Submitted to *Operations Research*
manuscript (Please, provide the mansucript number!)

Service-Level Differentiation in Many-Server Service Systems Via Queue-Ratio Routing

Itay Gurvich
Kellogg School of Management, Northwestern University, i-gurvich@kellogg.northwestern.edu
Ward Whitt
IEOR Department, Columbia University, ww2040@columbia.edu, www.columbia.edu/∼ww2040

Motivated by telephone call centers, we study large-scale service systems with multiple customer classes and multiple agent pools, each with many agents. To minimize staffing costs subject to service-level constraints, where we delicately balance the service levels (SLs) of the different classes, we propose a family of routing rules called *Fixed-Queue-Ratio* (FQR) rules. With FQR, a newly available agent next serves the customer from the head of the queue of the class (from among those he is eligible to serve) whose queue length most exceeds a specified proportion of the total queue length. The proportions can be set to achieve desired SL targets. The FQR rule achieves an important *state-space collapse* (SSC) as the total arrival rate increases, in which the individual queue lengths evolve as fixed proportions of the total queue length. In the current paper we consider a variety of service-level types and exploit SSC to construct asymptotically optimal solutions for the staffing-and-routing problem. The key assumption in the current paper is that the service rates depend only on the agent pool.

1. Introduction

Large call centers usually serve multiple classes of customers having different service requirements and different perceived value. The services provided by the call center agents usually require special skills, but it is usually not possible or cost-effective for all agents to have all skills. With current technology, call centers have the capability of routing calls to appropriate agents with the required skills, using some form of *Skill-Based Routing* (SBR), but it remains challenging to perform SBR effectively; see Section 5 of Gans et al. [9].

Call centers usually specify their operational objectives in the form of *Quality-of-Service* (QoS) *constraints*. Following common practice, we will focus on the *x-y service-level* (SL) *constraint*, which stipulates that $x\%$ of the calls should be answered within y seconds. We let the call center have

different SL constraints for different customer classes; e.g., with both regular and VIP customers, we might aim to respond to 80% of regular customers within 30 seconds, but 80% of VIP customers within 10 seconds.

In this context, the total problem has three components: design, staffing and routing. In the design phase, we start by grouping the customers into classes and the agents into service pools. (In doing so, we assume that the customers within classes are homogeneous, as are the agents within pools.) Then we must decide what skills should each pool have, i.e., which classes are they allowed to serve. In the staffing phase, we must decide how many agents should be in each service pool. Finally, in the routing phase, we must decide how the agents should be assigned to customers in real time.

The total problem is typically large and complex, so that it is unproductive to search for an optimal solution. Thus we look for a *good, simple solution*, which produces near-optimal performance in a relatively simple way. In particular, we hope to turn the large scale into an advantage instead of a disadvantage by finding relatively simple procedures that become more effective as the scale increases. Indeed, we want to find a relatively simple approach that is asymptotically optimal for specified problems as the scale increases. Our goal is to achieve *simplicity and asymptotic optimality.*

1.1. A Simple Intuitive Routing Rule: FQR

When considering possible controls, we think we should seek controls that are intuitive and structurally simple. Controls that lack any evident structure or insight are unlikely to be used by call center managers. A good example of a simple and intuitive control that is applicable to very general network structures (but essentially limited to single-agent service pools) is the *generalized-cμ* (*Gcμ*) *rule*, first introduced by Van-Meighem [17] for a model with multiple classes and a single server (also known as the V model), and generalized to more complicated networks by Mandelbaum and Stolyar [15]. A parallel to [15] in a many-server setting has been provided by Atar [3], who characterizes a family of controls that achieve asymptotically optimal performance in the QED

regime. (See [13] for more discussion.) While the controls in [3] can be implemented easily in a computerized environment, they are not nearly as simple as the $Gc\mu$ rule. Thus, it seems desirable to seek a family of controls for many-server systems that bridge the gap between the simple and intuitive $Gc\mu$ rule in [15] and the more complicated controls in [3].

With that goal in mind, we propose *Fixed-Queue-Ratio* (FQR) routing. We assume that there is a queue for each customer class. When an agent becomes free, he chooses the customer from the head of the line (from one of the classes he can serve) for which the queue length most exceeds a fixed proportion p_i of the total queue length (for all classes). The proportions p_i in turn are chosen to depend on the specified SL constraints. The FQR rule is a special case of the *Queue-and-Idleness-Ratio* (QIR) family of controls that we introduce in [12]. A consequence of [12] is that FQR makes the separate queue lengths *asymptotically proportional* to the total queue length. In other words, FQR produces a very important *state-space collapse* (SSC), causing the vector-valued queue-length process to evolve, asymptotically, as a one-dimensional process. In addition, FQR is a simple balancing rule like the $Gc\mu$ rule and, like the $Gc\mu$ rule, it is a highly decentralized control.

The key assumption that we make here is that the service rates are *pool-dependent*, i.e, that the service time of a customer depends on the type (pool) of the agent that provides the service but not on the customer class. Under this assumption, SSC reduces the multidimensional system dynamics to a tractable one-dimensional process. This reduction allows us to provide closed-form expressions for the staffing levels and prove that our proposed solution is not only feasible, but also asymptotically optimal.

Our solution stands on two pillars: (i) staffing via reduction of the multi-class multi-pool system to a single-class multi-pool system, and (ii) a simple routing rule that simultaneously makes the system perform as efficiently as the single-class multi-pool system and takes care of the service-level differentiation.

Our *reduction approach to staffing* illustrates our emphasis on simplicity: We propose first choosing the total number of agents by aggregating the service-level constraints and acting as if all

customers have access to all agents. Thus, we reduce the original SBR system into a single-class multi-pool call center, known as the inverted-V model; see Figure 1.

In the second step, we aim to satisfy the individual class-level QoS constraints by appropriately routing the customers in the system. Because the staffing and design decisions are highly interdependent, our proposed approach may seem naive and inadequate. However, we show that the FQR routing rule that we use in the second step guarantees that the two-step solution is nearly optimal, thus decomposing the joint optimization problem of design, staffing, and routing into more elementary problems, which can be addressed sequentially.

In our sequel [8], we remove the pool-dependence assumption and provide a general asymptotic feasibility (but not optimality) result. Based on the feasibility result, we then construct simple simulation-based optimization procedures to solve the design, staffing and routing problem for more general SBR systems.

1.2. Related Literature

Several recent papers have used the simplified reduction approach to staffing. Theoretical support is contained in Armony [1] and Gurvich et. al. [10]. These papers established asymptotic optimality of that staffing approach with appropriate routing for special classes of models as the total arrival rate increases. The first paper considered models with a single customer class and multiple agent types, while the second considered symmetric models with multiple customer classes but a single agent pool. Their asymptotic optimality follows Borst et al. [7], which formulated and established asymptotic optimality for the single-class, single-pool $M/M/N$ queue. The asymptotic framework is the now-familiar many-server heavy-traffic limiting regime, introduced by Halfin and Whitt [14], which is also known as the *Quality-and-Efficiency-Driven* (QED) *regime*. In the QED regime the arrival rate and numbers of servers both increase, while the service-time distribution remains unchanged. These two limits are coordinated so that the probability of delay approaches a limit strictly between 0 and 1. Borst et al. [7] showed that the QED regime arises naturally from economic considerations. We will be considering the QED regime throughout this paper.

The simplified reduction approach to staffing was also a central idea in Wallace and Whitt [19], which developed a simulation-based iterative algorithm for staffing an SBR call center that starts by choosing an initial total number of agents by acting as if the call center were a single-class single-skill call center. After initial skill requirements are assigned, simulation is used iteratively to find detailed staffing and skill requirements so that the SL and other QoS constraints are met. The approach in [19] has two shortcomings, which we address here. First, that approach requires an iterative simulation algorithm to adjust staffing levels and skill assignments in order to satisfy the class-dependent QoS constraints. Since service is performed in a relatively short time scale compared to staffing, we think it should be more effective to primarily rely on the routing rather than the staffing in order to achieve desired service differentiation. In this paper we provide a way to do that. Second, while the approach in [19] seems to become more effective as the scale increases, it has not yet been shown to be asymptotically feasible or optimal as the scale increases. Here, in contrast, we establish asymptotic optimality. In this paper we assume that the service rates are *pool-dependent*. A special case is the system with common service rates, e.g., as considered in [19].

The analysis in this paper relies heavily on our previous paper [12], which establishes SSC results for the QIR generalization of FQR. We establish asymptotic optimality for the case of convex holding costs in [13].

An important contribution here is simultaneously addressing the three problems of design, staffing and routing. Conventionally, these are treated separately and hierarchically. Wallace and Whitt [19] also addressed these three problems together, but the only previous work we are aware of that establishes asymptotic feasibility or optimality for all three problems is Bassamboo et. al. [4, 5, 6]. They establish asymptotic optimality for the problem of minimizing costs associated with waiting, abandonments and customer rejections. In [6] they also consider abandonment constraints but not tail-probability SL constraints. Their analysis is interesting because it focuses on uncertainty in the arrival rates. As a consequence, they consider a different limiting regime, the efficiency-driven (ED) regime. Their general setting allowing for uncertainty in the arrival rates comes at the price of having to restrict the analysis to a cruder notion of asymptotic optimality

than the one we use here. Our finer analysis, while more limited in its scope, allows us to identify key system characteristics and, in turn, to construct intuitive routing schemes. Moreover, it allows us to tackle directly waiting-time tail-probability SL constraints that are widely used in the industry. In addition, the routing scheme we propose in this paper is used in Gurvich, Luedtke and Tezcan [11] to construct a staffing and routing algorithm for a call center with uncertain arrival rates operating under SL constraints.

Within the context of single-server stations, several papers have tackled the problem of SL constraint satisfaction. Notable is Van-Meighem [18], which embeds the constraint-satisfaction problem into the convex-holding-cost setting of [17], rather than dealing with it directly.

Organization of the paper: In §2 we introduce the model and initial problem formulation. The proposed design, staffing and routing solution is introduced in §3 and the asymptotic optimality results are stated in §4. We introduce and solve additional problem formulations in §5. We state conclusions in §6. Some proofs and auxiliary results appear in the e-companion.

2. The SBR System and the Problem Formulation

We consider a system with a set $\mathcal{I} := \{1, \ldots, I\}$ of customer classes and a set $\mathcal{J} := \{1, \ldots, J\}$ of agent types. The number of agents of type j (which will be a decision variable) is denoted by N_j; let $N := (N_1, \ldots, N_J)$. Class-i customers arrive according to a Poisson process with rate λ_i and $\lambda := \sum_{i \in \mathcal{I}} \lambda_i$ is the aggregate arrival rate. If a type-j agent can serve a class-i customer, we let $\mu_{i,j}$ be the corresponding service rate. Alternatively, the mean handling time of a class-i customer by a type-j agent is $1/\mu_{i,j}$. **A key assumption in this paper is that the service rates are pool-dependent.** That is, for each $j \in \mathcal{J}$, we have $\mu_{i,j} = \mu_j$ for all $i \in \mathcal{I}$ that can be served by pool j. Throughout we will assume, without loss of generality, that the service rates are labelled in decreasing order:

$$\mu_1 \geq \ldots \mu_2 \geq \ldots \geq \mu_J.$$

At this point, we do not consider customer abandonment; see §5.2 for extensions to that case.

The possible-routing graph for this SBR system has a natural representation as a bipartite graph (see Figure 1) with vertices $V = \mathcal{J} \bigcup \mathcal{I}$; i.e., V is the union of the set of customer classes and the set of agent pools. Then, the only edges we consider connect customer classes to agent pools: $E = \{(i,j) \in \mathcal{I} \times \mathcal{J} : j \text{ can serve } i\}$. An edge (i,j) is present in the routing graph if class-i customers can be served by type-j agents. We let $I(j)$ be the set of customer classes that can be served by pool j, i.e, $I(j) := \{i \in \mathcal{I} : (i,j) \in E\}$ and, symmetrically, we define $J(i) := \{i \in \mathcal{I} : (i,j) \in E\}$ to be the set of pools that are qualified to serve class-i customers. In addition, we assume that agents of type-j incur a cost of c_j per unit of time. We also will allow the system to impose additional constraints on the staffing vector to reflect union contracts, hiring and training constraints or other managerial considerations; see §2.1 for the formal modeling of these constraints.

For the asymptotic analysis, we will construct a sequence of SBR systems indexed by the aggregate arrival rate λ. The service rates μ_j and the routing graph are held fixed. We will make the dependence on the index explicit by adding the superscript λ to all the relevant parameters and processes. We assume that the ratios $a_i := \lambda_i/\lambda$ remain constant for all λ. Also we let the SL target T_i^λ scale with λ to put the system into the QED regime.

Assumption 2.1 (**QED scaling for SL targets**) *The SL targets $T_i^\lambda, i \in \mathcal{I}$, are scaled so that $T_i^\lambda = \bar{T}_i/\sqrt{\lambda}$, for some strictly positive constants $\bar{T}_i,\ i \in \mathcal{I}$.*

The proposed staffing and routing solution will be completely defined in terms of the original targets T_i^λ; there will be no use of the constants $\bar{T}_i$. The scaling is only used for the proof of asymptotic optimality.

2.1. The Problem Formulation

To formulate our optimization problem, let $A_i^\lambda(t)$ be the number of class-i customers to arrive by time t. Let $\bar{W}^{\lambda,T}$ be the average waiting time of all customers that arrived up to time T; let $F_i^{\lambda,T}(\cdot)$ be the empirical distribution of the waiting time of class-i customers up to time T; and let $\bar{F}_i^{\lambda,T}(\cdot)$ be its complement; i.e.,

$$\bar{W}^{\lambda,T} := \frac{\sum_{i=1}^{I}\sum_{k=1}^{A_i^\lambda(T)} w_{i,k}^\lambda}{A^\lambda(T)} \text{ and } \bar{F}_i^{\lambda,T}(y) := \frac{\sum_{k=1}^{A_i^\lambda(T)} \mathbf{1}\{w_{i,k}^\lambda > y\}}{A_i^\lambda(T)} \ , \tag{1}$$

where $A^\lambda(T) := \sum_{i\in\mathcal{I}} A_i^\lambda(T)$, $w_{i,k}^\lambda$ is the realized waiting time of the k^{th} class-i customer to arrive to the system after time 0 and $\mathbf{1}B$ is the indicator of the event B, which is equal to 1 if B occurs and 0 otherwise. An initial formulation, representing common call center goals, can then be stated as follows:

$$\begin{array}{ll} \text{minimize} & \sum_{j\in\mathcal{J}} c_j N_j \\ \text{subject to:} & \limsup_{T\to\infty} \bar{F}_i^{\lambda,T}(T_i^\lambda) \le \alpha, \quad 1 \le i \le I, \\ & N \in \mathcal{A}^\lambda, \quad \pi \in \Pi, \end{array} \tag{2}$$

where $\mathcal{A}^\lambda$ is a subset of $\mathbb{Z}_+^J$ that is generated by linear constraints. Specifically, we allow constraints of the form $N \in \mathcal{A}^\lambda := \mathcal{A}_b^\lambda$, where $\mathcal{A}_b^\lambda = \{N \in \mathbb{Z}_+^J : A \cdot N \le b^\lambda\}$, for some matrix $A \in \mathbb{R}^{d\times J}$, $d \in \mathbb{Z}_+$ and $b^\lambda = \lambda\hat{b}$ for some $\hat{b} \in \mathbb{R}^d$. We also let $\tilde{\mathcal{A}}^\lambda := \tilde{\mathcal{A}}_b^\lambda$ be the set obtained from $\mathcal{A}^\lambda$ by relaxing the integrality assumptions. That is, $\tilde{\mathcal{A}}^\lambda = \{N \in \mathbb{R}_+^J : A \cdot N \le b^\lambda\}$. If the set of possible staffing vectors is unconstrained we set $\mathcal{A}^\lambda \equiv \mathbb{Z}_+$.

