Managing Flexibility: Optimal Sizing and Scheduling of Flexible Servers

Jinsheng Chen, Jing Dong
Columbia University, New York, NY 10027

Problem definition: We study the optimal joint staffing and scheduling problem in multi-class service systems, where there is an option to staff flexible servers who can handle multiple classes of customers. The specific feature we consider is that the flexible server may incur a higher cost or a loss of efficiency. We study how flexibility is best utilized in two scenarios: one with deterministic arrival rates and the other with random arrival rates. *Academic/practical relevance:* When managing resource flexibility in service systems, the conventional wisdom is that server flexibility is beneficial due to the resource pooling effect. However, in practice, flexibility often incurs some additional costs. Our work studies the interplay between the cost and benefit of flexibility in managing these systems. *Methodology:* We utilize a heavy-traffic asymptotic framework to develop structural insights. When there is no uncertainty in the arrival rates, we use a coupling argument and a diffusion approximation. When the arrival rates are random, we use a stochastic-fluid relaxation. *Results:* We derive asymptotically optimal joint staffing and scheduling rules for a two-class multi-server queue with both dedicated and flexible servers. *Managerial implications:* Our results show that the size of the flexible server pool is of a smaller order than the size of the dedicated pools, and the flexible servers are mostly used to hedge against system stochasticity or demand uncertainty, depending on which source of randomness dominates. The proposed staffing and scheduling policies are easy to implement and achieve near-optimal performance.

1. Introduction

Service systems typically involve multiple customer classes and server types. For example, in call centers, customers may require different types of service, and servers may be equipped with different skill sets (Gans et al. 2003). In hospitals, patients may be classified into differential specialties, each requiring a very different type of care, and nurses may be trained according to the care type (Best et al. 2015). Servers can sometimes be trained to handle multiple classes (types) of customers. We refer to these servers as flexible servers. Increasing the size of the flexible server pool can help balance the workload between different classes of customers, and improve system performance. Specifically, when managing queues with multiple classes of jobs, the benefit of load-balancing and capacity flexibility have been studied and demonstrated in various settings (see, for example, Andradóttir et al. (2003), Tsitsiklis and Xu (2012)).

However, flexibility may come at a cost. First, flexible servers who are capable of performing multiple types of tasks are typically more expensive to hire (Bassamboo et al. 2012). Second, multi-tasking may lead to a loss of efficiency. It is well-documented in the Psychology literature

that multi-tasking incurs cognitive switching costs which hinder productivity (Pashler 1994). A recent empirical study reveals that placing patients in the non-primary care ward can lead to worse patient outcomes including a longer length-of-stay (Song et al. 2019). Given the cost and benefit of flexible capacity, it is important to understand how to strike a balance in resource management.

When designing the service system, the service provider has to make multiple decisions. Chief among them are how many of each type of server to staff and how to match customers with servers. These problems are often referred to as the staffing and scheduling problems in the literature. In this paper, we study the joint staffing and scheduling problem in multi-class queues with both dedicated and flexible servers. In particular, to highlight the key tradeoff, we consider a stylized M-model with two classes of customers and three potential pools of servers: two dedicated pools and one flexible pool that can serve both classes of customers. To capture the cost of flexibility, we assume that the flexible servers may be more costly to staff and may serve at a slower rate than dedicated servers. The objective is to find the optimal staffing and scheduling policies that minimize the sum of the staffing cost, holding cost, and abandonment cost.

We consider two demand scenarios. One has deterministic arrival rates, which is the case when we have a very accurate estimate of customer demand. In this case, the flexible pool can be used to hedge against stochasticity, i.e., the stochastic fluctuation of interarrival times and service times. In particular, due to the stochasticity in system dynamics, one queue may incur a higher than average load while the other is at or below its normal load from time to time. In such situations, the flexible pool can be used to help the class with a heavier load, and thus balance the load between the two classes. The other scenario has random arrival rates, which is the case when there is a high degree of uncertainty in customer demand. In this case, the flexible pool is mainly used to hedge against parameter uncertainty. In particular, when the realized arrival rate of one class is higher than average while the realized arrival rate of the other class is at or below average, the flexible pool can be used to help the class with a higher realized arrival rate, and thus balance the load. The differences between the two scenarios described above give rise to different hedging mechanisms, which in turn lead to different sizes of the flexible pool in optimality. To see this, let λ denote the average arrival rate. When λ is large, the stochastic fluctuation of the system with a given arrival rate is in general of order $\sqrt{\lambda}$ (Garnett et al. 2002). The parameter uncertainty, on the other hand, can be of a different order than $\sqrt{\lambda}$ (Bassamboo et al. 2010b). Indeed, the case we are interested in is one where the standard deviation of the random arrival rate is of a larger order than $\sqrt{\lambda}$. Lastly, the different hedging mechanisms also lead to different scheduling policies in our developments.

Because staffing and scheduling decisions interact, the joint optimization problem can be very challenging. When arrival rates are deterministic and symmetric, we use a coupling construction

to derive the optimal scheduling policy for any staffing level. The scheduling policy prioritizes the dedicated servers (faster servers) when routing customers to servers, and prioritizes the class with more customers in the system when scheduling flexible servers, assuming the abandonment rate is less than the service rates. Given the optimal scheduling policy, we then optimize the staffing policy. To derive structural insights into the size of the flexible pool, we employ a heavy-traffic asymptotic approach, where we send the arrival rate to infinity and study how the size of the flexible pool scales with the arrival rate. Our result provides necessary and sufficient conditions for staffing rules to be asymptotically optimal. The key insight is that when flexibility comes at a cost, the optimal size of the flexible pool only leads to partial resource pooling. In particular, the flexible pool helps create some load-balancing, but the effect is not large enough to equalize the two queues asymptotically.

