inaccuracy of Markov's inequality. Nevertheless, for large systems, the thresholds that are calculated through (11) will be approximately optimal.

Remark 2.2. It should be intuitively clear that the staffing level suggested by this procedure is actually a lower bound on the required number of agents. To see this, note that for a fixed value of N, and because we have a single service rate μ, the overall average queue length (and by Little's law—also the overall average waiting time) is minimized by any work conserving policy, and in particular, by FCFS. Consequently, the number of agents needed to satisfy the global ASA constraint is at least as large as needed for the same purpose under FCFS. Our claim is that, using our policy, the lower bound is approximately achieved. That is, using the lower bound, one can approximately satisfy all of the constraints.

Remark 2.3. Note that $W^{\text{FCFS}}_{\lambda,\mu}(\cdot)$ is easily calculated through any of the available Erlang-C calculators. For our numerical experiments, we used the freeware *4CallCenters* (4CC)[4] which has a tool, *Advanced Queries*, that solves, among other problems, the problem given in (8). When using 4CC, one can get as part of the output the value $P\{W_J > 0\} \approx P\{W^{\text{FCFS}}_{\lambda,\mu}(N) > 0\}$, which is the probability of delay in the corresponding $M/M/N$ queue.

Remark 2.4. The family of threshold policies is rather large, including controls that use thresholds on the queue length (QTP), rather than on the number of idle agents (ITP). To illustrate what kind of controls belong to the class QTP, consider a two-class model; a possible control is then the following: serve class 2 as long as the queue length of class 1 is less than K, for some integer number K, otherwise, serve class 1. When setting $K = 0$, the resulting control is a static priority with high priority given to class 1. Alternatively, one might consider the following control scheme: give absolute priority to class 1 all the time and admit class 2 customers to service only when their queue is longer than some value K.

In the so-called *conventional heavy traffic* literature, thresholds on the queue lengths are widely used (e.g. Bell and Williams 2001 and Teh and Ward 2002). A natural question is then, why does our solution recommend using thresholds on the number of idle servers rather than thresholds on the queue lengths? Alternatively, one might ask if the same performance achieved through ITP can be achieved through QTP. The answer is no. In particular, for the simple two-class example introduced above, one can show that ITP can achieve better performance for class 1 than the best achievable performance using any of the suggested QTP controls. The implication of the above is that in a strong service-level differentiation setting, the use of a QTP-type control will be suboptimal. On the other hand, using ITP in the many-server heavy-traffic regime is natural, taking advantage of the fact that idling some servers might not affect the overall performance of the system. This is, of course, not the

[4] The software is available at http://iew3.technion.ac.il/serveng/4CallCenters/Downloads.htm.

case in conventional heavy traffic where the number of servers is fixed, and idling even one server will result in a significant loss of service capacity.

Before we present our numerical results, we would like to briefly discuss the model formulation and the restrictions on the set of admissible policies. We start with the model formulation and assumptions. First, note that our assumption that the service-time distribution for all customer classes is the same (represented by a single service rate μ). This assumption may be realistic in some contexts (e.g., a multilingual call center or centers providing similar service but to customers of different importance), but may be less realistic in others (e-mail versus phone conversation). If one assumes class-dependent service rates, the resulting problem is much more difficult (see Atar et al. 2004 and Harrison and Zeevi 2004), and the solution no longer possesses the great simplicity that the solution to our model does, which enables us to jointly solve the staffing and control problem.

With respect to model formulation, a common formulation used in the call center industry is actually slightly different than the one we propose in (1), and is given by a pure SL constraint formulation as follows:

$$\begin{aligned} &\text{minimize} && N \\ &\text{subject to} && P\{W_i > T_i\} \le \alpha_i, \quad i = 1, \ldots, J, \\ &&& N \in \mathbb{Z}_+, \quad \pi \in \Pi. \end{aligned} \tag{12}$$

Note the differences between (12) and (1). The formulation in (12) contains an SL constraint for *all* classes, including class J, while in our formulation (1) the constraint for class J is replaced with a global ASA constraint.

When considering the pure SL constraint formulation (12) in detail, one finds that a true optimal solution to this formulation might have characteristics that are highly undesirable from the practical point of view. Problems with this formulation have already been identified by Milner and Olsen (2008), and also by Koole (2003). We illustrate these issues through the following simple example which emphasizes the fact that in a multiclass setting, unlike the single-class $M/M/N$ model, optimal solutions to (12) can lead to extremely bad performance when measured by other performance measures such as the mean waiting time.

To this end, consider a two-class model with the following parameters: $\lambda_1 = \lambda_2 = 200$/hour, $\mu_1 = \mu_2 = 2$/hour, and assume we impose the following QoS constraints: $P\{W_1 \ge 1 \text{ minute}\} \le 0.6$ and $P\{W_2 \ge 1 \text{ minute}\} \le 0.6$. Then, because both classes have the same constraint, one might suggest to combine them into one queue and transform the problem into a single-class problem in which the constraint is $P\{W > 1 \text{ minute}\} \le 0.6$ (here W would be the steady-state waiting time of the merged customer class). Using any Erlang-C calculator, one can find that the minimum staffing level required to satisfy the constraint in the single-class model is 205 agents. Also, the average waiting time when using 205 agents will be less than four minutes. Note that so far we have approached the problem by merging the two customer classes into one class using a single-class staffing problem. We claim, however, that in the original multiclass setting, the minimal staffing level that will render the system stable will be optimal. In particular, the minimum staffing required for stability in this case, $N = 201$, will also be optimal. The reason for optimality is that one can use the following alternating priority scheme: at each excursion of the number of customers in system above 201, the system will give absolute priority to a different class. This way, half of the time class 1 will have absolute priority and half of the time class 2 will have absolute priority. Focusing on a particular class i—half of the time these customers experience a service level of high priority in a two-class multiserver queue, and the other half of the time they experience a service level of low priority. Using the expression for $M/M/N$ priority systems given by Kella and Yechiali (1985), one can conclude that under this scheme, indeed, $P\{W_i > 1 \text{ minute}\} \le 0.6$, $i = 1, 2$, so that both constraints are satisfied. It can be also shown, however, that under this "optimal" staffing level, the average waiting time of each class is approximately half an hour or 30 times the tail constraint we imposed, and this is clearly not acceptable. It is worthwhile mentioning that although the above example uses symmetric constraints, similar arguments can be constructed for nonsymmetric constraints.

The bottom line, as illustrated by the above example, is that considering a pure SL constraint formulation might lead to results in which there are extreme inconsistencies between the SL constraints we impose for a fixed T_i and other performance measures such as the average waiting times, or even tail constraints for other values $\widehat{T}_i \neq T_i$. Indeed, in the above example, 60% will wait less than one minute but more than 20% will wait more than 45 minutes!