We observe that, to consider optimality, (2) is not well formulated, because we have not yet sufficiently constrained the policies. So far, the formulation permits giving some customers satisfactory performance at the expense of giving other customers (in the proportion $1-\alpha$) arbitrarily poor performance. This problem is discussed extensively at the end of §2 of [10], so we will be brief here. To illustrate the difficulties, note that we could elect not to serve class-i customers who have waited longer than T_i. Even if we required first-come first-served (FCFS) service within each class, we could satisfy all the constraints with relatively limited staffing by disallowing any waiting, i.e., by using a pure-loss model. Clearly, in a loss model, all the customers that do enter the system do not experience any wait, and we may choose the number of agents so that the blocking probability is less than $1-\alpha$. That is clearly an undesirable outcome since many customers are blocked and do not receive service at all. Even when requiring that all customers be served, highly undesirable polices are possible such as the alternating-priority control discussed by Gurvich et. al. [10].

As a consequence, we modify (2) by adding an additional constraint; in particular, we initially consider the following **best-effort optimization problem:**

$$\begin{array}{ll} \text{minimize} & \sum_{j\in\mathcal{J}} c_j N_j \\ \text{subject to:} & \limsup_{T\to\infty} \bar{W}^{\lambda,T} \le T_I^\lambda, \\ & \limsup_{T\to\infty} \bar{F}_i^{\lambda,T}(T_i^\lambda) \le \alpha, \quad 1 \le i \le I-1, \\ & N \in \mathcal{A}^\lambda, \quad \pi \in \Pi. \end{array} \tag{3}$$

We emphasize that it is not sufficient to add the global average-waiting-time constraint. It is important also to remove the individual SL constraint of class I. Otherwise, the problems with formulation (2) remain unresolved. Indeed, if the global average-waiting-time constraint is very loose one can show that it is possible to construct alternating priority controls as illustrated in [10] to construct policies under which, in part of the time, each class experiences extremely low service levels.

The difficulties in formulation (2) can be avoided in other ways; e.g., considering only average-waiting-time constraints. Such a formulation and its corresponding solution are considered in §5. We first focus on the formulation (3).

We now define the set of admissible policies Π. To this end, we say that **all customers are served** if it is not allowed to block or overflow customers; i.e., we require that, for all $t \ge 0$, $Q_i(t) = A_i(t) - D_i(t) - Z_i(t)$, where $D_i(t)$ and $Z_i(t)$ are, respectively, the number of class-i departures from the system up to time t and the number of class-i customers in service at time t.

We say that a routing policy is **non-anticipative** if a decision at any time is based on the history up to that time and not upon future events. We say that a routing policy is **non-preemptive** if customers stay in service with the agent first assigned to them until their service is complete once an agent has been assigned.

Definition 2.1 (**admissible routing policies**) *We say that a routing policy π is **admissible** if:* (1) *it is non-anticipative,* (2) *it is non-preemptive, and* (3) *all customers are served. Let Π be the set of all admissible routing policies.*

We conclude this section with the definition of asymptotic feasibility.

Definition 2.2 (**asymptotic feasibility**) *A sequence of staffing vectors and routing policies*

$\{(N^\lambda, \pi^\lambda)\}$ *is asymptotically feasible for* (3) *if (a)* $\pi^\lambda \in \Pi$*, for all* λ*, and (b) for every* $\epsilon > 0$*, there exists* $T^*(\epsilon)$ *such that, for all* $T \geq T^*(\epsilon)$*,*

$$\limsup_{\lambda\to\infty} P\left\{\frac{\bar{W}^{\lambda,T}}{T_I^\lambda} \geq 1+\epsilon\right\} \leq \epsilon, \tag{4}$$

and

$$\limsup_{\lambda\to\infty} P\left\{\bar{F}_i^{\lambda,T}(T_i^\lambda) \geq \alpha+\epsilon\right\} \leq \epsilon, \quad 1 \leq i \leq I-1. \tag{5}$$

Asymptotic feasibility holds for (2) *instead of* (3) *if* (4) *and* (5) *above are replaced by*

$$\limsup_{\lambda\to\infty} P\left\{\bar{F}_i^{\lambda,T}(T_i^\lambda) \geq \alpha+\epsilon\right\} \leq \epsilon, \quad i \in \mathcal{I}. \tag{6}$$

Given two positive real-valued functions f and g, we say that $f(x)$ is $o(g(x))$ (as $x \to \infty$) if $f(x)/g(x) \to 0$ as $x \to \infty$; we say that $f(x)$ is $O(g(x))$ if $f(x)/g(x$ is bounded as $x \to \infty$. The definition of asymptotic optimality will be the same for all the formulations that we consider in this paper. Hence, we do not specify a specific formulation within the definition. For the rest of the paper asymptotic optimality will always be in the sense of Definition 2.3 below, where asymptotic feasibility will depend on the context. The $o(\sqrt{\lambda})$ in the asymptotic-optimality condition (7) below corresponds to asymptotic optimality in the diffusion scale, which is more refined than asymptotic optimality in the fluid scale, which would involve a larger bound on the error of $o(\lambda)$ as $\lambda \to \infty$. (The staffing levels will be $O(\lambda)$.)

Definition 2.3 (**asymptotic optimality**) *A sequence of staffing vectors and routing policies* $\{(N^\lambda, \pi^\lambda)\}$ *is asymptotically optimal if it is asymptotically feasible and,*

$$[c \cdot N^\lambda - c \cdot \tilde{N}^\lambda]^+ = o(\sqrt{\lambda}) \quad as \quad \lambda \to \infty \ , \tag{7}$$

for any other sequence $\{(\tilde{N}^\lambda, \pi^\lambda)\}$ *of asymptotically feasible staffing vectors and routing policies.*

We end this section with a brief discussion of the relation between our notions of asymptotic feasibility and optimality and the, more traditional, steady-state feasibility and optimality.

Remark 2.1 (**steady-state constraints**) While actual call center operations involve finite horizon decisions and constraints, the traditional way of call-center modeling would be to write both (2) and (3) with steady-state constraints instead of the finite horizon ones. To obtain the steady-state formulation, one would replace the individual tail constraints with $P\{W_i^\lambda(\infty) > T_i^\lambda\} \le \alpha$, where $W_i^\lambda(\infty)$ is the class-i steady-state waiting time. The global average delay constraint is replaced with the constraint $E[W^\lambda(\infty)] \le T_I^\lambda$ where $W^\lambda(\infty)$ is the steady-state global waiting time, i.e, $W^\lambda(\infty)$ is equal in distribution to $\sum_{i\in\mathcal{I}}(\lambda_i/\lambda)W_i^\lambda(\infty)$. Even though these alternative formulations differ little from a practical perspective, they are mathematically different. They are asymptotically equivalent only if one can establish a certain limit-interchange result. Such limit-interchange arguments are elementary for some models, such as the inverted-V model, but it is a complex task for the general SBR setting; In this paper we restrict the attention to the long-run average formulation in (2) and (3). What we do parallels what is done with simulation. ∎

2.2. A Lower Bound

Our solution will be based on a reduction of the SBR system to a more elementary model in which multiple agent types serve a single customer class, also known as the inverted-V (or $\bigwedge$) model. Given an SBR system, the associated $\bigwedge$ model has the same set of agent-pools $\mathcal{J}$, the same staffing levels $\{N_j,\ j\in\mathcal{J}\}$ and the same service rates $\{\mu_j,\ j\in\mathcal{J}\}$. In addition, the arrival rate of the single customer class is λ–the sum of the arrival rates in the SBR system. An example of an SBR system and its corresponding $\bigwedge$ model is given in Figure 1.

Clearly, the $\bigwedge$ model is not as simple as the $M/M/N$ queue. However, when it is optimally operated, its asymptotic performance leads to simple expressions for staffing, as has been shown by Armony [1]. We will exploit the results in [1] here. In particular, we will exploit a result for the $\bigwedge$ model, stating that

$$E[W^\lambda(\infty)] \le T_I^\lambda \quad \text{only if} \quad \sum_j \mu_j N_j^\lambda \ge \lambda + \beta^*\sqrt{\lambda} + o(\sqrt{\lambda}), \tag{8}$$

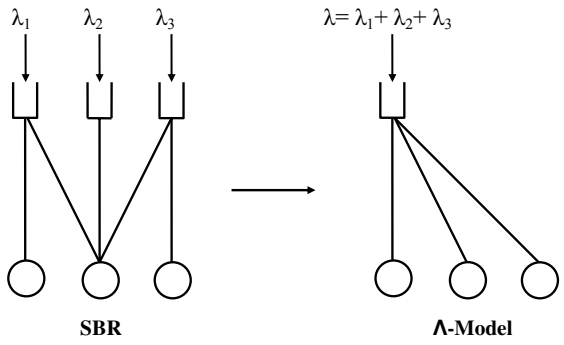

Figure 1 An SBR model and its corresponding $\bigwedge$ model

where β^* is the unique solution to

$$\frac{\mathbf{P}_{\mu_1}(\beta)}{\beta} = \sqrt{\lambda} T_I^\lambda =: \bar{T}_I, \quad \text{and} \quad \mathbf{P}_{\mu_1}(\beta) := \left[1 + \frac{\frac{\beta}{\sqrt{\mu_1}} \Phi\left(\frac{\beta}{\sqrt{\mu_1}}\right)}{\phi\left(\frac{\beta}{\sqrt{\mu_1}}\right)} \right]^{-1}, \tag{9}$$

with $\phi(\cdot)$ and $\Phi(\cdot)$ being, respectively, the standard normal pdf and cdf. Here, $\mathbf{P}_{\mu_1}(\beta)$ is the asymptotic delay probability in the $\bigwedge$ model operated under the Fastest-Server-First (FSF) policy as introduced by Armony [1]. That is, assuming that $\sum_{j \in \mathcal{J}} \mu_j N_j^\lambda = \lambda + \beta\sqrt{\lambda} + o(\sqrt{\lambda})$ and that FSF is used (plus additional technical conditions), [1] shows that $P\{W^\lambda(\infty) > 0\} \to \mathbf{P}_{\mu_1}(\beta)$, with $W^\lambda(\infty)$ being the steady-state waiting time in the $\bigwedge$ model.

We note that the necessary condition (8) is given in [1] for steady-state asymptotic feasibility, while we focus here on the, somewhat weaker, notion that appears in Definition 2.2. However, we find that the same necessary condition holds if one considers the same $\bigwedge$ model but with long-run average constraints (as in (3)) and with our notion of asymptotic feasibility. This is proved in Lemma EC.2.2 of the online appendix which, in turn, is a key step in the proof of Theorem 2.1 below.

We now use this necessary condition to construct a lower bound for the staffing of the SBR model. In doing this construction we will use two facts:

(i) In contrast to the SBR system, customers in the $\bigwedge$ model have access to all agent pools. It is intuitively clear, then, that if a given staffing vector is not sufficient for the given aggregate-waiting-time target in the $\bigwedge$ model, it will also not be sufficient in the less efficient SBR system.

Consequently, to meet the global waiting-time constraint it is necessary that $\sum_{j\in\mathcal{J}}\mu_j N_j^\lambda \geq \lambda + \beta^*\sqrt{\lambda} + o(\sqrt{\lambda})$.

(ii) As we have no abandonments in the system, the capacity should suffice to serve all customers (at least at a fluid scale). In particular, any feasible staffing vector must satisfy that

$$\sum_{j\in J(i)} \mu_j N_j y_{i,j} \geq \lambda_i, \quad i\in\mathcal{I},$$

for some vector y such that $\sum_{i\in I(j)} y_{i,j} \leq 1$ and $y_{i,j},\ (i,j)\in E$ are positive.

Together, these two informal arguments suggest that an asymptotic lower bound for the optimization problem (3) should be given by

$$\begin{array}{ll} \text{Minimize} & \sum_{j\in\mathcal{J}} c_j N_j \\ \text{Subject to:} & \sum_{j\in J} \mu_j N_j \geq \lambda + \beta^*\sqrt{\lambda}, \\ & \sum_{j\in J(i)} \mu_j N_j y_{i,j} \geq \lambda_i, \quad i\in\mathcal{I}, \\ & \sum_{i\in I(j)} y_{i,j} \leq 1, \quad j\in\mathcal{J}, \\ & N\in\mathcal{A}^\lambda, \quad y_{i,j}\geq 0, \quad (i,j)\in E\ . \end{array} \tag{10}$$

We call a staffing vector determined through the solution of (10) a **$\bigwedge$-based staffing**. A standard argument shows that the optimization problem (10) can be solved by solving an associated LP that yields the same set of optimal solutions. Since we are not concerned with the way in which (10) is solved, we will use (10) directly. The next theorem provides the formal lower-bound result. Its proof is given in the online appendix.

Theorem 2.1 (**lower-bound capacity**) *Consider the sequence of SBR systems and let* $\{(N^\lambda,\pi^\lambda)\}$ *be a sequence of asymptotically feasible staffing and routing rules such that*

$$\liminf_{\lambda\to\infty} \frac{N_j^\lambda}{\lambda} > 0,\ j\in\mathcal{J}.$$

Then, $\{N^\lambda,\ \lambda\geq 0\}$ *satisfies that*

$$\sum_{j\in\mathcal{J}} \mu_j N_j^\lambda \geq \lambda + \beta^*\sqrt{\lambda} + o(\sqrt{\lambda}), \tag{11}$$

where β^* *is the* $\bigwedge$*-model parameter in* (9).

Theorem 2.1 only provides a lower bound. We will next propose a solution that, we will prove, achieves this lower bound. The solution will be based on the $\bigwedge$-based staffing and the FQR routing rule.

3. The Proposed Solution

Our solution consists of a staffing component and a routing component. The staffing that we use is the $\bigwedge$-based staffing determined by an optimal solution to (10). For the routing component, we use FQR with ratios that will be explicitly determined as functions of the service level targets T_i^λ.

Let $Q_i^\lambda(t)$ be the number of class-i customers in queue. Let $Z_{i,j}^\lambda(t)$ be the number of type-j servers busy giving service to class-i customers, so that $X_i^\lambda(t) := Q_i^\lambda(t) + \sum_{j\in\mathcal{J}} Z_{i,j}^\lambda$ is the overall number of class-i customers present in the system at time t, and $I_j^\lambda(t) := N_j^\lambda - \sum_{i=1}^I Z_{i,j}^\lambda(t)$ be the number of idle agents in pool j at time t in the λ^{th} system. Accordingly, $I_\Sigma^\lambda(t) := \sum_{j=1}^J I_j^\lambda(t)$ is the total number of idle agents in the system. Let $X_\Sigma^\lambda(t)$ be the overall number of customers in the system (in service and in queue), i.e.,

$$X_\Sigma^\lambda(t) := \sum_{i=1}^I X_i^\lambda(t) = \sum_{i=1}^I \left(Q_i^\lambda(t) + \sum_{j=1}^J Z_{i,j}^\lambda(t) \right) ,$$

and let $N_\Sigma^\lambda(t) := \sum_{j\in\mathcal{J}} N_j^\lambda$ be the aggregate number of agents. Below we use $argmax$ and let it have the standard definition; i.e., given a function $f : A \mapsto \mathbb{R}$, with A a finite set, let $\arg\max f := \{y \in A : f(y) = \max_{x\in A} f(x)\}$.