When arrival rates are random and the magnitude of the parameter uncertainty dominates the system stochasticity, we employ a stochastic-fluid relaxation of the optimal staffing problem. In this relaxation, we ignore the stochasticity of the queueing dynamics and focus on the parameter uncertainty only. The stochastic-fluid optimization problem is a special case of the single-period multi-product inventory problem with demand substitution, for which we can characterize the optimal solution explicitly. The relaxation also motivates a simple scheduling rule that essentially decomposes the M-model into two independent inverted-V models for any realization of the arrival rates. When the average arrival rates grow to infinity, we show that the staffing and scheduling rules derived based on the stochastic-fluid relaxation are asymptotically optimal. The key insight is that when facing both parameter uncertainty and cost of flexibility, the optimal size of the flexible pool provides some hedging against the parameter uncertainty, and the cost saving, compared to the no-flexible resource case, is increasing with the magnitude of the uncertainty.

In addition to providing prescriptive solutions to managing flexibility, we also highlight the following contributions of our work.

1. When the arrival rates are symmetric and deterministic, we construct the optimal scheduling policy for any arrival rates and staffing levels. In contrast to most of the optimal scheduling literature for multi-server queues, our results do not rely on any asymptotic argument (for development on asymptotically optimal scheduling policies, see, for example, Atar (2005)). Instead, the proof uses a coupling argument that can be of interest to the analysis of other Markovian queueing systems. Our coupling technique also allows us to establish the optimality of a non-standard scheduling policy when the abandonment rate is larger than the service rates, see Theorem 5.

2. When the arrival rates are deterministic and the flexible pool is of the optimal order, we derive the diffusion limit of the M-model under heavy-traffic. The limit is a two-dimensional diffusion process. In particular, the complete resource pooling condition is not satisfied when the flexible

pool is optimally sized, i.e., the flexible pool size is not large enough to instantaneously balance the queue lengths between the two classes. Thus, we do not have state space collapse in the limit, i.e., the two-dimensional queue length process does not reduce to a one-dimensional process in the limit. This is in contrast to most of the optimal scheduling literature (see, for example, Dai and Tezcan (2011), Gurvich and Whitt (2009a)). On the other hand, the limiting process cannot be fully decomposed along each dimension, i.e., the drift terms of the two component diffusion processes are interconnected. Thus, we achieve partial resource pooling.

3. When the arrival rates are random and the parameter uncertainty is of a larger order than the stochasticity of the queueing dynamics, we quantify the optimality gap for policies derived based on the stochastic fluid approximation. This extends the results in Bassamboo et al. (2010b) from a multi-server queue with a single class of customers and a single pool of servers to a multi-class queue with multiple server types. We also allow the arrival rate distributions of the two classes to be asymmetric, i.e., they can have different means and different levels of uncertainty.

1.1. Literature review

We first review related works on queues with deterministic arrival rates. The M-model studied in this paper is a special case of parallel server systems (PSSs). Due to the interplay between staffing and scheduling decisions, the joint staffing and scheduling problem can be highly nontrivial for general PSSs. In the literature, most works only look at one of the two problems in isolation. However, there are a few exceptions. Noticeably, Armony and Mandelbaum (2011) consider the joint optimization problem for an inverted-V model where there is a single class of customers and multiple types of servers. Using a coupling argument, they establish the optimality of the fastest-server-first policy. Gurvich and Whitt (2009b) study the problem of staffing and scheduling PSSs to minimize total staffing costs subject to quality-of-service constraints. They establish that the queue-and-idleness-ratio control is asymptotically optimal in heavy-traffic. When dealing with a single class of customers and a single pool of servers, Borst et al. (2004) study the optimal staffing problem in an $M/M/n$ queue. They find that the quality-and-efficiency-driven (QED) regime, which is also known as the Halfin-Whitt regime (Halfin and Whitt 1981), arises naturally when staffing is set to balance the staffing cost and the system performance. The work is then extended by Mandelbaum and Zeltyn (2009) to allow for customer abandonment.

The work that is most related to ours is Bassamboo et al. (2012), which studies the sizing of flexible resources when service rates can be continuously chosen. They find that the linear staffing and holding costs often lead to an $O(\sqrt{\lambda})$ flexibility when flexible capacity is more expensive. The main difference between our work and theirs is the modeling of the service resources. They use a single-server mode of analysis and assume the service rate can be optimally chosen. This modeling

approach is reasonable for computer or manufacturing systems. Motivated mostly by large-scale service systems, our work adopts a many-server mode of analysis. As Bassamboo et al. (2012) point out, the many-server regime that we consider introduces substantial complexity to the analysis, and they leave this extension as a potential future research direction. In addition, Bassamboo et al. (2012) assumes a longest-queue-first scheduling and hypothesize that it is likely to be optimal. We establish the optimality of a scheduling policy that prioritizes the class with more customers in the system.

More broadly, optimal scheduling of various PSSs has been extensively studied in the literature. For example, Tezcan and Dai (2010) study the optimal scheduling of the N-model. They show that a $c\mu$-type of greedy policy is asymptotically optimal in the many-server QED regime. Atar (2005) studies the optimal scheduling problem of general PSSs, i.e., with multiple classes of customers and multiple pool of servers, and customer abandonment. The work establishes the asymptotical optimality of policies derived based on the corresponding optimal diffusion control problem in the many-server QED regime. Kim et al. (2018) study the optimal scheduling of V-model with general patience-time distributions. The main feature that distinguishes our work from the stream of works on PSSs in the QED regime is the size of our flexible server pool. In our analysis, the size of the flexible pool is asymptotically negligible in the fluid scale, whereas in the literature, it is almost universally assumed that the fluid-scaled pool sizes are non-negligible (see, for example, Assumption 1 in Atar (2005), Assumption 2.1 in Gurvich and Whitt (2009a), and equation (20) in Dai and Tezcan (2011)). Due to the difference in the size of our server pools, the asymptotic behavior (diffusion limit) of our system can be qualitatively different from what is observed in the literature.

