

Staffing Call-Centers With Uncertain Demand Forecasts:

A Chance-Constrained Optimization Approach

Itai Gurvich*

James Luedtke[†]

Tolga Tezcan[‡]

February 13, 2010

We consider the problem of staffing call-centers with multiple customer classes and agent types operating under quality-of-service (QoS) constraints and demand rate uncertainty. We introduce a formulation of the staffing problem that requires that the QoS constraints are met with high probability with respect to the uncertainty in the demand rate. We contrast this *chance-constrained* formulation with the average-performance constraints that have been used so far in the literature. We then propose a two-step solution for the staffing problem under chance constraints. In the first step, we introduce a Random Static Planning Problem (RSPP) and discuss how it can be solved using two different methods. The RSPP provides us with a first-order (or fluid) approximation for the true optimal staffing levels and a *staffing frontier*. In the second step, we solve a finite number of staffing problems with known arrival rates—the arrival rates on the optimal staffing frontier. Hence, our formulation and solution approach has the important property that it translates the problem with uncertain demand rates to one with known arrival rates. The output of our procedure is a solution that is feasible with respect to the chance constraint and nearly optimal for large call centers.

1 Introduction

We consider the problem of staffing call centers in which customers of different classes are served by agents with varying skills (types). The staffing problem is traditionally formulated as an optimization problem in which the objective is to minimize salary-related costs subject to meeting pre-specified Quality-of-Service (QoS) targets for the various customer classes. The input to this optimization problem is composed of the salary costs, the QoS constraints, and various system parameters such as arrival rates, service times and customers' patience. Solutions to this optimization problem specify two actions: (a) the required number of agents with each given skill, and

*Kellogg School of Management, Northwestern University. (i-gurvich@kellogg.northwestern.edu)

[†]Industrial and Systems Engineering, University of Wisconsin-Madison. (jrluedt1@wisc.edu)

[‡]Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign. (ttezcan@uiuc.edu)

(b) a dynamic routing policy that defines how customers are assigned to agents in real time. In general, coming up with optimal (or even nearly optimal) staffing and routing solutions for this optimization problem is an extremely complicated task. In this paper we address two important issues in call-center workforce optimization: (i) the arrival rates are forecasted in advance and, consequently, they are rarely precise, and (ii) different customer classes have different quality of service requirements so that they cannot be treated as a single customer class.

In various settings it is reasonable to assume that the stream of calls to the call center during a given day follows a non-homogenous Poisson process; see e.g. [11]. Forecasting procedures provide a point estimate for the (time-varying) rate of this process. While forecasts can be progressively updated during the day, the inability to instantaneously summon agents implies that the call center has to schedule agents to shifts in advance, before any information is obtained about the actual realization of the demand rates. Of course, if the arrival rates are *perfectly predictable*, in the sense that the point estimates precisely predict the demand rates, then staffing decisions are somewhat simpler. This is, however, rarely the case. Even with abundant historical data, it is expected that some level of uncertainty remains.

Naturally, the magnitude of the forecasting error depends both on the profile of the call center's customers and on the sophistication of the forecasting procedure. Financial-industry call centers can experience highly unpredictable surges in demand following unpredictable events in the stock market. Other call centers operate in a less volatile environment. Forecasting procedures need to take into account various factors in translating the historical data into demand predictions. Obvious factors are the day of the week and seasonality effects. The most sophisticated forecasting procedures provide, in addition to the point estimate, an estimate for the distribution of the forecasting error, i.e, the gap between the point estimate and the actual realization of the arrival rates (see §3). Such distribution estimates are extremely valuable and can be used in making the staffing decisions.

With the distribution estimates, the staffing problem for a call center with a single customer class and a single type of agents can be solved via a simple simulation-based search that finds the minimum number of servers that satisfies the QoS constraint. Under the assumption that service times are exponentially distributed the search can be replaced by an even simpler solution; see §2. In a multi-class multi-type setting, however, the staffing problem is significantly more complex. This is, of course, not surprising. Even with perfectly predictable rates the staffing and routing problem for multi-class multi-skill call centers is extremely complicated and closed-form solutions are not known for all but the simplest cases. Here, the need to meet differentiated service levels

for the different classes prevents treating these as one “super-class.” The complexity is further exacerbated by the fact that, with multiple caller streams (corresponding to the different classes), the forecasting error can be multi-dimensional with possible dependencies between the different classes. Simple and efficient search mechanisms cannot be applied here in a computationally efficient manner and more sophisticated solutions are needed.

We make it our objective to explicitly address these complexities by creating a procedure that is applicable to general forms of forecast uncertainty and relatively general network structures. Our solution explicitly models the uncertainty associated with arrival rate forecasts and considers the multi-class multi-type structure that is present in many call centers. While we are not the first to consider these two problems jointly, our approach has two distinguishing features: (a) we use a chance-constrained formulation that is different from the standard average-performance one (see §2), and (b) our solution exhibits a very desirable property in that, in a sense, it inherits properties from the staffing problem with perfectly predictable rates. Specifically, we provide a solution approach that translates (through mathematical programming) the problem of staffing with uncertain demand rates to one of finding a solution for a finite (and small) set of staffing problems with perfectly predictable rates.

Our point of departure in formulating the staffing problem is the observation that when explicitly modeling the arrival rate uncertainty, a choice has to be made with respect to the formulation. With perfectly predictable rates, a QoS constraint might require, for example, that at most 5% of the callers abandon before being served. If the arrival rates are known, such a constraint can be formally imposed by requiring that the steady-state fraction of abandonments is less than the target of 5%. In the presence of arrival rate uncertainty, however, the steady-state fraction of customers that abandon is itself a random variable that obtains different values, depending on the realization of the arrival rates. Hence, a different definition of service-level constraints is required. One possibility is to require that the *expected fraction of abandonments* is less than 5%, where the expectation is taken with respect to the distribution of the arrival rates. Another possibility is to require that the constraint is met on some pre-specified fraction of the arrival-rate values. This leads to the chance-constrained formulation that we adopt in this paper.

The chance-constrained formulation for the staffing problem is roughly as follows: the call-center’s management chooses a *risk level*, δ , and allows the QoS to be violated on at most a fraction δ of the arrival-rate realizations. For example, a chance-constrained version of the 5%-abandonment constraint would stipulate that the fraction of abandoning customers is less than 5% on a fraction $1 - \delta$ of the days in a month. In contrast, the expected-value-constraint approach

would require that the average fraction of abandonments over the month is less than 5%. The chance-constrained formulation has an advantage in that it lets the manager adapt her formulation to the way that she is measured and the risk-level that she is willing to absorb. Through properly setting the risk-level, δ , the manager may choose her own compromise between staffing costs and “safety” in terms of the likelihood with which the QoS constraints are met. We further discuss the distinction between the formulations in §2.

At the heart of our solution approach is a static approximation of the chance-constrained formulation that we refer to as the Random Static Planning Problem (RSPP). This problem is a chance-constrained analog to the so-called Static Planning Problem that is often used to obtain first-order estimates for the optimal staffing levels and system design when arrival rates are perfectly predictable (see §3). The RSPP does not explicitly model the QoS targets, and ignores the dynamics of the call center, and hence does not require the selection of a routing policy. Instead, the RSPP seeks a set of staffing levels that minimize staffing costs subject to the requirement that the staffing levels are sufficient to meet the demand of all classes with a probability that is $1 - \delta$ where δ is the risk level. The output of the RSPP is a staffing solution and a set of arrival rate vectors which we call the *staffing frontier*; Figure 1 illustrates such a frontier for a call center with two classes. The support of the distribution is the positive orthant and the frontier, which would be determined by the RSPP, is the set of arrival-rate vectors that lie on the solid boundary of the colored region. The RSPP chooses the colored region (and its boundary) so that the probability mass within the chosen region is greater than $1 - \delta$ and so that the chosen region is, in some sense, optimal with respect to staffing cost. In the second step we solve staffing problems for the arrival rate vectors on the staffing frontier and show how to use the output of this frontier-based staffing problem to generate a solution for the original chance-constrained staffing problem.

Most importantly, the *staffing frontier* approach reduces the complex staffing problem with uncertain rates to one of solving multiple problems with predictable rates. The output is a staffing and routing solution that is feasible with respect to the chance constraint and is nearly optimal (in fluid scale) for large call centers.

We end this introduction by pointing out that in this work we mostly focus on stationary (but uncertain) rates. The extension to one case with time-varying rates is discussed in §8 but the explicit modeling of time variation is postponed to subsequent work.

The rest of the paper is organized as follows: In §2 we contrast the chance-constrained formulation with the average-performance formulation. §3 contains a review of the relevant literature. The formal problem formulation is given in §4. We treat a certain idealized case in §5. The RSPP

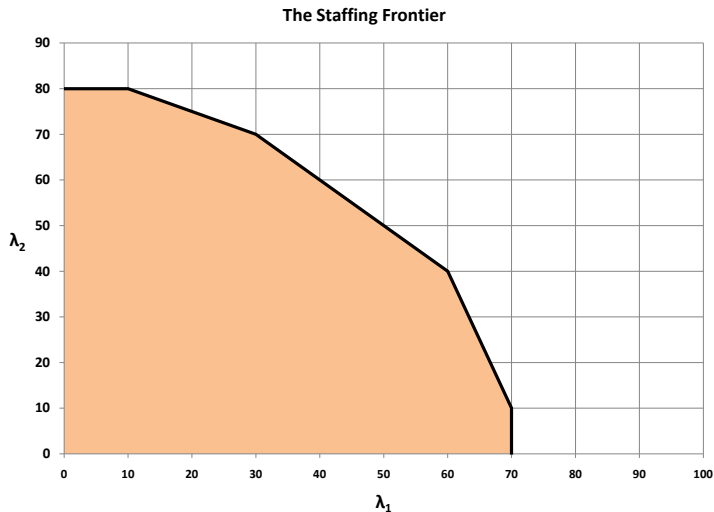


Figure 1

is introduced and analyzed in §6 where we also provide different solution approaches and computational results. Building on the RSPP, we then present in §7 a methodology to obtain feasible and nearly optimal solutions for the staffing problem with chance constraints. We conclude in §8 with some extensions and directions for future research. All the proofs are relegated to the e-companion.

2 Two alternative formulations of the staffing problem

In this section, we focus on a simple model of a call center in which there is a single class of customers and a single pool of agents. This simplified setting serves to illustrate the motivation for using a chance-constrained formulation for staffing call centers. We contrast the chance-constrained formulation with an average-constraint formulation and illustrate some basic properties of the former.

The single-class, single-pool call center is modeled as an $M(\Lambda)/M(\mu)/N+M(\theta)$ queue which, with perfectly predictable rate $\Lambda = \lambda$, is often referred to as the Erlang-A model. We assume that the service rate μ and the patience rate θ are known, but the arrival rate Λ is stationary but uncertain to the extent that we know the average arrival rate $\lambda = E[\Lambda]$ and we have an estimate of the distribution of Λ beyond its mean. The QoS constraint that we consider is one that limits the steady-state fraction of abandoning calls. If the arrival rate is perfectly predictable and equal to a constant $\lambda > 0$, the corresponding formulation is to minimize the number of agents, N , subject to the constraint that the long-run fraction of customers who abandon is at most α . Formally, we

would be looking for N^* such that

$$N^* = \min\{N \in \mathbb{Z}_+ : \lambda a(N, \lambda) \leq \alpha \lambda\},$$

where $a(N, \lambda)$ is the fraction of abandoning customers in steady state when there are N servers and the arrival rate is λ . In the presence of demand-rate uncertainty, however, the steady-state fraction of abandoning customers is itself a random variable, as different realizations of the demand rate Λ will lead to different abandonment fractions. In this setting, requiring that $a(N, \Lambda) \leq \alpha$ should be interpreted as requiring that the constraint holds for all realizations of Λ , which might be impossible or, at the very least, extremely conservative (and costly). Hence, an alternative formulation that takes into account the randomness of $a(N, \Lambda)$ is needed. A natural approach is to average the fraction of abandonments over the demand-rate distribution and put a constraint on that expected value. That is, the *average constraint* formulation is given by:

$$N^* := \min\{N \in \mathbb{Z}_+ : \mathbb{E}_\Lambda[\Lambda a(N, \Lambda)] \leq \alpha \mathbb{E}_\Lambda[\Lambda]\}, \quad (1)$$

where \mathbb{E}_Λ is the expectation with respect to the distribution of Λ , i.e.,

$$\mathbb{E}_\Lambda[\Lambda a(N, \Lambda)] = \int_0^\infty \lambda a(N, \lambda) dF_\Lambda(\lambda),$$

with $F_\Lambda(\cdot)$ being the cumulative distribution function of Λ . This problem is relatively easy to solve by means of a simulation-based search that finds the lowest feasible staffing level, N^* .

An alternative approach is to use a *chance-constrained* formulation. Here, we pre-specify a *risk level*, δ , for the probability that the constraint $a(N, \Lambda) \leq \alpha$ is violated. The chance-constrained formulation is then given by:

$$N^* := \min\{N \in \mathbb{Z}_+ : \mathbb{P}_\Lambda(\{a(N, \Lambda) \leq \alpha\}) \geq 1 - \delta\}, \quad (2)$$

where $\mathbb{P}_\Lambda(\{a(N, \Lambda) \leq \alpha\}) = \int_0^\infty 1\{a(N, \lambda) \leq \alpha\} dF_\Lambda(\lambda)$. The chance-constrained formulation is straightforward to solve for the Erlang-A queue. Let $\lambda^* := \inf\{\lambda \geq 0 : \mathbb{P}\{\Lambda \leq \lambda^*\} \geq 1 - \delta$ and

$$N(\lambda^*) := \inf\{N \in \mathbb{Z}_+ : a(N, \lambda^*) \leq \alpha\},$$

be the minimal staffing level required to satisfy the abandonment constraint when the arrival rate is λ^* . Since, with all the other parameters fixed, the abandonment rate is increasing in the arrival rate we have that $N(\lambda^*)$ is the optimal solution for the chance-constrained formulation (2).

There is a conceptual difference between the two formulations discussed above. We now illustrate this by looking at an example with specific numbers. Within the above Erlang-A setting, set

$\mu = \theta = 1$. Assume that the arrival rate, Λ , is normally distributed with $E[\Lambda] = Var[\Lambda] = 100$. We then vary α between 1% and 10%. For each value of α we compute the optimal staffing level in (1), $N(\alpha)$, via a simulation-based search. We then calculate the risk level implied by this staffing level, i.e, we calculate $\mathbb{P}_\Lambda (\{a(N(\alpha), \lambda) > \alpha\})$. This quantity is the fraction of realizations in which the abandonment constraint $a(N(\alpha), \lambda) \leq \alpha$ is violated when $N(\alpha)$ is used for staffing. The results are displayed in Table 1. The table shows that, while the constraint is met on average (as is expected by the definition of $N(\alpha)$), there is a significant portion of realizations on which the abandonment constraint is violated. To interpret this result, consider a call center in which the performance is measured daily. For such a call center, Table 1 would imply that the abandonment constraint is violated on more than 30% of the days in each month.

Moreover, on days in which the constraint is violated, the violation is not necessarily negligible. In the third row of the table, we calculate for each α (and corresponding $N(\alpha)$) the relative positive error $\mathbb{E}_\Lambda \left[\frac{(a-\alpha)^+}{\alpha} \right]$. Evidently, the relative errors are not negligible. In other words, under the average constraint formulation, not only will the constraint be violated on a significant fraction of the time intervals in consideration, the violation can be also significant.

| Target α | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.1 |
|-------------------------|------|------|------|------|------|------|------|------|------|------|
| Risk level | 0.25 | 0.28 | 0.31 | 0.3 | 0.34 | 0.38 | 0.34 | 0.38 | 0.4 | 0.39 |
| Relative positive error | 41% | 36% | 33% | 29% | 26% | 30% | 23% | 24% | 22% | 20% |

Table 1: Results under average constraint formulation

It is important to emphasize that this example *does not* imply necessarily that the average constraint formulation is inadequate for call centers, only that its adequacy should be judged with respect to the way the performance is measured in the call center. If, for example, the performance is measured only on a monthly average basis, then the average constraint would be appropriate. However, if the performance is measured over significantly shorter time intervals, using an average constraint is inadequate. Rather, the management of the call center should be able to pick its risk-level based on how it measures the performance. It can choose to be conservative by choosing small values for δ or to keep staffing costs lower by selecting a higher value of δ . Clearly, the smaller the risk level is, the higher is the staffing that will be required. The chance-constrained formulation lets the manager tradeoff these two attributes by choosing the risk level that is most appropriate for the call center.

3 Related Literature

The call center workforce management process has gained significant attention in the literature. For a detailed literature review on the subject we refer the reader to the two survey papers [1] and [22]. We focus here on two portions of that literature that are most relevant for our current work. The first is the one dealing with staffing and routing when arrival rates are perfectly predictable. The other stream is more recent and deals with staffing under uncertain arrival rates.

The problem of staffing multi-class multi-type call centers is notoriously hard even when demand rates are perfectly predictable. Consequently, most of the solutions in the literature are based on approximations of various types, simulation-based algorithms, or combinations of the two. In the context of approximations, there are numerous papers that use many-server approximation to solve the staffing problem. Pioneering this stream of literature is [10] which considers staffing the single-class single-pool $M/M/N$ queue under various types of costs and constraints while making use of many-server approximations to simplify that decision. That work was extended to the model with abandonments (the $M/M/N + G$ queue) in [32].

A multi-class (but single-pool) model of a call center, often referred to as the V model, is considered in [24]. That paper proposes a threshold priority policy together with a corresponding $M/M/N$ -based staffing rule. [2] and [3] consider the symmetric model in which a single customer class is served by multiple server classes, known as the inverted-V (or \wedge) model of Skill-Based-Routing (SBR). More recently [25] and [21] proposed the Fixed Queue Ratio (FQR) family of routing rules together with a corresponding staffing rule for much more general SBR systems. Under certain conditions, the FQR rule provides an asymptotically optimal solution. In general, it provides a means to construct good feasible solutions. [20, 21] propose a simulation-based optimization engine that utilizes FQR (or its waiting-time counterpart FWR) together with a Stochastic Approximations algorithm to solve the problems for which [25] does not provide closed form solutions. A simulation-based approach to staffing and scheduling is also introduced in [4, 13].

The approximations literature also includes numerous papers that focus on routing for given staffing under different objectives; see e.g. [15] and the references therein. The results of these papers are important in simplifying the search for good staffing solutions. Most notably, some of the routing rules that are proposed in the approximations literature have the desirable property of being independent of the arrival rate. This is an especially appealing property in our context of uncertain arrival rates; see our discussion of admissible routing rules in §4.

Different approaches to call arrival forecasting have been proposed in the literature. Some

relevant papers that focus on forecasting are [11, 42, 48, 41, 43, 46]. We will assume that the forecasting procedure provides a point estimate in addition to a distribution estimate for the forecast error around the point estimate; such forecasting procedures are developed and tested, for example, in [42] and [48]. Our paper takes these estimates as given and its performance depends, naturally, on the quality of the distributional estimates.

