Service-Level Differentiation in Many-Server Service Systems: A Solution Based on Fixed-Queue-Ratio Routing

Itay Gurvich* Ward Whitt †

October 13, 2007

Abstract

Motivated by telephone call centers, we study large-scale service systems with multiple customer classes and multiple agent pools, each with many agents. For the purpose of delicately balancing service levels of the different customer classes, we propose a family of routing controls called *Fixed-Queue-Ratio* (FQR) rules. 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. We show that the proportions can be set to achieve desired service-level targets for all classes; these targets are achieved asymptotically as the total arrival rate increases. The FQR rule is a special case of the Queue-and-Idleness-Ratio (QIR) family of controls which in a previous paper where shown to produce an important *state-space collapse* (SSC) as the total arrival rate increases. This SSC facilitates establishing asymptotic results. In simplified settings, SSC allows us to solve a combined design-staffing-and-routing problem in a nearly optimal way. Our analysis also establishes a diminishing-returns property of flexibility: Under FQR, very moderate cross-training is sufficient to make the call center as efficient as a single-pool system, again in the limit as the total arrival rate increases.

*Columbia Business School, 4I Uris Hall, 3022 Broadway, New York, NY 10027. (ig2126@columbia.edu)

†IEOR Department, Columbia University, 304 S. W. Mudd Building, 500 West 120th Street, New York, NY 10027-6699. (ww2040@columbia.edu)

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. [17].

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. Typically, the call center will have different SL objectives for different classes; e.g., a call center with both regular and VIP customers might aim to respond to 80% of regular customers within 30 seconds, but 80% of VIP customers within 10 seconds.

The Three Basic Problems. Three basic problems must be solved: (i) *What SBR system should be used?* Given the set of customer classes, which agent pools should be used? what skills should the agents have? (**design**) (ii) *How many agents with each different skill set does the call center need?* (**staffing**), and (iii) *How should agents be assigned to customers in real time?* (**routing** or on-line control).

The total problem is quite complex, so that it is unproductive to search for an optimal solution. The size of large call centers tends to rule out most candidate algorithms. Thus we look for a *good* 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 an approach that is asymptotically optimal for specified problems as the scale increases. Our goal is to achieve *simplicity and asymptotic optimality*.

A Simplified Aggregate Approach to Staffing. Our approach to staffing illustrates our emphasis on simplicity: We propose choosing the total number of agents by acting, as much as possible, as if all agents had all skills and the call center operated as a single-class single-skill call center. We use a linear program (LP) to allocate the total number of agents to service pools, having designated skill sets. That drastic simplification is clearly not appropriate in general. To make that approach feasible, we must carry out the remaining

design and routing steps appropriately. In the design phase, we provide the required flexibility by ensuring that the agents have enough skills. In basic settings, the routing graph (showing which customer classes can be served by which agent pools) can be a connected generalized M model (see Figure 6), where each class is served by two pools and each pool serves at most two classes. For the routing phase, we develop an approach that provides the required service-levels for the various classes. Indeed, our main contribution in this paper is to propose a routing scheme that helps make the simple aggregate approach to staffing work.

Several recent papers have proposed this simplified aggregate approach to staffing. Theoretical support is contained in Armony [2] and Armony et. al. [4]. These papers establish asymptotic optimality of that staffing approach with appropriate routing for special classes of models as the total arrival rate increases. The first paper considers models with a single customer class and multiple agent types, while the second considers symmetric models with multiple customer classes but a single agent pool. Their asymptotic optimality follows Borst et al. [11], which formulates and establishes asymptotic optimality for the single-class, single-pool $M/M/N$ queue. The asymptotic framework is the now-familiar many-server heavy traffic regime, introduced by Halfin and Whitt [22], 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. [11] show that the *QED* regime arises naturally from economic considerations. We will be considering the QED regime throughout this paper.

The simplified aggregate approach to staffing is also a central idea in Wallace and Whitt [30], which develops 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. For specified approaches to the design and routing, simulation results show that the performance requirements are met with a total number of agents that is very close to what would be required for a single-class single-skill call center (usually within a single agent out of about 100). That good performance is achieved with a conventional routing scheme based on priority matrices. The most important requirement is that there be sufficient cross-training, but it suffices to have only limited cross-training; specifically, it suffices for each agent to have only two skills. The simulation experiments show that only minimal cross-training provides the required flexibility. The diminishing-returns property of cross-training for call centers has been proved under certain conditions by

Akşin and Karaesmen [1].

The approach in Wallace and Whitt [30] 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 Wallace and Whitt [30] seems to become more effective as the scale increases, it has not yet been shown to be asymptotically optimal as the scale increases. Here, in contrast, we establish asymptotic optimality for the scheme we propose.

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* ($Gc\mu$) *rule*, first introduced by Van-Meighem [28] for the multi-class and single-agent V model (see Figure 1), and generalized to more complicated networks by Mandelbaum and Stolyar [25]. A parallel to Mandelbaum and Stolyar [25] in a many-server setting has been provided by Atar [6], who characterizes a family of controls that achieve asymptotically optimal performance in the QED regime. We refer the reader to our subsequent paper [20] for a more elaborate discussion of this literature. While the controls in [6] can be easily implemented in a computerized environment, they are not nearly as simple as the $Gc\mu$ rule. It seems desirable, if possible, to find a family of controls for many-server systems that will bridge the gap between the simple and intuitive $Gc\mu$ rule and the more complicated controls in [6].

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 introduced in our previous paper [19]. A consequence of our analysis in [19] is that FQR makes the separate queue lengths asymptotically equal to the fixed proportions of 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 particular, FQR is a simple balancing rule reminiscent of the $Gc\mu$ rule and, like the $Gc\mu$ rule, it is a highly decentralized control. Indeed, in our paper [20] we show that, when assuming pool dependent service rates, the QIR family of controls asymptotically minimizes convex holding costs, paralleling for many-server systems the role that $Gc\mu$ plays in the case of single-server stations. (In general, that optimality requires going beyond FQR to allow state-dependent proportions.)

In addition to proposing the FQR control for SBR call center models, we make several other contributions in this paper: On the routing front, we characterize network and parameter conditions under which one may use FQR to achieve service-level differentiation while asymptotically minimizing the staffing costs. On the design front, we use FQR to rigorously support the diminishing-return property of flexibility, by showing that it suffices to consider structures where most of the agents have a single skill and only a small proportion of the agents have two skills. In particular, there is no need for agents with more than two skills in their skill set. On the staffing front, we find asymptotic optimal staffing levels for several staffing optimization problems. The aggregate approach to staffing is mostly preserved, with the exception of the need to solve an additional LP. (For basic models, the LP is only used to allocate the total number of agents to the agent pools having designated skill sets.) Although the asymptotic optimality results are not for the most general models, the general feasibility result in this paper and the simulation-based solution algorithm that we propose in the subsequent paper [15], show that our FQR-based solution can be applied in very general settings to obtain an extremely simple and practical solution with only minor compromise in terms of cost optimality.

Related Literature. An important feature of our work is that we simultaneously address the three problems of design, staffing and routing. Conventionally, these are treated separately and hierarchically. Wallace and Whitt [30] also addressed the three problems of design, staffing and routing together, but the only previous work we are aware of that establishes asymptotic optimality for all three problems has been done by Bassamboo et. al. [9] and [10]. They consider a staffing problem for an SBR system and aim at minimizing a cost function which reflects the costs of waiting times, abandonments and customer rejections but not QoS constraints. They focus on uncertainty in the arrival rates. As a consequence, they consider a different asymptotic regime, the efficiency-driven (ED) regime instead of the QED 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 service-level constraints that are widely used in the industry but can not be covered in the framework of [9, 10].

It should be mentioned that within the context of single-server stations, several papers have tackled the problem of service-level constraint satisfaction. The most relevant is probably Van-Meighem [29] which embeds the constraint-satisfaction problem into the convex-holding-cost setting of his paper [28], rather than dealing with it directly.

The analysis in this paper relies heavily on our previous paper [19], which establishes SSC results for the QIR generalization of FQR; see [19] for additonal discussion about SSC and more references.

Organization. We begin in §2 by formulating our staffing-under-SL-constraints optimization problem for the V model – see (1) – and proposing, first a feasible solution, and then an asymptotically optimal solution, based on FQR. For asymptotic optimality, a key step is a modified best-effort formulation; see (7). The relatively simple V model serves well to illustrate the essential ideas in the application of FQR to the staffing problem, but the results are also interesting in their own right, because the formulation and results need to be altered for more complex models.

In §3 we consider more general models, present more general versions of FQR and, based on [19], identify conditions on the network structure under which the control admits the important SSC. Throughout we focus on Markovian models, assuming mutually independent Poisson arrival processes, exponential service times and exponential times to abandon (when we consider customer abandonment).

In §4 we consider the multi-pool SBR setting with pool-dependent service rates and discuss also the implications to the special case with common service rate, which is similar to the model of Wallace and Whitt [30]. In §5 we introduce a waiting-time based version of FQR, which we call the fixed-waiting-ratio (FWR) rule. We show that FWR is asymptotically equivalent to FQR. We state conclusions in §6 and discuss some important directions for future research. For the sake of readability, we place some proofs and auxiliary results in an appendix.

2 The V Model

We start with the most elementary SBR design, known as the V model; it is depicted in Figure 1. The V model consists of a set $\mathcal{I} := \{1, \ldots, I\}$ of customer classes. Class-i customers arrive to the system according to a Poisson process $\{A_i(t), t \geq 0\}$ with rate λ_i, independently of all other classes. There is a queue for each customer class with unlimited capacity. Let $\lambda := \sum_{i \in \mathcal{I}} \lambda_i$ be the aggregate arrival rate. All customers are served by agents from a single agent pool and all service times are assumed to be independent and identically distributed (i.i.d.) exponential random variables with common rate μ, regardless of the customer class. (We do not consider customer abandonment at this point.)

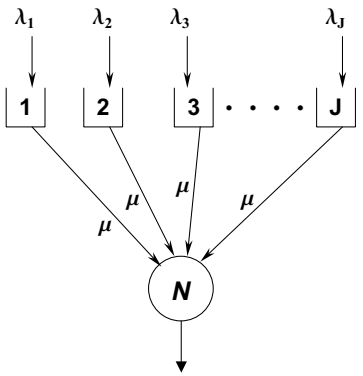

Figure 1: The V model

Our proposed staffing-and-routing optimization problem involves minimizing the total number of agents, N, subject to SL constraints. These constraints are expressed in terms of the proportion of class-i customers who wait more than T_i units of time before starting service. We call T_i the **class-i SL target**. Letting W_i^π be the steady-state waiting time of class-i customers under a control rule π (assuming it is well defined), we consider the **initial optimization problem**:

$$\begin{array}{ll} \text{minimize} & N \\ \text{subject to:} & P\{W_i^\pi > T_i\} \leq \alpha, \quad i \in \mathcal{I}, \\ & N \in \mathbb{Z}_+, \pi \in \Pi. \end{array} \tag{1}$$

where α is a specified probability satisfying $0 < \alpha < 1$ and $\mathbb{Z}_+$ is the set of positive integers. We call α the **SL probability**. The number of agents, N, and the routing control, π, are the decision variables. The

routing control π is assumed to belong to a family of admissible controls Π, which will be described in greater detail later in this section.

Remark 2.1 (**common SL probability** α) Our analysis exploits the requirement that the SL probability α be common for all classes. That formulation allows service differentiation only through the class-dependent SL targets T_i. We believe that this is not a serious restriction from a practical perspective. Indeed, we believe that the overall control should be easier to manage if only one of the SL parameters – T_i or α_i – is allowed to depend on i. In our formulation, with common SL probability α, having SL targets ordered by $T_1 < T_2$ clearly means better quality of service for class 1 than class 2. Of course, that will still be true provided we have consistent orderings for the SL targets and SL probabilities, such as $T_1 \leq T_2 \leq \cdots \leq T_I$ and $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_I$, but the consequences of inconsistent ordering, such as $T_1 < T_2$ and $\alpha_1 > \alpha_2$, are not so evident, and thus more difficult to manage. ∎

It should be noted that optimization problem (1) is not well formulated in its current form, because we have not 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 [4], 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, consistent with (1), such as the alternating-priority control discussed by Armony et. al. [4].

We will show later that these difficulties can be overcome by modifying the formulation, but the formulation (1) is of interest because it is commonly used in industry. Thus we begin by showing how to construct a feasible solution for optimization problem (1). Of course, we will want the solution to be simple and reasonable. Our solution relies upon FQR. We begin with a version of FQR that is tailored to the V model; we will introduce a more general version of FQR later. Let $Q_i(t)$ be the class-i queue length (not counting customers in service) at time t and let $Q_\Sigma(t) := \sum_{i \in \mathcal{I}} Q_i(t)$ be the aggregate queue length.

Definition 2.1 (**FQR for the V model**)

Given a probability vector $p := (p_1, \dots p_I)$, FQR for the V model is defined as follows:

- **Upon arrival**, *a class-i customer is admitted to service immediately if there are any idle agents, being assigned to any available agent; otherwise the customer waits in the class-i queue, to be served in order of arrival.*
- **Upon service completion**, *the newly available agent is assigned (and starts serving) the customer from the head of queue i^*, where i^* is the queue with the biggest imbalance among all the queues, i.e., for which the value of $Q_i(t) - p_i Q_\Sigma(t)$ is largest. If there are no waiting customers to serve, the agent goes idle (at the end of an agent queue, assuming we want to maintain fairness for the agents).*

Ties are broken in an arbitrary but consistent manner, so that the $(I+1)$-dimensional stochastic process $(Q_1(t), \dots, Q_I(t), I(t))$ is a continuous-time Markov chain (CTMC) with stationary transition probabilities. An example of a consistent decision rule is to always choose a customer from the class i with the largest index among the classes that maximize $Q_i(t) - p_i Q_\Sigma(t)$.

We observe that, as long as all p_i's are strictly positive, FQR is equivalent to choosing the class for which the value of $Q_i(t)/p_i$ is largest. Also, note that at each epoch when an agent becomes free, FQR tries to drive the system towards a balanced state in which $Q_i(t) \approx p_i Q_\Sigma(t)$ for all $i \in \mathcal{I}$. Having $Q_i(t)$ actually be asymptotically equivalent to $p_i Q_\Sigma(t)$ corresponds to the desired state-space collapse. We will prove that state-space collapse holds in great generality, including for this V model.

This state-space collapse is, in turn, a key component in establishing an asymptotically feasible solution to (1). There are two other components. The first component is the **aggregate queue behavior:** Because of the special structure of the V model, the steady-state overall queue length, Q_Σ, under any non-idling policy, has a probability law that is equivalent to the steady-state queue length in a single-class $M/M/N$ queue with arrival rate λ, service rate μ and FCFS discipline, which we denote by $Q^{FCFS}_{\lambda,\mu}(N)$. It is well known that $Q^{FCFS}_{\lambda,\mu}(N)$ has an exponential distribution except for an atom at the origin.

The second component is a **heavy-traffic Little's law:** In the QED asymptotic regime, we show that Little's law nearly holds for the time-dependent random variables (instead of just the steady-state means and time averages); i.e., $Q_i(t) \approx \lambda_i W_i(t)$, where $W_i(t)$ is the associated virtual waiting time at time t; see Theorem 6.1 in [19]. (More precisely, $W_i(t)$ is defined via $W_i(t) := \inf\{s \geq 0 : D_i(t+s) \geq Q(0) + A_i(t)\}$,

where $D_i(t)$ is the number of class-i customers that have entered service by time t.) As a consequence, under regularity conditions, $Q_i \approx \lambda_i W_i$, where Q_i is the steady state version of $Q_i(t)$ and W_i is the associated steady-state waiting time. (Since we have Poisson arrivals, the steady-state actual and virtual waiting times have the same distribution.) As a further consequence, $Q_\Sigma \approx \sum_{i\in\mathcal{I}} \lambda_i W_i$.

