Expanding Service Capabilities Through an On-Demand Workforce

Xu Sun
Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32603
Weiliang Liu
Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 119077

An on-demand workforce can greatly benefit a traditional call center by allowing it to adjust its service capacity on demand quickly. Despite its conceptual elegance, the operationalization of this process is challenging due to the various sources of randomness involved. The purpose of this paper is to help call centers enhance service levels while keeping operating expenses low by taking advantage of an on-call pool of temporary agents in day-to-day operations. For that purpose, we develop a two-stage decision model in which the first stage seeks the optimal mix of permanent and on-call staff, and the second stage seeks a joint on-demand staffing and call scheduling policy to minimize the associated cost given the base staffing level and the size of the on-call pool. Because the exact analysis of the two-stage decision model seems analytically intractable, we resort to an approximation in a suitable asymptotic regime. In that regime, we characterize the system dynamics of the service operation and derive an optimal joint on-demand staffing and call scheduling policy for the second-stage problem, which in turn is used to find an approximate solution to the first-stage problem. In particular, the derived policy for the second-stage problem involves tapping into the on-call pool to procure a team of on-demand agents when the number of calls to be processed exceeds a certain threshold and dismissing them when it falls below another threshold; additionally, the call scheduling rule shows an unusual pattern due to the interplay between staffing and scheduling decisions. Extensive numerical studies under realistic parameter settings show that the solution approach we propose can achieve significant cost savings.

Key words: on-demand staffing; call centers; many-server queues; dynamic scheduling; diffusion analysis

1. Introduction

Traditional call center staffing faces a difficult trade-off between having too few agents on hand (understaffing) and having too many agents on hand (overstaffing). Long call hold times, agent burnout, and call abandonment are just a few negative consequences of understaffing. Overstaffing, on the other hand, leads to high labor expenditures due to excessive idleness. An ideal solution would be to enable call centers to bring on a number of temporary agents on demand and sign them off when they are no longer needed to meet customer needs.

Meanwhile, the trend towards flexible work justifies the creation of temporary positions in service settings to achieve short-term staffing flexibility. For instance, nursing homes use part-time nurse aides from a "float" pool to meet patient-centered metrics (Slaugh et al. 2018). In our collaboration

with a large call center owned by a multinational American e-commerce company (referred to as Company A), we found that the call center historically struggled to adjust staffing levels quickly to handle unexpected call surges despite using short staffing periods (30-minute intervals). To address this issue, the management of the call center proposed two approaches: (i) creating an on-call pool of former employees (agents who previously worked at the call center and are willing to assist with unexpected workload spikes on demand); and (ii) partnering with a gig-work agency that offers additional freelance agents as needed. Both approaches enable service capacity to expand and shrink on demand while controlling labor costs. Company A ultimately decided to experiment with the first approach, which serves as the direct motivation of this work.

The inherent stochasticity in demand and service processes creates the intrinsic value of an on-call pool of temporary agents as a means to make short-term capacity adjustments on demand. This flexibility offers the manager an option to optimally utilize capacity as the system changes. However, deploying and using an on-call pool, as done by Company A, presents challenges. First, there are costs associated with using either on-demand or permanent agents, and thus determining the appropriate on-call pool size and base staffing level that maximizes economic benefits is essential. Second, to reap the full benefits of the on-demand workforce, the manager needs to optimize the timing of adding and removing on-demand capacity, termed as the on-demand staffing decision. These decisions are inherently path-dependent, requiring a rigorous stochastic optimization framework. Third, call centers cater to different customer groups and segment calls into various classes, making call scheduling an essential control level. As a result, there is a need to orchestrate on-demand staffing and call scheduling decisions. Our goal is to address these distinct challenges by developing a modeling framework that captures the economic trade-offs between utilizing on-demand and permanent agents, incorporates stochastic uncertainties and coordinates relevant control levers.

The quality of customer service provided by call centers can be evaluated using various metrics, such as the abandonment rate (the percentage of calls that are hung up before being answered by an agent). In this work, we use the abandonment cost as a way to assess the adverse effects of system congestion. We describe a call center as a multi-class many-server system, where different classes are distinguished based on their arrival rates, their abandonment rates, and the cost incurred if a call is abandoned. The second-stage problem involves decisions about both on-demand staffing and call scheduling given the number of permanent agents and the size of the on-call pool. On-demand staffing involves switching between two modes - "on" and "off" - where the call center uses only its permanent staff in the "off" mode, and both permanent and on-demand staff (from the on-call pool) in the "on" mode. When the call center wants to switch to "on" mode, it sends an "entry request" to the on-call pool, and any on-demand agents who accept the invitation are immediately admitted into the system, while those who do not wait for the next entry request. Switching from "off" to "on"

incurs a "setup" cost of C, and each on-demand agent who agrees to work is compensated at a rate of c_o until he or she signs off. There is no limit to the number of times the operating mode can be changed. Call scheduling, on the other hand, determines which customer will be served when an agent becomes available when there are customers from more than one class waiting, thereby constituting an additional decision strategy. The objective of the second-stage problem is to minimize the sum of abandonment costs and the costs associated with actually using the on-demand agent. The first-stage problem seeks to minimize the overall cost, which includes expenses for permanent staff and costs from the second-stage problem.

The two-stage decision problem does not lend itself to exact analysis, particularly because the second-stage problem (i.e., the problem of finding an optimal on-demand staffing and call scheduling policy) is generally complicated. For this reason, we use common approximation approaches found in the literature. We posit that the call center has a high call volume and a large number of permanent agents who collectively provide a total capacity that roughly equals the total demand volume. Formally, we consider the popular Halfin-Whitt regime (Halfin and Whitt 1981), in which a diffusion control problem (DCP) will emerge as an approximation for our second-stage problem. Then, it suffices to seek the solution to the DCP, as its solution readily translates into an admissible control strategy for the original stochastic service system, which solves the second-stage problem.

1.1. Contributions

We summarize our contributions as follows.

Modeling: To the best of our knowledge, this paper is one of the first to study a problem of jointly making short-term capacity adjustment and resource allocation decisions for a service system with many service units and multiple customer classes in a formal stochastic optimization framework. The solution to this problem is then integrated into a long-term capacity planning problem, which involves determining the size of the permanent staff and the size of the pool of standby service units providing on-demand capacity.

Methodology: One important contribution of this research is a judicious construction of a solution to the Bellman equation associated with the DCP that approximates the second-stage problem. Absent the on-demand capacity, our second-stage problem essentially reduces to the one in Kim et al. (2018), where the authors develop dynamic customer scheduling policies by formulating and solving a corresponding DCP. An important theoretical contribution of their work is the construction of a solution to the associated Bellman equation, which involves analyzing a parametric family of functions, each of which is a solution to a differential equation. In contrast, our DCP includes an additional dimension that captures staffing control, resulting in a bivariate value function. Thus, constructing a solution to our Bellman equation entails analyzing a parametric family of *function*

pairs and identifying a function pair from this family that intersects twice with an area of intersection equal to the setup cost C. This task in turn requires analyzing convexity/concavity properties of the pairs of functions, a procedure that is not required in Kim et al. (2018). Thus, establishing the well-posedness of our Bellman equation requires analytical tools that go beyond those used in Kim et al. (2018).

Managerial implications: (I) Our approximation solution to the second-stage problem reveals a threshold value, $\bar{c}$ (dependent on the decision variables of the first-stage problem), such that switching between modes does not result in economic gains if $c_o \geq \bar{c}$. However, if c_o is low enough, there exists a value $\bar{C}$ that ensures switching between modes is worthwhile when $C < \bar{C}$. In this case, the system should be in the "on" mode when the total number of customers in the system rises above a given level, $x_1^\star$, and in the "off" mode when the total number of customers falls below another level, $x_0^\star < x_1^\star$. The two policy thresholds, $x_0^\star$ and $x_1^\star$, can be easily computed by solving the Bellman equation.

(II) Solving the second-stage problem also reveals a key insight about call scheduling: a customer class may lose service priority as the system becomes moderately congested and then regain priority as the system becomes even more congested. This "double switching" is not possible when staffing levels are fixed and patience times are exponentially distributed; see Ata and Tongarlak (2013), Atar and Lev-Ari (2018); see also §4.1.1 of Kim et al. (2018). To provide a brief explanation of the phenomenon, it is instructive to first consider the situation where on-demand capacity is absent. As nicely explained in Kim et al. (2018), in this case, a more impatient yet more expensive class will receive priority when there are fewer customers, allowing the system to reduce the instantaneous abandonment cost, and then lose priority since by prioritizing the other class, the system gains an additional abandonment rate and can thus better reduce congestion. In the presence of on-demand staffing, however, the benefit of exploiting the high abandonment rate of the more expensive class to trim down congestion when congestion levels are high is weakened. This is because increasing service capacity (by the use of on-demand agents) serves as an alternative means to reduce congestion. The "abnormality" in the scheduling rule is thus attributed to the capacity expansion and shrinkage mechanisms, suggesting a rich interplay between the two types of decisions.

(III) Furthermore, we numerically demonstrate that solving the first-stage problem leads to significant cost savings compared to relying solely on permanent staff. Hence, call centers are likely to benefit significantly from combining operational and tactical decisions.

1.2. Organization

The paper is organized as follows. In Section 2, we provide a comprehensive review of relevant literature. Section 3 introduces our modeling framework and presents the two-stage decision problem. Section 4 presents the DCP, which serves as an approximation for the second-stage problem and

provides an approximation for the first-stage problem. In the same section, we solve the DCP by finding the solution to the associated Bellman equation and gain valuable structural insights from it. Section 5 describes our solution procedure, which is developed based on the results from Section 4. Specifically, we propose a joint on-demand staffing and scheduling policy based on our solution. In Section 6, we discuss some key points related to our modeling and solution framework. In Section 7, we explain our simulation setup, present our numerical findings, and discuss some qualitative insights. Finally, in Section 8, we provide a concluding summary. We provide proofs of our main results that were omitted in the main paper, as well as numerical schemes for determining various relevant quantities, such as the solution to the Bellman equation, in the e-companion.

2. Literature Review

This study builds upon several streams of research which we survey below.

Staffing. Numerous papers have addressed the problem of staffing service systems, especially call centers. A comprehensive review of earlier research in this area can be found in Gans et al. (2003) and Aksin et al. (2007). The typical method for staffing call centers is based on queueing theory, which assumes that work can be deferred if no servers are available. Recent papers have employed newsvendor models to provide effective staffing solutions for queues, such as Harrison and Zeevi (2005), Whitt (2006), and Bassamboo et al. (2010). In particular, Bassamboo et al. (2010) investigate a staffing problem in a service system modeled as a single-class queue where customers may abandon while waiting and the arrival rate of work is random. They show that rules-of-thumb, such as the square root safety staffing principle, may be dependent on the relationship between the incoming load and the extent of uncertainty in arrival rates. The study by Gurvich et al. (2019) is the first to consider staffing with contractors, examining a model in which a platform could determine the number of available providers, the wage paid to each one who decides to work, and a cap on the number of potential providers. In another paper, Ibrahim (2018) analyzes a firm with a two-period staffing dilemma, where each agent is guaranteed to work in some periods but prefers one over the other, and the firm must decide how many agents to encourage to work in each period given the possibility of customer abandonment. Dong and Ibrahim (2020) extend this approach by allowing for hybrid staffing models that combine employees and contractors. They develop two approximations for this queueing system: a fluid approximation and a stochastic fluid approximation, and evaluate their accuracy in the large system limit. However, these papers focus on situations where the staffing level is fixed (at least in the short run) and the customer population is homogeneous. In our study, we allow for short-term adjustments to staffing levels and serve consumers from various classes. This makes staffing level adjustments operational rather than tactical decisions. While our study shares some similarities with these papers in terms of the number of servers being subject to change, our model

differs in scope and technique. More recently, Lobel et al. (2023) investigate a platform's staffing problem, focusing on the choice between hiring employees and setting up a contractor marketplace. The authors seek to understand the operational differences between these two work models in the context of a demand-driven system that is both stochastic and evolving over time. Their research provides insights into the trade-offs between hiring employees and utilizing contractors in dynamic and uncertain demand environments, demonstrating the benefits of a contractor marketplace. In particular, they find that hybrid solutions that combine both employees and contractors have complex optimal solutions and offer relatively limited benefits compared to a contractor marketplace.

Scheduling queues with multiple customer classes. Our paper builds on the body of literature on optimal scheduling of many-server queues in heavy traffic. Atar et al. (2010, 2011) consider an overloaded multi-class multi-server Markovian queueing system with abandonment and show that $c\mu/\theta$ rule is asymptotically optimal for minimizing customer holding cost. To handle systems with nonlinear holding costs and general patience times, Long et al. (2020) extend the $c\mu/\theta$ rule to the $Gc\mu/h$ rule. Papers that seek to find optimal scheduling rules in the popular Halfin-Whitt regime include Harrison and Zeevi (2004), Atar et al. (2004), Gurvich et al. (2008), Kim et al. (2018) in the case of fully flexible servers and Tezcan and Dai (2010), Armony and Ward (2010) in the case of skill-based routing. Our paper differs from the prior studies in several ways. The most significant distinction, arguably, is that in our system staffing levels are allowed to change rapidly to a moderate degree. This distinguishing feature is not only fundamental to the decision problem under consideration in this study, but it also has a significant impact on scheduling decisions. Ata et al. (2019) investigate a volunteer staffing problem from the perspective of a crop gleaning operation. The authors translate the problem to an approximating drift rate control problem and derive a nested threshold policy as the optimal staffing policy using a heavy-traffic framework. One distinction of our paper from theirs is that, in our setting, flexible staffing has a fixed cost, resulting in a joint "mode-switching" and drift-rate control problem.

Rate control in service systems. Speaking of drift-rate control, our work is broadly related to the literature on managing queues through adjustable arrival and service rates. An adjustable arrival rate can be achieved through pricing strategies. For instance, Low (1974) investigate the optimal control of a Markovian queue with a finite buffer, where the objective is to maximize the long-run average reward by serving customers. The system manager can only control the arrival rates by adjusting prices, as they have no control over the capacity level. Yoon and Lewis (2004) explore the problem of dynamic pricing and admission control in a system where both arrival and service rates are nonstationary and customers are sensitive to prices. They establish several structural properties of the optimal policy, including the monotonicity of the optimal prices in the state of the system under various cost structures. Ata and Shneorson (2006) investigate the problem of dynamically controlling the arrival and service rates in a service facility to optimize long-run average system welfare. Their

study proposes a solution to determine the optimal dynamic prices and service rates a system manager should set when serving delay-sensitive, rational customers. Additionally, some papers model rate control in queues as a drift-rate control problem for Brownian systems, such as Ata et al. (2005), Ghosh and Weerasinghe (2010).

Optimal switching problems. Our work is situated in the literature that studies stochastic control problems involving sequential switching decisions. These problems involve switching costs that can be considered fixed investments required to realize the operational advantages of an appropriate regime. Intuitively, such costs force the controller to look beyond the immediate advantages to ensure that a regime switch will accrue sufficient benefit over time to merit the fixed investment. Duckworth and Zervos (2001) solve an optimal two-regime switching problem by first using dynamic programming to derive the Bellman equation and then establishing a verification theorem. Ly Vath and Pham (2007) investigate a two-regime switching problem in which a geometric Brownian motion is used as the underlying state process. In the context of stock investing, Zervos et al. (2013) consider and solve a two-regime switching problem. All the studies above adopt the discounted cost criterion, and hence their approaches and results do not directly apply to the DCP under consideration, which adopts an average cost criterion. Wu and Chao (2014) consider optimal switching of a Brownian production system under average cost criteria. In their study, the construction of a classical solution to the corresponding Bellman equation relies on showing that the graphs of two functions intersect exactly twice and that the area of intersection is precisely the switching cost. They achieve this by exploiting the fact that the two functions are solutions to two linear differential equations and thus have explicit expressions. Compared to Wu and Chao (2014), the major technical challenge here is that the two functions we analyze to construct a solution to the Bellman equation are solutions to two piecewise linear differential equations and do not have explicit expressions. To overcome this difficulty, we must derive the desired structural properties for the two functions from first principles, which is technically challenging.

Heavy-traffic approximations for many-server queues. Last but not least, our solution builds on modern heavy-traffic approximation techniques for queues with many servers originating from the seminal paper Halfin and Whitt (1981), where the authors prove a diffusion limit for the transient and steady state for a sequence of M/M/N queues staffed according to the so-called square-root staffing as the arrival rate approaches infinity. Extensions to include impatient customers are considered by Garnett et al. (2002), Zeltyn and Mandelbaum (2005), Reed and Tezcan (2012), Huang et al. (2016).

3. Model

We will now describe our modeling framework, which incorporates core problem features and fully exposes essential tactical and operational trade-offs. Although our work is primarily motivated by call

center operations, our framework can be easily adapted to assist decision-making in other service settings that employ an on-demand workforce. Therefore, we will use generic terms such as "jobs" and "servers" instead of "customer calls" and "agents" in the following discussion.

3.1. System Dynamics

Jobs are categorized into I different classes, and the arrival of class-i jobs follows a Poisson process with a rate of λ_i. Let $A_i(t)$ represent the total number of class-i jobs that have arrived in the system up to time t. Then, for each job class i, we have that

$$A_i(t) = \Pi_i^a(\lambda_i t) \quad \text{for} \quad t \geq 0,$$

where Π_i^a is a unit-rate Poisson process that is independent of any other process. Additionally, we define λ as the total arrival rate, which is the sum of all class rates, $\lambda = \sum_i \lambda_i$. Each class-$i$ job has a patience time that represents the maximum amount of time the job is willing to stay in the queue. This is a common scenario in call centers, where a customer may hang up if kept waiting too long. We assume that the patience times of class-i jobs are exponentially distributed with a rate of θ_i. Moreover, we use $Q_i(t)$ and $R_i(t)$ to denote the number of class-i jobs waiting in the queue at time t and the number of class-i jobs that have abandoned the queue up to time t, respectively. Then,

$$R_i(t) = \Pi_i^r\left(\theta_i \int_0^t Q_i(u)\mathrm{d}u\right) \quad \text{for} \quad t \geq 0,$$

where Π_i^r is a unit-rate Poisson process that is independent of any other process.

The service system has N_0 permanent servers, which represent full-time employees in call center operations. These permanent servers are homogeneous and fully flexible in the sense that they work at the same speed and can handle any job, regardless of its class. The service times at these permanent servers are assumed to be exponentially distributed with a common rate parameter μ. Although a more general setting in which different job classes have different service rate parameters could be modeled and analyzed, such an approach would significantly increase the dimension of the problem and blur the main insights provided by the proposed on-demand staffing mechanism.

In addition to the permanent servers, the system manager can draw capacity from a pool of K on-demand servers. When additional service capacity is needed, the manager requests that on-demand servers provide service, which we refer to as an "entry request." The system enters an "on" mode as soon as an entry request is made. However, the availability of on-demand workers cannot always be guaranteed, and the number of on-demand workers who show up can be random. To capture this uncertainty in the system dynamics, we adopt a modeling approach inspired by the influential paper by Ibrahim (2018). Specifically, we assume that when an entry request is made, each on-demand server

that is currently off duty agrees to work (and thus goes on duty) with probability p. Any on-demand server that is unavailable at the time the entry request is made gets to make their decision as to whether to work at the next entry request.

When the system is in "on" mode, the system manager can issue an "exit request," instructing all on-demand servers that are currently on duty to go off duty *en masse*. We say that the system enters "off" mode whenever an exit request is made. When the switch from "on" to "off" occurs (and the system remains in "off" mode), the following occurs: First, on-demand servers that are currently on duty but idle go off duty immediately. Second, if there are χ_1 idle permanent servers and χ_2 busy on-demand servers, $\min\{\chi_1, \chi_2\}$ of these on-demand servers are chosen to hand over their jobs to an idle permanent server and go off duty. Third, any on-demand servers that are still busy after these hand-overs will either finish the task at hand or be chosen to transfer the remaining work to the next available permanent server, whichever comes first, and then go off duty. Based on this arrangement, some on-demand servers may still be busy processing jobs even after the system enters "off" mode. When an incoming job sees both a permanent and an on-demand server idle, the job is sent to a permanent server. Finally, all on-demand servers are assumed to exhibit the same service pattern as permanent servers, so service times at those on-demand servers are also exponentially distributed with rate parameter μ.

To ensure a well-defined system model, it is essential to specify a job scheduling rule that will determine which job to serve when a server (either permanent or on-demand) becomes available while jobs from multiple classes are waiting. An *admissible* scheduling rule must be non-anticipating (unable to utilize future information), non-idling, and maintain the first come, first serve (FCFS) principle within each class. Specifically, non-idling means that if there is an available server (permanent or on-demand), no job should wait in the queue. Mathematically, this implies that the number of busy servers (permanent or on-demand) at time t can be expressed as $X(t) \wedge (N_0 + \hat{Y}(t))$, where $X(t)$ and $\hat{Y}(t)$ denote the number of jobs in the system (either being served or waiting) and the number of on-demand servers on duty at time t, respectively. Accordingly, the cumulative number of service completions up to time t, denoted as $D(t)$, can be expressed as

$$D(t) = \Pi^d \left(\mu \int_0^t \left[X(u) \wedge (N_0 + \hat{Y}(u)) \right] \mathrm{d}u \right),$$

where Π^d is a unit-rate Poisson process that is independent of any other process.

3.2. Costs and Decisions

The primary purpose for incorporating on-demand capacity into service systems is to balance high service quality with staffing costs. This requires the establishment of a cost structure that reflects this tradeoff. Specifically, we define c_p as the average cost of hiring a permanent server. In addition to

this cost, a variable cost of c_o is incurred only for the duration that an on-demand server is on duty. Moreover, a fixed cost of C is incurred each time the system transitions from "off" to "on" mode. In a call center, this fixed cost covers disutilities that arise from administrative effort associated with the on-boarding process, which includes activities such as sending invites and setting up payroll for those who accept the invite. Indeed, through conversations with the call center manager at Company A, we learned that while payroll setup can be automated, the mass invitation needs to be initiated manually by the system operator, who also monitors the onboarding process to ensure it is smooth and error-free. Based on this, we posit that the size of the on-call pool has a minimal impact on the fixed cost. However, the fixed cost may also be used to operationalize disutilities associated with the mental workload of on-call agents, as we learned from our conversation with the call center. To incorporate this, it is sufficient to link the fixed cost to the pool size (K) as well, but doing so will not fundamentally affect our solution approach. Lastly, to account for the disutility of job abandonment, a penalty cost of r_i is assigned to each class-i abandonment.

The two types of *operational* decisions are on-demand staffing and job scheduling. On-demand staffing involves switching between different operating modes. The transition between modes is instantaneous, and there is no limit to the number of times the operating mode can be changed. Job scheduling determines the order in which jobs are processed when a server (permanent or on-demand) becomes available while there are jobs from more than one class waiting. This constitutes an additional decision. If there is only one job class, the joint on-demand staffing and scheduling control is reduced to on-demand staffing only. Over a longer time horizon, the system manager must make a *tactical* decision about the number of permanent servers to hire and the size of the on-call pool.

3.3. Managerial Objective(s)

The system manager faces a two-stage decision problem. The first-stage problem seeks to find N_0 and K in order to minimize

$$c_p N_0 + \Gamma(N_0, K), \tag{1}$$

where $\Gamma(N_0, K)$ denotes the optimal objective value of the second-stage problem to be formally introduced below.

The second-stage problem aims to find a joint admissible on-demand staffing and job scheduling policy. Specifically, its objective is to minimize

$$\limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\left[\sum_{i=1}^{I} r_i R_i(t) + c_o \int_0^t \hat{Y}(u)\mathrm{d}u + C\Xi(t)\right], \tag{2}$$

where we have defined $\Xi(t)$ to be the total number of entry requests sent up to time t. It is worth emphasizing that the aforementioned quantities are associated with a system with N_0 permanent servers and an on-call pool of size K.

3.4. Exact Approach

When N_0 and K are fixed, the second-stage problem (2) can be modeled as a Markov decision process (MDP) with $I+2$ dimensions. Of these dimensions, I describe the queue contents and the total number of jobs in the system; one indicates the system mode; and one accounts for the number of on-demand servers on duty. §EC.3 of the e-companion provides the detailed formulation. In theory, we can solve these MDPs to obtain $\Gamma(N_0, K)$ for all choices of N_0 and K, and then select values of N_0 and K that minimize the mathematical expression in (1) to obtain the solution to the first-stage problem. However, solving each MDP can be computationally prohibitive, even for the single-class problem, due to the curse of dimensionality, not to mention the task of solving the entire two-stage decision problem effectively.

To address these challenges, we identify a limit regime in which the second-stage problem (2) can be approximated as a DCP, which provides greater analytical tractability and reduces computational complexity. In the next section, we will focus on the formulation and solution of this DCP, based on which we propose a joint on-demand staffing and scheduling policy for the original system in §5.

4. Diffusion Analysis

In §4.1, we first formulate and solve a DCP that approximates the second-stage problem. We then introduce the diffusion approximation of the first-stage problem in §4.2.

4.1. Approximating the Second-Stage Problem

In §4.1.1, we specify the appropriate scaling condition, which leads to a controlled diffusion that approximates the number of jobs in the system (with centering). We characterize the solution to the DCP via the associated Bellman equation in §4.1.2. We describe our recommended policy based on the DCP solution in §5 later.

4.1.1. The DCP formulation. To arrive at the DCP, we need to identify a limit regime under which the system state can be approximately described by a diffusion process. To that end, we set the number of permanent servers N_0 according to the square root staffing rule, namely,

$$N_0 = \lambda/\mu + \beta\sqrt{\lambda/\mu} \quad \text{for some} \quad \beta \in \mathbb{R}. \tag{3}$$

The scaling condition implies that by solely relying on the permanent servers, the service system can maintain a moderate level of congestion (Ward 2012). Therefore, the system manager needs to make only moderate capacity adjustments. More precisely, we view K to be $O(\sqrt{\lambda})$, so we consider on-demand capacity of the form

$$K = \kappa\sqrt{\lambda/\mu} \quad \text{for some} \quad \kappa \in \mathbb{R}_+. \tag{4}$$

To proceed, let $\bar{X}(t)$ be the number of jobs in the system *centered* by the number of *permanent* servers (i.e., $\bar{X}(t) = X(t) - N_0$). From the law of flow conservation, we find that

$$\bar{X}(t) = \bar{X}(0) + \sum_i A_i(t) - D(t) - \sum_i R_i(t), \tag{5}$$

where $\bar{X}(0)$ denotes the number of jobs at time zero centered by N_0. It may be worth pointing out that $[\bar{X}(t) - \hat{Y}(t)]^+$ and $[\bar{X}(t) - \hat{Y}(t)]^-$ represent, respectively, the number of waiting jobs and the number of idling servers at time t.

To arrive at the desired diffusion approximation for $\bar{X}$, we apply the strong approximations for Poisson processes, similar to those in Kim et al. (2018), to get

$$\begin{aligned} A_i(t) &= \lambda_i t + \sqrt{\lambda_i}\hat{A}_i(t) + \epsilon_i^a(t) \quad \text{for} \quad i = 1, \ldots, I, \quad \text{and} \\ D(t) &= \mu \int_0^t \left[\left(N_0 + \hat{Y}(u) \right) \wedge \left(N_0 + \bar{X}(u) \right) \right] \mathrm{d}u + \sqrt{\lambda}\hat{S}(t) + \epsilon^d(t), \end{aligned} \tag{6}$$

where $\hat{A}_i(t)$ and $\hat{S}(t)$ are independent standard Brownian motions; ϵ_i^a and ϵ^d are error terms arising from the strong approximations, and they are an order of magnitude smaller than $\sqrt{\lambda}$ over any finite time horizon. Moreover, it is well known from the heavy-traffic theory that the abandonment processes admit the following approximations:

$$R_i(t) = \theta_i \int_0^t Q_i(u)\mathrm{d}u + \epsilon_i^r(t) = \theta_i \int_0^t \left[\bar{X}(u) - \hat{Y}(u) \right]^+ g_i(u)\mathrm{d}u + \epsilon_i^r(t), \quad i = 1, \ldots, I, \tag{7}$$

where ϵ_i^r are again error terms that are an order of magnitude smaller than $\sqrt{\lambda}$, and $\boldsymbol{G} := (g_1, \ldots, g_I)$ is a I-dimensional random element such that $Q_i(t) = \left[\bar{X}(t) - \hat{Y}(t) \right]^+ g_i(t)$ for all $t \geq 0$ and $i = 1, \ldots, I$. Intuitively, $g_i(t)$ is the fraction of queue contents kept in class i at time t, so $\boldsymbol{G}$ reflects the job scheduling control process. With this intuition, it is plain to see that

$$\boldsymbol{G}(t) \in \mathcal{A} := \left\{ \boldsymbol{q} := (q_1, \ldots, q_I) \,\middle|\, \sum_{i=1}^I q_i = 1, q_i \geq 0 \quad \text{for} \quad i = 1, \ldots, I \right\}. \tag{8}$$

Because service completions by permanent servers take place at the rate $O(\lambda)$ and any of such service completions would allow a busy on-demand server (if any) to go off duty when the system is in "off" mode, busy on-demand servers depart at a rate in the order of λ. However, since there are only $O(\sqrt{\lambda})$ on-demand servers, the time required for the system to "remove" all on-demand servers when it goes "off" is approximately $1/\sqrt{\lambda}$. Hence, once the system enters "off" mode, all on-demand servers on duty are removed from the system almost instantly when λ is sufficiently large. It follows that when the system switches back to "on" mode, all K on-demand servers are expected to be in a dormant state as long as the system does not chatter between "on" and "off" mode. Assuming all K on-demand agents are currently off duty, the actual number of servers that will show up to work follows a binomial distribution with parameters K and p the moment the system enters "on" mode.

This binomial distribution has a mean of Kp and a standard deviation of $\sqrt{Kp(1-p)}$. Furthermore, our scaling condition on K shows that the mean and standard deviation of this binomial distribution are roughly of the order of $\sqrt{\lambda}$ and $\lambda^{1/4}$, respectively. As a diffusion analysis is intended to "wash out" anything that is an order of magnitude smaller than $\sqrt{\lambda}$, it is expected that roughly Kp on-demand servers will go on duty every time the system enters "on" mode. Formalizing these discussions, we can approximate the process $\hat{Y}$ by $\kappa p\sqrt{\lambda/\mu}Y$, where Y is a binary process that takes the value of one (zero) when the system is in "on" ("off") mode.