Definition 3.1 (**FQR for the SBR model**) *Given two probability vectors $v := \{v_j : j \in \mathcal{J}\}$ and $p := \{p_i : i \in \mathcal{I}\}$, FQR for the SBR model is defined as follows:*

- **Upon arrival of a class-i customer** *at time t, the customer will be routed to an available agent in pool j^*, where*

$$j^* \equiv j^*(t) \in \underset{j\in J(i), I_j^\lambda(t)>0}{\arg\max} \{I_j^\lambda(t) - v_j[X_\Sigma^\lambda(t) - N_\Sigma^\lambda]^-\};$$

i.e., the customer will be routed to an agent pool with the greatest idleness imbalance. If there are no such agents, the customer waits in queue i, to be served in order of arrival.

- **Upon service completion by a type-j agent** *at time t, the agent will admit to service the customer from the head of queue i^* where*

$$i^* \equiv i^*(t) \in \underset{i\in I(j), Q_i^\lambda(t)>0}{\arg\max} \left\{ Q_i^\lambda(t) - p_i[X_\Sigma^\lambda(t) - N_\Sigma^\lambda]^+ \right\};$$

i.e., the agent will admit a customer from the queue with the greatest queue imbalance. If there are no such customers, the agent will remain idle.

Ties are broken in an arbitrary but consistent manner, so that the vector-valued stochastic process (Q^λ, Z^λ) is a CTMC with stationary transition probabilities.

To explicitly express the dependence on the vectors p and v we will use the notation $FQR(p,v)$. We point out that, if $p_i > 0$ for all $i \in \mathcal{I}$, then FQR is equivalently given by having each newly available agent choose the customer from the head of queue i^*, where

$$i^* \equiv i^*(t) \in \underset{i\in I(j), Q_i^\lambda(t)>0}{\arg\max} \left\{ \frac{Q_i^\lambda(t)}{p_i} \right\}, \tag{12}$$

which makes the use of $[X_\Sigma^\lambda(t) - N_\Sigma^\lambda]^+$ unnecessary.

Choosing p and v: For the routing component of our solution we will be using FQR with the ratio vectors (p^*, v^*), where $v^* = (0, \ldots, 0, 1)$ and p^* is the unique solution to

$$\mathbf{P}_{\mu_1}(\beta^*) e^{-\frac{\lambda_{I-1}}{\lambda p_{I-1}} \beta^* \sqrt{\lambda} T_{I-1}^\lambda} = \alpha \quad \text{and} \quad \frac{p_i}{p_{I-1}} = \frac{\lambda_i T_i^\lambda}{\lambda_{I-1} T_{I-1}^\lambda}, \tag{13}$$

for $\mathbf{P}_{\mu_1}(\beta)$ defined in (9). Since $\sqrt{\lambda} T_I^\lambda = \bar{T}_i$ (see Assumption 2.1) and $\lambda_i/\lambda_{I-1} = a_i/a_{I-1}$, the value of p_{I-1}^* is independent of λ. Consequently, so are the values p_i^* for $i \neq I-1$. The choice of the ratio vector p^* will be justified by (i) a sample-path version of Little's law that holds for many-server service system (ii) the SSC that is induced by FQR and, (iii) the fact that it performs asymptotically as efficiently as the $\bigwedge$ model. Informally, SSC justifies the following sequence of approximations

$$\begin{aligned} P\{W_{I-1} > T_{I-1}\} &\approx P\left\{Q_{I-1} > \lambda_{I-1} T_{I-1}^\lambda\right\} \approx P\left\{p_{I-1} Q_\Sigma > \lambda_{I-1} T_{I-1}^\lambda\right\} \approx P\left\{Q_\Sigma^{\bigwedge} > \tfrac{\lambda_{I-1} T_{I-1}^\lambda}{p_{I-1}}\right\} \\ &\approx P\left\{Q_\Sigma^{\bigwedge} > 0\right\} e^{\frac{\left(\sum_{j\in\mathcal{J}} \mu_j N_j - \lambda\right)}{\lambda} \frac{\lambda_{I-1} T_{I-1}^\lambda}{p_{I-1}}}, \end{aligned}$$

where $Q_\Sigma^{\bigwedge}$ is the steady-state queue length in the $\bigwedge$ model that is constructed from the SBR system as in the beginning of §2.2. The last step follows from simple expressions for the distribution of the queue length for the $\bigwedge$ model. By the analysis of $\bigwedge$ model in [1], the probability $P\{Q_\Sigma^{\bigwedge} > 0\}$ converges to $\mathbf{P}_{\mu_1}(\beta^*)$ if we use the $\bigwedge$-based staffing. Also, $\sum_{j\in\mathcal{J}} \mu_j N_j = \lambda + \beta^*\sqrt{\lambda} + o(\sqrt{\lambda})$. Hence,

$$P\left\{W_{I-1} > T_{I-1}\right\} \approx P\left\{Q_\Sigma^{\bigwedge} > \frac{\lambda_{I-1} T_{I-1}^\lambda}{p_{I-1}}\right\} \approx \mathbf{P}_{\mu_1}(\beta^*) e^{-\frac{\lambda_{I-1}}{\lambda p_{I-1}}\beta^*\sqrt{\lambda} T_{I-1}^\lambda}. \tag{14}$$

Similar informal arguments can be repeated for each of the customer classes. Finally, as β^* was chosen so that the right hand side in (14) equals α, the $\bigwedge$-based staffing and FQR should provide an asymptotically feasible solution to (3). Since the $\bigwedge$-based staffing is a lower bound, this solution is also asymptotically optimal. This informal argument is formalized in the next section.

4. Asymptotic Feasibility and Optimality

We begin to consider the limiting behavior as $\lambda \to \infty$. We will show that the $\bigwedge$-based staffing and FQR, with appropriately chosen ratios, yield an asymptotically feasible and optimal solution for (3). First, however, we consider the design of the system. In order to identify the design, it suffices to look at (10) with one constraint removed, i.e, consider the following non-linear optimization problem:

$$\begin{array}{ll} \text{Minimize} & \sum_{j\in\mathcal{J}} c_j N_j \\ \text{Subject to:} & \sum_{j\in J(i)} \mu_j N_j y_{i,j} \geq \lambda_i, \quad i \in \mathcal{I}, \\ & \sum_{i\in I(j)} y_{i,j} \leq 1, \quad j \in \mathcal{J}, \\ & N \in \tilde{\mathcal{A}}^\lambda, \quad y_{i,j} \geq 0, \quad (i,j) \in E\ . \end{array} \tag{15}$$

The solution to the mathematical program (15) can be regarded as a first-order deterministic fluid approximation for the SBR system, as in [20]. From that point of view, given a selected solution $(\bar{N}, \bar{y})$, we would then use $\bar{N}$ to provide an initial estimate of the staffing and $\bar{y}$ to provide an initial estimate of the appropriate routing. We point out that the solution to (15) is independent of λ. Indeed, by the definition of $\tilde{\mathcal{A}}^\lambda$ and the assumption that $\lambda_i = a_i\lambda$, (15) is equivalent to the mathematical program:

$$\begin{array}{ll} \text{Minimize} & \sum_{j\in\mathcal{J}} c_j \nu_j \\ \text{Subject to:} & \sum_{j\in J(i)} \mu_j \nu_j x_{i,j} \geq a_i, \quad i \in \mathcal{I}, \\ & \sum_{i\in I(j)} x_{i,j} \leq 1, \quad j \in \mathcal{J}, \\ & A\nu \leq b, \\ & \nu_j \geq 0, \quad x_{i,j} \geq 0, \quad j \in \mathcal{J},\ (i,j) \in E\ . \end{array} \tag{16}$$

Both mathematical programs (15) and (16) can be replaced with linear programs (LP) that yield the same optimal solution. Henceforth we only refer to optimal solutions of (15), without considering how they are obtained. We denote an optimal solution to (15) by $(\bar{N}^\lambda, \bar{y}^\lambda)$. Our first assumption requires a weak form of uniqueness of optimal solutions to (15).

Assumption 4.1 (**uniquness of staffing**) *Fix λ and let $(\bar{N}^\lambda, \bar{y}^\lambda)$ and $(\tilde{N}^\lambda, \tilde{y}^\lambda)$ be two optimal solutions to* (15)*. Then $\bar{N}^\lambda = \tilde{N}^\lambda$.*

We note that, due to the equivalence between (15) and (16), if Assumption 4.1 holds for a given λ, then it holds for all λ. Similarly, the equivalence between (15) and (16) implies that, if $\bar{N}_j^\lambda > 0$ for one λ, then the same holds for all values of λ. Informally, Assumption 4.1 is required as we will want to use the optimal solutions of (15) as a first order (fluid) approximation for the staffing (and not only the staffing cost) that we get from the $\bigwedge$-based staffing as defined by the solution to (10). In some simplified settings, such as the one considered in §5.4, this assumption can be removed.

The second assumption is a critical loading assumption, needed to put the system in heavy traffic.

Assumption 4.2 (**critical loading**) *For any $\lambda \geq 0$, $\sum_{j\in\mathcal{J}} \mu_j \bar{N}_j^\lambda = \lambda$ for any optimal solution $(\bar{N}^\lambda, \bar{y}^\lambda)$ to* (15)*.*

Finally, we make the following structural assumption. Below, E and V are as defined in the beginning of §2. Also, we say that a graph is connected if there exists a path between every two nodes in the graph.

Assumption 4.3 (**connected routing graph**) *For any $\lambda \geq 0$, there exists an optimal solution $(\bar{N}^\lambda, \bar{y}^\lambda)$ for* (15) *such that the graph $\mathcal{E}(\bar{N}^\lambda, \bar{y}^\lambda) := \{(i,j) \in \mathcal{I}\times\mathcal{J} : \bar{y}_{i,j}^\lambda > 0\}$ is a connected subgraph of $G(V,E)$.*

As before, the equivalence between (15) and (16) guarantees that, if Assumption 4.3 holds for a given λ, then it holds for all λ. This connected-graph assumption is crucial for the ability to

instantaneously balance the system by re-directing capacity from one customer class to the other; see §2.7 of Atar [3] for elaboration. Assumptions 4.1-4.3 are assumed to hold throughout the rest of the paper. With the above definitions, we can state our asymptotic optimality result.

Theorem 4.1 (**asymptotic optimality for the SBR model with pool-dependent rates**) *Suppose that any optimal solution for* (15) *has* $\bar{N}_j^\lambda > 0$ *for all* $j \in \mathcal{J}$. *Let* N^λ *be determined through the* $\bigwedge$*-based staffing in* (10) *with* β^* *as in* (9). *Set* π^λ *to FQR*(p^*, v^*) *with* p^* *as in* (13) *and* $v^* = (0, \cdots, 0, 1)$. *Then, the sequence* $\{(N^\lambda, \pi^\lambda)\}$ *is asymptotically optimal for* (3).

Remark 4.1 (**choosing the ratio vector** v^*) In light of our SSC result in Theorem 3.1 of [12], the choice $v^* = (0, 0, \ldots, 1)$ will cause all the idleness to be concentrated in pool J, which is the slowest agent-pool. This choice guarantees that all the faster servers will be constantly busy, thus maximizing the depletion rate of customers from the system. Informally, then, this choice of v^* minimizes the aggregate queue length in the system by maximizing the depletion rate. Because this observation holds for any staffing level, this choice of v^*, is essential for the minimization of the number of agents required to achieve the aggregate waiting-time constraints. Once the aggregate queue length is minimized, it only remains to distribute it in a proper way to ensure that the SL constraints are met. The queue-ratio vector, p^*, takes care of this task. ∎

Theorem 4.1 illustrates one of the key benefits of FQR. Although, the $\bigwedge$ model is a more efficient system, FQR allows the SBR system to work as efficiently, asymptotically, making the staffing of the $\bigwedge$ model sufficient also for the SBR system.

Asymptotic feasibility for (2): We now discuss asymptotically feasible solution for the SBR problem (2). While this formulation is somewhat problematic, as discussed in §2, it is very common in industry. Hence, it is of interest to discuss the construction of feasible solutions for this problem. We now define p^* to be

$$p_i^* := \frac{\lambda_i T_i^\lambda}{\sum_{k\in\mathcal{I}} \lambda_k T_k^\lambda} = \frac{a_i \bar{T}_i}{\sum_{k\in\mathcal{I}} a_k \bar{T}_k} \ , \tag{17}$$

and re-define β^* to be the unique solution of

$$\mathbf{P}_{\mu_1}(\beta)e^{-\beta\sqrt{\lambda}\sum_{i\in\mathcal{I}}\frac{\lambda_i}{\lambda}T_i^\lambda} = \alpha \ , \tag{18}$$

where $\mathbf{P}_{\mu_1}(\beta)$ defined in (9). As before, we observe that the vector p^* is independent of λ, since $\lambda_i/\lambda = a_i$ and Assumption 2.1 holds.

Theorem 4.2 (**asymptotic feasibility for the SBR model with pool-dependent rates**) *Suppose that any optimal solution for* (15) *has* $\bar{N}_j^\lambda > 0$ *for all* $j \in \mathcal{J}$. *Let* N^λ *be determined through the* $\bigwedge$*-based staffing in* (10) *with* β^* *as in* (18). *Set* π^λ *to FQR*(p^*, v^*) *with* p^* *as in* (17) *and* $v^* = (0, \cdots, 0, 1)$. *Then, the sequence* $\{(N^\lambda, \pi^\lambda)\}$ *is asymptotically feasible for* (2).

Both Theorems 4.1 and 4.2 rely on the fact that FQR admits an important SSC result by which, asymptotically, the queues of class i is equal to the proportion p_i^* of the aggregate queue length, and the number of idle servers in pool j is equal to a proportion v_j^* of the aggregate number of idle servers. Somewhat informally, the SSC results guarantees that with the $\bigwedge$ staffing and FQR we will have that, for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$

$$\frac{Q_i^\lambda(t)}{\sqrt{\lambda}} - v_j\frac{Q_\Sigma^\lambda(t)}{\sqrt{\lambda}} \Rightarrow 0, \text{ as } \lambda\to\infty, \text{ and } \frac{I_j^\lambda(t)}{\sqrt{\lambda}} - p_i\frac{I_\Sigma^\lambda(t)}{\sqrt{\lambda}} \Rightarrow 0, \text{ as } \lambda\to\infty,$$

where $Q_\Sigma^\lambda(t)$ and $I_\Sigma^\lambda(t)$ are, respectively, the aggregate queue length and aggregate number-of-idle-agents at time t. The SSC result for this setting is a corollary of our more general result in Theorem 3.1. of [12].

5. Other Formulations

This section is dedicated to alternative formulation of the call-center optimization problem.