When combined, these three components suggest a very intuitive, yet feasible, solution to (1):

- **Staffing: Single-Class Staffing (SCS):** Choose the staffing level $\bar{N}_\Sigma$ by solving a simple $M/M/N$ problem given by

$$\bar{N}_\Sigma = min\left\{N \in \mathbb{Z}_+ : P\left\{Q_{\lambda,\mu}^{FCFS}(N) > \sum_{i\in\mathcal{I}} \lambda_i T_i\right\} \leq \alpha\right\}. \tag{2}$$

- **Control: Fixed-Queue-Ratio (FQR)** Use FQR as defined in Definition 2.1 with the p vector

$$p_i := \frac{\lambda_i T_i}{\sum_{k\in\mathcal{I}} \lambda_k T_k}\,, \quad i \in \mathcal{I}. \tag{3}$$

We refer to this overall candidate solution as the *Single-Class-Staffing* (SCS) and *Fixed-Queue-Ratio* (FQR) solution. We expect this SCS & FQR solution to be asymptotically feasible because we have the following chain of approximations:

$$\begin{aligned} P\{W_i > T_i\} &\approx P\{Q_i > \lambda_i T_i\} \approx P\{p_i Q_\Sigma > \lambda_i T_i\} \\ &\approx P\left\{Q_\Sigma > \sum_{k\in\mathcal{I}} \lambda_k T_k\right\} \approx P\left\{Q_{\lambda,\mu}^{FCFS}(N) > \sum_{k\in\mathcal{I}} \lambda_k T_k\right\} \leq \alpha\,. \end{aligned} \tag{4}$$

The state-space collapse produced by FQR allows us, then, to construct an extremely simple solution based on the specified target proportions $p_i,\ i \in \mathcal{I}$. The informal analysis in (4) also suggests that we could obtain an asymptotically equivalent routing control using waiting times instead of queue lengths. The analogous *Fixed-Waiting-Ratio* (FWR) *rule* stipulates that a newly available agent would select the customer from the head of the queue, from among the queues that he is eligible to serve, that has the largest ratio W_i/T_i, where W_i here denotes the elapsed waiting time for the customer at the head of the i^{th} queue at that instant. The special case in which we set a common target time for all customer classes, $T_i \equiv T$, is the *global first-come-first-served* rule. Through most of the paper we focus on FQR instead of FWR because it usually is easier to implement; knowledge of the current queue lengths is usually more readily available

than knowledge of the elapsed waiting times. We discuss FWR in §5.

Asymptotic feasibility (or near feasibility) of SCS & FQR is obtained through the QED asymptotic framework. To this end, we add the superscript λ to all the previous notation. For example, $\bar{N}_\Sigma^\lambda$ will be the indexed version of the quantity determined through (2), and Q_i^λ and W_i^λ will denote, respectively, the steady-state class-i queue length and waiting time under SCS & FQR, in the model with total arrival rate λ. We consider a family of V models indexed by the total arrival rate λ and let $\lambda \to \infty$. For simplicity, throughout the paper, we will let λ approach ∞ through a sequence of values and, in particular, consider sequences of systems. 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. **All assumptions stated in the text are assumed to hold thereafter throughout the paper.**

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 $\bar{T}_i,\ i \in \mathcal{I}$.

Assumption 2.1 is important for the asymptotics, but does not affect the solution as it would be used in applications, because our proposed solution is independent of the scaling and uses only the actual target values T_i^λ and not the $\bar{T}_i$'s. However, the assumption is critical for putting us in the QED regime. That can be expected because waiting times are known to be of order $1/\sqrt{\lambda}$ in the QED regime. Extensive experience shows that these QED approximations exhibit excellent performance; see [11], [4], [27]. We will substantiate that through a numerical example in §3.

Here is the first supporting theorem of this section:

Theorem 2.1 (**asymptotic feasibility and state-space collapse for the V model**)

If SCS and FQR are used (*and Assumption 2.1 holds*)*, then*

$$\limsup_{\lambda\to\infty} P\{W_i^\lambda > T_i^\lambda\} \le \alpha, \quad i \in \mathcal{I}, \tag{5}$$

and, for any $\eta > 0$,

$$\lim_{\lambda\to\infty} P\{|Q_i^\lambda - p_i Q_\Sigma^\lambda| > \eta\sqrt{\lambda}\} = 0, \quad i \in \mathcal{I}\,. \tag{6}$$

The bound in (5) establishes the feasibility (in an asymptotic sense) of SCS & FQR, while the limit in (6) establishes the important state-space collapse: It is the asymptotic version of the assertion that $Q_i(t) \approx p_i Q_\Sigma(t)$. Theorem 2.1 is given in steady-state terms. Consistent with the QED literature, the proof is based on a stochastic-process limit, followed by a limit-interchange argument that extends the state-space collapse to the corresponding steady-state statement.

We conclude this section with an asymptotic-optimality result, which requires a modification of formulation (1). Following [4], we propose a best-effort formulation closely related to (1), where class I is assumed (without loss of generality) to be the best-effort class. For this purpose, let π be the routing policy; let $W^{\pi,\lambda} := \sum_{i\in\mathcal{I}}(\lambda_i/\lambda)W_i^{\pi,\lambda}$ be the steady-state waiting time of the entire customer population in model λ using policy π; and let T_I^λ be a new average-speed-of-answer (ASA) target, applying to all classes, but remove the SL target for class I. The new **best-effort optimization problem** is:

$$\begin{array}{ll} \text{minimize} & N \\ \text{subject to:} & E[W^{\pi,\lambda}] \le T_I^\lambda, \\ & P\{W_i^{\pi,\lambda} > T_i^\lambda\} \le \alpha, \quad 1 \le i \le I-1, \\ & N \in \mathbb{Z}_+, \pi \in \Pi. \end{array} \tag{7}$$

Let N_*^λ be the **optimal-staffing level** for (7) in model λ. (Since the staffing level is discrete, an optimal value N_*^λ necessarily exists for each λ.) Just as in [4], formulation (7) prevents pathologies possible with (1). Constraining the average waiting time of the entire customer population prevents temporarily compromising the service level of one class for the benefit of another class, by imposing a certain consistency across all classes. Another formulation that prevents possible pathologies is one that constraints the average waiting time of each class rather than the tails of the waiting time distributions. While we discuss this formulation in Remark 2.2 we prefer to consider formulation (7) which is closer to the original formulation (1) that is widely used in industry.

The following is an adaptation of SCS and FQR to the setting of (7), which will be explained afterwards:

- **Staffing: Single-Class Staffing ($\widetilde{SCS}$)** Choose the staffing level $\bar{N}_\Sigma^\lambda$ by solving a simple $M/M/N$ problem given by

$$\bar{N}_\Sigma^\lambda := min\left\{N \in \mathbb{Z}_+ : E\left[Q_{\lambda,\mu}^{FCFS}(N)\right] \le \lambda T_I^\lambda\right\}. \tag{8}$$

- **Control: Fixed-Queue-Ratio ($\widetilde{FQR}$)** Use FQR with:

$$p_{I-1} : P\{Q^{FCFS}_{\lambda,\mu}(\bar{N}^{\lambda}_{\Sigma}) > 0\}e^{\{-(\lambda_{I-1}/\lambda p_{I-1})(\bar{N}^{\lambda}_{\Sigma}\mu-\lambda)T^{\lambda}_{I-1}\}} = \alpha, \text{ and}$$
$$\frac{p_i}{p_{I-1}} = \frac{\lambda_i T^{\lambda}_i}{\lambda_{I-1}T^{\lambda}_{I-1}}\ ,\ 1 \le i \le I-2. \tag{9}$$

To distinguish these definitions from the previously defined SCS and FQR, we use the notation $\widetilde{SCS}$ and $\widetilde{FQR}$. The new single-class staffing $\widetilde{SCS}$ follows from the constraint on the global waiting time. The new version of FQR, $\widetilde{FQR}$, is a bit more complex due to the best-effort structure of (7) which leaves an extra degree of freedom to be determined. That extra degree of freedom is applied to the last *best-effort class* I. Note that there is no individual SL constraint for class I.

Using our previous reasoning, we expect that under $\widetilde{FQR}$, $W^{\lambda}_{I-1} \approx Q^{\lambda}_{I-1}/\lambda_{I-1} \approx p_{I-1}Q^{FCFS}_{\lambda,\mu}(N)/\lambda_{I-1}$, where $N = \bar{N}^{\lambda}_{\Sigma}$. A similar Little's law reasoning applies also for the $M/M/N$ system, leading to $Q^{FCFS}_{\lambda,\mu}(N) \approx \lambda W^{FCFS}_{\lambda,\mu}(N)$, where $W^{FCFS}_{\lambda,\mu}(N)$ is the steady-state waiting time in an $M/M/N$ queue with arrival rate λ, service rate μ and N servers. In particular, we expect that

$$W^{\lambda}_{I-1} \approx \frac{\lambda}{\lambda_{I-1}} p_{I-1} W^{FCFS}_{\lambda,\mu}(N).$$

Hence, one should choose p_{I-1} so that

$$P\left\{\frac{\lambda}{\lambda_{I-1}} p_{I-1} W^{FCFS}_{\lambda,\mu}(N) > T^{\lambda}_{I-1}\right\} = \alpha,$$

which is equivalent to the first part of (9), using the known form for the steady-state waiting-time distribution in the $M/M/N$ model (exponential plus atom at the origin). Once p_{I-1} is determined, fixing the ratios p_i/p_{I-1} as proposed in the second line of (9), we expect to have

$$\begin{aligned} P\{W^{\lambda}_i > T_i\} &\approx P\{Q^{\lambda}_i > \lambda_i T^{\lambda}_i\} \approx P\{(p_i/p_{I-1})Q^{\lambda}_{I-1} > \lambda_i T^{\lambda}_i\} = P\{Q^{\lambda}_{I-1} > \lambda_{I-1}T^{\lambda}_{I-1}\} \\ &\approx P\{W^{\lambda}_{I-1} > T^{\lambda}_{I-1}\} = \alpha. \end{aligned}$$

Before we state our asymptotic-optimality result, we need to define the family Π of admissible controls (routing policies). To this end, we make the following definitions: We say that **class FCFS** holds if cus-

tomers are served FCFS (first-come first-served) within each class. We say that **all customers are served** if it is not allowed to block or overflow customers; i.e., we require that

$$Q_i(t) = A_i(t) - D_i(t) - Z_i(t) \quad \text{for all} \quad t \geq 0 \,,$$

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.

Let $Q_i^\pi(t)$ be the queue length and let $W_i^\pi(t)$ be the virtual waiting time of class i at time t under the control π, and denote by $(Q^\pi(t), W^\pi(t)) = ((Q_i^\pi(t), W_i^\pi(t)) : i \leq i \leq I)$ the corresponding vector-valued process including all classes. We say that **the control π admits a steady state** if there exists a random vector $(Q^\pi(\infty), W^\pi(\infty))$ such that

$$(Q^\pi(t), W^\pi(t)) \Rightarrow (Q^\pi(\infty), W^\pi(\infty)) \text{ in } \mathbb{R}^{2I} \quad \text{as} \quad t \to \infty \,,$$

where $\Rightarrow$ denotes convergence in distribution. Then the distribution of $(Q^\pi(\infty), W^\pi(\infty))$ is the steady-state distribution. As a consequence, there will be an associated steady-state distribution at arrival epochs, constructed using the classic Palm transformation, see Chapters 1 and 2 of [8], which agrees with the steady-state distribution at arbitrary times (above) by the PASTA (Poisson Arrivals See Time Averages) property, because we have exogenous Poisson arrival processes.

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. The detailed definition of Π is, then, as follows:

Definition 2.2 (**admissible routing policies**)

*For a given staffing level N, we say that a routing policy π is **admissible** if:* (1) *it is non-anticipative,* (2) *it is non-preemptive,* (3) *it satisfies class FCFS,* (4) *all customers are served, and* (5) *the control π admits a steady state. Let Π be the set of all admissible routing policies.*

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 $x \to \infty$, and $f(x)$ is $O(g(x))$ if $\limsup_{x\to\infty} f(x)/g(x) < \infty$. With the above definitions, we can state

our asymptotic optimality result. In passing, we note that our notion of asymptotic optimality is specified by the last line of (10) in Theorem 2.2.

Theorem 2.2 (**asymptotic optimality for the V model**)

If $\widetilde{SCS}$ *and* $\widetilde{FQR}$ *are used, then*

$$\begin{aligned} \limsup_{\lambda\to\infty} P\{W_i^{\pi,\lambda} > T_i^\lambda\} &\leq \alpha, \quad 1 \leq i \leq I-1, \\ \limsup_{\lambda\to\infty} \frac{E[W^{\pi,\lambda}]}{T_I^\lambda} &\leq 1, \\ [\bar{N}_\Sigma^\lambda - N_*^\lambda]^+ &= o(\sqrt{\lambda}) \quad as \quad \lambda \to \infty\,. \end{aligned} \tag{10}$$

Remark 2.2 (**an average-waiting-time formulation**) An alternative to formulation (7), which prevents the pathologies mentioned earlier in this section, is an average-waiting-time formulation in which all the SL constraints are replaced by average-waiting-time constraints. Specifically, we then minimize N subject to $E[W_i^{\pi,\lambda}] \leq T_i^\lambda$ for $i \in \mathcal{I}$. The arguments leading to Theorem 2.2 are easily adapted to show that it is then asymptotically optimal to staff with N_Σ^λ agents, where

$$N_\Sigma^\lambda := min\left\{N \in \mathbb{Z}_+ : E\left[Q_{\lambda,\mu}^{FCFS}(N)\right] \leq \sum_{i\in\mathcal{I}} \lambda_i T_i^\lambda\right\},$$

and, for control, to use FQR with

$$p_i = \frac{\lambda_i T_i^\lambda}{\sum_{k\in\mathcal{I}} \lambda_k T_k^\lambda}, \;\; i \in \mathcal{I}. \blacksquare$$

Theorems 2.1 and 2.2 will be proved as special cases of much more general results later. The intuition developed for the V model allows us next to take one step up in complexity and analyze a more general SBR model.

3 General SBR systems

In §2 we restricted ourselves both in terms of the design (the V model) and the underlying parameters (equal service rate). In this section we begin the description and construction of solutions for more general SBR

models. These models require a more detailed discussion of the design decision, the FQR control and the dependency between them. As before, FQR will be defined as in Definition 3.3, but we will also introduce another notion called Generalized FQR (GFQR) based on Atar [6].

Here is the outline for this section: after stating the general staffing problem, we formulate a fundamental mathematical program – see (16) – that does two important things: (i) it characterizes the allowed designs and (ii) it gives a first-order estimate of the optimal staffing levels. Based on the solution of this optimization problem, we characterize a set of structural and parametric conditions that guarantee that FQR will achieve some form of state-space collapse. This section is dedicated mainly to the analysis of FQR in general settings, with emphasis on its dependency on the design and staffing.

We begin by redefining the model: we consider a system with a set $\mathcal{I} = \{1, \ldots, I\}$ of customer classes and a set $\mathcal{J} = \{1, \ldots, I\}$ of agent types. The number of agents of type j (which will be a decision variable) is denoted by N_j and we let N be the corresponding vector $N = (N_1, \ldots, N_J)$. As before, 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. Let $\mu_{i,j}$ be the service rate at which a type-j agent can serve a class-i customer. Alternatively, the mean handling time of a class-i customer by a type-j agent is $1/\mu_{i,j}$. We set $\mu_{i,j} = 0$ whenever type-j agents do not have the required skill to serve class-i customers. The possible-routing graph for this SBR system has a natural representation as a bipartite graph 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: $E = \{(i, j) \in \mathcal{I} \times \mathcal{J} : \mu_{i,j} > 0\}$. An edge (i, j) is present in the routing graph if class-i customers can be served by type-j agents. In addition to the above agents of type-j will incur a cost of c_j per unit of time and we will allow the system to impose additional constraints on the staffing vector to reflect union contracts, hiring and training constraints or other managerial considerations. As before, for the asymptotic analysis we will construct a sequence of SBR systems indexed by the aggregate arrival rate λ. The service rates $\mu_{i,j}$ and the routing graph are held fixed. To make this structure explicit we define the general SBR model below:

Definition 3.1 (**the general SBR model**)

The general SBR model is defined through the following:

- ***Arrivals:*** *Class-i customers arrive according to a Poisson process with rate λ_i. Let $\lambda := \sum_{i \in \mathcal{I}} \lambda_i$ and $a_i := \lambda_i / \lambda$.*

- ***Customer Abandonment:*** *Class-i customers have an exponential patience with rate θ_i, so that setting $\theta_i \equiv 0$ reduces the model to the non-abandonment model.*

- ***Service Times:*** *The service time of a class-i customer by a type j agent has an exponential distribution with rate $\mu_{i,j}$. Let $\mu_{i,j} = 0$ if type-j agents are not allowed to serve class-i customers.*

- ***Routing Graph:*** *The routing graph is a sub-graph of $E := \{(i,j) \in \mathcal{I} \times \mathcal{J} : \mu_{i,j} > 0\}$.*

- ***Agents Costs:*** *Type-j agents incur a cost of c_j per unit of time.*

- ***System Constraints:*** *We allow constraints of the form $N \in \mathcal{A}^\lambda := \mathcal{A}^\lambda_b$, where $\mathcal{A}^\lambda_b = \{N \in \mathbb{Z}^J_+ : A \cdot N \leq b^\lambda\}$, for some matrix $A \in \mathbb{R}^{d \times J}$, $d \in \mathbb{Z}_+$ and $b^\lambda = \lambda \hat{b}$ for some $\hat{b} \in \mathbb{R}^d$. Let $\tilde{\mathcal{A}}^\lambda := \tilde{\mathcal{A}}^\lambda_b$ be the set obtained from $\mathcal{A}^\lambda$ by relaxing the integrality assumptions. That is, $\tilde{\mathcal{A}}^\lambda = \{N \in \mathbb{R}^J_+ : A \cdot N \leq b^\lambda\}$.*