As will be shown in §5, the best-effort formulation (1) leads to a solution in which consistency between different measures of performance is preserved. Specifically, classes that have small T_is, thus reflecting the management's desire to give them high quality of service, will experience high quality of service across different performance measures. In particular, a small T_i will lead to a small average waiting time.

As for the admissible set of policies, it is rather simple to see why one might expect undesirable outcomes in the absence of the assumption that *all customers are served*. To this end, consider the following staffing and control procedure. Staff with any number of agents. Upon a customer arrival, reject the customer if there are no agents available. This procedure transforms the system into a loss system in which the waiting time is identically zero and the constraints are formally met even if we set the number of agents to be zero. This is, of course, an extreme example because a large portion of the customers are not served. Note, however, that if one is willing to reject a certain portion of the customers, then one can choose the number of agents for the loss system so that only a fraction ϵ of the customers are not served and ϵ can be made arbitrarily small. Specifically, given an arrival rate λ and service rate μ, and using a loss system with $\lceil \lambda/\mu(1-\epsilon) \rceil$ agents, the percent of rejected customers will be approximately ϵ (see, for example, Whitt 1984). Our assumption that *all customers are served* is equivalent to saying that deliberate rejection is not acceptable and is intended to prevent outcomes of this sort. Finally, note that if we define the waiting time of a customer that does not get served to be ∞, then the assumption that all customers are served is redundant because the global ASA constraint will force the system to give service to all the customers.

The Class FCFS assumption seems to be natural from the practical point of view because it is considered fair (see, for example, Rafaeli et al. 2005) and it would seem inappropriate to prioritize customers within one homogenous class. It is not, however, optimal from a purely mathematical point of view. Indeed, for single-server settings, Towsley and Baccelli (1991) formally establish that the last-come first-served (LCFS) rule outperforms FCFS in terms of the fraction of customers that miss their deadlines. This is intuitive because LCFS tends to first serve customers who have not yet missed their deadlines. In particular, it is plausible that when removing the FCFS restriction, one can construct, based on LCFS, a feasible solution for (1) that will use fewer agents than under FCFS. This lower staffing level, however, will come at the price of extremely high mean and variance of the waiting time. That is, the policy might be inconsistent across performance measures, and one might meet the constraints in (1) for given values of T_i, while experiencing bad performance under other performance measures, such as the mean waiting time, or even SL constraints with other values of T_i. This should come as no surprise because it is well known that, within a large class of policies, FCFS minimizes the inconsistency or variance of the service level that customers experience (see, for example, Chap. 5 in Wolff 1989).

3. Numerical Study

Our proposed staffing and control described in the previous section is approximately optimal for large systems. Our purpose in this section is to show that our proposed solution performs extremely well, even for moderate-size systems. In particular, we show an example in which our proposed staffing differs by at most one agent from the staffing level associated with the optimal solution. Our numerical investigation is, by no means, an exhaustive one. However, based on our own experience as well as on other works that use similar methodologies to ours, such as Borst et al. (2004), we have a good reason to believe that our approximations perform extremely well, and that the numerical example given below is a representative one.

We apply our proposed solution to a simple three-class example. Specifically, consider a V-model with three classes such that $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}\lambda$, and an average handling time of three minutes, i.e., $\mu = 20$/hour. Assume that the overall average waiting time is required to be less than one minute. Also assume that a 80% of class 2 customers are expected to wait less than 20 seconds and 80% of class 1 customers are expected to wait less than 10 seconds. Formally, we consider the problem

$$\begin{aligned} \text{minimize} \quad & N \\ \text{subject to} \quad & E[W] \le 1 \text{ min}, \\ & P\{W_1 > 10 \text{ sec}\} \le 0.2, \\ & P\{W_2 > 20 \text{ sec}\} \le 0.2, \\ & N \in \mathbb{Z}_+. \end{aligned} \tag{13}$$

In this example, it is not completely obvious that the difference between one minute and 10 or 20 seconds is consistent with our assumption that $T_i \ll T$. Indeed, one might regard one minute and 10 or 20 seconds as being of the same order of magnitude. Still, we show that even in this seemingly ambivalent setting, our solution procedure works extremely well. By the argument given in Remark 2.2, for each value

Table 1 **Staffing Values of the $M/M/N$ Lower Bound**

Offered load	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
Staffing	17	22	27	32	37	43	48	53	58	63	68	73	78	83	88	93	98	103
Threshold	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2 Thresholds for Class 3 Using (11)

Offered load	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
Threshold	3	3	3	3	3	2	2	2	2	2	2	2	1	1	1	1	1	1

of λ a lower bound on the staffing level in (13) is given by considering an $M/M/N$ system with arrival rate λ, service rate μ, and FCFS service, in which we wish to find the minimal required staffing so that the average wait is less than one minute. The values of the required staffing levels for this simplified problem can be obtained easily by using any Erlang-C calculator. We used for this purpose the *Advanced Queries* feature of the 4CC software,[5] where we consider λ values between 500 arrivals to 2,000 arrivals per hour. This way, we start with a medium-size system of less than 20 agents. To emphasize the size of systems considered here, we give the numerical results as a function of the *offered load*, which is the amount of work arriving per unit of time, given by $R = \lambda/\mu$, which is also a lower bound on the number of agents required for stability. Because the $M/M/N$ staffing levels required to achieve the waiting-time constraints are typically close to R, this parameter gives a good indication of system size. Table 1 displays the output of the 4CC software, which indicates, for each value of R, the minimum required number of agents. Also, the third row gives the threshold for class 3 (K_3) that is recommended by our procedure, while the thresholds for classes 1 and 2 are recommended to be identically zero for all values of the offered load.

We now use inversion of the exact Laplace transforms given in Schaack and Larson (1986) to evaluate the performance experienced by the different customer classes. Figure 2(a) shows the performance levels experienced by classes 1 and 2 when using the staffing levels as given by the lower bound and employing a simple static priority rule with the highest priority given to class 1 and the lowest to class 3, that is, no thresholds are employed. As one can see, the policy is feasible for all values of $R > 35$.[6]

We next use the threshold priority control recommended by our procedure, with the same priority ordering but with a threshold of one applied to class 3 (for $R \leq 35$)—that is, a customer of class 3 will be admitted into service only if there is more than one free agent and queues 1 and 2 are empty. In this case, as depicted in Figure 2(b), we can see that feasibility holds for classes 1 and 2 for all values of R, but at the price of a slight violation of feasibility in terms of the global ASA constraint.

[5] http://ie.technion.ac.il/serveng.

[6] One can observe a jump down of the average waiting time when the offered load is 40. This is merely a result of the integral nature of the staffing levels and the transition, at $R = 40$, from staffing with two additional agents above R, to three additional agents above R.

The above might be satisfactory. However, we might do a great deal better by adding just a single agent. Figure 3(a) shows that, indeed, the addition of a single agent is sufficient for all values of R between 15 and 35, while using a static priority scheme (the same holds for a threshold policy).