5.1. Constraints on Average Delay

Here, we consider the formulation.

$$\begin{array}{ll} \text{minimize} & \sum_{j\in\mathcal{J}} c_j N_j \\ \text{subject to:} & \limsup_{T\to\infty} \bar{W}_i^{\lambda,T} \le T_i^\lambda, \ i\in\mathcal{I}, \\ & N \in \mathcal{A}^\lambda, \quad \pi \in \Pi, \end{array} \tag{19}$$

where

$$\bar{W}_i^{\lambda,T} := \frac{\sum_{k=1}^{A_i^\lambda(T)} w_{i,k}^\lambda}{A_i^\lambda(T)} \ .$$

Definition 5.1 (**asymptotic feasibility for (19)**) *A sequence of staffing vectors and routing policies* $\{(N^\lambda, \pi^\lambda)\}$ *is asymptotically feasible for* (19) *if (a)* $\pi^\lambda \in \Pi$*, for all* λ*, and (b) for every* $\epsilon > 0$*, there exists* $T^*(\epsilon)$ *such that, for all* $T \geq T^*(\epsilon)$*,*

$$\limsup_{\lambda\to\infty} P\left\{\frac{W_i^{\lambda,T}}{T_i^\lambda} \geq 1+\epsilon\right\} \leq \epsilon, \quad i \in \mathcal{I}. \tag{20}$$

The proposed solution in this case is as follows:

- **Staffing:** We use the optimal solution to (10) with β^* now given by the unique solution to

$$\frac{\mathbf{P}_{\mu_1}(\beta)}{\beta} = \sum_{i\in\mathcal{I}} a_i\sqrt{\lambda}T_i^\lambda = \sum_{i\in\mathcal{I}} a_i\bar{T}_i \ , \tag{21}$$

where again $\mathbf{P}_{\mu_1}(\beta)$ is defined in (9).

- **Routing:** FQR with ratio vectors $v^* = (0, 0, \ldots, 1)$ and p^* that is given by

$$p_i^* := \frac{\lambda_i T_i^\lambda}{\sum_{k\in\mathcal{I}} \lambda_k T_k^\lambda} = \frac{a_i\bar{T}_i}{\sum_{k\in\mathcal{I}} a_k\bar{T}_k} \ . \tag{22}$$

Theorem 5.1 (**asymptotic optimality for the average-waiting time formulation**) *Suppose that any optimal solution for* (15) *has* $\bar{N}_j^\lambda > 0$ *for all* $j \in \mathcal{J}$*. Let* N^λ *be determined through the* $\bigwedge$*-based staffing in* (10) *with* β^* *as in* (21)*. Set* π^λ *to* $FQR(p^*, v^*)$*, where* p^* *is as in* (22) *and* $v^* = (0, \cdots, 0, 1)$*. Then, the sequence* $\{(N^\lambda, \pi^\lambda)\}$ *is asymptotically optimal for* (19)*.*

5.2. Adding Customer Abandonment

In this section we augment our model by introducing customer abandonment. Specifically, we assume that a class-i customer has an exponential patience with rate θ_i. If his patience expires before he is admitted to service, the customer will abandon. Patience times of different customers are mutually independent. We will formulate an optimization problem for the abandonment model

and provide the corresponding asymptotic feasibility and optimality results. Our asymptotic optimality result for this augmented model is limited to the case in which all customers classes have the same abandonment rate, i.e, when $\theta_i \equiv \theta$. It is of practical interest, however, that we can provide asymptotically-feasible solutions for the case in which these rates are different.

To prove the asymptotic results for the abandonment model, we exploit Armony and Mandelbaum [2], which extends the results of [1] to the $\bigwedge$ model with customer abandonments. Using [2], we are able to provide proofs that parallel those for the non-abandonment case, repeating the analysis of (2), (3), and (19). Asymptotic-feasibility results can be obtained for general patience rates θ_i, while asymptotic-optimality can be obtained only for the homogeneous case.

Rather than repeating the analysis for formulations (2), (3), we introduce, and solve, a different formulation with constraint on the fraction of abandoning customers. To formulate this problem, let $L_i^\lambda(t)$ be the number of class-i customer that abandoned before being served up to time t. The fraction of customers that abandoned, among those that were initially in the queue (at time $t=0$) and those that arrived after time 0, is then given by

$$\mathrm{Ab}_i^{\lambda,T} := \frac{L_i^\lambda(T)}{Q_i^\lambda(0) + A_i^\lambda(T)} \ . \tag{23}$$

We then consider the formulation

$$\begin{array}{ll} \text{minimize} & \sum_{j\in\mathcal{J}} c_j N_j \\ \text{subject to:} & \limsup_{T\to\infty} \mathrm{Ab}_i^{\lambda,T} \le \alpha_i^\lambda, \ i\in\mathcal{I}, \\ & N \in \mathcal{A}^\lambda, \quad \pi\in\Pi, \end{array} \tag{24}$$

where we assume that $\alpha_i^\lambda = \bar{\alpha}_i/\sqrt{\lambda}$ for some strictly positive constants $\{\bar{\alpha}_i, \ i\in\mathcal{I}\}$, in order to place the system in the QED regime.

Definition 5.2 (**asymptotic feasibility for (24)**) *A sequence of staffing vectors and routing policies* $\{(N^\lambda,\pi^\lambda)\}$ *is asymptotically feasible for* (24) *if (a)* $\pi^\lambda\in\Pi$ *for all* λ *and (b) for every* $\epsilon>0$, *there exists* $T^*(\epsilon)$ *such that, for all* $T\ge T^*(\epsilon)$,

$$\limsup_{\lambda\to\infty} P\left\{\frac{Ab_i^{\lambda,T}}{\alpha_i^\lambda} \ge 1+\epsilon\right\} \le \epsilon, \qquad i\in\mathcal{I}. \tag{25}$$

We propose the following staffing and routing solution:

- **Staffing:** Use the optimal solution to (10) with β^* now given by the unique solution to

$$\bar{\alpha} = \sqrt{\bar{\theta}}\mathbf{P}_{\mu_1,\bar{\theta}}(\beta)\left[h\left(\frac{\beta}{\sqrt{\bar{\theta}}}\right) - \frac{\beta}{\sqrt{\bar{\theta}}}\right] \text{ and } \mathbf{P}_{\mu_1,\bar{\theta}}(\beta) := \left[1 + \frac{\sqrt{\bar{\theta}}h(\beta/\sqrt{\bar{\theta}})}{\sqrt{\mu_1}h(-\beta/\sqrt{\mu_1})}\right]^{-1}, \tag{26}$$

where $h(\cdot) := \phi(\cdot)/(1-\Phi(\cdot))$.

$$\bar{\theta} := \sum_{i\in\mathcal{I}} p_i\theta_i \text{ and } \bar{\alpha} = \sum_{i\in\mathcal{I}} a_i\alpha_i. \tag{27}$$

- **Routing:** FQR with ratio vectors $v^* = (0,0,\ldots,1)$ and p^* given by

$$p_i^* := \frac{\lambda_i\alpha_i^\lambda/\theta_i}{\sum_{k\in\mathcal{I}}\lambda_k\alpha_k^\lambda/\theta_k} = \frac{a_i\bar{\alpha}_i/\theta_i}{\sum_{k\in\mathcal{I}}a_k\bar{\alpha}_k/\theta_k}, \tag{28}$$

In (26), $\mathbf{P}_{\mu_1,\bar{\theta}}(\beta)$ is the asymptotic delay probability in the corresponding $\bigwedge$ model under the FSF policy, with the patience rate $\bar{\theta}$ and having $\sum_{j\in\mathcal{J}}\mu_j N_j = \lambda + \beta\sqrt{\lambda} + o(\sqrt{\lambda})$; see Propositions 4.5 and 4.6 in [2].

Theorem 5.2 (**asymptotic feasibility and optimality for the abandonment formulation**) *Suppose that any optimal solution for* (15) *has* $\bar{N}_j^\lambda > 0$ *for all* $j \in \mathcal{J}$. *Let* N^λ *be determined through the* $\bigwedge$*-based staffing in* (10) *with* β^* *in* (26). *Set* π^λ *to FQR*(p^*, v^*), *where* p^* *is as in* (28) *and* $v^* = (0,\cdots,0,1)$. *Then, the sequence* $\{(N^\lambda,\pi^\lambda)\}$ *is asymptotically feasible for* (24). *If, in addition,* $\theta_i \equiv \theta$ *for all* i, *then the sequence* $\{(N^\lambda,\pi^\lambda)\}$ *is also asymptotically optimal.*

We prove Theorem 5.2 in the e-companion. The required argument is similar to the previous case without abandonments. The key step is to show that, with FQR, the SBR model with abandonment is asymptotically equivalent (in terms of the aggregate number of customers in system) to a $\bigwedge$ model in which the customers' patience is exponential with a rate that is averaged using the ratio vector p of FQR in equation (27). For homogeneous patience rates, namely when $\theta_i \equiv \theta$, we show that a $\bigwedge$-based staffing (modified for the abandonment case) provides a lower bound on the staffing costs. Asymptotic optimality then follows from the asymptotic feasibility.

EXAMPLE 1. (**A two-class two-pool system**) We apply the proposed staffing-and-routing solution to a two-class two-pool system. We assume that pool-1 servers can serve only class-1 customers while pool-2 servers are cross-trained. The resulting N-model is depicted in Figure 2. The customer arrival and patience parameters are, respectively, $(\lambda_1, \lambda_2) = (100, 50)$ and $(\theta_1, \theta_2) = (2, 1)$. Since we are considering a fixed system, we omit the superscript λ from all notation. Since we are considering a setting with non-homogeneous patience rates, we only aim to show the feasibility of our solution.

To complete the model description, let the (pool-dependent) service rates be $(\mu_1, \mu_2) = (1.5, 1)$; let $c_1 = c_2 = c$; and let $\mathcal{A} = \{N \in \mathbb{Z}_+^2 : N_1 \leq 50\}$. In particular, the number of agents in the first pool can be at most 50. Finally, we assume that the abandonment constraints are 3% for class 1 and 5% for class 2, i.e, $(\alpha_1, \alpha_2) = (0.03, 0.05)$.

With these parameters, we construct our (asymptotically-feasible) $\bigwedge$-based staffing and FQR routing solution. The parameters for FQR are $p_1^* = 0.375$, $p_2^* = 0.625$, $\bar{\theta} = 1.375$ and $\bar{\alpha} = 0.3266$. From (26), for the $\bigwedge$-based staffing we have $\beta^* = 0.03926$, so that we need $\mu_1 N_1 + \mu_2 N_2 = 150 + \beta^* \sqrt{150} \geq 125.48$. Accordingly, we set $N_1 = 50$ and $N_2 = \lceil (125.4808 - 50)/\mu_2 \rceil = 76$.

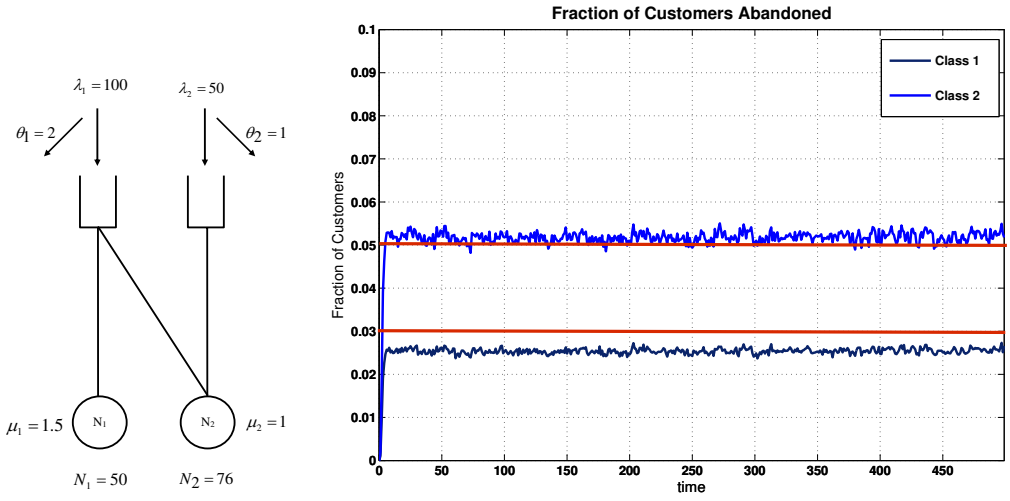

Figure 2 A two-class two-pool N model and the simulation results

We simulate this N model with the specified staffing and $FQR(p^*, v^*)$. We run 3000 replications of the system, each up to $T = 500$. The graph in Figure 2 displays the average proportion of

abandoning customers for each customer class and for each time unit (as a function of time). Evidently, the proportion of abandonments is below the target (for class 1) or only slightly above (for class 2). Also, we find that

$$P\left\{\frac{\mathrm{Ab}_i^T}{\alpha_i} \geq 1.02\right\} \leq 0.02\ , i=1,2, \tag{29}$$

where $P\{\cdot\}$ should be interpreted here as the empirical probability distribution over the 3000 replications, while (29) corresponds to the asymptotic-feasibility condition (25) in Definition 5.2 with ϵ there taken to be 0.02. ∎

5.3. Designated-Service Constraints

In practice, it is natural to require that most customers receive their *designated service*, i.e., service by the type of agent that the system designates for them. A good example is a multilingual call center, where one would prefer that Spanish-speaking customers be served by agents whose dominant language is Spanish, French-speaking customers be served by agents whose dominant language is French, and so forth. In other settings, the interpretation of designated service might be different. To treat designated-service constraints, we now assume that, for each customer class i, there is one agent type designated to provide service to that class; the agent type designated for class i is $j(\{i\})$; that pool of agents can only serve class i.

To impose a lower-bound constraint on the proportion of customers that receive their designated service, we let $D_i^{\lambda,T}$ be the proportion of the arriving class-i customers that are routed to their designated agents (the agents in pool $j(\{i\})$) by time T. Letting $\delta_i > 0$ be the non-designated-service proportion upper bound, the constraints are formulated as

$$\liminf_{T\to\infty} D_i^{\lambda,T} \geq 1-\delta_i, \quad i \in \mathcal{I}.$$

The optimization problem that we consider is then given by

$$\begin{array}{ll} \text{minimize} & \sum_{j\in\mathcal{J}} c_j N_j \\ \text{subject to:} & \limsup_{T\to\infty} \bar{W}^{\lambda,T} \le T_I^\lambda, \\ & \limsup_{T\to\infty} \bar{F}_i^{\lambda,T}(T_i^\lambda) \le \alpha, \quad 1\le i\le I-1, \\ & \liminf_{T\to\infty} D_i^{\lambda,T} \ge 1-\delta_i, \quad i\in\mathcal{I}, \\ & N\in\mathcal{A}^\lambda, \quad \pi\in\Pi. \end{array} \tag{30}$$

We now show that the designated service constraints can be incorporated into the framework of §4 by appropriately re-defining the set $\mathcal{A}^\lambda$. To this end, we extend the definition of asymptotic feasibility as follows:

Definition 5.3 (**asymptotic feasibility with designated-service constraints**) *A sequence of* $\{(N^\lambda,\pi^\lambda)\}$ *is asymptotically feasible for* (30) *if (a)* $\pi^\lambda\in\Pi$*, for all* λ*, and (b) for every* $\epsilon>0$*, there exists* $T^*(\epsilon)$ *such that, for all* $T\ge T^*(\epsilon)$*, equations* (4) *and* (5) *hold, as well as*

$$\limsup_{\lambda\to\infty} P\left\{D_i^{\lambda,T}\le 1-\delta_i-\epsilon\right\}\le\epsilon, \quad i\in\mathcal{I}. \tag{31}$$

The follow lemma provides a property that all asymptotically-feasible solutions must satisfy.