On substituting (6)–(7) into (5), ignoring all error terms, replacing $\hat{Y}$ by the above-mentioned approximation, and utilizing the scaling condition (3), we obtain a diffusion approximation for $\bar{X}$, denoted by Z, which is a solution to the following stochastic integral equation:

$$Z(t) = Z(0) + \int_0^t b\left(Y(u), Z(u), \boldsymbol{G}(u)\right) \mathrm{d}u + \sqrt{2\lambda}B(t), \tag{9}$$

where the drift-rate function b is defined by

$$b(y, z, \boldsymbol{q}) := -\beta\sqrt{\lambda\mu} - \kappa p\sqrt{\lambda\mu}y + \mu\left[z - \kappa p\sqrt{\lambda/\mu}y\right]^- - \sum_{i=1}^{I}\theta_i\left[z - \kappa p\sqrt{\lambda/\mu}y\right]^+ q_i, \tag{10}$$

and $B(t)$ is a standard Brownian motion. In the above, Y and $\boldsymbol{G}$ are on-demand staffing and job scheduling control processes that are progressively measurable with respect to the filtration generated by B; $\boldsymbol{G}$ satisfies (8) and

$$Y(t) \in \{0, 1\}. \tag{11}$$

It is worth noting that the standard theory on stochastic integration and differential equations guarantees a unique strong solution to (9) for a fixed stationary policy $(Y, \boldsymbol{G})$. See, for example, Theorem 7 in (Protter and Protter 2005, Chapter V).

Finally, upon replacing R_i in (2) with the right-hand side of (7) and ignoring all error terms, we can approximate the objective therein by

$$\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}\left[\sum_{i=1}^{I} r_i\theta_i\int_0^t\left[Z(u) - \kappa p\sqrt{\lambda/\mu}Y(u)\right]^+ g_i(u)\mathrm{d}u + c_o\kappa p\sqrt{\lambda/\mu}\int_0^t Y(u)\mathrm{d}u + C\sum_{u\le t}[\Delta Y(u)]^+\right]. \tag{12}$$

Thus, the DCP seeks to find some $(Y, \boldsymbol{G})$ that minimizes (12) subject to (8), (9) and (11).

4.1.2. Solution to the DCP. We now apply Bellman's principle of optimality to deduce the best control strategy for the DCP. To that end, let $v(y, z)$ denote the relative value function associated with the DCP, where $y \in \{0, 1\}$. With reference to the general control theory, we expect $v(y, z)$ to solve, in conjunction with some constant $\eta^\star$, the following Bellman equation:

$$\begin{aligned}\min\Bigg\{\lambda v_{zz}(y,z) + \min_{\boldsymbol{q}\in\mathcal{A}}\Bigg\{b(y,z,\boldsymbol{q})v_z(y,z) + \left[z - \kappa p\sqrt{\lambda/\mu}y\right]^+\sum_i r_i\theta_i q_i\Bigg\} \\ + c_o\kappa p\sqrt{\lambda/\mu}y - \eta^\star,\ v(1,z) + C - v(0,z),\ v(0,z) - v(1,z)\Bigg\} = 0\end{aligned} \tag{13}$$

subject to the boundary conditions, $\lim_{z\to-\infty} v_z(y,z)=0$ and $\lim_{z\to\infty} v_z(y,z)=r_*$ for $r_*:=\min_i r_i$. In average cost dynamic programming, $\eta^\star$ is interpreted as a guess for the optimal average cost.

In order to construct a solution to the Bellman equation, we hypothesize that at optimality, the control Y has the following structure: If the system is currently in "off" mode, then it is optimal to remain in that mode if Z is below a threshold, say $z_1^\star$, and switch to "on" mode once Z rises above $z_1^\star$. If, however, the system is currently in "on" mode, then it is optimal to remain in that mode if Z is above a certain level, say $z_0^\star$, and switch to "off" mode as soon as Z drops below $z_0^\star$. Clearly, this strategy is well-defined if $z_0^\star<z_1^\star$. Moreover, if this strategy, denoted as $Y^\star$, is indeed optimal, we should be able to find $v_0(z):=v(0,z), v_1(z):=v(1,z)$ and $\eta^\star$, such that

$$\lambda v_0''(z)+\min_{\boldsymbol{q}\in\mathcal{A}}\left\{b(0,z,\boldsymbol{q})v_0'(z)+[z]^+\sum_i r_i\theta_i q_i\right\}=\eta^\star \quad \text{for} \quad z<z_1^\star, \tag{14}$$

$$\lambda v_1''(z)+\min_{\boldsymbol{q}\in\mathcal{A}}\left\{b(1,z,\boldsymbol{q})v_1'(z)+\left[z-\kappa p\sqrt{\lambda/\mu}\right]^+\sum_i r_i\theta_i q_i\right\}+c_o\kappa p\sqrt{\lambda/\mu}=\eta^\star \quad \text{for} \quad z>z_0^\star, \tag{15}$$

$$v_0(z_0^\star)=v_1(z_0^\star) \quad \text{and} \quad v_0(z_1^\star)=v_1(z_1^\star)+C \tag{16}$$

subject to the boundary conditions

$$\lim_{z\to-\infty} v_0'(z)=0 \quad \text{and} \quad \lim_{z\to\infty} v_1'(z)=r_*, \tag{17}$$

plus a set of optimality conditions derived from the "principle of smooth fit":

$$v_0'(z_0^\star)=v_1'(z_0^\star) \quad \text{and} \quad v_0'(z_1^\star)=v_1'(z_1^\star). \tag{18}$$

Because (14) and (15) do not involve the zero-order term, they can be reduced to a pair of first-order differential equations by defining $f_y(z):=v_y'(z), y=0,1$. The observation leads to the consideration of the class of functions $\{f_0(\cdot,\eta);\eta\in\mathbb{R}\}$ where $f_0(\cdot,\eta)$ solves

$$\lambda f_0'(z)+\min_{\boldsymbol{q}\in\mathcal{A}}\left\{b(0,z,\boldsymbol{q})f_0(z)+[z]^+\sum_i r_i\theta_i q_i\right\}-\eta=0 \tag{19}$$

subject to the boundary condition $\lim_{z\to-\infty} f_0(z)=0$, and the function class $\{f_1(\cdot,\eta);\eta\in\mathbb{R}\}$ where $f_1(\cdot,\eta)$ is the solution to the following differential equation

$$\lambda f_1'(z)+\min_{\boldsymbol{q}\in\mathcal{A}}\left\{b(1,z,\boldsymbol{q})f_1(z)+\left[z-\kappa p\sqrt{\lambda/\mu}\right]^+\sum_i r_i\theta_i q_i\right\}+c_o\kappa p\sqrt{\lambda/\mu}-\eta=0 \tag{20}$$

subject to the boundary condition $\lim_{z\to\infty} f_1(z)=r_*$. For future reference, denote the two function classes by $\mathcal{F}_0$ and $\mathcal{F}_1$, respectively. We can also check that (16) is equivalent to

$$\int_{z_0^\star}^{z_1^\star}[f_0(z,\eta^\star)-f_1(z,\eta^\star)]\,\mathrm{d}z=C, \tag{21}$$

and appeal to (18) to obtain

$$f_0(z_0^\star,\eta^\star)=f_1(z_0^\star,\eta^\star)\quad\text{and}\quad f_0(z_1^\star,\eta^\star)=f_1(z_1^\star,\eta^\star). \tag{22}$$

Therefore, the task of constructing a solution to the Bellman equation (13) boils down to seeking variables $\eta^\star$, $z_0^\star$, and $z_1^\star$ such that both (21) and (22) hold.

To spell out the conditions for the existence of the triple $(\eta^\star,z_0^\star,z_1^\star)$, we begin by considering two related control problems, denoted as $\mathcal{P}_0$ and $\mathcal{P}_1$. Specifically, $\mathcal{P}_0$ (and $\mathcal{P}_1$) correspond to setting $Y\equiv 0$ (and $Y\equiv 1$) in the previously formulated DCP, and seeking an optimal drift-rate control as represented by $\boldsymbol{G}$. Problem $\mathcal{P}_0$, in particular, corresponds to the scenario where the service system relies solely on permanent servers, forgoing the option of using on-demand servers. Let $\tilde{v}_0$ and $\tilde{v}_1$ denote the relative value functions associated with $\mathcal{P}_0$ and $\mathcal{P}_1$, respectively. Similar to (13), we expect that the function $\tilde{v}_0$, along with some constant η_0, solves the following differential equation:

$$\lambda\tilde{v}_0''(z)+\min_{\boldsymbol{q}\in\mathcal{A}}\left\{b(0,z,\boldsymbol{q})\tilde{v}_0'(z)+[z]^+\sum_i r_i\theta_i q_i\right\}=\eta_0 \tag{23}$$

subject to the boundary conditions, $\lim_{z\to-\infty}\tilde{v}_0'(z)=0$ and $\lim_{z\to\infty}\tilde{v}_0'(z)=r_*$. Similarly, we expect that the function $\tilde{v}_1$ should satisfy, in conjunction with some constant η_1, the differential equation

$$\lambda\tilde{v}_1''(z)+\min_{\boldsymbol{q}\in\mathcal{A}}\left\{b(1,z,\boldsymbol{q})\tilde{v}_1'(z)+\left[z-\kappa p\sqrt{\lambda/\mu}\right]^+\sum_i r_i\theta_i q_i\right\}+c_o\kappa p\sqrt{\lambda/\mu}=\eta_1 \tag{24}$$

subject to the boundary conditions, $\lim_{z\to-\infty}\tilde{v}_1'(z)=0$ and $\lim_{z\to\infty}\tilde{v}_1'(z)=r_*$.

We interpret η_0 and η_1 as the optimal long-run average costs of $\mathcal{P}_0$ and $\mathcal{P}_1$, respectively. Also, from our definitions of $\mathcal{F}_0$ and $\mathcal{F}_1$, we can see that the two constants, η_0 and η_1, if they exist, are such that $\lim_{z\to\infty}f_0(z,\eta_0)=r_*$ and $\lim_{z\to-\infty}f_1(z,\eta_1)=0$. Our next result is concerned with the properties of the function classes, $\mathcal{F}_0$ and $\mathcal{F}_1$, as well as the solvability of Equations (23) and (24).

LEMMA 1. *(i) The function class $\mathcal{F}_0$ is well-defined, and $f_0(z,\eta)<f_0(z,\eta')$ for all z if $\eta<\eta'$; in particular, there exists a unique $\eta_0>0$ such that $f_0(z,\eta_0)$ is strictly increasing and $\lim_{z\to\infty}f_0(z,\eta_0)=r_*$, (ii) The function class $\mathcal{F}_1$ is also well-defined, and $f_1(z,\eta)>f_1(z,\eta')$ for all z if $\eta<\eta'$; in particular, there exists a unique η_1 such that $f_1(z,\eta_1)$ is strictly increasing and $\lim_{z\to-\infty}f_1(z,\eta_1)=0$.*

Intuitively, the constant $\eta^\star$, if it exists, should satisfy $\eta^\star\le\bar{\eta}:=\min(\eta_0,\eta_1)$, since both $Y\equiv 0$ and $Y\equiv 1$ are admissible strategies. The following result is concerned with the number of cross points that $f_0(\cdot,\eta)$ and $f_1(\cdot,\eta)$ can have for every $\eta<\bar{\eta}$.

PROPOSITION 1. *(i) There exists $\underline{\eta}\le\bar{\eta}$ such that $f_0(\cdot,\eta)$ and $f_1(\cdot,\eta)$ do not intersect on $[-\infty,\infty]$ for $\eta<\underline{\eta}$, intersect but do not cross (or touch) on $[-\infty,\infty]$ for $\eta=\underline{\eta}$, and cross at least twice for $\eta\in(\underline{\eta},\bar{\eta})$. In particular, when $\eta_0\le c_o\kappa p\sqrt{\lambda/\mu}-(\beta+\kappa p)\sqrt{\lambda\mu}r_*$, $f_0(\cdot,\eta)$ and $f_1(\cdot,\eta)$ do not intersect for all $\eta<\bar{\eta}$ (i.e., $\underline{\eta}=\bar{\eta}$). (ii) The two functions $f_0(\cdot,\eta)$ and $f_1(\cdot,\eta)$ can cross at most twice for $\eta\in(\underline{\eta},\bar{\eta})$.*

REMARK 1. In fact, the function graph of $f_1(\cdot,\eta)$ lies uniformly above that of $f_0(\cdot,\eta)$ for every $\eta<\underline{\eta}$. If, in addition, we have $\eta_0 \le c_o\kappa p\sqrt{\lambda/\mu}-(\beta+\kappa p)\sqrt{\lambda\mu}r_*$, then the above statement is extended to hold for every $\eta<\bar{\eta}$ (i.e., $\underline{\eta}=\bar{\eta}$).

REMARK 2. The proof of part (ii) of Proposition 1 relies on certain convexity/concavity properties of the function families $\mathcal{F}_0$ and $\mathcal{F}_1$.

According to Proposition 1, if $\underline{\eta}=\bar{\eta}$, then one can not find such $\eta<\bar{\eta}$ that the two functions, $f_0(\cdot,\eta)$ and $f_1(\cdot,\eta)$ will cross. We interpret this case to mean that, regardless of how small the fixed cost C is, switching is never optimal. However, if $\underline{\eta}<\bar{\eta}$, then $f_0(\cdot,\eta)$ and $f_1(\cdot,\eta)$ will cross at exactly two points for all $\eta\in(\underline{\eta},\bar{\eta})$. We interpret this case to be that, for a sufficiently small fixed cost, it is optimal to switch between the two modes from time to time. Also, we note that in the latter scenario, $f_0(\cdot,\bar{\eta})$ and $f_1(\cdot,\bar{\eta})$ not only touch at $z=-\infty$ or $z=\infty$ but also cross at some finite point if $\eta_0\neq\eta_1$; in the case that $\eta_0=\eta_1$, $f_0(\cdot,\bar{\eta})$ and $f_1(\cdot,\bar{\eta})$ must touch at both $z=-\infty$ and $z=\infty$. In either case, denote by $\bar{z}_0$ and $\bar{z}_1$ the two points of intersection and define $\bar{C}:=\int_{\bar{z}_0}^{\bar{z}_1}[f_0(z,\bar{\eta})-f_1(z,\bar{\eta})]\mathrm{d}z$. Assuming $\underline{\eta}<\bar{\eta}$ for non-triviality, the next result characterizes the optimal joint control strategy for the DCP (12).

THEOREM 1. *Suppose $C\le\bar{C}$. Then (i) there exists a triple $(\eta^\star,z_0^\star,z_1^\star)$ satisfying (21)–(22), and (ii) the joint control strategy $(Y^\star,\boldsymbol{G}^\star)$, where $Y^\star$ is characterized by $(z_0^\star,z_1^\star)$ and $\boldsymbol{G}^\star$ is of the form*

$$\boldsymbol{G}^\star(t):=\arg\min_{\boldsymbol{q}\in\mathcal{A}}\sum_i(r_i-v_z(Y^\star(t),Z(t)))\theta_i q_i, \tag{25}$$

is average-cost optimal for the DCP (12).

REMARK 3. Since $f_0(\cdot,\eta)$ is increasing in η and $f_1(\cdot,\eta)$ is decreasing in η by Lemma 1, the intersection area of these two functions (given that they cross) must be increasing in η. Hence, the intersection area achieves its maximum $\bar{C}$ when $\eta=\bar{\eta}$, and for any $C<\bar{C}$, there must exist $\eta^\star<\bar{\eta}$ such that (21) is satisfied. To provide an intuitive understanding of Theorem 1, we have displayed in Figure EC.4 of the e-companion the dynamics of these two functions and their intersected area as η increases to $\bar{\eta}$.

Assuming $C=\bar{C}$, we can infer from the proof of Theorem 1 that the optimal $\eta^\star=\bar{\eta}=\min(\eta_0,\eta_1)$. When $\bar{\eta}=\eta_0=\eta_1$, $\bar{z}_0=-\infty$ and $\bar{z}_1=\infty$, implying that the system should remain in its initial mode indefinitely. If $\bar{\eta}=\eta_1<\eta_0$, $\bar{z}_0=-\infty$ and $\bar{z}_1$ is finite, indicating that it is optimal to stay in the "on" mode, from the start, or to switch from "off" to "on" once Z exceeds $\bar{z}_1$, with a cost identical to $\mathcal{P}_1$. If $\bar{\eta}=\eta_0<\eta_1$, $\bar{z}_0$ is finite and $\bar{z}_1=\infty$, suggesting that it is optimal to stay in the "off" mode from the start or to switch from "on" to "off" once Z falls below $\bar{z}_0$, with a cost equivalent to $\mathcal{P}_0$. For $C>\bar{C}$, we interpret $\eta^\star=\bar{\eta}=\min(\eta_0,\eta_1)$ as the optimal strategy, and sequential switching is never optimal. Thus, we should adopt the optimal policy of either $\mathcal{P}_0$ (if $\eta_0<\eta_1$) or $\mathcal{P}_1$ (if $\eta_1<\eta_0$). The following proposition formalizes the last part of the foregoing discussion.

PROPOSITION 2. *If $C > \bar{C}$, then $\eta^\star = \bar{\eta} = \min(\eta_0, \eta_1)$. If $\eta_0 < \eta_1$, the optimal joint control strategy is characterized as $Y \equiv 0$ and $\boldsymbol{G}_0^\star := \arg\min_{\boldsymbol{q}\in\mathcal{A}} \sum_i (r_i - \tilde{v}_0'(Z(t)))\theta_i q_i$. On the other hand, if $\eta_0 \geq \eta_1$, the optimal joint control strategy is characterized as $Y \equiv 1$ and $\boldsymbol{G}_1^\star := \arg\min_{\boldsymbol{q}\in\mathcal{A}} \sum_i (r_i - \tilde{v}_1'(Z(t)))\theta_i q_i$.*

4.2. Approximating the First-Stage Problem

It noteworthy that the optimal objective value of the DCP (12) developed earlier depends on both β and κ. In this light, we write $\eta^\star(\beta, \kappa)$ to denote the optimal objective value of the DCP with given parameters β and κ. We can then approximate the first-stage problem defined by (1) as one that chooses β and κ to

$$\text{minimize} \quad c_p \beta \sqrt{\lambda/\mu} + \eta^\star(\beta, \kappa). \tag{26}$$

Specifically, the expression in (26) is obtained by first replacing N_0 and K in (1) with their expressions in (3) and (4), respectively, replacing $\Gamma(N_0, K)$ with its "diffusion approximation" $\eta^\star(\beta, \kappa)$, and finally subtracting the resulting expression by the constant $c_p\lambda/\mu$.

5. Policy Recommendation

In this section, we describe our solution procedure to the two-stage decision problem based on the results from the previous section. Specifically, we provide staffing and scheduling policy recommendations for the second-stage problem based on the DCP solution.

5.1. First Stage

Note that each (N_0, K) corresponds to a pair of parameters (β, κ) that can be directly plugged into (26) to obtain an approximate objective value of the first-stage problem (centered by $c_p\lambda/\mu$). Thus, given a pool of candidate values for (N_0, K), one can approximately evaluate their respective first-stage objective values and seek a pair that leads to the minimum objective value.

5.2. Second Stage

With a slight abuse of notation, through this subsection, we denote the pair of parameters obtained from the procedure as described in §5.1 by (β, κ).

5.2.1. Profitability of On-Demand Staffing. With the parameter pair (β, κ), along with other model primitives, the system manager can use a two-layer decision framework to decide the profitability of on-demand staffing:

- In the first layer, she computes the upper bound on wage rates as $\bar{c} := \sqrt{\mu/\lambda}\eta_0/(\kappa p) + (\beta + \kappa p) r_* \mu/(\kappa p)$, where η_0, the theoretical average cost associated with $\mathcal{P}_0$, is uniquely determined by the problem data and can be computed using the algorithm described in §EC.4.2 of the e-companion. If the wage rate satisfies $c_o \geq \bar{c}$, then she knows from Proposition 1 that it is not profitable to make short-term capacity adjustments, *no matter how small the fixed cost C is.*

- If $c_o < \bar{c}$, then she proceeds to the second layer of decision and calculates the upper bound on fixed cost, i.e., $\bar{C}$, using the algorithm described in §EC.4.2 of the e-companion. If $C \geq \bar{C}$, she knows from Theorem 1 that making short-term capacity adjustments is unprofitable at this fixed cost. However, if $C < \bar{C}$, then sequentially switching between different modes becomes profitable.

When on-demand staffing is unprofitable, based on Proposition 2, the system manager uses the static staffing control $Y \equiv 0$ or $Y \equiv 1$ and follows the associated scheduling control prescribed by $\boldsymbol{G}_0^\star$ or $\boldsymbol{G}_1^\star$, whichever produces the lower average cost; these scheduling controls (for the original system) are discussed in §5.2.3 below. Otherwise, she can obtain an optimal solution $(f_0(\cdot,\eta^\star), f_1(\cdot,\eta^\star), \eta^\star, z_0^\star, z_1^\star)$ of the Bellman equation using the algorithm described in §EC.4.3 of the e-companion. In what follows, we translate the optimal solution $(f_0(\cdot,\eta^\star), f_1(\cdot,\eta^\star), \eta^\star, z_0^\star, z_1^\star)$ to an implementable joint on-demand staffing and scheduling policy for the original system. Moreover, we use $x_0^\star := z_0^\star + N_0$ and $x_1^\star := z_1^\star + N_0$ to denote the switching thresholds for the number-in-system process, namely, X.

5.2.2. On-Demand Staffing Control. When the system is in "off" mode, it should remain in this mode until the number of jobs in the system, X, increases to $\lceil x_1^\star \rceil$, at which point the best action is to switch to "on" mode. Similarly, when the system is in "on" mode, the best action is to switch from "on" to "off" mode only when X decreases to $\lfloor x_0^\star \rfloor$.

5.2.3. Scheduling Control. When the system is in "off" mode and $X = x$, the *target* queue length distribution vector $q_0^\star(x)$ prescribed by the control $\boldsymbol{G}^\star$ in Theorem 1 is

$$q_{0,i}^\star(x) = \begin{cases} 1, & i = i_0^\star(x) \\ 0, & i \in \{1,\dots,I\} \setminus i_0^\star(x) \end{cases}, \quad \text{where} \quad i_0^\star(x) := \underset{i\in\{1,\dots,I\}}{\arg\min} \left\{ r_i\theta_i - \theta_i f_0(x - N_0, \eta^\star) \right\}. \tag{27}$$

Intuitively, $i_0^\star(x)$ is the "cheapest" class. When a server becomes available and there are jobs waiting, i.e., $x > N_0$, the system manager assigns processing priority to classes in descending order of the queue length discrepancy $Q_i - (x - N_0)q_{0,i}^\star(x)$, $i \in 1,\dots,I$. By the definition of $i_0^\star(x)$, the above scheduling control is equivalent to assigning lowest priority to the class $i_0^\star(x)$, and assigning priority to other classes in descending order of their respective queue length Q_i. Similarly, when the system is in "on" mode and $X = x$, the *target* queue length distribution vector $q_1^\star(x)$ is obtained by replacing $f_0(x - N_0, \eta^\star)$ in (27) by $f_1(x - N_0, \eta^\star)$; that is,

$$q_{1,i}^\star(x) = \begin{cases} 1, & i = i_1^\star(x) \\ 0, & i \in \{1,\dots,I\} \setminus i_1^\star(x) \end{cases}, \quad \text{where} \quad i_1^\star(x) := \underset{i\in\{1,\dots,I\}}{\arg\min} \left\{ r_i\theta_i - \theta_i f_1(x - N_0, \eta^\star) \right\}.$$

When a server becomes available and there are jobs waiting, the system manager gives the class $i_1^\star(x)$ the lowest priority and gives priority to other classes in descending order of their respective queue lengths Q_i. An actionable job scheduling rule can be derived in an analogous fashion based on the control $\boldsymbol{G}_0^\star$ or $\boldsymbol{G}_1^\star$ for the situation where the system is expected to stay in "off" mode ($Y \equiv 0$) or "on" mode ($Y \equiv 1$) forever, i.e., when on-demand staffing is unprofitable.

6. Discussions

In this section, we discuss some key points related to our modeling and solution framework.

6.1. Assumptions on Arrival and Service Processes

Up to this point, we have relied on several assumptions regarding the arrival of job requests and the completion of service processes. However, in terms of the diffusion analysis, some of these assumptions can be readily relaxed to achieve greater generality.

First, rather than assuming Poisson arrivals, we can assume that the job arrival process for each class follows a renewal process. Specifically, we can suppose that the inter-arrival times for class-i jobs are independent and identically distributed (i.i.d.) random variables with a mean of $1/\lambda_i$ and a coefficient of variation of c_i. With this new assumption in mind, we can use strong approximations for renewal processes to get

$$A_i(t) = \lambda_i t + c_i \sqrt{\lambda_i} \hat{A}_i(t) + \epsilon_i^a(t) \quad \text{for} \quad i = 1, \ldots, I,$$

where we have abused notation to let $A_i(t)$ denote the number of class-i jobs that arrived up to t and each $\hat{A}_i$ represent a standard Brownian motion. By making this generalization, we arrive at a new volatility parameter for our approximating diffusion, denoted by $\sigma := \sqrt{\lambda + \sum_i c_i \lambda_i}$. This parameter replaces the previous volatility parameter, $\sqrt{2\lambda}$, as seen in (9).

Second, all job classes are thus far assumed to have statistically identical service times. However, in reality, different classes may have unique service requirements. When studying scheduling problems in the Halfin-Whitt regime, papers such as Harrison and Zeevi (2004) and Atar et al. (2004) have shown that that allowing for class-dependent service rates can present an immense analytical challenge, as it involves solving a complex partial differential equation. This same observation has been made in other works, such as Gurvich et al. (2008) and Kim et al. (2018), where the authors assume class-independent services to gain clearer insights into the associated decision problems. In this paper, we share their intuition that identical service rates eliminate the need to distinguish between job classes during service, making the number of jobs in the system a good proxy for the system state. Nevertheless, to properly account for the possibility of varying service requirements across classes, we can specify that each class i has its own service rate μ_i, which satisfies

$$\mu_i = \mu + \hat{\mu}_i / \sqrt{\lambda}, \tag{28}$$

where $\hat{\mu}_i$ is a constant that does not change with λ. The scaling condition in (28) is inspired by the "perturbation" strategy considered in Maglaras and Zeevi (2004). Intuitively, this scaling condition suggests all service rates surround and are close to an "average" μ. Defining $a_i := \lambda_i / \lambda$, the difference

between μ_i and μ, as implied by (28), adds an additional term $-a_i\hat{\mu}_i\sqrt{\lambda}$ to the drift of the corresponding controlled diffusion. Thus, the extension (28) produces a controlled diffusion with a drift-rate function

$$\check{b}(y,z,\boldsymbol{q}) := -\beta\sqrt{\lambda\mu} - \sum_{i=1}^{I} a_i\hat{\mu}_i\sqrt{\lambda} - \kappa p\sqrt{\lambda\mu}y + \mu\left[z-\kappa p\sqrt{\lambda/\mu}y\right]^{-} - \sum_{i=1}^{I}\theta_i\left[z-\kappa p\sqrt{\lambda/\mu}y\right]^{+} q_i.$$

Alternatively, for the purpose of practical implementation, one can use "moment matching," as proposed in Kim et al. (2018) to approximate the service rate of each class by average weighted service across all classes. This involves selecting μ such that

$$\frac{1}{\mu} = \sum_i \frac{a_i}{\mu_i},$$

and then solving the DCP with the class-independent service rate μ. Numerical studies in Kim et al. (2018) validate the effectiveness of this approach. Our own numerical studies confirm their findings, showing that moment-matching is even more effective than the approximation from (28).

Third, we can allow on-demand servers to have a service rate $\tilde{\mu}$ that is smaller than or equal to μ, as opposed to assuming all servers to have identical service rates. This generalization reflects the prospect that on-demand agents may have less experience and, as a result, work at a slower pace than permanent staff. With approximately $\kappa p\sqrt{\lambda/\mu}$ on-demand servers available whenever the system is in the "on" mode, these servers collectively provide an additional service capacity of $\tilde{\mu}\kappa p\sqrt{\lambda/\mu}$ if they are busy. Suppose we always route incoming jobs to an on-demand server whenever both permanent and on-demand servers are idle. (Previously, we required incoming jobs to be routed to a permanent server if both types of servers were idle.) With this change, we can update the drift-rate function in the diffusion approximation. We denote the new function as $\tilde{b}$, which replaces b in (9) and is defined as

$$\tilde{b}(y,z,\boldsymbol{q}) := -\beta\sqrt{\lambda\mu} - \tilde{\mu}\kappa p\sqrt{\lambda/\mu}y + \mu\left[z-\kappa p\sqrt{\lambda/\mu}y\right]^{-} - \sum_{i=1}^{I}\theta_i\left[z-\kappa p\sqrt{\lambda/\mu}y\right]^{+} q_i.$$

Notably, none of the aforementioned extensions results in a Bellman equation that is fundamentally different from the one in (13). As a result, our analytical results for the base case model essentially carry over to those extensions.