Hence, we can summarize with Figure 3(b) that shows that the lower bound staffing given by the $M/M/N$ single-class model and used by our solution procedure differs by at most one from the feasible solution given by adding one agent, and in particular, it differs by at most one from the optimal solution.

Remark 3.1. In our calculation, we used the threshold as determined through Equation (9). As suggested in Remark 2.1, one can also use the less-precise formula given by Equation (11). When following this formula, the thresholds for class 3 are given in Table 2 (the thresholds for classes 1 and 2 are still zero).

The use of the these thresholds will, naturally, lead to a greater violation of the ASA global constraint for small systems. According to our calculations, however, feasibility is maintained for all values of offered

Figure 2 Constraint Satisfaction for Classes 1 and 2: (a) Using Static Priority and (b) Using Thresholds

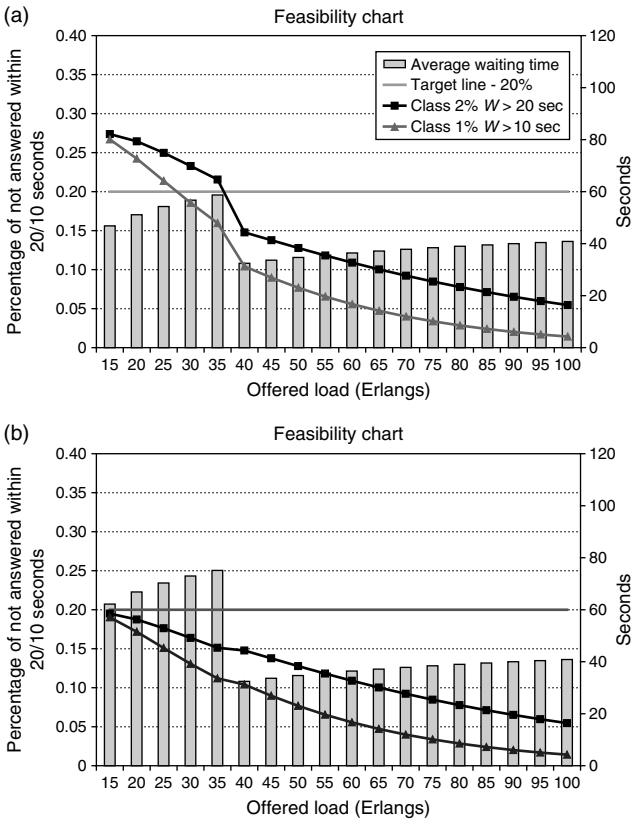
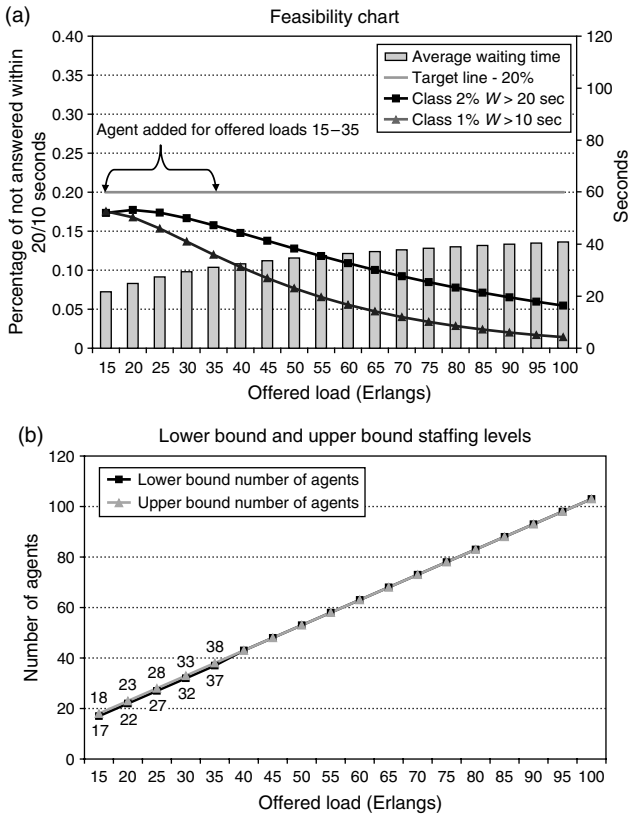


Figure 3 (a) Constraint Satisfaction After Staffing Refinement; (b) Optimality of Staffing Levels



load strictly greater than 35. Overall, the simplicity of the alternative threshold formula (9) comes at the price of being less precise for small systems. It will, however, work well for large systems and is proved to be approximately optimal for large systems.

4. The Asymptotic Framework

So far our references to the term "approximately optimal" have not been formally defined. To make these statements results formal, we need to first introduce our asymptotic framework. In this framework, we consider a sequence of systems with increasing arrival rates, and characterize staffing and control schemes which are asymptotically optimal, as the arrival rates increase to ∞. The original system of interest is assumed to be a member in this sequence. If the total arrival rate for this system is sufficiently large, then an asymptotically optimal policy is expected to be nearly optimal for this original system.

There are several reasons why it makes sense to consider an asymptotic approach to this problem instead of an exact one. First, it is clear from Yahalom and Mandelbaum (2004) that an optimal control policy that minimizes waiting costs must be highly dependent on system parameters and system state. Particularly, implementing such a control is difficult due to the large state space and the large number of system parameters. Even if attention is restricted to the threshold-priority (ITP) rule, the actual threshold values need to be determined. In addition, for staffing purposes, one would need to evaluate the system performance given different values of N. An exact approach would lead to very complicated expressions, and is not likely to provide useful and general insights.

Following the asymptotic approach, we consider a sequence of systems indexed by $r = 1, 2, \ldots$ (to appear as a superscript) with an increasing total arrival rate $\lambda^r = \sum_{i=1}^{I} \lambda_i^r$ and a fixed service rate $\mu^r \equiv \mu$. Let $R^r = \lambda^r/\mu$ be the total system load; then, without loss of generality, we assume that the index r is selected such that

$$r \equiv R^r. \tag{14}$$

The arrival rates to the different classes may be quite general. We only assume that the arrival rate of the lowest priority is comparable to λ^r for each r. More formally, we assume that there are J numbers $\xi_k \geq 0$, $k = 1, \ldots, J$, with $\sum_{k=1}^{J} \xi_k = 1$, such that the arrival rate of each class behaves according to the following rule:

$$\begin{aligned} &\lim_{r\to\infty} \frac{\lambda_k^r}{\lambda^r} = \xi_k, \quad k = 1, \ldots, J, \\ &\xi_J > 0,\ \xi_i \geq 0, \quad i = 1, \ldots, J-1. \end{aligned} \tag{15}$$