When the arrival rates are random, our work is related to works that look at staffing queues when facing parameter uncertainty. The stochastic-fluid relaxation was first proposed in Harrison and Zeevi (2005). Its efficacy has been studied in several subsequent works. Bassamboo et al. (2006) show that it leads to an asymptotically optimal staffing policy under a non-conventional asymptotic regime that features large arrival rates and short service times. The asymptotic framework is then extended in Bassamboo and Zeevi (2009), who consider the case when the arrival rate distribution is unknown and has to be estimated from data. Compared to these works, the analysis in this paper takes a different asymptotic approach. In particular, we increase the system demand, i.e., arrival rates, but do not scale other system parameters such as service rates and abandonment rates. The paper Bassamboo et al. (2010b) takes a similar heavy-traffic asymptotic approach as ours and establishes the optimality gap of the staffing policy derived from the stochastic-fluid relaxation for an Erlang-A model with a random arrival rate. We extend their results to a multi-class network setting, where in addition to the staffing decision, we also have to decide on the scheduling policy.

Whitt (2006) develops a different stochastic-fluid model that allows non-exponential service times and patience times, and studies the staffing problem with both random arrival rates and staffing levels (due to employee absenteeism). When facing demand uncertainty, Gurvich et al. (2010) study the staffing problem with a chance constraint for the quality of service. They first use mixed integer programming to obtain a first-order staffing solution, and then refine the staffing level using simulation. Koçağa et al. (2015) study the staffing and outsourcing problem when demand is random.

Our work contributes to this stream of literature in two key ways. First, we show that when dealing with random demand, it is the staffing, not the scheduling decision, that is of paramount importance. This supports why many papers tend to focus on the staffing instead of the scheduling decision in this setting (see, for example Gurvich et al. (2010), Bertsimas and Doan (2010)). Second, we quantify the benefit of flexibility. Specifically, we extend the notion that the order of flexibility should match the order of system stochasticity in Bassamboo et al. (2012) to the case where the order of flexibility should match the order of demand uncertainty. In general, the notion that just a small degree of flexibility is enough has been much investigated over the years in various different contexts (see, for example, Simchi-Levi and Wei (2012), Shi et al. (2019) for manufacturing systems, Tsitsiklis and Xu (2012), Wallace and Whitt (2005) for PSSs, etc.) Our work contributes to this literature as well.

1.2. Paper structure and notations

In Section 2, we introduce the queueing model and the optimization problem. In Section 3, we study the optimal scheduling and staffing policy for a symmetric M-model with deterministic arrival rates. The goal is to highlight the cost and benefit of flexibility in a classical setting with no parameter uncertainty. In Section 4, we study the staffing and scheduling problem for systems with random arrival rates. To highlight the effect of demand uncertainty, we focus on the regime where the demand uncertainty dominates the system stochasticity. We complement our theoretical analysis with numerical experiments in Section 5. In particular, the numerical analysis focuses on the pre-limit performance of our proposed staffing and scheduling rules. The proofs of all the theoretical results are delayed until the Appendix.

We next introduce some notations that are used throughout the paper. The set of non-negative integers is denoted by $\mathbb{N}_0$, and the set of real numbers is denoted by $\mathbb{R}$. We define $\eta(t) = 0$, $\chi(t) = t$, and $I(t) = 1$, for $t \geq 0$. Let D denote the space of functions from $[0, \infty)$ to $\mathbb{R}$ that are right-continuous with left limits and is endowed with Skorohod J_1 topology. Let e_i be a unit vector with the i-th element equal to 1. The dimension of e_i depends on the context. We write $1\{\cdot\}$ for the indicator function. A random variable A is said to be stochastically larger than a random variable

B, $A \geq_{st} B$, if $\mathbb{P}(A > x) \geq \mathbb{P}(B > x)$ for any $x \in \mathbb{R}$. For real sequences $\{a_n\}$ and $\{b_n\}$, we say that $a_n = O(b_n)$ if $\limsup_{n\to\infty} |a_n|/b_n < \infty$, $a_n = o(b_n)$ if $\limsup_{n\to\infty} |a_n|/b_n = 0$, and $a_n = \Theta(b_n)$ if $\liminf_{n\to\infty} |a_n|/b_n > 0$. For $a \in \mathbb{R}$, write $a^+ = \max(a, 0)$ and $a^- = \max(-a, 0)$.

2. The Model

We consider a classical M-model with possible demand uncertainty as depicted in Figure 1. In particular, the model has two customer classes, Class 1 and Class 2, and three pools of servers: two dedicated pools for the two customer classes and one flexible pool that can serve both classes. We allow the arrival rate for Class i, Λ_i, $i = 1, 2$, to be a random variable. For a given realization of Λ_i, i.e., $\Lambda_i = \lambda_i$, Class i arrivals follow a Poisson process with rate λ_i. Each server pool can have multiple servers. We write n_i for the number of servers in the dedicated pool for Class i, and n_F for the number of servers in the flexible pool. If a customer is served by the dedicated server, its service time follows an exponential distribution with rate μ. If a customer is served by the flexible server, its service time follows an exponential distribution with rate μ_F. We assume $\mu_F \leq \mu$ to account for the potential efficiency loss of flexible servers. Each customer has a patience time that follows an exponential distribution with rate θ. Once a customer's waiting time (in the queue) exceeds its patience time, it abandons the system.