3.1 The Optimization Problem and the Fundamental Mathematical Program

Our next step is to generalize the SBR optimization problems in (1) and (7) to this more general setting. While stability is almost trivial in the setting of §2, that is no longer the case in this SBR system. To avoid issues of stability, we now look at a slightly different formulation involving long-run averages instead of steady-state quantities. Towards that end, 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.,

$$\bar{W}^{\lambda,T} := \frac{\sum_{i=1}^{I} \sum_{k=1}^{A_i^\lambda(T)} w_{i,k}^\lambda}{A^\lambda(T)} \text{ and } \bar{F}_i^{\lambda,T}(y) := \frac{\sum_{k=1}^{A_i^\lambda(T)} \mathbf{1}\{w_{i,k}^\lambda > y\}}{A_i^\lambda(T)}\,, \tag{11}$$

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. The **initial SBR optimization problem** is given by:

$$\begin{array}{ll} \text{minimize} & \sum_{j \in \mathcal{J}} c_j N_j \\ \text{subject to:} & \limsup_{T \to \infty} \bar{F}_i^{\lambda,T}(T_i^\lambda) \leq \alpha, \quad 1 \leq i \leq I-1, \\ & N \in \mathcal{A}^\lambda, \quad \pi \in \Pi. \end{array} \tag{12}$$

The SBR analogue of the best-effort optimization problem (7) of §2 is obtained by replacing the SL

constraint of class I with a global waiting time constraint to get the **SBR best-effort optimization problem**:

$$\begin{array}{ll} \text{minimize} & \sum_{j\in\mathcal{J}} c_j N_j \\ \text{subject to:} & \limsup_{T\to\infty} \bar{W}^{\lambda,T} \le T_I^{\lambda}, \\ & \limsup_{T\to\infty} \bar{F}_i^{\lambda,T}(T_i) \le \alpha, \quad 1 \le i \le I-1, \\ & N \in \mathcal{A}^{\lambda}, \quad \pi \in \Pi. \end{array} \tag{13}$$

We begin to consider the limiting behavior as $\lambda \to \infty$. With that in mind, we observe that there is an evident connection between the problem (12) above and its natural deterministic counterpart:

$$\begin{array}{ll} \text{Minimize} & \sum_{j\in\mathcal{J}} c_j N_j \\ \text{Subject to:} & \sum_{j\in J} \mu_{i,j} N_j y_{i,j} \ge \lambda_i, \quad i \in \mathcal{I}, \\ & \sum_{i\in\mathcal{I}} y_{i,j} \le 1, \quad j \in \mathcal{J}, \\ & N \in \tilde{\mathcal{A}}^{\lambda}, \quad y_{i,j} \ge 0, \quad i \in \mathcal{I}, \quad j \in \mathcal{J}, \end{array} \tag{14}$$

where the service-level constraints have been replaced by simple flow constraints. By the definition of $\tilde{\mathcal{A}}^{\lambda}$ and recalling that $\lambda_i = a_i\lambda$, (14) then is equivalent to the mathematical program:

$$\begin{array}{ll} \text{Minimize} & \sum_{j\in\mathcal{J}} c_j \nu_j \\ \text{Subject to:} & \sum_{j\in J} \mu_{i,j} \nu_j x_{i,j} \ge a_i, \quad i \in \mathcal{I}, \\ & \sum_{i\in\mathcal{I}} x_{i,j} \le 1, \quad j \in \mathcal{J}, \\ & A\nu \le b, \quad \nu_j \ge 0, \quad x_{i,j} \ge 0, \quad i \in \mathcal{I}, \quad j \in \mathcal{J}\ . \end{array} \tag{15}$$

We claim that, if there exists an optimal solution $(\bar{\nu}, \bar{x})$ to (15), then there also exists an optimal solution $(\bar{\nu}, x')$ for which the constraints $\sum_{j\in\mathcal{J}} \mu_{i,j}\bar{\nu}_j x'_{i,j} \ge a_i$ hold as equalities. Indeed, if one of these inequalities is strict, then we can always decrease the value of $x_{i,j}$ for some $i \in \mathcal{I}$ and $j \in \mathcal{J}$ to replace this inequality with equality. Hence, we may consider, without loss of generality, the following **fundamental mathematical program**:

$$
\begin{array}{ll}
\text{Minimize} & \sum_{j\in\mathcal{J}} c_j\nu_j \\
\text{Subject to:} & \sum_{j\in J} \mu_{i,j}\nu_j x_{i,j} = a_i, \quad i\in\mathcal{I}, \\
& \sum_{i\in\mathcal{I}} x_{i,j} \le 1, \quad j\in\mathcal{J}, \\
& A\nu \le b, \quad \nu_j \ge 0, \quad x_{i,j}\ge 0, \quad i\in\mathcal{I}, \quad j\in\mathcal{J}\ .
\end{array}
\tag{16}
$$

Clearly, a feasible solution (N, y) for (14) corresponds to a solution $(\bar{\nu}, \bar{x})$ for (16) through $\bar{y} = \bar{x}$ and $N = \bar{\nu}\lambda$. If the problem (16) has multiple solutions, then we will select one, and make further assumptions about that solution. The first additional assumption is a critical-loading assumption, to be in force hereafter. This assumption is needed to place the system in the QED heavy-traffic limiting regime.

Assumption 3.1 (**critical loading**) *Let $(\bar{\nu}, \bar{x})$ be the selected optimal solution to (16). For any x' such that $(\bar{\nu}, x')$ is a feasible solution to (16), we require that $\sum_{i\in\mathcal{I}} x'_{i,j} = 1$ for all $j \in \mathcal{J}$.*

We now offer an intuitive explanation of this assumption: For a given feasible solution $(\bar{\nu}, x')$ to (16), we interpret $\sum_{i\in\mathcal{I}} x'_{i,j}$ as a first-order estimate of the proportion of time that pool j agents are busy giving service. Accordingly, $1 - \sum_{i\in\mathcal{I}} x'_{i,j}$ is the first-order approximation for the proportion of time they are idle. Intuitively, then, Assumption 3.1 ensures that, for the fixed staffing vector $\bar{\nu}$, there exists no allocation of the agents to customers so that one of the agent pools experiences significant idleness. In other words, Assumption 3.1 requires that all of the agents are busy almost all the time. Finally, we note that whenever the service rates are pool-dependent or class-dependent, i.e., when $\mu_{i,j} = \mu_i$ for all $j \in \mathcal{J}$ or $\mu_{i,j} = j$ for all $i \in \mathcal{I}$, the constraints $\sum_{j\in\mathcal{J}} \mu_{i,j}\bar{\nu}_j x'_{i,j} = a_i,\ i\in\mathcal{I}$ will automatically guarantee that $\sum_{i\in\mathcal{I}} x'_{i,j} = 1$ for all $j \in \mathcal{J}$.

Remark 3.1 (**alternative fomulations**) If we replace (12) with a best-effort formulation as in (7), then we still obtain the same deterministic counterpart (14) and in turn to the same mathematical program (16). Hence, all our results in this section apply equally well to both formulations. ∎

Remark 3.2 (**replacing (16) with an LP**) In the introduction we mentioned the need to solve linear programs (LP's), but both optimization problems (14) and (16) are clearly nonlinear in their stated form. However, with the restriction to optimal solutions that satisfy (3.1), they can be converted into LP's. First, the set of optimal solutions to (16) that satisfy Assumption 3.1 is obtained simply by replacing the inequality constraint $\sum_{i\in\mathcal{I}} x_{i,j} \le 1$ with an equality constraint. Henceforth in this remark assume that has been done.

The LP is obtained by introducing the variables $N_{i,j} = N_j y_{i,j},\ i \in \mathcal{I}, j \in \mathcal{J}$ in (14) and the variables $\nu_{i,j} = \nu_j x_{i,j},\ i \in \mathcal{I}, j \in \mathcal{J}$ in (16). For example, assuming the dimensions of the matrix A are $d \times J$, we define the LP

$$\begin{array}{ll} \text{Minimize} & \sum_{j\in\mathcal{J}} c_j \left(\sum_{i\in\mathcal{I}} \nu_{i,j}\right) \\ \text{Subject to:} & \sum_{j\in J} \mu_{i,j}\nu_{i,j} = a_i, \quad i \in \mathcal{I}, \\ & \sum_{j=1}^{J} A_{k,j} \left(\sum_{i\in\mathcal{I}} \nu_{i,j}\right) \le b_j,\ \forall k = 1, \ldots, d, \\ & \nu_{i,j} \ge 0, \quad i \in \mathcal{I}, \quad j \in \mathcal{J}. \end{array} \tag{17}$$

This LP is equivalent, in terms of the set of optimal solutions, to (16). To see this, note that from an optimal solution $\nu_j,\ j \in \mathcal{J}$ and $x_{i,j},\ i \in \mathcal{I}, j \in \mathcal{J}$, for (16), we can construct a solution for (17) with the same objective function value by setting $\nu_{i,j} = \nu_j x_{i,j},\ i \in \mathcal{I}, j \in \mathcal{J}$. Also, from any feasible solution, $\nu_{i,j},\ i \in \mathcal{I}, j \in \mathcal{J}$, for (17), we can construct a feasible solution, $\nu_j,\ j \in \mathcal{J}$ and $x_{i,j},\ i \in \mathcal{I}, j \in \mathcal{J}$ for (16) by setting $\nu_j = \sum_{i\in\mathcal{I}} \nu_{i,j}$ and $x_{i,j} = \nu_{i,j}/\nu_j$ whenever $\sum_{i\in\mathcal{I}} \nu_{i,j} > 0$ and setting ν_j and $x_{i,j}$ to 0 otherwise. Observe that in the last step we used the fact that we are only interested in optimal solutions with $\sum_{j\in\mathcal{J}} x_{i,j} = 1$. As a consequence, there is an equivalence between the optimal solutions for the optimization problems (16) that satisfy Assumption 3.1 and the set of optimal solution for (17). A similar construction and equivalence argument works for (14) and its associated LP. Hence, there is no need to solve a nonlinear program to obtain these allocation values. Greater computational efficiency can be obtained by solving the LP and then converting its solution. ∎

Henceforth we only refer to optimal solutions of (16); the way they were obtained is immaterial for our discussion. When the choice of solution $(\bar{\nu}, \bar{x})$ is clear from the context we will occasionally use the notation $\mathcal{J}$ when actually referring to the subset of $\mathcal{J}$ with elements j such that $\bar{\nu}_j > 0$.

The solution to the mathematical program (16) can be regarded as a first-order deterministic fluid approximation for the SBR system, as in [33]. From that point of view, given a selected solution $(\bar{\nu}, \bar{x})$, we would then use $\bar{\nu}$ to provide an initial estimate of the staffing and $\bar{x}$ to provide an initial estimate of the appropriate routing. For the staffing, that is precisely what we do, but the story for the routing is very different.

We start by explaining what happens with the staffing: With N_*^λ being the optimal staffing vector solution for (12) when the arrival rate is λ, we expect that $|c \cdot N_*^\lambda - c \cdot \bar{\nu}\lambda| = o(\lambda)$. This intuitive result will be rigorously established in Theorem 3.3. Actually, we will prove that $|c \cdot N_*^\lambda - c \cdot \nu\lambda| = O(\sqrt{\lambda})$. Toward that end, we define the following square-root-safety-staffing rule:

Definition 3.2 (**square-root-safety-staffing rule: SRSS**)

Given an optimal solution $(\bar{\nu}, \bar{x})$ to (16), we say that the system is staffed according to the square-root-safety-staffing rule $SRSS(\bar{\nu})$, if for $j \in \mathcal{J}$ with $\bar{\nu}_j > 0$, we set $N_j^\lambda = \lceil \bar{\nu}_j \lambda + \gamma_j \sqrt{\lambda} \rceil$ for some $\gamma_j \in (-\infty, \infty)$, and $N_j^\lambda = 0$ otherwise. We define $N_\Sigma^\lambda = \sum_{j \in \mathcal{J}} N_j^\lambda$.

We now indicate how we apply $\bar{x}$ from the chosen optimal solution $(\bar{\nu}, \bar{x})$ for the fundamental mathematical program (16). Since we intend to use FQR for the routing, we do not use $\bar{x}$ for the routing, but $\bar{x}$ plays a critical role in the design. Specifically, **we omit all edges with $\bar{\mathbf{x}}_{\mathbf{i,j}} = \mathbf{0}$ from the network routing graph**; i.e., we do not allow any class i customers to be routed to pool j if $\bar{x}_{i,j} = 0$. **The routing graph design includes all edges with $\bar{\mathbf{x}}_{\mathbf{i,j}} > \mathbf{0}$;** we stipulate that the routing graph is $\{(i,j) \in \mathcal{I} \times \mathcal{J} : \bar{x}_{i,j} > 0\}$. If, a priori, pool j is unable to serve class i or if we do not want pool j to serve class i, then we enforce that by imposing the constraint $x_{i,j} \leq 0$ in the mathematical program. When fixing a solution $(\bar{\nu}, \bar{x})$ for (16) we define $I(j)(\bar{\nu}, \bar{x}) = \{i \in \mathcal{I} : \bar{x}_{i,j} > 0\}$. We will often just use $I(j)$ when the solution $(\bar{\nu}, \bar{x})$ that is being used is clear from the context.

The system design, which is determined by $\bar{x}$, is in turn closely linked to the dynamic control. The characterization of that dependence is the main subject of this section. To facilitate the discussion, we should have a clear notion of the network graphs under consideration. Our network graphs are simple undirected bipartite graphs, i.e., with at most one edge connecting any two nodes, and with edges only between a customer class and an agent pool. Beyond this basic feature, the first important structural assumption we impose is the following:

Assumption 3.2 (**connected routing graph**) *The selected optimal solution $(\bar{\nu}, \bar{x})$ for (16) produces a routing graph determined by the edges $\mathcal{E}(\bar{\nu}, \bar{x}) := \{(i,j) \in \mathcal{I} \times \mathcal{J} : \bar{x}_{i,j} > 0\}$ that is a connected subgraph of $G(V, E)$.*

This connected-routing-graph assumption is assumed to hold throughout the rest of the paper. By saying that the graph is connected, we follow the common graph theory terminology. That is, a graph is connected if there exists a path between every two nodes in the graph. This connected-graph assumption is crucial for the ability to instantaneously balance the system; see Section 2.7 of Atar [6] for elaboration. We will actually need a finer characterization of the network graph beyond its connectedness. More specifically, connected graphs can be cyclic or acyclic (following common terminology - a graph is acyclic if there is a

unique path between each pair of nodes). This distinction will be important for our results, because FQR will work well with cyclic networks only when certain parametric conditions hold; see Theorem 3.1. For a more elaborate discussion we refer the reader to Remark 3.1 in [19].

3.2 Definition of FQR

In order to define FQR and state the necessary results, let $Z^{\lambda}_{i,j}(t)$ be the number of type-j servers busy giving service to class-i customers, so that $X^{\lambda}_i(t) := Q^{\lambda}_i(t) + \sum_{j\in\mathcal{J}} Z^{\lambda}_{i,j}$ is the overall number of class-i customers present in the system at time t, and $I^{\lambda}_j(t) = N^{\lambda}_j - \sum_{i=1}^{I} Z^{\lambda}_{i,j}(t)$ be the number of idle agents in pool j at time t in the λ^{th} system. Accordingly, $I^{\lambda}_{\Sigma}(t) := \sum_{j=1}^{J} I^{\lambda}_j(t)$ is the total number of idle agents in the system. Let $X^{\lambda}_{\Sigma}(t)$ be the overall number of customers in the system (in service and in queue), i.e.,

$$X^{\lambda}_{\Sigma}(t) := \sum_{i=1}^{I} X^{\lambda}_i(t) = \sum_{i=1}^{I} \left(Q^{\lambda}_i(t) + \sum_{j=1}^{J} Z^{\lambda}_{i,j}(t) \right) .$$

In work-conserving models, where idle agents and waiting customers do not co-exist, the identities $Q^{\lambda}_{\Sigma}(t) = [X^{\lambda}_{\Sigma}(t) - N^{\lambda}_{\Sigma}]^+$ and $I^{\lambda}_{\Sigma}(t) = [X^{\lambda}_{\Sigma}(t) - N^{\lambda}_{\Sigma}]^+$, hold. The V-model from §2 is a classical example of such a system. These identities need not hold in more complex systems, but we expect them to hold approximately for efficient systems; i.e., we expect $[X^{\lambda}_{\Sigma}(t) - N^{\lambda}_{\Sigma}]^+$ and $[X^{\lambda}_{\Sigma}(t) - N^{\lambda}_{\Sigma}]^-$ to be reasonable proxies for the overall queue length, $Q^{\lambda}_{\Sigma}(t)$, and the overall number of idle agents $I^{\lambda}_{\Sigma}(t)$. Our general version of FQR exploits the proxy for $I^{\lambda}_{\Sigma}(t)$ instead of its actual value. In doing so, we note that the intuitive reasoning and simplicity of FQR are preserved. In fact, the resulting state-space collapse can be equivalently given either in terms of $(Q^{\lambda}_{\Sigma}(t), I^{\lambda}_{\Sigma}(t))$ or $[X^{\lambda}_{\Sigma}(t) - N^{\lambda}_{\Sigma}]^+$ and $[X^{\lambda}_{\Sigma}(t) - N^{\lambda}_{\Sigma}]^-$, since FQR will guarantee that $[X^{\lambda}_{\Sigma}(t) - N^{\lambda}_{\Sigma}]^+ \approx Q^{\lambda}_{\Sigma}(t)$ and $[X^{\lambda}_{\Sigma}(t) - N^{\lambda}_{\Sigma}]^- \approx I^{\lambda}_{\Sigma}(t)$.