For every fixed r, the service-level differentiation between classes $1, \ldots, J-1$ is mathematically imposed through the following asymptotic formulation:

$$\begin{aligned} \text{minimize} \quad & N \\ \text{subject to} \quad & E[W^r] \leq T^r, \\ & P\{W_i^r > T_i^r\} \leq \alpha_i, \quad i = 1, \ldots, J-1, \\ & N \in \mathbb{Z}_+, \quad \pi \in \Pi, \end{aligned} \tag{16}$$

where we assume, w.l.o.g, that classes $i = 1, \ldots, J-1$ are ordered in nondecreasing order of T_i^r, with $\alpha_i < \alpha_{i+1}$ whenever $T_i^r = T_{i+1}^r$. Also, the informal assumption that the constraint for classes $i = 1, \ldots, J-1$ are of smaller order of magnitude than the global constraint, is formally given through the following:

Assumption 4.1. *$T^r = \widehat{T}/r^\gamma$, $T_i^r = \widehat{T}_i/r^{\gamma_i}$, and $\gamma_i > \gamma$ for all $i < J$ and $\gamma \in (0, \infty)$.*

In what follows, we assume that Assumption 4.1 is satisfied. The results will be given for arbitrary $\gamma \in (0, \infty)$. Our experience, as reflected also by the evidence given in Borst et al. (2004), indicates, however, that the results obtained from fixing $\gamma = 1/2$ are typically extremely close to the true, nonasymptotic optimal results.

In terms of the asymptotic framework, the SCS and ITP rules are given as follows: Consider a sequence of

systems, indexed by r, with service rate μ and aggregate arrival rate λ^r for the rth system, and such that (15) holds. Then, the SCS staffing rule and the ITP control rule are given as follows:

• **Staffing (SCS):** Find the staffing level through the single-class $M/M/N$ (or Erlang-C) model with arrival rate λ^r, service rate μ, and FCFS service. Specifically, let

$$N^{*r} = \text{Min}\{N \in \mathbb{Z}_+ \colon E[W^{\text{FCFS}}_{\lambda^r,\mu}(N)] \le T^r\}. \tag{17}$$

• **Control:** Use the ITP rule with the differences $\{K^r_{j+1} - K^r_j\}_{j\le J-1}$ chosen recursively for $j = J-1, \dots, 1$ in the following manner:

Compute

$$K^r_{j+1} - K^r_j = \left\lceil \frac{\ln(\alpha_j/[P\{W^r_{j+1} > 0\}\bar{F}(N^{*r}\cdot T^r_j;\, \sigma^r_j,\, \sigma^r_{j-1})])}{\ln(\sigma^r_j)} \right\rceil \vee 0, \quad j = J-1, \dots, 1. \tag{18}$$

Set

$$P\{W^r_j > 0\} = P\{W^r_{j+1} > 0\}(\sigma^r_j)^{K^r_{j+1}-K^r_j}. \tag{19}$$

In the above, we set $P\{W^r_J > 0\} := P\{W^{\text{FCFS}}_{\lambda^r,\mu}(N^{*r}) > 0\}$ and $\sigma^r_j = \sum_{k=1}^{j} \rho^r_k = \sum_{k=1}^{j}(\lambda^r_k/N^r\mu)$. The actual threshold values are then determined by setting $K^r_1 = 0$.

Remark 4.1. Note that the staffing and control rules above are defined through the true parameters T^r and T^r_i, α_i, $i = 1, \dots, J-1$, and independently of the scaling parameters γ_i, $i = 1, \dots, J$. In particular, the implementation of the policy is straightforward and there is no need to know or guess the scaling factors.

5. Asymptotic Feasibility of SCS and ITP

In this section, we establish that our proposed SCS rule and ITP control are asymptotically feasible for (16).

We start by defining asymptotic feasibility. For $r = 1, 2, \dots$, consider a sequence of systems with a fixed number of customer classes J and a fixed service rate. Let $\bar{\lambda}^r = \{\lambda^r_1, \dots, \lambda^r_J\}$ be a sequence of arrival rates with a total arrival rate $\lambda^r = \sum_{i=1}^{J} \lambda^r_i$, which is increasing to ∞ as $r\to\infty$. Let (N^r, π^r) be a joint staffing and control pair associated with the rth system.

Definition. The sequence $\{(N^r, \pi^r)\}$ is *asymptotically feasible* with respect to $\bar{\lambda}^r$ and $\bar{\alpha}^r = (\alpha^r_1, \dots, \alpha^r_J)$ if the following conditions apply:

- $\limsup_{r\to\infty} E[W^r]/T^r \le 1$, and
- $\limsup_{r\to\infty} P\{W^r_i > T^r_i\} \le \alpha_i \ \forall\, i = 1, \dots, J-1$.

From now on, when using ITP and SCS, we refer to the asymptotic version as given in Equations (17) and (18). The asymptotic feasibility of ITP and SCS is stated in Theorem 5.1, which is given at the end of this section. This theorem is based on Propositions 5.1 and 5.2 that are given below. In what follows, for two sequences $\{a^r\}$ and $\{b^r\}$, we say that $a^r \approx b^r$ if $a^r/b^r \to 1$ as $r \to \infty$.

Proposition 5.1. *Consider a sequence of systems indexed by $r = 1, 2, \dots$, with service rate μ for all classes, and class i arrival rates λ^r_i, $i = 1, \dots, J$, which satisfy* (15). *Fix the values of $\widehat{T}$, γ, $\widehat{T}_i$, $i = 1, \dots, J-1$ and γ_i, $i = 1, \dots, J-1$, and assume that N^r is determined according to SCS and ITP is used with thresholds K^r_i, $i = 1, \dots, J$ determined through* (18). *Then,*

$$P\{W^r_J > 0\} \approx P\{W^{\text{FCFS}}_{\lambda^r,\mu}(N^r) > 0\} \tag{20}$$

and

$$P\{W^r_i > 0\} \approx P\{W^{\text{FCFS}}_{\lambda^r,\mu}(N^r) > 0\} \cdot \prod_{j=i}^{J-1} (\sigma^r_j)^{K^r_{j+1}-K^r_j}, \quad i = 1, \dots, J-1. \tag{21}$$

Proposition 5.1 evaluates the delay probability for the different customer classes under the SCS and ITP policies. But what about the actual waiting time, given that a customer is indeed delayed? Proposition 5.2 provides expressions for the limiting distribution of the normalized waiting times (conditional on a positive wait). In this proposition and throughout, we use $\Rightarrow$ to denote weak convergence.