Lemma 5.1 *If* $\{(N^\lambda,\pi^\lambda)\}$ *is a sequence of asymptotically-feasible staffing and routing rules in the sense of Definition 5.3, then,*

$$\liminf_{\lambda\to\infty} \frac{\mu_{j(\{i\})}N^\lambda_{j(\{i\})}}{\lambda_i} \ge (1-\delta_i),\ i\in\mathcal{I}. \tag{32}$$

Lemma 5.1 suggests that one may incorporate the setting with designated-service constraints within the framework of §2 by replacing the optimization problem (30) with

$$\begin{array}{ll} \text{minimize} & \sum_{j\in\mathcal{J}} c_j N_j \\ \text{subject to:} & \limsup_{T\to\infty} \bar{W}^{\lambda,T} \le T_I^\lambda \\ & \limsup_{T\to\infty} \bar{F}_i^{\lambda,T}(T_i^\lambda) \le \alpha, \quad 1\le i\le I-1, \\ & N\in\mathcal{C}^\lambda, \quad \pi\in\Pi, \end{array} \tag{33}$$

where

$$\mathcal{C}^\lambda := \mathcal{A}^\lambda \bigcap \mathcal{B}^\lambda \text{ and } \mathcal{B}^\lambda := \left\{ N\in\mathbb{Z}_+^J : N_{j(\{i\})} \ge (1-\delta_i)\frac{\lambda_i}{\mu_{j(\{i\})}},\ i\in\mathcal{I}\right\}\ . \tag{34}$$

Note that the set $\mathcal{C}^\lambda$ fits in the framework of §2.1 as it is defined by linear constraints. In particular, (33) is a special case of (3) obtained by letting the set A^λ there be equal to $\mathcal{C}^\lambda$. The formal asymptotic feasibility result for this section appears in the next proposition.

Theorem 5.3 *Suppose that the conditions of Theorem 4.1 hold with respect to formulation* (33)*. Let N^λ be the asymptotically optimal $\bigwedge$-based staffing based on* (10) *with $\mathcal{A}^\lambda$ replaced by $\mathcal{C}^\lambda$ in* (34) *and set π^λ to FQR(p^*, v^*) with p^* as in* (13) *and $v^* = (0, \ldots, 0, 1)$. Then, $\{(N^\lambda, \pi^\lambda)\}$ is asymptotically feasible for* (30) *in the sense of Definition* 5.3*. It is asymptotically optimal for* (30) *if one of the following holds: (i) $A^\lambda \supseteq B^\lambda$, or (ii) $c_j \equiv c$ and $\mu_j \equiv \mu$.*

We emphasize that the asymptotically optimal solution to (33), while asymptotically feasible for (30), it need not be, in general, asymptotically optimal. The inequalities imposed by the set $\mathcal{B}^\lambda$ might be too restrictive. Actually, as the proof of Proposition 5.3 reveals, we could have $\mathcal{B}^\lambda$ defined through

$$\mathcal{B}^\lambda = \left\{ N \in \mathbb{Z}_+^J : N_{j(\{i\})} \geq (1-\delta_i)\frac{\lambda_i}{\mu_{j(\{i\})}} - K\sqrt{\lambda},\ i \in \mathcal{I} \right\},$$

for some $K > 0$. This would be less restrictive but would suffice to generate an asymptotically feasible solution for (30).

In the next section we consider a setting in which $c_j \equiv c$ and $\mu_j \equiv \mu$. There, as in the second part of Proposition 5.3, we will be able to replace (30) with (33) without compromising asymptotic optimality.

5.4. Common Service Rates

This section is devoted to a simple setting: a common service rate μ, no abandonments, a common cost c for all agents and $\mathcal{A}^\lambda = \mathbb{Z}_+$, as considered in Wallace and Whitt [19]. We have three purposes: first, to contrast the FQR-based solution with the simulation-based approach of [19], second, to illustrate an explicit construction of a system design when costs and system constraints do not pose significant restrictions and, third, to illustrate the diminishing-return property of flexibility. The optimization problem that we consider in this section is (30) but with $c_j \equiv c$, $j \in \mathcal{J}$, and $\mathcal{A}^\lambda \equiv \mathbb{Z}_+^J$.

Since we have a common service rate, we can staff using (non-asymptotic) formulas for the $M/M/N$ queue instead of the asymptotic expressions associated with the $\bigwedge$-based staffing in §4. (In this setting, these staffing methods are asymptotically equivalent.) To specify the staffing method here, let

$$N_\Sigma^\lambda := \min\left\{N \in \mathbb{Z}_+ : E[W_{\lambda,\mu}^{FCFS}] \le \lambda T_I^\lambda\right\}, \tag{35}$$

where $W_{\lambda,\mu}^{FCFS}$ is the steady-state waiting time in an $M/M/N$ queue with arrival rate λ and service rate μ. For routing, we also can use non-asymptotic expressions. Specifically, we can use FQR with ratio vector p given by

$$P\{W_{\lambda,\mu}^{FCFS} > 0\} e^{-\mu\left(\bar{N}_\Sigma - \frac{\lambda}{\mu}\right)\frac{\lambda_{I-1}T_{I-1}^\lambda}{\lambda p_{I-1}}} = \alpha \quad \text{and} \quad \frac{p_i}{p_{I-1}} = \frac{\lambda_i T_i^\lambda}{\lambda_{I-1}T_{I-1}^\lambda}, \tag{36}$$

Note that the ratio vector p here does depend on λ.

Rather than assuming that the design is given, we allow ourselves in this section to choose the design. In doing this, we will take into account the designated-service constraint. We will incorporate this constraint from §5.3. The specific design we suggest here is based on a concatenation of M systems, which we call the **generalized M (GM) model**. An example of a GM model with three customer classes is depicted in Figure 3. The GM model has a routing graph constructed by allowing only edges of form $(i, j(\{i\}))$, $i \in \mathcal{I}$, and $(i, j(\{i, i+1\}))$, $i = \mathcal{I}\backslash I$ ($\mathcal{I}$ excluding the element I). The GM model is relatively inexpensive in terms of cross-training, since it uses agents with at most two skills, and only a limited number with two skills.

The solution we propose is as follows:

- **Design: generalized M model (GM).** Use a GM model.
- **Staffing: Single-Class Staffing (SCS).** Determine the overall number of agents, $\bar{N}_\Sigma$ using (35). Then allocate agents to the pools by

  — $N_{j(\{i\})} = (1 - \delta/2)\left(\frac{\lambda_i}{\lambda}\right)\bar{N}_\Sigma$, $\quad i \in \{1, I\}$, $\quad N_{j(\{i\})} = (1-\delta)\left(\frac{\lambda_i}{\lambda}\right)\bar{N}_\Sigma$ for all $i = 2, \ldots, I-1$, and

  — $N_{j(\{i,i+1\})} = \left(\frac{\delta(\lambda_i + \lambda_{i+1})}{2\lambda}\right)\bar{N}_\Sigma \quad$ for all $\quad i = 1, \ldots, I-1$,

where $\delta := \min\{\delta_i : 1 \le i \le I\} > 0$.

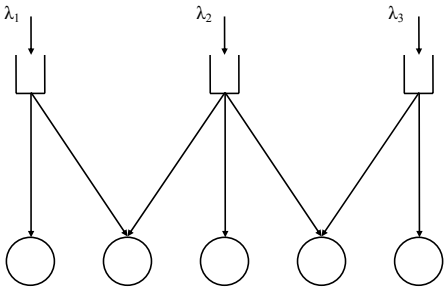

Figure 3 The generalized M model for 3 classes

- **Control: Fixed-Queue-Ratio (FQR).** Use FQR with p as defined in (36) and v defined by

$$v_{j(\{i,i+1\})} := \frac{1}{I-1} \quad \text{for all} \quad i \in \mathcal{I} \backslash I \ . \tag{37}$$

The common service rate allows us to use any vector v in the control step above. This stands in contrast to the case of different service rates, where we needed to use $v^* = (0, \ldots, 0, 1)$. The specific vector in (37) is designed to increase the amount of designated service by forcing the system to route customers that find agents idle in both pools $j(\{i\})$ and $j(\{i, i+1\})$ to the designated agents in pool $j(\{i\})$. One could also modify FQR so that all customers that find agents idle in more than one agent pool that can serve them will go to the designated agent pool $j(\{i\})$. This modification is guaranteed to achieve, asymptotically, the same performance as FQR. Using the results in §4, the above combined design-staffing-and-control solution can be shown to be asymptotically optimal as the arrival rate grows:

Theorem 5.4 (**asymptotic optimality for the SBR model with common service rates**) *Consider the simple SBR model specified above and assume that the GM design is used. Let N^λ be determined by SCS staffing and set π^λ to FQR(p^*, v^*) with p^* as in* (36) *and v^* as in* (37)*. Then, the sequence $\{(N^\lambda, \pi^\lambda)\}$ is asymptotically optimal for* (30) *with $c_j \equiv c$ and $\mathcal{A}^\lambda \equiv \mathbb{Z}_+^J$.*

6. Conclusions

In this paper we have proposed the *Fixed-Queue-Ratio* (FQR) routing scheme for the real-time routing of customers in call centers with multiple customer classes and multiple agent types oper-

ating under QoS constraints. FQR routing facilitates the construction of *combined staffing-design-and-routing solutions* for some settings of the complicated *skill-based-routing* (SBR) problem, with precisely specified goals. In this paper, we used FQR to to construct an asymptotically optimal solution for the staffing and routing problem subject to QoS constraints. The key assumption that we made is that the service rates are pool-dependent. However, as discussed in §2, this is not enough and we need to be careful about the formulation; to get asymptotic optimality, we need to replace the initial formulation (2) with the best-effort formulation in (3); Theorem 4.1 shows that FQR is asymptotically optimal in this setting. FQR also produces asymptotic optimality for other important formulations, specified in §5. Some modification is needed for each new formulation, but a version of FQR applies in each case. A key component of the proof was showing that our SBR problem is asymptotically equivalent to the $\bigwedge$ model previously analyzed by Armony [1].

It is especially instructive to see what can be done in the special case of a common service rate μ and a common agent cost c considered in §5.4, which was previously considered by Wallace and Whitt [19] using an iterative simulation-based staffing algorithm. In that case, we need neither the $\bigwedge$-based staffing in (10), nor the optimization problem (15). Consequently, for this special case we do not need Assumptions 4.1-4.3. This simple model illustrates how FQR simplifies tremendously the construction of the joint design-staffing-and-control solution, by allowing one to ignore the SL constraints when making the design and staffing decisions. The FQR routing will take care of those through a simple choice of the ratio vector. This essential decoupling of the design, staffing and control decisions is beneficial for applications, because in practice the design and staffing decisions are indeed often made in advance, and cannot easily be adjusted in real time. This stands in contrast to [19], where the numbers of agents in the service pools need to be fine-tuned through simulation to meet the SL constraints.

Finally, it is significant that the GM design used in §5.4 uses only limited flexibility. In particular, it uses agents that have at most two skills, and then only a limited number with two skills. Still, with this limited amount of flexibility, the SBR system performs, asymptotically, as efficiently as the

single-class single-pool $M/M/N$ queue. While this idea was communicated in [19], no mathematical results were established there.

Acknowledgments. This research is based on the first author's doctoral dissertation at Columbia University. The authors are grateful to Avi Mandelbaum and Mor Armony for fruitful discussions, and to Zohar Feldman for contributions to the simulation. The second author was supported by NSF Grant DMI-0457095.

References

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

[2] Armony, M., A. Mandelbaum. 2008. Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. Working Paper. New York University, New York, NY and Technion - The Israeli Institute of Technology, Haifa, Israel.

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

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

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

[6] A. Bassamboo, A. Zeevi. 2008. Staffing telephone call centers subject to service-level constraints: An approximate approach via constraint dualization. Working Paper. Northwestern University, Evanston, IL, and Columbia University, New York, NY.

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

[8] Feldman, Z., I. Gurvich and W. Whitt. 2007. Managing quality of service in call centers via Queue-Ratio Routing: Asymptotic analysis and simulation-based optimization. Working paper, Columbia University, New York, NY.

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

[10] Gurvich, I., M. Armony, A. Mandelbaum. 2008. Service-level differentiation in call centers with fully flexible servers. *Management Science* **54**(2) 279-294.

[11] Gurvich, I., J. Luedtke, T. Tezcan. 2008. Staffing call centers with uncertain demand forecasts: A chance-constraints approach. Working Paper. Northwestern University, Evanston, IL.

[12] Gurvich, I., W. Whitt. 2007. Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.*, forthcoming.

[13] Gurvich, I., W. Whitt. 2007. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management*, forthcoming.

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

[15] Mandelbaum, A., A. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the Generalized $c\mu$-Rule. *Oper. Res.* **52**(6) 836 - 855.

[16] Tezcan, T. 2006. Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Math. Oper. Res.*, forthcoming.

[17] Van Mieghem J.A. 1995. Dynamic scheduling with convex delay costs: the generalized $c\mu$ rule. *Ann. Appl. Probab.* **5**(3) 809-833.

[18] Van Mieghem J.A. 2003. Due date scheduling: asymptotic optimality of generalized longest queue and generalized largest delay rules. *Oper. Res.*, **51**(1) 113-122.

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

[20] Whitt, W. 2006. A Multi-class fluid model for a contact center with skill-based routing. *International Journal of Electronics and Communications (AEU)* **60**(2) 95-102.

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Asymptotic Optimality of Queue-Ratio Routing for Many-Server Service Systems

e-Companion

Itay Gurvich Ward Whitt

In this e-companion we provide the proofs of all major results in the main paper [7]. Customary numerical equation numbers refer to equations in [7], while alphanumerical equations refer to equations here.

EC.1. Relevance of the SSC Paper

We start by indicating why we can apply the state-space-collapse (SSC) and stochastic-process-limit results from our SSC paper [9]. We will need the following lemma, which relies strongly on Assumption 4.1.

Lemma EC.1.1 *Let N^λ be a sequence of staffing vectors determined by the $\bigwedge$-based staffing in* (10)*. Suppose that Assumptions 4.1-4.3 hold with respect to solutions of* (15)*. Then,*

$$\frac{N_j^\lambda}{\lambda} \to \nu_j \text{ as } \lambda \to \infty, \tag{EC.1}$$

where $\nu := (\nu_1, \ldots, \nu_J)$ is the first component of any optimal solution (ν, x) to (16)*.*

The proof of Lemma EC.1.1 can be found in [8]. By Assumption 4.1, all optimal solutions (ν, y) to (16) share the same ν. That ν is the limit in (EC.1). Before proving Lemma EC.1.1, we explain why it allows us to apply [9]. First, we will want to use the SSC result in Theorem 3.1 of [9] under condition C-1 there (pool-dependent service rates). However, in Assumption 2.1 of [9] it is assumed that, in addition to (EC.1), there exist constants $\gamma_j,\ j \in \mathcal{J}$ so that

$$\frac{(N_j^\lambda - \nu_j \lambda)}{\sqrt{\lambda}} \to \gamma_j \text{ as } \lambda \to \infty. \tag{EC.2}$$

However, a careful reading of the proof of Theorem 3.1 of [9] under condition C-1 reveals that (EC.2) is not required in the pool-dependent case. We only require that (EC.1) holds in addition to $\sum_{j\in\mathcal{J}}\mu_j N_j^\lambda = \lambda+\beta\sqrt{\lambda}+o(\sqrt{\lambda})$, which is implied by the definition of the $\bigwedge$-based staffing.