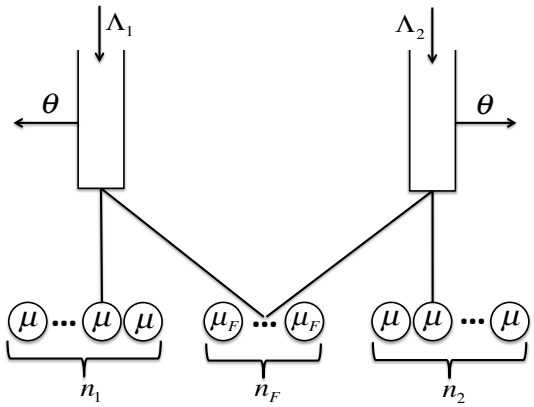

Figure 1 The M-model

For $i = 1, 2$, let $X_i(t)$ denote the number of Class i customers in the system at time t. We denote $Z_i(t)$ and $Z_{Fi}(t)$ as the number of dedicated servers and flexible servers serving Class i customers at time t respectively. Note that

$$Z_i(t) \leq n_i, Z_F(t) := Z_{F1}(t) + Z_{F2}(t) \leq n_F, \text{ and } Z_i(t) + Z_{Fi}(t) \leq X_i(t). \tag{1}$$

Let $Q_i(t)$ denote the number of Class i customers waiting in the queue at time t. Then $Q_i(t) = X_i(t) - Z_i(t) - Z_{Fi}(t)$. Let $X(t) := (X_1(t), X_2(t))$, $Z(t) := (Z_1(t), Z_2(t), Z_{F1}(t), Z_{F2}(t))$, and $Q(t) :=$

$(Q_1(t), Q_2(t))$. We also define the total number of customers in the system and the total queue length processes as $X_\Sigma(t) = X_1(t) + X_2(t)$ and $Q_\Sigma(t) = Q_1(t) + Q_2(t)$ respectively. Let A_i, S_i, S_{Fi}, G_i be independent unit-rate Poisson processes, which will be used to represent the arrival, departure, and abandonment events respectively. At the beginning of the planning horizon Λ_i is realized. Given $\Lambda_i = \lambda_i$, $X_i(t)$ satisfies the following dynamics:

$$X_i(t) = X_i(0) + A_i(\lambda_i t) - G_i\left(\theta \int_0^t Q_i(s)\, ds\right) - S_i\left(\mu \int_0^t Z_i(s)\, ds\right) - S_{Fi}\left(\mu_F \int_0^t Z_{Fi}(s)\, ds\right).$$

To fully describe dynamics of the system, we need to specify the scheduling policy – how to allocate the servers. We restrict ourselves to preemptive deterministic Markovian policies, where the allocation of servers $Z(t)$ can be viewed as a function of the current state of the system $X(t)$ (Harrison and Zeevi 2004). Let ν denote such a policy (mapping), i.e., $Z(t) = \nu(X^\lambda(t))$ and it satisfies the feasibility conditions listed in (1).

Let $Q_\Sigma(\infty; n_1, n_2, n_F; \nu)$ be the steady-state total queue length given staffing level (n_1, n_2, n_F) and scheduling policy ν. If the system is not stable under a certain staffing and scheduling rule, we define $Q_\Sigma(\infty; n_1, n_2, n_F; \nu) \equiv \infty$. Our goal is to jointly choose the staffing levels for each pool and the scheduling policy to minimize the sum of the staffing costs and the steady-state average holding and abandonment costs:

$$\min_{n_1, n_2, n_F, \nu} \Pi(n_1, n_2, n_F; \nu) := c(n_1 + n_2) + c_F n_F + (h + a\theta)\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)], \tag{2}$$

where $c > 0$ is the per server per unit time staffing cost for the dedicated pools, $c_F > 0$ is the per server per unit time staffing cost for the flexible pool, $h > 0$ is the per customer per unit time holding cost, and $a > 0$ is the per customer abandonment cost. Note that the abandonment cost is $a\theta\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)]$ because $\theta\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)]$ is the rate at which customers abandon in stationarity. We also note that

$$\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)] = \mathbb{E}[\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)|\Lambda]],$$

where $\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)|\Lambda = (\lambda_1, \lambda_2)]$ is the steady-state average queue length of an M-model with arrival rates (λ_1, λ_2), and the outer expectation is taken with respect to the random arrival rates Λ, i.e., the demand uncertainty.

In order to avoid trivial situations, we impose the following condition on the rates and cost parameters:

$$c/\mu < c_F/\mu_F < h/\theta + a. \tag{3}$$

The first inequality ensures that flexible servers have some disadvantage over dedicated servers. Otherwise, we would never staff dedicated servers. The second inequality ensures that the cost of

serving a customer using a flexible server is less than the cost of letting the customer wait and abandon. Otherwise, we would never staff flexible servers.

We highlight two challenges in solving (2). First, even for a given staffing level, characterizing the optimal scheduling policy can be highly nontrivial. Second, even after pinning down the optimal scheduling policy, it remains difficult to solve for the optimal staffing level due to the lack of an analytical characterization of $\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)]$. We will address these challenges in subsequent sections. In particular, a lot of our developments rely on a heavy-traffic asymptotic mode of analysis, in which our goal is to characterize how the optimal decisions scale with the arrival rate (average arrival rate) λ as $\lambda \to \infty$. To explicitly mark the dependence of the policies and system dynamics on λ, we use the superscript λ. For example, ν^λ and $(n_1^\lambda, n_2^\lambda, n_F^\lambda)$ are the scheduling policy and staffing levels for the system with the arrival rate parameter λ (i.e., the λ-th system). Similarly, X^λ, Z^λ, and Q^λ are the number-in-system, number-in-service, and number-in-queue processes of the λ-th system.

3. The Case with Deterministic Arrival Rate

In this section, we study a special case of the system where the arrival rate is deterministic. In particular, we assume $\Lambda_1 = \Lambda_2 = \lambda$ with probability 1. In this case, we have a symmetric M-model. The goal is to highlight how to strike a balance between the cost and benefit of flexibility.

We start by providing an overview of how we address the two challenges listed in Section 2 to derive the optimal staffing and scheduling rules jointly. In Section 3.1, we use a coupling argument to derive the optimal scheduling policy for any given staffing level. This optimal scheduling rule turns out to have a very neat and intuitive structure. In particular, the policy prioritizes the faster servers (the dedicated servers) when routing customers to servers, and the flexible servers prioritize the class with more customers in the system (the larger $X_i(t)$). This is similar in structure to the queue-idleness ratio policy (Gurvich and Whitt 2009a), the fastest server first policy (Armony and Mandelbaum 2011), and the max-pressure policy (Dai and Lin 2008). However, we emphasize that using the coupling argument, we are able to establish exact optimality instead of asymptotic optimality. Moreover, in the staffing regime we are interested in, there is no state-space collapse. Thus, the asymptotic optimality framework leveraged in the literature can no longer be applied.