6.2. About Patience-Time Distributions

In their study of the job scheduling problem within the framework of a multi-class many-server queue, Kim et al. (2018) consider approximations for class-specific reneging processes. These approximations enable the development of a tractable DCP that accommodates general distributions for the patience times of each class. Notably, these approximations can be effectively applied to our second-stage problem as well. Specifically, if the patience times of jobs in class i are distributed according to a well-defined hazard-rate function h_i, then the reneging process for class i can be approximated using

hazard rate scaling techniques as developed in Reed and Ward (2008), Reed and Tezcan (2012) to obtian

$$R_i(t) \approx \lambda_i \int_0^t \zeta_i \left(\frac{Q_i(u)}{\lambda_i} \right) \mathrm{d}u,$$

where $\zeta_i(\cdot)$ is defined as $\zeta_i(x) := \int_0^x h_i(y)\mathrm{d}y$. This results in a DCP whose solution can be described by the Bellman equation:

$$\begin{aligned} \min \Bigg\{ \lambda v_{zz}(y,z) + \min_{\boldsymbol{q}\in\mathcal{A}} \Bigg\{ \bar{b}(y,z,q) v_z(y,z) + \sum_i r_i \lambda_i \zeta_i \left(\left[z - \kappa p \sqrt{\lambda/\mu} y \right]^+ q_i/\lambda_i \right) \Bigg\} \\ + c_o \kappa p \sqrt{\lambda/\mu} y - \eta^\star,\ v(1,z) + C - v(0,z),\ v(0,z) - v(1,z) \Bigg\} = 0 \end{aligned} \tag{29}$$

subject to $\lim_{z\to-\infty} v_z(y,z) = 0$ and $\lim_{z\to\infty} v_z(y,z) = r_*$ for $r_* := \min_i r_i$, where

$$\bar{b}(y,z,\boldsymbol{q}) := -\beta\sqrt{\lambda\mu} - \kappa p \sqrt{\lambda\mu} y + \mu \left[z - \kappa p \sqrt{\lambda/\mu} y \right]^- - \sum_{i=1}^I \lambda_i \zeta_i \left(\left[z - \kappa p \sqrt{\lambda/\mu} y \right]^+ q_i/\lambda_i \right).$$

As in §4.1.2, we can identify two functions from (29) denoted as $\bar{f}_0$ and $\bar{f}_1$, which fulfill roles similar to those of f_0 and f_1. However, we are unable to achieve a result similar to that in Proposition 1 for $\bar{f}_0$ and $\bar{f}_1$. Consequently, we are unable to draw any analytical conclusions about the potential structure of the switching rule. Indeed, the proof of Proposition 1 relies on the assumption of exponential patience times, which enables the domain of the functions f_0 and f_1 to be partitioned into a finite number of sets. These equations can be solved in closed form on each set because they belong to the class of linear first-order differential equations on each set. Unfortunately, it is unclear how to extend the arguments presented in the proof of Proposition 1 to obtain a similar result for $\bar{f}_0$ and $\bar{f}_1$.

Nonetheless, our numerical explorations in §EC.5.1 seem to indicate positive outcomes. In that section, we employed numerical methods to solve the DCP solutions under various patience time distributions, all of which belong to the Weibull family and share a common mean value. These distributions have constant, decreasing, and increasing hazard rates. Our results show that the functions $\bar{f}_0$ and $\bar{f}_1$ maintain the desired structure, enabling us to solve the Bellman equation numerically. We also observed that when the patience time exhibits an increasing (or decreasing) hazard rate, the average cost of the system and the profitability of on-demand staffing tend to be lower (or higher) and higher (or lower), respectively, than when the patience time follows an exponential distribution with a constant hazard rate. Therefore, exploring the effects of incorporating more general patience-time distributions on on-demand staffing decisions could be a promising direction for future research.

6.3. Incorporating Holding Costs

Apart from reducing call abandonment, call centers may also care about the negative experiences that customers may encounter while waiting for service. To effectively model these negative experiences, a holding cost can be introduced to each job in the queue. Atar et al. (2010, 2011) consider a multi-class, multi-server Markovian queueing system with class-specific linear holding costs, abandonment penalties, and abandonment rates. They show that this cost structure is equivalent to one that only involves class-specific linear holding costs. This insight carries over to the present study, allowing us to incorporate holding costs through a slight modification to the solution framework, thereby saving the need to overhaul the entire analysis.

Suppose that each job in queue i incurs a holding cost of γ_i per unit of time. Thus, the objective of the second-stage problem is to establish an on-demand staffing and job scheduling policy that minimizes the following:

$$\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}\left[\sum_{i=1}^{I}\int_0^t \gamma_i Q_i(u)\mathrm{d}u+\sum_{i=1}^{I} r_i R_i(t)+c_o\int_0^t \hat{Y}(u)\mathrm{d}u+C\Xi(t)\right],$$

where the processes Q_i, R_i, $\hat{Y}$ and Ξ are defined as before. Since the addition of the holding cost term to the objective does not affect the queueing dynamics, the approximating diffusion remains the same as in (9) when holding costs are included. However, the objective of the corresponding DCP is now to choose $(Y,\boldsymbol{G})$ to minimize

$$\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}\Bigg[\sum_{i=1}^{I}(\gamma_i+r_i\theta_i)\int_0^t\left[Z(u)-\kappa p\sqrt{\lambda/\mu}Y(u)\right]^+ g_i(u)\mathrm{d}u \\ +c_o\kappa p\sqrt{\lambda/\mu}\int_0^t Y(u)\mathrm{d}u+C\sum_{u\le t}[\Delta Y(u)]^+\Bigg].$$

Accordingly, the Bellman equation associated with the corresponding DCP seeks to find some pair (v,η) that satisfies

$$\min\Bigg\{\lambda v_{zz}(y,z)+\min_{\boldsymbol{q}\in\mathcal{A}}\Bigg\{b(y,z,\boldsymbol{q})v_z(y,z)+\left[z-\kappa p\sqrt{\lambda/\mu}y\right]^+\sum_i(\gamma_i+r_i\theta_i)q_i\Bigg\} \\ +c_o\kappa p\sqrt{\lambda/\mu}y-\eta^\star,\ v(1,z)+C-v(0,z),\ v(0,z)-v(1,z)\Bigg\}=0$$

subject to $\lim_{z\to-\infty}v_z(y,z)=0$ and $\lim_{z\to\infty}v_z(y,z)=\tilde{r}_*$ for $\tilde{r}_*:=\min_i(r_i+\gamma_i/\theta_i)$. It follows that adding holding costs has no fundamental effect on the structure of the Bellman equation, other than replacing each congestion-related cost rate $r_i\theta_i$ in the equation with an effective cost rate $\gamma_i+r_i\theta_i$ and slightly modifying the right boundary condition (to ensure the value function has a polynomial growth rate). As a result, our solution approach and prior analytical results again carry over to the extension with holding costs.

6.4. About the Asymptotic Regime

Although we do not establish a formal notion of asymptotic optimality for our proposed approach, it is based on a diffusion approximation framework that assumes an asymptotic regime where the total demand rate and base capacity, built through permanent staff, grow proportionally to infinity. In this regime, parameters such as μ and θ_i, as well as cost parameters r_i, c_p, and c_o, remain constant. To set the base capacity, we employ the square-root staffing principle given in (3), which is a widely-used rule-of-thumb in call centers. This principle is based on the fact that the stochastic fluctuation of the number of jobs in an $M/M/\infty$ queue is in the order of the square root of the offered load, and thus, square-root staffing serves as a stochastic uncertainty hedge. This insight is formalized in Halfin and Whitt (1981), whose results have been extended to include customer abandonment (Garnett et al. 2002) and other useful extensions.

Borst et al. (2004) demonstrate how cost-related considerations can drive the square-root staffing principle. Specifically, they show that a square-root staffing rule can optimally balance staffing costs against service quality when the staffing cost and delay cost are comparable. Our use of the square-root staffing formula aligns with their findings, since c_p and r_i are comparable and do not scale with λ. Moreover, under the scaling condition (3) and the assumption of exponential patience times, the rate at which a job abandons a queue is proportional to the queue length, and queue lengths are typically in the order of $\sqrt{\lambda}$. This implies that the abandonment cost is also of the order $\sqrt{\lambda}$. Furthermore, because the on-demand capacity created by a pool of on-demand servers serves as an additional hedge against stochastic uncertainty (in addition to the square-root safety staffing), the associated cost of using the on-demand capacity should also be in the order of $\sqrt{\lambda}$. This justifies the scaling condition (4) and implies that C should be of the order $\sqrt{\lambda}$, which we can formalize as $C := \sqrt{\lambda}\hat{C}$, where $\hat{C}$ is a constant that does not scale with λ.

Various authors have proposed alternative staffing rules that lead to different asymptotic regimes. For example, Borst et al. (2004) show that in cases where the staffing cost outweighs (in magnitude) the delay cost, the lowest total cost is achieved in the efficiency-driven or overloaded regime. In this regime, service capacity falls short of demand by an order of magnitude greater than what the square-root staffing principle suggests, and job delays are much longer. An efficiency regime can also arise under constraint satisfaction, as in Mandelbaum and Zeltyn (2009), where the consideration of the tail probability of delay leads to an alternative staffing rule that makes both fluid- and diffusion-based analyses relevant; see also Liu and Whitt (2012). Several studies have suggested that demand uncertainty may motivate deviations from the square-root staffing principle in capacity planning decisions. For instance, Whitt (2006) investigates a fluid-based staffing approach that accommodates arrival-rate uncertainty and server absenteeism. Meanwhile, Bassamboo and Randhawa

(2010) propose a newsvendor-based method that is effective when the magnitude of arrival-rate uncertainty exceeds stochastic variability. Their staffing rule aims to meet average demand while guarding against arrival-rate uncertainty. More recently. Hu et al. (2021) develop a two-stage staffing problem that permits larger arrival-rate uncertainty than stochastic variability. They utilize a solution method that incorporates the stochastic fluid approximation introduced by Harrison and Zeevi (2005), which accounts for system stochasticity during the surge stage. If we incorporate demand uncertainty into our model, similar to the approach taken by Hu et al. (2021), it may result in an asymptotic regime that differs from what is suggested in (3). Therefore, a different analytical approach, such as fluid analysis, may be necessary. Nonetheless, given the significant empirical evidence indicating the potential existence of inherent stochastic uncertainty in the demand rate, exploring this direction could be a valuable avenue for future research.

6.5. About Policy Recommendations

Our methodology for seeking an asymptotically optimal control policy follows a well-established procedure pioneered by Harrison (1988), which has sparked extensive research in the field. This procedure can be summarized in three steps in the context of the present study: (1) approximate a queueing control problem by a DCP, (2) solve the DCP, and (3) interpret the solution of the DCP in the context of the original queueing system to obtain an effective control policy. The first step of the procedure, in particular, requires a condition of balanced heavy loading, which allows for the replacement of properly scaled processes, typically using a scaling parameter n, with diffusion processes that are believed to represent the heavy-traffic limits of their pre-limit counterparts as n approaches infinity.

Generally, two approaches can be used to justify the asymptotic correctness of a solution informed by the DCP. The first approach rigorously proves asymptotic optimality by formally justifying the interchange of n and expectation. This method is often highly technical and may warrant a separate paper. Examples of papers that have employed this approach include Atar and Lev-Ari (2018) and Gao and Huang (2022), among others. The second approach involves numerically comparing the solution informed by the DCP to the solution of the original problem. This approach is commonly used in studies following Harrison's (1988) work.

Similar to Kim et al. (2018), our work utilizes a modern approach known as "universal approximation," which places less emphasis on scaling and re-scaling to obtain and interpret the diffusion-based solution as an implementable policy for the original system. The theoretical foundations of this approach are thoroughly discussed by Huang and Gurvich (2018). Our proposed policy is consistent with the solution to the DCP. The policy emerging from the DCP aims to maintain the queue contents at their respective targets, which are solutions to the minimization problem in (25). Although

queue lengths may temporarily deviate from their targets in the actual system, queues that exceed their targets do not experience heavy traffic under our proposed scheduling rule (as these queues exclusively seize all service resources, which are not shared with classes whose queue lengths fall short of the targets). As a result, these queues have a net output rate of order λ, which allows the "queue imbalance", which is of order $\sqrt{\lambda}$ to be corrected at a rate of order $\sqrt{\lambda}$. Therefore, for large enough λ, corrections can be made almost instantly, showing that the proposed scheduling rule effectively operationalizes the solution to the DCP.

7. Numerical Studies

In §7.1, we present experiment results for a single-class example to demonstrate the accuracy of our proposed diffusion approximation with respect to the exact MDP and to highlight the significant value of on-demand staffing in terms of cost savings in comparison to static staffing. In §7.2 we report numerical findings for a two-class example to show how the on-demand staffing decision and the job scheduling decision can be successfully integrated. We further show that when on-demand staffing is present, the corresponding scheduling decision exhibits an unusual double-switching pattern. In §7.3, we conduct a case study using real data from a US bank call center to illustrate how our proposed framework can be implemented in practice, where we relax several modelling assumptions.

Our simulations are based on 100 i.i.d. replications, which run for a time interval T of length 10^4 units of time. To allow the system to approach steady state, we include a warm-up period of length $T/5$. To estimate the mean of a random variable ξ, we use the sample averages of the 100 values, which are expected to be Gaussian, allowing us to construct a 95% confidence interval (CI).The simulated mean is followed by the width of the CI in parentheses.

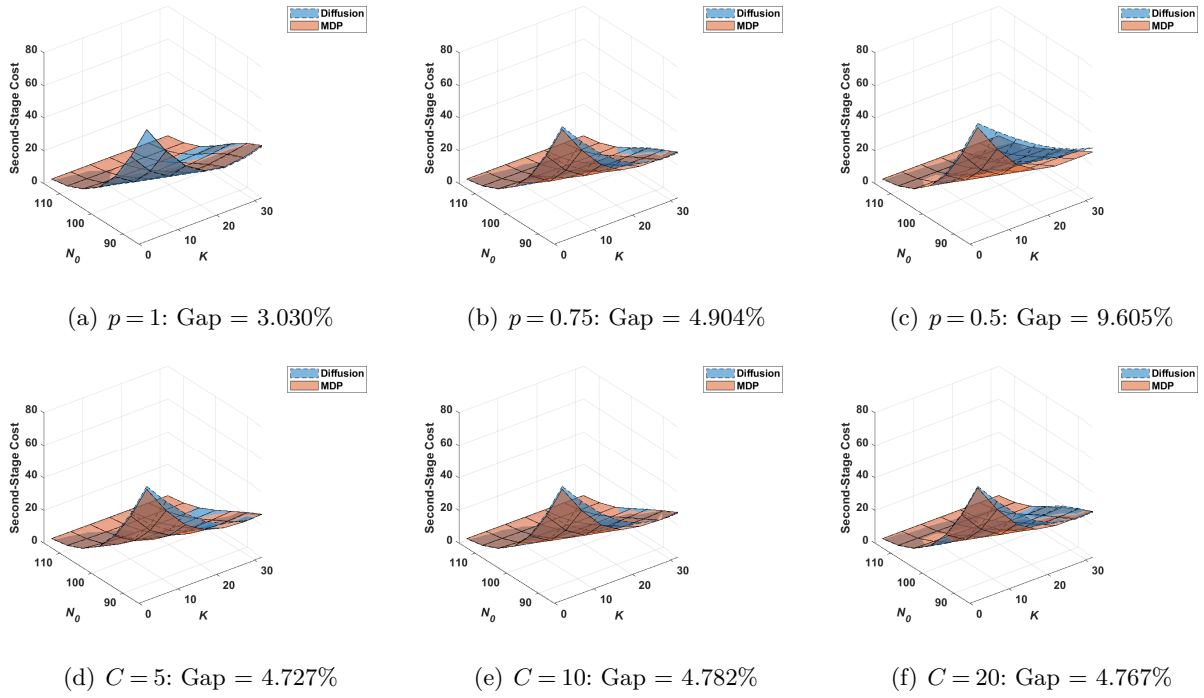
7.1. A Single-Class Example

We assume, without loss of generality (by way of scaling), that the mean service time is $1/\mu = 1$, and the wage rate for each on-demand server is $c_o = 1$. In this single-class scenario, jobs arrive at the system according to a Poisson process with a rate $\lambda = 100$. Each job has an exponentially distributed patience time with a mean of $1/\theta = 2$, and each abandonment incurs a penalty cost of $r = 5$. The base value for an on-demand server's show-up probability is set at $p = 0.75$, while the base switching cost is $C = 15$. To gain more insights, we vary p (or C) while keeping C (or p) fixed at its base value. Finally, we set the time-average cost of hiring a permanent server as $c_p = 1$. Since the nominal load of this system is $\lambda/\mu = 100$, we choose the candidate set for the number of permanent servers N_0 as $\{85, 90, \ldots, 115\}$, and the candidate set for the on-call pool size K as $\{2, 7, \ldots, 32\}$ when solving for the two-stage decision problem defined by equations (1) and (2).

7.1.1. Comparing with exact MDP. For single-class job systems, it is feasible to solve the second-stage problem (2) using the exact MDP approach. The optimal long-run average cost and staffing policy obtained from the MDP serve as a benchmark for testing the accuracy of our proposed diffusion approach in approximating the second-stage problem.

Figure 1 presents the diffusion costs and MDP costs of the second-stage problem for all candidate (N_0, K) under different show-up probability p and switching cost C. Overall, the proposed diffusion approach performs well and produces a small cost gap with respect to the exact MDP in most cases. However, a closer inspection of Figure 1 reveals that the diffusion approximation is less accurate either when N_0 is significantly different from the nominal load or when the show-up behavior of on-demand servers becomes uncertain, particularly when $p = 0.5$. Inaccuracy in the former scenario can be attributed to deviations from the limit regime (3) under which the diffusion process is derived. When N_0 is far from the nominal load, the service system behaves more like it is operating in the efficiency-driven or quality-driven regime. However, this inaccuracy should not impede the diffusion approach from generating high-quality solutions to the first-stage decision problem. This is because it is unlikely that the optimal N_0 would deviate significantly from the nominal load, as demonstrated in Table 1 below.

On the other hand, the inaccuracy in approximation in the latter scenario is due to an unexpected structure of the optimal MDP staffing policy concerning the uncertain show-up behavior of on-demand servers. We report this finding in Figure 2 and discuss it next. Compared to the diffusion staffing policy, which is entirely determined by two policy thresholds, $\lfloor x_0^\star \rfloor$ and $\lceil x_1^\star \rceil$, monitoring the number of jobs in the system, the optimal MDP staffing policy is more complex and depends on both the number of jobs and the number of on-demand servers in the system. The proposed diffusion staffing policy is almost identical to the optimal MDP policy when $p = 1$, as shown in Figures 2(a) and 2(d). However, when $p < 1$, the optimal MDP policy has a more complex structure that is not captured by the diffusion staffing policy, as shown in Figures 2(e) and 2(f). Specifically, the optimal MDP policy includes a consecutive switching policy, where the system switches from "on" to "off" and then back to "on" when the number of on-demand servers who show up to work is deemed insufficient. Although this policy is optimal for the MDP, which assumes that each on-demand server will appear with probability p in each round of invitation, it may perform poorly in practice because an on-demand server that is unavailable in one round of invitation is likely to remain unavailable in the immediately following time instance. In contrast, the diffusion approach leads to a highly interpretable policy with structural insights that is more robust to modeling assumptions. Furthermore, the small cost gap in Figure 1 indicates that the diffusion staffing policy is able to reap most of the benefits of the more complex MDP policy.

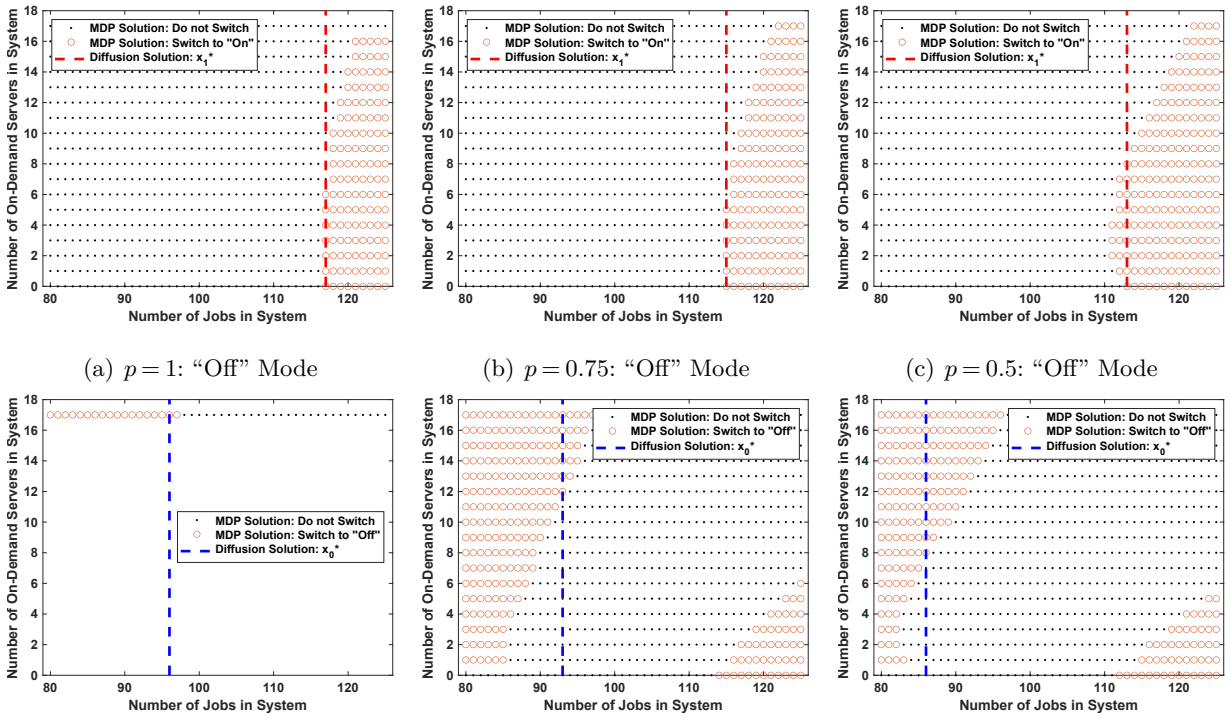


(a) $p = 1$: Gap = 3.030% (b) $p = 0.75$: Gap = 4.904% (c) $p = 0.5$: Gap = 9.605%

(d) $C = 5$: Gap = 4.727% (e) $C = 10$: Gap = 4.782% (f) $C = 20$: Gap = 4.767%

Figure 1 Comparison between the theoretical second-stage costs computed from the proposed diffusion approach and the exact MDP under different show-up probabilities $p \in \{1, 0.75, 0.5\}$ and switching costs $C \in \{5, 10, 15, 20\}$ in the single-class example. The gap represents the absolute difference between the diffusion cost and the MDP cost normalized by the MDP cost, averaged over all (N_0, K) combinations.

After studying the accuracy of the diffusion approximation, we compare the optimal first-stage solutions obtained from the two approaches. Firstly, we compute the theoretical first-stage costs for each candidate (N_0, K) based on the theoretical second-stage costs obtained in Figure 1. Then, we choose the cost-minimizing (N_0, K) under each approach. To investigate the actual cost performances of these obtained solutions, we implement the on-demand staffing policies prescribed by either the diffusion or MDP solution via simulations. The results are summarized in Table 1. We observe that the diffusion approach produces optimal solutions that are similar or identical to those of the MDP approach in all cases. Even when the optimal solutions do not coincide, the resulting simulated first-stage costs are almost the same, with a gap of less than 0.1%. This phenomenon can be attributed to the relatively "flat" first-stage cost surface as a function of (N_0, K) near its minimum point, which makes it insensitive to small perturbations around the optimal solution. However, such flatness is only a local property around the minimum point since the difference between the best-case and worst-case first-stage costs can be quite significant, up to 40% as reported in Figure EC.4 of the e-companion. These findings demonstrate the value of solving the two-stage decision problem and the effectiveness of our proposed diffusion approach in providing near-optimal solutions.



(a) $p=1$: "Off" Mode (b) $p=0.75$: "Off" Mode (c) $p=0.5$: "Off" Mode

(d) $p=1$: "On" Mode (e) $p=0.75$: "On" Mode (f) $p=0.5$: "On" Mode

Figure 2 Comparison between on-demand staffing policies prescribed by the diffusion solution and the MDP solution when the system is operating in the "off" mode (i.e., (a)(b)(c)) and "on" mode (i.e., (d)(e)(f)) under $(N_0, K) = (100, 17)$ and $p \in \{1, 0.75, 0.5\}$ in the single-class example.

Table 1 The optimal first-stage solutions (N_0, K) obtained from the diffusion and MDP approaches, together with the corresponding simulated first-stage costs under different p and C in the single-class example. The simulated first-stage problem cost is a sum of the deterministic capacity planning cost $c_p N_0$ and the simulated second-stage cost. When simulating the second-stage cost under the diffusion (resp. MDP) approach, on-demand staffing policy prescribed by the diffusion (resp. MDP) solution is implemented.

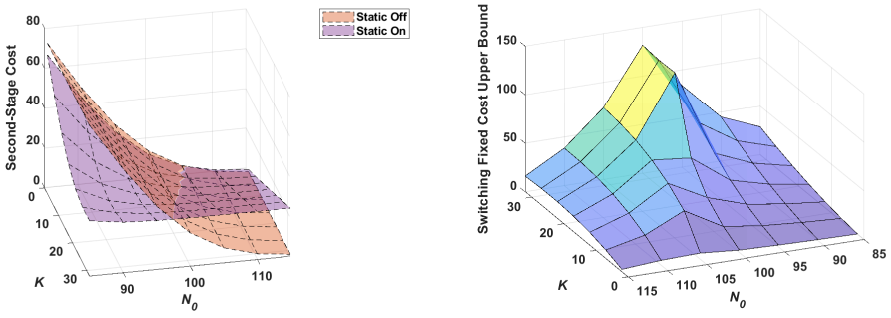
p	C	Optimal First-Stage Solution (N_0, K)		Simulated First-Stage Problem Cost Under Optimal Solution		
		Diffusion	MDP	Diffusion	MDP	Gap (%)
1	15	$(100, 12)$	$(100, 17)$	$100 + 11.160(.0359)$	$100 + 11.170(.0352)$	< 0.1
0.75	5	$(100, 17)$	$(100, 17)$	$100 + 9.271(.0344)$	$100 + 9.174(.0335)$	< 0.1
0.75	10	$(100, 17)$	$(100, 17)$	$100 + 10.354(.0317)$	$100 + 10.321(.0355)$	< 0.1
0.75	15	$(100, 17)$	$(105, 22)$	$100 + 11.211(.0329)$	$105 + 6.228(.0303)$	< 0.1
0.75	20	$(105, 22)$	$(105, 22)$	$105 + 6.581(.0342)$	$105 + 6.540(.0319)$	< 0.1
0.5	15	$(100, 27)$	$(105, 32)$	$100 + 11.247(.0430)$	$105 + 6.222(.0338)$	< 0.1

As an additional note, we experimented with different choices of c_p, the time-average cost of hiring a permanent server, and found that the optimal first-stage solution may prescribe $K = 0$ when c_p is small. This implies that if the cost of a permanent workforce is lower than that of an on-demand workforce, a company should forego the on-demand workforce option entirely. Although we presented the most interesting parameter setting in this study, the decision-maker could readily use our proposed

two-stage decision framework to determine the profitability of an on-demand workforce at the tactical capacity planning level.

7.1.2. Comparing with static staffing policies. At the operational level, on-demand staffing can yield significant cost reductions, regardless of whether the capacity decision (N_0, K) is optimal or not. To demonstrate this, we compare on-demand staffing with two meaningful benchmarks: the static "off" policy and the static "on" policy. The former policy relies solely on the N_0 permanent servers, while the latter always exerts $[Kp]$ on-demand servers in addition to the N_0 permanent servers, where $[x]$ denotes the nearest integer to $x \in \mathbb{R}_+$. We fix p as its base value 0.75. For each candidate (N_0, K), we use the numerical algorithms described in §EC.4.2 of the e-companion to compute the theoretical second-stage costs η_0 and η_1 of the two static policies, as well as the maximum switching cost $\bar{C}$ for on-demand staffing to be worthwhile.

The results are reported in Figure 3. We observe that the static "on" policy has better performance than the static "off" policy when N_0 is small. However, when N_0 is large, the static "off" policy becomes better than the static "on" policy, as the latter continuously incurs additional labor costs at the rate $c_o[Kp]$. For a fixed on-call pool size K, we observed that the maximum switching cost $\bar{C}$ first increases and then decreases as N_0 increases. The reason for this trend is that when N_0 is too small (resp. large), it is preferable to stick to the static "on" (resp. "off") policy to prevent the system from being too congested (resp. idling). Therefore, to make the on-demand staffing policy worthwhile, the switching cost needs to be sufficiently small in these cases.

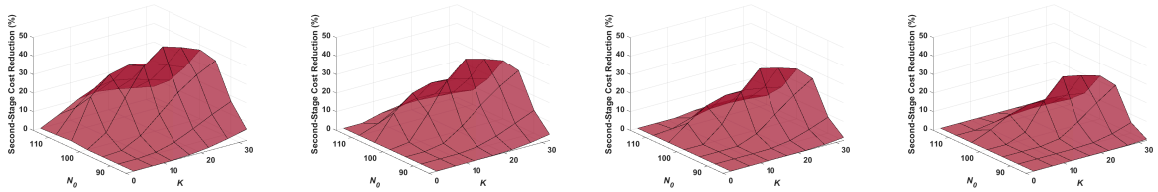


(a) (b)

Figure 3 (a) Theoretical second-stage costs of the static "off" policy and the static "on" policy; and (b) the maximum switching cost $\bar{C}$ for on-demand staffing to be worthwhile in the single-class example.

We utilize the numerical algorithm outlined in §EC.4.3 of the e-companion to calculate the theoretical second-stage cost $\eta^\star$ of the on-demand staffing policy, along with the policy thresholds $\lfloor x_0^\star \rfloor$ and $\lceil x_1^\star \rceil$, for each candidate (N_0, K) where on-demand staffing is found to be worthwhile. Figure 4 displays the percentage reduction in second-stage cost achieved by our proposed policy under different switching

costs. Our findings demonstrate that the on-demand staffing policy results in significant cost savings over a broad range of candidate (N_0, K), even reaching up to 40% in some instances. However, it is not surprising that the cost savings diminish as the switching cost increases.



(a) $C = 5$ (b) $C = 10$ (c) $C = 15$ (d) $C = 20$

Figure 4 Reduction in second-stage cost (%) achieved by the proposed on-demand staffing policy compared to the best possible static staffing policies under different (N_0, K) and switching costs $C \in \{5, 10, 15, 20\}$.

Table 2 Structure and simulated performances of the proposed on-demand staffing policy under $(N_0, K) = (100, 17)$ and different switching cost in the single-class example in comparison to the static staffing policies.