Proposition 5.2. *Under the assumptions of Proposition* 5.1 *and assuming that SCS are ITP are used, both $r^\gamma W^{\text{FCFS}}_{\lambda^r,\mu}(N^r)$ and $r^\gamma W^r_J$ converge weakly to the same limit. That is, both*

$$\begin{gathered} r^\gamma W^{\text{FCFS}}_{\lambda^r,\mu}(N^r) \ \Rightarrow\ W \quad \text{as } r\to\infty, \quad \text{and} \\ r^\gamma W^r_J \ \Rightarrow\ W \quad \text{as } r\to\infty, \end{gathered} \tag{22}$$

where the limit W is a proper random variable. In addition, the steady-state waiting times of the higher priorities $i = 1, \dots, J-1$ satisfy

$$N^r \cdot [W^r_i \mid W^r_i > 0] \ \Rightarrow\ [W_i \mid W_i > 0] \quad \text{as } r\to\infty, \tag{23}$$

where the limit W_i is a proper random variable and the density of $[W_i \mid W_i > 0]$ has the Laplace transform

$$\psi(s;\, \sigma_i,\, \sigma_{i-1}) = \begin{cases} \dfrac{\mu(1-\sigma_1)}{s(s+\mu(1-\sigma_1))}, & i = 1, \\ \dfrac{\mu(1-\sigma_i)(1-\tilde{\gamma}_i(s))}{s(s-\hat{\lambda}_i+\hat{\lambda}_i\tilde{\gamma}_i(s))}, & i = 2, \dots, J-1, \end{cases} \tag{24}$$

with

$$\sigma_i = \lim_{r\to\infty} \sum_{j=1}^{i} \frac{\lambda_i^r}{N^r\mu}, \quad \sigma_0 = 0, \quad \hat{\lambda}_i = \lim_{r\to\infty} \frac{\lambda_i^r}{N^r},$$

and

$$\tilde{\gamma}_i(s) = \frac{s+\mu}{2\sigma_{i-1}\mu} + \frac{1}{2} - \sqrt{\left(\frac{s+\mu}{2\sigma_{i-1}\mu} + \frac{1}{2}\right)^2 - \frac{1}{\sigma_{i-1}}}. \tag{25}$$

Also, for $i = 1, \ldots, J-1$, the limits of the first and second moments of the conditional waiting time satisfy

$$\begin{aligned} &N^r E[W_i^r \mid W_i^r > 0] \to [\mu(1-\sigma_i)(1-\sigma_{i-1})]^{-1} \\ &\qquad\qquad \text{as } r\to\infty, \quad \text{and} \\ &(N^r)^2 E[(W_i^r)^2 \mid W_i^r > 0] \\ &\quad \to 2(1-\sigma_i\sigma_{i-1})\left[(\mu)^2(1-\sigma_i)^2(1-\sigma_{i-1})^3\right]^{-1} \\ &\qquad\qquad \text{as } r\to\infty. \end{aligned} \tag{26}$$

In particular,

$$E[W^r] \approx E[W^{\mathrm{FCFS}}_{\lambda^r,\mu}(N^r)]. \tag{27}$$

Remark 5.1. Note that the Laplace transform depends only on the global parameter μ and on σ_i and σ_{i-1}. Hence, we can use a common function $F(\cdot;\cdot,\cdot)$ to describe the approximate distribution function of $N^r[W_i^r \mid W_i^r > 0]$ for different i. In particular, $[W_i^r \mid W_i^r > 0]$ will have approximately a distribution function $F(N^r\cdot;\sigma_{i-1},\sigma_i)$ that has the Laplace transform $\psi(s/N^r;\sigma_i,\sigma_{i-1})/s$. This Laplace transform can be numerically inverted using any numerical inversion package.

As a direct consequence of Proposition 5.2, one can conclude that the order of magnitude of the queue lengths associated with the higher-priority classes, $i = 1, \ldots, J-1$, is $\Theta(1)$, where for two nonnegative sequences $\{a_n\}_{n\ge1}$ and $\{b_n\}_{n\ge1}$, we say that $a_n = \Theta(b_n)$ if $\limsup_{n\to\infty} a_n/b_n < \infty$ and $\liminf_{n\to\infty} a_n/b_n > 0$. The details are stated in the following corollary.

Corollary 5.1. *Under the assumptions of Proposition* 5.1, *and assuming that SCS and ITP are used, the class-level queue lengths for the high-priority classes satisfy $E[Q_i^r \mid Q_i^r > 0] = \Theta(\lambda_i^r/N^r)$, $i = 1, 2, \ldots, J-1$. In particular, for $i = 1, \ldots, J-1$, and using the notation of Proposition* 5.2,

$$\begin{aligned} &E[Q_i^r \mid Q_i^r > 0] = \lambda_i^r E[W_i^r \mid W_i^r > 0] \\ &\quad \to \hat{\lambda}_i[\mu(1-\sigma_i)(1-\sigma_{i-1})]^{-1} \quad \text{as } r\to\infty, \quad \text{and} \\ &E[Q_i^r] \approx \frac{\lambda_i^r}{N^r} P\{W_i^r > 0\}[\mu(1-\sigma_i)(1-\sigma_{i-1})]^{-1}. \end{aligned} \tag{28}$$

An important implication of Proposition 5.2 and Corollary 5.1 is that the queue length of class J is of order $r^{1-\gamma}$, while the queue lengths of other classes are of smaller order. Hence, if queue lengths are scaled by $r^{1-\gamma}$, only the queue length of the lowest priority J does not disappear in the limit as $r\to\infty$. This essentially implies that, when r is very large, it is sufficient to know the total queue length to deduce the class-level queue lengths. This result is summarized in the following proposition.

Proposition 5.3 (State-Space Collapse). *Under the assumptions of Proposition* 5.1, *and assuming that SCS and ITP are used,*

$$\frac{1}{r^{1-\gamma}} Q_i^r \;\Rightarrow\; 0, \quad i = 1, \ldots, J-1. \tag{29}$$

The following proposition formally states the asymptotic feasibility of SCS and ITP and is a summary of Propositions 5.1 and 5.2 above.

Theorem 5.1. *Consider a sequence of systems indexed by $r = 1, 2, \ldots,$ with service rate μ for all classes, and class i arrival rate λ_i^r, $i = 1, \ldots, J$, which satisfy* (15). *Fix the values of $\widehat{T}$, γ, $\widehat{T}_i$, $i = 1, \ldots, J-1$ and γ_i, $i = 1, \ldots, J-1$, and assume that N^r is determined according to SCS and ITP is used with thresholds K_i^r, $i = 1, \ldots, J$, determined through* (18). *Then, we have asymptotic feasibility, i.e.,*

- $\limsup_{r\to\infty} E[W^r]/T^r \le 1$, *and*
- $\limsup_{r\to\infty} P\{W_i^r > T_i^r\} \le \alpha_i \;\forall\, i = 1, \ldots, J-1$.

Having the feasibility of SCS and ITP, Remark 2.2 can be used to argue that SCS and ITP are actually optimal. We make this assertion formally in the next section.

6. Asymptotic Optimality of SCS and ITP

In this section, we establish the asymptotic optimality of SCS and ITP as a joint staffing and control solution to problem (16).