Next, in Section 3.2, we take a heavy-traffic asymptotic approach to derive a necessary and sufficient characterization of the optimal staffing rules. In particular, we gradually send the arrival rate λ to infinity and study how the optimal staffing level scales with λ. Our analysis shows that the optimal staffing rule leads the system to operate in the QED regime, and the optimal size of the flexible pool is $O(\sqrt{\lambda})$. This extends the insights developed in Bassamboo et al. (2012) to the many-server setting.

Due to the symmetry of the system, we assume, without loss of optimality, that $n_1^\lambda = n_2^\lambda = n^\lambda$. Thus, our decision variables for the staffing rule reduce to n^λ and n_F^λ. For the model analyzed in this section, we allow $\theta = 0$, i.e., no abandonment. When $\theta = 0$, we need to put more restrictions on the staffing levels to ensure system stability. In particular, we define

$$\Omega^\lambda(0) := \left\{(n^\lambda, n_F^\lambda) \in \mathbb{N}_0^2 : 2\lambda < 2n^\lambda\mu + n_F^\lambda\mu_F\right\}.$$

The following lemma show that when $\theta = 0$, having $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(0)$ ensures that the system is stable under the optimal scheduling rule.

LEMMA 1. *If $\theta = 0$, for any $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(0)$ there exists a scheduling policy ν^λ, under which the stochastic process X^λ has a unique stationary distribution.*

To ensure consistent notation, for $\theta > 0$ we define

$$\Omega^\lambda(\theta) = \{(n^\lambda, n_F^\lambda) \in \mathbb{N}_0^2\}.$$

3.1. Optimal Scheduling Rule

Intuitively, a good scheduling policy should reduce the queues as fast as possible and balance the queues of the two classes. This motivates the following scheduling rule. For the dedicated pool of servers,

$$Z_i^\lambda(t) = \min\{n^\lambda, X_i^\lambda(t)\} \text{ for } i = 1, 2; \tag{4}$$

and for the flexible pool of servers, if $X_1^\lambda(t) \geq X_2^\lambda(t)$,

$$Z_{F1}^\lambda(t) = \min\{n_F^\lambda, (X_1^\lambda(t) - n^\lambda)^+\}, \quad Z_{F2}^\lambda(t) = \min\{n_F^\lambda - Z_{F1}^\lambda(t), (X_2^\lambda(t) - n^\lambda)^+\}; \tag{5}$$

otherwise,

$$Z_{F1}^\lambda(t) = \min\{n_F^\lambda - Z_{F2}^\lambda(t), (X_1^\lambda(t) - n^\lambda)^+\}, \quad Z_{F2}^\lambda(t) = \min\{n_F^\lambda, (X_2^\lambda(t) - n^\lambda)^+\}. \tag{6}$$

Note that under this policy, we first try to assign as many customers to the dedicated pools as possible, i.e., (4). Then, for the flexible pool, we give priority to the class with more customers in the system, i.e., (5) and (6). We comment that for our scheduling policy, ties can be broken in an arbitrary way. For simplicity of exposition, we assume that when $X_1^\lambda(t) = X_2^\lambda(t)$, the flexible pool gives priority to Class 1. We denote the policy defined in (4) - (6) as $\nu^{\lambda,*}$.

The next theorem shows that when $\theta \leq \mu_F$, for any fixed staffing level (n^λ, n_F^λ), $\nu^{\lambda,*}$ is optimal.

THEOREM 1. *Suppose $\theta \leq \mu_F$. For any Markovian scheduling policy ν^λ,*

$$\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \nu^\lambda)] \geq \mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \nu^{\lambda,*})],$$

which implies that $\Pi^\lambda(n^\lambda, n_F^\lambda; \nu^\lambda) \geq \Pi^\lambda(n^\lambda, n_F^\lambda; \nu^{\lambda,*})$.

Note that for $\theta \le \mu_F$, the policy $\nu^{\lambda,*}$ tries to equalize X_1^λ and X_2^λ at the maximum rate. Due to the symmetric in the system structure, we expect this policy to perform well. We prove the theorem by developing a coupling construction based on the transition rates of the underlying Markov processes (see Appendix B.2 for more details). We also comment that the condition $\theta \le \mu_F$ is necessary for $\nu^{\lambda,*}$ to be optimal. If $\theta > \mu_F$, $\nu^{\lambda,*}$ no longer equalizes X_1^λ and X_2^λ at the maximum rate, because a larger rate can be attained by keeping customers waiting in the queue instead of sending them to the flexible severs. Indeed, when $\theta \ge \mu_F = \mu$, we can show that a scheduling rule that prioritizes the class with fewer customers in the system is optimal (see Theorem 5 in Appendix B.3).

3.2. Asymptotically Optimal Staffing Rule

Based on the analysis in Section 3.1, the scheduling policy $\nu^{\lambda,*}$ is optimal for any λ and (n^λ, n_F^λ) when $\theta \le \mu_F$. In subsequent analysis, we assume without loss of optimality that the policy $\nu^{\lambda,*}$ is employed. When there is no confusion, we will omit the scheduling policy from the notation of the corresponding stochastic processes. Now, the problem of jointly optimizing staffing and scheduling rules, i.e., (2), reduces to optimizing the staffing levels only:

$$\min_{(n^\lambda, n_F^\lambda)\in\Omega^\lambda(\theta)} \Pi^\lambda(n^\lambda, n_F^\lambda) := 2cn^\lambda + c_F n_F^\lambda + (h + a\theta)\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda)]. \tag{7}$$

Solving (7) analytically is still challenging due to the lack of a closed-form expression for $\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda)]$. In this section, we study the structure of the optimal staffing levels under heavy traffic. In particular, we send $\lambda \to \infty$ while keeping the service rates and abandonment rates fixed. Our analysis reveals how the optimal sizes of the dedicated pool and flexible pool scale with the arrival rate λ.