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 $\operatorname{argmax} f := \{y \in A : f(y) = \max_{x\in A} f(x)\}$.

Definition 3.3 (**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

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

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*

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

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

$$(Q^\lambda, Z^\lambda) := (Q_i^\lambda(t), Z_{i,j}^\lambda(t);\ i \in \mathcal{I}, j \in \mathcal{J})) \tag{18}$$

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 $i \in \mathcal{I}$, FQR is equivalently given by choosing upon service completion to serve the customer from the head of queue i^* where

$$i^* \equiv i^*(t) \in \underset{i \in I(j), Q_i^\lambda(t)>0}{\operatorname{argmax}} \left\{\frac{Q_i^\lambda(t)}{p_i}\right\}, \tag{19}$$

which makes the use of $[X_\Sigma^\lambda(t) - N_\Sigma^\lambda]^+$ unnecessary. Also, note that our definition of consistency here is somewhat different than it was in the case of the V model. The reason is that in a system with multiple customer classes, multiple agent pools and general service rates, it is not guaranteed that there exists a control under which the process (Q^λ, I^λ) will be a Markov process, since we will have to take account of the number of agents busy with each customer class to determine the transition rates. The process (Q^λ, I^λ) might still be a Markov process in an SBR system with a common service rate, like the one we analyze in this section, but we choose to make the more general definition for future use.

3.3 State-Space Collapse Under FQR

We next define the scaled and normalized processes of interest:

Definition 3.4 (**scaled and normalized processes using** $SRSS(\bar{\nu})$) *Fix an optimal solution* $(\bar{\nu}, \bar{x})$ *for* (16) *for which the edges in* $\mathcal{E}(\bar{\nu}, \bar{x})$ *induce a connected routing graph. Fix a staffing vector* N^λ *determined through* $SRSS(\bar{\nu})$*. Then, we define the following scaled processes:*

$$\hat{X}^\lambda_\Sigma(t) := \frac{X^\lambda_\Sigma(t) - N^\lambda_\Sigma}{\sqrt{\lambda}}; \quad \hat{I}^\lambda_\Sigma(t) := \frac{I^\lambda_\Sigma(t)}{\sqrt{\lambda}}; \quad \hat{X}^\lambda_i(t) := \frac{X^\lambda_i(t) - \sum_{j\in\mathcal{J}} \bar{x}_{i,j} N^\lambda_j}{\sqrt{\lambda}},\ i \in \mathcal{I};$$
$$\hat{Q}^\lambda_i(t) := \frac{Q^\lambda_i(t)}{\sqrt{\lambda}},\ i \in \mathcal{I}; \quad \hat{I}^\lambda_j(t) := \frac{I^\lambda_j(t)}{\sqrt{\lambda}},\ j \in \mathcal{J}; \quad \hat{Z}^\lambda_{i,j}(t) := \frac{Z^\lambda_{i,j}(t) - \bar{x}_{i,j} N^\lambda_j}{\sqrt{\lambda}},\ (i,j) \in \mathcal{I}\times\mathcal{J}.$$

Towards the presentation of the state-space collapse result, we let $D^d := D^d[0,\infty)$ be the space of all RCLL (Right Continuous with Left Limits) functions with values in d-dimensional Euclidean space $\mathbb{R}^d$, equipped with the Skorohod J_1 metric; e.g., see [32]. Below, we will denote by 0 the function in D^d that is identically 0. We will also consider a weaker notion of convergence, using the space $D^d_- := D^d(0,\infty)$, where the domain is treated as open at the left instead of closed. We again let convergence (to continuous limits) be characterized by uniform convergence over bounded intervals. The restriction to the domain $(0,\infty)$ means that we exclude uniform convergence for intervals of the form $[0,b]$. We have $Y^\lambda(t) \Rightarrow 0$ in $D^d(0,\infty)$ if and only if, for each $0 < s < T < \infty$, $\|Y^\lambda\|^*_{s,T} \Rightarrow 0$. Our state-space-collapse result under FQR is:

Theorem 3.1 (**state-space collapse under FQR for general SBR models**)

Fix an optimal solution $(\bar{\nu}, \bar{x})$ *for* (16) *for which the edges in* $\mathcal{E}(\bar{\nu}, \bar{x})$ *induce a connected routing graph. Fix the two probability vectors* p *and* v*. Let FQR and* $SRSS(\bar{\nu})$ *be used, following Definitions 3.3 and 3.2. Suppose that at least one of the following conditions holds with respect to* $(\bar{\nu}, \bar{x})$*:*

- ***C- 1*** **Only one pool has cross-trained agents:** *There exists at most one* $j \in \mathcal{J}$ *with skill set* $I(j)(\bar{\nu}, \bar{x})$ *containing more than one element; denote this pool by* j^**. Also, we require that* $v = e_{j^*}$*.*
- ***C- 2*** **The service rates depend only on the agent type:** *For all* $(i,j) \in \mathcal{E}(\bar{\nu}, \bar{x})$*,* $\mu_{i,j} = \mu_j$*.*
- ***C- 3*** **The service rates depend only on the customer class:** *For all* $(i,j) \in \mathcal{E}(\bar{\nu}, \bar{x})$*,* $\mu_{i,j} = \mu_i$*.*

If, in addition, $(\hat{X}^\lambda(0), \hat{Z}^\lambda(0)) \Rightarrow (\hat{X}(0), \hat{Z}^\lambda(0))$ *in* $\mathbb{R}^{I+I\cdot J}$*, then we have state-space collapse:*

$$\hat{Q}_i^\lambda(t) - p_i \hat{Q}_\Sigma^\lambda(t) \Rightarrow 0 \quad \textit{in} \quad D_- \quad \textit{as} \quad \lambda \to \infty, \quad i \in \mathcal{I}, \tag{20}$$

and

$$\hat{I}_j^\lambda(t) - v_j \hat{I}_\Sigma^\lambda(t) \Rightarrow 0 \quad \textit{in} \quad D_- \quad \textit{as} \quad \lambda \to \infty, \quad j \in \mathcal{J}\ . \tag{21}$$

The convergence is replaced with convergence in D *if we assume, in addition, that*

$$\hat{Q}_i^\lambda(0) - p_i \hat{Q}_\Sigma^\lambda(0) \Rightarrow 0,\ i \in \mathcal{I},\ \textit{and}\ \hat{I}_j^\lambda(0) - v_j \hat{I}_\Sigma^\lambda(0) \Rightarrow 0,\ j \in \mathcal{J}.$$

Finally, if condition C-1 holds, then,

$$\frac{1}{\sqrt{\lambda}} \hat{Z}_{i,j}^\lambda(t) \Rightarrow 0 \quad \textit{in} \quad D \quad \textit{as} \quad \lambda \to \infty, \quad i \in \mathcal{I}, \quad j \in \mathcal{J}. \tag{22}$$

If both conditions C-2 and C-3 fail to hold, state-space collapse is guaranteed under FQR only for acyclic networks that satisfy condition C-1. Primary examples of such networks are given in Figure 2. Atar constructs a control, which we denote by GFQR, that achieves the same state-space collapse when generalizing condition C-1 to the weaker requirement of a tree structure. The price of the applicability of GFQR to arbitrary tree structures is the complexity of the control when compared to FQR. The definition of GFQR depends heavily on the precise tree structure through the solution of a specific linear program. We refer the reader to §4.2 of [19] for a precise definition of the control and only state here the corresponding state-space collapse result for completeness.

Theorem 3.2 (**state-space collapse under GFQR**)

Fix two probability vectors p *and* v *and an optimal solution* $(\bar{\nu}, \bar{x})$ *for (16). Suppose that GFQR and* $SRSS(\nu)$ *are used, and that the following holds with respect to* $(\bar{\nu}, \bar{x})$*:*

- ***C- 4*** *Under* $(\bar{\nu}, \bar{x})$*, the graph induced by the edges* $\mathcal{E}(\bar{\nu}, \bar{x})$ *is a tree.*

If, in addition $(\hat{X}^\lambda(0), \hat{Z}^\lambda(0)) \Rightarrow (\hat{X}(0), \hat{Z}(0))$ *in* $\mathbb{R}^{I+I\cdot J}$*, then the conclusions of Theorem* 3.1 *hold and*

$$\frac{1}{\sqrt{\lambda}} \hat{Z}_{i,j}^\lambda(t) \Rightarrow 0 \quad \textit{in} \quad D \quad \textit{as} \quad \lambda \to \infty, \quad i \in \mathcal{I}, \quad j \in \mathcal{J}. \tag{23}$$

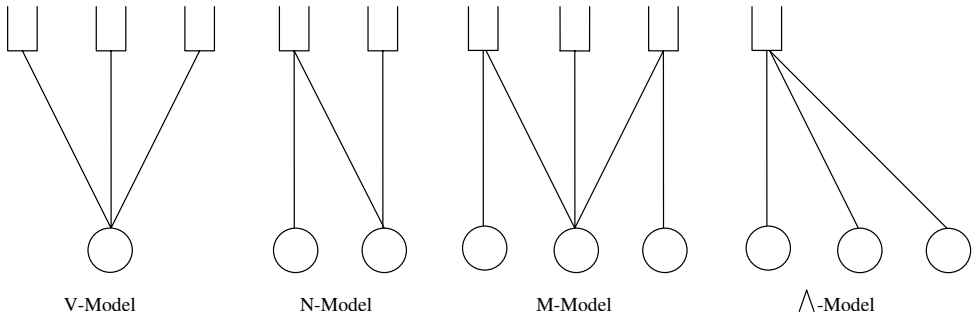

Figure 2: The V, N, M and $\bigwedge$ Models

Remark 3.3 (**decomposition of excluded models**) At first, it might seem that FQR and GFQR combined should cover arbitrary network structures, but that is not the case. With arbitrary service rates, when neither of the conditions 2 or 3 holds, some rather basic models - like the X model depicted in the left hand side of *Figure 3)* - are ruled out. This is not a merely a limitation of our proof technique, but a true limitation of these controls. Simulations (initially conducted by Ohad Perry) show that there exist parameter combinations such that the X model with FQR will lead the system to explode.

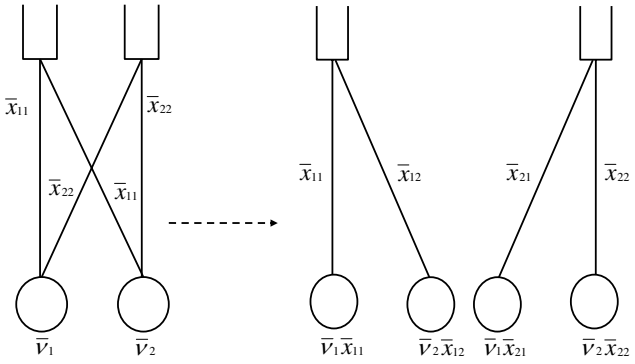

Figure 3: The X model

While the X model itself does not satisfy the conditions of FQR, we can decompose the original X model to construct an alternative model with two separate components, such that the overall model still has the same first order cost $c_1\bar{\nu_1} + c_2\bar{\nu}_2$. Even though the original model does not satisfy Assumption 3.2, each of the components of the new model do satisfy the conditions of FQR. To be concrete, assume that $(\bar{x}, \bar{\nu})$

is the optimal solution of (16) that induces the X model, and consider the following alternative: assume in addition to agent type 1 and 2 we add to the model agents types 1_1 and 1_2, with agent type $1_i,\ i = 1, 2$ being able to serve only type i customers. Costs and rates remain the same, so that $c_{1_i} = c_1,\ i = 1, 2$. Similarly, we define agents types $2_i,\ i = 1, 2$. We re-solve (16) with the additional agents pools. Then, an optimal solution to the new problem (although not the only one) would be given (in terms of the original optimal solution) by $\bar{\nu}_{j^i} = \bar{\nu}_j \bar{x}_{i,j},\ i = 1, 2,\ j = 1, 2$, suggesting the pair of inverted V models on the right hand side of Figure 3.

Thus we have obtained two models, the initial X model and an alternative pair of inverted V systems, with essentially different structures but the same associated deterministic costs. Moreover, FQR is applicable in each component of the new model. However, this, by no means, makes FQR more general, because we still can not apply FQR directly to the X model. But it does have important implications: If FQR is important, then we can find a way to apply it by changing the system design. Care is needed, however, because changing the design in order to be able to apply FQR may come at the price of diminished system flexibility, as is indeed the case in this simple X-model example.

What are the practical implications then? Clearly, when neither of the conditions 2 or 3 holds, FQR should not be used in the X model. We thus may elect to change the design. However, while we can evaluate the performance in the decomposed system with FQR, in the absence of a reasonable theory for the X model in the QED asymptotic regime, we do not have a control for the X model to use as a benchmark for comparison to the decomposed system. In particular we can not evaluate the price of the diminished flexibility. Flexibility, seems especially important when arrival-rate estimates may be inaccurate. However, in the presence of reliable arrival-rate estimates, it seems reasonable to use the decomposed system with the simple FQR control. We have this shown that it may be possible to expand the domain of applicability of FQR. ∎

Recall that we use optimization problem (12) for asymptotic feasibility and optimization problem (13) for asymptotic optimality. Focusing on formulation (12), we have the following result, where we fix p^* to be as defined by $p_i^* = \lambda_i T_i^\lambda / \sum_{i \in \mathcal{I}} \lambda_i T_i^\lambda$. Also, $\gamma_j,\ j \in \mathcal{J}$ are the coefficients of the square-root terms in the definition of SRSS; see Definition 3.2.

Theorem 3.3 (**feasibility of a SRSS($\bar{\nu}$)**) *Consider a sequence of SBR systems staffed with SRSS($\bar{\nu}$), and using $FQR(p^*, v)$ with any probability vector v. Then, there exists $\beta > 0$ such that, if $\sum_{j \in \mathcal{J}} \gamma_j = \beta$, all*

constraints in (12) hold asymptotically, i.e, for each $\epsilon > 0$, there exists $T^*(\epsilon)$ such that, for all $T \geq T^*(\epsilon)$,

$$\limsup_{\lambda \to \infty} P\left\{ \bar{F}_i^{\lambda,T}(T_i^\lambda) \geq \alpha + \epsilon \right\} \leq \epsilon, \quad i \in \mathcal{I}, \tag{24}$$

In particular, determining N^λ through $SRSS(\bar{\nu})$ with $\sum_{j \in \mathcal{J}} \gamma_j = \beta$, we have that

$$[c \cdot N^\lambda - c \cdot N_*^\lambda]^+ = O(\sqrt{\lambda}),$$

where N_*^λ is an optimal staffing vector for (12).

Remark 3.4 (**steady-state quantities and alternatives**) Our notion of asymptotic feasibility is embedded in Theorem 3.3 above. In particular, the limits in (24) imply that we regard a combination of design, staffing and control as asymptotically feasible if for any $\epsilon > 0$ there exists a time $t^*(\epsilon)$ after which we are asymptotically ϵ away from feasibility. Although mathematically different, this is practically the same as the steady-state asymptotic feasibility notion used in §2. To see this, note that for fixed λ requiring that $P\{W_i^\lambda(\infty) > T_i\}$ be less than α is like requiring the existence of a function $T^*(\epsilon)$ as above. Still, although the notion above and the one used in §2 should be regarded as the same for all practical purposes, they are mathematically different and a limit-interchange argument is required to connect the two. Specifically, in the presence of an interchange argument the two notions will be equivalent. The limit-interchange argument for the V-Model analyzed in the previous section is quite simple, but that is not the case for more general SBR models, so that the ability to interchange the limits remains an open problem. For an example of techniques that may apply, see Gamarnik and Zeevi [16], Budhiraja and Lee [13], and Gurvich and Zeevi [21]. Here we settle for the weaker notion of feasibility expressed in Theorem 3.3 and in Theorem 4.2 below. ∎

Theorem 3.3 implies that one can search for the feasible β through simulation while keeping the vector p fixed and pre-determined through the SL targets independently of the resulting β or the network structure. This observation can simplify significantly the search of practical solutions to the general SBR problem. We take that approach in the subsequent paper [15] . There are some cases, however, when more can be said and we can solve, in an asymptotically optimal way, the best-effort formulation introduced in §2, but now for more general SBR models with multiple agents types. Such is the case of pool-dependent service rates of which a specific case is the case of equal service rates. We address these two cases in §4 below. First, however, we end this section with an example.