The importance of taking the uncertainty of arrival rates into account is underscored in [14] which analyzes the effect of arrival rate uncertainty on performance. In our work we assume that the uncertainty model is given and focus only on the optimization of the call center. In that respect most relevant are the works on performance analysis and optimization of (mostly single-class single-pools) call centers under random arrival rates; see e.g. [33, 18]. Most recently, both [23] and [37] both consider a stochastic programming approach to shift scheduling under uncertainty. All of the above focus on a single-class, single-pool call center with the average constraint formulation (see §2).

Some of the recent work on uncertainty uses many-server heavy-traffic approximations. [31] consider the single-class single-type call center with uncertain arrival rates and is, in a sense, an extension of [32] to a setting with uncertain arrival rates. An important contribution to the study of multi-type multi-skill call centers with uncertain demand is made in the sequence of papers [5, 6, 8] and the more recent [7]. All of these use a fluid model approach to provide solutions for the staffing problem when arrival rates are time varying and uncertain. Most relevant among these is [8] that considers constraints on the fraction of abandonments (while the others consider abandonment costs). The formulation used in [8] is the average-constraint formulation (see §2 above) and the solution approach is a fluid-model based procedure that translates the staffing problem—via dualization of the constraints and a fluid approximation—to a newsvendor network problem. The fluid model approach is likely to produce good results in very large call centers in which the demand uncertainty is significant. For call centers of medium size and tight abandonment constraints (of the order of 5%) the fluid model might be too crude.

Our work has several distinguishing features: (a) we use a chance-constrained formulation, (b) our solution approach is applicable for general forms of forecast uncertainty and small bounds on the fraction of abandonments, (c) the solution approach can be used for various quality of service constraints while the fluid model approach seems to be limited to constraints on the average queues or the average fraction of abandonments, and finally (d) our *staffing frontier* approach has the desirable property of reducing the complexity of the staffing problem with arrival rate uncertainty to that of solving the perfectly-predictable case.

It is important to emphasize that, in contrast to [8], we do not cover explicitly time-varying rates. Our framework is extendable to the time varying case under some models of time variation; see §8. In the greatest generality, a practical approach is to apply our solution to each time interval during the day on which the arrival rate can be assumed to be stationary. This, together with other directions for future research, is discussed in §8.

4 Network model and problem formulation

We consider a call center with I customer classes and J server pools. Servers in the same pool have the same skills in terms of the set of customer classes that they are capable of serving. We set $\mathcal{I} = \{1, \dots, I\}$ and $\mathcal{J} = \{1, \dots, J\}$. We model the call center as a *parallel server system*; see Figure 2. In the parallel server system customers go through a single stage of service before departing from the system.

The arrival rate to each customer class is fixed during the time interval on which we analyze the system, i.e, class- i customers arrive according to a stationary Poisson process with rate Λ_i , where $\Lambda = (\Lambda_1, \dots, \Lambda_I)$. When making the staffing decisions the vector Λ is a random variable with known distribution and its actual realization is not known. Our model should be interpreted as focusing on a single time interval within the workforce scheduling process as described in the introduction.

If a customer is not admitted to service immediately upon his arrival (or call) he is queued. Customers from the same class are served in a First Come First Served (FCFS) manner but we allow customers to abandon while waiting in the queue. We model this by assigning exponential patience with rate θ_i for class- i customers. A customer whose patience expires before entering services abandons. We assume that all customers have finite patience so that $\theta_i > 0$ for all $i \in \mathcal{I}$.

Agents of different pools have different skills where a skill corresponds to the subset of the customer classes that an agent can serve. In terms of the network representation, if pool- j agents have the skill to serve class- i customers, an edge connecting class i to pool j will appear and we say that pool j has skill i . We denote by $J(i)$ the set of server pools with skill i and by $I(j)$ the set of skills that pool- j agents have. We let E be the set of edges (i, j) such that pool- j servers can serve class- i customers. In Figure 2, for example, pool-1 has both skills 1 and 2 while pool-2 has all skills, 1 to 3. It should be noted that, for various reasons, we might choose to make some of the edges inactive. Using the standard terminology from queueing-networks literature, if class- i customers can be served by pool- j servers, the pair (i, j) is referred to as an activity. We assume

that the service-time of a class- i customer with a pool- j agent is exponentially distributed with rate μ_{ij} .

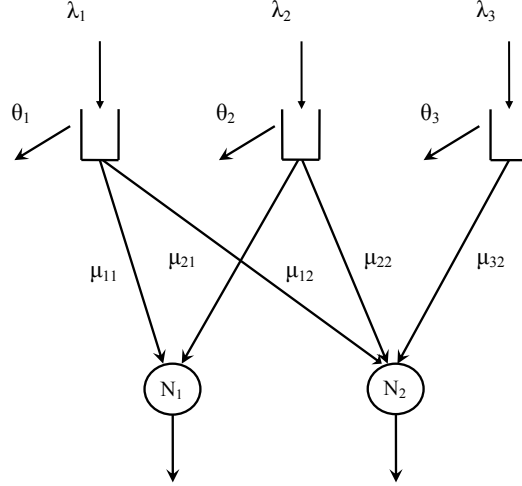


Figure 2: A multi-class multi-pool call center

The basic structure discussed thus far is rather standard to the literature on skill-based routing in many-server parallel server systems; see e.g. [15, 16] and the references therein. Our model is different from these papers in terms of the uncertainty that we associate with the arrival rates Λ .

We assume that we are given a point estimate $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_I)$ for the arrival rates. To avoid trivialities we assume λ is strictly positive. The arrival-rate vector is then a random variable $\Lambda = (\Lambda_1, \dots, \Lambda_I)$, whose mean is that point estimate λ . Specifically, we assume that $\Lambda = \lambda + Z$, where Z is an I -dimensional zero-mean random variable truncated to ensure that Λ obtains only positive values. We use the notation $\mathbb{P}_Z(\cdot)$ for the measure induced by Z , i.e, for a set $\mathcal{B} \in \mathbb{R}^I$, $\mathbb{P}_Z(\mathcal{B}) := \mathbb{P}\{Z \in \mathcal{B}\}$.

The dynamics of the underlying queueing system are a function of its primitives—network structure, service times, patience rates and arrival rates—and of the routing rule that is used to assign servers to customers in real time. A routing rule specifies two decisions: (a) which server should be assigned to an arriving customer, if there are multiple agents that are available and capable of serving that customer, and (b) which customer should be assigned to a newly available server given that there are customers waiting in several of the queues that this agent can serve.

In general, the control rule should be optimized jointly with the staffing levels. As discussed above, even with known arrival rates and fixed staffing levels the optimal control policy is very

difficult to characterize and one may prefer to optimize the staffing when fixing the routing rule. As will be evident in the subsequent sections, our approach can be applied to various control rules. We will use one such routing rule for illustration purposes.

Quality-of-Service constraints: We focus on constraints on the fraction of abandoning customers. Given the risk level $\delta > 0$ and the point estimate λ , our QoS constraint is given by

$$\mathbb{P}_Z \left(z : a_i(\lambda + z, N, \pi) \leq \alpha_i, i \in \mathcal{I} \right) \geq 1 - \delta,$$

where $a_i(\lambda + z, N, \pi)$ is the long run fraction of class- i customers that abandon before being served when the arrival-rate vector is $\lambda + z$, the staffing vector is N and the routing rule is π .¹

Formally,

$$a_i(\lambda, N, \pi) := \limsup_{T \rightarrow \infty} \frac{R_i^T(\lambda, N, \pi)}{A_i^T(\lambda)},$$

where $R_i^T(\lambda, N, \pi)$ is the number of customer abandonments from queue i by time T under the triplet (λ, N, π) and $A_i^T(\lambda)$ is the number of class- i calls by time T when the arrival rate is λ . When the routing rule π is clear from the context we will omit it and use $a_i(\lambda, N)$. When the routing rule π admits a steady-state distribution, $R_i^T(\lambda, N, \pi)/A_i^T(\lambda)$ will converge almost surely to a constant. If not, the inequality $a_i(\lambda, N, \pi) \leq \alpha_i$ should be interpreted as

$$P \left\{ \limsup_{T \rightarrow \infty} \frac{R_i^T(\lambda, N, \pi)}{A_i^T(\lambda)} > \alpha_i \right\} = 0, \text{ for all } i \in \mathcal{I}. \quad (3)$$

The staffing and routing problem: Our objective is to minimize the total staffing costs subject to the QoS constraint specified above. We assume that agents of type j incur a salary cost c_j for the time interval in consideration. Hence, given the number of agents $N = (N_1, \dots, N_J)$ the operational cost is given by $c \cdot N := \sum_{j \in \mathcal{J}} c_j N_j$. Clearly, the elements of N must be integers.

Our optimization problem is then given by:

$$\begin{aligned} \min \quad & c \cdot N \\ \text{s.t.} \quad & \mathbb{P}_Z \left(z : a_i(\lambda + z, N, \pi) \leq \alpha_i, i \in \mathcal{I} \right) \geq 1 - \delta, \\ & N \in \mathbb{Z}_+^J, \pi \in \Pi, \end{aligned} \quad (4)$$

The decision variables in this optimization problem are the staffing level, $N \in \mathbb{Z}_+^J$ and the routing rule $\pi \in \Pi$, where Π is the family of admissible routing rules which we will define shortly. It is

¹In (4) we use a joint chance constraint that requires all service targets to met simultaneously with high probability. An alternative is to introduce individual chance constraints. Namely, to require that class i has its abandonment target met with a probability of at least $1 - \delta_i$. In this paper we consider only the joint-constraint and our solution approach is tailored to that formulation.

simple to construct a feasible solution (π, N) for (4), for example, by fixing some non-idling policy and taking N to be sufficiently large for that policy. In particular, the set of feasible solutions is non-empty. Moreover, as N obtains only integer values, the infimum cost over the feasible family must be attained as a minimum for some pair (π^*, N^*) .²

The problem (4) can be further simplified by truncating the support of Λ . Indeed, (4) shares its optimal solution with the following optimization problem for sufficiently large b :

$$\begin{aligned} \min \quad & c \cdot N \\ \text{s.t.} \quad & \mathbb{P}_Z \left(z : \max_i z_i \leq b, a_i(\lambda + z, N, \pi) \leq \alpha_i, i \in \mathcal{I} \right) \geq 1 - \delta, \\ & N \in \mathbb{Z}_+^J, \pi \in \Pi. \end{aligned} \tag{5}$$

We provide a proof of this claim in the online supplement (see § EC1), but the intuition is simple: given an upper bound on the optimal solution to (4) (obtained by any feasible solution), b can be chosen large enough so that if z is such that $z_i > b$ for some i , then *any* staffing vector N that satisfies $a_i(\lambda + z, N, \pi) \leq \alpha_i$ will necessarily be costlier than the upper bound.

We end this section with a definition of the family, Π , of admissible routing rules. Our discussion below is somewhat informal but sufficient for the purposes of this paper.

A routing rule is a process $\{U(t), t \geq 0\}$ with $U(t) := (r_{ij}(t), s_{ij}(t); i \in \mathcal{I}(j), j \in \mathcal{J})$ such that a class- i arrival at time t will be routed to pool j only if $r_{ij}(t) = 1$. Similarly, a service completion at time t in pool j is followed by an admission of a class- i customer to service only if $s_{ij}(t) = 1$.

The routing rule may depend on the system primitives—service times, arrival rates, patience rates, staffing level, network design and QoS constraints—as well as on the evolution of the system, which is captured at time t by some stated descriptor $X(t)$. Slightly informally, we would have that $U(t) = f(\mu, \theta, N, E, (X(s), 0 \leq s < \infty))$ where f is some function and, as before, E is the set of edges in the network graph. We say that a routing rule π *has no a priori knowledge of the arrival rates* if $U(t)$ is not a function of the arrival rate vector or, in other words, λ is not an argument of the function $f(\cdot)$ above. Also, following standard convention, we say that π is *non-anticipative* if it does not depend on the future evolution of the system. Combined, these imply that $U(t) = f(\mu, \theta, N, E, (X(s), 0 \leq s < t))$. Finally, we say that π is a *monotone routing rule* if, fixing a staffing vector N , and two arrival-rate vectors λ^1 and λ^2 such that $\lambda^1 \leq \lambda^2$ componentwise, we have that $a_i(\lambda^1, N, \pi) \leq a_i(\lambda^2, N, \pi)$.

²The proof of the existence of such a minimizer is very similar to that of Lemma EC1.1 in the e-companion. The set of staffing vectors N for which a feasible routing rule exists is a discrete and non-empty set that can be truncated to be finite without compromise in cost. Thus, an optimal solution N^* is guaranteed to exist and π^* is taken to be any policy in Π that makes (4) feasible – such a π^* must exist otherwise N^* would not be feasible.

Definition 4.1 (admissible routing rules) We say that a routing rule π is admissible if (i) it is non-anticipative, (ii) it is monotone, and (iii) it has no a priori knowledge of the arrival rates realization.

Definition 4.1 is a very natural one. The fact that π does not see the arrival rates is not restrictive. It requires that the router does not have a priori knowledge of the arrival rates, but it allows for learning of the arrival rate by using, at time t , the information about the evolution of the system as reflected in the process $(X(s), 0 \leq s < t)$. As examples of routing rules that are admissible by our Definition 4.1 we mention the routing rule in [8] that uses learning of the demand rate, the routing rule used in [15], and the FWR routing rule that was proposed in [25] and that we use for illustration purposes in §7.

Outline of the solution procedure: Our proposed procedure consists of two main phases.

1. **A Random Static Planning Problem:** We formulate a chance-constrained optimization problem that provides a first-order approximation for (4). This optimization problem, which we refer to as the *RSPP*, can be regarded as a fluid approximation of the original staffing problem. The RSPP is, in a sense, a random version of the *static-planning problem* that is often used in the queueing network literature. The RSPP yields three outputs: (i) a lower bound on the optimal staffing cost in (4), (ii) a staffing vector, \bar{N} , that corresponds to that lower bound, and finally (iii) a set of arrival rate vectors, \mathcal{F} , referred to as a *staffing frontier*. The RSPP and approaches towards solving it are covered in §6
2. **Simulation-based search on the RSPP frontier:** Fixing a routing rule and using the staffing vector from the RSPP as our starting point, we perform a simulation-based search for a staffing solution that is feasible to (4). The simulation searches for a staffing vector that is, in a sense, close to the RSPP staffing vector, and one that satisfies the abandonment constraints for each of arrival-rate vectors on the RSPP frontier. Namely, we will search for a staffing vector N such that $a_i(\lambda, N, \pi) \leq \alpha_i$ for all $\lambda \in \mathcal{F}$ with \mathcal{F} being the staffing frontier that we obtained in the first step. The staffing vector that is found by this search is shown to be feasible for the original staffing problem (4). This procedure is the subject of §7.

Before laying out the two steps above we discuss an idealized setting that serves to highlight some of the complexities associated with solving (4) directly.

5 When solutions are known for the predictable case–inherited optimality

In this section we focus on the (ideal) case in which one has an oracle that, given a staffing level N , an arrival rate vector λ and a routing rule π can detect whether the QoS constraints are met. We show that, when such an oracle is available, one can construct nearly optimal solutions for the case with uncertain rates by means of a discretized chance-constrained problem, with the optimality gap being a function of the discretization resolution. The discussion of this idealized scenario serves three purposes: (a) it illustrates the close relationship between the predictable case and uncertain one, (b) the need for an oracle underscores the difficulty in solving the staffing problem with uncertainty and motivates the search for alternative procedures that do not impose such a strict requirement, and finally (c) the discretization approach prepares the ground for the alternative procedure which is the subject of the rest of the paper.

We start by constructing a discretized version of (5), the truncated version of (4). To that end, let $\mathcal{A}^b := \{x \in \mathbb{R}_+^I : \max_i x_i \leq b\}$ where b is the truncation constant from (5). We define a parameter Δ which we refer to as the *resolution* of the approximation and assume, without loss of generality, that $b = L\Delta$ for some $L > 0$ (otherwise, we can always increase b). We then divide the region \mathcal{A}^b into regions as follows: let $\mathcal{L}(\Delta) = \{k \in \mathbb{Z}_+^I : k_i < L \text{ for all } i \in \mathcal{I}\}$ so that $|\mathcal{L}(\Delta)| = L^I$ and define

$$\begin{aligned} A_k &:= \{x \in \times_{i=1}^I [\Delta \cdot k_i, \Delta \cdot (k_i + 1))\}, k \in \mathcal{L}(\Delta), i \in \mathcal{I}, \\ A_\infty &:= \{x \in \mathbb{R}_+^I : \max_i x_i \geq b\}. \end{aligned}$$

The sets A_k are I -dimensional hypercubes with edge-length Δ that partition the region $[0, b]^I$, and A_∞ covers the remainder of \mathbb{R}_+^I . Next, for $k \in \mathcal{L}(\Delta)$, let $\lambda_i(k) = \Delta(k_i + 1)$ for $i \in \mathcal{I}$ and $\lambda(k) := (\lambda_1(k), \dots, \lambda_I(k))$. In addition, let $p_k = \mathbb{P}_Z(\Lambda \in A_k)$ for $k \in \mathcal{L}(\Delta)$ and for $k = \infty$. We then define a discrete random vector $\hat{\Lambda}$ by letting

$$\mathbb{P}_{\hat{\Lambda}}(\hat{\Lambda} = \lambda(k)) = p_k \text{ for } k \in \mathcal{L}(\Delta) \text{ and } \mathbb{P}_{\hat{\Lambda}}(\hat{\Lambda} = \lambda(\infty)) = p_\infty, \quad (6)$$

where $\lambda(\infty)$ is an arbitrary point in A_∞ .

To relate the staffing and routing to the service level constraint, we define the function

$$g(\lambda, N, \pi) := \begin{cases} 0 & \text{if } \max_{i \in \mathcal{I}} (a_i(\lambda, N, \pi) - \alpha_i) \leq 0, \\ 1 & \text{otherwise.} \end{cases} \quad (7)$$

In other words, $g(\cdot, \cdot, \cdot)$ is a function that—given an arrival-rate vector λ and a routing rule π —identifies whether or not the staffing vector N is feasible with respect to the service-level constraints. Following the key steps in §6 we then define the following discretized version of (4).

$$\begin{aligned}
\min \quad & c \cdot N \\
\text{s.t.} \quad & y_k g(\lambda(k), N, \pi) \leq 0, \quad k \in \mathcal{L}(\Delta), \\
& \sum_{k \in \mathcal{L}(\Delta)} y_k p_k \geq 1 - \delta, \\
& y_k \in \{0, 1\} \quad k \in \mathcal{L}(\Delta), \\
& N \in \mathbb{Z}_+^J, \pi \in \Pi.
\end{aligned} \tag{8}$$

To solve (8) we must be able to compute the function $g(\lambda, N, \pi)$ by some means, which requires that we must identify a *function* not just a number. In the single-class, single-pool $M/M/N + M$ queue of §2, the routing rule is $\pi = FCF S$ and $g(\cdot, \cdot, \cdot)$ is given by $g(\lambda, N, \pi) = 0$ if $N \geq N^*(\lambda)$ and otherwise $g(\lambda, N, \pi) = 1$ where $N^*(\lambda) = \min\{N \in \mathbb{Z}_+ : a(N, \lambda) \leq \alpha\}$.

For the following theorem, let $\bar{c} := \max_{j \in \mathcal{J}} c_j$.