Let

$$\Pi^{\lambda,*} := \min_{(n^\lambda, n_F^\lambda)\in\Omega^\lambda(\theta)} \Pi^\lambda(n^\lambda, n_F^\lambda) \text{ and } (n^{\lambda,*}, n_F^{\lambda,*}) \in \operatorname*{arg\,min}_{(n^\lambda, n_F^\lambda)\in\Omega^\lambda(\theta)} \Pi^\lambda(n^\lambda, n_F^\lambda).$$

Define $R^\lambda := \lambda/\mu$, which is the offered load of Class i, $i = 1, 2$.

LEMMA 2. *Suppose* (3) *holds. Then,* $\Pi^{\lambda,*} = 2cR^\lambda + O(\sqrt{\lambda})$. *Moreover, for* $(n^{\lambda,*}, n_F^{\lambda,*})$,

$$-\infty < \liminf_{\lambda\to\infty} \frac{n^{\lambda,*} - R^\lambda}{\sqrt{\lambda}} \le \limsup_{\lambda\to\infty} \frac{n^{\lambda,*} - R^\lambda}{\sqrt{\lambda}} < \infty$$

and

$$\limsup_{\lambda\to\infty} \frac{n_F^{\lambda,*}}{\sqrt{\lambda}} < \infty.$$

Motivated by Lemma 2, our goal in subsequent analysis is to close the $O(\sqrt{\lambda})$ optimality gap. In particular, we employ the following notion of asymptotic optimality.

DEFINITION 1. A sequence of staffing levels (n^λ, n_F^λ) (indexed by λ) is asymptotically optimal if

$$\Pi^\lambda(n^\lambda, n_F^\lambda) = \Pi^{\lambda,*} + o(\sqrt{\lambda}).$$

The key question we would like to address is how much flexibility is optimal. We first note that if there is no 'cost' of flexibility, then we would want as much flexibility as possible. This is because flexible servers create resource pooling in the system. To be more precise, we have the following result:

LEMMA 3. *When $\mu = \mu_F \geq \theta$, we have*

$$\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda)] \geq \mathbb{E}[Q_\Sigma^\lambda(\infty; 0, 2n^\lambda + n_F^\lambda)].$$

In practice, flexibility often comes at a cost. Here, we consider two forms of cost: a higher staffing cost, i.e., $c_F \geq c$, and an efficiency cost, i.e., $\mu_F \leq \mu$. In this case, Lemma 2 indicates that, overall, it is optimal to follow the square-root staffing rule, i.e., $2n^{\lambda,*} + n_F^{\lambda,*} = 2R^\lambda + O(\sqrt{\lambda})$. More importantly, $n_F^{\lambda,*}$ cannot be too large, i.e., $n_F^{\lambda,*} = O(\sqrt{\lambda})$.

To derive an asymptotically optimal staffing rule, we need to have a good approximation of $\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda)]$ in (7). In the many-server heavy-traffic analysis, there are two commonly used approximations: the fluid approximation and the diffusion approximation. To achieve $o(\sqrt{\lambda})$ optimality, we need to use the finer-scale diffusion approximation. We next define some diffusion-scaled processes. Let

$$\hat{X}_i^\lambda(t) = \frac{X_i^\lambda(t) - n^\lambda}{\sqrt{\lambda}},\ \hat{Q}_i^\lambda(t) = \frac{Q_i^\lambda(t)}{\sqrt{\lambda}},\ \text{ and }\ \hat{Z}_i^\lambda(t) = \frac{Z_i^\lambda(t) - n^\lambda}{\sqrt{\lambda}} \text{ for } i = 1, 2.$$

We also write $\hat{X}^\lambda = (\hat{X}_1^\lambda, \hat{X}_2^\lambda)$, and define

$$\hat{X}_\Sigma^\lambda(t) = \frac{X_\Sigma^\lambda(t) - 2n^\lambda - n_F^\lambda}{\sqrt{\lambda}} \ \text{ and }\ \hat{Q}_\Sigma^\lambda(t) = \frac{Q_\Sigma^\lambda(t)}{\sqrt{\lambda}}.$$

In our subsequent development, for any stochastic process $Y(t)$, we write $Y(\infty)$ as its stationary distribution.

Recall that $n_F^{\lambda,*} = O(\sqrt{\lambda})$. The following theorem characterizes the diffusion limit of the number-in-system processes in this case.

THEOREM 2. *For $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$, suppose $n^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ and $n_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$, where $\beta \in \mathbb{R}, \beta_F \geq 0$, and if $\theta = 0$, $2\beta\mu + \beta_F\mu_F > 0$. Then, if $\hat{X}^\lambda(0) \Rightarrow \hat{X}(0)$ as $\lambda \to \infty$,*

$$\hat{X}^\lambda \Rightarrow \hat{X} \text{ in } D^2 \text{ as } \lambda \to \infty,$$

where $\hat{X}$ is a two-dimensional diffusion process with

$$d\hat{X}_i(t) = \left(-\beta\sqrt{\mu} + \mu\hat{X}_i(t)^- - (\mu_F - \theta)f_i\left(\hat{X}_1(t), \hat{X}_2(t)\right) - \theta\hat{X}_i(t)^+\right)dt + \sqrt{2}\,dB_i(t),$$

for $i=1,2$, B_1 and B_2 are independent standard Brownian motions, and

$$f_1(x_1,x_2)=\begin{cases} x_1^+ \wedge \frac{\beta_F}{\sqrt{\mu}} & \text{if } x_1 \geq x_2, \\ x_1^+ \wedge \left(\frac{\beta_F}{\sqrt{\mu}} - x_2^+\right)^+ & \text{if } x_1 < x_2; \end{cases} \quad f_2(x_1,x_2)=\begin{cases} x_2^+ \wedge \left(\frac{\beta_F}{\sqrt{\mu}} - x_1^+\right)^+ & \text{if } x_1 \geq x_2, \\ x_2^+ \wedge \frac{\beta_F}{\sqrt{\mu}} & \text{if } x_1 < x_2. \end{cases}$$