Example 3.1 (optimization through simulation in a two-class model) We now present an example that emphasizes three important points: First, even though the initial network configuration may not satisfy the conditions for FQR to be effective, there may exist an optimal solution to the fundamental mathematical program that has the desired properties. Second, state-space collapse is clearly evident in systems of medium size. Finally, we will use this example to illustrate Theorem 3.3.

Toward these ends, consider a system with two customer classes, $\mathcal{I} = \{1, 2\}$, and three possible agents types, $\mathcal{J} = \{1, 2, 3\}$. Let the arrival rates be $\lambda_1 = \lambda_2 = 200$. Assume there is no abandonment. Agents of type 1 can serve both class-1 and class-2 customers. They serve class-1 customers at rate $\mu_{11} = 1$ and class-2 customers at rate $\mu_{21} = 3$. Agents of type 2 can only serve class 2, and do so with rate $\mu_{22} = 3$. Agents of type 3 are cross-trained agents and can give service to both classes, and they do so with rates $\mu_{13} = 2$ and $\mu_{23} = 3$.

Suppose that system managers have imposed an extra constraint requiring that at least half of class-1 customers are served by the type-1 agents. This is imposed by adding the constraint $\mu_{11}N_1 \geq \lambda_1/2$. This constraint comes at a price, because these agents are more expensive. In particular, it is assumed that $c_1 = 22$ dollars/hour, while $c_2 = c_3 = 8$ dollars/hour.

Before solving the mathematical program in (16), it looks like this system is not covered by our theorems, as certain staffing decisions may lead to a cyclic system with service rates that are both class-dependent and pool-dependent. Solving the mathematical program (16) reveals, however, a different story. Since $c_2 = c_3$, the mathematical program (16) has multiple optimal solutions. For example, $(\bar{\nu}_1, \bar{\nu}_2, \bar{\nu}_3) = (1/4, 3/40, 13/60)$ is one optimal solution for (16), which does not satisfy the condition for state-space collapse under FQR as given in Theorem 3.1. Fortunately, however, there exists an optimal solution that does satisfy these conditions. Specifically, a better optimal solution for us is $(\bar{\nu}_1, \bar{\nu}_2, \bar{\nu}_3) = (1/4, 0, 7/24)$ with corresponding $\bar{x}$ values given by $\bar{x}_{1,1} = 1$, $\bar{x}_{1,3} = 3/7$, $\bar{x}_{2,3} = 4/7$, and $\bar{x}_{i,j} = 0$ otherwise. This solution translates to an N model (see Figure 2). We staff this N model using $SRSS(\bar{\nu})$ with $\bar{\nu} = (1/4, 0, 7/24)$ and $\gamma = (0, 0, 0)$. We see, then, that by a judicious choice of the solution to the mathematical program, we obtain a routing graph that satisfies our theorem, and in particular one that allows us to use the simple FQR control in order to achieve state-space collapse.

In Example 3.1 in [19] we simulate a single realization of this same N model under FQR for given ratio vectors p and v. The simulation illustrates how strongly state-space collapse holds even for the medium-size system in consideration; see Figure 3 in [19].

We now illustrate Theorem 3.3 through the same N-model. We use the SL targets $T_1 = T_2 = 0.2$ (corresponding to $\bar{T}_1 = \bar{T}_2 = 4$ in Assumption 2.1) and the SL probability 0.2. The corresponding SL constraints are then $\limsup_{T\to\infty} F_i^T(0.2) \leq 0.2$ for $i = 1, 2$. We use FQR with the ratio vector $p = (0.5, 0, 5)$ as recommended in Theorem 3.3. Also, we fix $v = (0.5, 0, 5)$ which implies that the idleness will be divided evenly between the two pools. To construct the SRSS staffing, we let $\bar{\nu} = (1/4, 0, 7/24)$ and $N_1 = 0.25\lambda$ (corresponding to $\gamma_1 = 0$), and find γ_2 such that with $N_2 = 7/24\lambda + \gamma_2\sqrt{\lambda}$ the SL constraints will be satisfied. Specifically, we use simulation to to search over values of $N_2 \geq 7/24\lambda$, while keeping the ratios p and v fixed. We find that $N_2 = 120$ suffices, see the right-hand graph in Figure (4), in which the probability of violating the SL target, 0.2, at time t, $P\{W(t) > T\}$ is plotted as a function of time for each of the customer classes ($W(t)$ should be interpreted as the virtual waiting time at time t). We observe that, at least for $\lambda = 400$, $N_2 = 120$ corresponds to $\beta = \gamma_2 = 1/6$.

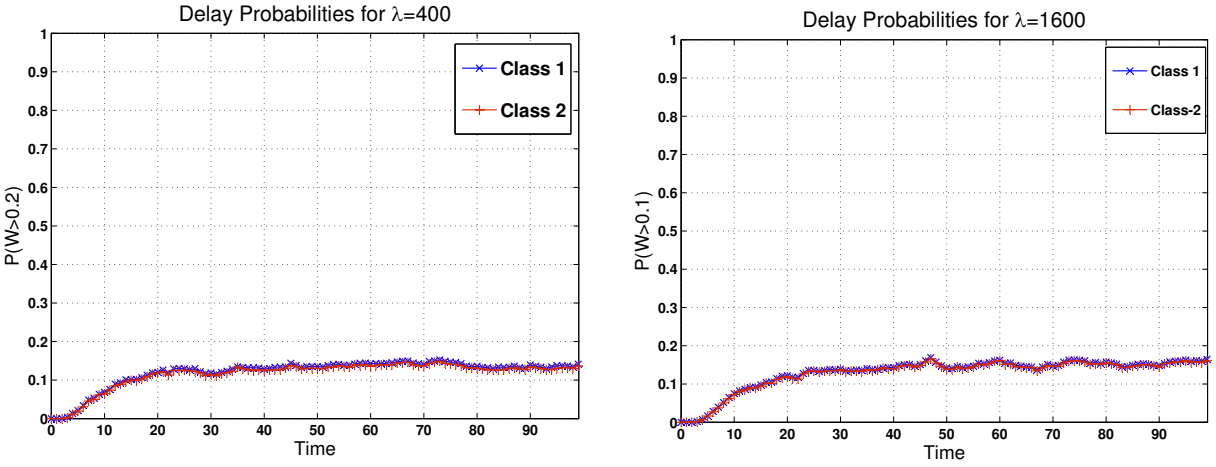

Figure 4: Delay probabilities under FQR for $\lambda = 400$ and $\lambda = 1600$

Theorem 3.3 claims, however, that β should be invariant to λ for all λ large enough. To show this, we consider next the same N model but now with arrival rates $\lambda_1 = \lambda_2 = 800$, so that $\lambda = 1600$, and using the targets $T_i^\lambda = \bar{T}_1/\sqrt{\lambda} = 4/\sqrt{1600} = 0.1$. By Theorem 3.3, we should be able to use the same value of β for this larger system. Accordingly, we set $N_1 = 0.25 * 1600 = 400$ and $N_2 = \lfloor 7/24 * \lambda + 1/6\sqrt{\lambda} \rfloor = 473$. The right-hand graph in Figure (4) shows that the constraints are indeed satisfied. ∎

4 Pool-Dependent Service Rates

In this section we show that the aggregate approach to staffing, as applied in §2 to the simplified V-model setting, can be extended to more general SBR models, as long as the service rates are pool-dependent, i.e., if $\mu_{i,j} = \mu_j$ for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$. Throughout we will assume, without loss of generality, that the service rates are labelled in decreasing order:

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

We further restrict the general model of §3 by assuming that customers do not abandon; we discuss how to partially remove this assumption in Remark 4.2. All other assumptions from §3 remain unchanged. We consider the two formulations (12) and (13). Following the discussion in §2, we establish asymptotic optimality only for the best-effort formulation (13), while restricting the discussion of the initial optimization problem (12) to asymptotic feasibility.

asymptotic optimality. Our solution in §2 was based on a reduction of the V model to the associated $M/M/N$ queue. Such a reduction is no longer possible now, but again our solution will be based on a reduction of the SBR system to a simpler model, namely, the inverted-V (or $\bigwedge$) model, in which multiple agent types serve a single customer class. Specifically, given an SBR system, the associated $\bigwedge$ model is one with the same set of agent-pools $\mathcal{J}$, the same staffing levels $\{N_j,\ j \in \mathcal{J}\}$ and the same service rates $\{\mu_j,\ j \in \mathcal{J}\}$, but with all the I customer classes merged into a single super-class with arrival rate $\lambda = \sum_{i \in \mathcal{I}} \lambda_i$. An example of an SBR system and its corresponding $\bigwedge$ model is given in Figure 5.

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 [2]. We will exploit the results in Armony [2] here. We begin with the best-effort formulation (12), and establish a lower bound on the feasible staffing vectors.

Theorem 4.1 (**lower-bound capacity**) *Fix a sequence of staffing vectors* $\{N^\lambda\}$ *that satisfy*

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

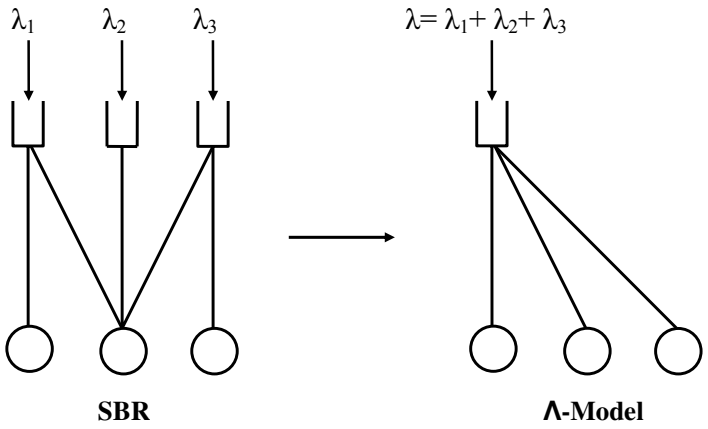

Figure 5: An SBR model and its corresponding $\bigwedge$ model

Then,

$$\limsup_{\lambda\to\infty} P\left\{\frac{\bar{W}^{\lambda,T}}{T_I^\lambda} \geq 1+\epsilon\right\} \leq \epsilon \tag{25}$$

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

$$\sum_{j\in\mathcal{J}} \mu_j N_j^\lambda \geq \lambda + \beta^*\sqrt{\lambda} + o(\sqrt{\lambda}), \tag{26}$$

where β^* *is the unique solution to*

$$\frac{\mathbf{P}_{\mu_1}(\beta)}{\beta} = \sqrt{\lambda}T_I^\lambda =: \bar{T}_I, \tag{27}$$

and

$$\mathbf{P}_{\mu_1}(\beta) := \left[1 + \frac{\frac{\beta}{\sqrt{\mu_1}}\Phi\left(\frac{\beta}{\sqrt{\mu_1}}\right)}{\phi\left(\frac{\beta}{\sqrt{\mu_1}}\right)}\right]^{-1}.$$

Theorem 4.1 is a consequence of a feasibility result for the $\bigwedge$ model and a stochastic-order result comparing the SBR model to the $\bigwedge$ model. 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. However, building on [2], we will show that a staffing vector that is feasible for the $\bigwedge$ model in terms of the aggregate-waiting-time constraint must satisfy (26). This in turn implies the result for the SBR system. The

detailed proof appears in the appendix.

We now turn to the asymptotic optimality. We first observe that any staffing solution to (13) must also satisfy (14). Theorem 4.1 implies that an asymptotic lower bound on the optimal cost in (13) is given by the solution of the following mathematical program:

$$\begin{array}{ll} \text{Minimize} & \sum_{j\in\mathcal{J}} c_j N_j \\ \text{Subject to:} & \sum_{j\in J} \mu_j N_j \geq \lambda + \beta^* \sqrt{\lambda}, \\ & \sum_{j\in J} \mu_j N_j y_{i,j} \geq \lambda_i, \quad i \in \mathcal{I}, \\ & \sum_{i\in\mathcal{I}} y_{i,j} \leq 1, \quad j \in \mathcal{J}, \\ & N \in \tilde{\mathcal{A}}^\lambda, \quad y_{i,j} \geq 0, \quad i \in \mathcal{I}, \quad j \in \mathcal{J}, \end{array} \tag{28}$$

We observe that (28) is equivalent to (14) if one replaces β^* with zero. Consequently, as (28) adds only a $O(\sqrt{\lambda})$ terms to solutions of (14), the intuitive "fluid" counterpart for both math programs is identical and is given by (16).

We denote a staffing vector determined through the solution of (28) as a **$\bigwedge$-based staffing**. We will show that using the $\bigwedge$-based staffing is asymptotically optimal as long as it is used in combination with FQR with the ratio vectors (p^* and v^*), where p^* is the unique solution to

$$p_{I-1} = \mathbf{P}_{\mu_1}(\beta^*) e^{-\frac{\lambda_{I-1}}{\lambda p_{I-1}} \beta^* \sqrt{\lambda} T^\lambda_{I-1}} = \alpha \quad \text{and} \quad \frac{p_i}{p_{I-1}} = \frac{\lambda_i T^\lambda_i}{\lambda_{I-1} T^\lambda_{I-1}}, \tag{29}$$

and $v^* = (0, 0, \ldots, 1)$.

Since $\sqrt{\lambda} T^\lambda_I = \bar{T}_i$ (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 $i \neq I-1$. We remark that the construction of p is more subtle than in (9), which was based on the reduction of the V model to the $M/M/N$ queue. The current construction is based on asymptotic expressions for the delay probability in the $\bigwedge$ model under the control proposed in [2]; see Lemma B.2 in the appendix as well as Propositions 4.4 and 4.6 in [2]. Still, the two constructions are consistent as they are asymptotically equivalent under the assumption of equal service rates; see §4.1.

Here is the asymptotic optimality result for this section:

Theorem 4.2 (**asymptotic optimality for the SBR model with pool-dependent rates**) *Suppose that any optimal solution for (16) has $\bar{\nu_j} > 0$ for all $j \in \mathcal{J}$. Suppose that N^λ is determined through the $\bigwedge$-based*

staffing and $FQR(p^*, v^*)$ is used, where p^* is as in (29) *and $v^* = (0, \cdots, 0, 1)$. Then, for each $\epsilon > 0$, there exists $T^*(\epsilon)$ such that, for all $T \geq T^*(\epsilon)$,*

$$\limsup_{\lambda \to \infty} P\left\{ \bar{F}_i^{\lambda,T}(T_i^\lambda) \geq \alpha + \epsilon \right\} \leq \epsilon, \quad 1 \leq i \leq I-1, \tag{30}$$

$$\limsup_{\lambda \to \infty} P\left\{ \frac{\bar{W}^{\lambda,T}}{T_I^\lambda} \geq 1 + \epsilon \right\} \leq \epsilon. \tag{31}$$

Finally, any other sequence $(\tilde{N}^\lambda, \pi^\lambda)$ of staffing vectors and routing rules that satisfy (30) *and* (31)*, must satisfy,*

$$[c \cdot N^\lambda - c \cdot \tilde{N}^\lambda]^+ = o(\sqrt{\lambda}) \quad \textit{as} \quad \lambda \to \infty\,. \tag{32}$$

Remark 4.1 (**choosing the ratio vector** v^*) In light of our state-space collapse result Theorem 3.1, 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. As 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 so as to guarantee the service-level constraints. The queue-ratio vector, p^*, takes care of this task. ∎

Theorem 4.2 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 inverted-V model sufficient also for the SBR system.

asymptotic feasibility. We now establish asymptotic feasibility for the initial SBR optimization problem (12). We re-define p^* in accordance with (3) to be

$$p_i^* = \frac{\lambda_i T_i^\lambda}{\sum_{i \in \mathcal{I}} \lambda_i T_i^\lambda},$$

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

$$\mathbf{P}_{\mu_1}(\beta)e^{-\beta\sqrt{\lambda}\sum_{i\in\mathcal{I}}\frac{\lambda_i}{\lambda}T_i^\lambda} = \alpha. \tag{33}$$

As before, we observe that the vector p^* is independent of λ, since $\lambda_i/\lambda = a_i$ and Assumption 2.1 holds.