Theorem 5.1 *Fix $\Delta > 0$. Let (N^*, π^*) and $(\check{N}(\Delta), \pi(\Delta))$ be an optimal solution to (5) and (8) respectively. Then, there exists a constant $C > 0$ (independent of Δ and α) such that*

$$|c \cdot \check{N}(\Delta) - c \cdot N^*| \leq \bar{c}I \vee C\Delta. \tag{9}$$

We note that the right hand side in (9) can be as large as a constant times the number of customer classes. Ideally, one would like the difference to decrease with Δ to 0. This, however, is not possible because of the integrality of the staffing levels. There are settings in which increasing the demand by Δ for each class might require the addition of one server for each class.

The proof of Theorem 5.1 appears in the e-companion to this paper. A key step in the proof is to establish linear-growth of the optimal cost (5) in the arrival rate vector, i.e, that a shift of the whole demand distribution by Δ does not increase the optimal cost by more than $C\Delta$. The main challenge in establishing this property is that the optimal routing rule π^* is abstract rather than specifically given. To overcome this difficulty, the proof follows a constructive argument, in which, given an optimal solution to the staffing problem (5) we construct a feasible solution for (8) with a cost that is, at most, $\bar{c}I \vee C\Delta$ higher than the optimal cost. In that construction we make explicit use the fact that π^* is admissible in the sense of Definition 4.1 and is hence monotone.

The requirement in (8) that we have an available characterization of the mapping $g(\cdot, \cdot, \cdot)$ is a significant restriction. To be able to handle general cases, we introduce in the next section the

Random Static Planning Problem (RSPP) and analyze its properties. The RSPP will be a powerful tool in constructing simple and feasible solutions for (4) while maintaining reasonable optimality gaps.

6 The random static planning problem (RSPP)

When the arrival rates are perfectly predictable, a so-called *static-planning problem* (SPP) is often used to provide first-order approximations for the optimal staffing levels and allocations of customer classes to agent pools; see e.g. §4 of [25]. Specifically, given the arrival rate vector $\lambda = (\lambda_1, \dots, \lambda_I)$, the SPP is given by³

$$\begin{aligned} \min \quad & c \cdot N \\ \text{s.t.} \quad & \sum_{j \in \mathcal{J}(i)} \mu_{ij} \nu_{ij} \geq \lambda_i (1 - \alpha_i), \quad i \in \mathcal{I}, \\ & \sum_{i \in \mathcal{I}(j)} \nu_{ij} \leq N_j, \quad j \in \mathcal{J}, \\ & N \in \mathbb{R}_+^J, \quad \nu \in \mathbb{R}_+^{IJ}. \end{aligned} \tag{10}$$

The quantity ν_{ij} can be thought of as the long-run number of servers in server pool j that are allocated to serve class i customers. The objective of the SPP is to minimize the total staffing costs subject to meeting the (approximate) service level target. The SPP ignores the effect of all the randomness caused by inter-arrival and service times by only considering a deterministic (fluid) version of the actual queueing system.

Since we allow for up to a fraction α_i of class- i customers to abandon, we require, in first-order, that the capacity of the system be only sufficient to serve $\lambda_i(1 - \alpha_i)$ customers from class i per unit of time. Our use of the term static planning problem (SPP) for problem (10) is somewhat non-standard; the commonly used SPP only allocates the fluid input between pools with pre-specified staffing levels while (10) also optimizes the staffing levels. With a slight abuse of terminology we refer to this as the *deterministic static-planning problem* to distinguish it from the random version that we discuss next.

For the case of random arrival rates we construct the following *Random Static-Planning Problem* (RSPP):

$$\begin{aligned} \min \quad & c \cdot N \\ \text{s.t.} \quad & \mathbb{P}_Z(\Lambda \in \mathcal{B}(N)) \geq 1 - \delta, \\ & N \in \mathbb{R}_+^J, \end{aligned} \tag{11}$$

³This SPP is slightly different than the one in [25] and takes explicitly into account the permissible abandonment fractions α_i .

where

$$\mathcal{B}(N) := \left\{ \lambda \in \mathbb{R}_+^I : \exists \nu \in \mathbb{R}_+^{I \times J} \text{ with } \sum_{j \in \mathcal{J}(i)} \mu_{ij} \nu_{ij} \geq \lambda_i (1 - \alpha_i), i \in \mathcal{I}, \sum_{i \in \mathcal{I}(j)} \nu_{ij} \leq N_j, j \in \mathcal{J} \right\}. \quad (12)$$

To simplify exposition of the results, we will assume throughout the paper that an optimal solution to (11) exists. The optimization problem (11) is a chance-constrained optimization problem that requires the staffing level N to be such that $\Lambda \in \mathcal{B}(N)$ with probability at least $1 - \delta$. The problem (11) is a *linear* chance-constrained optimization problem. Furthermore, it is a two-stage problem: for a solution N to be feasible for (11), we require that, with probability $1 - \delta$, there exists a (second stage) *recourse* solution which satisfies the linear constraints parameterized by N .

Chance-constrained optimization problems have been studied extensively in the optimization literature, see e.g. [35]. The difficulty in such problems comes from the fact that, in general, the feasible region defined by a chance constraint is not convex. Moreover, evaluating the probability $\mathbb{P}_Z(\Lambda \in \mathcal{B}(N))$ for a candidate solution N can be, in itself, computationally challenging.

Significant progress has been made in handling chance constraints with particular structure, most notably when the randomness is restricted to the right-hand side of the constraint, i.e. when the chance constraint takes the form $\mathbb{P}(Tx \geq \xi) \geq 1 - \delta$, where x is the decision vector, T is a deterministic matrix and ξ is a random vector; see e.g. [35, 17, 30, 26]. Unfortunately, the formulation (11) of the RSPP is a two-stage chance-constrained optimization problem, and the research into such problems is relatively scarce. Notable exceptions are [39]—in which a finite discrete distribution is assumed and the resulting mixed-integer programming formulation is strengthened using precedence constraints—as well as [12], [36], [19] and [34] which consider *conservative* sample approximations of general chance-constrained problems (including the two-stage case). Recent approaches for finding exact solutions to more general chance-constrained problems having a discrete distribution include [28, 45].

Intuitively, a solution of the RSPP identifies, in addition to the staffing vector N , an “optimal” subset of the support of Λ , having probability mass at least $1 - \delta$, for which a feasible recourse decision (with respect to N) exists. It might seem initially that this can be achieved by optimally selecting a single point $\lambda' \in \mathbb{R}_+^I$ such that $\mathbb{P}_Z(\Lambda \leq \lambda') \geq 1 - \delta$ and then requiring a feasible recourse decision to exist for λ' (and hence also for all $\lambda \leq \lambda'$). While this naive approach would result in a feasible solution for the RSPP, limiting attention to solutions generated this way is overly restrictive as the optimal subset will generally not have the shape $\{\lambda : \lambda \leq \lambda'\}$ for some λ' . Figure 5, which depicts the optimal frontier for Example 6.1, will illustrate this point.

We will use the RSPP to guide our search for solutions to (4). It is also useful that the RSPP provides a lower bound on the optimal value of (4), and hence can be used to establish the quality of a proposed solution.

Theorem 6.1 *Let (π^*, N^*) be an optimal solution to (4) and suppose that the RSPP (11) has an optimal solution with value z_{RSPP} . Then*

$$z_{RSPP} \leq c \cdot N^*. \tag{13}$$

Theorem 6.1 is proved by showing that the constraints in the RSPP are necessary constraints for the original staffing problem (4). In turn, any staffing vector N that is feasible for (4) is necessarily feasible for (11).

We emphasize that the assumption of existence of an optimal solution to the RSPP is made for purposes of simplifying the exposition and all subsequent statements in the paper can be made with respect to the infimum over all feasible solutions rather than with respect to the optimum, if such an optimum does not exist.

Our approach to solving the RSPP is to use a discrete approximation of the random variable Λ , after which the RSPP can be formulated as a mixed-integer program (MIP) and solved by an off-the-shelf MIP solver. We consider two approaches for generating the discrete approximation. The approach in §6.1 uses a fixed grid and has the advantage of giving a deterministic a priori bound on the approximation error, and is computationally viable for call centers with a small number of customer classes. The approach in §6.2 uses Monte Carlo sampling to generate the discrete approximation and can be used to find feasible solutions regardless of the number of customer classes, but provides statistical error bounds as opposed to the deterministic guarantees obtained using the fixed-grid approach.

As discussed in the introduction, the output of the RSPP will also include a set of arrival rate vectors (a staffing frontier) that will we will subsequently use in our procedure for obtaining a solution to the original problem (4). In §6.3 we give the formal definition of a staffing frontier, and specify how we obtain such a frontier from either the fixed grid or sample-based approximations.

6.1 Fixed grid approximation

In many practical applications, the number of customer classes is small. In such cases, a fixed grid can be employed to yield an approximate solution to RSPP.

The fixed grid discretization is identical to the one in §5. First, as in the discussion following (4) we may, without loss of generality, truncate the distribution to a subset $[0, b]^I$ of \mathbb{R}_+^I , so that we can replace $\mathcal{B}(N)$ in RSPP with $\mathcal{B}(N) \cap \{\lambda \in \mathbb{R}_+^I : \max_i \lambda_i \leq b\}$. With a slight abuse of notation we use the notation $\mathcal{B}(N)$ for this intersection. The constant b might be different from the one in (5) but we can just take the largest of the two constants and use it for both the staffing problem and the RSPP. We fix the resolution $\Delta > 0$ and define the set $\mathcal{L}(\Delta)$ as before. We then approximate the RSPP with the following discrete version (D-RSPP):

$$\begin{aligned} \min \quad & c \cdot N \\ \text{s.t.} \quad & \mathbb{P}_{\hat{\Lambda}}(\hat{\Lambda} \in \mathcal{B}(N)) \geq 1 - \delta, \\ & N \in \mathbb{R}_+^J, \end{aligned} \tag{14}$$

where $\mathbb{P}_{\hat{\Lambda}}(\cdot)$ is as defined in (6).

Formulation as a Mixed-Integer Program (MIP): A chance-constrained optimization problem can be formulated as a mixed-integer program when the underlying distribution has finite support, see e.g. [39]. Following this approach, we exploit the finite support of $\hat{\Lambda}$ to formulate (14) as a mixed-integer program. To do so, we introduce binary variables y_k for $k \in \mathcal{L}(\Delta)$ where $y_k = 1$ indicates that $\lambda(k) \in \mathcal{B}(N)$. This leads to the following MIP formulation of D-RSPP:

$$\begin{aligned} \min \quad & c \cdot N \\ \text{s.t.} \quad & \sum_{j \in \mathcal{J}(i)} \nu_{ij}^k \geq y_k \lambda_i(k)(1 - \alpha_i), \quad i \in \mathcal{I}, k \in \mathcal{L}(\Delta), \end{aligned} \tag{15a}$$

$$\sum_{i \in \mathcal{I}(j)} \nu_{ij}^k \leq N_j, \quad j \in \mathcal{J}, k \in \mathcal{L}(\Delta), \tag{15b}$$

$$\sum_{k \in \mathcal{L}(\Delta)} p_k y_k \geq 1 - \delta, \tag{15c}$$

$$N \in \mathbb{R}_+^J, y_k \in \{0, 1\}, \nu^k \in \mathbb{R}_+^{JJ}, k \in \mathcal{L}(\Delta).$$

The constraints (15a) and (15b) ensure that $\lambda(k) \in \mathcal{B}(N)$ if $y_k = 1$. The constraint (15c) ensures that N is such that $\lambda(k) \in \mathcal{B}(N)$ for sufficiently many points so that the chance constraint is met.

Next, we propose several additions to the initial formulation of D-RSPP which will improve its computational efficiency.

Adding dominance constraints: For a given N , if $\lambda(k) \in \mathcal{B}(N)$ then $\lambda(l) \in \mathcal{B}(N)$ for any $\lambda(l) \leq \lambda(k)$. Thus, we can strengthen this formulation by adding the inequalities $y_l \geq y_k$ for all $l, k \in \mathcal{L}(\Delta)$ with $\lambda(l) \leq \lambda(k)$. Moreover, using the grid structure of our discrete distribution

we can obtain the same effect with many fewer inequalities by adding the inequality only for immediate neighbors. To that end, we define $d(k)$ to be the index set of the smaller “immediate neighbors” of cell k . Formally,

$$d(k) := \{l \in \mathcal{L}(\Delta) : \exists j \text{ s.t. } \lambda_i(l) = \lambda_i(k) \forall i \neq j \text{ and } \lambda_j(l) = \lambda_j(k) - \Delta\}.$$

The number of elements in $d(k)$ is at most I . We then add to D-RSPP the constraints $y_l \geq y_k$ only for $l \in d(k)$ for each k . It is clear that these constraints are as strong as having the constraints $y_l \geq y_k$ for all $l, k \in \mathcal{L}(\Delta)$ with $\lambda(l) \leq \lambda(k)$. Indeed, if $\lambda(l) \leq \lambda(k)$ there will exist a sequence of points between $\lambda(l)$ and $\lambda(k)$ which are immediate neighbors of each other, whose corresponding inequalities which we do enforce would imply $y_l \geq y_k$.

Fixing and removing variables: It is also possible to a priori fix some of the binary variables y_k , $k \in \mathcal{L}$ to 1 and use this to remove some of the second-stage variables ν . To do so, define

$$\mathcal{G} := \left\{ k \in \mathcal{L}(\Delta) : \sum_{l \in \mathcal{L}(\Delta) : \lambda(l) \geq \lambda(k)} p_l > \delta \right\}.$$

A lattice point is in the set \mathcal{G} if and only if the probability that a random arrival rate dominates it is larger than δ . Thus, if a point $k \in \mathcal{G}$ is not covered, the chance constraint cannot hold, and hence we can fix $y_k = 1$. (Formally, if $y_k = 0$, then $y_l = 0$ for all $l \in \mathcal{L}(\Delta)$ with $\lambda(l) \geq \lambda(k)$ and hence the constraint (15c) cannot be satisfied.) Thus, we only need binary variables y_k for $k \in \mathcal{D} := \mathcal{L}(\Delta) \setminus \mathcal{G}$, and for $k \in \mathcal{G}$ we can set $y_k = 1$ and replace the constraints (15a) with

$$\sum_{j \in \mathcal{J}(i)} \nu_{ij}^k \geq \lambda_i(k)(1 - \alpha_i), \quad i \in \mathcal{I}. \quad (16)$$

Hence, we can remove the variables y_k for $k \in \mathcal{G}$ and replace the inequality (15c) with:

$$\sum_{k \in \mathcal{D}} p_k y_k \geq 1 - \delta - \gamma,$$

where $\gamma := \sum_{k \in \mathcal{G}} p_k$. Finally, let \mathcal{E} be the set of nondominated points in \mathcal{G} , i.e.,

$$\mathcal{E} := \left\{ l \in \mathcal{G} : \nexists k \in \mathcal{G} \text{ s.t. } \lambda(k) \geq \lambda(l), \lambda(k) \neq \lambda(l) \right\}.$$

We can then eliminate the variables ν_{ij}^l together with the constraints (16) and (15b) for all $l \in \mathcal{G} \setminus \mathcal{E}$. Indeed, for each $l \in \mathcal{G} \setminus \mathcal{E}$, there will be a point $k \in \mathcal{E}$ with $\lambda(k) \geq \lambda(l)$ for which the variables ν_{ij}^k and constraints (16) and (15b) will be in the model, and because $\lambda(k) \geq \lambda(l)$ these imply feasibility of the corresponding constraints for the point l .

Final improved formulation: Combining the above observations we arrive at the following improved MIP formulation of D-RSPP which we refer to as MIP-RSPP.

$$\begin{aligned}
& \min c \cdot N \\
& \text{s.t.} \quad \sum_{j \in \mathcal{J}(i)} \mu_{ij} \nu_{ij}^k \geq y_k \lambda_i(k)(1 - \alpha_i), \quad i \in \mathcal{I}, k \in \mathcal{D}, \tag{17a} \\
& \quad \sum_{j \in \mathcal{J}(i)} \mu_{ij} \nu_{ij}^k \geq \lambda_i(k)(1 - \alpha_i), \quad i \in \mathcal{I}, k \in \mathcal{E}, \tag{17b} \\
& \quad y_l \geq y_k, \quad l \in d(k), k \in \mathcal{D}, \tag{17c} \\
& \quad \sum_{i \in \mathcal{I}(j)} \nu_{ij}^k \leq N_j, \quad j \in \mathcal{J}, k \in \mathcal{D} \cup \mathcal{E}, \tag{17d} \\
& \quad \sum_{k \in \mathcal{D}} p_k y_k \geq 1 - \delta - \gamma, \tag{17e} \\
& \quad N \in \mathbb{R}_+^J, y_k \in \{0, 1\}, k \in \mathcal{D}, \nu^k \in \mathbb{R}_+^{IJ}, k \in \mathcal{D} \cup \mathcal{E}.
\end{aligned}$$

MIP-RSPP has an optimal solution because it is a binary mixed-integer program with objective bounded from below. We now provide an estimate of the gap between the RSPP and its approximation in MIP-RSPP. As in Theorem 6.1 we assume the RSPP (11) has an optimal solution with cost z_{RSPP} .

Theorem 6.2 *Fix $\Delta > 0$ and an optimal solution $\bar{N}(\Delta)$ for the MIP-RSPP with resolution Δ . Then, $\bar{N}(\Delta)$ is feasible for the RSPP and there exists a constant C (independent of Δ and of $\bar{N}(\Delta)$) such that*

$$c \cdot \bar{N}(\Delta) \leq z_{RSPP} + C\Delta.$$

The proof of Theorem 6.2 is rather intuitive. Roughly speaking, compared to the RSPP, the MIP-RSPP has a demand distribution which is shifted up by (at most) Δ . Because of the linear constraints within the chance constraint one expects that shifting the RSPP solution by some (carefully chosen) multiple of Δ will produce a feasible solution for MIP-RSPP.

We discuss the computational properties of MIP-RSPP in §6.4. First, in the next section we introduce a different solution approach to the RSPP.

6.2 Sample-based approximation

Relying on [29], we use Monte Carlo sampling to obtain a discrete approximation of Λ . In particular, we generate an independent sample of size K from the distribution of Λ . Let this sample be $\lambda(1), \dots, \lambda(K)$ and assign a probability mass of $1/K$ to each point (as in constructing

the empirical distribution). Each of these sample points is itself a vector of dimension I , i.e., $\lambda(k) = (\lambda_1(k), \dots, \lambda_I(k))$. Then, as in the fixed-grid approximation, we construct a MIP formulation of the approximate problem by introducing binary variables y_k where $y_k = 1$ will indicate that $\lambda(k) \in \mathcal{B}(N)$. This leads to the following sample-based formulation which we refer to as **S-RSPP**:

$$\hat{z}^K := \min c \cdot N$$

$$\text{s.t. } \sum_{j \in \mathcal{J}(i)} \mu_{ij} \nu_{ij}^k \geq y_k \lambda_i(k) (1 - \alpha_i), \quad i \in \mathcal{I}, k = 1, \dots, K, \quad (18a)$$

$$\sum_{i \in \mathcal{I}(j)} \nu_{ij}^k \leq N_j, \quad j \in \mathcal{J}, k = 1, \dots, K, \quad (18b)$$

$$\sum_{k=1}^K y_k \geq K(1 - \delta), \quad (18c)$$

$$N \in \mathbb{R}_+^J, y_k \in \{0, 1\}, \nu^k \in \mathbb{R}_+^{IJ}, k = 1, \dots, K.$$

Because the resulting optimal value \hat{z}^K and the corresponding staffing level \hat{N}^K are a function of the sample, they are *random*. Given the sample, the above formulation is almost identical to the MIP-RSPP that we introduced for the fixed grid approximation with the exception that we do not include the dominance constraints because the sample points $\lambda(k)$ no longer have a grid structure, and hence the concept of immediate neighbors does not apply. (A dominance constraint could still be included between pairs of realizations where one dominates the other, but we did not find this beneficial in our experiments.)