Moreover,

$$\mathbb{E}[\hat{Q}_\Sigma^\lambda(\infty)] \to \mathbb{E}[(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu})^+] \text{ as } \lambda \to \infty.$$

We make two important observations from Theorem 2. First, to characterize $\hat{X}_\Sigma^\lambda$, we need to keep track of a two-dimensional diffusion process $\hat{X}$ in the limit. In this sense, we do not achieve complete resource pooling. On the other hand, the drift terms of $\hat{X}$ cannot be fully decomposed along each dimension, i.e., $f_i(x_1,x_2)$ depends on both x_1 and x_2. Thus, we achieve partial resource pooling. Second, $\mathbb{E}[(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu})^+]$ serves as a good approximation for $\mathbb{E}[\hat{Q}_\Sigma^\lambda(\infty)]$, which suggests approximating $\min_{(n^\lambda, n_F^\lambda)\in\Omega^\lambda(\theta)}(\Pi^\lambda(n^\lambda,n_F^\lambda) - 2cR^\lambda)/\sqrt{\lambda}$ by the following optimization problem:

$$\begin{aligned} \min_{(\beta,\beta_F)\in\hat{\Omega}(\theta)} \hat{V}_p(\beta,\beta_F) :=& 2c\beta/\sqrt{\mu} + c_F\beta_F/\sqrt{\mu} \\ &+ (h+a\theta)\mathbb{E}\left[\left(\hat{X}_1(\infty;\beta,\beta_F)^+ + \hat{X}_2(\infty;\beta,\beta_F)^+ - \beta_F/\sqrt{\mu}\right)^+\right], \end{aligned} \tag{8}$$

where, if $\theta=0$,

$$\hat{\Omega}(0) := \{(\beta,\beta_F) : \beta\in\mathbb{R}, \beta_F \geq 0, 2\beta\mu + \beta_F\mu_F > 0\},$$

and, if $\theta>0$,

$$\hat{\Omega}(\theta) := \{(\beta,\beta_F) : \beta\in\mathbb{R}, \beta_F \geq 0\}.$$

We do not have a closed-form expression for $\mathbb{E}\left[\left(\hat{X}_1(\infty;\beta,\beta_F)^+ + \hat{X}_2(\infty;\beta,\beta_F)^+ - \beta_F/\sqrt{\mu}\right)^+\right]$. Thus, (8) can only be solved numerically. In Figure 2, we plot $\hat{V}_p(\beta,\beta_F)$ for different values of β and β_F. We observe that $\hat{V}_p(\beta,\beta_F)$ is convex and is minimized at $(0.5,0.5)$ in this example.

We next characterize the optimal staffing rule by rigorously drawing the connection between the solution of the optimal staffing problem (7) and the diffusion optimization problem (8). Due to the lack of an analytical solution for $\hat{V}_p(\beta,\beta_F)$, we impose the following technical assumption.

ASSUMPTION 1. *The set* $\arg\min_{(\beta,\beta_F)\in\hat{\Omega}(\theta)} \hat{V}_p(\beta,\beta_F)$ *is non-empty and finite.*

THEOREM 3. *For $\theta \leq \mu_F \leq \mu$, under Assumption 1, a sequence of staffing policies (n^λ, n_F^λ) is asymptotically optimal if and only if the following two conditions hold:*

1. $n^\lambda = R^\lambda + \beta^\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$
2. $n_F^\lambda = \beta_F^\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$

where $(\beta^\lambda, \beta_F^\lambda) \in \arg\min_{(\beta,\beta_F)\in\hat{\Omega}(\theta)} \hat{V}_p(\beta,\beta_F)$.

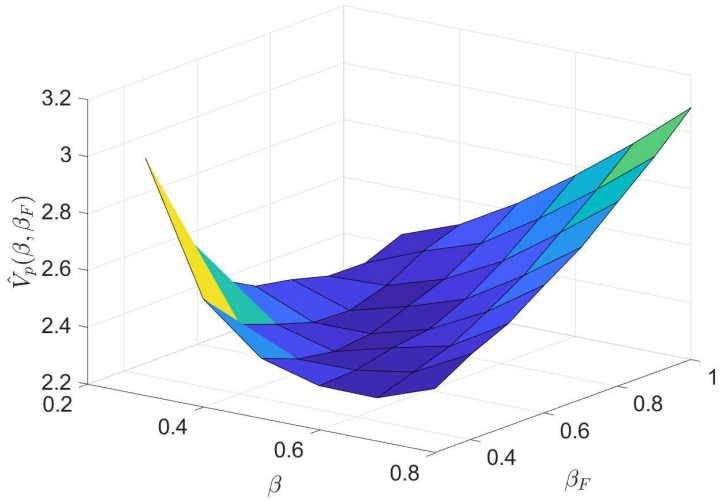

Figure 2 $\hat{V}_p(\beta, \beta_F)$ **as a function of** β **and** β_F**.** ($\mu = 1, \mu_F = 0.85, \theta = 0, c = 1, c_F = 1.4, h = 1$)

REMARK 1. If $\hat{V}_p(\beta, \beta_F)$ has a unique minimizer (β^*, β_F^*), then the asymptotically optimal staffing levels satisfy $n^\lambda = R^\lambda + \beta^* \sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ and $n_F^\lambda = \beta_F^* \sqrt{R^\lambda} + o(\sqrt{R^\lambda})$.