Theorem 4.3 (**asymptotic feasibility for the SBR model with pool-dependent rates**) *Suppose that any optimal solution for (16) has $\bar{\nu_j} > 0$ for all $j \in \mathcal{J}$. Suppose that the $\bigwedge$-based staffing is used as well as $FQR(p^*, v^*)$, where p^* and v^* are as in* 4.2*. Then, for each $\epsilon > 0$, there exists $T^*(\epsilon)$ such that, for all $T \geq T^*(\epsilon)$,*

$$\limsup_{\lambda\to\infty} P\left\{\bar{F}_i^{\lambda,T}(T_i^\lambda) \geq \alpha + \epsilon\right\} \leq \epsilon, \quad i \in \mathcal{I}. \tag{34}$$

Remark 4.2 (**adding customer abandonment**) When there is customer abandonment, it is natural to consider quality-of-service constraints that limit the number of customer abandonments, in addition to the quality-of-service constraints we have used so far. However, assuming that the same formulations are maintained, the asymptotic feasibility and optimality results of Theorems 4.2 and 4.3 hold partially in an extension of this SBR model where we allow customer abandonment, provided that the times to abandon are i.i.d. and have an exponential distribution with a common rate θ. However, this extension is only partial, because asymptotic optimality can be established only within the set of work-conserving policies; i.e., policies, that do not idle an agent on purpose when there are customers waiting in any of the queues this agent is eligible to serve. With this restriction, the results of Theorems 4.2 and 4.3 continue to hold after the asymptotic expressions for the $\bigwedge$ model are replaced with those for a $\bigwedge$ model with abandonment. These expressions can be obtained from Armony and Mandelbaum [3]. If all service rates are equal, i.e, $\mu_j \equiv \mu,\ j \in \mathcal{J}$, then more can be said; see Remark 4.3. ∎

adding 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.

One can impose a lower-bound constraint on the proportion of customers that receive their designated service. For that purpose, we let $DS_i[T]$ 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. Then, letting $\delta_i > 0$ be the non-designated-service proportion upper bound, the constraints are formulated as

$$\liminf_{T\to\infty} DS_i^{\pi}[T] \geq 1 - \delta_i, \quad i \in \mathcal{I}$$

These constraints can be incorporated in the framework of (12) by choosing A and $\hat{b}$ so that

$$\left\{ N \in \mathbb{Z}_+^J : N_{j(\{i\})} \geq (1-\delta)\frac{\lambda_i}{\mu},\ i \in \mathcal{I} \right\} \subseteq \mathcal{A}^{\lambda}, \tag{35}$$

where $\delta = \min_{i\in\mathcal{I}} \delta_i$. This definition of the set $\mathcal{A}^{\lambda}$ is guaranteed to achieve the designated service constraints asymptotically. This is proved in the following Proposition

Proposition 4.1 *Suppose that the set $\mathcal{A}^{\lambda}$ satisfies (35). Then, under the conditions of Theorem 4.2 (respectively 4.3), equations (30)-(32) (respectively (34)) hold and, in addition, for any $\epsilon > 0$ there exists $T^*(\epsilon)$ such that for all $T > T^*(\epsilon)$,*

$$\liminf_{\lambda\to\infty} P\left\{ DS_i^{\lambda}[T] \leq 1 - \delta_i - \epsilon \right\} \leq \epsilon, \quad i \in \mathcal{I}. \tag{36}$$

4.1 One Common Service Rate

When all service rates are equal, i.e., when $\mu_{i,j} \equiv \mu$ for all i and j, the staffing solution is simplified even further. In particular, the $\bigwedge$-based staffing and the choice of the ratio vector p^* are asymptotically equivalent to $\widetilde{SCS}$ and $\widetilde{FQR}$ as defined in equations (8) and (9). The equivalence is in the sense that if N^{λ} is the

sequence of vectors obtained through the $\bigwedge$-based staffing, then

$$\frac{\sum_{j\in\mathcal{J}} N_j^\lambda - \bar{N}_\Sigma^\lambda}{\sqrt{\lambda}} \to 0, \text{ as } \lambda \to \infty,$$

where $\bar{N}_\Sigma^\lambda$ is obtained through $\widetilde{SCS}$. This conclusion is deduced from the following lemma:

Lemma 4.1 (**square-root 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 SCS - see equation (2) (respectively $\widetilde{SCS}$ - see equation (8)), then*

$$\frac{\mu\bar{N}_\Sigma^\lambda - \lambda}{\sqrt{\lambda}} \to \beta^* > 0, \text{ as } \lambda \to \infty, \tag{37}$$

where β^* is the unique solution of

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

respectively of

$$\mathbf{P}_\mu(\beta^*)\, \frac{1}{\beta^*} = \bar{T}_I.$$

Setting $\beta^{**} := \beta^*/\sqrt{\mu}$, the result of the lemma can also be interpreted as

$$\bar{N}_\Sigma^\lambda = \frac{\lambda}{\mu} + \beta^{**}\sqrt{\frac{\lambda}{\mu}} + o\left(\sqrt{\frac{\lambda}{\mu}}\right),$$

which is the widely known form of the square-root safety staffing rule; see Borst et. al. [11]. We now claim that, in the case of equal service rates, the two definitions β^* in (33) and (27) are equivalent to those in Lemma 4.1. Indeed, this follows from the definitions $T_i^\lambda = \bar{T}_i/\sqrt{\lambda}$ and $\lambda_i/\lambda = a_i$. A similar observation shows that the definition of the vector p in (29) is equivalent to that under $\widetilde{FQR}$. Consequently, in the case of equal service rates, one can use the non-asymptotic expressions in (8) and (9) to determine the staffing level and the ratio vector p. In particular, instead of using the constraint $\sum_{j\in\mathcal{J}} \mu_j N_j \geq \lambda + \beta^*\sqrt{\lambda}$, to bound

the aggregate capacity one can use the constraint

$$\sum_{j\in\mathcal{J}}\mu_j N_j = \mu\sum_{j\in\mathcal{J}} N_j^{\lambda} \geq \bar{N}_{\Sigma}^{\lambda},$$

where $\bar{N}_{\Sigma}^{\lambda}$ is as defined by (2) in the case of formulation (12) and by (8) in the case of formulation (13).

Remark 4.3 (**adding customer abandonment**) In this model, with equal service rates, the addition of abandonments with common abandonment rate θ is quite straightforward. Specifically, we replace $Q_{\lambda,\mu}^{FCFS}(N)$ in the definition of SCS and $\widetilde{SCS}$ as well as in the definition of p in (9) by $Q_{\lambda,\mu,\theta}^{FCFS}(N)$, where $Q_{\lambda,\mu,\theta}^{FCFS}(N)$ is the steady-state queue length in an $M/M/N+M$ system ($+M$ means Markovian abandonment) with arrival rate λ, service rate μ and abandonment rate θ. For example, we replace (2) with

$$\bar{N}_{\Sigma} = min\left\{N\in\mathbb{Z}_+ : P\left\{Q_{\lambda,\mu,\theta}^{FCFS}(N) > \sum_{i\in\mathcal{I}}\lambda_i T_i\right\}\leq\alpha\right\}. \tag{38}$$

■

4.2 The Setting of Wallace and Whitt

This section is dedicated to the setting of Wallace and Whitt [30], i.e, a setting with a common service rate μ and equal costs for all agents, i.e, $c_j \equiv c,\ j\in\mathcal{J}$. The discussion of this simplified model has three purposes: first, to contrast between the FQR-based solution and the simulation based approach of [30]; second, to illustrate an explicit construction of a system design when costs or system constraints do not pose significant restrictions, and, third, to illustrate the diminishing-return property of flexibility.

Towards these ends, we start by specifying the optimization problem. We will consider the problem

$$\begin{array}{ll}
\text{minimize} & \sum_{j\in\mathcal{J}} N_j \\
\text{subject to:} & \limsup_{T\to\infty} \bar{W}^{\lambda,T} \leq T_I^{\lambda} \\
& \limsup_{T\to\infty} \bar{F}_i^{\lambda,T}(T_i) \leq \alpha, \quad 1\leq i\leq I-1, \\
& \liminf_{T\to\infty} DS_i^{\pi}[T] \geq 1-\delta_i, \quad i\in\mathcal{I}, \\
& N\in\mathbb{Z}_+^J, \quad \pi\in\Pi.
\end{array} \tag{39}$$

The optimization problem (39) is obtained from (13) by adding a designated-service constraint and removing

any system constraints.

The specific design we suggest here is based on a concatenation of M systems, which we call the **generalized M (GM) model**. An example for a system with three customer classes is depicted in Figure 6. The generalized M 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 M model is relatively inexpensive in terms of cross-training, since it uses agents with at most two skills.

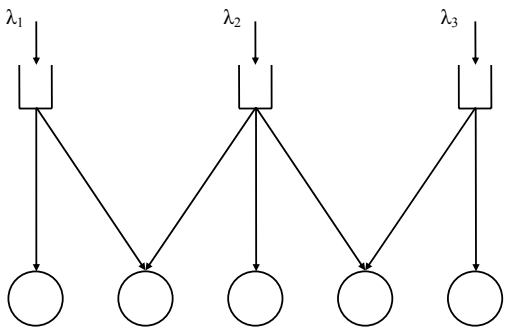

Figure 6: The generalized M model for 3 classes

The solution we propose for (39) 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 the Single-Class Staffing (SCS) rule given by (2). Then allocate agents to the pools by
  - $N_{j(\{i\})} = (1 - \delta/2)\left(\frac{\lambda_i}{\lambda}\right)\bar{N}_\Sigma, \quad i \in \{1, I\},$
    $N_{j(\{i\})} = (1 - \delta)\left(\frac{\lambda_i}{\lambda}\right)\bar{N}_\Sigma \quad \text{for all} \quad i = 2, \ldots, I-1, \quad \text{and}$
  - $N_{j(\{i,i+1\})} = \left(\frac{\delta(\lambda_i+\lambda_{i+1})}{2\lambda}\right)\bar{N}_\Sigma \quad \text{for all} \quad i = 1, \ldots, I-1,$

  where $\delta := \min\{\delta_i : 1 \le i \le I\} > 0$.
- **Control: Fixed-Queue-Ratio (FQR).** Use FQR with p as defined in (9) and v defined by

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

We could have used any vector v in the control step above. The specific vector we suggested 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 and §4.1, the above combined design-staffing-and-control solution can be easily shown to be asymptotically optimal as the arrival rate grows.

This simple model illustrates, then, 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. FQR 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, as the design and staffing decisions are often made in advance and can not be easily adjusted in real-time to accommodate changing environment. This also stands in contrast to Wallace and Whitt [30] where the realized service-levels are sensitive to the number of agents in each pool, and these need to be fine-tuned through simulation to meet that SL constraints.

We end this section by observing that the GM design uses only limited flexibility. In particular, it uses agents that have at most two skills. Still, with this limited amount of flexibility the SBR system performs, asymptotically, as efficiently as the V model from §2 in which all customers have access to all agents.

5 The Waiting-Time-Based Control - FWR

In this section we introduce a modification of FQR that uses waiting-time information instead of queue-length information. Such a control is relevant only when the system manager has access to the information regarding the accumulated waiting time of the customers waiting at the head-of-the-line in each queue. FWR will be identical to FQR in terms of the action upon customer arrival and differ only in terms of the action upon service completion. To contrast this control with FQR, we denote it by Fixed-Waiting-Ratio (FWR). To define FWR we let $W_{h,i}(t)$ be the accumulated waiting time of the customer waiting at the head of the class-i queue at time t. The definition of FWR is inspired by Little's law ($L = \lambda W$). It replaces the queue length $Q_i(t)$ in FQR with its approximation $\lambda W_{h,i}(t)$. Specifically, we define FWR as follows:

Definition 5.1 (**FQR for the SBR model**)

Given two probability vectors $v := \{v_j : j \in \mathcal{J}\}$ and $p := \{p_i : i \in \mathcal{I}\}$, FWR 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*

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

  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*

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

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

Ties are broken by choosing an agent pool with the largest index among the classes that maximize $I_j^\lambda(t) - v_j[X_\Sigma^\lambda(t) - N_\Sigma^\lambda]^-$ and choosing a customer from the class i with the largest index among the classes that maximize $W_{h,i}^\lambda(t) - p_i[X_\Sigma^\lambda(t) - N_\Sigma^\lambda]^+$.

In contrast to FQR, the process $(Q^\lambda, I^\lambda, Z^\lambda)$ need not be a Markov chain under FWR, regardless of the tie-breaking rule. To avoid complications, we have hence explicitly defined the tie-breaking rule. As in the case of FQR, if p is a strictly positive vector, then we can choose, upon service completion, the customer from the head the class i^* queue where,

$$i^* \equiv i^*(t) \in \underset{i \in I(j), Q_i^\lambda(t) > 0}{\operatorname{argmax}} \left\{\frac{W_{h,i}^\lambda(t)}{p_i}\right\}$$

This fact also leads to a nice feature of FWR in the context of SL targets. Specifically, if $T_1, \ldots, T_I$ are strictly positive SL targets, i.e., $T_i > 0,\ i \in \mathcal{I}$, then FWR with the ratio vector p as defined in equation (3)

reduces to choosing class i^* with

$$i^* \equiv i^*(t) \in \underset{i \in I(j), Q_i^\lambda(t) > 0}{\operatorname{argmax}} \left\{ \frac{W_{h,i}^\lambda(t)}{T_i} \right\}.$$

This is a rather intuitive rule as it suggests serving customers in decreasing order of the violations of their associated SL targets. We point out that, in contrast to FQR (see (3)), the FWR rule with SL constraints is completely independent of the arrival rates.

We end this section by establishing the asymptotic equivalence of FQR and FWR. This theorem draws on the strong from of Little's law, which was mentioned in §2. Rather than tackling directly the equivalence, we show that Theorem 3.1 holds when FQR is replaced with FWR with the same ratio vectors. As all the results in §2-4 build on this state-space collapse, they all hold for FWR just as they do for FQR. For the following, we let $\hat{W}_{h,i}^\lambda(t) = \sqrt{\lambda} W_{h,i}^\lambda(t)$. Instead of directly assuming that the waiting times of all the customers initially in the system at time 0 scale properly, we will make the stronger assumption that all queues are empty at time 0. We believe that the results extend to more general initial conditions, but the treatment of the augmented hydrodynamic model equations (see §5 of [19]) used to prove state-space collapse become significantly more complicated. The following is proved in Theorem 4.4 of [20].

Theorem 5.1 (**state-space collapse under FWR**) *In addition to the conditions of Theorem 3.1, assume that* $\hat{Q}_i^\lambda(0) = 0$ *for all* λ *and all* $i \in \mathcal{I}$*. Then, in addition to the results of Theorem 3.1,*

$$\hat{Q}_i^\lambda(t) - a_i \hat{W}_{h,i}^\lambda(t) \Rightarrow 0 \quad \textit{in} \quad D \quad \textit{as} \quad \lambda \to \infty, \quad i \in \mathcal{I}. \tag{40}$$

6 Conclusions and Future Research

In this paper we have proposed the *fixed-queue-ratio* (FQR) routing scheme, specified in Definition 3.3, which typically reduces to the decentralized control in (19). The FQR control is important because it provides a way to achieve desired *service-level differentiation* in call centers with multiple customer classes and multiple agent types, operating under service-level constraints. We think that FQR has great appeal because of its simplicity and transparency. The key is to be able to produce the appropriate ratio probabil-

ity vector p in (3), exploiting the approximate relations in (4). As a consequence, FQR routing facilitates tremendously the construction of *combined staffing-design-and-routing solutions* for some settings of the complicated *skill-based-routing* (SBR) problem, with precisely specified goals.

From both the engineering and mathematical perspectives, an important contribution of this paper is to develop useful asymptotic formulations. As illustrated by our first results, Theorems 2.1 and 2.2, we consider both asymptotic feasibility and asymptotic optimality. The asymptotic results show that some care is needed in formulating both the constraints and the notion of asymptotic optimality. The theorems in turn substantiate the benefits of FQR and the overall approach for large systems.

The ability to demonstrate effectiveness of the proposed FQR solution for large systems, via the limit for the sequence of models as $\lambda \to \infty$, largely depends on the asymptotic state-space collapse that was proved in [19] and holds under certain network and parameter conditions. This state-space collapse is achieved in considerable generality, going well beyond the simple models in §2 and §4 of this paper. This wide applicability, in turn, suggests that we might be able to use FQR to construct simple solutions for much more general cases, perhaps using decomposition techniques as in Remark 3.3. On the other hand, we have observed that applying FQR when the conditions for good asymptotic performance are not satisfied can lead to serious problems.