C	**Diffusion Cost**	**Simulated Cost**	**Cost Reduction (%)**	**Staffing Cost**	**Switching Rate**	$(\lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil)$
5	9.077	9.271(.0344)	36.561	5.519(.0196)	0.229(.007)	$(97, 112)$
10	10.196	10.354(.0317)	29.147	6.120(.0198)	0.174(.006)	$(95, 114)$
15	11.060	11.211(.0329)	23.281	6.805(.0232)	0.146(.005)	$(93, 115)$
20	11.505	11.906(.0399)	18.528	7.361(.0254)	0.124(.005)	$(91, 116)$
Static "Off"	16.525	16.496(.0898)	-12.882	0(0)	0(0)	–
Static "On"	14.327	14.614(.0188)	0	13(0)	0(0)	–

We now zoom in the case where $(N_0, K) = (100, 17)$ to conduct a detailed investigation of the structure and simulated (second-stage) performances of both the on-demand and static staffing policies, as presented in Table 2. Column 3 displays the simulated average costs of the three policies, which closely align with their respective theoretical values $\eta^\star, \eta_0$ and η_1 shown in column 2. Column 4 confirms the substantial cost reduction achieved by the proposed on-demand staffing policy. Furthermore, as depicted in columns 5, 6, and 7, we observe that the threshold interval $(\lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil)$ expands as C increases, resulting in less frequent switching between operating modes. However, the simulated staffing costs (including both the wage and switching cost) continue to increase as C increases.

7.2. A Two-Class Example

We examine a two-class system in the following analysis. Class 1 jobs have the same attributes as in the single-class scenario, except that the arrival rate is equally divided between the two classes, such that $\lambda_1 = \lambda_2 = 50$. For class 2 jobs, we assume $\theta_2 = 1.2$ and $r_2 = 3$, making them more expensive $(r_2\theta_2 > r_1\theta_1)$ and more impatient $(\theta_2 > \theta_1)$. The on-demand server's show-up probability and switching

cost are fixed at their base values of $p = 0.75$ and $C = 15$, respectively. All other problem data remains the same as in §7.1.

We first replicate the procedure in §7.1 to solve the first-stage decision problem approximately, using the diffusion approach. This approach returns an optimal solution of $(N_0, K) = (100, 17)$. However, in this scenario, obtaining the MDP-optimal solution becomes challenging for two-class systems. We also repeat the comparison with static staffing policies. As in §7.1, on-demand staffing policies yield significant cost savings over a broad range of (N_0, K) values. Figure EC.5 in the e-companion summarizes these results.

Next, we focus on the two-class system operating under optimal capacity $(N_0, K) = (100, 17)$. Our goal is to examine the joint on-demand staffing and job scheduling decision of this system. We begin by calculating the Bellman equation solution $(\eta^\star, x_0^\star, x_1^\star, f_0(\cdot, \eta^\star), f_1(\cdot, \eta^\star))$ for this system. The on-demand staffing decision is straightforwardly determined by the thresholds $(\lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil)$. To obtain the job scheduling decision, let $\ell := (\theta_2 r_2 - \theta_1 r_1)/(\theta_2 - \theta_1)$. From §5, we know that for a given number of jobs in the system, x, class 2 jobs should be prioritized in the "on" mode if $f_0(x - N_0, \eta^\star) < \ell$. In "off" mode, class 2 should be given priority if $f_1(x - N_0, \eta^\star) < \ell$, while class 1 should be given priority if not. Figures 5(a) and 5(b) display the relationship between the on-demand staffing and scheduling decisions and the Bellman equation solution when the system is in "off" and "on" mode, respectively, where the graphs of the functions $f_0(\cdot, \eta^\star)$ and $f_1(\cdot, \eta^\star)$ have been shifted to the right by N_0, and the green dashed horizontal line represents the line $\ell = (\theta_2 r_2 - \theta_1 r_1)/(\theta_2 - \theta_1)$. For comparison, we also show the scheduling decisions of the two static staffing policies in relation to their respective Bellman equation solutions $f_0(\cdot, \eta_0)$ and $f_1(\cdot, \eta_1)$ in Figures 5(c) and 5(d).

One intriguing discovery in Figure 5 is the unconventional structure of the proposed scheduling rule when the system is in "off" state, as displayed in Figure 5(a). In situations where there is no on-demand staffing, it is widely known that the optimal scheduling rule for exponential abandonment and linear penalty settings is a dynamic index rule. In this rule, service priority may move from one job class to another as the number of jobs in the system rises, and this type of priority switching, if it occurs, can occur only once in a two-class system. This is precisely what is shown in Figures 5(c) and 5(d), where the system utilizes a static staffing policy. However, in the presence of on-demand staffing, priority switching can occur twice when the system is in "off" mode, even though there are only two job classes in the system. In fact, as seen in Figure 5(a), when the system is in "off" mode, priority shifts from class 2 to class 1 when the number of jobs in the system grows to 102, and then switches back to class 2 from class 1 when x reaches 112.

From a mathematical standpoint, the double switching is caused by the non-monotonic behavior of the function $f_0(\cdot, \eta^\star)$, making it possible for the horizontal line $\ell = (\theta_2 r_2 - \theta_1 r_1)/(\theta_2 - \theta_1)$ to intersect $f_0(\cdot, \eta^\star)$ twice. This type of double switching is not possible in a static staffing scenario, since the

functions $f_0(\cdot, \eta_0)$ and $f_1(\cdot, \eta_1)$ are guaranteed to be monotonically increasing. Such double switching suggests a rich interplay between the staffing and scheduling decisions. The managerial insights behind this interplay can be summarized as follows: In a system without on-demand staffing control, the proposed scheduling rule dictates that priority be shifted away from the more expensive and impatient class and toward a less expensive but more patient class in the hopes of the queue being drained faster due to abandonment. However, in the presence of on-demand staffing, the first priority switching occurs for the same reason as before. The second priority switching, which involves shifting priority from the cheaper to the more expensive class, occurs because it is anticipated that as the queue grows longer, a group of on-demand servers will join the system, resulting in a faster depletion of the queue. Therefore, the optimal scheduling rule anticipates this increase in workforce and shifts the priority back to the more expensive but impatient class.

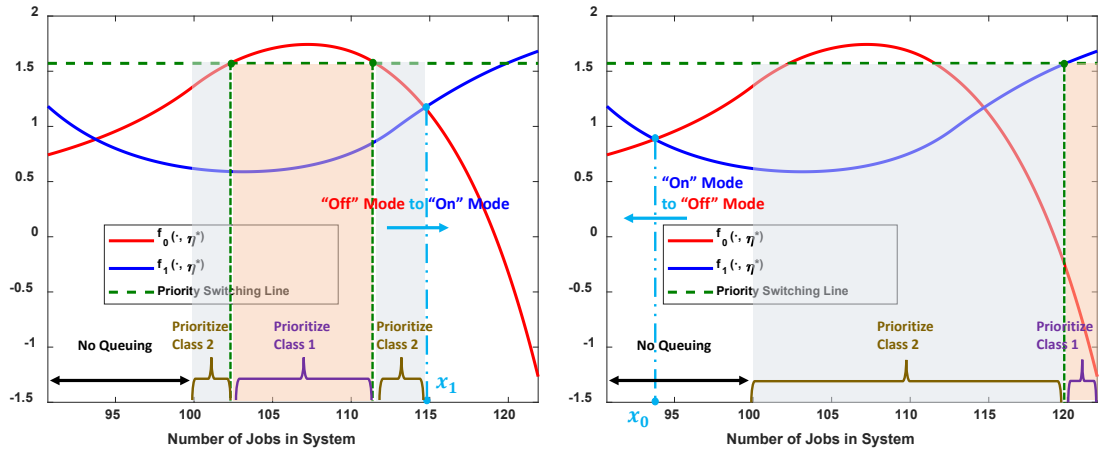
Finally, we validate the benefit of this unusual double-switching scheduling rule by comparing it with a competitive benchmark. This benchmark adopts the same staffing rule as the proposed on-demand staffing policy but utilizes a different scheduling rule. Specifically, it applies the static "off" (resp. "on") scheduling policy when the system is "off" (resp. "on"), which is henceforth referred to as the static scheduling rule. This enables the benchmark to isolate the impact of the staffing decision and attribute any performance differences to the difference in scheduling rule. The comparison outcomes are presented in Table 3. As in §7.1, simulated costs of the proposed policy and static staffing polices are similar to their respective theoretical diffusion costs $\eta^\star = 10.906, \eta_0 = 12.514$ and $\eta_1 = 14.275$. The first two rows of the table reveal that the double-switching scheduling rule yields a slight but noteworthy reduction in abandonment cost when compared to the static scheduling rule. Additionally, as the last two rows illustrate, the proposed policy outperforms the static staffing policies.

Table 3 Structure and simulated performances of (i) the proposed joint on-demand staffing and scheduling policy, (ii) a benchmarking policy that uses the same staffing rule as the proposed one but a different scheduling rule, and (iii) the two static staffing policies under $(N_0, K) = (100, 17)$ in the two-class example.

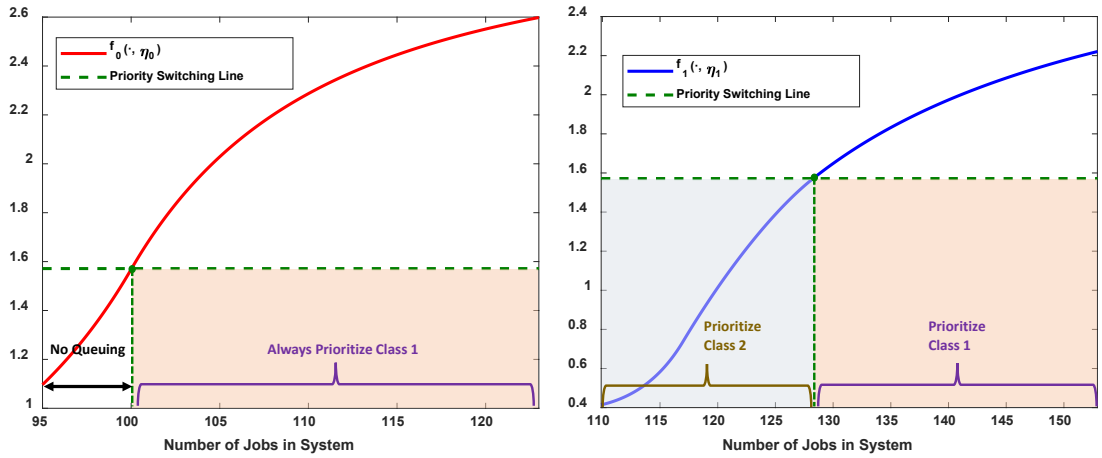
Policy	Simulated Cost	Abandonment Cost	Staffing Cost	$(\lfloor x_0^\dagger \rfloor, \lceil x_1^\dagger \rceil)$	"Off" Scheduling	"On" Scheduling
Joint On-Demand Staffing and Scheduling	10.992(.0364)	5.065(.0200)	5.927(.0248)	(93,115)	Class 2→1 at 102, 1→2 at 112	Class 2→1 at 120
On-Demand Staffing + Static Scheduling	11.094(.0342)	5.156(.0222)	5.938(.0219)	(93,115)	Always prioritize class 1	Class 2→1 at 129
Static "Off"	12.731(.0505)	12.731(.0505)	0(0)	–	Always prioritize class 1	
Static "On"	14.587(.0167)	1.587(.0167)	13(0)	–	Switch priority from class 2→1 at 129	

7.3. A Case Study Using Real Data

We devote the remainder of this section to a case study using a real data set sourced from the Service Enterprise Engineering (SEE) lab at the Technion Institute of Technology. The data set is about a US bank call center that operates 24/7 and receives, on average, 53,237 calls per day on weekdays while receiving an average of 19,742 calls per day on weekends. The data set contains approximately 59



(a) On-Demand Staffing and Scheduling: "Off" Mode (b) On-Demand Staffing and Scheduling: "On" Mode



(c) Static "Off" Scheduling Policy (d) Static "On" Scheduling Policy

Figure 5 Graphical display of (i) the on-demand staffing and scheduling policy, (ii) the static "Off" scheduling policy, and (iii) the static "On" scheduling policy, and their relationships with the Bellman equation solutions $(\eta^\star, x_0^\star, x_1^\star, f_0(\cdot,\eta^\star), f_1(\cdot,\eta^\star))$, $f_0(\cdot,\eta_0)$ and $f_1(\cdot,\eta_1)$ in the two-class example. The double switching in scheduling priority observed in (a) is due to the non-monotonicity of the function $f_0(\cdot,\eta^\star)$, which is never possible in a static staffing scenario because the functions $f_0(\cdot,\eta_0)$ and $f_1(\cdot,\eta_1)$ plotted in (c) and (d), respectively, must be monotonically increasing.

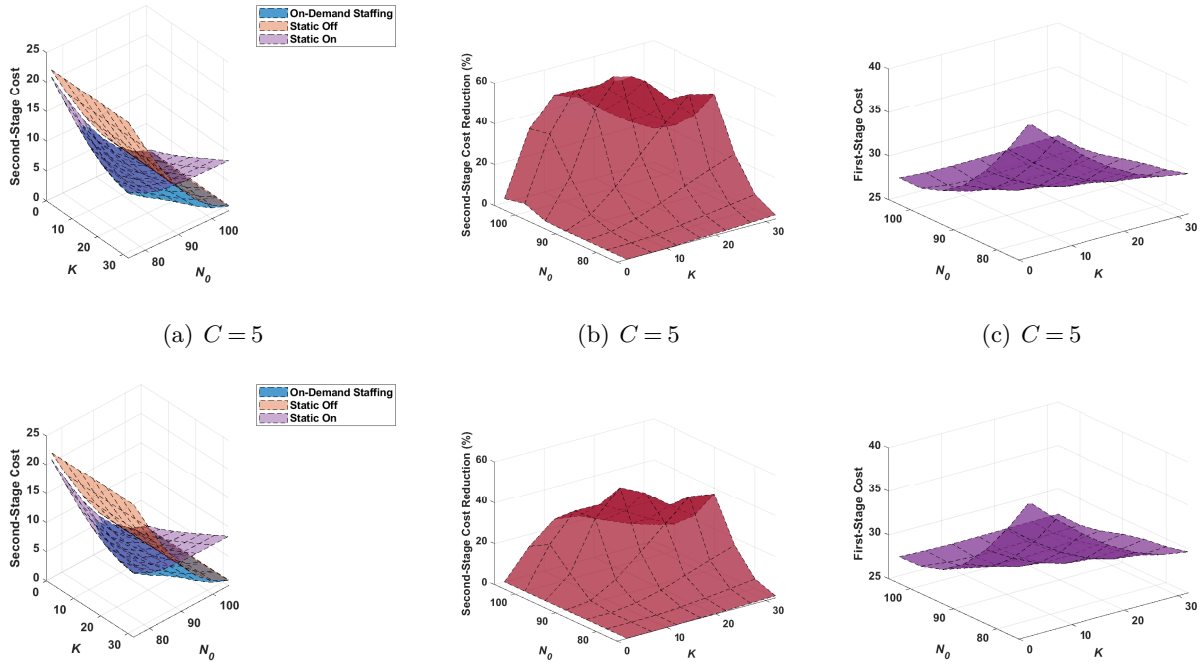
million records of call-level data from March 1, 2001, to December 31, 2001. For each call, this data set records its waiting time in queue, service duration (which is zero if it is abandoned), and service type from one of the six categories. These are Retail, Online Banking, Premier, Business, Consumer Loans, and Telesales.

To illustrate the implementation of our proposed framework, we focus on a two-class setting, where classes 1 and 2 represent retail banking and online banking, respectively. We fit abandonment time distributions for the two classes based on abandoned calls with a waiting time greater than 5 seconds, which occurs in 99.9% of the observations. Figures EC.6(a) and EC.6(b) of the e-companion

demonstrate that the exponential distribution provides a good fit. The estimated abandonment rates are $\theta_1 = 0.66$ and $\theta_2 = 0.96$ calls per minute. Similarly, we use exponential distributions to fit the historical service times shown in Figures EC.6(c) and EC.6(d) of the e-companion, yielding $\mu_1 = 1/4.326$ and $\mu_2 = 1/5.654$ calls per minute. To handle this non-identical service rate setting, we use the moment-matching method, which is discussed in §6.1. Given the significant difference in call volumes between weekdays and weekends, we solve the proposed two-stage decision problem separately for weekdays and weekends. However, we only present the results for weekdays in the following. The arrival rate to the call center on weekdays can be estimated using a Poisson process with a rate of $\lambda = 18.485$ calls per minute. We assume that $\lambda_1 = \lambda_2 = 0.5\lambda$. Since the nominal load of the system is around 92, we choose the candidate set for N_0 as $75, 80, \ldots, 105$, while using the same candidate set for K as in §7.1. The average salary of a call center agent in the USA is approximately \$15 per hour (TalentCom 2023). Therefore, we set $c_p = 0.25$ dollars per minute and $c_o = 1.4c_p$ to reflect the higher wage rate of an on-demand agent compared to a permanent agent. We assume that the show-up probability of on-demand servers is $p = 0.75$. Finally, we set the cost for each abandoned call as $r_1 = r_2 = 18c_o$, which is equivalent to 0.3 hours of salary for an on-demand agent, to reflect that only 1.54% of customers abandoned the queue before an agent was free to serve them in the dataset.

We present results for two different switching costs, $C = 5$ and $C = 10$. Figures 6(a) and 6(d) illustrate that the proposed on-demand staffing policy can significantly reduce costs when the number of permanent servers, N_0, is close to the nominal load. However, when N_0 is too large or too small, the on-demand policy performs similarly to the static "off" or "on" policies, respectively. Figures 6(b) and 6(e) further demonstrate the benefits of on-demand staffing across a range of candidate (N_0, K), particularly when N_0 is near the nominal load. We also report the theoretical first-stage costs of the proposed policy obtained from the diffusion approach in Figures 6(c) and 6(f). We find that the optimal first-stage solutions under both switching costs, $C = 5$ and $C = 10$, are $(N_0, K) = (100, 12)$.

Next, we focus on the capacity setting $(N_0, K) = (100, 12)$ to investigate the performance of the proposed on-demand staffing and scheduling policy in solving the second-stage problem. To this end, we compare the proposed policy, as well as the two static staffing policies, via simulations. We relax the assumption that an on-demand agent will enter the system immediately by testing with different levels of show-up delay after he or she accepts the invitation. The results are summarized in Table 4. Column 3 shows that the simulated costs of different policies are still close to their respective theoretical diffusion costs $[\eta^\star | C = 5] = 1.452, [\eta^\star | C = 10] = 1.677, \eta_0 = 2.225$ and $\eta_1 = 3.530$. However, the cost gaps are not as small as what we have observed in §7.1 and §7.2, due to the extra approximation error resulting from the non-identical service rate setting. From columns 4 and 5, we see that abandonment costs increase significantly as show-up delay increases, while staffing costs slightly decrease. This is because show-up delay may hinder on-demand agents from dealing with unexpected demand surges



(a) $C = 5$ (b) $C = 5$ (c) $C = 5$

(d) $C = 10$ (e) $C = 10$ (f) $C = 10$

Figure 6 (a)(d) Theoretical diffusion second-stage costs (dollars per minute) of the on-demand staffing and scheduling policy and the two static staffing policies, (b)(e)percentage of reduction in second-stage cost achieved by the proposed policy compared to the best possible static staffing police, and (c)(f) theoretical first-stage cost of the proposed policy under different switching costs $C = 5$ and $C = 10$ in the real-data case study. The optimal first-stage solutions when $C = 5$ and $C = 10$ are both $(N_0, K) = (100, 12)$.

in a timely manner. With the presence of show-up delay, it is possible that surges in demand have already passed because of massive abandonment at the moment on-demand agents are able to join the system, which inflates the abandonment cost. On the other hand, on-demand agents tend to stay in the system for a shorter time due to show-up delay. This explains the slight decrease in staffing cost since less wage is paid. Overall, show-up delay harms the effectiveness of on-demand staffing in terms of cost savings. Nonetheless, the proposed on-demand policy is still able to outperform static staffing policies under moderate levels of show-up delay. Compared with the best possible static staffing policies, on-demand staffing leads to an annual savings of up to \$450,964 and \$315,360 when the switching costs $C = 5$ and $C = 10$, respectively. Furthermore, column 6 shows that, on average, on-demand agents are summoned every 19 minutes and 32 minutes when the switching costs $C = 5$ and $C = 10$, respectively. The decrease in the summoning frequency is aligned with the expansion of the threshold interval $(\lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil)$ as C increases, as observed in the last column and in Table 2.

Table 4 Structure and simulated performances of the proposed joint on-demand staffing and scheduling policy under different switching costs and show-up delays compared with the two static staffing policies under $(N_0, K) = (100, 12)$ in the real-world case study. The optimal scheduling rule is to prioritize class 2 for all policies. The cost unit is dollars per minute.

Policy		Simulated Cost	Abandonment Cost	Staffing Cost	Switching Rate	$(\lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil)$
Proposed Policy $C=5$	Zero Delay	1.558(.0167)	0.776(.0111)	0.782(.00796)	0.0527(< .001)	(96,105)
	30-Second Delay	1.861(.0228)	1.187(.0157)	0.674(.00807)	0.0537(< .001)	
	60-Second Delay	2.039(.0220)	1.461(.0169)	0.578(.00647)	0.0540(< .001)	
	90-Second Delay	2.205(.0217)	1.695(.0176)	0.511(.00531)	0.0542(< .001)	
Proposed Policy $C=10$	Zero Delay	1.816(.0188)	1.049(.0128)	0.768(.00977)	0.0313(< .001)	(94,107)
	30-Second Delay	2.047(.0214)	1.354(.0152)	0.693(.00834)	0.0315(< .001)	
	60-Second Delay	2.248(.0240)	1.604(.0176)	0.645(.00759)	0.0323(< .001)	
	90-Second Delay	2.300(.0243)	1.721(.0185)	0.579(.00710)	0.0317(< .001)	
Static "Off"		2.416(.0300)	2.416(.0300)	0(0)	0	–
Static "On"		3.612(.0113)	0.462(.0113)	3.150(0)	0	–

8. Conclusion

A call center's ability to quickly adjust staffing levels with an on-demand workforce presents an interesting problem of optimally balancing labor costs with service levels. This paper advances a two-stage decision model that fully addresses this tradeoff. Since the second-stage problem is particularly challenging to solve exactly, we resort to approximation techniques. Specifically, we analyze the call center's operations under heavy traffic and obtain, as an approximate solution to the second-stage problem, an almost explicit joint on-demand staffing and scheduling strategy. The on-demand staffing rule is a switching policy that determines when to call in or dismiss on-demand agents based on a set of predetermined switching boundaries. The scheduling rule, on the other hand, is a nested threshold rule that specifies the relative urgency of each customer class. In addition, the approximation for the second-stage problem leads to a convenient approximation for the first-stage problem as well. Through extensive numerical experiments, we demonstrate that solving the full two-stage problem can achieve significant cost savings compared to not using on-demand staffing at all.

Our model has thus far assumed a single on-call pool and restricted the manager's choice to two operating modes. However, additional flexibility can be created by allowing different numbers of on-demand agents in the system depending on the current system state. One way to accomplish this is by dividing the total number of on-demand agents into sub-pools, such as Tier-1, Tier-2, and so on. The manager can then send invites first to the Tier-1 pool and, as congestion levels increase, to the Tier-2 pool, and so on. Conversely, as congestion levels drop, agents can be removed from the pool with the highest tier index. We demonstrate this approach in §EC.6 through a numerical study in which we divide a single on-demand pool into two sub-pools and adapt our solution approach to compute switching boundaries that allow the manager to switch between three different modes: permanent staff only, permanent staff plus Tier-1 pool, and permanent staff plus both pools.

This numerical study reveals two persistent patterns that emerge across the various parameter configurations we have tested. First, operating with more than two modes, where on-demand agents

are added or removed in batches based on congestion levels, generates additional value compared to using only two modes. Second, the marginal benefit obtained from additional modes decreases dramatically as the number of modes increases. This observation aligns with the well-known "power of two" phenomenon that arises routinely in the operations research literature. For instance, in a supermarket model that involves load balancing, it has been shown that systems with two choices can perform almost as effectively as a perfect load balancing system with global load knowledge (Mitzenmacher 2001). Additionally, in studying the problem of using dynamic pricing to maximize revenues in queueing systems with price- and delay-sensitive customers, Kim and Randhawa (2018) shows that a simple policy of using only two prices (known as the two-price scheme) can achieve the majority of the benefits of dynamic pricing. Therefore, we suggest that utilizing only two staffing levels may be the most advantageous approach, as it can achieve significant cost savings compared to fixed staffing and is straightforward to implement.

From a theoretical perspective, solving an optimal switching problem with multiple regimes (or modes) is a notoriously difficult problem, even with the heavy-traffic approximation. It has long been observed that for problems that involve more than two regimes, "the analytical approach becomes cumbersome" (Chernoff and Petkau 1978). Recently, Vande Vate (2021) proposed a methodological advance that involves optimally controlling the drift of a Brownian motion with a finite set of drift rates to minimize long-term average cost. Despite benefiting from explicit formulas for the value function and nice cost structure assumptions, the mathematical analysis involved in this approach is still highly complex. Replicating this analysis in our setting with multiple modes may prove even more difficult, primarily because our value function does not even admit an explicit expression.

Through conversations with the management of Company A's call center, we learned that, while additional flexibility is potentially beneficial, it is important to consider the flexibility-usability tradeoff. This design principle posits that as a system's flexibility increases, its usability decreases. Therefore, our recommendation for the use of two modes may not only stem from tractability considerations but also from the need to sacrifice some flexibility for simplicity, ultimately improving the usability of the solution. Nonetheless, an interesting avenue for future research is to investigate the impact of multiple operating modes on the setting under consideration and explore how additional flexibility can help reduce operating costs.

References

Aksin Z, Armony M, Mehrotra V (2007) The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* 16(6):665–688.

Armony M, Ward AR (2010) Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research* 58(3):624–637.

Ata B, Harrison J, Shepp L (2005) Drift rate control of a brownian processing system. *Annals of Applied Probability* 1145–1160.

Ata B, Lee D, Sönmez E (2019) Dynamic volunteer staffing in multicrop gleaning operations. *Operations Research* 67(2):295–314.

Ata B, Shneorson S (2006) Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Management Science* 52(11):1778–1791.

Ata B, Tongarlak MH (2013) On scheduling a multiclass queue with abandonments under general delay costs. *Queueing Systems* 74(1):65–104.

Atar R, Giat C, Shimkin N (2010) The cμ/θ rule for many-server queues with abandonment. *Operations Research* 58(5):1427–1439.

Atar R, Giat C, Shimkin N (2011) On the asymptotic optimality of the cμ/θ rule under ergodic cost. *Queueing Systems* 67(2):127–144.

Atar R, Lev-Ari A (2018) Workload-dependent dynamic priority for the multiclass queue with reneging. *Mathematics of Operations Research* 43(2):494–515.

Atar R, Mandelbaum A, Reiman MI (2004) Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability* 14(3):1084–1134.

Bassamboo A, Randhawa RS (2010) On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations Research* 58(5):1398–1413.

Bassamboo A, Randhawa RS, Zeevi A (2010) Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science* 56(10):1668–1686.

Borst S, Mandelbaum A, Reiman MI (2004) Dimensioning large call centers. *Operations Research* 52(1):17–34.

Chernoff H, Petkau AJ (1978) Optimal control of a brownian motion. *SIAM Journal on Applied Mathematics* 34(4):717–731.

Dong J, Ibrahim R (2020) Managing supply in the on-demand economy: Flexible workers, full-time employees, or both? *Operations Research* 68(4):1238–1264.

Duckworth K, Zervos M (2001) A model for investment decisions with switching costs. *Annals of Applied Probability* 239–260.

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

Gao X, Huang J (2022) Optimal control of make-to-stock systems, available at `https://hjfcuhk.github.io/papers/BCP.pdf`.

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

Ghosh AP, Weerasinghe AP (2010) Optimal buffer size and dynamic rate control for a queueing system with impatient customers in heavy traffic. *Stochastic Processes and Their Applications* 120(11):2103–2141.

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

Gurvich I, Lariviere M, Moreno A (2019) Operations in the on-demand economy: Staffing services with self-scheduling capacity. *Sharing Economy: Making Supply Meet Demand* 249–278.

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

Harrison JM (1988) Brownian models of queueing networks with heterogeneous customer populations. *Stochastic differential systems, stochastic control theory and applications*, 147–186 (Springer).

Harrison JM, Zeevi A (2004) Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Operations Research* 52(2):243–257.

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

Hu Y, Chan CW, Dong J (2021) Prediction-driven surge planning with application in the emergency department. *Submitted to Management Science* .

Huang J, Gurvich I (2018) Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue. *Operations Research* 66(4):1168–1188.

Huang J, Zhang H, Zhang J (2016) A unified approach to diffusion analysis of queues with general patience-time distributions. *Mathematics of Operations Research* 41(3):1135–1160.

Ibrahim R (2018) Managing queueing systems where capacity is random and customers are impatient. *Production and Operations Management* 27(2):234–250.

Kim J, Randhawa RS (2018) The value of dynamic pricing in large queueing systems. *Operations Research* 66(2):409–425.

Kim J, Randhawa RS, Ward AR (2018) Dynamic scheduling in a many-server, multiclass system: The role of customer impatience in large systems. *Manufacturing & Service Operations Management* 20(2):285–301.

Liu Y, Whitt W (2012) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations Research* 60(6):1551–1564.

Lobel I, Martin S, Song H (2023) Employees versus contractors: An operational perspective, forthcoming in Management Science.

Long Z, Shimkin N, Zhang H, Zhang J (2020) Dynamic scheduling of multiclass many-server queues with abandonment: The generalized cμ/h rule. *Operations Research* 68(4):1218–1230.

Low DW (1974) Optimal dynamic pricing policies for an M/M/s queue. *Operations Research* 22(3):545–561.

Ly Vath V, Pham H (2007) Explicit solution to an optimal switching problem in the two-regime case. *SIAM Journal on Control and Optimization* 46(2):395–426.

Maglaras C, Zeevi A (2004) Diffusion approximations for a multiclass markovian service system with "guaranteed" and "best-effort" service levels. *Mathematics of Operations Research* 29(4):786–813.

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

Mitzenmacher M (2001) The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems* 12(10):1094–1104.

Protter PE, Protter PE (2005) *Stochastic differential equations* (Springer).

Reed J, Tezcan T (2012) Hazard rate scaling of the abandonment distribution for the GI/M/n+ GI queue in heavy traffic. *Operations Research* 60(4):981–995.

Reed J, Ward AR (2008) Approximating the GI/GI/1+ GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Mathematics of Operations Research* 33(3):606–644.

Slaugh VW, Scheller-Wolf AA, Tayur SR (2018) Consistent staffing for long-term care through on-call pools. *Production and Operations Management* 27(12):2144–2161.

TalentCom (2023) Call Center Agent average salary in the usa. `https://www.talent.com/salary?job=call+center+agent`, accessed: 2023-03-03.