First, to ensure stability, a reasonable staffing level would be of at least the order of λ^r. Hence, different staffing-level propositions are expected to all be of the same order of magnitude. Therefore, to obtain a meaningful form of asymptotic optimality, one must compare normalized staffing costs that measure the difference between the actual staffing costs and a base cost of the order of λ^r, which is a lower bound of the staffing cost.

To define asymptotic optimality, let $\bar{K}^r = \{K_1^r, \ldots, K_J^r\}$ and $\bar{\lambda}^r = \{\lambda_1^r, \ldots, \lambda_J^r\}$ be the thresholds and arrival rates in the rth system. Note that $R^r = \lambda^r/\mu$ is a lower bound on the value of the objective function in (16) because at least R^r servers are required for stability.

Definition. An asymptotically feasible sequence $\{N^r, \pi^r\}$ is *asymptotically optimal* with respect to

$\bar{\lambda}^r, \bar{\alpha}^r = (\alpha_1^r, \ldots, \alpha_J^r)$, if for any other asymptotically feasible sequence of policies $\{\widehat{N}^r, \widehat{\pi}^r\}$, we have

$$\liminf_{r\to\infty} \frac{\widehat{N}^r - R^r}{N^r - R^r} \geq 1.$$

We will now turn to the solution of (16). The following theorem states the asymptotic optimality of SCS and ITP as a solution for (16).

THEOREM 6.1. *Under* (15) *and Assumption* 4.1, *the SCS and ITP staffing and control as defined in* (17) *and* (18) *are asymptotically optimal for problem* (16).

REMARK 6.1 (INTUITIVE EXPLANATION OF THEOREM 6.1). This theorem is an immediate consequence of Theorem 5.1. To see this, let us consider the following reasoning: an intuitive lower bound for the required staffing would be to solve a different constraint satisfaction problem, in which all classes are treated as a single class with a constraint imposed only on the overall average waiting time. Hence, as suggested also by Remark 2.2, the staffing levels determined by SCS are a lower bound for (16). Theorem 5.1 ensures, then, that using the same lower bound staffing level for the original multiclass system, together with appropriately chosen thresholds, is asymptotically feasible. Hence, the lower bound is achieved and the policy is asymptotically optimal.

6.1. Discussion of Operational Regimes

So far, in an attempt to state our results in the highest degree of generality, the choice of a specific asymptotic regime has not been specified. In particular, our SCS staffing rule is given in terms of an associated $M/M/N$ system without specifying in what regime the $M/M/N$ system operates. For $M/M/N$ systems, however, there is a precise characterization of asymptotic operational regimes, which is fully given in Borst et al. (2004). In particular, the regime spectrum is divided into three possible outcomes: the efficiency-driven (ED) regime, the quality-and-efficiency-driven (QED) regime, and the quality-driven (QD) regime. The different regimes may be characterized by the probability of delay experienced by different customers. In particular, in the ED regime the probability of delay is close to one, and in the QD regime it is close to zero, while in the QED regime there is a delicate combination of high efficiency with a probability of delay that is strictly between zero and one. Can we construct a parallel characterization for the V-model operating under an ITP policy (and denoted by $M/M/N/\{K_i\}$)? The answer is yes, and the characterization is actually given implicitly in Proposition 5.1. Specifically, one can show that if the $M/M/N/\{K_i\}$ model is used with $K_J^r \ll r^\gamma$, then, analogously to Halfin and Whitt (1981), we have that

$$P\{W_J^r > 0\} \to \alpha, \quad 0 < \alpha < 1, \tag{30}$$

if and only if

$$\sqrt{N^r}(1-\rho^r) \to \beta > 0, \tag{31}$$

which corresponds to $\gamma = 1/2$. In which case, we would have $\alpha = \alpha(\beta)$, where the Halfin-Whitt delay function $\alpha(\cdot)$ is given by

$$\alpha(\beta) \triangleq \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}. \tag{32}$$

Here $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the standard normal density and distribution functions.

Moreover, because our staffing solution to (16) is strongly related to an optimization problem for an associated single-class $M/M/N$ queue, it can be stated in terms of orders of magnitude along the lines of Borst et al. (2004). In particular, according to Borst et al. (2004), the SCS rule reduces to *staff with* $N = R + \beta^r\sqrt{R}$, *where* β^r *is the unique solution to*

$$\alpha_\gamma(\beta^r)\frac{1}{\beta^r\mu\sqrt{R}} = T^r, \tag{33}$$

where

$$\alpha_\gamma(\beta^r) = \begin{cases} 1 & \text{if } \gamma < 1/2, \\ \dfrac{\phi(\beta^r)}{\beta^r} & \text{if } \gamma > 1/2, \\ \alpha(\beta^r) & \text{if } \gamma = 1/2. \end{cases} \tag{34}$$

Here, $\phi(\cdot)$ *is the standard normal density function.*

Note that for $\gamma > 1/2$, we will have $\beta^r \to \infty$, as $r \to \infty$, and the probability of delay (which is approximated using the tail of the normal distribution) will converge to zero as $r \to \infty$. If $\gamma = 1/2$, the procedure results in the well-known square-root safety staffing rule, and the optimal staffing level is given by $N = R + \beta\sqrt{R}$ for some $\beta > 0$.

Our results on the convergence of the waiting time of class J can also be stated in terms of the operational regime as follows:

$$r^\gamma W_J^r \;\Rightarrow\; W, \tag{35}$$

where W has the distribution function

$$P\{W \leq t\} = \begin{cases} 1 - \alpha(\tilde{\beta}) & \text{if } t = 0, \\ \alpha(\tilde{\beta})(1 - e^{-\xi_J\tilde{\beta}\mu}) & \text{otherwise.} \end{cases} \tag{36}$$

Here, $\tilde{\beta} = \lim_{r\to\infty}\beta^r$ and ξ_J are defined in (15).

To conclude this section, note that Equation (34) maps γ into the corresponding operational regime. Specifically, $\gamma = 1/2$ leads to the QED regime, while $\gamma > 1/2$ leads to the QD regime and $\gamma < 1/2$ leads to the ED regime. It is also important to note that Equation (36) implies that, under ITP, γ determines the order of magnitude of the waiting time of class J (which is of order $1/r^\gamma$), while for the other customer classes, ITP allows us to achieve extremely good service levels.

7. Adding Abandonments

Our results can be extended to the case where customers might abandon the system if their service does not begin within a certain time. Due to space considerations, we state here only our proposed solution while relegating the detailed analysis of this model to the online appendix. Consider, then, the same multiclass model studied so far, with the addition that class i customers have a finite patience which is exponential with rate θ_i. We assume the following ordering with respect to the abandonment rates θ_i, $i=1,\dots,J$.

Assumption 7.1. *The best-effort classes are the most patient ones, i.e.,* $\theta_J = \min_i \theta_i$.