Therefore, it is useful to summarize the practical implications of the two main state-space collapse results: Theorems 3.1 and 3.2. Combined, these two results suggest a simple rule of thumb:

- If there exists an optimal solution $(\bar{\nu}, \bar{x})$ for the fundamental mathematical program (16) for which the FQR conditions hold - use **FQR** and **SRSS**$(\bar{\nu})$.
- Otherwise, if the conditions above do not hold, but there exists an optimal solution $(\bar{\nu}, \bar{x})$ for (16) such that condition 4 holds, then use **GFQR** and **SRSS**$(\bar{\nu})$.
- Alternatively, if **FQR** and **SRSS**$(\bar{\nu})$ cannot be used directly, consider alternative designs, as in Remark 3.3, that would allow FQR and SRSS$(\bar{\nu})$ to be applied.

In [15], we use the above recipes to propose a simple heuristic for the combined design-staffing-and-routing problem in more general SBR networks, which do not satisfy the condition of pool-dependent service rates.

The state-space-collapse results and other results used in the appendix here can be found in [19]. That paper introduces the more general QIR controls, which allow the ratios to be state-dependent. The more general QIR rule is also used in in [20], under the assumption of pool-dependent service rates, to minimize convex holding costs over a finite time horizon, thus extending partially the $Gc\mu$ result in [25] to the QED regime.

Among the many important issues remaining unaddressed in our work, we regard model uncertainty as an important one. Throughout this paper we have assumed that the arrival rates λ_i, the numbers of available servers N_j, and the service rates $\mu_{i,j}$ are known and fixed. That is clearly an idealization; uncertainty and estimation errors often lead to unexpected system load. In systems with model uncertainty, more attention needs to be paid to design issues, as the system flexibility is now expected to play a more important role. While certain forms of model uncertainty may be treated within the framework in this paper, others will require different analysis. It is of interest, then, to examine the performance of FQR in settings with different forms of model uncertainty. That would parallel recent work on model uncertainty by Atar [7], Bassamboo et al. [9, 10] and Whitt [34].

Acknowledgments: The authors are grateful to Avi Mandelbaum and Mor Armony for fruitful discussions, and to Zohar Feldman and Ohad Perry for contributions to the simulation, including the use of their codes. The second author was partially supported by NSF grant DMI-0457095.

References

[1] Akşin O.Z., F. Karaesmen. 2005. Characterizing the performance of process flexibility structures. Working Paper, Koç University, Istanbul, Turkey.

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

[3] Armony M. A. Mandelbaum. 2004. Design, staffing and control of large service systems: The case of a single customer class and multiple server types. Working Paper. New York University, New York, NY and Technion - The Israeli Institute of Technology, Haifa, Israel.

[4] Armony M., I. Gurvich, A. Mandelbaum. 2006. Service-level differentiation in call Centers with fully flexible servers. *Management Science* forthcoming.

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

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

[7] Atar R. 2007. Central limit theorem for a many-server queue with random service rates. working paper, the Technion.

[8] Baccelli, F., P. Bremaud. 1994. *Elements of Queueing Theory: Palm-Martingale Calculus and Stochastic Recurrences*. Springer-Verlag, New York.

[9] 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.

[10] 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.

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

[12] Browne S., W. Whitt. 1995. Piecewise-Linear Diffusion Processes. J. Dshalalow, ed. *Advances in Queueing*. CRC Press, Boca Raton, FL, 463-480.

[13] Budhiraja, A., C. Lee. 2007. Stationary distribution convergence for generalized Jackson networks in heavy traffic. Working paper, The University of North Carolina at Chapel Hill.

[14] Dai J.G., T. Tezcan. 2005. State space collapse in many server diffusion limits of parallel server systems. Working Paper, Georgia Institute of Technology, Atlanta, GA.

[15] Feldman Z., I. Gurvich and W. Whitt. 2007. Constraint satisfaction in call-centers: an algorithm based on the FQR control. Working paper, Columbia University, New York, NY.

[16] Gamarnik D., A. Zeevi. 2006. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *Annals of Applied Probability* **16** 56-90.

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

[18] Garnett O., A. Mandelbaum, M. Reiman. 2002. Designing a Call Center with Impatient Customers. *Manufacturing Service Oper. Management*, **4**(3), 208-227.

[19] Gurvich I., W. Whitt. 2007. Queue-and-idleness-ratio controls in many-server service systems. Working paper, Columbia University, New York, NY.

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

[21] Gurvich I., A. Zeevi. 2007. Validity of heavy-traffic steady-state Approximations in open queueing networks: Sufficient conditions involving state-space collapse. Working paper, Columbia University, New York, NY.

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

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

[24] 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.

[25] 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.

[26] Puhalskii A. 1994. On the invariance principle for the first passage time, *Math. Oper. Res.* **19**(4) 946 - 954.

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

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

[29] 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.

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

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

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

[33] 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.

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

Appendix

In this appendix we provide the proofs of all major results in the paper. The proofs for §3 and §4 appear in §A and §B respectively. Finally, Remark B.1 covers the results of §2.

A Proofs for §3

A.1 Proof of Theorems 3.1 and 3.2

Theorems 3.1 and 3.2 follow, respectively, from Theorems 3.1 and 4.2 of [19], but an additional argument is needed because the conditions of the Theorems here and in [19] are expressed slightly differently. In particular, the conditions in [19] are stated in terms of solution to the LP (4) there, whereas Theorem 3.1 here uses the mathematical program (16) here.

Towards this end, fix an optimal solution $(\bar{\nu}, \bar{x})$ of (16) for which the edges of $\mathcal{E}(\bar{\nu}, \bar{x})$ induce a connected graph. By Assumption 3.1, we have that $\sum_{j \in J} \mu_{i,j} \bar{\nu}_j \bar{x}_{i,j} = a_i$ for all $i \in \mathcal{I}$. Considering (4) in [19] with

ν fixed to $\bar{\nu}$, we have that $\bar{x}$ together with $\rho = 1$ constitute a feasible solution. Moreover, it is an optimal one. Indeed, if given $\bar{\nu}$, there exists an optimal solution (x, ρ) for (4) in [19] with $\rho < 1$, then the solution $(\bar{\nu}, x)$ is optimal for (16) contradicting Assumption 3.1. Finally, with $(\bar{x}, 1)$ being the solution for (4) in [19] associated with $\bar{\nu}$, the routing graph $\mathcal{E}(\bar{x})$ as defined in Assumption 2.3 of is the routing graph $\mathcal{E}(\bar{\nu}, \bar{x})$. Consequently, the former is connected if and only if the latter is.

Also, we note that with the selected solution $(\bar{\nu}, \bar{x})$, the staffing determined through $SRSS(\bar{\nu})$ satisfies the heavy-traffic conditions in Assumption 2.1 of [19]. Finally, we observe that FQR is a special case of QIR with the constant ratio functions $p(\cdot) \equiv (p_1, \ldots, p_I)$ and $v(\cdot) \equiv (v_1, \ldots, p_I)$ that are clearly admissible in the sense of Definition 2.2 in [19].

Theorem 3.2 follows from Theorem 4.2 in [19] using similar arguments.

A.2 Proof of Theorem 3.3

The proof relies on Proposition 5.1 in [19] that connects the tail of the waiting time distribution to the the limit, $\hat{X}_\Sigma(t)$, of the scaled number-of-customers in the system, $\hat{X}^\lambda_\Sigma(t)$. The result of the Theorem will then follow by proving that the family of processes $\{\hat{X}^\beta(t)\}_{\beta \geq 0}$, where the superscript denotes the dependence on β, is, in some sense, monotone decreasing in β.

First, to apply Proposition 5.1 from [19], we observe that, under the assumptions of Theorem 3.1, the sequence $\hat{X}^\lambda_\Sigma(t)$ is C-Tight. We explain now how this conclusion follows from [19]. First, under condition C-1, we show in §4.2 in [19], that FQR is equivalent to the GFQR control used in [6]. The C-Tightness under this condition follows then from Proposition 1 in [6]. For conditions C-2 and C-3, the C-Tightness follows, respectively, from Theorems 5.1 and 5.3 in [19].

We now fix a convergent subsequence of $\hat{X}^\lambda_\Sigma(t)$. As our results below will hold for any convergent subsequence, we assume, without loss of generality, that the whole sequence converges to a limit with continuous sample paths. Let $\hat{X}_\Sigma(t)$ be this limit. Fixing $t > 0$, and a ratio vector p, we have by Proposition 5.1 in [19] that

$$\bar{F}_i^{\lambda,T}(y/\sqrt{\lambda}) \Rightarrow \frac{1}{T}\int_0^T \mathbf{1}\left\{\frac{p_i}{a_i}\left[\hat{X}_\Sigma(t)\right]^+ > y\right\} dt \quad \text{in} \quad [0,1] \quad \text{as} \quad \lambda \to \infty. \tag{A1}$$

Equation (24) is now equivalently stated as

$$P\left\{\frac{1}{T}\int_0^T \mathbf{1}\left\{\frac{p_i^*}{a_i}\left[\hat{X}_\Sigma^\beta(t)\right]^+ > \bar{T}_i\right\} dt \geq \alpha+\epsilon\right\} \leq \epsilon,\ i\in\mathcal{I}$$

where we have added the superscript β to reflect the dependence on β. The proof will be complete, then, if we show that

$$\lim_{\beta\to\infty} P\left\{\frac{1}{T}\int_0^T \mathbf{1}\left\{\frac{p_i^*}{a_i}\left[\hat{X}_\Sigma^\beta(t)\right]^+ > \bar{T}_i\right\} dt \geq \alpha+\epsilon\right\} = 0,\ i\in\mathcal{I}. \tag{A2}$$

This follows from the following Lemma by choosing any $\delta < \alpha$.

Lemma A.1 *For any $\delta > 0$,*

$$\sup_{\delta\leq t\leq T}\left[\hat{X}_\Sigma^\beta(t)\right]^+ \xrightarrow{P} 0,\ \textit{as}\ \beta\to\infty.$$

Proof: We start by observing that under any of the conditions C-1, C-3 and C-3, the limit process $\hat{X}_\Sigma^\beta(t)$ satisfies

$$\hat{X}_\Sigma^\beta(t) \leq \hat{X}_\Sigma(0) - \beta t + c_1\int_0^t [\hat{X}_\Sigma^\beta(s)]^- ds - c_2\int_0^t [\hat{X}_\Sigma^\beta(s)]^+ ds + \sqrt{2}\hat{B}(t), \tag{A3}$$

for some constants $c_1 > 0$, $c_2 > 0$ and where $\hat{B}(t)$ is a standard Brownian motion. This follows from Theorems 5.1 and 5.3 in [19] under conditions C-2 and conditions C-3 respectively. For condition C-1, this is a consequence of the diffusion limits in [6] and the equivalence of FQR to GFQR under this condition; see §4.2 of [19]. Consequently,

$$\hat{X}_\Sigma^\beta(t) \leq \hat{X}_\Sigma^\beta(0) - \beta t + c_1\int_0^t [\hat{X}_\Sigma^\beta(s)]^- ds + \sqrt{2}\hat{B}(t).$$

Define now $\tau_0^\beta = \inf\{t\geq 0 : \hat{X}_\Sigma^\beta(t) \leq \epsilon\}$. We claim that

$$\tau_0^\beta \xrightarrow{P} 0,\ \text{as}\ \beta\to\infty. \tag{A4}$$

Indeed, assume that $\hat{X}_\Sigma^\beta(0) = x > \epsilon$. Then, for all $t \leq \tau_0^\beta$ we have that

$$\hat{X}_\Sigma^\beta(t) \leq x - \beta t + \sqrt{2}\hat{B}(t).$$

Given $\eta > 0$, $P\{\tau_0^\beta \geq \eta\} = P\{\inf_{0\leq t\leq\eta} x - \beta t + \sqrt{2}\hat{B}(t) \geq \epsilon\}$ so that $\tau_0^\beta \xrightarrow{P} 0$ readily follows from basic

properties of Brownian motion. For $\delta, \epsilon > 0$ we now define

$$\tau_2^\beta = \inf\{t \geq \tau_0^\beta : [\hat{X}_\Sigma^\beta(t)]^+ \geq 2\epsilon\}, \text{ and } \tau_1^\beta = \sup\{t \geq \tau_0^\beta, t \leq \tau_2^\beta : [\hat{X}_\Sigma^\beta(t)]^+ \leq \epsilon\}.$$

Then,

$$\hat{X}_\Sigma^\beta(t) \leq \hat{X}_\Sigma^\beta(\tau_2^\beta) - \beta(t - \tau_2^\beta) + \sqrt{2}\hat{B}(t) - \sqrt{2}\hat{B}(\tau_2^\beta), \tag{A5}$$

for all $\tau_1^\beta \leq t \leq \tau_2^\beta$. Now,

$$P\left\{\sup_{\tau_0^\beta \leq t \leq T} [\hat{X}_\Sigma^\beta(t)]^+ \geq 2\epsilon\right\} \leq P\left\{\sup_{\tau_1^\beta \leq s \leq t \leq \tau_2^\beta \wedge T} [\hat{X}_\Sigma^\beta(t)]^+ - [\hat{X}_\Sigma^\beta(s)]^+ \geq \epsilon\right\} \tag{A6}$$

Using (A5) and noting that $[\hat{X}_\Sigma^\beta(t)]^+ = \hat{X}_\Sigma^\beta(t)$ for $\tau_1^\beta \leq t \leq \tau_2^\beta$ it is now straightforward to show that

$$P\left\{\sup_{\tau_1^\beta \leq s \leq t \leq \tau_2^\beta \wedge T} [\hat{X}_\Sigma^\beta(t)]^+ - [\hat{X}_\Sigma^\beta(s)]^+ \geq \epsilon\right\} \to 0, \text{ as } \beta \to \infty. \tag{A7}$$

The result now follows by noting that for each $\delta, \epsilon > 0$,

$$P\left\{\sup_{\delta \leq t \leq T} \left[\hat{X}_\Sigma^\beta(t)\right]^+ \geq 2\epsilon\right\} \leq P\{\tau_0^\beta \geq \delta\} + P\left\{\sup_{\tau_0^\beta \leq t \leq T} [\hat{X}_\Sigma^\beta(t)]^+ \geq 2\epsilon\right\},$$

and applying (A4) and (A7). ■

B Proofs for §4

B.1 Proof of Theorem 4.1

The proof consists of two steps. First, a coupling argument, Lemma B.1 below, shows that the $\bigwedge$ model constitutes a lower bound for the SBR system in terms of the aggregate queue length. Lemma B.2 builds on [2] to argue that any staffing vector for the $\bigwedge$ model that satisfies the constrain (25) must satisfy the capacity constraint (26). Together, these steps imply the result of the Theorem.

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 an inverted-V 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 5.

Given admissible controls π_1 and π_2 for the SBR and $\bigwedge$ model, respectively, we let $Z^{\pi_1}_{j,SBR}(t)$ and $Z^{\pi_2}_{j,\bigwedge}(t)$ 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. We add a superscript to explicitly express the dependence on the control.

Lemma B.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 $\bigwedge$ 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 B.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 we $\bar{W}^{\lambda,T}_{\bigwedge}$ be the aggregate averaged waiting time, as defined in (11) but add the superscript $\bigwedge$ to denote that it corresponds to the $\bigwedge$ model rather than with the general SBR system.

Lemma B.2 (**a feasibility result for the $\bigwedge$ model**) *Fix a sequence of $\bigwedge$ models with capacity vectors that*

satisfy

$$\liminf_{\lambda\to\infty} \frac{N_j^\lambda}{\lambda} > 0,\ j \in \mathcal{J}. \tag{A8}$$

Then,

$$\limsup_{\lambda\to\infty} P\left\{ \frac{\bar{W}_{\bigwedge}^{\lambda,T}}{T_I^\lambda} \geq 1+\epsilon \right\} \leq \epsilon \tag{A9}$$

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

$$\sum_{j\in\mathcal{J}} \mu_j N_j^\lambda \geq \lambda + \beta^*\sqrt{\lambda} + o(\sqrt{\lambda}), \tag{A10}$$

where β^* *is the unique solution to*

$$\frac{\mathbf{P}_{\mu_1}(\beta)}{\beta} = \sqrt{\lambda} T_I^\lambda =: \bar{T}_I, \tag{A11}$$

and

$$\mathbf{P}_{\mu_1}(\beta) := \left[1 + \frac{\frac{\beta}{\sqrt{\mu_1}} \Phi\left(\frac{\beta}{\sqrt{\mu_1}}\right)}{\phi\left(\frac{\beta}{\sqrt{\mu_1}}\right)} \right]^{-1}.$$

Proof: First, we have that

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

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 [31]). 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

$$\limsup_{\lambda\to\infty} P\left\{ \frac{1}{T} \int_0^T \frac{Q_{\Sigma,\bigwedge}^\lambda(t)}{\lambda T_I^\lambda} dt \geq 1+\epsilon \right\} \leq \epsilon. \tag{A12}$$

We now apply the results of [2]. 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 [2] 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. The argument is extended to arbitrary controls in the end of the proof. Equation (A8) allows us to apply Proposition 4.2 in [2], and the continuity of the integral map to have that

$$\lim_{\lambda\to\infty} \frac{1}{T}\int_0^T \frac{Q^\lambda_{\Sigma,\bigwedge}(t)}{\lambda T_I^\lambda} dt = \frac{1}{T}\int_0^T \frac{[\hat{X}_{\Sigma,\bigwedge}(t)]^+}{\bar{T}_I} dt, \tag{A13}$$

Where $\hat{X}_{\Sigma,\bigwedge}(t)$ is the limit in Proposition 4.2 of [2]. Here, we also used Assumption 2.1 to write $T_I^\lambda = \bar{T}_I/\sqrt{\lambda}$.