The sample approximation approach is attractive because, under mild assumptions, the sample size that is required for a reasonable approximation grows only linearly with the dimension of the decision space [29]. The results in [29] can be used to derive a priori estimates on the sample size required to obtain solutions with a desired level of accuracy. However, these a priori estimates are conservative leading to a suggested sample size that is too large for practical use. Thus, we instead use a modified approach that was introduced by [34] and experimented with in [29]. The idea is to solve S-RSPP multiple times, each time with a different sample. Specifically, we generate M different samples, each of a size K , which is fixed in advance. For each sample $r = 1, \dots, M$, we solve the corresponding S-RSPP problem, obtaining an optimal value $\hat{z}^{r,K}$, and solution $\hat{N}^{r,K}$. We then define the value

$$\hat{L}^M = \min\{\hat{z}^{r,K} : r = 1, \dots, M\}, \quad (19)$$

and, as discussed in [29], this value provides a lower bound to the true problem RSPP with probability at least $1 - (1/2)^M$. So, for $M = 10$, \hat{L}^M is a valid lower bound with probability at least

0.999. Because the solutions $\hat{N}^{r,K}$ are based on a random sample, they are not guaranteed to be feasible for the original RSPP. In the next section, we will discuss how we use these solutions as a starting point to construct a staffing frontier and corresponding solution that are feasible with high confidence.

6.3 RSPP staffing frontier

We now define a δ -feasible staffing frontier which, along with the approximately optimal RSPP staffing vector \bar{N} , will be useful in constructing staffing solutions for (4). First, for any set of points $\mathcal{F} \subset \mathbb{R}^I$ we introduce the notation

$$\mathcal{M}(\mathcal{F}) = \bigcup_{\lambda' \in \mathcal{F}} \{\lambda \in \mathbb{R}^I : \lambda \leq \lambda'\},$$

which represents the set of points dominated by some point in \mathcal{F} .

Definition 6.3 (a δ -feasible staffing frontier.) *A set of points $\mathcal{F} = \{\lambda(k) : k \in F\}$, where F is a finite index set, is called a δ -feasible staffing frontier if*

$$\mathbb{P}_Z\{\Lambda \in \mathcal{M}(\mathcal{F})\} \geq 1 - \delta,$$

and for every point $\lambda \in \mathcal{F}$ there does not exist another $\lambda' \in \mathcal{F}$ such that $\lambda' \geq \lambda$.

The δ -feasible frontier will be key in our solution for the original staffing problem (4). To check feasibility of a candidate solution (N, π) to (4), one simply checks via simulation whether (3) holds for *all* of the arrival rate vectors in the staffing frontier. If so, the monotonicity requirement of admissible routing rules implies the solution is feasible for the chance constraint in (4); see §7.1. The simulation work required for this verification is proportional to the size of the frontier, and hence it is advantageous to keep the frontier as small as possible. We next discuss how we obtain a δ -feasible staffing frontier from both solution approaches to the RSPP.

Fixed grid approximation. Let (\bar{N}, \bar{y}) be a solution for (17). For such a solution, define \mathcal{F}^* as follows

$$\mathcal{F}^* = \{\lambda(k) \mid k \in \mathcal{E} \text{ or } k \in \mathcal{D} \text{ and } \bar{y}_k = 1\}.$$

Then, the associated δ -feasible staffing frontier is constructed by removing the dominated points in \mathcal{F}^* :

$$\mathcal{F} = \{\lambda \in \mathcal{F}^* \mid \nexists \lambda' \in \mathcal{F}^* \text{ with } \lambda' \neq \lambda, \lambda' \geq \lambda\}.$$

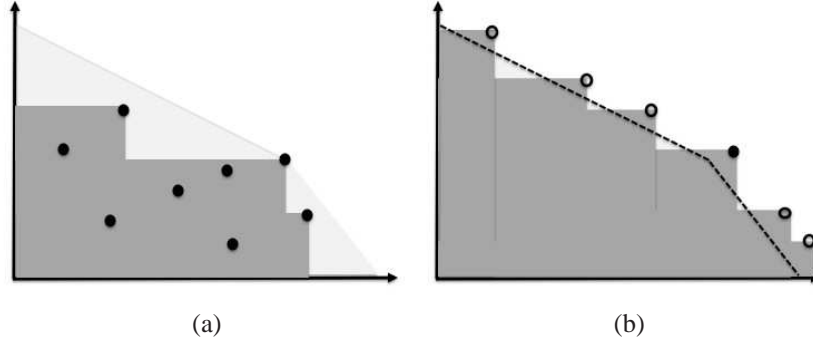


Figure 3: Construction of a staffing frontier from sample approximation solution.

Feasibility of the solution \bar{y} to the inequality (17e) implies that \mathcal{F} is a δ -feasible staffing frontier. In the typical case when δ (in the definition of the chance constraint) is small, the size of the staffing frontier will be much smaller than the size of the discretization used to approximate RSPP, that is, we expect that $|\mathcal{F}| \ll |\mathcal{L}(\Delta)|$.

Sample-based approximation. Constructing a δ -feasible staffing frontier from a solution (\hat{N}, \hat{y}) to the sample approximation S-RSPP requires more care. Consider a feasible solution (\hat{N}, \hat{y}) of the S-RSPP that satisfies the chance constraint:

$$\mathbb{P}(\mathcal{B}(\hat{N})) \geq 1 - \delta. \quad (20)$$

It then seems natural to construct a staffing frontier $\hat{\mathcal{F}}_1$ by including every $\lambda(k)$ such that $\hat{y}_k = 1$, and then discarding the dominated points (similar to the construction based on a solution of the fixed grid approximation). Since $\hat{y}_k = 1$ implies that $\lambda(k) \in \mathcal{B}(\hat{N})$, we have that

$$\mathcal{M}(\hat{\mathcal{F}}_1) \subseteq \mathcal{B}(\hat{N}).$$

Unfortunately, *this inclusion may be strict* since there can be points $\lambda \in \mathcal{B}(\hat{N})$ which do not satisfy $\lambda \leq \lambda(k)$ for any k with $\hat{y}_k = 1$. Thus, even if (20) holds, this does not imply that $\hat{\mathcal{F}}_1$ is a δ -feasible staffing frontier. Figure 3(a) illustrates this difficulty; in the figure, the dots represent the selected points in the random sample, the light shaded area represents the set $\mathcal{B}(\hat{N})$, and the dark shaded area represents the region $\mathcal{M}(\hat{\mathcal{F}}_1)$.

One can overcome this difficulty using a simple update mechanism that will generate a δ -feasible staffing frontier, as the one illustrated in Figure 3(b). In the appendix we describe one relatively simple approach for obtaining a feasible frontier, i.e, for moving from 3(a) to 3(b). The

approach is based on adding points that lie in the “gaps” between $\mathcal{M}(\hat{\mathcal{F}}_1)$ and $\mathcal{B}(\hat{N})$ and then scaling the points up until we obtain a set \mathcal{F} with $\mathbb{P}(\mathcal{M}(\mathcal{F})) \geq 1 - \delta$. The approach allows one to trade-off solution accuracy with time; by adding more points a better approximation is obtained at the cost of more expensive simulations when using the staffing frontier to find a feasible solution. We remark that there may be better ways to find a δ -feasible staffing frontier that approximates $\mathcal{B}(\hat{N})$; the important point is only that we obtain such an approximation.

Recall from §6.2 that when using the sample-based approximation we solve M instances based on independent random samples to obtain a statistical lower bound. We run the above procedure on each of the resulting solutions, and for each we obtain a corresponding staffing solution \bar{N} by scaling up the original solution \hat{N} by the same amount that we scaled up the points in the frontier. We then select the δ -feasible staffing frontier such that the corresponding solution \bar{N} has minimum cost. The selected δ -feasible staffing frontier and corresponding solution are what we subsequently use to obtain a solution to the original problem (4).

Statistical test of δ -feasibility of a staffing frontier. Checking whether a set of points \mathcal{F} is a δ -feasible staffing frontier can be done using a simple statistical test using a single very large sample of arrival rate vectors $\hat{\lambda}(k)$, $k = 1, \dots, K'$. If $\hat{\lambda}(k) \in \mathcal{M}(\mathcal{F})$ for at least a fraction $1 - \delta + \epsilon(K', \beta, \delta)$ of the points $\hat{\lambda}(k)$ for $k = 1, \dots, K'$, then we can say with a confidence level $1 - \beta$ that the set of points \mathcal{F} is indeed a δ -feasible staffing frontier, where $\epsilon(K', \beta, \delta) = [(2\delta/K') \ln(1/\beta)]^{1/2}$. (This confidence estimate is based on the Chernoff inequality, see [40] page 394.) In our experiments, for example, we use $K' = 200,000$, $\beta = 0.01$, $\delta = 0.1$, and obtain $\epsilon(K, \beta, \delta) = 0.0022$.

6.4 Computational experience with solving RSPP

Solving the RSPP is a crucial step in our solution approach for the real staffing and routing problem of interest (4). Hence, while computational efficiency is not the focus of this paper, we show via some numerical examples that the discrete approximations of the RSPP we have proposed can be solved in a reasonable amount of time using an off-the-shelf MIP solver.

Example 6.1 Consider a system with two customer classes, i.e. $\mathcal{I} = \{1, 2\}$, and three agent groups $\mathcal{J} = \{1, 2, F\}$ such that agents in pool i , $i = 1, 2$ are trained to serve only class- i customers but so that agents in pool F are generalists and can serve both customer classes; see Figure 4. This setup is known as the M model of skill-based routing. We take $\mu_{ij} \equiv 1$ and $\theta_1 = \theta_2 = 1$. Also, we

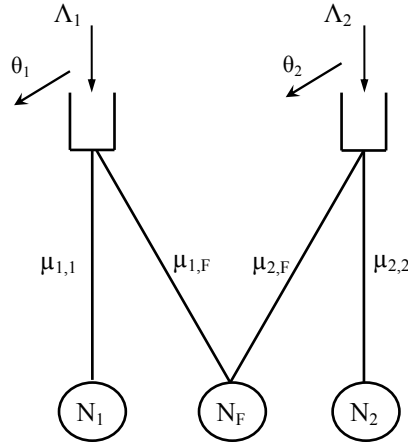


Figure 4: The M model of call centers

fix the same abandonment constraint $\alpha_1 = \alpha_2 = 4\%$ for both customer classes and let $\delta = 0.1$ so that the chance constraint bound, $1 - \delta$, is equal to 0.9. We assume that the point estimate is given by $(\lambda_1, \lambda_2) = (120, 80)$ and that the error, Z , has a bivariate normal distribution (truncated so that $\Lambda = \lambda + Z$ is positive) with $\sigma_1^2 = 820$, $\sigma_2^2 = 460$ and correlation $\rho = -0.25$.⁴

Finally, we assume that the agent costs are given by $c_1 = c_2 = 1$ and $c_F = 1.1$ so that the flexible agents cost 10% more.

Table 2 gives results for different values of Δ for example 6.1 solving the problem using the original MIP formulation (15) and improved formulation MIP-RSPP given by (17). These results were obtained on a 2.13 GHz machine with 2 GB RAM using CPLEX 9.0 to solve the MIP formulations. We used a time limit of one hour; a ‘Lim’ entry in the table indicates the time limit was reached, in which case the final percent gap between the best solution and lower bound is reported. It is clear from Table 2 that the improved formulation yields significantly improved solution times, especially as the resolution parameter Δ gets small. Also, for relatively large resolution Δ the obtained solutions are more costly and conservative (low violation probability) but as Δ is decreased, we obtain solutions with reduced cost, and with violation probability closer to the target $\delta = 0.1$. (The exception is $\Delta = 1$, where a significant optimality gap remains, suggesting that the best solution found within the time limit of one hour is significantly suboptimal.) We also observe from Table 2 that the number of points in the staffing frontier, $|\mathcal{F}|$, remains small, even for small Δ when $|\mathcal{L}(\Delta)|$ gets large.

⁴The choice of negative correlation has no meaning in our experiment. We use both negative correlation and positive correlations in our examples—see Example 6.2).

| Δ | Original MIP (15) | | MIP-RSPP (17) | | | | |
|----------|-------------------------|-------------|-----------------|------------|-------|-------------|-----------------|
| | $ \mathcal{L}(\Delta) $ | Time (s) | $ \mathcal{D} $ | Time (s) | Cost | Viol. Prob. | $ \mathcal{F} $ |
| 10 | 169 | 2.4 | 70 | 0.2 | 244.8 | 0.084 | 5 |
| 8 | 256 | 17.3 | 110 | 0.5 | 243.5 | 0.082 | 8 |
| 6 | 441 | 2393.5 | 197 | 2.8 | 243.3 | 0.086 | 9 |
| 4 | 961 | Lim (1.3%) | 451 | 17.4 | 240.0 | 0.097 | 15 |
| 2 | 3721 | Lim (4.6%) | 1831 | Lim (0.3%) | 238.7 | 0.100 | 29 |
| 1 | 14641 | Lim (38.0%) | 7384 | Lim (8.0%) | 241.0 | 0.087 | 34 |

Table 2: Computational results for solving RSPP for example 6.1.

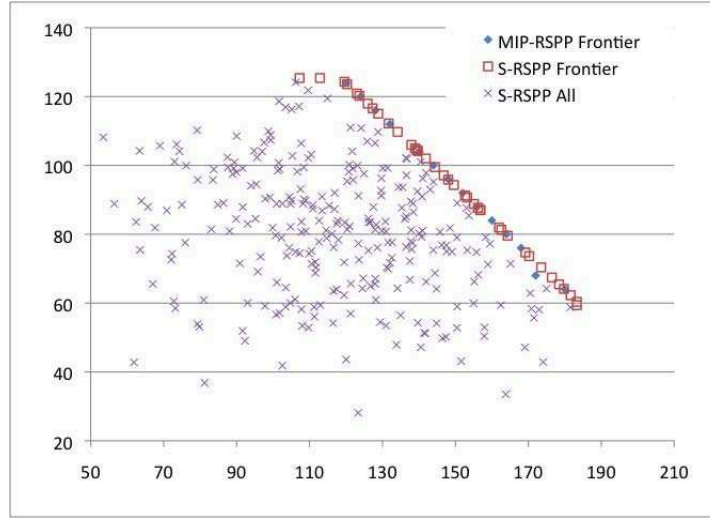


Figure 5: Staffing frontier for example 6.1.

For comparison, we also solve this example using the sample approximation S-RSPP as in §6.2. We used a sample size of $K = 300$, and solved $M = 10$ problems. The total time to solve all 10 problems was 200.3 seconds, and the total time to construct δ -feasible staffing frontiers from each of these solutions was less than 90 seconds. The δ -feasible staffing frontier with lowest cost corresponding solution had 38 points, and the corresponding solution had cost 240.3. The minimum of the ten optimal values \hat{L}^{10} is equal to 232.5 and, as explained in §6.2, provides a lower bound on the true optimal value of the RSPP with probability at least 0.999.

Figure 5 depicts the staffing frontier selected by the fixed-grid approach with $\Delta = 4$, and the δ -feasible staffing frontier obtained from the sample-approximation approach, along with the complete set of $\lambda(k)$ points that were selected in the solution of the S-RSPP instance that yielded this frontier. ■

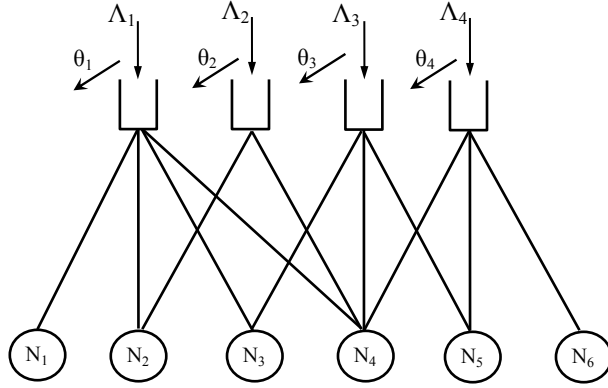


Figure 6: A call center with 4 customer classes and 6 agent groups.

Example 6.2 We now provide an example with 4 customer classes and 6 agent types so that $\mathcal{I} = \{1, \dots, 4\}$ and $\mathcal{J} = \{1, \dots, 6\}$. In this example, the service rates are independent of agent class, i.e., $\mu_{ij} \equiv \mu_i$ for all $j \in J(i)$. The service rate vector is given by $\mu = (1, 1.5, 1.4, 1.3)$. The targets α_i , $i = 1, \dots, 4$ are given by $\alpha = (0.05, 0.04, 0.03, 0.05)$ and the chance constraint risk-level δ is 0.1. The costs are given by the vector $c = (1, 1.1, 1.2, 1.25, 1.1, 1)$. Finally, we assume that the point estimate is $\lambda = (120, 80, 100, 150)$, the variance are given by $\sigma^2 = (820, 460, 700, 400)$ and the correlations are $\rho_{1,2} = \rho_{2,1} = 0.25$, $\rho_{3,4} = \rho_{4,3} = -0.35$ and $\rho_{i,k} = 0$ otherwise. The connectivity in the network is as depicted in Figure 6.

We solved this example using the sample approximation approach again using $K = 300$ and $M = 10$. The minimal objective value of all ten instances was 415.0 which, in turn, by our discussion in §6.2 and by Theorem 6.1 is a lower bound on the optimal value of the original staffing problem (4) with probability 0.999. Solving the ten instances took a total of about 63 minutes. The time for solving each individual instance ranged from 3–20 minutes, with most of them solved in less than six minutes. The time to generate a δ -feasible staffing frontier using the procedure described in §6.3 was about one minute for each instance, leading to a total of another ten minutes of computation time. Of the ten δ -feasible staffing frontiers constructed, the one with the least cost was selected; this frontier had 182 points in it and the cost of the corresponding feasible staffing vector was 442.0, about 6.5% higher than the statistical lower bound. The cost of the best feasible staffing vector found—as tested by the a priori test, but before applying the procedure to construct a feasible frontier—was about 427.5, suggesting that more than half of the 6.5% error gap can be attributed to the approximation introduced by the need to find a δ -feasible staffing frontier.

Finally, to demonstrate that the additional procedure to guarantee feasibility is indeed necessary, we observe that the frontier $\hat{\mathcal{F}}_1$ obtained (without our additional procedure) from the instance that yielded the least cost feasible solution has a violation probability of $\mathbb{P}\{\Lambda \in \mathcal{M}(\hat{\mathcal{F}}_1)\} \approx 0.34$, so this set would have been far from a δ -feasible staffing frontier. ■

These computational results indicate that the RSPP problem can indeed be solved approximately in a reasonable amount of time with an off-the-shelf MIP solver, using a fixed-grid approach if $|\mathcal{I}| = 2$, or using the sample approximation approach with more customer classes. In addition, certain specialized new techniques for solving chance-constrained optimization problems could be used if necessary to reduce the computational burden of this step. For example, the approach in [28] was able to solve instances of the RSPP with as many as 40 customer classes and sample size $K = 3000$ to optimality in less than two minutes. A final important point we make is that the number of points in the staffing frontier produced by the RSPP is small when $|\mathcal{I}|$ is small, and can be kept reasonable (potentially with an accuracy trade-off) for larger $|\mathcal{I}|$. We will return to a discussion of computational efficiency in §7.2 after adding the last step of our approach, which uses the frontier information to construct a solution for the original staffing problem (4).