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

Vande Vate JH (2021) Average cost Brownian drift control with proportional changeover costs. *Stochastic Systems* 11(13):218–263.

Ward AR (2012) Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys in Operations Research and Management Science* 17(1):1–14.

Weerasinghe A (2015) Optimal service rate perturbations of many server queues in heavy traffic. *Queueing Systems* 79(3):321–363.

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

Wu J, Chao X (2014) Optimal control of a Brownian production/inventory system with average cost criterion. *Mathematics of Operations Research* 39(1):163–189.

Yoon S, Lewis ME (2004) Optimal pricing and admission control in a queueing system with periodically varying parameters. *Queueing Systems* 47(3):177–199.

Zeltyn S, Mandelbaum A (2005) Call centers with impatient customers: many-server asymptotics of the M/M/n+ G queue. *Queueing Systems* 51(3):361–402.

Zervos M, Johnson TC, Alazemi F (2013) Buy-low and sell-high investment strategies. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* 23(3):560–578.

Electronic Companion

EC.1. Proofs of Key Results

For notational convenience, we will simply let $p=1$ throughout this section (§EC.1) and the next section (§EC.2). For $p<1$, simply replace all κ's below with κp, which has no effect on the validity of the arguments or procedure used.

Proof of Lemma 1. To prove Lemma 1, we need the following auxiliary results whose proofs are deferred to Appendix EC.2. Part (i) of Lemma 1 follows directly from Lemma EC.1-EC.2.

LEMMA EC.1. *If $\eta<\eta'$, then $f_0(z,\eta)<f_0(z,\eta')$ for all $z\in\mathbb{R}$.*

LEMMA EC.2. *There exists a unique $\eta_0>0$ such that $\lim_{z\to\infty} f_0(z,\eta_0)=r_*$ and $f_0'(z,\eta_0)>0$ for all $z\in\mathbb{R}$.*

For part (ii) of Lemma 1, the existence of η_1 follows directly from Lemma EC.3.

LEMMA EC.3. *There exists a unique $\eta_1>0$ such that $\lim_{z\to-\infty} f_1(z,\eta_1)=0$ and $f_1'(z,\eta_1)>0$ for all $z\in\mathbb{R}$.*

To prove the desired monotonicity of $f_1(z,\eta)$ in η, we consider a class of functions $\{W_\alpha(z,\eta),\eta\in\mathbb{R}\}$ that are solutions to (20) on $[\kappa\sqrt{\lambda/\mu},\infty)$ with no restriction to the right boundary condition but with the left boundary condition $W_\alpha(\kappa\sqrt{\lambda/\mu},\eta)=\alpha$.

LEMMA EC.4. *For each α, there exists a constant $\eta(\alpha)$ such that $\lim_{z\to\infty} W_\alpha(z,\eta(\alpha))=r_*$; the function $W_\alpha(z,\eta(\alpha))$ is unique and continuous. Moreover, $W_\alpha(z,\eta(\alpha))$ is strictly increasing on $[\kappa\sqrt{\lambda/\mu},\infty)$ if $\alpha<r_*$, strictly decreasing on $[\kappa\sqrt{\lambda/\mu},\infty)$ if $\alpha>r_*$, and a constant function r_* if $\alpha=r_*$.*

LEMMA EC.5. *If $\alpha<\alpha'$, then $W_\alpha(z,\eta(\alpha))<W_\alpha(z,\eta(\alpha'))$ for all $z\geq\kappa\sqrt{\lambda/\mu}$.*

LEMMA EC.6. *The mapping $\eta(\cdot)$ is continuous and strictly decreasing. Moreover, $\lim_{\alpha\to\infty}\eta(\alpha)=-\infty$ and $\lim_{\alpha\to-\infty}\eta(\alpha)=\infty$.*

To see that $f_1(z,\eta)$ is decreasing in η, consider $\eta_1<\eta_2$. Let $\eta^{-1}(\cdot)$ denote the inverse mapping, which is well defined by Lemma EC.6, and we have $\eta^{-1}(\eta_1)>\eta^{-1}(\eta_2)$. That $f_1(z,\eta_1)>f_1(z,\eta_2)$ for all $z\geq\kappa\sqrt{\lambda/\mu}$ follows from the fact that $W_{\eta^{-1}(\eta_1)}(z,\eta_1)>W_{\eta^{-1}(\eta_2)}(z,\eta_2)$ for all $z\geq\kappa\sqrt{\lambda/\mu}$ by Lemma EC.5. To show the desired monotonicity of $f_1(z,\eta)$ in η for $z\leq\kappa\sqrt{\lambda/\mu}$, we next consider another class of functions $\{U(z,\eta),\eta\in\mathbb{R}\}$ that are solutions to (20) on $(-\infty,\kappa\sqrt{\lambda/\mu}]$ with the right boundary condition $U(\kappa\sqrt{\lambda/\mu},\eta)=\eta^{-1}(\eta)$. Since $\eta_1<\eta_2$ and $\eta^{-1}(\eta_1)>\eta^{-1}(\eta_2)$, that $U(z,\eta_1)>U(z,\eta_2)$ for all $z\leq\kappa\sqrt{\lambda/\mu}$ follows from the comparison theorem for ordinary differential equations. We concludes that $f_1(z,\eta_1)>f_1(z,\eta_2)$ for all $z\in\mathbb{R}$. □

Proof of Proposition 1. To begin, we need the following results that characterize the behaviors of $f_0(z,\eta)$ and $f_1(z,\eta)$; their proofs are deferred to EC.2. For convenience, define $\bar{\beta} := \sqrt{\lambda\mu}(\beta+\kappa)$.

LEMMA EC.7. *For $0<\eta<\eta_0$, $f_0(z,\eta)$ is strictly increasing on $(-\infty, m_\eta)$ and strictly decreasing on (m_η,∞) for some $m_\eta>0$; for $\eta=0$, $f_0(z,\eta)$ is a constant function 0 on $(-\infty,0]$ and is strictly decreasing on $(0,\infty)$; for $\eta<\eta_0$, $\lim_{z\to\infty} f_0(z,\eta)=-\infty$.*

PROPOSITION EC.1. *For $0<\eta<\eta_0$, $f_0(z,\eta)$ is concave on $[0,\infty]$.*

LEMMA EC.8. *For $c_o\kappa\sqrt{\lambda/\mu}-\bar{\beta}r_*<\eta<\eta_1$, $f_1(z,\eta)$ is strictly decreasing on $(-\infty,n_\eta)$ and strictly increasing on (n_η,∞) for some $n_\eta<\kappa\sqrt{\lambda/\mu}$; for $\eta=c_o\kappa\sqrt{\lambda/\mu}-\bar{\beta}r_*$, $f_1(z,\eta)$ is strictly decreasing on $(-\infty,\kappa\sqrt{\lambda/\mu})$ and is a constant function r_* on $[\kappa\sqrt{\lambda/\mu},\infty)$; for $\eta<\eta_1$, $\lim_{z\to-\infty} f_1(z,\eta)=\infty$.*

PROPOSITION EC.2. *For $c_o\kappa\sqrt{\lambda/\mu}-\bar{\beta}r_*<\eta<\eta_1$, $f_1(z,\eta)$ is convex on $(-\infty,\kappa\sqrt{\lambda/\mu}]$.*

We note that the stated concavity/convexity properties in Propositions EC.1–EC.2 are crucial to the proof of Proposition 1 part (ii), whose proof rely on Lemmas EC.7–EC.8.

Continuing our proof of Proposition 1, suppose $\eta_0 \le c_o\kappa\sqrt{\lambda/\mu}-\bar{\beta}r_*$. From Lemmas EC.7–EC.8 it follows that $f_0(\cdot,\eta)<r_*<f_1(\cdot,\eta)$ on $[-\infty,\infty]$ for $\eta<\bar{\eta}=\min\{\eta_0,\eta_1\}$. Therefore, $f_0(\cdot,\eta)$ and $f_1(\cdot,\eta)$ do not intersect for $\eta<\bar{\eta}$. Thus, without loss of generality, in the remainder of the proof we shall assume $\eta_0 > c_o\kappa\sqrt{\lambda/\mu}-\bar{\beta}r_*$.

We observe that the two functions $f_0(\cdot,\bar{\eta})$ and $f_1(\cdot,\bar{\eta})$ must coincide at $z=-\infty$ or $z=\infty$. If letting $\hat{\eta} := \max\left\{c_o\kappa\sqrt{\lambda/\mu}-\bar{\beta}r_*, 0\right\}$, by Lemmas EC.7–EC.8, $f_1(z,\hat{\eta})-f_0(z,\hat{\eta})>0$ for any $z\in[-\infty,\infty]$. Also, from Lemma 1 it follows that for any z, $f_0(z,\eta)$ is increasing in η and $f_1(z,\eta)$ is decreasing in η. Therefore, there exists some $\underline{\eta}\in(\hat{\eta},\bar{\eta}]$ such that $f_0(\cdot,\eta)$ and $f_1(\cdot,\eta)$ do not intersect for $\eta<\underline{\eta}$, touch for $\eta=\underline{\eta}$, and cross at least twice for $\eta\in(\underline{\eta},\bar{\eta})$. This completes the proof of part (i)

To establish part (ii), we will write f_0 and f_1 in place of $f_0(\cdot,\eta)$ and $f_1(\cdot,\eta)$, respectively, and let $\tilde{f} := f_0-f_1$. We intend to argue that f_0 and f_1 will cross at most twice for any $\eta\in(\underline{\eta},\bar{\eta})$. Suppose, by way of contradiction, that the two functions cross more than twice. Because $f_0(-\infty)<f_1(-\infty)$ and $f_0(\infty)<f_1(\infty)$ due to Lemmas EC.7–EC.8, there must exist four zeros of $\tilde{f}$, $\hat{z}_1<\hat{z}_2<\hat{z}_3<\hat{z}_4$, such that $\tilde{f}$ crosses the horizontal line $y=0$ from below at $\hat{z}_1$ and $\hat{z}_3$ and from above at $\hat{z}_2$ and $\hat{z}_4$. On the other hand, from (19) and (20) it follows that

$$\lambda\tilde{f}'(z)=\begin{cases} c_o\kappa\sqrt{\lambda/\mu} & \text{for all } z \text{ such that } z\le 0 \text{ and } \tilde{f}(z)=0,\\ \lambda f_0'(z)-\lambda f_1'(z) & \text{for } z\in(0,\kappa\sqrt{\lambda/\mu}),\\ c_o\kappa\sqrt{\lambda/\mu}-\kappa\sqrt{\lambda/\mu}\varphi(f_1(z)) & \text{for all } z \text{ such that } z\ge\kappa\sqrt{\lambda/\mu} \text{ and } \tilde{f}(z)=0,\end{cases} \tag{EC.1}$$

where $\varphi(x) := \min_{\boldsymbol{q}\in\mathcal{A}}\{\sum_i[(\mu-\theta_i)x+r_i\theta_i]\,q_i\}$. It is easy to see that $-\varphi$ is convex. Also, since f_1 is monotonic on $[\kappa\sqrt{\lambda/\mu},\infty)$ by Lemma EC.8, $-\varphi\circ f_1$ is quasi-convex on $[\kappa\sqrt{\lambda/\mu},\infty)$. Furthermore, for each fixed $\eta\in(\underline{\eta},\bar{\eta})$, f_0 is concave and f_1 is convex on $(0,\kappa\sqrt{\lambda/\mu})$. Therefore, $\tilde{f}'$ is decreasing on

$(0, \kappa\sqrt{\lambda/\mu})$. These, along with (EC.1), imply that for each $\eta \in (\underline{\eta}, \bar{\eta})$, the derivative function $\tilde{f}'$ can change signs at most twice when restricted to the smaller domain comprising all its zeros. This means that if $\tilde{f}$ crosses the horizontal line $y = 0$ at some point that is greater than $\hat{z}_3$, it must have done so from below, which leads to a contradiction. Therefore, f_0 and f_1 can cross at most twice for any $\eta \in (\underline{\eta}, \bar{\eta})$. This completes the proof of part (ii). □

Proof of Theorem 1. First, let us prove part (i). For any $\eta \in (\underline{\eta}, \bar{\eta})$, by Proposition 1, the functions f_0 and f_1 cross at two points. Let $z_0(\eta) < z_1(\eta)$ denote two points where the functions f_0 and f_1 cross. When $\eta \to \underline{\eta}$, $z_1(\eta) - z_0(\eta) \to 0$ and $\int_{z_0(\eta)}^{z_1(\eta)} [f_0(z,\eta) - f_1(z,\eta)]\mathrm{d}z \to 0$. When $\eta = \bar{\eta}$, $z_0(\eta) = \bar{z}_0$ and $z_1(\eta) = \bar{z}_1$. Since for any z, $f_1(z,\eta)$ is decreasing in η and $f_0(z,\eta)$ is increasing in η, $z_0(\eta)$ is decreasing in η and $z_1(\eta)$ is increasing in η. As a result, $\int_{z_0(\eta)}^{z_1(\eta)} [f_0(z,\eta) - f_1(z,\eta)]\mathrm{d}z$ is increasing in η. Hence, as η increases from $\underline{\eta}$ to $\bar{\eta}$, $\int_{z_0(\eta)}^{z_1(\eta)} [f_0(z,\eta) - f_1(z,\eta)]\mathrm{d}z$ increases from 0 to $\bar{C} = \int_{\bar{z}_0}^{\bar{z}_1} [f_0(z,\bar{\eta}) - f_1(z,\bar{\eta})]\mathrm{d}z$, which is greater than C by the assumption. Hence, we conclude that there exists some $\eta^\star \in (\underline{\eta}, \bar{\eta})$ such that $\int_{z_0(\eta^\star)}^{z_1(\eta^\star)} [f_0(z,\eta^\star) - f_1(z,\eta^\star)]\mathrm{d}z = C$. Denoting $z_0^\star = z_0(\eta^\star)$ and $z_1^\star = z_1(\eta^\star)$, we complete the proof of part (i).

To establish part (ii), let $v(0,\cdot)$ and $v(1,\cdot)$ be such that (a) $v(0, z_0^\star) = v(1, z_0^\star)$, (b) $v_z(0,z) = f_0(z,\eta^\star)1_{\{z<z_1^\star\}} + f_1(z,\eta^\star)1_{\{z\geq z_1^\star\}}$, and (c) $v_z(1,z) = f_0(z,\eta^\star)1_{\{z\leq z_0^\star\}} + f_1(z,\eta^\star)1_{\{z>z_0^\star\}}$. By our construction, $v(y,z)$ and the constant $\eta^\star$ collective solve the Bellman equation (13). Using the Itô–Tanaka formula, we can see that

$$\begin{aligned} v(Y(t),Z(t)) = v(y,z) &+ \int_0^t \left(\lambda v_{zz}(Y(u),Z(u)) + b(Y(u),Z(u),G(u))v_z(Y(u),Z(u))\right)\mathrm{d}u \\ &+ \sqrt{2\lambda}\int_0^t v_z(Y(u),Z(u))\mathrm{d}B(u) + \sum_{u\leq t}\left(v(Y(u+),Z(u)) - v(Y(u),Z(u))\right) \end{aligned}$$

This implies

$$\begin{aligned} &\sum_{i=1}^I r_i\theta_i \int_0^t \left[Z(u) - \kappa\sqrt{\lambda/\mu}Y(u)\right]^+ g_i(u)\mathrm{d}u + c_o\kappa\sqrt{\lambda/\mu}\int_0^t Y(u)\mathrm{d}u + C\sum_{u\leq t}[\Delta Y(u)]^+ \\ &\quad = v(y,z) - v(Y(t),Z(t)) + \int_0^t \Big(\lambda v_{zz}(Y(u),Z(u)) + b(Y(u),Z(u),G(u))v_z(Y(u),Z(u)) \\ &\qquad + \sum_{i=1}^I r_i\theta_i\left[Z(u) - \kappa\sqrt{\lambda/\mu}Y(u)\right]^+ g_i(u) + c_o\kappa\sqrt{\lambda/\mu}Y(u)\Big)\mathrm{d}u \\ &\qquad + \mathcal{M}(t) + \sum_{u\leq t}\left(v(Y(u+),Z(u)) - v(Y(u),Z(u)) + C[\Delta Y(u)]^+\right), \end{aligned}$$

where $\mathcal{M}(t)$ is a stochastic integral defined as

$$\mathcal{M}(t) = \sqrt{2\lambda}\int_0^t v_z(Y(u),Z(u))\mathrm{d}B(u).$$

Because v satisfies (13), we obtain

$$v(y,z)-v(Y(t),Z(t))+\eta^\star t \le \sum_{i=1}^{I} r_i\theta_i \int_0^t \left[Z(u)-\kappa\sqrt{\lambda/\mu}Y(u)\right]^+ g_i(u)\mathrm{d}u + c_o\kappa\sqrt{\lambda/\mu}\int_0^t Y(u)\mathrm{d}u + C\sum_{u\le t}[\Delta Y(u)]^+ - \mathcal{M}(t).$$

Taking expectations, dividing both sides by t, and noting that $\mathcal{M}(t)$ has expectation 0, we get

$$\eta^\star \le \frac{1}{t}\mathbb{E}\left[\sum_{i=1}^{I} r_i\theta_i \int_0^t \left[Z(u)-\kappa\sqrt{\lambda/\mu}Y(u)\right]^+ g_i(u)\mathrm{d}u + c_o\kappa\sqrt{\lambda/\mu}\int_0^t Y(u)\mathrm{d}u + C\sum_{u\le t}[\Delta Y(u)]^+\right] + \frac{1}{t}\mathbb{E}\left[v(Y(t),Z(t))\right]. \tag{EC.2}$$

To prove that the last term on the right-hand side of (EC.2) vanishes as $t\to\infty$, we need the following lemma, whose proof is deferred to EC.2.

LEMMA EC.9. *Regardless of the choice of* (Y,G)*, we have* $\limsup_{t\to\infty} t^{-1}\mathbb{E}\left[Z(t)\right]=0$.

Now, by sending $t\to\infty$ in (EC.2) and using Lemma EC.9 plus the Lipschitz continuity of v in z, we conclude

$$\eta^\star \le \frac{1}{t}\mathbb{E}\left[\sum_{i=1}^{I} r_i\theta_i \int_0^t \left[Z(u)-\kappa\sqrt{\lambda/\mu}Y(u)\right]^+ g_i(u)\mathrm{d}u + c_o\kappa\sqrt{\lambda/\mu}\int_0^t Y(u)\mathrm{d}u + C\sum_{u\le t}[\Delta Y(u)]^+\right].$$

If (Y,G) is replaced by $(Y^\star,G^\star)$, all the preceding inequalities hold as an equality. By following the same arguments as before, we can see that $\eta^\star$ is the long-run average cost under the joint control rule $(Y^\star,G^\star)$. Therefore, the proof is complete. □

Proof of Proposition 2. We will only prove the case where $\bar\eta=\eta_0\le\eta_1$. The proof for the case where $\bar\eta=\eta_1\le\eta_0$ is similar.

Suppose $\bar\eta=\eta_0\le\eta_1$. Let $v(0,\cdot)$ and $v(1,\cdot)$ be defined as follows: (a) $v(0,\bar z_0)=v(1,\bar z_0)$, (b) $v_z(0,z)=f_0(z,\bar\eta)$, and (c) $v_z(1,z)=f_0(z,\bar\eta)1_{\{z\le\bar z_0\}}+f_1(z,\bar\eta)1_{\{z>\bar z_0\}}$. From the proof of part (ii) of Theorem 1, we know that it suffices to show that $v(0,\cdot)$ and $v(1,\cdot)$ constructed in this way satisfy (13) with $\eta^\star$ therein being $\bar\eta$. With respect to $v(0,z)$, we have by construction:

$$\lambda v_{zz}(0,z)+\min_{\boldsymbol{q}\in\mathcal{A}}\left\{b(0,z,\boldsymbol{q})v_z(0,z)+[z]^+\sum_i r_i\theta_i q_i\right\}=\eta_0=\bar\eta \quad \text{for all} \quad z\in\mathbb{R}.$$

With respect to $v(1,\cdot)$, we have by construction:

$$\lambda v_{zz}(1,z)+\min_{\boldsymbol{q}\in\mathcal{A}}\left\{b(1,z,\boldsymbol{q})v_z(1,z)+\left[z-\kappa p\sqrt{\lambda/\mu}\right]^+\sum_i r_i\theta_i q_i\right\}+c_o\kappa p\sqrt{\lambda/\mu}=\eta_0=\bar\eta \quad \text{for} \quad z\le\bar z_0,$$

and

$$\lambda v_{zz}(1,z)+\min_{\boldsymbol{q}\in\mathcal{A}}\left\{b(1,z,\boldsymbol{q})v_z(1,z)+\left[z-\kappa p\sqrt{\lambda/\mu}\right]^+\sum_i r_i\theta_i q_i\right\}+c_o\kappa p\sqrt{\lambda/\mu}=\eta_1\ge\bar\eta \quad \text{for} \quad z>\bar z_0.$$

In addition, by our construction, we have $v(0,z)=v(1,z)$ for all $z\le\bar{z}_0$,

$$v(0,z)-v(1,z)=\int_{\bar{z}_0}^{z}[f_0(u,\bar{\eta})-f_1(u,\bar{\eta})\mathrm{d}u]\le\int_{\bar{z}_0}^{\bar{z}_1}[f_0(u,\bar{\eta})-f_1(u,\bar{\eta})\mathrm{d}u]=\bar{C}$$

for all $\bar{z}_0<z\le\bar{z}_1=\infty$. Therefore,

$$-C<-\bar{C}\le v(1,z)-v(0,z)\le 0\quad\text{for all}\quad z\in\mathbb{R}.$$

Taken together, we can conclude that $v(0,\cdot)$ and $v(1,\cdot)$ constructed as above satisfy (13) with $\eta^\star$ therein being $\bar{\eta}$. □

EC.2. Proofs of Auxiliary Results

Proof of Lemma EC.1. Towards proving part (i), we note that (19) admits an explicit expression for $z\le 0$, and it is given by

$$f_0(z,\eta)=\frac{\eta}{\lambda}\sqrt{\frac{\lambda}{\mu}}\frac{\Phi\left(\sqrt{\frac{\mu}{\lambda}}\left(z+\beta\sqrt{\frac{\lambda}{\mu}}\right)\right)}{\Phi'\left(\sqrt{\frac{\mu}{\lambda}}\left(z+\beta\sqrt{\frac{\lambda}{\mu}}\right)\right)},\tag{EC.3}$$

where Φ denotes the cumulative distribution function of a standard normal random variable. It is plain to see f_0 is strictly increasing in η for each $z\le 0$. In particular, $f_0(0,\eta)<f_0(0,\eta')$ for $\eta<\eta'$. Then existence of solution for $f_0(\cdot,\eta)$ follows from the general theory of ordinary differential equations, and $f_0(z,\eta)<f_0(z,\eta')$ for all $z>0$ thanks to the comparison theorem for ordinary differential equations. □

Proof of Lemma EC.2. The existence of η_0 satisfying conditions stated in the proposition has been established in Theorem 1 of Kim et al. (2018). Our equation is a special case of theirs. □

Proof of Lemma EC.3. The existence proof for η_1 can be accomplished by mimicking the proof of Theorem 1 in Kim et al. (2018). More precisely, we can construct a solution to (24) subject to the two boundary conditions by pasting two functions together at $z=\kappa\sqrt{\lambda/\mu}$ instead of at the origin; a careful examination of the analysis in Kim et al. (2018) indicates that the proof does not rely on the specific value of the pasting point. However, we must have $\eta_1>c_o\kappa\sqrt{\lambda/\mu}$ since $c_o\kappa\sqrt{\lambda/\mu}$ is the inherent running cost for the system to operate in "on" mode, regardless of any abandonment penalty. □

Proof of Lemmas EC.4 - EC.5. The proofs to these two lemmas are essentially contained in the the proof of Proposition 6.1 in Weerasinghe (2015) and therefore omitted. □

Proof of Lemma EC.6. To show that the mapping $\eta(\cdot)$ is continuous, consider an increasing sequence $\{\alpha_n\}$ with α being the limit. For ease of notation, we write $W_{\alpha_n}(\cdot,\eta(\alpha_n))$ as $W_n(\cdot)$ and $\eta(\alpha_n)$ as η_n. We aim to show that $\eta_n\to\eta(\alpha)$ as $n\to\infty$. By Lemma EC.5, $\{W_n(x)\}$ is an increasing

sequence satisfying $W_n(z) < W_\alpha(z)$ for each fixed z. Hence $W_\infty(z) := \lim_{n\to\infty} W_n(z)$ is well defined for each fixed z. From (20), note that $(W_n(\cdot), \eta_n)$ satisfies

$$\lambda W_n'(z) - \bar{\beta} W_n(z) + \phi(z, W_\alpha(z)) + c_o\kappa\sqrt{\lambda/\mu} - \eta_n = 0 \tag{EC.4}$$

subject to boundary conditions $W_n(\kappa\sqrt{\lambda/\mu}) = \alpha_n$ and $\lim_{z\to\infty} W_n(z) = r_*$, where $\bar{\beta} := \sqrt{\lambda\mu}(\beta+\kappa)$ and $\phi(z,w) := (z - \kappa\sqrt{\lambda/\mu}) \inf_{\boldsymbol{q}\in\mathcal{A}} \{\sum_i q_i\theta_i(r_i - w)\}$. Let $a := \kappa\sqrt{\lambda/\mu}$. Integrating (EC.4) over $[a, z]$ yields

$$\lambda(W_n(z) - \alpha_n) = \int_a^z \left(\bar{\beta} W_n(s) - \phi(s, W_n(s)) - c_o\kappa\sqrt{\lambda/\mu} + \eta_n\right) \mathrm{d}s. \tag{EC.5}$$

Sending $n \to$ in the above equation yields

$$\lambda(W_\infty(z) - \alpha) = \int_a^z \left(\bar{\beta} W_\infty(s) - \phi(s, W_\infty(s)) - c_o\kappa\sqrt{\lambda/\mu} + \eta_\infty\right) \mathrm{d}s.$$

On the other hand, by the definition of W_α and η_α we know that

$$\lambda(W_\alpha(z) - \alpha) = \int_a^z \left(\bar{\beta} W_\alpha(s) - \phi(s, W_\alpha(s)) - c_o\kappa\sqrt{\lambda/\mu} + \eta(\alpha)\right) \mathrm{d}s.$$

By the uniqueness result from Lemma EC.4, it follows that (W_∞, η_∞) must coincide with (W_α, η_α). The case where $\{\alpha_n\}$ decreases and converges to α can be analyzed in exactly the same way. Thus, we have shown that $\eta(\cdot)$ is continuous.

We next show that $\eta(\alpha)$ is decreasing in α. To this end, we consider $\alpha_1 > \alpha_2$ and assume for the sake of contradiction that $\hat{\eta}_1 \geq \hat{\eta}_2$ for $\hat{\eta}_1 := \eta(\alpha_1)$ and $\hat{\eta}_2 := \eta(\alpha_2)$. By Lemma EC.4, we know that there exist $W_1 := W_{\alpha_1}$ and $W_2 := W_{\alpha_2}$ such that

$$\lambda W_1'(z) - \bar{\beta} W_1(z) + \phi(z, W_1(z)) + c_o\kappa\sqrt{\lambda/\mu} - \hat{\eta}_1 = 0, \tag{EC.6}$$

$$\lambda W_2'(z) - \bar{\beta} W_2(z) + \phi(z, W_2(z)) + c_o\kappa\sqrt{\lambda/\mu} - \hat{\eta}_2 = 0, \tag{EC.7}$$

and $\lim_{z\to\infty} W_{\alpha_1}(z) = \lim_{z\to\infty} W_{\alpha_2}(z) = r_*$. Because

$$\int_a^\infty W_1'(z)\mathrm{d}z = r_* - \alpha_1 < r_* - \alpha_2 = \int_a^\infty W_2'(z)\mathrm{d}z,$$

there must exists $\bar{z} > a$ such that $W_1'(\bar{z}) < W_2'(\bar{z})$. Upon subtracting (EC.7) from (EC.6) and evaluating the resulting equation at $z = \bar{z}$, we obtain

$$\phi(\bar{z}, W_1(\bar{z})) - \phi(\bar{z}, W_2(\bar{z})) = -\lambda\left(W_1'(\bar{z}) - W_2'(\bar{z})\right) + \bar{\beta}\left(W_1(\bar{z}) - W_2(\bar{z})\right) + \hat{\eta}_1 - \hat{\eta}_2. \tag{EC.8}$$

Assume for the moment that $\bar{\beta} \geq 0$. By our hypothesis, the right-hand side of the above equation is positive. On the other hand, $\phi(z,w)$ is non-increasing in w for each $z > a$. Therefore, $\phi(\bar{z}, W_1(\bar{z})) - \phi(\bar{z}, W_2(\bar{z})) \leq 0$, due to the fact that $W_1(\bar{z}) > W_2(\bar{z})$. This however leads to a contraction. Therefore, we must have $\eta_1 < \eta_2$ when $\bar{\beta} \geq 0$. If $\bar{\beta} < 0$, then we can define $a' := a - \bar{\beta}/\theta_*$ for $\theta_* := \min_i \theta_i$. It is

straightforward to verify that $\tilde{\phi}(z,w)$ defined via $\tilde{\phi}(z,w) := -\bar{\beta}w + \phi(z,w)$ is non-increasing in w for each $z > a'$. Also, because

$$\int_{a'}^{\infty} W_1'(z)\mathrm{d}z = r_* - W_1(a') < r_* - W_2(a') = \int_{a'}^{\infty} W_2'(z)\mathrm{d}z,$$

there must exists $\tilde{z} > a'$ such that $W_1'(\tilde{z}) < W_2'(\tilde{z})$. From (EC.8), we see that

$$\tilde{\phi}(\tilde{z}, W_1(\tilde{z})) - \tilde{\phi}(\tilde{z}, W_2(\tilde{z})) = -\lambda\left(W_1'(\tilde{z}) - W_2'(\tilde{z})\right) + \hat{\eta}_1 - \hat{\eta}_2.$$

By our hypothesis, the right-hand side is strictly positive whereas the left-hand side is non-positive, which is again a contradiction. To summarize, we have proved that $\eta(\alpha_1) < \eta(\alpha_2)$ if $\alpha_1 > \alpha_2$.