By Proposition 4.4 in [2], $\hat{X}_{\Sigma,\bigwedge}(t)$ 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 [23]. 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 [5] for the general regenerative-process result and Theorem 3.1 of Khas'miniskii [24] for the diffusion-process result. Through simple integration, Proposition 4.4 in [2] also shows that $E\left[\left[\hat{X}_\Sigma(\infty)\right]^+\right] = \mathbf{P}_{\mu_1}(\hat{\beta})/\hat{\beta}\bar{T}_I$. Using Lemma B.1 in Borst et. al. [11], it follows that this expectation is decreasing as a function of its argument, β, and we have that for some $\epsilon > 0$,

$$\frac{1}{T}\int_0^T \frac{\left[\hat{X}_{\Sigma,\bigwedge}(t)\right]^+}{\bar{T}_I} dt \to \frac{E\left[\left[\hat{X}_{\Sigma,\bigwedge}(\infty)\right]^+\right]}{\bar{T}_I} = \mathbf{P}_{\mu_1}(\hat{\beta})\frac{1}{\hat{\beta}\bar{T}_I} > 1+\epsilon \quad \text{as } T\to\infty \quad \text{w.p.1}\,, \tag{A14}$$

where the last inequality easily follows from the fact that $\hat{\beta} < \beta^*$. Combining equations (A12)-(A14) 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 (A9) can not hold. It still remains to argue what happens when $\hat{\beta} \leq 0$. This, however, follows from our previous argument. Indeed, considering the diffusion process $\hat{X}_{\Sigma,\bigwedge}(t)$ of Proposition 4.2 in [2], it is a matter of a simple comparison result for diffusions (see e.g. Proposition 5.2.18 in [23]) to show that if for

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

for some $\epsilon > 0$ and all T large enough, then the same holds for all $\beta \leq \hat{\beta}$ and in particular for $\beta \leq 0$. We proved, then, that if FSF is used any staffing level that satisfies (A9) and (A8) must also satisfy (A11). To rule out the possibility that there is a different control under which (A11) is not necessary, we point out that by Theorem 4.1 of [20] we have that for each $T > 0$,

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

where π is some control other than FSF and $Q_{\Sigma,\bigwedge}^{\pi,\lambda}(t)$ is the queue-length process under this control. Consequently, repeating all the arguments in the proof we have that if (A12) is violated under FSF, it is also violated under any other control. The proof is now complete. ■

Proof of Theorem 4.1 Given Lemmas B.1 and B.2 the proof the Theorem is straightforward. Specifically, we can construct the sample paths so that

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

where the first inequality is as in the proof of Lemma B.2 and the second inequality follows from Lemma B.1. One now repeats the proof of Lemma B.2 for the right hand side to obtain the result of the Theorem.

■

B.2 Proof of Theorems 4.2 and 4.3

In Theorem 4.1, we established a lower bound on the staffing vector as given by a $\bigwedge$-based staffing. Towards asymptotic optimality and feasibility it remains only to show that such a 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 (30) and (31) in the case of Theorem 4.2 and (34) in the case of Theorem 4.3. These results will be a consequence of two Lemmas. The first, Lemma (B.3, connects the performance measures for the SBR system with performance measures for the $\bigwedge$ model. As before, $\hat{Q}^{\lambda}_{\Sigma,\bigwedge}(t)$ is the queue-length process in the $\bigwedge$ model operated under the Faster-Server-First (FSF) policy defined in [2].

Lemma B.3 (**asymptotic equivalence with the $\bigwedge$ model**) *Under the conditions of Theorem 4.2 (respectively Theorem 4.3, we have that for every $T \geq 0$ and $y \geq 0$*

$$\lim_{\lambda\to\infty} \frac{1}{T}\int_0^T \mathbf{1}\left\{Q^{\lambda}_{\Sigma,SBR}(t) > y\sqrt{\lambda}\right\} dt \stackrel{d}{=} \lim_{\lambda\to\infty} \frac{1}{T}\int_0^T \mathbf{1}\left\{Q^{\lambda}_{\Sigma,\bigwedge}(t) > y\sqrt{\lambda}\right\} dt, \tag{A15}$$

and in particular,

$$\lim_{\lambda\to\infty} \bar{F}_i^{\lambda,T}(y/\sqrt{\lambda}) \stackrel{d}{=} \lim_{\lambda\to\infty} \frac{1}{T}\int_0^T \mathbf{1}\left\{\frac{p_i}{a_i} Q^{\lambda}_{\Sigma,\bigwedge}(t) > y\sqrt{\lambda}\right\} dt,\ i \in \mathcal{I}. \tag{A16}$$

Also,

$$\lim_{\lambda\to\infty} \sqrt{\lambda}\bar{W}^{\lambda,T} \stackrel{d}{=} \lim_{\lambda\to\infty} \sqrt{\lambda}\bar{W}^{\lambda,T}_{\bigwedge}, \tag{A17}$$

and in particular,

$$\lim_{\lambda\to\infty} \frac{\bar{W}^{\lambda,T}}{T^{\lambda}_I} \stackrel{d}{=} \lim_{\lambda\to\infty} \frac{\bar{W}^{\lambda,T}_{\bigwedge}}{T^{\lambda}_I}. \tag{A18}$$

Proof: By Theorem 1.1 in [19] 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 5.1 in [19] 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 5.1 in [19]. Some care is required in applying this integral mapping and we refer the reader to the proof of Proposition 5.1 in [19] for further details.

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 [2]. 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 5.1 in [19] in the same manner. ■

The next Lemma shows that the performance measures for the $\bigwedge$ model that appear in Lemma B.3 indeed satisfy the required bounds.

Lemma B.4 *Consider a sequence of $\bigwedge$ models operated under FSF and satisfying (A8) and*

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

with β^* as given (27) (respectively (33)). Assume that

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

and

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

We then have the following:

1. *Suppose that β^* is determined through (27). Then, for each $\epsilon > 0$, there exists $T^*(\epsilon)$ so that for all $T \geq T^*(\epsilon)$,*

$$\lim_{\lambda\to\infty} P\left\{\frac{\bar{W}^{\lambda,T}_\bigwedge}{T^\lambda_I} \geq 1+\epsilon\right\} \leq \epsilon, \tag{A19}$$

$$P\left\{\frac{1}{T}\int_0^T \mathbf{1}\left\{p_{I-1}Q^\lambda_{\Sigma,\bigwedge}(t) > \lambda_{I-1}T^\lambda_{I-1}\right\} dt \geq \alpha+\epsilon\right\} \leq \epsilon, \tag{A20}$$

   and

$$P\left\{\frac{1}{T}\int_0^T \mathbf{1}\left\{p_i Q^\lambda_{\Sigma,\bigwedge}(t) > \lambda_i T^\lambda_i\right\} dt \geq \alpha+\epsilon\right\} \leq \epsilon. \tag{A21}$$

2. *Suppose that β^* is determined through (33). Then, for each $\epsilon > 0$ there exists $T^*(\epsilon)$ so that, for all*

$T \geq T^*(\epsilon)$,

$$\lim_{\lambda\to\infty} P\left\{\frac{1}{T}\int_0^T \mathbf{1}\left\{Q^\lambda_{\Sigma,\bigwedge}(t) > \sum_{i\in\mathcal{I}} \lambda_i T_i^\lambda\right\} dt \geq \alpha + \epsilon\right\} \leq \epsilon. \tag{A22}$$

Proof: Equation (A19) is actually proved within the proof of Lemma B.2. In particular, it is proved there that

$$\limsup_{\lambda\to\infty} \frac{\bar{W}^{\lambda,T}_{\bigwedge}}{T_I^\lambda} \Rightarrow 1, \text{ as } T \to \infty.$$

Equation (A20) follow by applying the argument in the proof of Lemma B.3 to show that for each $y > 0$,

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

where $\hat{X}_\Sigma(t)$ is the diffusion limit from Proposition 4.2 in [2] with δ there set equal to β^*. By that same proposition, and applying the theory of regenerative processes as in the proof of Lemma B.2, we have that

$$\frac{1}{T}\int_0^T \mathbf{1}\left\{\hat{X}_\Sigma(t) > \sum_{i\in\mathcal{I}} a_i\bar{T}_i\right\} \Rightarrow \mathbf{P}_{\mu_1}(\beta)e^{-\beta^*\sum_{i\in\mathcal{I}} a_i\bar{T}_i} = \alpha.$$

Here, the equality follows from the definition of p_{I-1} in (29) Equation (A21) now follows from the definition of p_i for $i \neq I-1$. Equation (A22) is proved in a similar manner using the definition of p in (3). ■

Proof of Theorem 4.2 Equation (30) follows from the choice of p in (29) and combining (A16), (A20) and (A21). Equation (31) is obtained by combining (A18) and (A19). Finally, equation (32) follows from the fact that $c \cdot N^\lambda$ is an asymptotic lower bound by Theorem 4.1. ■

Proof of Theorem 4.3 Equation (34) follows readily by combining (A16) and (A22) and using the definition of the ratio vector p in (3). ■

Proof of Proposition 4.1 Let $A^\lambda_{ij}(t)$ be the aggregate number class-i customers routed (upon arrival or service completion) to agents in pool j by time t. By definition, then,

$$DS_i^\lambda[T] = \frac{\Phi^\lambda_{ij}(T) + A^\lambda_{ij}(T)}{A^\lambda_i(T)}$$

for $j = j(\{i\})$. For $j = j(\{i\})$, however, we have that $Z_j^\lambda(t) = Z_j^\lambda(0) + A_{ij}^\lambda(T) - D_j^\lambda(t)$, where $D_j^\lambda(t)$ is the number of service completion in pool j by time t. In particular, $A_{ij}^\lambda(T) = Z_j^\lambda(t) - Z_j^\lambda(0) + D_j^\lambda(t) \geq -N_j^\lambda + D_j^\lambda(t)$. Hence,

$$DS_i^\lambda[T] = \frac{1}{A_i^\lambda(T)}(A_{ij}^\lambda(T)) \geq \frac{-N_j^\lambda + D_j^\lambda(T)}{A_i^\lambda(T)}.$$

The result is now established using basic renewal theory, the stochastic boundedness of $I_j^\lambda(t)$ from Corollary 5.3 in [19], and the definition of N_j^λ. In particular, a direct consequence of Corollary 5.3 in [19] is that

$$\frac{\mu \int_0^T Z_j^\lambda(s)ds}{\lambda} \Rightarrow \mu\tilde{\nu}_j T,$$

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 [32]), to conclude that

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

In particular,

$$\frac{D_j^\lambda(T)}{A_i^\lambda(T)} \Rightarrow \frac{\mu}{a_i}\tilde{\nu}_j = 1-\delta, \text{ as } \lambda \to \infty. \tag{A23}$$

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

$$\limsup_{\lambda\to\infty} \frac{N_j^\lambda}{A_i^\lambda(T)} \leq \epsilon/2. \tag{A24}$$

Combining (A23) and (A24), we conclude that there exists T large enough so that

$$P\left\{DS_i^\lambda[T] \leq 1-\delta-\epsilon\right\} \leq \epsilon.$$

Equation (36) now follows by recalling that $\delta = \min_i \delta_i$. ■

B.3 Proof of Lemma 4.1

We prove the result for $\widetilde{SCS}$. The proof is similar for SCS. First, note that with a sequence of staffing levels $N^\lambda = \lambda/\mu + \beta\sqrt{\lambda/\mu} + o(\sqrt{\lambda/\mu})$, for some $\beta > 0$, we have by Theorem 1 of [22] that

$$\frac{Q^{FCFS}_{\lambda,\mu}(N^\lambda)}{\sqrt{N^\lambda}} \Rightarrow \tilde{Q}^{FCFS}(\beta), \text{ in } \mathbb{R}, \text{ as } \lambda \to \infty,$$

where $P\{\tilde{Q}^{FCFS}(\beta) > x\} = \mathbf{P}_\mu(\beta/\sqrt{\mu})e^{-\frac{\beta}{\sqrt{\mu}}x}$ and $Q^{FCFS}_{\lambda,\mu}(N^\lambda)$ is the steady-state queue length in the given $M/M/N$ queue. Since $N^\lambda/\lambda \to 1/\mu$ as $\lambda \to \infty$ we also have that

$$\frac{Q^{FCFS}_{\lambda,\mu}(N^\lambda)}{\sqrt{\lambda}} \Rightarrow \hat{Q}^{FCFS}(\beta), \tag{A25}$$

where $\hat{Q}^{FCFS}(\beta) := \tilde{Q}^{FCFS}(\beta)/\sqrt{\mu}$. The convergence in (A25) holds also in expectation by Corollary 1 of [22]. Moreover, $E[\hat{Q}^{FCFS}(\beta)]$ is a continuous strictly decreasing function of β for $\beta \in (0, \infty)$ with $E[\hat{Q}^{FCFS}(\beta)] \uparrow \infty$ as $\beta \to 0$ and $E[\hat{Q}^{FCFS}(\beta)] \downarrow 0$ as $\beta \to \infty$ (see [22] as well as Lemmas B.1 and C.1 of [11]). As a consequence, we can choose β^* so that

$$E[\hat{Q}^{FCFS}(\beta^*)] = \bar{T}_I. \tag{A26}$$

We are now ready to prove (37). We will give a proof by contradiction. First, suppose that

$$\limsup_{\lambda\to\infty} \frac{\mu\bar{N}^\lambda_\Sigma - \lambda}{\sqrt{\lambda}} \geq \beta^* + \epsilon,$$

for some $\epsilon > 0$. Then, with $N^\lambda = \lceil \lambda/\mu + (\beta^* + \epsilon/2)\sqrt{\lambda}/\mu \rceil$ we have by equation (A25) that

$$E\left[\frac{Q^{FCFS}_{\lambda,\mu}(N^\lambda)}{\sqrt{\lambda}}\right] \to E\left[\hat{Q}^{FCFS}(\beta^* + \epsilon/2)\right], \text{ as } \lambda \to \infty.$$

Together with (A26), the above implies that there exist λ large enough with $N^\lambda < \bar{N}^\lambda_\Sigma$ and

$$E\left[\frac{Q^{FCFS}_{\lambda,\mu}(N^\lambda)}{\sqrt{\lambda}}\right] \leq \bar{T}_I,$$

contradicting the definition of $\bar{N}_\Sigma^\lambda$. To complete the argument assume that

$$\liminf_{\lambda\to\infty} \frac{\mu\bar{N}_\Sigma^\lambda - \lambda}{\sqrt{\lambda}} \leq \beta^* - \epsilon,$$

for some $\epsilon > 0$. Then, repeating a similar argument we have that for λ large enough

$$E\left[\frac{Q_{\lambda,\mu}^{FCFS}(\bar{N}_\Sigma^\lambda)}{\sqrt{\lambda}}\right] > \bar{T}_I,$$

contradicting the definition of $\bar{N}_\Sigma^\lambda$ once again. Overall, we have established that

$$\mu\bar{N}_\Sigma^\lambda = \lambda + \beta^*\sqrt{\lambda} + o\left(\sqrt{\lambda}\right). \tag{A27}$$

■

Remark B.1 (**proofs of Theorems 2.1 and 2.2**) For the V-Model, the proofs are significantly easier. First, it is immediate that the conditions of Corollary 6.4 in [19] apply for the V-Model. Indeed, the overall number of customers in the V-Model under any work-conserving policy is identical to the corresponding number in the associated $M/M/N$ queue, for which existence of steady-state distributions and tightness of the sequence of stationary distributions is known from [22]. In particular, in the V Model we can consider the optimization problem in steady state form as in (1) and (7). Once this is done, the asymptotic feasibility and optimality are obtained through significant simplification of the proofs given for Theorem 4.2 and 4.3. More specifically, a detailed proof would be a very simple formalization of the intuitive argument given in §2 through the use of state-space collapse and the equivalence of the aggregate queue-length in the V model and the $M/M/N$ model. ■