7 Back to the staffing problem

The RSPP serves as a first-order approximation for the staffing problem but cannot be used directly to determine the staffing levels for the call center as it ignores the effect of the stochastic behavior of the queue length. In this section we propose a simple and computationally efficient simulation-based search procedure that builds on the RSPP outputs to obtain a feasible solution to the original staffing problem (4). The resulting feasible solution can be shown to have provably low optimality gaps (see the discussion in §8).

The methods we present here are intentionally simple. Our purpose is to show that feasible solutions can easily be obtained from the solution and staffing frontier obtained from the RSPP. In particular, the most important feature of the suggested approach is that, through the RSPP, it translates the staffing problem with uncertain rates to one of solving a number of staffing problems with predictable rates corresponding to each of the arrival rate vectors of the RSPP staffing frontier. It is plausible that one can find more sophisticated search methods that build on the RSPP solution frontier as the starting point, but are more computationally efficient than our simple heuristic below.

7.1 A feasible solution via the RSPP's frontier

In this section we take as given an approximately optimal solution to the RSPP problem (11), \bar{N} , that satisfies

$$c \cdot \bar{N} \leq z_{RSPP} + \epsilon \quad (21)$$

for some $\epsilon > 0$, as well as a δ -feasible staffing frontier $\mathcal{F} = \{\lambda(k) : k \in F\}$ that corresponds to \bar{N} . The pair (\bar{N}, \mathcal{F}) may be obtained using one of the approaches that we described in §6. If the fixed-grid approximation with resolution Δ is used, this pair is obtained as described in §6.1 with $\epsilon = C\Delta$; see Theorem 6.2. If the sample-based approximation is used, \mathcal{F} is the best δ -feasible staffing frontier and \bar{N} is the solution corresponding to this frontier. In this case, the approximation error is given by $\epsilon = c \cdot \bar{N} - \hat{L}^M$, where \hat{L}^M is the statistical lower bound given by (19). Recall that the lower bound and the feasibility of the δ -feasible staffing frontier are statistically verified and hence do not hold deterministically. To simplify exposition, we will assume for the remainder of this section that we are in the (highly probable) case in which the feasibility of the staffing frontier and validity of the lower bound do hold.

Our point of departure for the algorithm that we propose is that—as the RSPP serves as a *fluid approximation* of the original staffing problem—we expect the optimal solution to (4) to be a perturbation, in some sense, of the RSPP solution. This is evidently the case for the simple case of the $M/M/N + M$ queue. In this case, discussed in §2, the RSPP is given by

$$\begin{aligned} \min \quad & N \\ \text{s.t.} \quad & \mathbb{P}_Z(\mu N \geq \Lambda(1 - \alpha)) \geq 1 - \delta, \\ & N \in \mathbb{R}_+. \end{aligned}$$

Trivially then, the solution of the RSPP is given by $\bar{N} = \lambda^*(1 - \alpha)/\mu$ where $\lambda^* := \inf\{\lambda \geq 0 : \mathbb{P}_Z\{\Lambda \leq \lambda\} \geq 1 - \delta\}$. Thus the staffing frontier is composed of the single point λ^* and $N^* = \min\{N \in \mathbb{Z}_+ : a(\lambda^*, N) \leq \alpha\}$ is feasible for (4). Moreover, in this simple case N^* is also optimal for (4). Because $N^* \geq \bar{N}$, N^* is equivalently defined by

$$N^* = \min\{N \in \mathbb{Z}_+, N \geq \bar{N} : a(\lambda^*, N) \leq \alpha\}. \quad (22)$$

In the multi-class multi-pool setting the single frontier point λ^* in the above example is replaced by a set of points – a frontier. As in the single-class, single-pool case, a staffing level that satisfies the service-level constraints for each point in the frontier will also be feasible for the staffing problem (4). This single-class case also suggests that, without a significant compromise in costs,

one may restrict attention to staffing levels N that are in the neighborhood of the RSPP staffing level \bar{N} . Hence, a multi-class analogue of (22) is the following formulation:

$$\begin{aligned}
\min \quad & c \cdot N \\
\text{s.t.} \quad & \alpha_i(\lambda(k), N, \pi) \leq \alpha_i, \quad i \in \mathcal{I}, k \in F, \\
& N_j \geq \bar{N}_j, \quad j \in \mathcal{J}, \\
& N \in \mathbb{Z}_+^J, \pi \in \Pi.
\end{aligned} \tag{23}$$

It is important that, unlike the single-class single-pool case, it is no longer guaranteed that (23) will be optimal for (4). Hence, one should regard (23) as a way to generate reasonably good solutions for (4). In the following, N^* is an optimal staffing solution to (4), and recall that we have assumed \bar{N} satisfies (21).

Theorem 7.1 *Let $(\check{N}, \check{\pi})$ be a feasible solution to (23). Then, $(\check{N}, \check{\pi})$ is feasible for (4). Moreover,*

$$|c \cdot \check{N} - c \cdot N^*| \leq \epsilon + \sum_{j \in \mathcal{J}} c_j |\check{N}_j - \bar{N}_j|. \tag{24}$$

The bound in (24) can be interpreted as saying that the optimality gap is composed of two components: (1) the error, ϵ , introduced by solving the RSPP approximately, and (2) the perturbation around the RSPP solution required to obtain a solution that satisfies the service level targets. In particular, one can judge the quality of a solution for (23) by using the RSPP cost $c \cdot \bar{N}$ as a benchmark. Theorem 7.1 is a direct consequence of Theorem 6.1 and the definition of the frontier—the detailed argument appears in the e-companion.

Optimizing with a fixed routing rule: In general, optimal solutions for (23) are not known, even when the frontier consists of a single point in which case (23) reduces to a staffing problem with perfectly predictable rates. Consequently, one needs to use heuristics to solve (23).

An issue in solving (23) is finding the routing rule π . If one starts from a fixed routing rule simulation-based optimization can be used to find the staffing; see e.g. [4, 13].⁵ We will follow this approach by fixing the routing rule to be the Fixed-Waiting-Ratio (FWR) rule that was proposed in [25]. It is important to emphasize that the scheme we present is generic and can be applied to other routing rules such as priority overflow rules [47, 13], the shadow-routing rule of [44], occupancy-tracking policies [8, 9] or a variety of index-based routing rules that are implemented in call-center

⁵Although, in contrast with these papers, we will have a set of arrival-rate vectors rather than a single one.

software. A useful property of FWR in our context is its simplicity and independence of the rule's parameters and the actual realization of the arrival rates.

The FWR rule has several versions, each designed to meet a different type of service-level target (see [25, 21]). We focus on the version aimed to satisfy abandonment constraints. For the introduction of FWR, let $W_i^h(t)$ be the accumulated waiting time of the customer at the head of the class- i queue and $Q_i(t)$ denote the number of customers in queue i at time t . In the following, whenever the set argmax contains more than one element we pick the one with the highest index.

- **A type- j agent that becomes available at time t** will serve the next the customer from queue i^* such that

$$i^* = \max \operatorname{argmax}_{i \in I(j): Q_i(t) > 0} \frac{W_i^h(t)}{\alpha_i / \theta_i},$$

- **A class- i customer that arrives at time t** will be routed to the server in a pool $j \in J(i)$ that has been idle for the longest time.
- Customers within each class are served in a FCFS manner.

Fixing $\pi = FWR$, the simulation-based procedure is iterative; it starts with \bar{N} —the staffing levels found from the RSPP and, for each $k \in F$, it simulates the call center with staffing levels \bar{N} and the given scheduling policy, π . As soon as it finds a $k \in F$ (and the corresponding arrival-rate vector $\lambda(k)$) such that $a_i(\lambda(k), N, \pi) > \alpha_i$ for some $i \in \mathcal{I}$ (i.e. the abandonment constraints are not met), it increases the staffing level using a simple update mechanism and then re-runs the simulations. The algorithm stops when a staffing level N is found such that $a_i(\lambda(k), N, \pi) \leq \alpha_i$ for all $i \in \mathcal{I}$ and $k \in F$. This staffing level N is necessarily feasible for (23).

The update mechanism of the staffing level is again simple: a server is added to a single pool $j \in \mathcal{J}$ so that the relative sizes of the server pools remain as close as possible to those suggested by the RSPP, i.e., to the vector $w(\bar{N})$ given by $w_j(\bar{N}) := \bar{N}_j / \sum_l \bar{N}_l$. Namely, let N^k be the staffing vector at the beginning of the k^{th} iteration. Then, N^{k+1} is constructed by adding an agent to one of the agent pools so as to minimize the difference between $w(\bar{N})$ and $w(N^{k+1}) := (N_1^{k+1} / \sum_l N_l^{k+1}, \dots, N_J^{k+1} / \sum_l N_l^{k+1})$. We will refer to this procedure as the *frontier-based simulation search procedure* to distinguish it from the exhaustive search by simulation that we perform in the following example to obtain some benchmarks.

Example 7.1 (An M model with common service rates) To be able to assess the performance of our proposed solution in terms of optimality we require a setting in which, at the very least,

reasonable lower and upper bounds on the optimal performance can be found. Using these bounds, we can then perform an exhaustive search to find the optimal solution. That optimal solution will serve as a benchmark for our RSPP-based solution.

For that purpose, we use the M-model from Example 6.1 but simplify it by setting the service rates to be common to all pools and customer classes, so that $\mu_{i,j} \equiv 1$, and letting both customer classes have the same patience rate, i.e., $\theta_1 = \theta_2 = 1$. Also, we set the abandonment targets to $\alpha_1 = \alpha_2 = 4\%$. As in example 6.1, we set $\sigma_1^2 = 820$, $\sigma_2^2 = 460$ and $\rho = -0.25$.

With the common service rates, patience rates, and abandonment constraints, a clear lower bound (which can be easily formalized by a coupling argument) is the $M/M/N + M$ staffing problem in (2) with arrival rate $\Lambda = \Lambda_1 + \Lambda_2$, service rate $\mu = 1$, patience rate $\theta = 1$ and $\alpha = 0.04$. Following the solution for the $M/M/N + M$ outlined in §2, we find the number 235. In turn, a lower bound on the number of servers (but not on the cost) for our M model is 235 servers and this bound is independent of the routing rule. To obtain an upper bound one can use a simple heuristic that determines the total staffing by considering individually each of the classes (the detailed description of the simple heuristic appears in Appendix B). For this model, the heuristic generates the solution $N_1 = 165$ and $N_2 = 115$ so that the total cost is 280 (recall that $c_1 = c_2 = 1$ and $c_F = 1.1$). Clearly, the true optimal solution must lie in between.

To find the best solution (with routing rule fixed to FWR), it now suffices to perform an exhaustive simulation-based search over all possible combinations of staffing levels (N_1, N_F, N_2) in $\mathcal{N} := \{N \in \mathbb{Z}_+^3 : N_1 + N_F + N_2 \geq 235, N_1 + N_F \geq 155, N_F + N_2 \geq 108\}$. To further decrease the running time of the exhaustive search, we let the staffing levels change in steps of 2 rather than one, i.e., the staffing levels we check have at least two more or two less agents than the next closest point. Therefore, the best solution found is at most 1.1 away from the optimal staffing level under the FWR routing rule. It is important that even after these simplifications, the running time of the exhaustive search is several days. This is because, to check the feasibility of the chance constraint for a given staffing vector, one has to randomize a large number of arrival-rate vectors and simulate the underlying queueing system for each of these.

The exhaustive search procedure yields two staffing vector solutions, $N = (98, 79, 60)$ and $N = (82, 99, 54)$. This in turn, implies that the cost of the optimal solution under FWR cannot be less than 243.8.

We now use the optimal solution of the exhaustive search to benchmark our frontier-based solution. We first solved the RSPP using the fixed-grid approach with $\Delta = 4$ to get $\bar{N} = (115.2, 57.6, 61.44)$ with a cost of 240 (with $\Delta = 2$ the cost is 238.2). Using the frontier points

from the RSPP with $\Delta = 4$, fixing the routing rule to FWR, and using the frontier-based simulation search procedure we obtain the staffing vector $\tilde{N} = (123, 62, 65)$ with an associated cost of 256.1. This cost is 5% away from the lower bound of 243.8 and 8.5% better than the solution obtained from the simple heuristic considering the customer classes independently. The running time of this procedure on a 1.6GHz machine with 4GB of RAM is 2 minutes. The overall running time (together with the time to solve the RSPP) is less than 3 minutes using the MIP-RSPP.

Finally, if we apply directly the RSPP solution for staffing and simulate the system, we find that the QoS constraint is satisfied on less than 80% of the realizations, indicating that the RSPP does not produce a solution that is feasible directly to (4) (at least not under FWR) and the additional steps that we introduced in this section are required. ■

Example 7.2 (Back to Example 6.2) We return to the network in Example 6.2. The network data remains unchanged but we specify in addition the patience parameters $\theta = (1, 1.3, 1.1, 1.4)$ (which are not needed for specifying the RSPP). As in Example 7.1, we first used a single-class based heuristic (see Appendix B) to obtain an upper bound with a cost of 524.2. The corresponding staffing vector is $(171, 82, 0, 0, 110, 142)$.

Applying our approach, we used \bar{N} and \mathcal{F} obtained using the sample-based approach to the solution of the RSPP and applied the frontier-based simulation procedure using FWR as the routing rule. This yielded the staffing vector $(141, 64, 27, 32, 79, 116)$ which is feasible to (4) and has a total cost of 482.2, an 8% improvement over the simple heuristic solution. Recall that, by Theorem 6.1, the RSPP serves as a lower bound for the true optimal solution of (4) and can be thus used as a crude benchmark. The true optimal solution would lie in between the RSPP solution and our proposed solution. The feasible solution we constructed for the RSPP using the sample approach had a cost of 442—i.e., within 8.3% of our staffing solution. It is important to recall that this cost of 442 corresponds to the feasible RSPP solution obtained after we “inflated” the solution to have a feasible frontier; see §6.3. The *infeasible* lower bound we obtained for the RSPP was 415.0 (see Example 6.2) which is 13.9% from the cost of the staffing solution we obtained. Hence, it is hard to know what is the true gap between our staffing solution and the optimal solution, but it must be smaller than 13.9%.

In this problem the RSPP δ -feasible staffing frontier (corresponding to the cost of 442) consists of 182 points. Letting the time horizon for each realization be 100 time units (which translates to an order of magnitude of 10000 arrivals per simulation), the run time of the simulation part is less

than 15 minutes on a 1.6 GHz machine with 4GB RAM⁶. This running time should be added to the approximately 75 minutes it takes to solve the RSPP and generate a staffing frontier using the sample-based approach to get a total of 90 minutes. The total running time is dominated by the time it takes to solve the RSPP. In particular, expected improvements in the running time of the RSPP (see the discussion at the end of §6.4) will lead to a significant decrease in the total running time. ■

Remark 7.2 (A decomposition approach) Yet another approach to obtaining feasible solutions to (23) is to decompose it into $|F|$ individual staffing problems such that the k^{th} problem is one with predictable rate $\lambda(k)$:

$$\begin{aligned} \min \quad & c \cdot N \\ \text{s.t.} \quad & a_i(\lambda(k), N, \pi) \leq \alpha_i, \quad i \in \mathcal{I}, \\ & N_j \geq \tilde{N}_j, \quad j \in \mathcal{J}, \\ & N \in \mathbb{Z}_+^J, \end{aligned} \tag{25}$$

Let (\check{N}^k, π^k) be a solution to (25) for $k \in F$. If simple (e.g. closed-form) solutions are available for (25), such a decomposition can be valuable. To be concrete, if $\pi^k \equiv \pi \in \Pi$, then any staffing vector that dominates each of $\{\check{N}^k, k \in F\}$ is necessarily feasible for (23). In particular, letting $\tilde{N}_j = \max_{k \in F} \check{N}_j^k$, we have that (\tilde{N}, π) is feasible for (23). While \tilde{N} might be somewhat conservative, there could be significant efficiency gains if one can use distributed computing to solve the problems (25) in parallel, or if simple solutions are indeed known to each of the problems (25). ■

7.2 Complexity of the overall procedure

We now discuss the complexity of our solution procedure, and in particular compare this with the complexity of performing an exhaustive simulation-based search on the possible staffing solutions (using a fixed routing policy).

Recall that our procedure consists of two phases: (i) formulation and optimization of the RSPP, and (ii) a simulation-based search procedure that finds a feasible solution for (4) using the RSPP frontier as an input. The RSPP is a two-stage chance-constrained optimization problem that is approximated by solving a mixed-integer program based on a discretization of the arrival-rate distribution. Solving this MIP has worst-case exponential complexity, but in practice it can often be solved efficiently provided that the size of the discretization is not too large by using specialized optimization approaches as in [28]. We presented two discretization approaches. The fixed-grid

⁶The simulation is implemented using a Matlab search code that calls a C++ simulation code to test the various staffing levels during the run of the procedure.

approach yields a discretization with size that grows exponentially in the number of customer classes I , and is hence practical only for systems with a small number of classes. The sample-based approximation partly overcomes this limitation because the required sample size grows only linearly with the size of the decision space (number of agents) and is independent of the number of customer classes. The main challenge in the sample-based approximation is the construction of the δ -feasible staffing frontier. Here it is possible to trade-off accuracy and computational efficiency by placing a bound on the number of points in this frontier. As the dimension of the arrival-rate vector grows such a bound will lead to less accurate approximations for the true optimal value of the RSPP. Thus, our approach towards the solution of the RSPP can be applied safely to systems with a moderate number of customer classes but may lead to crude approximations if the number of customer classes is very large. However, most call centers that have been modeled in the literature have fewer than 10 classes in each connected sub-system (see e.g. [47, 8, 9]), which would be considered to be of moderate size for our purposes.

In the second phase of our approach, we search for a feasible staffing solution in the neighborhood of the RSPP solution. The key operation in this search is checking whether a candidate solution satisfies the chance constraint on the service targets. Given the staffing frontier from the RSPP, we can perform this search efficiently by simulating the system with each of the arrival rate vectors on the frontier; e.g., in Example 7.2 we had to simulate with 182 arrival rate vectors.

To put the above in perspective it is important to emphasize that, in contrast with our approach, in a direct simulation-based optimization, the *simulation* complexity is significant. Indeed, checking feasibility requires first sampling a large number of arrival rate vectors, then simulating the system with each of these. For example, obtaining the same feasibility confidence and precision as we obtained for our δ -feasible staffing frontier requires a sample of 200,000 arrival rate vectors (see the end of §6.3). This large sample does not constitute a problem in the context of finding a feasible staffing frontier for the RSPP because we only need to check whether each of the sampled vectors is dominated by a vector in the frontier. Such a huge number is very significant when a simulation, which requires more than a second to perform, is required for each vector.