We will also want to use the diffusion limit result from Theorem 4.1 in [9]. The proof of that Theorem does require (EC.1), as it relies on the SSC result which itself requires this condition. Beyond that, however, the only restriction on the vector N^λ is that it satisfies $\sum_{j\in\mathcal{J}}\mu_j N_j^\lambda = \lambda+\beta\sqrt{\lambda}+o(\sqrt{\lambda})$ for some $\beta\in(-\infty,\infty)$.

Finally, we will be applying some results from Armony [1]. We emphasize that, there too, the only restrictions on the sequence $\{N^\lambda\}$ is that (EC.1) holds in addition to $\sum_{j\in\mathcal{J}}\mu_j N_j^\lambda = \lambda+\beta\sqrt{\lambda}+o(\sqrt{\lambda})$; see equations (A1) and (A2) there.

EC.2. Proof of Theorem 2.1

The proof consists of two steps. First, a coupling argument, Lemma EC.2.1 below, shows that the $\bigwedge$ model constitutes a lower bound for the SBR system in terms of the aggregate queue length. Lemma EC.2.2 builds on [1] to argue that any staffing vector for the $\bigwedge$ model that satisfies the constraint (4) must satisfy the capacity constraint (11). Together, these steps imply Theorem 2.1.

We start, then, with a stochastic comparison result. The lemma holds for each λ, so the superscript is omitted from the notation. For the stochastic ordering results, we let $SBR(\mathcal{I},\overrightarrow{\lambda},\mathcal{J},E,N,\mu)$ denote an SBR system with a set $\mathcal{I}$ of customer classes, arrival-rate vector $\overrightarrow{\lambda}=(\lambda_1,\ldots,\lambda_I)$, a set $\mathcal{J}$ of agent pools, a routing graph E, staffing vector N and pool-dependent service rates given by the vector $\mu=(\mu_1,\ldots,\mu_j)$. The corresponding $\bigwedge$ model is denoted by $\bigwedge(\lambda,\mathcal{J},N,\mu)$ and stands for a $\bigwedge$ model with arrival rate $\lambda=\sum_{i\in\mathcal{I}}\lambda_i$, a set $\mathcal{J}$ of agent pools with staffing vector N and service-rate vector μ. The set of admissible policies for the $\bigwedge$ model is the set of non-anticipating policies. An example of an SBR system and its corresponding $\bigwedge$ model was given in Figure 1 of [7].

Given admissible controls π_1 and π_2 for the SBR and $\bigwedge$ model, respectively, we let $Z_{j,SBR}^{\pi_1}(t)$ and $Z_{j,\bigwedge}^{\pi_2}(t)$ be the corresponding number of busy agents in agent pool j in each of the systems under

their respective controls. Similarly, we let $Q^{\pi_1}_{\Sigma,SBR}(t)$ and $Q^{\pi_2}_{\Sigma,\bigwedge}(t)$ be the corresponding aggregate queue length processes. Here, the subscript Σ is used also for the $\bigwedge$ model only for purposes of notational consistency and the reader is reminded that the $\bigwedge$ model has only a single queue.

Lemma EC.2.1 (**the $\bigwedge$ model as a lower bound**) *Fix the data* $(\mathcal{I},\mathcal{J},E,N,\overrightarrow{\lambda},\mu)$. *Assume that* $(Z_{j,SBR}(0),Q_{\Sigma,SBR}(0);j\in\mathcal{J})=(Z_{j,\bigwedge}(0),Q_{\Sigma,\bigwedge};j\in\mathcal{J})$. *Then, given any admissible policy* π_1 *for the SBR system, there exists an admissible policy* π_2 *for the* $\bigwedge$ *model and a construction of the sample paths such that almost surely*

$$\{Q^{\pi_1}_{\Sigma,SBR}(t),\ t\geq 0\}=\{Q^{\pi_2}_{\Sigma,\bigwedge}(t),\ t\geq 0\}.$$

Consequently,

$$\{Q^{\pi_1}_{\Sigma,SBR}(t),\ t\geq 0\}\stackrel{d}{=}\{Q^{\pi_2}_{\Sigma,\bigwedge}(t),\ t\geq 0\}.$$

The proof of Lemma EC.2.1 follows a very simple coupling argument based on the observation that any customer assignment that can be made in the SBR system also can be made in the corresponding $\bigwedge$ system. The complete formal argument is omitted. We proceed to prove a result about the $\bigwedge$ model. For the following, let $\bar{W}^{\lambda,T}_{\bigwedge}$ be the aggregate averaged waiting time, as defined in (1), but add the subscript $\bigwedge$ to denote that it corresponds to the $\bigwedge$ model rather than with the general SBR system.

Lemma EC.2.2 (**asymptotic feasibility for the $\bigwedge$ model**) *Fix a sequence of* $\bigwedge$ *models with capacity vectors that satisfy*

$$\liminf_{\lambda\to\infty}\frac{N^\lambda_j}{\lambda}>0,\ j\in\mathcal{J}. \tag{EC.3}$$

Then,

$$\limsup_{\lambda\to\infty}P\left\{\frac{\bar{W}^{\lambda,T}_{\bigwedge}}{T^\lambda_I}\geq 1+\epsilon\right\}\leq\epsilon\ , \tag{EC.4}$$

for all $\epsilon>0$ **only if** *the sequence* N^λ *satisfies*

$$\sum_{j\in\mathcal{J}}\mu_j N^\lambda_j\geq\lambda+\beta^*\sqrt{\lambda}+o(\sqrt{\lambda}), \tag{EC.5}$$

where β^* *is the unique solution to* (9).

Proof: First, we have

$$\bar{W}_{\bigwedge}^{\lambda,T} \geq \frac{1}{A^{\lambda}(T)} \int_0^T Q_{\Sigma,\bigwedge}^{\lambda}(t)dt. \tag{EC.6}$$

This inequality follows from a simple inequality for queueing systems which is used often to prove Little's law (see for example equation (2.1) in Whitt [14]). By the renewal strong law

$$\lim_{\lambda\to\infty} \frac{\sqrt{\lambda}}{A^{\lambda}(T)} \int_0^T Q_{\Sigma,\bigwedge}^{\lambda}(t)dt = \lim_{\lambda\to\infty} \frac{1}{\sqrt{\lambda}} \int_0^T Q_{\Sigma,\bigwedge}^{\lambda}(t)dt,$$

and the equality holds almost surely. Consequently,

$$\limsup_{\lambda\to\infty} P\left\{ \frac{\bar{W}_{\bigwedge}^{\lambda,T}}{T_I^{\lambda}} \geq 1+\epsilon \right\} \leq \epsilon,$$

only if

$$\limsup_{\lambda\to\infty} P\left\{ \frac{1}{T} \int_0^T \frac{Q_{\Sigma,\bigwedge}^{\lambda}(t)}{\lambda T_I^{\lambda}} dt \geq 1+\epsilon \right\} \leq \epsilon. \tag{EC.7}$$

We now apply the results of [1]. Specifically, to reach a contradiction, assume that there exists $\tilde{\beta} < \beta^*$ such that $\sum_{j\in\mathcal{J}} \mu_j N_j^{\lambda} < \lambda + \tilde{\beta}\sqrt{\lambda} + o(\sqrt{\lambda})$. Fix a subsequence λ^k such that

$$\lim_{k\to\infty} \frac{\sum_{j\in\mathcal{J}} \mu_j N_j^{\lambda} - \lambda}{\sqrt{\lambda}} = 0 < \hat{\beta} \leq \tilde{\beta}.$$

Observe that we have restricted the attention to $\hat{\beta} > 0$. We will later deal with the cases $\hat{\beta} \leq 0$. We now focus on this subsequence. As the following arguments will apply to any convergent subsequence we assume, without loss of generality, that this holds for the whole sequence. Armony [1] introduces the Fast-Server-First (FSF) for the $\bigwedge$ model and shows that it is asymptotically optimal in that it minimizes the steady-state scaled queue length. Henceforth, we assume that FSF is used for control for the $\bigwedge$ model. The argument is extended to arbitrary controls in the end of the proof. Equation (EC.3) allows us to apply Proposition 4.2 in [1], and the continuity of the integral map to have that

$$\lim_{\lambda\to\infty} \frac{1}{T} \int_0^T \frac{Q_{\Sigma,\bigwedge}^{\lambda}(t)}{\lambda T_I^{\lambda}} dt = \frac{1}{T} \int_0^T \frac{\left[\hat{X}_{\Sigma,\bigwedge}(t)\right]^+}{\bar{T}_I} dt, \tag{EC.8}$$

Where $\hat{X}_{\Sigma,\bigwedge}(t)$ is the limit in Proposition 4.2 of [1]. Here, we also used Assumption 2.1 to write $T_I^{\lambda} = \bar{T}_I/\sqrt{\lambda}$.

By Proposition 4.4 in [1], $\hat{X}_{\Sigma,\bigwedge}(t)$ has a unique stationary distribution. The same conditions that guarantee this existence also guarantee that $\hat{X}_{\Sigma,\bigwedge}(t)$ is positive recurrent; see exercise 5.5.40 in page 352 of [12]. The general theory of regenerative processes then implies that

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T g\left(\hat{X}_{\Sigma,\bigwedge}(t)\right)dt = E\left[g\left(\hat{X}_{\Sigma,\bigwedge}(\infty)\right)\right] \quad \text{w.p.1}$$

for any function $g(\cdot)$ that is integrable under the unique invariant measure of $\hat{X}_{\Sigma,\bigwedge}(t)$; see §VI.3 of Asmussen [3] for the general regenerative-process result and Theorem 3.1 of Khas'miniskii [13] for the diffusion-process result. Through simple integration, Proposition 4.4 in [1] also shows that $E\left[\left[\hat{X}_{\Sigma,\bigwedge}(\infty)\right]^+\right] = \mathbf{P}_{\mu_1}(\hat{\beta})/\hat{\beta}\bar{T}_I$. Using Lemma B.1 in Borst et. al. [5], it follows that this expectation is decreasing as a function of its argument, β, and we have that for some $\epsilon > 0$,

$$\frac{1}{T}\int_0^T \frac{\left[\hat{X}_{\Sigma,\bigwedge}(t)\right]^+}{\bar{T}_I}dt \to \frac{E\left[\left[\hat{X}_{\Sigma,\bigwedge}(\infty)\right]^+\right]}{\bar{T}_I} = \mathbf{P}_{\mu_1}(\hat{\beta})\frac{1}{\hat{\beta}\bar{T}_I} > 1+\epsilon \quad \text{as } T\to\infty \quad \text{w.p.1}\ , \qquad \text{(EC.9)}$$

where the last inequality easily follows from the fact that $\hat{\beta} < \beta^*$. Combining equations (EC.7)-(EC.9) we have reached a contradiction and in particular that if $\sum_{j\in\mathcal{J}}\mu_j N_j^\lambda = \lambda + \hat{\beta}^*\sqrt{\lambda} + o(\sqrt{\lambda})$ for $0 < \hat{\beta} < \beta^*$, then (EC.4) can not hold. It still remains to argue what happens when $\hat{\beta} \le 0$. This, however, follows from our previous argument. Indeed, considering the diffusion process $\hat{X}_{\Sigma,\bigwedge}(t)$ of Proposition 4.2 in [1], it is a matter of a simple comparison result for diffusions (see e.g. Proposition 5.2.18 in [12]) to show that if, for $\hat{\beta}$,

$$\frac{1}{T}\int_0^T \frac{\left[\hat{X}_{\Sigma,\bigwedge}(t)\right]^+}{\bar{T}_I} > 1+\epsilon,$$

for some $\epsilon > 0$ and all T large enough, then the same holds for all $\beta \le \hat{\beta}$ and in particular for $\beta \le 0$. We proved, then, that if FSF is used any staffing level that satisfies (EC.3) and (EC.4) must also satisfy (9). To rule out the possibility that there is a different control under which (9) is not necessary, we point out that by Theorem 4.1 of [10] we have that for each $T > 0$,

$$\liminf_{\lambda\to\infty}\frac{1}{T}\int_0^T \frac{Q_{\Sigma,\bigwedge}^{\pi,\lambda}(t)}{\lambda T_I^\lambda}dt \ge_{st} \frac{1}{T}\int_0^T \frac{\left[\hat{X}_{\Sigma,\bigwedge}(t)\right]^+}{\bar{T}_I}dt,$$

where π is some control other than FSF and $Q_{\Sigma,\bigwedge}^{\pi,\lambda}(t)$ is the queue-length process under this control. Consequently, if (EC.7) is violated under FSF, it is also violated under any other control. The proof is now complete. ■

Proof of Theorem 2.1: Given Lemmas EC.2.1 and EC.2.2 the proof the theorem is straightforward. Specifically, we can construct the sample paths so that

$$\bar{W}^{\lambda,T} \geq \frac{1}{A^\lambda(T)} \int_0^T Q^\lambda_{\Sigma,SBR}(t)dt \geq \frac{1}{A^\lambda(T)} \int_0^T Q^\lambda_{\Sigma,\bigwedge}(t)dt,$$

where the first inequality is as in (EC.6) and the second inequality follows from Lemma EC.2.1. One now repeats the proof of Lemma EC.2.2 for the right hand side to obtain the result of the theorem. ■

EC.3. Proofs for §4

In Theorem 2.1, we established that the $\bigwedge$-based staffing-vector constitutes a lower bound for the staffing of the SBR system. To prove asymptotic optimality it remains to show that this staffing vector is not only a lower bound, but is also asymptotically feasible for the SBR system. Towards that end, we need to show that, together with FQR and the appropriate ratio vectors, the $\bigwedge$-based staffing satisfies (4) and (5) in the case of Theorem 4.1 and (6) in the case of Theorem 4.2. These results will be a consequence of two lemmas. The first, Lemma (EC.3.1), relates the performance measures for the SBR system to performance measures for the $\bigwedge$ model. As before, $Q^\lambda_{\Sigma,\bigwedge}(t)$ is the queue-length process in the $\bigwedge$ model operated under the Faster-Server-First (FSF) policy defined in [1].