To show that $\eta_\alpha \to -\infty$ as $\alpha \to \infty$, suppose by way of contradiction that $\eta(\alpha) \to \eta_* > -\infty$ when $\alpha \to \infty$. Using (EC.4) and the definition of ϕ, we find that

$$\lambda W_\alpha'(z) \geq \left((z-a)\theta_* + \bar{\beta}\right) W_\alpha(z) - (z-a)\nu - c_o a + \eta_* \quad \text{for} \quad z \geq a,$$

where $\nu := \max_i \theta_i r_i$. By applying the comparison principle for ordinary differential equations, we get that

$$W_\alpha(z) \geq e^{\frac{1}{\lambda}\left(\theta_*\left(\frac{z^2}{2} - az\right) + \bar{\beta}z\right)}\left[\alpha e^{\frac{1}{\lambda}\left(\theta_*\frac{a^2}{2} - \bar{\beta}a\right)} + \frac{1}{\lambda}\int_a^z e^{-\frac{1}{\lambda}\left(\theta_*\left(\frac{u^2}{2} - au\right) + \bar{\beta}u\right)}\left(\eta_* - (u-a)\nu - c_o a\right)\mathrm{d}u\right]$$

for $z \geq a$. If follows that $\lim_{z\to\infty} W_\alpha(z) = \infty$ for all large enough α. This, however, contradicts our requirement that $\lim_{z\to\infty} W_\alpha(z) = r_*$. Therefore, we must have $\eta(\alpha) \to -\infty$ as $\alpha \to \infty$. That $\eta(\alpha) \to \infty$ as $\alpha \to -\infty$ can be proved in an analogous fashion. We leave it as an exercise. □

Proof of Lemma EC.7. We first show that for each $\eta < \eta_0$, if $f_0'(z,\eta) = 0$ for some $z \geq 0$, then z is a local maximum. When there is no confusion, we suppress the dependence of f_0 on η. Note that for $z \geq 0$, (19) is equivalent to

$$\lambda f_0'(z) - \beta\sqrt{\lambda\mu} f_0(z) + z\hat{\phi}(f_0(z)) = \eta, \tag{EC.9}$$

where $\hat{\phi}(w) := \inf_{\boldsymbol{q}\in\mathcal{A}}\left\{\sum_i q_i\theta_i\left(r_i - w\right)\right\}$ is a piecewise linear function that is differentiable at all but finite break points. Let $\mathcal{I}^\star(z) := \arg\min_i\left\{\theta_i(r_i - f_0(z))\right\}$. In the case that $\mathcal{I}^\star(z)$ only contains one element, say, $i^\star(z)$, $w = f_0(z)$ is not one of the break points; directly differentiating (EC.9) yields

$$\lambda f_0''(z) - \left(\beta\sqrt{\lambda\mu} + z\theta_{i^\star(z)}\right) f_0'(z) + \theta_{i^\star(z)}\left(r_{i^\star(z)} - f_0(z)\right) = 0.$$

When $f_0'(z) = 0$, we have $f_0''(z) = \theta_{i^\star(z)}\left(f_0(z) - r_{i^\star(z)}\right) < 0$, since $f_0(z,\eta) < r_*$ for each $\eta < \eta_0$ by Lemma 1, and thus z is a local maximum. In the case that $\mathcal{I}^\star(z)$ contains more than one elements, we can

always identify two elements therein, say, $i_1^\star(z)$ and $i_2^\star(z)$, so that $\hat{\phi}(f_0(z)) = \theta_{i_1^\star(z)}\left(r_{i_1^\star(z)} - f_0(z)\right) = \theta_{i_2^\star(z)}\left(r_{i_2^\star(z)} - f_0(z)\right)$, and that

$$\hat{\phi}(f_0(z-\epsilon)) = \theta_{i_1^\star(z)}\left(r_{i_1^\star(z)} - f_0(z-\epsilon)\right)$$
$$\hat{\phi}(f_0(z+\epsilon)) = \theta_{i_2^\star(z)}\left(r_{i_2^\star(z)} - f_0(z+\epsilon)\right)$$

for all sufficiently small $\epsilon > 0$; note that $i_1^\star(z)$ and $i_2^\star(z)$ may coincide if $f_0(\cdot)$ is non-monotonic at z. With the above, one can evaluate (EC.9) at $z-\epsilon$ and $z+\epsilon$ and subtract from one another before dividing both sides by 2ϵ and letting $\epsilon \to 0^+$ to get

$$\lambda\left(f_0''\right)^+(z) = \frac{\theta_{i_1^\star(z)}\left(f_0(z) - r_{i_1^\star(z)}\right)}{2} + \frac{\theta_{i_2^\star(z)}\left(f_0(z) - r_{i_2^\star(z)}\right)}{2} + \left(\beta\sqrt{\lambda\mu} + \left(\frac{\theta_{i_1^\star(z)}}{2} + \frac{\theta_{i_2^\star(z)}}{2}\right)z\right) f_0'(z).$$

Using a similar argument on $\epsilon < 0$, one can show that $\left(f_0''\right)^-(z) = \left(f_0''\right)^+(z)$. Therefore, we have shown that z is a local maximum when $f_0'(z) = 0$ for both cases. It follows that if $f_0'(\hat{z}) \le 0$ for some $\hat{z} \ge 0$, then $f_0'(z)$ must be strictly decreasing on $(\hat{z}, \infty)$. From the proof of Lemma EC.1 we know that f_0 admits an explicit expression (EC.3) for $z \le 0$, which allows us to find $\lambda f_0'(0) = \left(1 + \beta\frac{\Phi(\beta)}{\Phi'(\beta)}\right)\eta$. Since the function $g(\beta) := 1 + \beta\frac{\Phi(\beta)}{\Phi'(\beta)}$ is positive for all β, $f_0'(0)$ shares the same sign with η. Hence, $f_0(z,\eta)$ must be strictly decreasing on $(0,\infty)$ for all $\eta \le 0$. For $\eta > 0$, we observe that $f_0(z,\eta)$ can not be strictly increasing on $(0,\infty)$, since otherwise $f_0(z,\eta)$ would grow to infinity or converge from below to r_*, which contracts the assumption that $\eta < \eta_0$. Therefore, there must exist a maximum point $m_\eta > 0$ so that $f_0(z,\eta)$ is strictly increasing on $(0, m_\eta)$ and decreasing on (m_η, ∞). The other properties of $f_0(z,\eta)$ on $(-\infty, 0)$ come directly from the explicit expression (EC.3). That $\lim_{z\to\infty} f_0(z,\eta) = -\infty$ for $\eta < \eta_0$ can be proven by contradiction upon sending $z \to \infty$ in (EC.9), because we have $\lim_{z\to\infty} f_0'(z,\eta) = 0$ but $\lim_{z\to\infty} z\hat{\phi}(f_0(z)) = -\infty$. □

Proof of Proposition EC.1. In what follows we fix the value of $\eta \in (0, \eta_0)$ and suppress the dependence of f_0 on η when there is no confusion. We aim to show that $f_0(z)$ is concave on $(0,\infty)$. To begin, we first consider a single-class setting, under which (EC.9) simplifies to

$$\lambda f_0'(z) - \beta\sqrt{\lambda\mu} f_0(z) + z\theta(r - f_0(z)) = \eta. \tag{EC.10}$$

By defining $\tilde{f}_0(z) := r - f_0(z)$, it follows from (EC.10) that $\tilde{f}_0$ obeys the ODE

$$\lambda\tilde{f}_0'(z) - \beta\sqrt{\lambda\mu}\tilde{f}_0(z) - \theta z\tilde{f}_0(z) = \tilde{\eta}, \tag{EC.11}$$

where $\tilde{\eta} := -\eta - \beta\sqrt{\lambda\mu} r$. For a given boundary condition $f_0(0) := \alpha_\eta$, $\tilde{f}_0(z)$ admits the explicit expression

$$\tilde{f}_0(z) = \frac{(r-\alpha_\eta)\Phi'(\sqrt{\mu/\theta}\beta) + \tilde{\eta}/\lambda\int_0^z \Phi'(\sqrt{\theta/\lambda}y + \sqrt{\mu/\theta}\beta)dy}{\Phi'(\sqrt{\theta/\mu}z + \sqrt{\mu/\theta}\beta)} \tag{EC.12}$$

$$= \frac{\sqrt{\lambda\theta}(r-\alpha_\eta)\Phi'(\sqrt{\mu/\theta}\beta) - \tilde{\eta}\Phi(\sqrt{\mu/\theta}\beta) + \tilde{\eta}\Phi(\sqrt{\theta/\lambda}z + \sqrt{\mu/\theta}\beta)}{\sqrt{\lambda\theta}\Phi'(\sqrt{\theta/\lambda}z + \sqrt{\mu/\theta}\beta)} \tag{EC.13}$$

$$:= \frac{\delta_\eta + \tilde{\eta}\Phi(\sqrt{\theta/\lambda}z + \sqrt{\mu/\theta}\beta)}{\sqrt{\lambda\theta}\Phi'(\sqrt{\theta/\lambda}z + \sqrt{\mu/\theta}\beta)}, \tag{EC.14}$$

for $z \in [-\infty, \infty]$, where $\Phi(\cdot)$ is the normal CDF and $\delta_\eta := \sqrt{\lambda\theta}(r-\alpha_\eta)\Phi'(\sqrt{\mu/\theta}\beta) - \tilde{\eta}\Phi(\sqrt{\mu/\theta}\beta)$. Since $\eta \in (0, \eta_0)$, we have $f_0(z) < f_0(z, \eta_0) < r$ for all $z \geq 0$ and $f_0'(0) > 0$ by Lemma EC.7. In particular, since $\tilde{f}_0(\infty) > 0$, it follows from (EC.14) that (C1) $\delta_\eta + \tilde{\eta} > 0$. Suppose now we extend (EC.10) to $[-\infty, 0]$ subject to the boundary condition $f_0(0) = \alpha_\eta$. It can be shown that if $f_0'(z) = 0$ for some $z \leq 0$, then z is a local maximum. Since $f_0'(0) > 0$, it follows that $f_0(z)$ defined by (EC.10) is strictly increasing on $[-\infty, 0]$. Therefore, we have $f_0(-\infty) < \alpha_\eta < r$, or equivalently, $\tilde{f}_0(-\infty) > 0$, which implies (C2) $\delta_\eta > 0$ in view of (EC.14).

From (EC.11) and its differential form, we obtain

$$\lambda^2 \tilde{f}_0''(z) = \left((\beta\sqrt{\lambda\mu} + \theta z)^2 + \lambda\theta\right) \tilde{f}_0(z) + (\beta\sqrt{\lambda\mu} + \theta z)\tilde{\eta}. \tag{EC.15}$$

To show that f_0 is concave on $(0, \infty)$, it suffices to show that (EC.15) is nonnegative for all $z > 0$. Using the expression of $\tilde{f}_0(z)$ in (EC.14) and defining $t := \sqrt{\theta/\lambda}z + \sqrt{\mu/\theta}\beta$, one can verify that (EC.15) being nonnegative for all $z > 0$ is equivalent to

$$\underbrace{(t^2+1)\left(\delta_\eta + \tilde{\eta}\Phi(t)\right) + t\tilde{\eta}\Phi'(t)}_{\square} \geq 0 \quad \text{for all} \quad t > \sqrt{\mu/\theta}\beta. \tag{EC.16}$$

To establish (EC.16), we consider the case $\tilde{\eta} \geq 0$ and $\tilde{\eta} < 0$ separately. When $\tilde{\eta} < 0$, we use (C1) to see $\square > -\tilde{\eta}\left((t^2+1)(1-\Phi(t)) - t\Phi'(t)\right) > 0$ for all t, since the function $p(t) := (t^2+1)(1-\Phi(t)) - t\Phi'(t) > 0$ is positive for all t. When $\tilde{\eta} \geq 0$, we use (C2) to see $\square > \tilde{\eta}\left((t^2+1)\Phi(t) + t\Phi'(t)\right) \geq 0$, since the function $h(t) := (t^2+1)\Phi(t) + t\Phi'(t)$ is positive for all t.

With the above intuition, we next generalize the proof to accommodate a multi-class setting. Since the function $\hat{\phi}(w)$ is piecewise linear, for a given initial value $f_0(0)$ we can find $z^{(0)} := 0 < z^{(1)} < \cdots < z^{(\bar{J}-1)} < z^{(\bar{J})} := \infty$ that partitions the positive real line into $\bar{J}$ segments, and in the j th segment, i.e., $z \in \left(z^{(j-1)}, z^{(j)}\right)$, $\hat{\phi}(f_0(z))$ follows the same rule $\hat{\phi}(f_0(z)) = \theta^{(j)}(r^{(j)} - f_0(z))$ for $z \in \left(z^{(j-1)}, z^{(j)}\right)$. Intuitively, the "cheapest" class $i^\star(z)$ remains the same in each segment, which allows us to define $\theta^{(j)} := \theta_{i^\star(z)}$ and $r^{(j)} := r_{i^\star(z)}$ for $z \in \left(z^{(j-1)}, z^{(j)}\right)$. Let $\alpha_\eta^{(j)} := f_0(z^{(j-1)}), j = 1, \ldots, \bar{J}$. It follows that in the j th segment $f_0(z)$ evolves according to the ODE

$$\lambda f_0'(z) - \beta\sqrt{\lambda\mu} f_0(z) + z\theta^{(j)}(r^{(j)} - f_0(z)) = \eta \tag{EC.17}$$

subject to the boundary condition $f_0(z^{(j-1)}) = \alpha_\eta^{(j)}$. We intend to show that $f_0(z)$ is concave in each segment. To begin, let $\tilde{f}_0^{(j)}(z) := r^{(j)} - f_0(z)$ for $z \in (z^{(j-1)}, z^{(j)})$ and define $\tilde{\eta}^{(j)} := -\eta - \beta\sqrt{\lambda\mu}r^{(j)}$, $j = 1, \ldots, \bar{J}$. It follows from (EC.12)-(EC.14) that $\tilde{f}_0^{(j)}(z)$ admits the explicit expression

$$\tilde{f}_0^{(j)}(z) = \frac{\delta_\eta^{(j)} + \tilde{\eta}^{(j)}\Phi\left(\sqrt{\theta^{(j)}/\lambda}z + \sqrt{\mu/\theta^{(j)}}\beta\right)}{\sqrt{\lambda\theta^{(j)}}\Phi'\left(\sqrt{\theta^{(j)}/\lambda}z + \sqrt{\mu/\theta^{(j)}}\beta\right)} \tag{EC.18}$$

for $z \in (z^{(j-1)}, z^{(j)})$, where we have defined

$$\delta_\eta^{(j)} := \sqrt{\lambda\theta^{(j)}}\left(r^{(j)} - \alpha_\eta^{(j)}\right)\Phi'\left(\sqrt{\theta^{(j)}/\mu}z^{(j-1)} + \sqrt{\mu/\theta^{(j)}}\beta\right) - \tilde{\eta}^{(j)}\Phi\left(\sqrt{\theta^{(j)}/\mu}z^{(j-1)} + \sqrt{\mu/\theta^{(j)}}\beta\right).$$

We first show that (C1') $\delta^{(j)} + \tilde{\eta}^{(j)} > 0$. From the proof of the single-class case, (C1') establishes the concavity of $f_0(z)$ in segment j when $\tilde{\eta}^{(j)} < 0$. To establish (C1'), we extend the ODE (EC.17) to infinity while satisfying the boundary condition $f_0(z^{(j-1)}) = \alpha_\eta^{(j)}$; we denote its solution by $f_0^{(j,\infty)}(z, \eta)$. The solution family $\left\{f_0^{(j,\infty)}(z, \eta), \eta \in \mathbb{R}\right\}$ is closely related to the following admissible scheduling rule: it coincides with the optimal rule when Z is less that $z^{(j-1)}$, but when Z is above $z^{(j-1)}$, it becomes a static priority rule that keeps all the work content in class $i^\star(\hat{z})$, where $\hat{z} \in (z^{(j-1)}, z^{(j)})$. Using similar arguments, one can see that $f_0^{(j,\infty)}(z, \eta)$ must be increasing in η and there exists a unique $\eta_0^{(j)}$ so that $\lim_{z\to\infty} f_0^{(j,\infty)}(z, \eta_0^{(j)}) = r^{(j)}$, where $\eta_0^{(j)}$ is to be interpreted as the long-run average cost of this admissible rule. Since $\eta < \eta_0$ by our assumption and $\eta_0^{(j)} \geq \eta_0$ due to the optimality of η_0 when $Y \equiv 0$, we have $f_0^{(j,\infty)}(\infty, \eta) < f_0^{(j,\infty)}(\infty, \eta_0^{(j)}) = r^{(j)}$. Therefore, we have $\tilde{f}_0^{(j)}(\infty) > 0$, which implies (C1') in view of (EC.18). Hence, $f_0(z)$ is concave in segment j when $\tilde{\eta}^{(j)} < 0$. The case when $\tilde{\eta}^{(j)} \geq 0$ needs more careful treatment. Ideally, we want to prove the analog of (C2), i.e., the condition (C2') $\delta_\eta^{(j)} > 0$. This condition (C2') can be proven in a similar fashion as the single-class proof when $f_0'(z^{(j-1)}) \geq 0$: one extends the ODE (EC.17) to the negative infinity subject to $f_0(z^{(j-1)}) = \alpha_\eta^{(j)}$ and observes that the solution must be strictly increasing on $[-\infty, z^{(j-1)})$ since $f_0'(z^{(j-1)}) \geq 0$, and thus we have $\tilde{f}_0^{(j)}(-\infty) > 0$, which implies (C2') in view of (EC.18). Next, we consider the case $f_0'(z^{(j-1)}) < 0$. From Lemma EC.7, there exists $0 < m_\eta < \infty$ such that $f_0(z)$ is strictly increasing on $(0, m_\eta)$ and strictly decreasing on (m_η, ∞), and we have $f_0'(m_\eta) = 0$ and $f_0''(m_\eta) < 0$. It follows that $z^{(j-1)} > m_\eta$ since $f_0'(z^{(j-1)}) < 0$. We next show that $f_0''(z)$ can not be positive on (m_η, ∞), which covers the interval $(z^{(j-1)}, z^{(j)})$. To this end, we differentiate (EC.9) twice on all but finite non-differential points to see that

$$\lambda f_0'''(z) = \beta\sqrt{\lambda\mu}f_0''(z) - 2\hat{\phi}'(f_0(z))f_0'(z) - z\hat{\phi}'(f_0(z))f_0''(z),$$

where we used $\hat{\phi}''(x) = 0$; the case of non-differential points can be treated in a similar way by resorting to left and right limits as in the proof of Lemma EC.7. Since $\hat{\phi}'(x) < 0$ for all $x \in \mathbb{R}$ and $f_0'(z) < 0$ for all $z \in (m_\eta, \infty)$, it holds that $f_0'''(z) < 0$ whenever $f_0''(z) = 0$. Combined with $f_0''(m_\eta) < 0$, it follows that $f_0''(z) \leq 0$ for all $z \in (m_\eta, \infty)$, and thus $f_0(z)$ is concave. □

Proof of Lemma EC.8. The proof for Lemma EC.8 is analogous to the proof for Lemma EC.7. Note that when $z \le \kappa\sqrt{\lambda/\mu}$, (20) is equivalent to

$$\lambda f_1'(z) - \beta\sqrt{\lambda\mu} f_1(z) - \mu z f_1(z) = \eta - c_o\kappa\sqrt{\lambda/\mu}. \tag{EC.19}$$

By differentiating (EC.19) and utilizing the fact that $f_1(z,\eta) > f_1(z,\eta_1) > 0$ for each $\eta < \eta_1$, we can show that if $f_1'(z,\eta) = 0$ for some $z \le \kappa\sqrt{\lambda/\mu}$, then z is a local minimum. It follows that if $f_1'(\hat{z}) \le 0$ for some $\hat{z} \le \kappa\sqrt{\lambda/\mu}$, then $f_1(z)$ must be decreasing on $(-\infty, \hat{z})$. By Lemmas EC.4 and EC.6, we know that for each $\eta \le \eta(r_*) = c_o\kappa\sqrt{\lambda/\mu} - \bar{\beta} r_*$, $f_1(z,\eta)$ is either strictly decreasing or a constant function r_* on $(\kappa\sqrt{\lambda/\mu}, \infty)$. Hence, $f_1(z,\eta)$ must be strictly decreasing on $(-\infty, \kappa\sqrt{\lambda/\mu})$ for each $\eta \le c_o\kappa\sqrt{\lambda/\mu} - \bar{\beta} r_*$. For $\eta_1 > \eta > c_o\kappa\sqrt{\lambda/\mu} - \bar{\beta} r_*$, since $f_1'(c_o\kappa\sqrt{\lambda/\mu}, \eta) > 0$ by Lemma EC.4 and $f_1'(z,\eta)$ can not be strictly increasing on $(-\infty, \kappa\sqrt{\lambda/\mu})$ (contradicting the assumption $\eta < \eta_1$ if otherwise), there must exists a minimum point $n_\eta < \kappa\sqrt{\lambda/\mu}$ so that $f_1(z,\eta)$ is strictly decreasing on $(-\infty, n_\eta)$ and increasing on (n_η, ∞). □

Proof of Proposition EC.2. In what follows we fix the value of $\eta \in (c_o\kappa\sqrt{\lambda/\mu} - \bar{\beta} r_*, \eta_1)$ and suppress the argument η in f_1 when there is no confusion. We intend to prove the claim that $f_1(z)$ is convex on $(-\infty, \kappa\sqrt{\lambda/\mu})$ for $\eta \in (c_o\kappa\sqrt{\lambda/\mu} - \bar{\beta} r_*, \eta_1)$. To begin, we solve (EC.19) subject to the boundary condition $f_1(\kappa\sqrt{\lambda/\mu}) = \alpha_\eta$ to obtain the following explicit expression for $z \le \kappa\sqrt{\lambda/\mu}$

$$f_1(z) = \frac{\sqrt{\lambda\mu}\alpha_\eta\Phi'(\beta+\kappa) - \hat{\eta}\Phi(\beta+\kappa) + \hat{\eta}\Phi(\sqrt{\mu/\lambda}z + \beta)}{\sqrt{\lambda\mu}\Phi'(\sqrt{\mu/\lambda}z + \beta)}, \tag{EC.20}$$

where $\hat{\eta} := \eta - c\kappa\sqrt{\lambda\mu}$ and $\Phi(\cdot)$ is the normal CDF. By Lemmas 1 and EC.8, it must holds that $f_1(\kappa\sqrt{\lambda/\mu}) > f_1(\kappa\sqrt{\lambda/\mu}, \eta_1) > 0$ and $f_1'(\kappa\sqrt{\lambda/\mu}) > 0$, which translates to (C3) $\alpha_\eta > 0$ and (C4) $\hat{\eta} + \sqrt{\lambda\mu}(\beta+\kappa)\alpha_\eta > 0$. Moreover, because $f_1(z,\eta) > f_1(z,\eta_1) > 0$ for all $z \le \kappa\sqrt{\lambda/\mu}$ by Lemma 1, we must have (C5) $\sqrt{\lambda\mu}\alpha_\eta\Phi'(\beta+\kappa) - \hat{\eta}\Phi(\beta+\kappa) > 0$, since otherwise $f_1(-\infty, \eta) \le f_1(-\infty, \eta_1) = 0$.

From (EC.19) and its differential form we obtain

$$\lambda^2 f_1''(z) = \left((\beta\sqrt{\lambda\mu} + \mu z)^2 + \lambda\mu\right) f_1(z) + (\beta\sqrt{\lambda\mu} + \mu z)\hat{\eta}. \tag{EC.21}$$

The remaining task is then to show that under conditions (C3)(C4)(C5), (EC.21) is nonnegative for all $z \in (-\infty, \kappa\sqrt{\lambda/\mu})$. Using the expression of $f_1(z)$ in (EC.20) and defining $t := \sqrt{\mu/\lambda}z + \beta$, one can verify that (EC.21) being nonnegative for all $z \in (-\infty, \kappa\sqrt{\lambda/\mu})$ is equivalent to

$$\underbrace{(t^2+1)\left(\sqrt{\lambda\mu}\alpha_\eta\Phi'(\beta+\kappa) - \hat{\eta}\Phi(\beta+\kappa) + \hat{\eta}\Phi(t)\right) + t\hat{\eta}\Phi'(t)}_{\triangle} \ge 0 \quad \text{for all} \quad t \in (-\infty, \beta+\kappa). \tag{EC.22}$$

To establish (EC.22), we consider the case $\hat{\eta} \ge 0$ and $\hat{\eta} < 0$ separately. When $\hat{\eta} \ge 0$, we use condition (C5) to see that $\triangle > (t^2+1)\hat{\eta}\Phi(t) + t\hat{\eta}\Phi'(t) \ge 0$, since the function $h(t) := (t^2+1)\Phi(t) + t\Phi'(t)$ is

positive for all t. When $\hat{\eta}<0$, we further consider the case $\beta+\kappa\leq 0$ and $\beta+\kappa>0$ separately. When $\hat{\eta}<0$ and $\beta+\kappa\leq 0$, we use condition (C3) to see

$$\triangle > (t^2+1)\left(-\hat{\eta}\Phi(\beta+\kappa)+\hat{\eta}\Phi(t)\right)+t\hat{\eta}\Phi'(t).$$

Since $-\hat{\eta}>0$, it suffices to show that the function $k_1(t):=(t^2+1)\Phi(\beta+\kappa)-(t^2+1)\Phi(t)-t\Phi'(t)$ is nonnegative for all $t<\beta+\kappa$, which, follows from the fact that $k_1(t)$ is decreasing on $(-\infty,\beta+\kappa)$ and thus $k_1(t)>k_1(\beta+\kappa)=-(\beta+\kappa)\Phi'(\beta+\kappa)\geq 0$ for all $t<\beta+\kappa$. Finally, when $\hat{\eta}<0$ and $\beta+\kappa>0$, we use condition (C4) to see that $\sqrt{\lambda\mu}\alpha_\eta>-\hat{\eta}/(\beta+\kappa)$, so we have

$$\triangle > (t^2+1)\left(-\hat{\eta}\frac{\Phi'(\beta+\kappa)}{\beta+\kappa}-\hat{\eta}\Phi(\beta+\kappa)+\hat{\eta}\Phi(t)\right)+t\hat{\eta}\Phi'(t).$$

Since $-\hat{\eta}>0$, it suffices to show that the function $k_2(t):=(t^2+1)\left(\frac{\Phi'(\beta+\kappa)}{\beta+\kappa}+\Phi(\beta+\kappa)\right)-(t^2+1)\Phi(t)-t\Phi'(t)$ is non-negative for all $t<\beta+\kappa$, which, follows from the fact that $k_2(t)$ is decreasing on $(-\infty,\beta+\kappa)$ and thus $k_2(t)>k_2(\beta+\kappa)=\Phi'(\beta+\kappa)/(\beta+\kappa)>0$ for all $t<\beta+\kappa$. Since we have addressed all the possible scenarios, the proof is complete. □

Proof of Lemma EC.9. Our goal is to seek two processes $\underline{Z}$ and $\bar{Z}$ such that

$$\underline{Z}(t)\overset{s.t.}{\leq} Z(t)\overset{s.t.}{\leq}\bar{Z}(t). \tag{EC.23}$$

If both $\underline{Z}$ and $\bar{Z}$ have a stationary distribution with a finite mean, then we are done with the proof. To this end, let $\underline{Z}$ and $\bar{Z}$ be two piecewise linear diffusion processes defined as

$$\underline{Z}(t)=Z(0)-(\beta+\kappa)\sqrt{\lambda\mu}t+\int_0^t\left\{\mu\left[\underline{Z}(u)-\kappa\sqrt{\lambda/\mu}\right]^-\!\!-\theta^*\left[\underline{Z}(u)-\kappa\sqrt{\lambda/\mu}\right]^+\right\}\mathrm{d}u+\sqrt{2\lambda}B(t)$$

and

$$\bar{Z}(t)=Z(0)-\beta\sqrt{\lambda\mu}t+\int_0^t\left\{\mu\left[\bar{Z}(u)\right]^-\!\!-\theta_*\left[\bar{Z}(u)\right]^+\right\}\mathrm{d}u+\sqrt{2\lambda}B(t),$$

respectively. Then the stochastic relations in (EC.23) follow immediate by our construction. That $\underline{Z}$ and $\bar{Z}$ have a stationary distribution with a finite mean follows from Theorem 3 of Garnett et al. (2002). This concludes the proof. □

EC.3. Exact MDP Formulation for the Second-Stage Problem

For fixed values of N_0 and K, the second stage problem can be formulated as a continuous-time Markov Decision Process (CTMDP) with a dimension of $I+2$. It requires I dimensions to describe the total number of jobs in the system and queue contents, one dimension of binary process to describe the staffing mode of the system (i.e., 0 for "off" mode and 1 for "on" model), and one dimension to describe the number of on-demand servers in the system.

The second-stage problem can be expressed as a continuous-time Markov Decision Process (CTMDP) with $I+2$ dimensions, where N_0 and K are fixed values. I dimensions are allocated for the total number of jobs in the system and the queue contents, and one for a binary process to define the staffing mode (0 for "off" and 1 for "on"). Additionally, one dimension is allocated to indicate the number of on-demand servers in the system.

EC.3.1. System States, Decisions and Cost Structure

At time $t \geq 0$, the state of the queuing system is represented by a $(I+1)$-dimensional vector, denoted as $(X(t), \boldsymbol{Q}(t), Y(t), n(t))$. Here, $X(t)$ represents the total number of jobs in the system, $\boldsymbol{Q}(t) := (Q_1(t), Q_2(t), \ldots, Q_{I-1}(t))^\top$ represents the queue lengths of the first $I-1$ classes, $Y(t) \in 0, 1$ denotes the staffing mode indicator, and $n(t)$ tracks the number of on-demand servers at time t. When $I = 1$, the state descriptor $\boldsymbol{Q}(t)$ can be omitted, as discussed in Appendix EC.3.4.