The way in which the efficiency depends on the number of agent and customer types also differs when comparing direct simulation-based search with our approach. The complexity of direct simulation-based search would grow exponentially with the number of agent types. In contrast, our approach has little dependence on the number of agent types. On the other hand, a direct simulation-based search does not depend significantly on the number of customer classes (although more customer classes will increase the already burdensome simulation time discussed

above) whereas the need to construct a feasible staffing frontier limits our approach to a moderate number of customer classes.

8 Discussion

In this paper we propose a new chance-constrained formulation for the problem of staffing a multi-class multi-type call center facing demand uncertainty. We provide a detailed solution procedure for this complex problem. Our solution approach for this formulation is based on the ability to translate the problem of staffing-with-uncertain-demand-forecasts to one of solving a small set of problems with perfectly predictable demand-rate vectors. The “translation” is achieved via the introduction of a random version of the static-planning problem—the RSPP introduced in §6.

This reduction of the uncertain case to the predictable case is important. It is plausible that optimal or nearly optimal solutions for the predictable-rates case can be translated via our approach (or a modification thereof) to optimal or nearly optimal solutions for the case with uncertainty. For example, it is plausible that diffusion-scale asymptotically-optimal staffing solutions for the perfectly-predictable case, as in [25], can be used to construct diffusion-scale asymptotically optimal solutions for the uncertain case. Unfortunately, the results in [25] are restricted to a certain subset of models, and consequently, an extension to the uncertain case is likely to share this restriction. This underscores the importance of improving the understanding of the (simpler) case with predictable rates. Improved solutions for these can then be translated to stronger optimality results for the case with uncertainty.

In terms of the formulation, we have argued that in certain situations the chance-constrained formulation can be more appropriate than the formulation based on an average constraint. However, one may also conceive of alternative formulations that are in a sense a compromise between these two approaches. There has been significant work, especially in the context of financial engineering, about alternatives to the Value at Risk (VaR) approach - which is essentially a chance constraint applied in a portfolio optimization context. A popular alternative is known as Conditional Value at Risk (CVaR) (see e.g. [38]) which has two advantages over VaR: (1) it yields a convex formulation and (2) it limits the *magnitude* of the losses instead of just the fraction of time losses occur as in VaR. We leave this approach as an interesting area for future research.

Time varying arrival rates The arrival-rates during the day are, of course, not stationary but rather time varying. The standard approach to staffing a call center is, however, to divide the day

into time intervals (of 15 or 30 minutes) and staff each time interval as if arrivals are stationary during that interval. Our approach in this paper can be interpreted as corresponding to the staffing problem of such a single interval.

Applying our approach repeatedly, each time for one interval can be too computationally expensive. Fortunately, for one model of time varying and uncertain arrival rates, our approach can be applied directly with a solution efficiency that is almost as good as the single-interval case.

Specifically, the literature suggests (see, e.g., [11]) that it is reasonable to assume that, over the time horizon $[0, T]$ (which may be regarded as a single day) the multi-dimensional arrival process is doubly stochastic Poisson with (random) rate-function vector $(\Lambda(t); 0 \leq t \leq T) := (\Lambda_1(t), \Lambda_2(t), \dots, \Lambda_I(t); 0 \leq t \leq T)$. In other words, given the realization of the multi-dimensional function $\Lambda(t)$, the arrivals of different classes follow independent non-homogeneous Poisson processes with the respective realized rate functions.

We will assume a specific case of the above general model which has been itself used in the literature (see, e.g., [23]). There exist (known deterministic) functions $f(t) := (f_1(t), \dots, f_I(t))$ such that $\Lambda_i(t) = \Lambda_i f_i(t)$ and so that $\int_0^T f_i(t) dt = 1$. Λ_i is interpreted as the daily class- i volume and the function $f_i(t)$ represent how this volume is distributed throughout the day. In this special model, then, the variability is decomposed into the predictable variability (captured by the functions $f_i(t)$, $i \in \mathcal{I}$) and the stochastic variability that is captured by the vector $\Lambda = (\Lambda_1, \dots, \Lambda_I)$. This model has been shown to be valid for various datasets and also used in [23]. It is thus encouraging that our approach can be applied in this case with relative simplicity and efficiency.

To apply the RSPP-based approach to this time-varying arrivals model one divides the day into a discrete set of smaller time intervals, $\mathcal{T} = \{1, \dots, T\}$. We then have a matrix $A = (a_{il}, i \in \mathcal{I}, l \in \mathcal{T})$ where $a_{il} = \sup_{(l-1)L \leq t \leq lL} f_i(t)$ and L is the size of the time interval (say half and hour). Because the uncertainty is still captured by only an I -dimensional vector Λ (recall that I is the number of customer classes), the RSPP-based solution requires only little modification. In particular, the dimension of the set \mathcal{L} in the fixed-grid approximation and the size of the sampled vectors in the sample-based approach are still I . The changes are restricted to the definition of the set $\mathcal{B}(N)$ which we define differently to take care of the various time intervals. Specifically, let N be now an $J \times T$ matrix where N_{jl} will represent the number of agents of type j required in

interval l . The set $\mathcal{B}(N)$ is now given by

$$\mathcal{B}(N) := \left\{ \lambda \in \mathbb{R}_+^I : \exists \nu \in \mathbb{R}_+^{I \times J \times T} \text{ with } \sum_{j \in \mathcal{I}(i)} \mu_{ij} \nu_{ijl} \geq \lambda_i a_{il} (1 - \alpha_i), i \in \mathcal{I}, l \in \mathcal{T}, \right. \\ \left. \sum_{i \in \mathcal{I}(j)} \nu_{ijl} \leq N_{jl}, j \in \mathcal{J}, l \in \mathcal{T} \right\}. \quad (26)$$

Importantly, we are still choosing λ from \mathbb{R}^I (and not from a higher dimension). Given a solution $\bar{N} := \{\bar{N}_{jl}, j \in \mathcal{J}, l \in \mathcal{T}\}$ of the RSPP one can use, as before, a simulation based approach, as in §7.1 using \bar{N} as a starting point.

We now highlight some potential directions for future research.

Other sources of uncertainty In this paper we restricted our attention to uncertainty in the arrival rates. One may wish to explicitly take into account other sources of uncertainty such as service times—which corresponds to uncertainty in estimating the parameters $(\mu_{ij}, i \in \mathcal{I}(j), j \in \mathcal{J})$ —or the magnitude of agent absenteeism. The latter corresponds to the fact that the number of agents who show up for work will often be different from the number scheduled to a shift. Conceptually, these forms of uncertainty (and potentially also others) can be incorporated in our framework at the expense of increasing the dimension of the problem and, in turn, the computational burden.

Asymptotic optimality in heavy-traffic The RSPP based approach can be shown to be asymptotically optimal in fluid scale, in the sense used in [8], when the arrival rates and systems size grow large. Hence, the RSPP-based approach provides a solution that is feasible and nearly optimal in an appropriate sense. A refined analysis of asymptotic optimality—in both fluid and diffusion scale—will most likely provide insight into the choice of routing rules as well as into the choice of various parameters in our approach such as the size of the discretization factor Δ in the construction of the fixed-grid discretized version of the RSPP.

Acknowledgments. The authors are grateful to Zohar Feldman for his significant help with the simulation. The authors also thank the anonymous referees for their careful review of this paper and many suggestions that helped significantly improve the presentation.

References

- [1] Z. Akşin, M. Armony, and V. Mehrotra. A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):655–688, 2008.
- [2] M. Armony. Dynamic routing in large-scale service systems with heterogenous servers. *Queueing Systems*, 51(3-4):287–329, 2005.
- [3] M. Armony and A. Mandelbaum. Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. *Oper. Res.*, 2009. Forthcoming.
- [4] J. Atlason, M.A. Epelman, and S.G. Henderson. Optimizing call center staffing using simulation and analytic center cutting plane methods. *Management Science*, 54(2):295–309, 2008.
- [5] A. Bassamboo, J.M. Harrison, and A. Zeevi. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.*, 54(3):419–435, 2006.
- [6] A. Bassamboo, J.M. Harrison, and A. Zeevi. Dynamic routing and admission control in high volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems*, 52(3-4):249–285, 2006.
- [7] A. Bassamboo, R. Randhawa, and A. Zeevi. Capacity planning in service systems with arrival rate uncertainty: Safety staffing principles revisited. Working paper, Northwestern University, Evanston, IL, 2008.
- [8] A. Bassamboo and A. Zeevi. Staffing telephone call centers subject to service-level constraints: An approximate approach via constraint dualization. Working paper, Northwestern University, Evanston, IL, and Columbia University, New York, NY, 2008.
- [9] D. Bertsimas and X. V. Doan. Data-driven and robust optimization approaches to call centers. Working paper, MIT, 2009.
- [10] S. Borst, A. Mandelbaum, and M. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, 2004.
- [11] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50, 2005.
- [12] G.C. Calafiore and M.C. Campi. Uncertain convex programs: randomized solutions and confidence levels. *Math. Program.*, 102:25–46, 2005.
- [13] M.T. Cezik and P. L'ecuyer. Staffing multiskill call centers via linear programming and simulation. *Management Science*, 54(2):310–323, 2008.
- [14] Bert P.K. Chen and Shane G. Henderson. Two issues in setting call centre staffing levels. *Annals of Operations Research*, 108(1–4):175–192, 2001.
- [15] J.G. Dai and T. Tezcan. Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems*, 59:59–134, 2008.
- [16] J.G. Dai and T. Tezcan. State space collapse in many server diffusion limits of parallel server systems. *Math. Oper. Res.*, 2008. Forthcoming.
- [17] D. Dentcheva, A. Prékopa, and A. Ruszczyński. Concavity and efficient points of discrete distributions in probabilistic programming. *Math. Program.*, 89:55–77, 2000.
- [18] A. Deslauriers, P. L'ecuyer, J. Pichitlamken, A. Ingolfsson, and A.N. Avramidis. Markov chain models of a telephone call center in blend mode. *Computers and Operations Research*, 34(6):16161645, 2007.
- [19] E. Erdoğan and G. Iyengar. On two-stage convex chance constrained problems. *Math. Meth. Oper. Res.*, 65:115–140, 2007.

- [20] Z. Feldman. Optimal staffing of systems with skills-based-routing. Msc. Thesis, Technion-Israel Institute of Technology, 2008.
- [21] Z. Feldman, I. Gurvich, and W. Whitt. Managing quality of service in call centers via queue-ratio routing: Asymptotic analysis and simulation-based optimization. *Working Paper. Northwestern University, Evanston, IL.*, 2009.
- [22] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.
- [23] N. Gans, H. Shen, and Y-P Zhou. Stochastic shift scheduling with recourse. Working paper, University of Pennsylvania, Philadelphia, PA, 2008.
- [24] I. Gurvich, M. Armony, and A. Mandelbaum. Service-level differentiation in call centers with fully flexible servers. *Manage. Sci.*, 54(2):279–294, 2008.
- [25] I. Gurvich and W. Whitt. Service-level differentiation in many-server service systems via queue-ratio routing. *Operations Research*, 2009. Forthcoming.
- [26] S. Küçükyavuz. On mixing sets arising in chance-constrained programming. Available at www.optimization-online.org, 2009.
- [27] T.G. Kurtz. Strong approximation theorems for density dependent Markov chains. *Stochastic Process. Appl.*, 6:223–240, 1978.
- [28] J. Luedtke. An integer programming and decomposition approach to general chance-constrained mathematical programs. In *IPCO 2010*, LNCS. Springer, 2010. Forthcoming.
- [29] J. Luedtke and S. Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM J. Optim.*, 19:674–699, 2008.
- [30] J. Luedtke, S. Ahmed, and G.L. Nemhauser. An integer programming approach for linear programs with probabilistic constraints. *Math. Program.*, 122:247–272, 2010.
- [31] S. Maman, A. Mandelbaum, and S. Zeltyn. Uncertainty in the demand for service: The case of call centers and emergency departments. Working paper, Technion - The Israeli Institute of Technology, Haifa, Israel, 2008.
- [32] A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.*, 57(5):1189–1205, 2009.
- [33] V. Mehrotra, O. Ozluk, and R. Saltzman. Intelligent procedures for intra-day updating of call center agent schedules. *Production and Operations Management (to appear)*, 2009.
- [34] A. Nemirovski and A. Shapiro. Scenario approximation of chance constraints. In G. Calafiore and F. Dabbene, editors, *Probabilistic and Randomized Methods for Design Under Uncertainty*, pages 3–48. Springer, London, 2005.
- [35] A. Prékopa. Probabilistic programming. In A. Ruszczyński and A. Shapiro, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, page 267. Elsevier Science B.V., 2003.
- [36] D. Pucci de Farias and B. Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Math. Oper. Res.*, 29:462–478, 2004.
- [37] T.R. Robins and T.P. Harrison. Call center scheduling with uncertain arrivals and global service level agreements. Working paper, Pennsylvania State University, University Park, PA, 2008.
- [38] R.T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.
- [39] A. Ruszczyński. Probabilistic programming with discrete distributions and precedence constrained knapsack polyhedra. *Math. Program.*, 93:195–215, 2002.

- [40] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelphia, 2009.
- [41] H. Shen and J. Z. Huang. Forecasting time series of inhomogeneous poisson processes with application to call center workforce management. *The Annals of Applied Statistics*, 2:601–623, 2008.
- [42] H. Shen and J. Z. Huang. Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management*, 10(3):391–410, 2008.
- [43] R. Soyer and M. M. Tarimcilar. Modeling and analysis of call center arrival data: A bayesian approach. *Management Science*, 54:266–278, 2008.
- [44] A. L. Stolyar and T. Tezcan. Control of systems with flexible multi-server pools: A shadow routing approach. Technical report, University of Illinois at Urbana-Champaign, 2008.
- [45] M.W. Tanner and L. Ntaimo. IIS branch-and-cut for joint chance-constrained programs with random technology matrices. Under second review, 2008.
- [46] J. W. Taylor. A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science*, 54(2):253–265, 2008.
- [47] R.B. Wallace and W. Whitt. A staffing algorithm for call centers with skill-based routing. *Manufacturing & Service Operations Management*, 7(4):276–294, 2005.
- [48] J. Weinberg, L. D. Brown, and J. R. Stroud. Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *Journal of the American Statistical Association*, 102:1185–1198, 2007.

Appendix A Update procedure for feasible frontier

To overcome the difficulty highlighted in §6.3 with respect to getting a δ -feasible frontier from the sample approximation, we make use of the condition (20) which is likely to hold at least approximately for a solution to a sample approximation problem, and attempt to find a δ -feasible staffing frontier that is a close approximation to $\mathcal{B}(\hat{N})$. Here there is a trade-off between computational time and solution accuracy. If we use more points in the frontier, we can get a better approximation to $\mathcal{B}(\hat{N})$ but at the expense of requiring more work to subsequently check solution feasibility using the frontier. Our procedure for constructing a δ -feasible staffing frontier that approximates $\mathcal{B}(\hat{N})$ begins with the set of points $\hat{\mathcal{F}}_1$ and proceeds in three steps.

1. The first step, illustrated in Figure 7(b), is to scale up each of the points in $\hat{\mathcal{F}}_1$ as much as possible while still remaining in the set $\mathcal{B}(\hat{N})$. Specifically, for each $\lambda' \in \hat{\mathcal{F}}_1$ we calculate $\beta' = \max\{\beta \mid \beta\lambda' \in \mathcal{B}(\hat{N})\}$ (by solving a simple linear program) then replace λ' by $\beta'\lambda'$, yielding a set $\hat{\mathcal{F}}_2$ with the same cardinality as $\hat{\mathcal{F}}_1$, and such that $\mathcal{M}(\hat{\mathcal{F}}_2) \subseteq \mathcal{B}(\hat{N})$ still holds.
2. In this step, illustrated in Figure 7(c), we add more points to obtain a closer “inner” approximation of $\mathcal{B}(\hat{N})$. The result will be a new set of points $\hat{\mathcal{F}}_3$ which is initialized as $\hat{\mathcal{F}}_2$. To

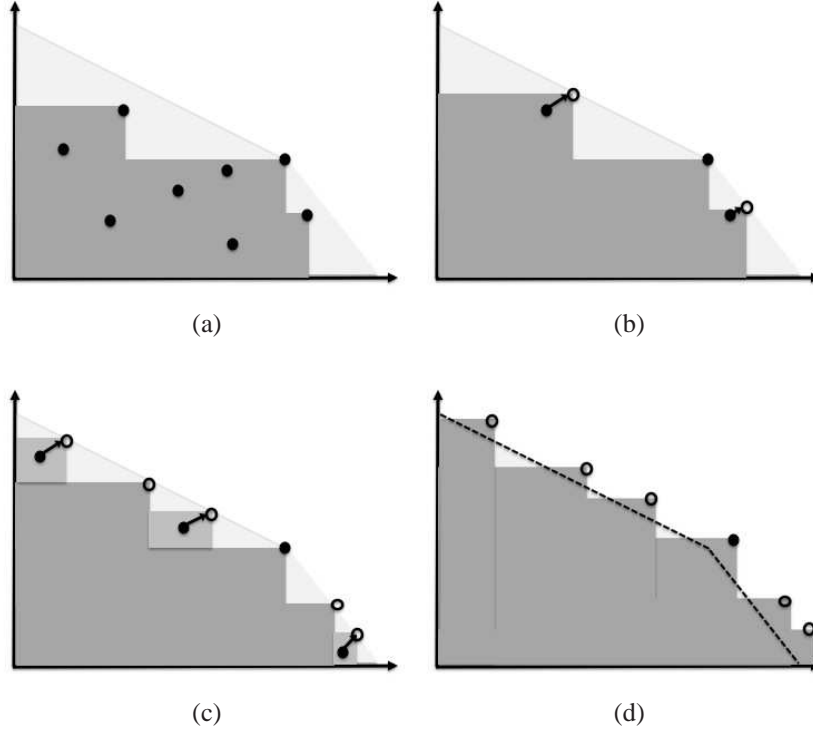


Figure 7: Construction of a staffing frontier from sample approximation solution.

limit the size of the resulting frontier, we fix in advance the number, p , of points we will add; we used $p = |\hat{\mathcal{F}}_2|$ so that we double the number of points. We obtain these points by taking random samples of Λ . For each sampled point $\hat{\lambda}$, we check whether $\hat{\lambda} \in \mathcal{B}(\hat{N}) \setminus \mathcal{M}(\hat{\mathcal{F}}_3)$; if not, we reject $\hat{\lambda}$ and continue, if yes, we scale up $\hat{\lambda}$ as much as possible while remaining in $\mathcal{B}(\hat{N})$ (as in the previous step), and add it to the set $\hat{\mathcal{F}}_3$. We terminate once we have added the maximum number of points. As long as $\phi := \mathbb{P}\{\Lambda \in \mathcal{B}(\hat{N}) \setminus \mathcal{M}(\hat{\mathcal{F}}_3)\}$ is not very small, we will not need to sample too many times before successfully adding a point. On the other hand, if we do sample a very large number of times without successfully adding a point, this would provide evidence that ϕ is very small, and hence we could stop because we already have a good approximation to $\mathcal{B}(\hat{N})$. Because we restricted the number of points added to be small, this latter stopping condition was not needed in our experiments.