Lemma EC.3.1 (**asymptotic equivalence with the $\bigwedge$ model**) *Under the conditions of Theorem 4.1 (respectively Theorem 4.2), for every $T \geq 0$ and $y \geq 0$,*

$$\lim_{\lambda\to\infty} \frac{1}{T} \int_0^T \mathbf{1}\left\{Q^\lambda_{\Sigma,SBR}(t) > y\sqrt{\lambda}\right\} dt \stackrel{d}{=} \lim_{\lambda\to\infty} \frac{1}{T} \int_0^T \mathbf{1}\left\{Q^\lambda_{\Sigma,\bigwedge}(t) > y\sqrt{\lambda}\right\} dt, \tag{EC.10}$$

and, in particular,

$$\lim_{\lambda\to\infty} \bar{F}_i^{\lambda,T}(y/\sqrt{\lambda}) \stackrel{d}{=} \lim_{\lambda\to\infty} \frac{1}{T} \int_0^T \mathbf{1}\left\{\frac{p_i}{a_i} Q^\lambda_{\Sigma,\bigwedge}(t) > y\sqrt{\lambda}\right\} dt, \ i \in \mathcal{I}. \tag{EC.11}$$

Also,

$$\lim_{\lambda\to\infty} \sqrt{\lambda}\bar{W}^{\lambda,T} \stackrel{d}{=} \lim_{\lambda\to\infty} \sqrt{\lambda}\bar{W}^{\lambda,T}_{\bigwedge}, \quad \textit{so that} \quad \lim_{\lambda\to\infty} \frac{\bar{W}^{\lambda,T}}{T^\lambda_I} \stackrel{d}{=} \lim_{\lambda\to\infty} \frac{\bar{W}^{\lambda,T}_{\bigwedge}}{T^\lambda_I}. \tag{EC.12}$$

Proof: By Theorem 4.1 in [9] we have that $\hat{Q}^{\lambda}_{\Sigma}(t) \Rightarrow [\hat{X}_{\Sigma}(t)]^{+}$ as $\lambda \to \infty$. As in the beginning of the proof of Proposition EC.2.1 in [6], we apply the mapping $f_a(x) = \frac{1}{T}\int_0^T \mathbf{1}\{x(t) > a\}$ from $D[0,\infty)$ to $\mathbb{R}$, to get

$$\lim_{\lambda\to\infty} \frac{1}{T}\int_0^T \mathbf{1}\left\{Q^{\lambda}_{\Sigma,SBR}(t) > y\sqrt{\lambda}\right\} \Rightarrow \frac{1}{T}\int_0^T \mathbf{1}\left\{\left[\hat{X}_{\Sigma}(t)\right]^{+} dt > y\right\} \quad \text{in} \quad [0,1] \quad \text{as} \quad \lambda \to \infty,$$

where $\hat{X}_{\Sigma}(t)$ is the diffusion process from Theorem 4.1 in [9]. Some care is required in applying this integral mapping as in the proof of Proposition EC.2.1 in [6].

With the choice of $v^* = (0, 0 \ldots, 1)$ it is evident that this diffusion process has the same law as the one corresponding to the limit of the $\bigwedge$ model as given Proposition 4.2 in [1]. Consequently, we also have that

$$\lim_{\lambda\to\infty} \frac{1}{T}\int_0^T \mathbf{1}\left\{Q^{\lambda}_{\Sigma,\bigwedge}(t) > y\sqrt{\lambda}\right\} \Rightarrow \frac{1}{T}\int_0^T \mathbf{1}\left\{\left[\hat{X}_{\Sigma}(t)\right]^{+} dt > y\right\} \quad \text{in} \quad [0,1] \quad \text{as} \quad \lambda \to \infty.$$

The other parts of the lemma follow from Proposition EC.2.1 in [6]. ■

The next Lemma shows that the performance measures for the $\bigwedge$ model that appear in Lemma EC.3.1 indeed satisfy the required bounds.

Lemma EC.3.2 *Consider a sequence of $\bigwedge$ models operated under FSF and satisfying* (EC.3) *and*

$$\sum_{j\in\mathcal{J}} \mu_j N^{\lambda}_j = \lambda + \beta^*\sqrt{\lambda} + o(\sqrt{\lambda}),$$

with β^* as given (9) *(respectively* (18)*). Assume that*

$$\hat{X}^{\lambda}_{\Sigma,\bigwedge}(0) \Rightarrow \hat{X}_{\Sigma,\bigwedge}(0) \quad \textit{and} \quad \hat{Q}^{\lambda}_{\Sigma,\bigwedge}(0) \Rightarrow \hat{Q}_{\Sigma,\bigwedge}(0) \quad \textit{as } \lambda \to \infty.$$

We then have the following:

1. *Suppose that β^* is determined through* (9)*. Then, for each $\epsilon > 0$, there exists $T^*(\epsilon)$ so that for all $T \geq T^*(\epsilon)$,*

$$\lim_{\lambda\to\infty} P\left\{\frac{\bar{W}^{\lambda,T}_{\bigwedge}}{T^{\lambda}_I} \geq 1 + \epsilon\right\} \leq \epsilon, \tag{EC.13}$$

$$P\left\{\frac{1}{T}\int_0^T \mathbf{1}\left\{p_{I-1}Q^{\lambda}_{\Sigma,\bigwedge}(t) > \lambda_{I-1}T^{\lambda}_{I-1}\right\} dt \geq \alpha + \epsilon\right\} \leq \epsilon, \tag{EC.14}$$

and

$$P\left\{\frac{1}{T}\int_0^T \mathbf{1}\left\{p_i Q^\lambda_{\Sigma,\wedge}(t) > \lambda_i T_i^\lambda\right\} dt \geq \alpha+\epsilon\right\} \leq \epsilon,\ 1 \leq i \leq I-2. \tag{EC.15}$$

2. *Suppose that β^* is determined through* (18)*. Then, for each $\epsilon > 0$ there exists $T^*(\epsilon)$ so that, for all $T \geq T^*(\epsilon)$,*

$$\lim_{\lambda\to\infty} P\left\{\frac{1}{T}\int_0^T \mathbf{1}\left\{Q^\lambda_{\Sigma,\wedge}(t) > \sum_{i\in\mathcal{I}} \lambda_i T_i^\lambda\right\} dt \geq \alpha+\epsilon\right\} \leq \epsilon. \tag{EC.16}$$

Proof: Equation (EC.13) is proved within the proof of Lemma EC.2.2. In particular, it is proved there that $\limsup_{\lambda\to\infty} \bar{W}^{\lambda,T}_\wedge / T^\lambda_I \Rightarrow 1$ as $T \to \infty$. Equation (EC.14) follows by applying the argument in the proof of Lemma EC.3.1 to show that for each $y > 0$,

$$\frac{1}{T}\int_0^T \mathbf{1}\left\{Q^\lambda_{\Sigma,\wedge}(t) > y\sqrt{\lambda}\right\} \Rightarrow \frac{1}{T}\int_0^T \mathbf{1}\left\{[\hat{X}_\Sigma(t)]^+ > y\right\},\ \text{as } \lambda\to\infty,$$

where $\hat{X}_\Sigma(t)$ is the diffusion limit from Proposition 4.2 in [1] with δ there set equal to β^*. By that same proposition, and applying the theory of regenerative processes as in the proof of Lemma EC.2.2, we have that, as $t \to \infty$,

$$\frac{1}{T}\int_0^T \mathbf{1}\left\{[\hat{X}_\Sigma(t)]^+ > \frac{a_{I-1}}{p_{I-1}}\bar{T}_{I-1}\right\} \Rightarrow \mathbf{P}_{\mu_1}(\beta) e^{-\beta^* \frac{a_{I-1}}{p_{I-1}}\bar{T}_{I-1}} = \alpha,$$

where the convergence is almost surely. The equality on the right-hand side follows from the definition of p_{I-1} in (13). Equation (EC.15) now follows from the definition of p_i for $i \neq I-1$. Finally, equation (EC.16) is proved in a similar manner using the definition of p in (17). ■

Proof of Theorem 4.1: Equation (4) follows from the choice of p in (13) and combining (EC.11), (EC.14) and (EC.15). Equation (5) is obtained by combining (EC.12) and (EC.13). Finally, equation (7) follows from the fact that $c \cdot N^\lambda$ is an asymptotic lower bound by Theorem 2.1. ■

Proof of Theorem 4.2: Equation (6) follows readily by combining (EC.11) and (EC.16) and using the definition of the ratio vector p in (17). ■

EC.4. Proofs for §5.2

We start by providing an outline of the proof of Theorem 5.2. Notably, the proof follows closely the steps in the proofs of asymptotic feasibility and optimality for the non-abandonment model. First, in Lemma EC.4.2, we state and prove an asymptotic feasibility result for the $\bigwedge$ model with abandonment. This result builds on the steady-state feasibility result in [2]. We then show that, with FQR and the $\bigwedge$-based staffing–both with the appropriate parameters–the SBR system is asymptotically equivalent, in terms of the aggregate-queue-length process, to the $\bigwedge$ model. The asymptotic feasibility then follows because FQR distributed the queues properly. For the case of homogeneous patience rates, $\theta_i \equiv \theta$, we use a couple argument to show that the $\bigwedge$-based staffing is an asymptotic lower bound. As this bound is achieved, our proposed solution is asymptotically optimal.

We start, as in the non-abandonment case, with a stochastic comparison result. The lemma holds for each λ, so the superscript is omitted from the notation. We augment our notation from the non-abandonment case to let $SBR(\mathcal{I}, \overrightarrow{\lambda}, \mathcal{J}, E, N, \mu, \bar{\theta})$ be as $SBR(\mathcal{I}, \overrightarrow{\lambda}, \mathcal{J}, E, N, \mu)$ but with the addition of a homogeneous patience rate $\bar{\theta}$ for all customer classes. The corresponding $\bigwedge$ model is denoted by $\bigwedge(\lambda, \mathcal{J}, N, \mu, \bar{\theta})$ and is the same as $\bigwedge(\lambda, \mathcal{J}, N, \mu)$ with the exception of the single customer class now having finite exponential patience with rate $\bar{\theta}$. The sets of admissible policies as well as the notation introduced prior to Lemma EC.2.1 remain the same except for the following addition: For $t \geq 0$, We let $Ab^{\pi_1,t}_{\Sigma,SBR}$ (respectively $Ab^{\pi_2,t}_{\Sigma,\bigwedge}$) be the global fraction of abandoning customers by time t under policy π_1 (respectively π_2) in the λ^{th} SBR system (respectively $\bigwedge$ model). Also, we let $\alpha^\lambda = \bar{\alpha}/\sqrt{\lambda}$, with $\bar{\alpha}$ ad defined in equation (27).

Lemma EC.4.1 (**the $\bigwedge$ model as a lower bound**) *Fix the data $(\mathcal{I}, \mathcal{J}, E, N, \overrightarrow{\lambda}, \mu, \theta)$. Assume that $(Z_{j,SBR}(0), Q_{\Sigma,SBR}(0); j \in \mathcal{J}) = (Z_{j,\bigwedge}(0), Q_{\Sigma,\bigwedge}; j \in \mathcal{J})$. Then, given any admissible policy π_1 for the SBR system, there exists an admissible policy π_2 for the $\bigwedge$ and a construction of the sample paths such that almost surely $\{Ab^{\pi_1,t}_{\Sigma,SBR}, t \geq 0\} = \{Ab^{\pi_2,t}_{\Sigma,\bigwedge},\ t \geq 0\}$. Consequently,*

$$\{Ab^{\pi_1,t}_{\Sigma,SBR},\ t \geq 0\} \stackrel{d}{=} \{Ab^{\pi_2,t}_{\Sigma,\bigwedge},\ t \geq 0\}.$$

Lemma EC.4.2 (**a feasibility result for the $\bigwedge$ model with abandonments**) *Fix a sequence of $\bigwedge$ models (with abandonments) with capacity vectors that satisfy*

$$\liminf_{\lambda\to\infty} \frac{N_j^\lambda}{\lambda} > 0,\ j\in\mathcal{J}. \tag{EC.17}$$

Then,

$$\limsup_{\lambda\to\infty} P\left\{\frac{Ab_{\bigwedge}^{\lambda,T}}{\alpha^\lambda}\geq 1+\epsilon\right\}\leq\epsilon\ , \tag{EC.18}$$

for all $\epsilon>0$ **only if** *the sequence N^λ satisfies*

$$\sum_{j\in\mathcal{J}}\mu_j N_j^\lambda \geq \lambda+\beta^*\sqrt{\lambda}+o(\sqrt{\lambda}), \tag{EC.19}$$

where β^* is the unique solution to (26)*.*

Proof: Let $L^\lambda(t)$ be the number of abandonments by time t in the λ^{th} $\bigwedge$ model. First, we use a construction of the sample path through unit-rate Poisson process (see §6.1 of [9]). In particular, we write $L^\lambda(t)=R\left(\bar\theta\int_0^t Q^\lambda_{\Sigma,\bigwedge}(s)ds\right)$, where $\{R(t),\ t\geq 0\}$ is a unit-rate Poisson process. We can now apply the analysis in [9] by assuming that the SBR model in consideration is a $\bigwedge$ model. First, we write $R\left(\bar\theta\int_0^t Q^\lambda_{\Sigma,\bigwedge}(s)ds\right)=\bar\theta\int_0^t Q^\lambda_{\Sigma,\bigwedge}(s)ds+M^\lambda_{R_i}(t)$, where $M^\lambda_{R_i}(t)$ is the martingale defined in equation (19) of [9]. By Lemma 7.1 of [9], we have that $M^\lambda_{R_i}(t)/\sqrt{\lambda}\Rightarrow 0$ as $\lambda\to\infty$ in D. By Theorem 4.1 of [9] we have that $Q^\lambda_{\Sigma,\bigwedge}(t)/\sqrt{\lambda}\Rightarrow[\hat X_{\Sigma,\bigwedge}(t)]^+$ where $\hat X_{\Sigma,\bigwedge}(t)$ is the diffusion process defined in Theorem 4.1 of [9]. Also, by assumption we have that $Q^\lambda(0)/\lambda\xrightarrow{P}0$. Consequently, we have that

$$\lim_{\lambda\to\infty} Ab^{\lambda,T}=\lim_{\lambda\to\infty}\sqrt{\lambda}\frac{R^\lambda(T)}{Q^\lambda(0)+A^\lambda(T)}=\lim_{\lambda\to\infty}\frac{\lambda+Q^\lambda(0)}{A^\lambda(T)}\frac{R^\lambda(T)}{\sqrt{\lambda}}\overset{d}{=}\lim_{\lambda\to\infty}\frac{1}{T}\bar\theta\int_0^T\frac{Q^\lambda_{\Sigma,\bigwedge}(s)}{\sqrt{\lambda}}ds, \tag{EC.20}$$

and, from the convergence of $Q^\lambda_{\Sigma,\bigwedge}(t)/\sqrt{\lambda}$, we have that

$$\sqrt{\lambda}\frac{R^\lambda(T)}{A^\lambda(T)}\Rightarrow\frac{1}{T}\bar\theta\int_0^T[\hat X_\Sigma(s)]^+ds,\ \text{as}\ \lambda\to\infty. \tag{EC.21}$$

From here the proof follows closely that of Lemma EC.2.2 with simple modifications. For completeness, we provide the detailed proof. To reach a contradiction to the statement of the lemma, assume that there exists $\tilde\beta<\beta^*$ such that $\sum_{j\in\mathcal{J}}\mu_jN_j^\lambda<\lambda+\tilde\beta\sqrt{\lambda}+o(\sqrt{\lambda})$. Fix a subsequence λ^k such that $\lim_{k\to\infty}(\sum_{j\in\mathcal{J}}\mu_jN_j^\lambda-\lambda)/(\sqrt{\lambda})=\hat\beta\leq\tilde\beta$. Note that in the abandonment case we do not

need to consider $\hat{\beta} < 0$ separately as the diffusion limits will have a steady-state even for $\hat{\beta} < 0$. We now focus on the convergent subsequence. As the following arguments will apply to any convergent subsequence, we assume, without loss of generality, that this holds for the whole sequence. Armony and Mandelbaum [2] extend the results of [1] to show that the Fast-Server-First (FSF) minimizes the steady-state fraction of abandonments in the $\bigwedge$ model. Henceforth, we assume that FSF is used for control for the $\bigwedge$ model. The argument is extended to arbitrary controls in the end of the proof.