Although the queue length of class-I at time t, denoted by $Q_I(t)$, is not considered as a system state, it can be obtained using the relation $Q_I(t) = (X(t) - N_0 - n)^+ - \sum_{i=1}^{I-1} Q_i(t)$, due to the non-idling scheduling policy. The state space, which we denote by $\mathcal{S}$, is given by

$$\mathcal{S} := \left\{ (X, \boldsymbol{Q}, Y, n) : 0 \leq X \leq M, \boldsymbol{Q} \geq \boldsymbol{0}, \sum_{i \in [I-1]} Q_i \leq (X - N_0 - n)^+, Y \in \{0,1\}, n \in \{0, 1, \ldots, K\} \right\}.$$

In the above, $\sum_{i \in [I-1]} Q_i = (X - N_0 - n)^+ - Q_I$ must be no greater than $(X - N_0 - n)^+$ since $Q_I \geq 0$. Also, to ensure that the state space remains finite in the numerical study, we stipulate that the total number of jobs in the system, denoted by X, is no greater than a constant value, M. This truncation does not sacrifice realism as long as M is chosen to be large enough because queue lengths are extremely unlikely to be very large in a queueing system with abandonment. Finally, n takes value in the set $\{0, 1, \ldots, K\}$ due to the random show-up behavior of on-demand servers.

The decision epoch coincides with each state transition. The scheduling decision is represented by a priority list $\boldsymbol{l} = (l_1, l_2, \ldots, l_I)$, a permutation of the class indices $1, 2, \ldots, I$, where class-l_1 (resp. class-l_I) has the highest (resp. lowest) priority. The on-demand staffing decision is captured by $y \in \{0, 1\}$, where $y = 0$ (resp. $y = 1$) maintains (resp. switches) the current staffing mode. The scheduling decision $\boldsymbol{l}$ affects the system dynamics in two ways. First, when a server becomes available due to service completion and there are jobs waiting from multiple classes, the newly available server is paired with the highest-priority class specified by $\boldsymbol{l}$. Second, when the system transits to the "on" mode, newly entering on-demand servers deplete the existing queue contents based on the class priority specified by $\boldsymbol{l}$. Specifically, suppose the current system state is $(X, \boldsymbol{Q}, Y, n)$. Define $j^* := \arg\min_j \{j \in [I] | Q_{l_j} > 0\}$ and $i^* := l_{j^*}$. Here, $i^* \in [I]$ is the highest-priority class among classes with positive queue lengths. When there is no confusion, we drop the dependence of i^* on $(\boldsymbol{Q}, \boldsymbol{l})$ for clarity. Hence, when a server becomes

available due to service completion, it will be paired with the job at the head of the class-i^* queue. If the set $\{j \in [I] | Q_{l_j} > 0\}$ is empty, we let $i^* = 0$ to indicate that there is no job waiting, and thus the newly available server remains idle. Next, when the system transits to the "on" mode and brings in n^a on-demand servers, these servers deplete the queue contents $\boldsymbol{Q}$ based on the scheduling decision $\boldsymbol{l}$. Let $T(\cdot|\boldsymbol{l}, n^a) : \mathbb{R}_+^{I-1} \mapsto \mathbb{R}_+^{I-1}$ be a function that maps the current queue contents to the queue contents after depletion by n^a on-demand servers under scheduling decision $\boldsymbol{l}$. Suppose $\boldsymbol{Q}' = T(\boldsymbol{Q}|\boldsymbol{l}, n^a)$. The mapping T is defined by setting $Q'_{l_1} = (Q_{l_1} - n^a)^+$ and $Q'_{l_j} = (Q_{l_j} - (n^a - \sum_{k=1}^{j-1} Q_{l_k})^+)^+$ for $j \geq 2$. Intuitively, the n^a on-demand servers first deplete the queue content of the highest-priority class, i.e., class-l_1. If there are on-demand servers still available after this depletion, i.e., $(n^a - Q_{l_1})^+ > 0$, they continue to deplete the second highest-priority class, i.e., class-l_2, and so on.

In terms of the cost structure, when the system is in state $(X, \boldsymbol{Q}, Y, n)$, the congestion cost is continuously incurred at a rate of $\sum_{i=1}^{I} Q_i \theta_i r_i$. Additionally, the cost of staffing on-demand servers is continuously incurred at a rate of nc_o. Finally, whenever the system transitions from the "off" staffing mode to the "on" staffing mode, a fixed cost of C is incurred.

EC.3.2. Uniformization: The Embeded Discrete-Time MDP

We utilize the uniformization technique to convert the above CTMDP to its embedded discrete-time MDP (DTMDP). Let $\alpha(X, \boldsymbol{Q}, Y, n)$ denote the transition rate when the system is in state $(X, \boldsymbol{Q}, Y, n)$. It follows that $\alpha(X, \boldsymbol{Q}, Y, n) = \lambda + \sum_{i \in [I]} \theta_i Q_i + \min\{X, N_0 + n\}\mu$. Letting $\bar{\theta} := \max_{i \in [I]}\{\theta_i\}$, we choose the uniformization constant $\bar{\alpha}$ as

$$\bar{\alpha} := \max_{(X, \boldsymbol{Q}, Y, n) \in \mathcal{S}} \alpha(X, \boldsymbol{Q}, Y, n) = \begin{cases} \lambda + M\bar{\theta} + N_0(\mu - \bar{\theta}), & \text{if} \quad \mu \leq \bar{\theta} \\ \lambda + M\bar{\theta} + (N_0 + K)(\mu - \bar{\theta}), & \text{if} \quad \mu > \bar{\theta} \end{cases}.$$

We next detail the state transitions in the embedded CTMDP. Throughout, let $\boldsymbol{e}_i$ denote a vector with the ith component equal to one and all others equal to zero. The dimension of $\boldsymbol{e}_i$ will be clear from the context. Suppose the current system state is $(X, \boldsymbol{Q}, Y, n)$.

Job Arrival. With probability $\lambda_i/\bar{\alpha}$, a class-i job arrives in the system for $i \in [I]$. Note that the quantity $(X - N_0 - n)^-$ tracks the number of idling servers.

- If the staffing decision is to remain in the current mode (i.e., $y = 0$), the system state becomes $(X + 1, \boldsymbol{Q} + \boldsymbol{e}_i \mathbb{1}\{(X - N_0 - n)^- = 0, i \neq I\}, Y, n)$;
- If the system is currently in the "off" mode (i.e., $Y = 0$) and the staffing decision is switching to the "on" mode (i.e., $y = 1$), each on-demand servers (of number $K - n$) in the pool accepts to join the system with probability p, and thus the probability that the number of on-demand servers change from n to n' after the staffing mode transition is given by

$$p_{n,n'} := \binom{K-n}{n'-n} p^{n'-n}(1-p)^{K-n'} \quad \text{for} \quad n' = n, n+1, \ldots, K.$$

Conditional on the event that the number of on-demand servers changes to n' and the scheduling decision is $\boldsymbol{l}$, the system state becomes $(X+1, T(\boldsymbol{Q}+\boldsymbol{e}_i\mathbb{1}\{(X-N_0-n)^-=0, i\neq I\}|\boldsymbol{l}, n'-n), \mathbb{1}\{n'>0\}, n')$;

- If the system is currently in the "on" mode (i.e., $Y=1$) and the staffing decision is switching to the "off" mode (i.e., $y=1$), permanent servers that are idle start to take over jobs from the hands of on-demand servers who are currently busy, and the number of on-demand servers who are still in the system after this takeover is $\min\{(X+1-N_0)^+, n\}$. Therefore, the system state becomes $(X+1, \boldsymbol{Q}+\boldsymbol{e}_i\mathbb{1}\{(X-N_0-n)^-=0, i\neq I\}, 0, \min\{(X+1-N_0)^+, n\})$.

Job Abandonment. With probability $\theta_i Q_i/\bar{\alpha}$, a class-$i$ job abandons the system for $i\in[I]$.

- If the staffing decision is to remain in the current mode (i.e., $y=0$), the system state becomes $(X-1, \boldsymbol{Q}-\boldsymbol{e}_i\mathbb{1}\{i\neq I\}, Y, n)$;
- If the system is currently in the "off" mode (i.e., $Y=0$) and the staffing decision is switching to the "on" mode (i.e., $y=1$), conditional on the event that the number of on-demand servers changes to n' (with probability $p_{n,n'}$) and the scheduling decision is $\boldsymbol{l}$, the system state becomes $(X-1, T(\boldsymbol{Q}-\boldsymbol{e}_i\mathbb{1}\{i\neq I\}|\boldsymbol{l}, n'-n), \mathbb{1}\{n'>0\}, n')$;
- If the system is currently in the "on" mode (i.e., $Y=1$) and the staffing decision is switching to the "off" mode (i.e., $y=1$), the system state becomes $(X-1, \boldsymbol{Q}-\boldsymbol{e}_i\mathbb{1}\{i\neq I\}, 0, \min\{(X-1-N_0)^+, n\})$.

Service Completion. With probability $\min\{X, N_0+n\}\mu/\bar{\alpha}$, a server finishes its on-hand service. If the system is in the "off" mode but there are still on-demand servers on duty (i.e., $Y=0$ but $n>0$), this service completion will cause a on-demand server to leave the system either by leaving directly or after a job handover. Otherwise, suppose the scheduling decision is $\boldsymbol{l}$. Recall that $i^*=i^*(\boldsymbol{Q}|\boldsymbol{l})$ is the highest-priority class among classes with positive queue lengths, and $i^*=0$ if all queues are empty.

- If the system is currently in the "off" mode (i.e., $Y=0$) and the staffing decision is to remain in the current mode (i.e., $y=0$), the system state becomes $(X-1, \boldsymbol{Q}-\boldsymbol{e}_{i^*}\mathbb{1}\{i^*\neq 0, i^*\neq I, n=0\}, 0, (n-1)^+)$;
- If the system is currently in the "off" mode (i.e., $Y=0$) and the staffing decision is switching to the "on" mode (i.e., $y=1$), the system state becomes $(X-1, T(\boldsymbol{Q}-\boldsymbol{e}_{i^*}\mathbb{1}\{i^*\neq 0, i^*\neq I, n=0\}|\boldsymbol{l}, n'-(n-1)^+), \mathbb{1}\{n'>0\}, n')$ with probability $p_{(n-1)^+,n'}$ for $n'=(n-1)^+, (n-1)^++1, \ldots, K$;
- If the system is currently in the "on" mode (i.e., $Y=1$) and the staffing decision is to remain in the current mode (i.e., $y=0$), the system state becomes $(X-1, \boldsymbol{Q}-\boldsymbol{e}_{i^*}\mathbb{1}\{i^*\neq 0, i^*\neq I\}, 1, n)$;
- If the system is currently in the "on" mode (i.e., $Y=1$) and the staffing decision is switching to the "off" mode (i.e., $y=1$), the system state becomes $(X-1, \boldsymbol{Q}-\boldsymbol{e}_{i^*}\mathbb{1}\{i^*\neq 0, i^*\neq I\}, 0, \min\{(X-1-N_0)^+, n\})$.

Self Transition. With probability $1-\alpha(X, \boldsymbol{Q}, Y, n)/\bar{\alpha}$, the state fictitiously transits to itself due to uniformization.

EC.3.3. Bellman Equation

Let $V:\mathcal{S}\mapsto\mathbb{R}$ and $g\in\mathbb{R}_+$ be the relative value function and long-run average cost, respectively, for the MDP formulation of the second-stage problem. Based on the DTMDP derived in the previous section and with reference to the Bellman's principle of optimality, we expect the solution pair (V,g) to satisfy the following equation:

$$
\begin{aligned}
&V(X,\boldsymbol{Q},0,n)+g/\bar{\alpha}\\
=&\sum_{i\in[I]}\lambda_i/\bar{\alpha}\times\min\Big\{V(X+1,\boldsymbol{Q}+\boldsymbol{e}_i\mathbb{1}\{(X-N_0-n)^-=0,i\neq I\},0,n),\\
&C+\sum_{n'=n}^{K}p_{n,n'}\min_{\boldsymbol{l}}\{V(X+1,T(\boldsymbol{Q}+\boldsymbol{e}_i\mathbb{1}\{(X-N_0-n)^-=0,i\neq I\}|\boldsymbol{l},n'-n),\mathbb{1}\{n'>0\},n')\}\Big\}\\
&+\sum_{i\in[I]}\theta_iQ_i/\bar{\alpha}\times\min\Big\{V(X-1,\boldsymbol{Q}-\boldsymbol{e}_i\mathbb{1}\{i\neq I\},0,n),\\
&\quad C+\sum_{n'=n}^{K}p_{n,n'}\min_{\boldsymbol{l}}\{V(X-1,T(\boldsymbol{Q}-\boldsymbol{e}_i\mathbb{1}\{i\neq I\}|\boldsymbol{l},n'-n),\mathbb{1}\{n'>0\},n')\}\Big\}\\
&+\min\{X,N_0+n\}\mu/\bar{\alpha}\times\min\Big\{\min_{\boldsymbol{l}}\{V(X-1,\boldsymbol{Q}-\boldsymbol{e}_{i^*}\mathbb{1}\{i^*\neq0,i^*\neq I,n=0\},0,(n-1)^+)\},\\
&\quad C+\sum_{n'=(n-1)^+}^{K}p_{(n-1)^+,n'}\min_{\boldsymbol{l}}\{V(X-1,T(\boldsymbol{Q}-\boldsymbol{e}_{i^*}\mathbb{1}\{i^*\neq0,i^*\neq I,n=0\}|\boldsymbol{l},n'-(n-1)^+),\\
&\qquad\qquad\mathbb{1}\{n'>0\},n')\}\Big\}\\
&+(1-\alpha(X,\boldsymbol{Q},0,n)/\bar{\alpha})\times V(X,\boldsymbol{Q},0,n)+\sum_{i\in[I]}Q_i\theta_ir_i/\bar{\alpha}+c_on/\bar{\alpha},
\end{aligned}
\tag{EC.24}
$$

and the equation

$$
\begin{aligned}
&V(X,\boldsymbol{Q},1,n)+g/\bar{\alpha}\\
=&\sum_{i\in[I]}\lambda_i/\bar{\alpha}\times\min\Big\{V(X+1,\boldsymbol{Q}+\boldsymbol{e}_i\mathbb{1}\{(X-N_0-n)^-=0,i\neq I\},1,n),\\
&\quad V(X+1,\boldsymbol{Q}+\boldsymbol{e}_i\mathbb{1}\{(X-N_0-n)^-=0,i\neq I\},0,\min\{(X+1-N_0)^+,n\})\Big\}\\
&+\sum_{i\in[I]}\theta_iQ_i/\bar{\alpha}\times\min\Big\{V(X-1,\boldsymbol{Q}-\boldsymbol{e}_i\mathbb{1}\{i\neq I\},1,n),\\
&\quad V(X-1,\boldsymbol{Q}-\boldsymbol{e}_i\mathbb{1}\{i\neq I\},0,\min\{(X-1-N_0)^+,n\})\Big\}\\
&+\min\{X,N_0+n\}\mu/\bar{\alpha}\times\min\Big\{\min_{\boldsymbol{l}}\{V(X-1,\boldsymbol{Q}-\boldsymbol{e}_{i^*}\mathbb{1}\{i^*\neq0,i^*\neq I\},1,n)\},\\
&\quad\min_{\boldsymbol{l}}\{V(X-1,\boldsymbol{Q}-\boldsymbol{e}_{i^*}\mathbb{1}\{i^*\neq0,i^*\neq I\},0,\min\{(X-1-N_0)^+,n\})\}\Big\}\\
&+(1-\alpha(X,\boldsymbol{Q},1,n)/\bar{\alpha})\times V(X,\boldsymbol{Q},1,n)+\sum_{i\in[I]}Q_i\theta_ir_i/\bar{\alpha}+c_on/\bar{\alpha},
\end{aligned}
\tag{EC.25}
$$

where recall that the highest-priority class index $i^*=i^*(\boldsymbol{Q}|\boldsymbol{l})$ depends on both the queue contents $\boldsymbol{Q}$ and the scheduling decision $\boldsymbol{l}$.

EC.3.4. Special Case for $I=1$

When $I=1$, the system state can be described by $(X(t),Y(t),n(t))$, where $0\le X(t)\le M$ denotes the number of jobs in the system, $Y(t)\in 0,1$ indicates the staffing mode, and $n(t)$ records the number of servers currently on duty in response to on-demand requests. The queue length $Q_1(t)$ at time t can be obtained as $Q_1(t)=(X-N_0-n)^+$. In a single-class system, there is only one feasible scheduling decision, which is to always prioritize class-1 jobs. Therefore, we have $l=(1)$ and $i^*(Q_1|l)=\mathbb{1}\{Q_1>0\}$. This observation greatly simplifies the equations (EC.24) and (EC.25), which are now given by

$$\begin{aligned}
&V(X,0,n)+g/\bar{\alpha}\\
=&\frac{\lambda_1}{\bar{\alpha}}\min\Big\{V(X+1,0,n),C+\sum_{n'=n}^{K}p_{n,n'}V(X+1,\mathbb{1}\{n'>0\},n')\Big\}\\
&+\frac{\theta_1 Q_1}{\bar{\alpha}}\min\Big\{V(X-1,0,n),C+\sum_{n'=n}^{K}p_{n,n'}V(X-1,\mathbb{1}\{n'>0\},n')\Big\}\\
&+\frac{\min\{X,N_0+n\}\mu}{\bar{\alpha}}\min\Big\{V(X-1,0,(n-1)^+),C+\sum_{n'=(n-1)^+}^{K}p_{(n-1)^+,n'}V(X-1,\mathbb{1}\{n'>0\},n')\Big\}\\
&+\frac{\bar{\alpha}-\alpha(X,0,n)}{\bar{\alpha}}V(X,0,n)+\frac{Q_1\theta_1 r_1+c_o n}{\bar{\alpha}},
\end{aligned}\tag{EC.26}$$

and

$$\begin{aligned}
V(X,1,n)+g/\bar{\alpha}=&\frac{\lambda_1}{\bar{\alpha}}\min\Big\{V(X+1,1,n),V(X+1,0,\min\{(X+1-N_0)^+,n\})\Big\}\\
&+\frac{\theta_1 Q_1+\min\{X,N_0+n\}\mu}{\bar{\alpha}}\min\Big\{V(X-1,1,n),V(X-1,0,\min\{(X-1-N_0)^+,n\})\Big\}\\
&+\frac{\bar{\alpha}-\alpha(X,1,n)}{\bar{\alpha}}V(X,1,n)+\frac{Q_1\theta_1 r_1+c_o n}{\bar{\alpha}}.
\end{aligned}\tag{EC.27}$$

When $I=1$, the above Bellman equations can be solved using standard value iteration algorithm to within $\epsilon-$optimality. Let g^ϵ and $\tilde{V}^\epsilon(X,Y,n)$ denote the algorithm return. Then, the $\epsilon-$optimal MDP cost is given by $\bar{\alpha}g^\epsilon$; the $\epsilon-$optimal staffing decision is to switch from "off" to "on" when $\tilde{V}^\epsilon(X,0,n)>C+\sum_{n'=n}^{K}p_{n,n'}\tilde{V}^\epsilon(X,\mathbb{1}\{n'>0\},n')$, and to switch from "on" to "off" when $\tilde{V}^\epsilon(X,1,n)>\tilde{V}^\epsilon(X+1,0,\min\{(X-N_0)^+,n\})$.

EC.4. Numerical Algorithms

In §EC.4.1 we present a numerical algorithm for computing $f_0(\cdot,\eta)$ and $f_1(\cdot,\eta)$ for any given $\eta\in\mathbb{R}$. This numerical algorithm will serve as the backbone in the algorithms for computing η_0,η_1 and $\bar{C}$, which we introduce in §EC.4.2. Finally, with η_0,η_1 and $\bar{C}$ as inputs, we introduce in §EC.4.3 a binary search algorithm in η for finding the optimal Bellman equation solution $(\eta^\star,z_0^\star,z_1^\star)$. As in §EC.1 and EC.2, we will use $p=1$ throughout this section to ease notation. To account for $p<1$, one can

simply replace all instances of κ below with κp, which does not affect the validity of the arguments or procedures used.

For the algorithmic parameters M, δ, Δ that appear below, we have set $M = 2\lceil \lambda/\mu \rceil$, $\delta = 0.001$ and $\Delta = 0.005$ for all problem instances in the present study.

EC.4.1. Finite Difference Method for Computing $f_0(\cdot,\eta)$ and $f_1(\cdot,\eta)$

For a given η, we next present a numerical algorithm for computing $f_0(\cdot,\eta)$ and $f_1(\cdot,\eta)$, which is based on the Finite Difference Method. In what follows, we fix the value of η and suppress the argument η in f_0 and f_1.

We consider a large enough interval $[-M, M]$ of the real line, over which $f_0(\cdot)$ and $f_1(\cdot)$ is solved. We discretize the interval $[-M, M]$ using $2N+1$ points $(z_n)_{n=-N}^{N}$ defined by $z_n = n\delta$, where $\delta = M/N$ is the discretization precision, so that the point z_{-N} (resp. z_N) corresponds to $-M$ (resp. M) in the real line.

Instead of solving $f_0(\cdot)$ and $f_1(\cdot)$ over the continuous interval, we solve the discretized approximation $(f_0(z_n))_{n=-N}^{N}$ and $(f_1(z_n))_{n=-N}^{N}$. For easy reference, we write $f_{0,n} = f_0(z_n)$ and $f_{1,n} = f_1(z_n)$. We approximate the first derivative of a function at point z_n by $f'(z_n) = \frac{f_n - f_{n-1}}{\delta}$. Hence, the discretized version of (19) and (20) subject to their respective boundary conditions become

$$\lambda \frac{f_{0,n} - f_{0,n-1}}{\delta} + \min_{\boldsymbol{q}\in\mathcal{A}} \left\{ b(0, n\delta, q) f_{0,n} + [n\delta]^+ \sum_i r_i \theta_i q_i \right\} = \eta, \ n \in [N], \tag{EC.28}$$

$$f_{0,-N} = \frac{\eta}{\sqrt{\lambda\mu}} \frac{\Phi(-\sqrt{\mu/\lambda} N\delta + \beta)}{\Phi'(-\sqrt{\mu/\lambda} N\delta + \beta)}, \tag{EC.29}$$

$$\lambda \frac{f_{1,n} - f_{1,n-1}}{\delta} + \min_{\boldsymbol{q}\in\mathcal{A}} \left\{ b(1, n\delta, q) f_{1,n} + \left[n\delta - \kappa\sqrt{\lambda/\mu} \right]^+ \sum_{i=1}^{I} r_i \theta_i q_i \right\} + c_o \kappa \sqrt{\lambda/\mu} = \eta, \ n \in [N]. \tag{EC.30}$$

$$f_{1,N} = r_*. \tag{EC.31}$$

where (EC.29) follows from the closed form (EC.3); (EC.31) serves as an approximation to the boundary condition $f_1(\infty) = r_*$, and we have defined $[N] = \{-N, -N+1, \ldots, N\}$.

Suppose we have knowledge of the optimal queue length distribution vector $q_0^\star(n\delta) = \left(q_{0,1}^\star(n\delta), \ldots, q_{0,I}^\star(n\delta) \right)$ that achieves the minimum in Equation (EC.28) for $n \in [N]$. We can use Equations (EC.28) and (EC.29) to compute $(f_0(z_n))_{n=-N}^{N}$ in a forward manner. Rearranging terms in Equation (EC.28), we obtain the iterative equation

$$f_{0,n} = \frac{\eta - [n\delta]^+ \sum_i r_i \theta_i q_{0,i}^\star(n\delta) + \frac{\lambda}{\delta} f_{0,n-1}}{\lambda/\delta + b(0, n\delta, q_0^\star(n\delta))}, \ n \in \{-N+1, \ldots, N\}. \tag{EC.32}$$

However, as $q_0^\star(n\delta)$ is usually unknown, we begin with an initial guess $q_0^{(0)}(n\delta) \in \mathcal{A}$ and replace $q_0^\star(n\delta)$ in Equation (EC.32) with $q_0^{(0)}(n\delta)$ to obtain the sequence $\left(f_0^{(0)}(z_n)\right)_{n=-N}^{N}$. We then update the queue length distribution vector as follows:

$$q_0^{(1)}(n\delta) = \arg\min_{\boldsymbol{q}\in\mathcal{A}} \left\{ b(0, n\delta, q) f_{0,n}^{(0)} + [n\delta]^+ \sum_i r_i \theta_i q_i \right\},\ n \in [N],$$

with which we can compute $\left(f_0^{(1)}(z_n)\right)_{n=-N}^{N}$ and update for $q_0^{(2)}(n\delta), n \in [N]$. This procedure repeats until the consecutive queue length distribution vectors stop changing, at which point we obtain $q_0^\star(n\delta)$ and the desired sequence $(f_0(z_n))_{n=-N}^{N}$. Similarly, we can use Equations (EC.30) and (EC.31) to solve for $(f_1(z_i))_{i=-N}^{N}$ in a backward manner. Rearranging Equation (EC.30), we get the iterative equation

$$f_{1,n-1} = \frac{\delta}{\lambda}\left[\left(\frac{\lambda}{\delta} + b(1, n\delta, q_1^\star(n\delta))\right) f_{1,n} + \left[n\delta - \kappa\sqrt{\lambda/\mu}\right]^+ \sum_i r_i\theta_i q_{1,i}^\star(n\delta) + c_o\kappa\sqrt{\lambda/\mu} - \eta\right]. \quad \text{(EC.33)}$$

This iterative equation holds for n in the range of $\{-N+1, \ldots, N\}$. Here, $q_1^\star(n\delta)$ is the optimal queue length distribution vector in "on" mode, which is defined as $q_1^\star(n\delta) = \left(q_{1,1}^\star(n\delta), \ldots, q_{1,I}^\star(n\delta)\right)$. We can find $q_1^\star(n\delta)$ by using a similar iterative updating approach as we did for $q_0^\star(n\delta)$.

EC.4.2. Algorithms for Computing η_0, η_1 and $\bar{C}$

Algorithm for computing η_0. We begin by introducing an algorithm to compute η_0. First, we select two interval endpoints $\eta_{\mathrm{L}} < \eta_{\mathrm{R}}$ such that $f_0(M, \eta_{\mathrm{L}}) < r_* - \Delta$ and $f_0(M, \eta_{\mathrm{R}}) > r_* + \Delta$.These function values can be computed using the algorithm described in §EC.4.1. In practice, we can set $\eta_{\mathrm{L}} = 0$ and increase η_{R} from $\sum_i r_i\theta_i/I$ by the power of two until the above condition is met. Due to the monotonicity result from Lemma 1, we know that $\eta_0 \in (\eta_{\mathrm{L}}, \eta_{\mathrm{R}})$. Next, we compute the midpoint $\eta_{\mathrm{m}} := (\eta_{\mathrm{L}} + \eta_{\mathrm{R}})/2$ and evaluate $f_0(M, \eta_{\mathrm{m}})$. If $f_0(M, \eta_{\mathrm{m}}) > r_* + \Delta$, then by the monotonicity result from Lemma 1, we know that $\eta_0 < \eta_{\mathrm{m}}$, and we shrink the interval by updating $\eta_{\mathrm{R}} \leftarrow \eta_{\mathrm{m}}$. If, instead, $f_0(M, \eta_{\mathrm{m}}) < r_* - \Delta$, then we know that $\eta_0 > \eta_{\mathrm{m}}$ by the monotonicity result, and we shrink the interval by updating $\eta_{\mathrm{L}} \leftarrow \eta_{\mathrm{m}}$. This process guarantees that the interval $(\eta_{\mathrm{L}}, \eta_{\mathrm{R}})$ contains η_0, and the interval is halved in each iteration. The iteration terminates either when $|f_0(M, \eta_{\mathrm{m}}) - r_*| \le \Delta$ or when $\eta_{\mathrm{R}} - \eta_{\mathrm{L}} < \Delta$. In the former case, we take η_{m} as the solution for η_0, while in the latter case, we take $(\eta_{\mathrm{L}} + \eta_{\mathrm{R}})/2$ as the solution.

Algorithm for computing η_1. The algorithm for computing η_1 follows a similar approach. We begin with two interval endpoints $\eta_{\mathrm{L}} < \eta_{\mathrm{R}}$ such that $f_1(M, \eta_{\mathrm{L}}) > \Delta$ and $f_1(M, \eta_{\mathrm{R}}) < -\Delta$. Using the monotonicity result from Lemma 1, we know that $\eta_1 \in (\eta_{\mathrm{L}}, \eta_{\mathrm{R}})$. Next, we calculate the function value at the midpoint of the interval, $\eta_{\mathrm{m}} := (\eta_{\mathrm{L}} + \eta_{\mathrm{R}})/2$, by computing $f_1(M, \eta_{\mathrm{m}})$. If $f_1(M, \eta_{\mathrm{m}}) > \Delta$, then we know that $\eta_1 > \eta_{\mathrm{m}}$ by the monotonicity result from Lemma 1, and we update the interval by setting $\eta_{\mathrm{L}} \leftarrow \eta_{\mathrm{m}}$. Conversely, if $f_1(M, \eta_{\mathrm{m}}) < -\Delta$, then we know that $\eta_1 < \eta_{\mathrm{m}}$ by the same monotonicity

result, and we update the interval by setting $\eta_{\mathrm{R}} \leftarrow \eta_{\mathrm{m}}$. Similarly, the iteration terminates either if $|f_1(M, \eta_{\mathrm{m}})| \le \Delta$ or if $\eta_{\mathrm{R}} - \eta_{\mathrm{L}} < \Delta$.

Algorithm for computing $\bar{C}$. To compute $\bar{C}$, we start by setting $\bar{\eta}$ as the minimum of η_0 and η_1. Next, we use the algorithm described in §EC.4.1 to obtain the entire discretized sequences $(f_0(z_n, \bar{\eta}))_{n=-N}^{N}$ and $(f_1(z_n, \bar{\eta}))_{n=-N}^{N}$. It is possible that the two functions $f_0(\cdot, \bar{\eta})$ and $f_1(\cdot, \bar{\eta})$ do not intersect, for instance, when $c_o \ge \bar{c}$. When this happens, we set $\bar{C} = 0$ because on-demand staffing is never worthwhile. Otherwise, we compute $\bar{C}$ by the numerical integration $\bar{C} = \sum_{n=-N}^{N} \delta[f_0(z_n, \bar{\eta}) - f_1(z_n, \bar{\eta})]^+$.

EC.4.3. Algorithm for Finding $(\eta^\star, z_0^\star, z_1^\star)$

Given the inputs η_0, η_1, and $\bar{C}$, we can find the optimal Bellman equation solution $(\eta^\star, z_0^\star, z_1^\star)$ for any $C \le \bar{C}$. Note that if $C > \bar{C}$, then it can be concluded that the solution $(\eta^\star, z_0^\star, z_1^\star)$ does not exist.