3. The final step is illustrated in Figure 7(d). Here, we simply scale up all points by a value $\beta \geq 1$ such that $\mathbb{P}\{\Lambda \in \mathcal{M}(\beta\hat{\mathcal{F}}_3)\} \geq 1 - \delta$ holds with high confidence, as checked using a statistical test based on a very large sample (described at the end of §6.3). We choose a minimal β such that this condition holds (which can be found by a binary search). The set

$\hat{\mathcal{F}} = \beta \hat{\mathcal{F}}_3$ is the final result of the procedure and is a δ -feasible staffing frontier with high confidence.

The motivation for using random sampling in step 2 to choose the points to add is that this will tend to add points in regions of $\mathcal{B}(\hat{N}) \setminus \mathcal{M}(\hat{\mathcal{F}}_3)$ that have more probability mass. The staffing frontier $\hat{\mathcal{F}}$ constructed by the above procedure does not satisfy $\mathcal{M}(\hat{\mathcal{F}}) \subseteq \mathcal{B}(\hat{N})$. However, it does satisfy $\mathcal{M}(\hat{\mathcal{F}}) \subseteq \mathcal{B}(\beta \hat{N})$ where β is from step three of the procedure. We therefore define the solution $\bar{N} = \beta \hat{N}$ to be the solution *corresponding* to the constructed staffing frontier. Because $\mathcal{M}(\hat{\mathcal{F}}) \subseteq \mathcal{B}(\bar{N})$, the solution \bar{N} is feasible to RSPP with confidence at least as high as the confidence that $\hat{\mathcal{F}}$ is a δ -feasible staffing frontier.

Appendix B A simple heuristic and an upper bound

In practice, assuming that service rates depend only on the class of the customers, i.e, such that $\mu_{ij} \equiv \mu_i$, initial staffing calculations are often made for each customer class separately. With a perfectly predictable arrival-rate vector λ , this corresponds to finding for each class i

$$\underline{N}_i := \min\{N \in \mathbb{Z}_+ : a(N, \lambda_i, \mu_i, \theta_i) \leq \alpha_i\},$$

where $a(N, \lambda_i, \mu_i, \theta_i)$ is the fraction of abandoning customers in an $M/M/N + M$ queue with N servers, arrival rate λ_i , service rate μ_i and patience rate θ_i .

In a second step, one solves an allocation problem that determines how to allocate the total required capacity $\sum_{i \in \mathcal{I}} \underline{N}_i$ between the agent pools. This can be done by solving a version of the static planning problem with the requirements \underline{N}_i

$$\begin{aligned} \min \quad & c \cdot N \\ \text{s.t.} \quad & \sum_{j \in \mathcal{J}(i)} \mu_i \nu_{ij} \geq \underline{N}_i, \quad i \in \mathcal{I}, \\ & \sum_{i \in \mathcal{I}(j)} \nu_{ij} \leq N_j, \quad j \in \mathcal{J}, \\ & N \in \mathbb{R}_+^J, \quad \nu \in \mathbb{R}_+^{IJ}. \end{aligned} \tag{27}$$

In real time, one would then exploit the benefits of Skill-Based Routing to gain some of the efficiencies of cross-training. A possible translation of this to the chance constraint setting is to use

$$\underline{N}_i := \min \left\{ N \in \mathbb{Z}_+ : \mathbb{P}_Z \left(a(N, \lambda_i + z_i, \mu_i, \theta_i) \leq \alpha_i \right) \geq 1 - \delta/I \right\},$$

where δ/I is used to guarantee that the joint chance-constraint holds.

e-companion for:

Staffing Call-Centers With Uncertain Demand Forecasts:

A Chance-Constrained Optimization Approach

This e-companion contains the proofs for the theorems and lemmas in the paper. The proofs of the different results appear in their order of appearance in the paper.

EC1 Proofs

Proof of the transition from (4) to (5): We need to prove that (4) and (5) share the same set of optimal solutions.

To this end, fix $\bar{z} \in \mathbb{R}_+^I$ be such that $\mathbb{P}_Z(z : z \leq \bar{z}) \geq 1 - \delta$. Then, consider the optimization problem

$$\begin{aligned} \min \quad & c \cdot N \\ \text{s.t.} \quad & a_i(\lambda + \bar{z}, N, \pi) \leq \alpha_i, \quad i \in \mathcal{I} \\ & N \in \mathbb{Z}_+^J, \pi \in \Pi. \end{aligned} \tag{EC1}$$

Let $(N(\bar{z}), \pi)$ be an optimal solution. Such an optimal solution exists by the following lemma.

Lemma EC1.1 *There exist an optimal solution $(N(\bar{z}), \pi)$ for the optimization problem (EC1).*

The simple and detailed proof is postponed to the end of this e-companion. Obviously, $c \cdot N(\bar{z})$ is an upper bound for the optimal solution to (4). Let $\bar{N} = \max\{\sum_j N_j : N \in \mathbb{Z}_+^J, c \cdot N \leq c \cdot N(\bar{z})\}$. Then, any optimal solution, N^* , to (5) will have $\sum_{j \in \mathcal{J}} N_j^* \leq \bar{N}$. We now use this observation to bound the region of the arrival rates that we need to consider.

To this end, let $\bar{\mu} = \max_{i \in \mathcal{I}, j \in \mathcal{J}} \mu_{i,j}$. Then, the output rate under any optimal solution N^* for (4) is at most $\bar{\mu} \bar{N}$. In particular, any arrival rate λ that is covered under the optimal solution (i.e. such that $a_i(\lambda, N^*, \pi^*) \leq \alpha_i, i \in \mathcal{I}$) must satisfy $\sum_{i \in \mathcal{I}} \lambda_i (1 - \alpha_i) \leq \bar{N} \bar{\mu}$. Putting, $\bar{\alpha} = \max_{i \in \mathcal{I}} \alpha_i$, we then have that any arrival rate λ that is covered by (N^*, π^*) must satisfy $\|\lambda\| \leq b$ for all $i \in \mathcal{I}$ with $b = \bar{\mu} \bar{N} / (1 - \bar{\alpha})$. Consequently, any optimal solution (N^*, π^*) for (4) is feasible for (5). Since (5) is an upper bound for (4)—because we need to satisfy the chance constraint by using only arrival rate vectors in $[0, b]^I$ —we can conclude that they share the same set of optimal solutions. ■

Proof of Theorem 5.1: Fix $\Delta > 0$. Let N^* and \check{N} be, respectively, optimal solutions for (5) and (8). Obviously, \check{N} is also feasible for (5) so that $c \cdot N^* \leq c \cdot \check{N}$. Consider now a perturbed version

of (5) in which the chance constraint is replaced by a “ Δ -perturbed” chance constraint, i.e, consider the problem

$$\begin{aligned} \min \quad & c \cdot N \\ \text{s.t.} \quad & \mathbb{P}_Z(z : \max_i z_i \leq b, a_i(\lambda + z + \Delta, N) \leq \alpha_i, i \in \mathcal{I}) \geq 1 - \delta, \\ & N \in \mathbb{Z}_+^J, \pi \in \Pi. \end{aligned} \quad (\text{EC2})$$

An optimal solution $(N^*(\Delta), \pi^*(\Delta))$ exists for (EC2) by the same arguments that guarantee the existence of such solutions for (5); see §4. The outline of the rest of the proof is as follows: we will first show that $(N^*(\Delta), \pi^*(\Delta))$ is feasible for (8). This will imply that $c \cdot N^* \leq c \cdot \check{N} \leq c \cdot N^*(\Delta)$. We will then conclude the proof by showing that $|c \cdot N^*(\Delta) - c \cdot N^*| \leq C\Delta \vee \bar{c}I$ for some constant C as in the statement of the theorem.

First, we show that $(N^*(\Delta), \pi^*(\Delta))$ is feasible for (8). To do this we have to show that fixing the staffing to $N^*(\Delta)$ and the routing to $\pi^*(\Delta)$ we can choose a vector $\{y_k, k \in \mathcal{L}(\Delta)\}$ so that $\sum_{k \in \mathcal{L}(\Delta)} y_k p_k \geq 1 - \delta$ and such that $g(\lambda(k), N^*(\Delta), \pi^*(\Delta)) \leq 0$ for all k with $y_k = 1$. To this end, let

$$\mathcal{A}(\Delta) := \{x \in \mathbb{R}_+^I : a_i(x + \Delta, N^*(\Delta)) \leq \alpha_i, i \in \mathcal{I}\}.$$

In words, $\mathcal{A}(\Delta)$ is the set of arrival-rate vectors that are covered by $(N^*(\Delta), \pi^*(\Delta))$. We construct a vector $(y_k, k \in \mathcal{L}(\Delta))$ by setting $y_k = 1$ if $A_k \cap \mathcal{A}(\Delta) \neq \emptyset$ and $y_k = 0$ otherwise (with A_k as defined in §6). To see that the constructed vector y has the desired properties, fix k with $y_k = 1$. Since $A_k \cap \mathcal{A} \neq \emptyset$, the assumed monotonicity of the routing rule implies that $\lambda(k) \in \mathcal{A}(\Delta)$ and, in particular, that $a_i(\lambda(k), N^*(\Delta), \pi^*(\Delta)) \leq \alpha_i$, for all $i \in \mathcal{I}$. Finally, we can assume, without loss of generality that, for each $x \in \mathcal{A}(\Delta)$, $x_i \leq b$ for all $i \in \mathcal{I}$; see §4. In particular, by the construction of the vector y we also have that $\sum_{k \in \mathcal{L}(\Delta)} y_k p_k \geq \mathbb{P}(\mathcal{A}(\Delta)) \geq 1 - \delta$.

Hence, we have shown that $(\pi^*(\Delta), N^*(\Delta))$ is feasible for (8) and, in particular, that

$$c \cdot N^* \leq c \cdot \check{N} \leq c \cdot N^*(\Delta) \quad (\text{EC3})$$

To obtain (9), it remains to bound the distance $|c \cdot N^* - c \cdot N^*(\Delta)|$. To establish this bound we will show that, given an optimal solution (π^*, N^*) for (5), we can construct a feasible solution for its Δ -perturbed version in (EC2) with a cost that is larger than $c \cdot N^*$ by at most $C\Delta$. Equation (9) will then follow from (EC3).

We will initially construct the feasible solution for (EC2) under the assumption that, with (N^*, π^*) , the chance constraint in (5) is met with a strict inequality, i.e, that

$$\mathbb{P}_Z\left(z : \max_i z_i \leq b, a_i(\lambda + z, N^*, \pi^*) \leq \alpha_i, i \in \mathcal{I}\right) > 1 - \delta. \quad (\text{EC4})$$

We will remove this assumption at the end of the proof.

To construct now a feasible solution for (EC2) from (N^*, π^*) , we will create virtual server pools and thin the arrival streams so that some customers are routed immediately upon arrival to these server pools. Specifically, for each class i we arbitrarily pick a pool $j \in J(i)$ and denote this pool by $j(i)$. We then define a new staffing vector \tilde{N} by setting $\tilde{N}_j = N_j^* + \lceil \kappa \Delta / \mu_{i,j} \rceil$ if $j = j(i)$ for some i , and $\tilde{N}_j = N_j^*$ otherwise. Here, $\kappa > 0$ is a constant that we will later choose explicitly and it will be independent of Δ . We will not use the added servers in the same way that we use the original N_j^* servers. Rather, we will separate these additional servers from their pool to create a new pool, denote by $\tilde{j}(i)$, that consists of $\lceil \kappa \Delta / \mu_{i,j} \rceil$ servers. We will let this pool have its own queue. We then augment π^* as follows: when a class- i customer arrives, he is routed to the queue in front of pool $\tilde{j}(i)$ with a certain probability η_i (to be defined shortly) and is sent to the (regular) class- i queue with probability $1 - \eta_i$.

The thinning probabilities are determined from an estimate of the realized arrival rate. For this, we fix a time $T > 0$. Until time T we use the original routing rule π^* . At time T we register the number of class- i arrivals up to that moment, denoted by $A_i(T)$. We let $\hat{\lambda}_i(T) := A_i(T)/T$ be an estimate of the arrival rate. We then set the thinning probability for class- i to be $\eta_i = \min(2\Delta/\hat{\lambda}_i, 1)$. Beyond this thinning of the streams, we keep using π^* for routing. Customers in the added queues, $\{\tilde{j}(i), i \in \mathcal{I}\}$ are served on a FCFS basis.

Denote by $\hat{\pi}$ be the resulting routing rule. The actual value of η_i only depends on the number of arrivals until time T , hence the new policy $\hat{\pi}$ is admissible, because π^* is admissible. If we show that $(\hat{\pi}, \tilde{N})$ is feasible for (EC2), the proof is complete since this would imply that there exists a constant $C > 0$ such that $|c \cdot N^* - cN^*(\Delta)| \leq C\Delta$.

To establish the required feasibility, fix $\epsilon > 0$ and let

$$\xi(T, \epsilon) := \sup_{\lambda \in [\epsilon, b]^I} P\{2\lambda_i \Delta / \hat{\lambda}_i(T) \notin [\Delta, 4\Delta], \text{ for some } i \in \mathcal{I}\}.$$

Given $\lambda \in [\epsilon, b]^I$, with probability $1 - \xi(T, \epsilon)$ the arrival rate of class- i customers to the (new) queue in front of pool $\tilde{j}(i)$ is at least Δ so that the arrival rate of these customers to the (original) class- i queue is at most λ_i . Hence, given $\lambda \in [\epsilon, b]^I$,

$$a_i(\lambda + \Delta, \tilde{N}, \hat{\pi}) \leq (1 - \xi(T, \epsilon)) [a_i(\lambda, N^*, \pi^*) + \delta_i^\Delta(\kappa, \lambda)], \quad (\text{EC5})$$

where $\delta_i^\Delta(\kappa, \lambda)$ is the fraction of class- i customers that abandon from the (new) queue $\tilde{j}(i)$. We used here the assumed monotonicity of the routing rule to conclude that the fraction of abandoning class- i customers from the (original) class- i queue is at most like the fraction of abandonments if

the arrival rate to that queue is λ_i . Applying probabilities with respect to the distribution of the arrivals, we have that

$$\begin{aligned} \mathbb{P}_Z(a_i(\lambda + \Delta, \hat{N}, \hat{\pi}) \leq \alpha_i, i \in \mathcal{I}) &\geq \\ (1 - \xi(T, \epsilon))\mathbb{P}_Z(\lambda \in [\epsilon, b]^I; a_i(\lambda, N^*, \pi^*) \leq \alpha_i; \delta_i^\Delta(\kappa, \lambda) \leq \alpha_i, i \in \mathcal{I}) &. \end{aligned} \quad (\text{EC6})$$

We claim that $P(\lambda \in \delta_i^\Delta(\kappa, \lambda) \leq \alpha_i) \rightarrow 1$ as $\kappa \rightarrow \infty$, uniformly in $\lambda \in [0, b]^I$. Intuitively, the thinned pools have an arrival rate of at most 4Δ served by capacity of $\kappa\Delta/\mu_{i,j}$ which for κ large enough entails a small fraction of abandonments; this is formally proved in Lemma EC1.3 below. Also, in Lemma EC1.2 we show that $\xi(T, \epsilon) \rightarrow 0$ as $T \rightarrow \infty$. Hence, given $\epsilon > 0$, we can choose $T(\epsilon)$ and $\kappa(\epsilon)$ large enough so that

$$\begin{aligned} (1 - \xi(T(\epsilon), \epsilon))\mathbb{P}_Z(\lambda \in [\epsilon, b]^I; a_i(\lambda, N^*, \pi^*) \leq \alpha_i; \delta_i^\Delta(\kappa(\epsilon), \lambda) \leq \alpha_i, i \in \mathcal{I}) \\ \geq \mathbb{P}_Z(\lambda \in [\epsilon, b]^I; a_i(\lambda, N^*, \pi^*) \leq \alpha_i, i \in \mathcal{I}) - \frac{\epsilon}{4} \end{aligned} \quad (\text{EC7})$$

Finally, from the assumed slack in (EC4) of the statement of the theorem it follows that

$$\begin{aligned} \mathbb{P}_Z(\lambda \in [\epsilon, b]^I; a_i(\lambda, N^*, \pi^*) \leq \alpha_i, i \in \mathcal{I}) &\stackrel{\epsilon \rightarrow 0}{\rightarrow} \mathbb{P}_Z(\lambda \in [0, b]^I; a_i(\lambda, N^*, \pi^*) \leq \alpha_i, i \in \mathcal{I}) \\ &> 1 - \delta, \end{aligned} \quad (\text{EC8})$$

Combining (EC6)-(EC8) and choosing $\epsilon > 0$ small enough we have that

$$\mathbb{P}_Z(a_i(\lambda + z + \Delta, \hat{N}) \leq \alpha_i, i \in \mathcal{I}) \geq 1 - \delta.$$

This shows that $(\hat{\pi}, \hat{N})$ is feasible for (EC2) and concludes the proof of the theorem under the slack-assumption (EC4).