Assumption 7.1 holds trivially if $\theta_i \equiv \theta$ for some $\theta \geq 0$. For the abandonment model, we consider the following formulation:

$$\begin{aligned} \text{minimize} \quad & N \\ \text{subject to} \quad & P\{Ab\} \leq \alpha, \\ & P\{W_i > T_i\} \leq \alpha_i, \quad i=1,\dots,J-1, \\ & N \in \mathbb{Z}_+, \quad \boldsymbol{\pi} \in \Pi, \end{aligned} \tag{37}$$

where $P\{Ab\}$ is defined as the steady-state fraction of customers that abandon before receiving service and $0<\alpha<1$.

Remark 7.1. Note that (37) differs from the nonabandonment formulation given in (1) with respect to the global constraint, which in (37) is associated with the fraction of abandoning customers rather than with the average waiting time in (1). This formulation is very natural in an environment that includes customer abandonment. Specifically, the fraction of abandoning customers is a very important measurement in call centers with impatient customers because it reflects, in some sense, the way that customers perceive the waiting time they experience. Hence, it is only natural to bound the fraction of abandoning customers rather than the waiting time itself. Moreover, in systems where each service rendered is associated with revenue, the number of abandoning customers becomes a measurement of economic importance.

Remark 7.2 (Admissible Policies). Naturally, in a finite patience setting, one can no longer expect all customers to be served because some will abandon (unless we have an infinite-server system). Instead, we require that customers cannot be blocked or routed somewhere else, and, formally, that $Q(t) = A(t) - D(t) - Z(t) - L(t)\ \forall t \geq 0$, where, in addition to the previously defined notation, $L(t)$ is the number of customers that abandoned by time t. That is, we require that all nonabandoning customers are served. Having this, we modify Π by replacing the requirement that all customers are served by the requirement that all nonabandoning customers are served. As before, we still require that the policies be nonanticipative and nonpreemptive, and that customers are served FCFS within each class.

Recalling that our staffing solution was based on a staffing problem for an associated $M/M/N$ queue, one would expect that in the abandonment case the solution would be based on a staffing problem for some $M/M/N+M$ (Erlang-A) system. This is indeed the case. To be able to formulate our solution, we define $P\{Ab\}^{\mathrm{FCFS}}_{\lambda,\mu,\theta_J}(N)$ to be the steady-state probability of abandonment in a single-class FCFS $M/M/N+M$ queue with arrival rate λ, service rate μ, patience rate θ_J, and N agents. We also redefine $W^{\mathrm{FCFS}}_{\lambda,\mu,\theta_J}(N)$ to be the steady-state waiting time for an $M/M/N+M$ single-class FCFS queue with arrival rate λ, service rate μ, patience rate θ_J, and N agents. Then, considering formulation (37) for the abandonment case, we have the following approximately optimal solution:

- **Staffing:** Find the staffing level through the single-class $M/M/N+M$ (or Erlang-A) model with arrival rate λ, service rate μ, abandonment rate θ_J, and FCFS service. Specifically, let

$$N^* = \mathrm{Min}\{N \in \mathbb{Z}_+ \colon P\{Ab\}^{\mathrm{FCFS}}_{\lambda,\mu,\theta_J}(N) \leq \alpha\}.^{7} \tag{38}$$

- **Control:** Use the ITP rule with the differences $\{K_{j+1}-K_j\}_{j\leq J-1}$ chosen recursively for $j=J-1,\dots,1$ in the following manner:

Compute

$$K_{j+1}-K_j = \left\lceil \frac{\ln(\alpha_j T_j/[P\{W_{j+1}>0\}\widehat{w}(N^*,\sigma_j,\sigma_{j-1})])}{\ln(\sigma_j)} \right\rceil \vee 0, \quad j=J-1,\dots,1, \tag{39}$$

where $\widehat{w}(N^*,\sigma_j,\sigma_{j-1}) = [N^*\mu(1-\sigma_j)(1-\sigma_{j-1})]^{-1}$.

Set

$$P\{W_j>0\} = P\{W_{j+1}>0\}\sigma_j^{K_{j+1}-K_j}. \tag{40}$$

In the above, we set $P\{W_J>0\} = P\{W^{\mathrm{FCFS}}_{\lambda,\mu,\theta_J}(N^*)>0\}$, and for two real numbers x and y, $x\vee y =: \max\{x,y\}$. The actual threshold values are then determined by setting $K_1=0$.

Note that in the abandonment setting, we only have the version of the threshold formula that uses the function $\widehat{w}$ (as in Remark 2.1 for the nonabandonment case) rather than any distribution function. The reason is that in the abandonment case, we do not have precise approximations for the waiting-time distributions of classes $i=1,\dots,J-1$. The threshold formula is based on Markov's inequality and is hence less precise than the formula we gave for the nonabandonment case. Still, using the thresholds given above is proved to be approximately optimal.

[7] This quantity can be calculated using the 4CC freeware introduced in Remark 2.3.

To conclude this section, we should point out that by using the relation $\lambda P\{Ab\} = \theta E[Q]$ (which holds for queues with exponential patience), one can easily generalize our results above to a variant of formulation (37) in which the individual waiting-time constraints are replaced with abandonment constraints of the form $P_i\{Ab\} \le \alpha_i$. Here, $P_i\{Ab\}$ is the steady-state fraction of abandoning customers from class i. Analogously to the waiting-time constraints, requiring that α_i is smaller in orders of magnitude than α, the ITP and SCS solution will be asymptotically optimal also for this variant of the formulation. Specifically, one would use SCS for staffing and choose the thresholds for classes $i = 1, \ldots, J-1$ so that $P_i\{Ab\} = \theta_i E[Q_i]/\lambda_i \le \alpha_i$.

8. Conclusions and Further Research

We study large-scale service systems with multiple customer classes and fully flexible servers. For such systems, we investigate the question of how many servers are needed and how to match them with customers so as to minimize staffing costs subject to service level, average speed of answer, and abandonment probability constraints. We find that a single-class staffing (SCS) rule and an idle-server-based threshold-priority (ITP) control are asymptotically optimal in the many-server heavy-traffic limiting regime. While the asymptotic optimality is established as the number of agents grows indefinitely, our proposed solution performs extremely well even for small and medium-sized systems.

The staffing level determined by the SCS rule is shown to depend on the overall system demand, and a global quality-of-service constraint only. This implies that because the staffing level does not depend on the class-level constraints, one may say that service-level differentiation is obtained "for free," in the sense that no additional servers are needed to satisfy those class-level constraints. Moreover, even if the class-dependent arrival rates or performance targets are unknown at the time when staffing decisions are made, the staffing levels remain unchanged.