By Proposition 4.5 in [2], $\hat{X}_{\Sigma,\bigwedge}(t)$ has a unique stationary distribution. By Corollary 4.3 in [2] and, in particular, the right-hand side of equation 4.36 there, we have that

$$\bar{\theta} E\left[[\hat{X}_{\Sigma,\bigwedge}(\infty)]^+\right] = \sqrt{\bar{\theta}} \mathbf{P}_{\mu_1,\bar{\theta}}(\beta)\left[h\left(\frac{\beta}{\sqrt{\bar{\theta}}}\right) - \frac{\beta}{\sqrt{\bar{\theta}}}\right], \tag{EC.22}$$

A simple differentiation shows that the right-hand side of (EC.22) is strictly decreasing in β, and we have that for some $\epsilon > 0$, as $T \to \infty$,

$$\frac{1}{T}\int_0^T \bar{\theta}\Big[\hat{X}_{\Sigma,\bigwedge}(t)\Big]^+ dt \to \bar{\theta} E\left[\Big[\hat{X}_{\Sigma,\bigwedge}(\infty)\Big]^+\right] = \sqrt{\bar{\theta}} \mathbf{P}_{\mu_1,\bar{\theta}}(\hat{\beta})\left[h\left(\frac{\hat{\beta}}{\sqrt{\bar{\theta}}}\right) - \frac{\hat{\beta}}{\sqrt{\bar{\theta}}}\right] > (1+\epsilon)\bar{\alpha} \quad \text{w.p.1 ,.} \tag{EC.23}$$

Here, the convergence follows from the same regenerative argument as in the proof of Lemma EC.2.2 and last inequality easily follows from the fact that $\hat{\beta} < \beta^*$. Combining equations (EC.22)-(EC.23) we have reached a contradiction and in particular that if $\sum_{j\in\mathcal{J}} \mu_j N_j^\lambda = \lambda + \hat{\beta}^* \sqrt{\lambda} + o(\sqrt{\lambda})$ for $0 < \hat{\beta} < \beta^*$, then (EC.4) can not hold. We proved, then, that if FSF is used any staffing level that satisfies (EC.17) and (EC.18) must also satisfy (EC.19). To rule out the possibility that there is a different control under which (EC.19) is not necessary, we point out that, by Theorem 4.1 of [10], we have that for each $T > 0$,

$$\liminf_{\lambda\to\infty} \frac{1}{T}\bar{\theta}\int_0^T \frac{Q_{\Sigma,\bigwedge}^{\pi,\lambda}(t)}{\lambda T_I^\lambda} dt \geq_{st} \frac{1}{T}\bar{\theta}\int_0^T \frac{[\hat{X}_{\Sigma,\bigwedge}(t)]^+}{\bar{T}_I} dt,$$

where π is some control other than FSF and $Q_{\Sigma,\bigwedge}^{\pi,\lambda}(t)$ is the queue-length process under this control. Using (EC.20), we have that if (EC.18) is violated under FSF, it is also violated under any other control. The proof is now complete. ■

We now turn to treat the SBR model. Unlike Lemma EC.4.1, the following lemma does allow for different (non-homogeneous) patience rates θ_i, $i \in \mathcal{I}$.

Lemma EC.4.3 (**asymptotic equivalence with the $\bigwedge$ model: with abandonments**) *Under the conditions of Theorem 5.2, for every $T \geq 0$,*

$$\lim_{\lambda\to\infty} Ab_i^{\lambda,T} \stackrel{d}{=} \lim_{\lambda\to\infty} \frac{\theta_i p_i}{\bar{\theta}} Ab_{\bigwedge}^{\lambda,T}, \ i \in \mathcal{I}. \tag{EC.24}$$

Proof: Relying on [9] and repeating the argument in the proof of Lemma EC.4.2 but now for the SBR model, we have that for all,

$$\left(\frac{L_1^\lambda(t)}{\sqrt{\lambda}}, \ldots, \frac{L_I^\lambda(t)}{\sqrt{\lambda}}\right) \Rightarrow \left(p_1\theta_1 \int_0^t [\hat{X}_\Sigma(s)]^+ ds, \ldots, p_I\theta_I \int_0^t [\hat{X}_\Sigma(s)]^+ ds\right) \text{ in } D^I, \text{ as } \lambda \to \infty,$$

where the limit $\hat{X}_\Sigma(s)$ is as in Theorem 4.1 (when applied to the SBR model) in [9] and it is the same limit obtained when applying that Theorem 4.1. to the $\bigwedge$ model. The proof is completed by applying (EC.21). ■

Proof of Theorem 5.2: The asymptotic feasibility part follows from Lemmas EC.4.2 and EC.4.3. Indeed, using the definitions of p_i, $\bar{\theta}$ and $\bar{\alpha}$ we have that $(\theta_i p_i)/(\bar{\theta}) = \bar{\alpha}_i/\bar{\alpha}$ so that asymptotic feasibility of the SBR model now follows by combining equations (EC.24) and (EC.18). Towards proving asymptotic optimality, note first that (25) requires, at least, that $P\left\{\mathrm{Ab}_i^{\lambda,T}/\alpha_i^\lambda \geq 1+\epsilon, \ i \in \mathcal{I}\right\} \leq \epsilon$ which in turn requires, at least, that

$$P\left\{\sum_{i\in\mathcal{I}} \frac{\lambda_i}{\lambda} \mathrm{Ab}_i^{\lambda,T} \geq (1+\epsilon) \sum_{i\in\mathcal{I}} \frac{\lambda_i}{\lambda} a_i^\lambda\right\} = P\left\{\frac{\mathrm{Ab}_\Sigma^{\lambda,T}}{\bar{\alpha}/\sqrt{\lambda}} \geq 1+\epsilon\right\} \leq \epsilon.$$

With this, the asymptotic optimality follows from the (already proved) asymptotic feasibility and from Lemma EC.4.1. ■

EC.5. Proofs for §5.3

Proof of Theorem 5.3: Let $A_{ij}^\lambda(t)$ be the aggregate number class-i customers routed (upon arrival or service completion) to agents in pool j by time t. By definition, then, $D_i^{\lambda,T} = (\Phi_{ij}^\lambda(T) + A_{ij}^\lambda(T))/A_i^\lambda(T)$ for $j = j(\{i\})$. For $j = j(\{i\})$, however, we have that $Z_j^\lambda(t) = Z_j^\lambda(0) + A_{ij}^\lambda(T) - S_j^\lambda(t)$,

where $S_j^\lambda(t)$ is the number of service completion in pool j by time t. In particular, $A_{ij}^\lambda(T) = Z_j^\lambda(t) - Z_j^\lambda(0) + S_j^\lambda(t) \geq -N_j^\lambda + S_j^\lambda(t)$. Hence, $D_i^{\lambda,T} \geq (-N_j^\lambda + S_j^\lambda(T))/A_i^\lambda(T)$. The result is now established using basic renewal theory, the stochastic boundedness of $I_j^\lambda(t)$ from Corollary 6.5 in [9], and the definition of N_j^λ. In particular, a direct consequence of Corollary 6.5 in [9] is that $(\mu \int_0^T Z_j^\lambda(s)ds)/\lambda \Rightarrow \mu\tilde{\nu}_j T$, where $\tilde{\nu}_j \geq (1-\delta)a_i/\mu$. Letting $\Psi^\lambda(T) := (\mu \int_0^T Z_j^\lambda(s)ds)/(\lambda T)$, we can apply the renewal strong law of large number and the random-time-change theorem (see Theorem 13.2.1 in [15]), to conclude that

$$\frac{S_j^\lambda(T)}{\lambda_i T} \stackrel{d}{=} \frac{\mathcal{N}_j\left(\mu \int_0^T Z_j^\lambda(s)ds\right)}{\lambda_i T} \Rightarrow \frac{\mu}{a_i}\tilde{\nu}_j, \text{ as } \lambda \to \infty,$$

where $\mathcal{N}_j(\cdot)$ is a unit-rate Poisson process. In particular,

$$\frac{S_j^\lambda(T)}{A_i^\lambda(T)} \Rightarrow \frac{\mu}{a_i}\tilde{\nu}_j \geq 1 - \delta_i, \quad \text{as } \lambda \to \infty. \tag{EC.25}$$

Finally, since $N_j^\lambda/\lambda \to \tilde{\nu}_j$, we can choose T large enough so that

$$\limsup_{\lambda\to\infty} \frac{N_j^\lambda}{A_i^\lambda(T)} \leq \epsilon/2. \tag{EC.26}$$

Combining (EC.25) and (EC.26), we conclude that there exists T large enough so that $P\left\{D_i^{\lambda,T} \leq 1 - \delta_i - \epsilon\right\} \leq \epsilon$. To complete the proof we need to show that, when $c_j \equiv c$ and $\mu_j \equiv \mu$, an asymptotically optimal solution to (33) is asymptotically optimal to (30). To this end, note that the objective function value of (3) constitutes a lower bound on the value of (30) as we have added a constraint in the transition from (3) to (30). Also, note that, for both (3) and (33), the $\bigwedge$-based staffing has a cost of $c \times [(\lambda/\mu) + (\beta^*/\mu)\sqrt{\lambda}]$ with β^* as in equation (9). Consequently, as the $\bigwedge$-based staffing for (33) has been proved to be asymptotically feasible for (30) it must also be asymptotically optimal. ■

EC.6. Proofs for §5.4

In this section we prove Theorem 5.4. We start by proving asymptotic feasibility of the proposed solution and then continue to asymptotic optimality. First, we have the following lemma which relates the non-asymptotic formulas for the staffing in (35) and

$$\tilde{N}_\Sigma^\lambda := \min\left\{N \in \mathbb{Z}_+ : P\left\{W_{\lambda,\mu}^{FCFS} > \sum_{i\in\mathcal{I}} \frac{\lambda_i}{\lambda} T_i^\lambda\right\} \leq \alpha\right\} \tag{EC.27}$$

to the condition, $\sum_{j\in\mathcal{J}}\mu_j N_j \geq \lambda+\beta^*\sqrt{\lambda}$ in (8), that we used to construct the $\bigwedge$-based staffing. We note that in the common-service-rates setting this condition is given by $\mu N_\Sigma^\lambda \geq \lambda+\beta^*\sqrt{\lambda}$.

Lemma EC.6.1 (**square-root safety staffing rule**) *Consider a sequence of $M/M/N$ queues with arrival rate λ for the λ^{th} queue, all with service rate μ. Also, assume that Assumption 2.1 holds and that the λ^{th} queue uses $\bar{N}_\Sigma^\lambda$ servers. If $\bar{N}_\Sigma^\lambda$ is defined through* (35) (*respectively* (EC.27)) *then*

$$\frac{\mu\bar{N}_\Sigma^\lambda-\lambda}{\sqrt{\lambda}}\to\beta^*>0,\ \textit{as}\ \lambda\to\infty, \tag{EC.28}$$

where β^* is the unique solution of

$$\frac{\mathbf{P}_\mu(\beta^*)}{\beta^*}=\bar{T}_I\ ,$$

(*respectively of* $\mathbf{P}_\mu(\beta^*)\,exp\left\{-\beta^*\sum_{i\in\mathcal{I}}a_i\bar{T}_i\right\}=\alpha$).

The proof of Lemma EC.6.1 can be found in [8] and we use it here to prove Theorem

Proof of Theorem 5.4: We start by proving asymptotic feasibility. First, by Lemma EC.6.1, we have that $\mu N_\Sigma^\lambda=\lambda+\beta^*\sqrt{\lambda}+o(\sqrt{\lambda})$ where β^* is determined by (9). Then, by our proposed SCS and GM staffing and design solution we have that, as $\lambda\to\infty$,

$$\frac{N^\lambda_{j(\{i\})}}{\lambda}\to(1-\delta/2)\frac{a_i}{\mu},\ i\in\{1,I\}\ \text{and}\ \frac{N^\lambda_{j(\{i\})}}{\lambda}\to(1-\delta)\frac{a_i}{\mu},\ i=2,\ldots,I-1,$$

and

$$\frac{N^\lambda_{j(\{i,i+1\})}}{\lambda}\to\left(\frac{\delta(a_i+a_{I+1})}{2\mu}\right),\ i=1,\ldots,I-1.$$

In particular, (EC.3) holds and we can apply Lemmas EC.3.1 and EC.3.2 to conclude the asymptotic feasibility of the proposed GM, SCS and FQR solution. Some care, however, is needed in applying Lemma EC.3.1 as the proof of the Lemma relies on the results in [9]. Those results, in turn, assume that the ratio vectors p and v are fixed and, in particular, do not scale with λ. In §5.4, however, we use non-asymptotic expressions do determine the ratio vector p^λ. Still, we claim that the results of Lemma EC.3.1 can be applied in this setting. First, using Lemma EC.6.1 and Proposition 1 in [11], it is easily verifiable that, as $\lambda\to\infty$, $p^\lambda\to p$ where p is determined as in (13)

with $\mu_1 = \mu$. (17). Then, a careful reading of the proof of Theorem 3.1 in [9] would reveal that the state-space collapse result there is still valid when, instead of a fixed ratio vector, we have a convergent sequence of ratio vectors. As Theorem 4.1 in [9] relies on Theorem 3.1 there, Theorem 4.1 there is also valid and can be used in the proof of Lemma EC.3.1 here. The proof of asymptotic feasibility is hence complete.

To prove asymptotic optimality, note that, by Lemma EC.2.2, we must have that $N_\Sigma^\lambda \geq \lambda + \beta^*\sqrt{\lambda} + o(\sqrt{\lambda})$ with β^* as in (9). In particular, as we have here that $c_j \equiv 1$, we have that a lower bound on the cost is $\lambda + \beta^*\sqrt{\lambda} + o(\sqrt{\lambda})$. Since, by Lemma EC.6.1, $N_\Sigma = \lambda + \beta^*\sqrt{\lambda}$ we have that the proposed solution achieves the lower bound asymptotically. Since it is also asymptotically feasible, it is asymptotically optimal. ■

References

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

[2] Armony, M., A. Mandelbaum. 2008. Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. Working Paper. New York University, New York, NY and Technion - The Israeli Institute of Technology, Haifa, Israel.

[3] Asmussen, S. 2003. *Applied Probability and Queues*, second ed., Springer, New York.

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

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

[6] Feldman, Z., I. Gurvich and W. Whitt. 2007. Managing quality of service in call centers via Queue-Ratio Routing: Asymptotic analysis and simulation-based optimization. Working paper, Columbia University, New York, NY.

[7] Gurvich, I., W. Whitt. 2008. Service-Level Differentiation in Many-Server Service Systems Via Queue-Ratio Routing. Working paper, Columbia University, New York, NY.

[8] Gurvich, I., W. Whitt. 2008. Service-Level Differentiation in Many-Server Service Systems Via Queue-Ratio Routing, Unabridged Version. Working paper, Columbia University, New York, NY. donwloadable at http://www.columbia.edu/∼ww2040

[9] Gurvich, I., W. Whitt. 2007. Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.*, forthcoming.

[10] Gurvich, I., W. Whitt. 2007. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management*, forthcoming.

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

[12] Karatzas, I., S.E. Shreve. 1991. *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer-Verlag, New York.

[13] Khas'minskii, R.Z. 1960. Ergodic properties of recurrent diffusion processes and stabilitization of the solution to the cauchy problem for parabolic equations. *Theory of Probability and its Applications* **5**(2) 179-196.

[14] Whitt, W. 1991. A Review of $L = \lambda W$ and extensions. *Queueing Systems* **9**(3) 235-268.

[15] Whitt, W. 2002. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer-Verlag, New York.