To complete the proof it remains then to remove the assumption in (EC4). To this end, define the staffing vector \hat{N} by $\hat{N}_j = N_j^* + 1$ if $j = j(i)$ and $\hat{N} = N_j^*$ otherwise. Then, $|c \cdot \hat{N} - c \cdot N^*| \leq \bar{c}I$. Moreover, using the same construction with thinning as above, we can now show that there exist a routing rule $\hat{\pi}$ so that

$$\mathbb{P}_Z\left(z : \max_i z_i \leq b, a_i(\lambda + z + \phi, \hat{N}, \hat{\pi}) \leq \alpha_i, i \in \mathcal{I}\right) \geq 1 - \delta,$$

for some $\phi > 0$. Using the monotonicity of the routing rule (and the fact that error is Normally distributed), we then have that

$$\mathbb{P}_Z\left(z : \max_i z_i \leq b, a_i(\lambda + z, \hat{N}, \hat{\pi}) \leq \alpha_i, i \in \mathcal{I}\right) > 1 - \delta.$$

Hence, a strict inequality holds for $(\hat{N}, \hat{\pi})$. We can now repeat our proof using $(\hat{N}, \hat{\pi})$ as our reference solution (replacing (N^*, π^*)) to obtain that $|c \cdot N^*(\Delta) - c \cdot \hat{N}| \leq \bar{c}I \vee C\Delta$ which, in turn, implies, that $|c \cdot N^* - c \cdot N^*(\Delta)| \leq \bar{c}I + \bar{c}I \vee C\Delta$. ■

Lemma EC1.2 *Given $\epsilon > 0$, $\xi(T, \epsilon) \rightarrow 0$ as $T \rightarrow \infty$.*

Proof: Let $\mathcal{N}_i(\cdot)$ be a unit-rate Poisson process. Using strong approximations (see e.g. Lemma 3.1 in [27]), we have that, for given $T, x_1, x_2 > 0, \lambda_i \geq \epsilon$,

$$P\left(\frac{\mathcal{N}_i(\lambda_i T) - \lambda_i T}{1 \vee \sqrt{\lambda_i T}} \notin [-x_1(1 \vee \sqrt{\lambda_i T}), x_2(1 \vee \sqrt{\lambda_i T})]\right) \leq ce^{-(1 \vee \sqrt{\lambda_i T})x_1 \vee x_2} \leq ce^{-(1 \vee \sqrt{\epsilon T})x_1 \vee x_2}, \quad (\text{EC9})$$

for all T large enough and some constant $c > 0$ that is independent of λ_i, T, x_1 and x_2 . By the definition of $\hat{\lambda}_i$,

$$P(2\lambda_i \Delta / \hat{\lambda}_i(T) \notin [\Delta, 4\Delta]) \leq P\left(\frac{\mathcal{N}_i(\lambda_i T) - \lambda_i T}{1 \vee \sqrt{\lambda_i T}} \notin \left[-\frac{1}{2}\sqrt{\lambda_i T}, 3\sqrt{\lambda_i T}\right]\right)$$

Combining this with (EC9) we have that (as $T \rightarrow \infty$)

$$\sup_{\lambda \in [\epsilon, b]^I} P\left(2\lambda_i \Delta / \hat{\lambda}_i(T) \notin [\Delta, 4\Delta]\right) \rightarrow 0 \text{ as } \kappa \rightarrow \infty. \quad \blacksquare$$

Lemma EC1.3 *Fix $\Delta > 0$. Then,*

$$\sup_{\lambda \in [0, b]^I} \delta_i^\Delta(\kappa, \lambda) \rightarrow 0, \text{ as } \kappa \rightarrow \infty.$$

Proof: The argument is straightforward and we provide only the sketch. Observe that $\delta_i^\Delta(\kappa, \lambda)$ is the fraction of customers abandoning from an $M/M/s + M$ queue with arrival rate that is at most 4Δ , service rate equal to $\mu_{i,j}$, abandonment rate θ_i , and $s = \lceil \kappa \Delta / \mu_{i,j} \rceil$ servers. Also the probability of abandonment from this queue is at most as the probability of delay from an $M/M/s$ queue with $\rho \leq 4\Delta / \kappa \Delta$. Assume that $\kappa > 4$, this $M/M/s$ queue is stable and the probability of delay is smaller than ρ (this follows from basic properties of the stationary distribution for the $M/M/s$ queue). Because the throughout, ρ , is independent of Δ and λ , the fraction of abandonments decreases to zero as κ increases uniformly in λ . ■

Proof of Theorem 6.1: We show that N^* , the optimal solution to (4), is a feasible solution of RSPP (11). To do so, we prove that, for any non-anticipating policy π , fixed arrival rate vector λ , and staffing level N , if (3) holds then

$$\mathcal{B}(N, \lambda) = \left\{ \nu \in \mathbb{R}_+^{I \times J} : \sum_{j \in \mathcal{J}} \mu_{ij} \nu_{ij} \geq \lambda_i (1 - \alpha_i), i \in \mathcal{I}, \sum_{i \in \mathcal{I}} \nu_{ij} \leq N_j, j \in \mathcal{J} \right\} \neq \emptyset. \text{(EC10)}$$

Hence, as (N^*, π^*) satisfy (3) with probability at least $1 - \delta$ (according to the constraint in (4)), N^* is a feasible solution for RSPP.

We prove this claim by contradiction. Assume that for all classes (3) holds but $\mathcal{B}(N, \lambda)$ is empty. Let $D_{ij}^{(n)}(t)$ be the number of class i customers that are served by server “ n ” in pool j and $T_{ij}^{(n)}(t)$ denote the total time spent by the n th server in pool i by time t . For all $t \geq 0$

$$D_{ij}^{(n)}(t) = S_{ij}^{(n)} \left(T_{ij}^{(n)}(t) \right),$$

where S_{ij}^n is a Poisson process with rate μ_{ij} . Also, we define

$$D_i(t) = \sum_{j \in \mathcal{J}} \sum_{n=1}^{N_j} D_{ij}^{(n)}(t).$$

We have, for any $t \geq 0$,

$$Q_i(t) = Q_i(0) + A_i(t) - R_i(t) - D_i(t). \text{(EC11)}$$

Because the number of customers in the system is bounded above by a $M/M/\infty$ system with service rate equal to $\min_{i,j} \mu_{ij} \wedge \min_i \gamma_i$, and the latter is clearly positive recurrent, we have

$$t^{-1} Q_i(t) \rightarrow 0 \text{(EC12)}$$

a.s. as $t \rightarrow \infty$. Also, by functional strong law of large numbers, as $t \rightarrow \infty$

$$t^{-1} A(t) \rightarrow \lambda \text{ and } t^{-1} S_{ij}^{(n)}(t) \rightarrow \mu_{ij} \text{(EC13)}$$

for all $i \in \mathcal{I}$, $j \in \mathcal{J}$, and $n \leq N_j$. Letting Ω be the sample space, we restrict our attention to $\omega \in \Omega$ that satisfy (EC12) and (EC13).

By our assumption (3), for $\omega \in \tilde{\Omega} \subset \Omega$, with $P(\tilde{\Omega}) = 1$, there exists a subsequence of $\{t_k\}$, which we denote again by $\{t_k\}$, such that $t_k \rightarrow \infty$ as $k \rightarrow \infty$ and

$$R_i(t_k) / (A_i(t_k) \vee 1) = \tilde{\alpha}_i \leq \alpha_i.$$

Also, since $A_i(t_k)/t_k \rightarrow \lambda_i$ as $k \rightarrow \infty$, we have

$$R_i(t_k)/t_k \rightarrow \lambda_i \tilde{\alpha}_i \leq \lambda_i \alpha_i \quad (\text{EC14})$$

as $k \rightarrow \infty$.

Let t_k be an increasing sequence such that $t_k \rightarrow \infty$. Let $T = (T_{ij}^{(n)} : i \in \mathcal{I}, j \in \mathcal{J}, n \leq N_j)$. Since $\frac{T_{ij}^{(n)}(t)}{t} \leq 1$, for all $t \geq 0$, for any $\omega \in \tilde{\Omega}$, we can find a subsequence denoted by $\{t_k\}$ again for notational simplicity, such that

$$\frac{T(t_k)}{t_k} \rightarrow x$$

for some $x = (x_{ij}^{(n)} : i \in \mathcal{I}, j \in \mathcal{J}, n \leq N_j)$, as $k \rightarrow \infty$. This with SLLN implies that

$$\frac{D_{ij}^{(n)}(t_k)}{t_k} = \frac{S_{ij}^{(n)}(T_{ij}^{(n)}(t_k))}{t_k} \rightarrow \mu_{ij} x_{ij}^{(n)} \quad (\text{EC15})$$

a.s. as $k \rightarrow \infty$. Also, because $\sum_i T_{ij}^{(n)}(t) \leq t$, for all t , $\sum_i x_{ij}^{(n)} \leq 1$ and so

$$\sum_{i \in \mathcal{I}} \sum_{n=1}^{N_j} x_{ij}^{(n)} \leq N_j$$

for all $j \in \mathcal{J}$. This with the fact that $\mathcal{B}(N, \lambda)$ is empty, there exists at least one class, say i , such that

$$\sum_{j \in \mathcal{J}} \mu_{ij} \sum_{n=1}^{N_j} x_{ij}^{(n)} < \lambda_i (1 - \alpha_i) - \epsilon \quad (\text{EC16})$$

for some $\epsilon > 0$.

For any $\omega \in \tilde{\Omega}$, by (EC11), (EC15) and (EC16)

$$\lim_{k \rightarrow \infty} t_k^{-1} R_i(t_k) = \lim_{k \rightarrow \infty} t_k^{-1} A_i(t_k) - \lim_{k \rightarrow \infty} t_k^{-1} D_i(t_k) \geq \lambda_i \alpha_i + \epsilon.$$

This clearly contradicts (EC14). ■

Remark EC1.4 *The above proof would be valid also if the definition of feasibility for given λ is a weaker version of (3). Specifically, it suffices to assume that*

$$\limsup_{t \rightarrow \infty} P \left\{ \frac{R_i(t)}{A_i(t)} > \alpha_i \right\} = 0, \text{ for all } i \in \mathcal{I}. \quad (\text{EC17})$$

In this case we can find a sequence $\{t_n\} \uparrow \infty$ such that

$$\left(\frac{R_i(t_n)}{A_i(t_n)} - \alpha_i \right)^+ \rightarrow 0, \text{ for all } i \in \mathcal{I}, \quad (\text{EC18})$$

a.s. as $t \rightarrow \infty$. Then the same arguments in the proof can be used to complete the proof.

Proof of Theorem 6.2: We first show that $\hat{N}(\Delta)$ is feasible to the RSPP (11). Let $\mathcal{F} = \{k \in \mathcal{L}(\Delta) : \lambda(k) \in \mathcal{B}(\hat{N}(\Delta))\}$. Feasibility of $\hat{N}(\Delta)$ to (MIP-RSPP) implies $\sum_{k \in \mathcal{F}} p_k \geq 1 - \delta$. Then, since $\lambda(k) \in \mathcal{B}(\hat{N}(\Delta))$ implies $\lambda \in \mathcal{B}(\hat{N}(\Delta))$ for all $\lambda \in A_k$ we obtain

$$\mathbb{P}_Z(\Lambda \in \mathcal{B}(\hat{N}(\Delta))) \geq \sum_{k \in \mathcal{F}} \mathbb{P}_Z(A_k) = \sum_{k \in \mathcal{F}} p_k \geq 1 - \delta$$

and so $\hat{N}(\Delta)$ is feasible to (RSPP).

Next, let N^* be an optimal solution to (11), so that $z_{RSPP} = c \cdot N^*$. Let $\mathcal{L}^*(\Delta) = \{k \in \mathcal{L}(\Delta) : A_k \cap \mathcal{B}(N^*) \neq \emptyset\}$. Now define \bar{N} by $\bar{N}_j = N_j^* + \Delta \sum_{i \in \mathcal{I}} (1/\mu_{ij})$ for $j \in \mathcal{J}$. We show that \bar{N} is feasible to (MIP-RSPP). To do so, we first demonstrate that $\lambda(k) \in \mathcal{B}(\bar{N})$ for each $k \in \mathcal{L}^*(\Delta)$. Consider any $k \in \mathcal{L}^*(\Delta)$ and let $\lambda' \in A_k \cap \mathcal{B}(N^*)$. As $\lambda' \in \mathcal{B}(N^*)$ there exists $\nu \in \mathbb{R}_+^{\mathcal{I} \times \mathcal{J}}$ such that

$$\sum_{j \in \mathcal{J}(i)} \mu_{ij} \nu_{ij} \geq \lambda'_i (1 - \alpha_i), \quad i \in \mathcal{I}, \quad \sum_{i \in \mathcal{I}(j)} \nu_{ij} \leq N_j^*, \quad j \in \mathcal{J}.$$

Let ν' be defined by $\nu'_{ij} = \nu_{ij} + \Delta/\mu_{ij}$. Then

$$\sum_{j \in \mathcal{J}} \mu_{ij} \nu'_{ij} = \sum_{j \in \mathcal{J}} \mu_{ij} \nu_{ij} + \Delta \geq \lambda'_i (1 - \alpha_i) + \Delta \geq \lambda_i(k) (1 - \alpha_i),$$

for each $i \in \mathcal{I}$ and

$$\sum_{i \in \mathcal{I}} \nu'_{ij} = \sum_{i \in \mathcal{I}} \nu_{ij} + \Delta \sum_{i \in \mathcal{I}} \frac{1}{\mu_{ij}} \leq N_j^* + \Delta \sum_{i \in \mathcal{I}} \frac{1}{\mu_{ij}} = \bar{N}_j$$

for each $j \in \mathcal{J}$. Thus, $\lambda(k) \in \mathcal{B}(\bar{N})$. Then we have

$$\mathbb{P}_{\hat{\Lambda}}(\hat{\Lambda} \in \mathcal{B}(\bar{N})) \geq \sum_{k \in \mathcal{L}^*(\Delta)} p_k = \sum_{k \in \mathcal{L}^*(\Delta)} \mathbb{P}_Z(\Lambda \in A_k) \geq \mathbb{P}_Z(\Lambda \in \mathcal{B}(N^*)) \geq 1 - \delta,$$

where the last inequality follows from feasibility of N^* to (11) and the second-to-last follows because $\mathcal{B}(N^*) \subseteq \cup_{k \in \mathcal{L}^*(\Delta)} A_k$. Thus, we have proved that \bar{N} is feasible to (MIP-RSPP) and therefore

$$c \cdot \bar{N} \geq c \cdot \hat{N}(\Delta)$$

by optimality of $\hat{N}(\Delta)$ to (MIP-RSPP). Therefore,

$$\sum_{j \in \mathcal{J}} c_j \hat{N}_j(\Delta) - z^* \leq \sum_{j \in \mathcal{J}} c_j \bar{N}_j - \sum_{j \in \mathcal{J}} c_j \cdot N_j^* = \sum_{j \in \mathcal{J}} c_j \Delta \sum_{i \in \mathcal{I}} \frac{1}{\mu_{ij}} = C \Delta,$$

where $C := \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} c_j / \mu_{ij}$. ■

Proof of Theorem 7.1: First we show the feasibility of \check{N} . For $k \in \mathcal{F}$ we define

$$\mathcal{B}_k = \{\lambda : \lambda \leq \lambda_k\}.$$

Because \check{N} is feasible if the arrival rate vector is equal to λ_k , it is feasible for all $\lambda \in \mathcal{B}_k$ by the assumed monotonicity of admissible policies. Hence, the probability that \check{N} is feasible is

$$\mathbb{P}(\cup_{k \in \mathcal{F}} \mathcal{B}_k) \geq 1 - \delta,$$

by the definition of the frontier so that \check{N} is feasible with respect to the chance-constraint. Finally, using Theorem 6.1, we have

$$\begin{aligned} c \cdot (\check{N} - N^*) &= c \cdot (\check{N} - \bar{N}) + c \cdot (\bar{N} - N^*) \\ &\leq c(\check{N} - \bar{N}) + (z_{RSPP} - z^*) + \epsilon \leq c(\check{N} - \bar{N}) + \epsilon, \end{aligned}$$

where the first inequality follows because \bar{N} satisfies (21). Feasibility of \check{N} then implies that $c \cdot (\check{N} - N^*) \geq 0$ completing the proof of (24). \blacksquare

Proof of Lemma EC1.1: Let $\mathcal{Z}(\bar{z}) \subset \mathbb{Z}_+^J$ be the set of staffing vectors, N , for which a feasible routing rule π exists. In other words, we say that N is in $\mathcal{Z}(\bar{z})$ if there exists a policy π such that $a_i(\lambda + \bar{z}, N, \pi) \leq \alpha_i$ for all $i \in \mathcal{I}$. We will now show that \mathcal{Z} is non-empty. Choosing then $\check{N} \in \mathcal{Z}$, we have that solving (EC1) is equivalent to solving the problem $\min_{N \in \mathcal{Z}} c \cdot N$ subject to the constraint $c \cdot N \leq c \cdot \check{N}$. This is an optimization problem over a compact set and hence it has an optimal solution $N(\bar{z})$.

Hence, it only remains to identify a pair (N, π) that is feasible for (EC1). This will guarantee then that \mathcal{Z} is non empty. We do this as follows: let $\tilde{\pi}$ be a policy that satisfies the following two properties: (i) it is work conserving policy, i.e, it does not idle an agent while there is a customer waiting in one of the queues that can be served by that agent), and (ii) it is Markovian with respect to the state-descriptor $\Xi(t) = (Q_i(t), Z_{ij}(t); i \in \mathcal{I}(j), j \in \mathcal{J})$ where $Q_i(t)$ is the class- i queue length at time t and $Z_{ij}(t)$ is the number of type- j server serving class- i customers at time t . A simple example for a potential choice is a policy that sends an arriving customer to any available server that can serve this customer and under which a newly available server serves the customer from the longest queue among those queues that he is capable of serving.

Let \check{N} be such that $\check{N}_1 = \check{N}_2 = \dots = \check{N}_J = M$ and put $\check{N}(M) := (M, M, \dots, M)$. We will show that

$$a_i(\lambda + \bar{z}, \check{N}(M), \tilde{\pi}) \rightarrow 0, \text{ as } M \rightarrow \infty, \text{ for all } i \in \mathcal{I}. \quad (\text{EC19})$$

Choosing choose M_0 large enough, $(\tilde{N}(M_0), \tilde{\pi})$ will be such that $\tilde{N}(M_0) \in \mathcal{Z}$ thus establishing that \mathcal{Z} is non-empty.

Hence, we turn to establish (EC19). To that end, let $X_i(t) := Q_i(t) + \sum_{j \in \mathcal{J}(i)} Z_{ij}(t)$ be the number of class- i customers in the system at time t . One can construct sample paths of $X_i(t)$ so that $X_i(t) \leq Y_i(t)$ for all $t \geq 0$ where $Y_i(t)$ is the number of customers in an $M/M/\infty$ queue with arrival rate $\lambda_i + \bar{z}_i$ and service rate $\min\{\theta_i, \underline{\mu}_i\}$ where $\underline{\mu}_i := \min_{j \in \mathcal{J}(i)} \mu_{ij}$. Moreover, one can construct the sample path so that the bounding infinite-server processes $Y_1(t), \dots, Y_I(t)$ are independent.

$$E \left[\left(\sum_{i \in \mathcal{I}(j)} X_i(\lambda + \bar{z}, \tilde{N}(M), \tilde{\pi}) - M \right)^+ \right] \leq E \left[\left(\sum_{i \in \mathcal{I}(j)} Y_i - M \right)^+ \right],$$

where $X_i(\lambda + \bar{z}, \tilde{N}(M), \tilde{\pi})$ and Y_i have the steady-state distribution of $X_i(t)$ and $Y_i(t)$. Since $(\sum_{i \in \mathcal{I}} Y_i - M)^+ \xrightarrow{P} 0$ as $M \rightarrow \infty$ and since $(Y_i - M)^+ \leq Y_i$ we have by dominated convergence that $E \left[(\sum_{i \in \mathcal{I}(j)} Y_i - M)^+ \right] \rightarrow 0$ as $M \rightarrow \infty$. In turn,

$$E \left[\left(\sum_{i \in \mathcal{I}(j)} X_i(\lambda + \bar{z}, \tilde{N}(M), \tilde{\pi}) - M \right)^+ \right] \rightarrow 0, \text{ as } M \rightarrow \infty,$$

and this holds for each $j \in \mathcal{J}$. Next we relate this result to the abandonments. Because of the work conservation of $\tilde{\pi}$, we have that $\sum_{i \in \mathcal{I}(j)} Q_i(\lambda + \bar{z}, \tilde{N}(M), \tilde{\pi}) \leq (\sum_{i \in \mathcal{I}(j)} X_i(\lambda + \bar{z}, \tilde{N}(M), \tilde{\pi}) - M)^+$, therefore

$$\sum_{i \in \mathcal{I}(j)} E \left[Q_i(\lambda + \bar{z}, \tilde{N}(M), \tilde{\pi}) \right] \rightarrow 0, \text{ as } M \rightarrow \infty, \quad (\text{EC20})$$

for all $j \in \mathcal{J}$. By an application of Little's law, we have

$$(\lambda_i + \bar{z}_i) a_i(\lambda + \bar{z}, \tilde{N}(M), \tilde{\pi}) = \theta_i E \left[Q_i(\lambda + \bar{z}, \tilde{N}(M), \tilde{\pi}) \right],$$

so that (EC20) implies, in particular, that

$$a_i(\lambda + \bar{z}, \tilde{N}(M), \tilde{\pi}) \rightarrow 0, \text{ as } M \rightarrow \infty.$$

This establishes (EC19) and concludes the proof of the lemma. ■