Practically, the demand uncertainty becomes even more of an issue when the service is performed by a third party who has no access to demand information. This problem appears to be of increasing importance due to the proliferation of call center outsourcing. As it is often the case with subcontracting, the uncertainty associated with future demand together with information asymmetry can cause incentive misalignments between the two parties, which may result in system inefficiencies (e.g., Cachon and Lariviere 2001). To resolve these inefficiencies, a mechanism needs to be designed that would enforce the multidimensional demand information to be shared truthfully. But such multidimensional signalling problems are notoriously hard. Our insight reduces the problem into a one-dimensional one, that may be more tractable.

8.1. Directions for Future Research

While we believe that the managerial insights obtained through our analysis of a relatively simple model extend to more general settings, staffing and control solutions are still out of reach for the more general case where service rates depend on the customer class as well as the server pool. Even for the simplest extension of the V-model studied in this paper, one where the service rates are class-dependent, asymptotically optimal staffing and control are unknown at this point.

If one assumes further that not all servers can serve all customers, the problem then becomes even more complicated. Partial cross-training is not only a realistic scenario, but it was also shown by Wallace and Whitt (2005) and Pinker and Shumsky (2000) to be sufficient in obtaining satisfactory performance. This more general problem is difficult, and our paper is a step toward solving it, assuming servers are fully flexible. Our initial investigation suggests that the insights gained from studying the V-model are useful in analyzing more complicated network structures (Gurvich 2004). Other researchers have also tackled this general skill-based routing (SBR) problem (e.g., Bassamboo et al. 2006a, 2006b; Atar 2005; Tezcan and Dai 2007; Gurvich and Whitt 2006), but many issues remain unresolved.

Finally, our solution in this paper assumes that the global quality-of-service constraint is much less restrictive than the class-level constraints. But what if some of the classes have constraints associated with them that are as loosely restrictive as the global constraint? Or, what if the constraints are of a different form (e.g., $P\{\text{exists } i\text{: } W_i > T_i\} \le \alpha$)? In these scenarios, it is unclear what are asymptotically optimal staffing and control. We propose such problems as a topic for further investigation.

Electronic Companion

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

Acknowledgments

The authors thank the referees, associate editor, and special issue editor, whose careful reviews have lead to an essentially rewritten and much improved paper. This research was supported in part by BSF (Binational Science Foundation) Grant 2001685/2005175. The last author was also supported by ISF (Israeli Science Foundation) Grants 388/99, 126/02, and 1046/04, and by the Technion funds for the promotion of research and sponsored research.

References

Armony, M. 2005. Dynamic routing in large-scale service systems with heterogenous servers. *Queueing Systems* **51** 287–329.

Armony, M., C. Maglaras. 2004a. Contact centers with a call-back option and real-time delay information. *Oper. Res.* **52**(4) 527–545.

Armony, M., C. Maglaras. 2004b. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* **52**(2) 271–292.

Armony, M., A. Mandelbaum. 2007. Routing and staffing in large-scale service systems with heterogeneous servers and impatient customers. Working paper, New York University, New York.

Atar, R. 2005. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* **15**(4) 2606–2650.

Atar, R., A. Mandelbaum, M. Reiman. 2004. Scheduling a multi-class queue with many exponential servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* **14**(3) 1084–1134.

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* **51** 249–285.

Bell, S. L., R. J. Williams. 2001. Dynamic scheduling of a system with two parallel servers in heavy traffic with complete resource pooling: Asymptotic optimality of a continuous review threshold policy. *Ann. Appl. Probab.* **11** 608–649.

Bhulai, S., G. Koole. 2003. A queueing model for call blending in call centers. *IEEE Trans. Automatic Control* **48** 1434–1438.

Borst, S., A. Mandelbaum, A. M. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52**(1) 17–34.

Cachon, G. P., M. A. Lariviere. 2001. Contracting to assure supply: How to share demand forecasts in a supply chain. *Management Sci.* **47**(5) 629–646.

Deshpande, V., M. A. Cohen, K. Donohue. 2003. A threshold inventory rationing policy for service-differentiated demand classes. *Management Sci.* **49**(6) 683–703.

de Véricourt, F., J. P. Gayon, F. Keraesmen. 2004. Stock rationing in a multi-class make-to-stock queue with information on the production status. Working paper, Fuqua School of Business, Duke University, Durham, NC.

Gans, N., Y. P. Zhou. 2003. A call-routing problem with service-level constraints. *Oper. Res.* **51**(2) 255–271.

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

Gurvich, I. 2004. Design and control of the $M/M/N$ queue with multi-class customers and many servers. Masters thesis, Technion, Israel Institute of Technology, Haifa, Israel.

Gurvich, I., W. Whitt. 2006. Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing. Working paper, Columbia University, New York.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–587.

Harrison, J. M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the Halfin and Whitt heavy traffic regime. *Oper. Res.* **52**(2) 243–257.

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

Kella, O., U. Yechiali. 1985. Waiting times in the non-preemptive priority $M/M/c$ queue. *Stochastic Models* **1**(2) 257–262.

Koole, G. 2003. Redefining the service level in call centers. Technical report, Department of Stochastics, Vrije Universiteit, Amsterdam.

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. 2006. Staffing many server queues with impatient customers: Constraint satisfaction in call centers. Working paper, Technion, Israel Institute of Technology, Haifa, Israel.

Massey, A. W., B. R. Wallace. 2006. An optimal design of the $M/M/C/K$ queue for call centers. *Queueing Systems*. Forthcoming.

Milner, J. M., T. L. Olsen. 2008. Service-level agreements in call centers: Perils and prescriptions. *Management Sci.* **54**(2) 238–252.

Pinker, E. J., R. A. Shumsky. 2000. The efficiency-quality trade-off of cross-trained workers. *Manufacturing Service Oper. Management* **2**(1) 32–48.

Puhalskii, A., M. Reiman. 1998. A critically loaded multirate link with trunk reservation. *Queueing Systems* **28** 157–190.

Rafaeli, A., E. Kedmi, D. Vashdi, G. Barron. 2005. Queues and fairness: A multiple study investigation. Technical report, Technion, Israel Institute of Technology, Haifa, Israel.

Schaack, C., R. Larson. 1986. An N-server cutoff priority queue. *Oper. Res.* **34**(2) 257–266.

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

Tezcan, T., J. G. Dai. 2007. Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Oper. Res.* Forthcoming.

Towsley, D., F. Baccelli. 1991. Comparisons of service disciplines in a tandem queueing network with real-time constraints. *Oper. Res. Lett.* **10** 49–55.

Wallace, R. B., W. Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manufacturing Service Oper. Management* **7**(4) 276–294.

Whitt, W. 1984. Heavy traffic approximations for service systems with blocking. *AT&T Bell Laboratories Tech. J.* **63**(5) 689–708.

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

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

Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, Englewood Cliffs, NJ.

Yahalom, T., A. Mandelbaum. 2004. Optimal scheduling of a multi-server multi-class non-preemptive queueing system. Working paper, Technion, Israel Institute of Technology, Haifa, Israel.