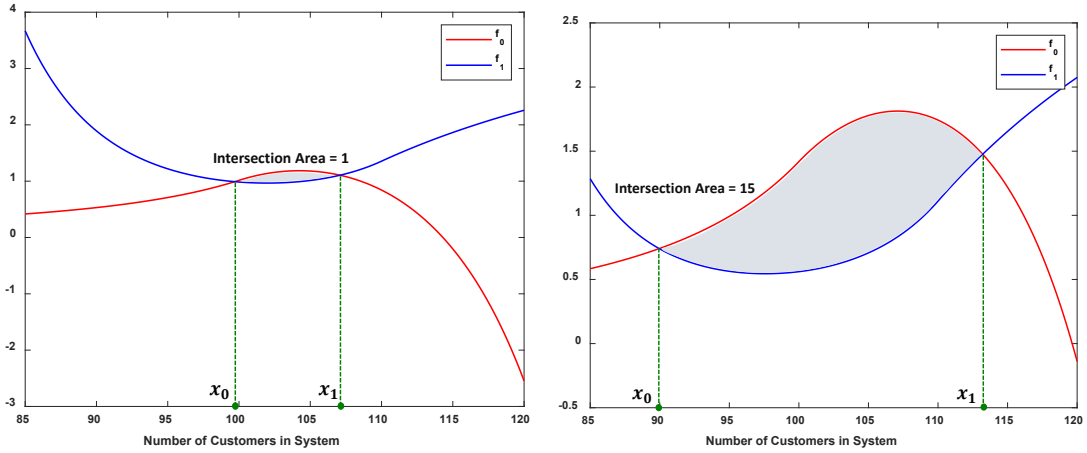
Similar to what we did in §EC.4.2, we start by selecting two interval endpoints $\eta_{\mathrm{L}} < \eta_{\mathrm{R}}$ such that $\sum_{n=-N}^{N} \delta[f_0(z_n, \eta_{\mathrm{L}}) - f_1(z_n, \eta_{\mathrm{L}})]^+ < C - \Delta$ and $\sum_{n=-N}^{N} \delta[f_0(z_n, \eta_{\mathrm{R}}) - f_1(z_n, \eta_{\mathrm{R}})]^+ > C + \Delta$. If the two functions $f_0(\cdot, \eta)$ and $f_0(\cdot, \eta)$ do not intersect for $\eta \in \{\eta_{\mathrm{L}}, \eta_{\mathrm{R}}\}$, we treat the corresponding numerical integration $\sum_{n=-N}^{N} \delta[f_0(z_n, \eta) - f_1(z_n, \eta)]^+$ as zero. In practice, we can set $\eta_{\mathrm{L}} = 0$ and $\eta_{\mathrm{R}} = \bar{\eta}$.

Since the area of the intersected region is monotonically decreasing in η, we can use a binary search on η to find the optimal $\eta^\star$ such that the integration $\sum_{n=-N}^{N} \delta[f_0(z_n, \eta^\star) - f_1(z_n, \eta^\star)]^+$ is very close to C, e.g., $|\sum_{n=-N}^{N} \delta[f_0(z_n, \eta^\star) - f_1(z_n, \eta^\star)]^+ - C| \le \Delta$.

Once we have $\eta^\star$, we can find the points of intersection by setting $z_0^\star := \min\{z_n, n \in [N] || f_0(z_n, \eta^\star) - f_1(z_n, \eta^\star)| \le \Delta\}$ and $z_1^\star := \max\{z_n, n \in [N] || f_0(z_n, \eta^\star) - f_1(z_n, \eta^\star)| \le \Delta\}$.

EC.5. Supporting Materials for the Numerical Studies

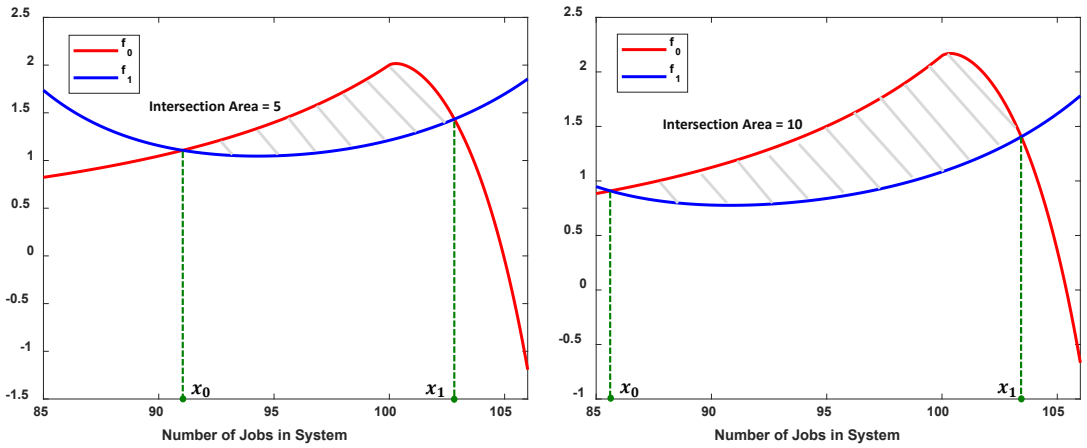
EC.5.1. DCP Solutions under Constant, Decreasing and Increasing Patience-time Hazard Rate Functions



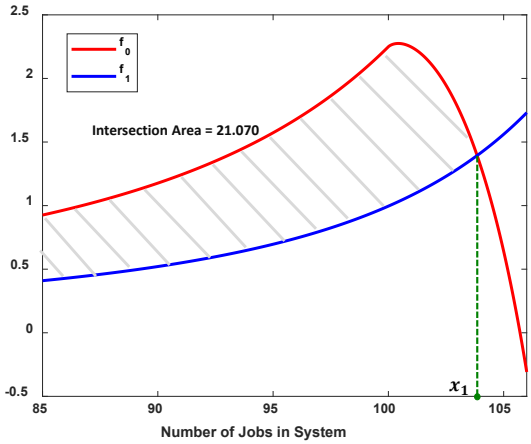
(a) $C = 1 : (\eta^\star, \lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil) = (8.096, 99, 108)$ (b) $C = 15 : (\eta^\star, \lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil) = (11.320, 89, 114)$

(c) $C = 30 : (\eta^\star, \lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil) = (12.709, 78, 117)$ (d) $C = 35.186 : (\eta^\star, \lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil) = (12.848, -\infty, 117)$

Figure EC.1 Graphical display of the Bellman equation solution $(f_0(\cdot, \eta^\star), f_1(\cdot, \eta^\star), \eta^\star, x_0^\star, x_1^\star)$ under different switching cost $C \in \{1, 15, 30, 35.158\}$ in the single-class example. In this example, We have fixed $N_0 = 100, K = 10, p = 1$, under which the maximum switching cost is solved to be $\bar{C} = 35.186$, $\eta_0 = 16.525$ and $\eta_1 = 12.848$.

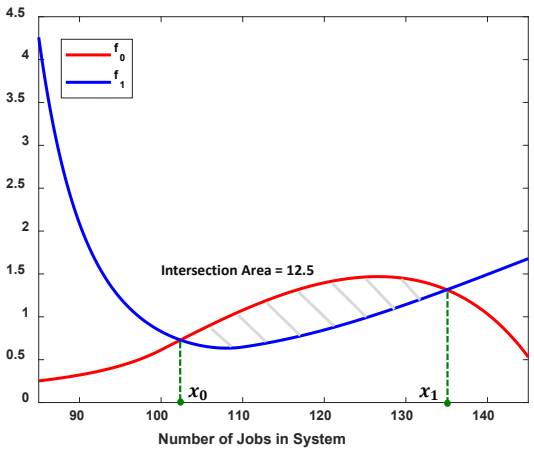


(a) $C = 5 : (\eta^\star, \lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil) = (15.967, 90, 103)$ (b) $C = 10 : (\eta^\star, \lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil) = (17.142, 85, 104)$

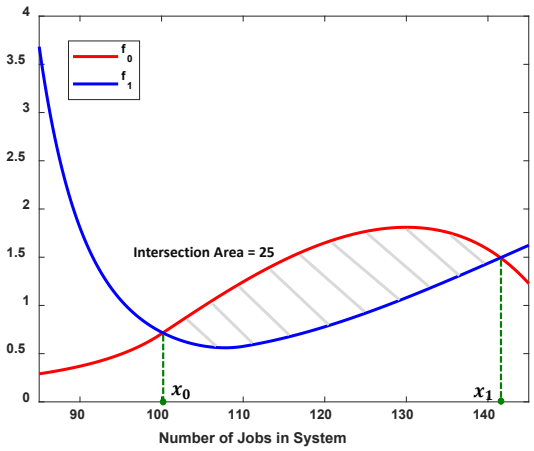


(c) $C = 21.070 : (\eta^\star, \lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil) = (17.947, -\infty, 104)$

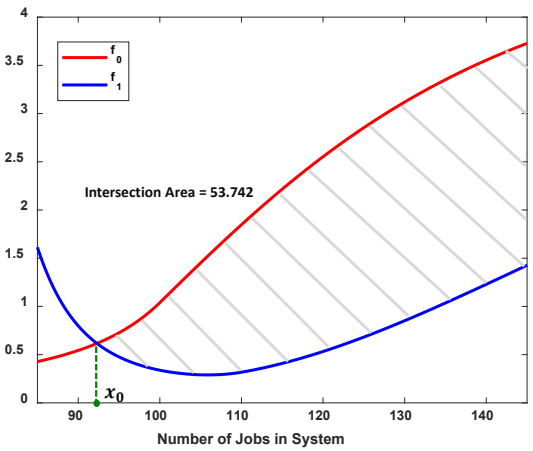
Figure EC.2 Graphical display of the Bellman equation solution $(f_0(\cdot, \eta^\star), f_1(\cdot, \eta^\star), \eta^\star, x_0^\star, x_1^\star)$ under different switching cost $C \in \{5, 10, 21.070\}$ in the single-class example, where the abandonment time has a decreasing hazard rate function $h(x) := k\hat{\theta}^k x^{k-1}$ with $k = 0.5$ and $\hat{\theta} := \theta\,\text{gamma}(1+k)$. In this example, we have fixed $N_0 = 100, K = 10, p = 1$, under which the maximum switching cost is solved to be $\bar{C} = 21.070$, $\eta_0 = 27.742$ and $\eta_1 = 17.947$.



(a) $C = 12.5 : (\eta^\star, \lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil) = (4.898, 102, 136)$



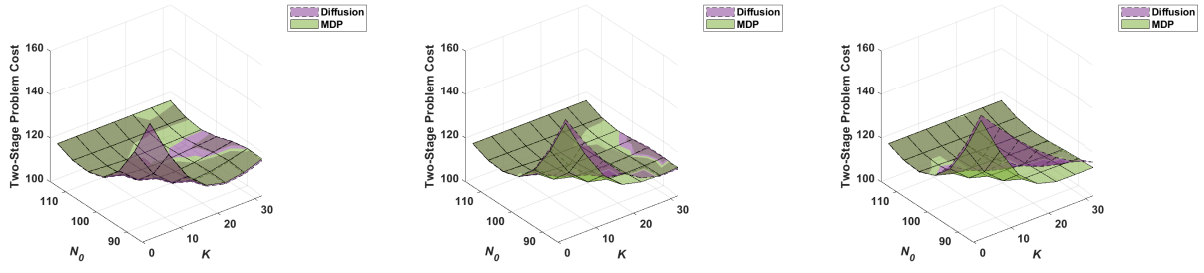
(b) $C = 25 : (\eta^\star, \lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil) = (5.638, 100, 142)$



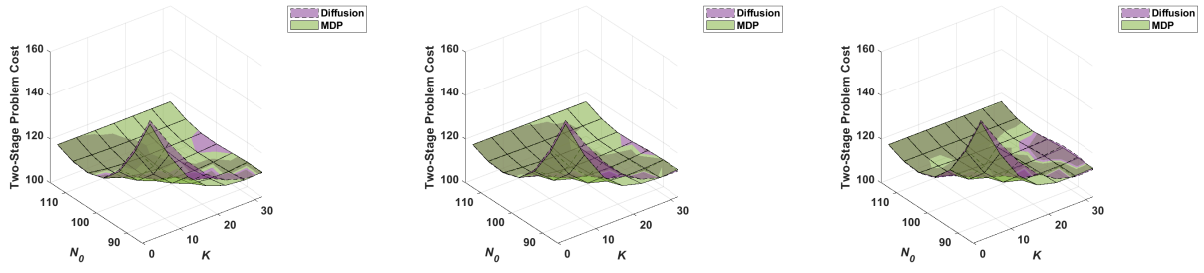
(c) $C = 53.742 : (\eta^\star, \lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil) = (8.296, 92, \infty)$

Figure EC.3 Graphical display of the Bellman equation solution $(f_0(\cdot, \eta^\star), f_1(\cdot, \eta^\star), \eta^\star, x_0^\star, x_1^\star)$ under different switching cost $C \in \{12.5, 25, 53.742\}$ in the single-class example, where the abandonment time has an increasing hazard rate function $h(x) := k\hat{\theta}^k x^{k-1}$ with $k = 2$ and $\hat{\theta} := \theta\,\text{gamma}(1+k)$. In this example, we have fixed $N_0 = 100, K = 10, p = 1$, under which the maximum switching cost is solved to be $\bar{C} = 53.742$, $\eta_0 = 8.296$ and $\eta_1 = 10.344$.

EC.5.2. Additional Numerical Results for §7

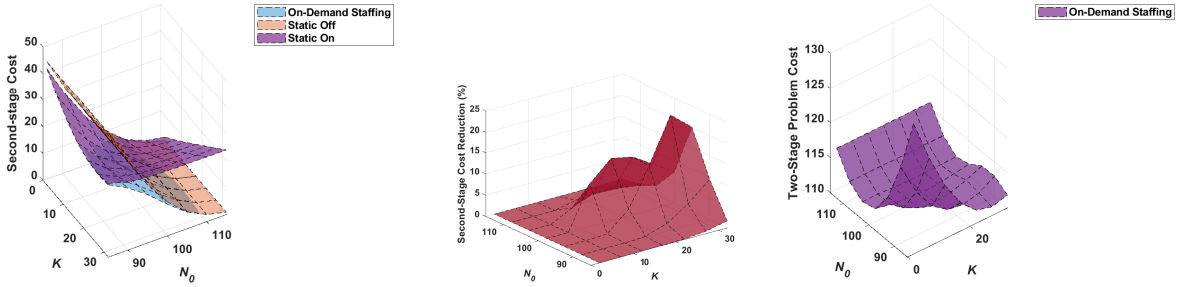


(a) $p=1$: Reduction $=37.409\%$ (b) $p=0.75$: Reduction $=37.377\%$ (c) $p=0.5$: Reduction $=38.539\%$



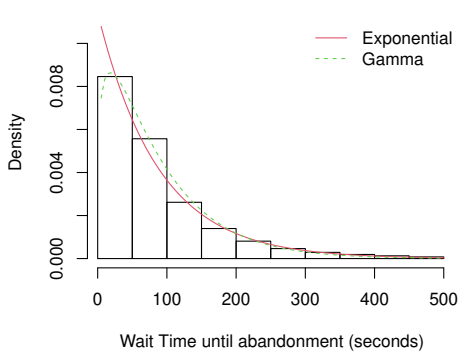
(d) $C=5$: Reduction $=39.947\%$ (e) $C=10$: Reduction $=38.440\%$ (f) $C=20$: Reduction $=37.699\%$

Figure EC.4 Theoretical first-stage problem costs computed using the diffusion approach and the exact MDP approach in the single-class example. The first-stage problem cost (or two-stage problem cost) is computed by adding the capacity planning cost $c_p N_0$ to the second-stage cost found in Figure 1. The reduction captures the difference between the largest first-stage MDP cost (given by the worst solution) and the minimal first-stage MDP cost (given by the best solution).

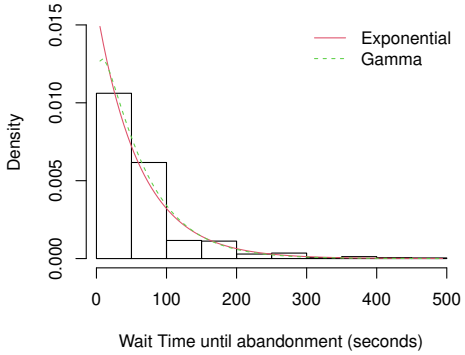


(a) (b) (c)

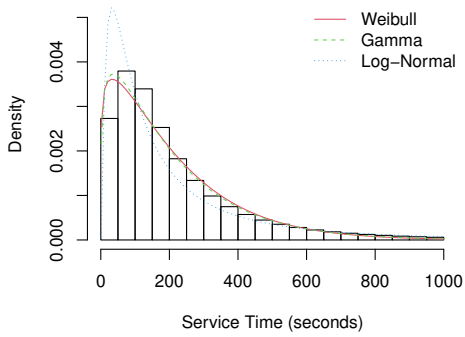
Figure EC.5 (a) Theoretical second-stage costs of the on-demand staffing policy and the two static staffing policies; (b) reduction in second-stage cost achieved by the proposed on-demand staffing policy compared to the best possible static staffing police; and (c) theoretical first-stage problem cost of the on-demand staffing policy in the two-class example.



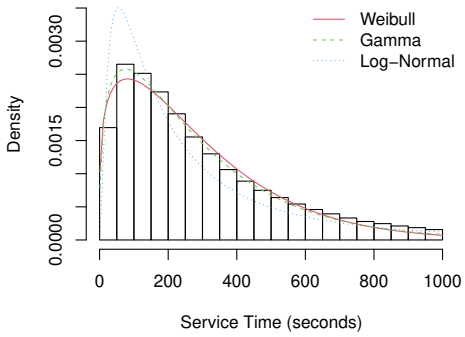
(a) Retail Banking: Abandonment Time



(b) Online Banking: Abandonment Time



(c) Retail Banking: Service Time



(d) Online Banking: Service Time

Figure EC.6 Historical (a)(b) abandonment times and (c)(d) service times of customers whose service types are retail and online banking.

EC.6. Numerical Results for Three Operating Modes

As alluded to in §8, we can enhance the system's flexibility by dividing the total number of on-demand agents, K, into two sub-pools: the Tier-1 pool and the Tier-2 pool. The sizes of these pools are K_1 and $K - K_1$, respectively. This division enables the system to switch between three operating modes, namely, mode 0: permanent staff only (N_0 agents), mode 1: permanent staff plus Tier-1 pool (at most $N_0 + K_1$ agents), and mode 2: permanent staff plus both pools (at most $N_0 + K$ agents). We next demonstrate how our solution approach can be adapted to determine the optimal switching boundaries for the three-mode staffing problem. To set up the corresponding Bellman equation, we require additional notation.

Firstly, to reflect the maximum possible number of on-demand agents in each mode, we define κ_y for $y=0,1,2$ as

$$\kappa_0 := 0, \kappa_1 := K_1/\sqrt{\lambda/\mu} \quad \text{and} \quad \kappa_2 := K/\sqrt{\lambda/\mu}.$$

We also modify the drift-rate function (10) as follows:

$$\bar{b}(y,z,\boldsymbol{q}) := -\beta\sqrt{\lambda\mu} - \kappa_y p\sqrt{\lambda\mu} + \mu\left[z-\kappa_y p\sqrt{\lambda/\mu}\right]^- - \sum_{i=1}^{I}\theta_i\left[z-\kappa_y p\sqrt{\lambda/\mu}\right]^+ q_i.$$

Secondly, to make a fair comparison between a two-mode and a three-mode system later on, we assume that the switching costs between the three operating modes, denoted as C_{01}, C_{12} and C_{02}, respectively, satisfy

$$C_{01}/(\kappa_1-\kappa_0) = C_{12}/(\kappa_2-\kappa_1) = C_{02}/(\kappa_2-\kappa_0).$$

Additionally, we set $C_{02}=C$. It is important to note that the above condition corresponds to the proportional-changeover-cost condition, which is adopted in Vande Vate (2021). This condition leads to a strongly ordered switching policy in terms of optimally controlling a simple Brownian motion, as demonstrated in that paper. While our underlying diffusion process is more complex, we conjecture that the optimal three-mode staffing policy under this condition should also have a similar ordered structure. Specifically, we conjecture that, under proper parameter configurations, the optimal policy is characterized by four switching boundaries $z_{10}^\star, z_{01}^\star, z_{21}^\star, z_{12}^\star$ satisfying the following order relationships: $z_{10}^\star < z_{01}^\star$, $z_{21}^\star < z_{12}^\star$, $z_{01}^\star < z_{12}^\star$ and $z_{10}^\star < z_{21}^\star$. This policy involves switching the system's staffing mode from 0 to 1 when the diffusion process Z, which approximates the number of jobs in the system centered around N_0, increases to $z_{01}^\star$. Similarly, it switches the mode from 1 to 2 as Z increases further to $z_{12}^\star$. Conversely, the system transitions from mode 2 to 1 as Z decreases to $z_{21}^\star$, and to mode 0 as Z decreases further to $z_{10}^\star$.

If the above conjectured policy is indeed optimal, analogous to equations (15)-(18), we should be able to find relative value functions $\bar{v}_y(\cdot)$ for $y=0,1,2$ and a constant $\bar{\eta}^\star$, such that

$$\begin{aligned}
&\lambda\bar{v}_0''(z) + \min_{\boldsymbol{q}\in\mathcal{A}}\left\{\bar{b}(0,z,\boldsymbol{q})\bar{v}_0'(z) + [z]^+\sum_i r_i\theta_i q_i\right\} = \bar{\eta}^\star \quad \text{for} \quad z < z_{01}^\star,\\
&\lambda\bar{v}_1''(z) + \min_{\boldsymbol{q}\in\mathcal{A}}\left\{\bar{b}(1,z,\boldsymbol{q})\bar{v}_1'(z) + \left[z-\kappa_1 p\sqrt{\lambda/\mu}\right]^+\sum_i r_i\theta_i q_i\right\} + c_o\kappa_1 p\sqrt{\lambda/\mu} = \bar{\eta}^\star \quad \text{for} \quad z_{10}^\star < z < z_{12}^\star,\\
&\lambda\bar{v}_2''(z) + \min_{\boldsymbol{q}\in\mathcal{A}}\left\{\bar{b}(1,z,\boldsymbol{q})\bar{v}_2'(z) + \left[z-\kappa_2 p\sqrt{\lambda/\mu}\right]^+\sum_i r_i\theta_i q_i\right\} + c_o\kappa_2 p\sqrt{\lambda/\mu} = \bar{\eta}^\star \quad \text{for} \quad z > z_{21}^\star,\\
&\bar{v}_0(z_{10}^\star) = \bar{v}_1(z_{10}^\star) \quad \text{and} \quad \bar{v}_0(z_{01}^\star) = \bar{v}_1(z_{01}^\star) + C_{01},\\
&\bar{v}_1(z_{21}^\star) = \bar{v}_2(z_{21}^\star) \quad \text{and} \quad \bar{v}_1(z_{12}^\star) = \bar{v}_2(z_{12}^\star) + C_{12}
\end{aligned}$$

subject to the boundary conditions

$$\lim_{z\to-\infty}\bar{v}_0'(z) = 0 \quad \text{and} \quad \lim_{z\to\infty}\bar{v}_2'(z) = r_*,$$

plus a set of optimality conditions derived from the "principle of smooth fit":

$$\bar{v}_0'(z_{10}^\star) = \bar{v}_1'(z_{10}^\star) \quad \text{and} \quad \bar{v}_0'(z_{01}^\star) = \bar{v}_1'(z_{01}^\star), \tag{EC.34}$$

$$\bar{v}_1'(z_{21}^\star) = \bar{v}_2'(z_{21}^\star) \quad \text{and} \quad \bar{v}_1'(z_{12}^\star) = \bar{v}_2'(z_{12}^\star). \tag{EC.35}$$

If the constant $\bar{\eta}^\star$ exists, it can serve as a guess for the average cost of the conjectured policy that switches between the three operating modes. However, we currently do not have an obvious way to formally establish the optimality of the conjectured policy or the existence of the switching boundaries $(z_{10}^\star, z_{01}^\star, z_{21}^\star, z_{12}^\star)$ and the constant $\bar{\eta}^\star$. Nonetheless, by extending the algorithms developed in §EC.4, we can numerically solve for $(z_{10}^\star, z_{01}^\star, z_{21}^\star, z_{12}^\star)$ and $\bar{\eta}^\star$ using a two-layer binary search algorithm over η and the initial boundary condition of $\bar{v}_1'(\cdot)$ to find solutions that simultaneously satisfy conditions (EC.34) and (EC.35).

Using this algorithm, we can compare the performances of (i) static staffing policies, (ii) on-demand staffing policies using a single on-call pool, and (iii) on-demand staffing policies using Tier-1 and Tier-2 pools of different sizes through numerical experiments. To illustrate this, we consider the single-class example in §7.1 and focus on the setting where the on-demand agent's show-up probability is $p = 0.75$, and the switching cost between mode 0 (permanent staff only) and mode 2 (permanent staff plus both pools) is $C_{02} = C = 15$. In this setting, the optimal first-stage solution is $(N_0, K) = (100, 17)$, as shown in the fifth row of Table 1. When $(N_0, K) = (100, 17)$, the best static staffing policy yields an average cost of $\eta_1 = 14.327$, while the proposed on-demand staffing policy using a single on-call pool produces an average cost of $\eta^\star = 11.060$. Furthermore, the optimal switching boundaries of this policy are $(\lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil) = (93, 115)$, as shown in the fourth and last rows of Table 2.

Next, we consider Tier-1 pool sizes of $K_1 = 1, 2, \ldots, K-1$ and numerically solve for the switching boundaries $(z_{10}^\star, z_{01}^\star, z_{21}^\star, z_{12}^\star)$ and the long-run average cost $\bar{\eta}^\star$ for each K_1. The cost comparison between η_1, $\eta^\star$, and $\bar{\eta}^\star$ under different K_1 is shown in Figure EC.7. We observe that the on-demand staffing policy utilizing two sub-pools leads to cost reduction compared to the policy that utilizes a single pool, and this reduction is most significant when the two sub-pools are of roughly equal sizes. The intuition behind this is that when the Tier-1 pool size is too small or large, mode 1 becomes similar to mode 0 or mode 2, and thus the three-mode staffing policy behaves more like the two-mode staffing policy. This observation is supported by the last two columns of Table EC.1, from which we see that the switching boundaries $(\lfloor x_{21}^\star \rfloor, \lceil x_{12}^\star \rceil)$ (resp. $(\lfloor x_{10}^\star \rfloor, \lceil x_{01}^\star \rceil)$) almost coincides with the optimal two-mode switching boundaries $(\lfloor x_0^\star \rfloor, \lceil x_1^\star \rceil) = (93, 115)$ when K_1 is too small (resp. large).

A more important piece of information revealed by Figure EC.7 is that the marginal benefit of having more operating modes decreases dramatically. To see this, we calculate the improvement of the three-mode staffing policy over the best static staffing policies by $(\eta_1 - \bar{\eta}^\star)/\eta_1 \times 100\%$ and over the

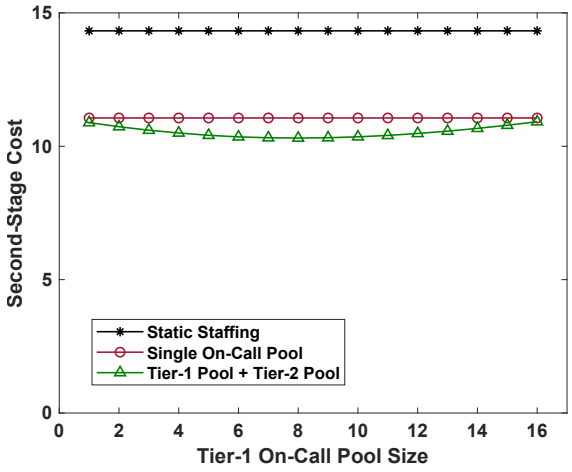


Figure EC.7 Second-stage costs of (i) the best possible static staffing policy, (ii) the on-demand staffing policy using a single on-call pool, and (iii) the on-demand staffing policy using Tier-1 and Tier-2 pools of different sizes.

two-mode staffing policy by $(\eta^\star - \bar{\eta}^\star)/\eta^\star \times 100\%$. We also calculate the benefit of having one more operating mode of flexibility (in terms of reducing the static staffing cost) by $(\eta^\star - \bar{\eta}^\star)/\eta_1 \times 100\%$. These quantities are reported in columns 2,3 and 4 of Table EC.1. From the small numbers in columns 3 and 4 compared to those in columns 2, we can see that a simpler policy that utilizes only two operating modes can reap most of the benefits of on-demand staffing.

Table EC.1 Cost improvement (improv.) of the three-mode staffing policy (i) over the best static staffing polices and (ii) over the two-mode staffing policy utilizing a single on-call pool; (iii) the benefit of having one more operating mode of flexibility in terms of reducing the static staffing cost, and (iv) the optimal switching boundaries of the three-mode staffing policy under different Tier-1 pool sizes.

Tier-1 Pool Size	Improv. Over Static Staffing(%)	Improv. Over Single Pool(%)	Benefit of Additional Flexibility(%)	$(\lfloor x^\star_{10} \rfloor, \lceil x^\star_{01} \rceil)$	$(\lfloor x^\star_{21} \rfloor, \lceil x^\star_{12} \rceil)$
1	24.010	1.566	1.209	(83,106)	(94,115)
2	25.089	2.963	2.288	(84,106)	(95,116)
3	26.005	4.151	3.204	(85,107)	(95,116)
4	26.728	5.087	3.927	(86,107)	(96,117)
5	27.334	5.872	4.533	(86,108)	(97,117)
6	27.744	6.402	4.943	(87,108)	(98,118)
7	27.982	6.711	5.181	(88,109)	(98,119)
8	28.052	6.802	5.251	(88,110)	(99,119)
9	27.970	6.696	5.169	(89,110)	(100,120)
10	27.730	6.385	4.929	(90,111)	(100,120)
11	27.365	5.912	4.564	(90,111)	(101,121)
12	26.857	5.253	4.056	(91,112)	(102,121)
13	26.254	4.473	3.453	(91,113)	(102,122)
14	25.518	3.520	2.717	(92,113)	(103,122)
15	24.701	2.461	1.900	(92,114)	(103,123)
16	23.788	1.278	0.987	(93,114)	(103,124)