To illustrate why $n_F^\lambda = O(\sqrt{\lambda})$ is necessary for asymptotic optimality, we plot $\mathbb{E}[Q_\Sigma^\lambda(\infty; (60 - n_F^\lambda)/2, n_F^\lambda)]$ as a function of n_F^λ in Figure 3. We set $\lambda = 25$ and test two different scenarios for the service rate when $\theta = 0$. In the left plot, $\mu = \mu_F = 1$. In the right plot, $\mu = 1$ while $\mu_F = 0.85$. The stationary queue lengths are estimated through simulation. The simulation errors (estimated using the batch means method) are less than 0.01 and hence omitted. When $\mu = \mu_F$ (left plot in Figure 3), we observe that increasing n_F^λ beyond $2\sqrt{\lambda} = 10$ has almost no effect on the stationary total queue length. In this case, if $c < c_F$, the staffing cost increases linearly with n_F^λ while the holding cost does not decrease much as n_F^λ increases beyond 10. Thus, the optimal n_F^λ cannot be too large. When $\mu > \mu_F$ (right plot in Figure 3), the stationary total queue length is not monotone in n_F^λ. The minimum is achieved at a relatively small value of n_F^λ, i.e., $n_F^\lambda = 6$. Therefore, the optimal n_F^λ cannot be too large in this case as well.

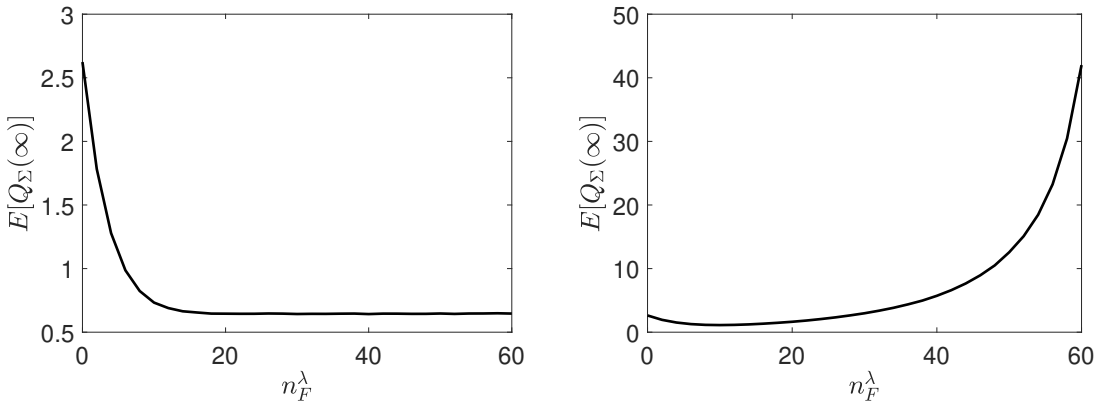

Figure 3 $\mathbb{E}[Q_\Sigma(\infty;(60-n_F^\lambda)/2,n_F^\lambda)]$ **as a function of** n_F^λ**. Left:** $\mu=\mu_F=1$**; Right** $\mu=1$**,** $\mu_F=0.85$**.** $(\lambda=25,\theta=0)$

We conclude this section with some sensitivity analysis on β^* and β_F^*. Let $h=c=1$, $\mu=1$, and $\theta=0$. Note that setting $c=1$ and $\mu=1$ is without loss of generality as it is equivalent to choosing units for cost and time. We first test how (β^*,β_F^*) varies with c_F, when $\mu_F=0.85$. Table 1 shows one such experiment. We observe that β^* is increasing in c_F while β_F^* is decreasing in c_F. When c_F is large, i.e., $c_F\geq 1.6$, $\beta_F^*=0$, suggesting it becomes too expensive to use the flexible servers then.

c_F	1	1.2	1.4	1.6	1.8
β^*	-0.2	0.2	0.5	0.9	0.9
β_F^*	1.9	1.1	0.5	0	0

Table 1 **Sensitivity of** (β^*,β_F^*) **with respect to** c_F

We next test how (β^*,β_F^*) varies with μ_F, when $c_F=1.4$. Table 2 shows one such experiment. We observe that β^* is decreasing in μ_F while β_F^* is increasing in μ_F. For very small values of μ_F, i.e., $\mu_F\leq 0.55$, flexible servers are too inefficient to be staffed.

μ_F	0.55	0.65	0.75	0.85	0.95
β^*	0.8	0.8	0.7	0.5	0.4
β_F^*	0	0.1	0.2	0.5	0.6

Table 2 **How optimal** (β,β_F) **varies with** μ_F

4. The Case with Demand Uncertainty

In this section, we study the joint staffing and scheduling optimization problem (2) with random arrival rates. We assume $\Lambda_i=p_i\lambda+\lambda^{\alpha_i}Y_i$, where $p_i>0$, $1/2<\alpha_i\leq 1$, and Y_i is a random variable with $\mathbb{E}[Y_i]=0$ and $\mathrm{Var}(Y_i)=\sigma_i^2<\infty$. As Λ_i is an arrival rate, we assume $\Lambda_i\geq 0$ with probability 1. For example, when $\alpha_i=1$, we assume $Y_i\geq -p_i$ with probability 1. For ease of exposition, we also assume Y_i's are continuous random variables with strictly increasing marginal cdf on their domains

of definition. We allow Y_1 and Y_2 to be dependent and denote by g their joint density. Without loss of generality, we assume $\alpha_1 \geq \alpha_2$. For the analysis in this section, we also require $\theta > 0$ to ensure system stability regardless of the realized arrival rates.

We next make some comments about our modeling assumptions for this section. First, we allow quite some asymmetry between the two classes. In particular, p_i's, α_i's, and the marginal distribution of Y_i's can be different for the two classes. This implies that the optimal n_1 and n_2 might be different. Second, the mean of Λ_i is of order λ while the standard deviation of Λ_i is of order λ^{α_i}. For queues with deterministic arrival rate λ, our analysis in Section 3 reveals that the stochastic fluctuation of the system is of order $\lambda^{1/2}$. In this section, we are interested in the case where $\alpha_i > 1/2$, so that the demand uncertainty is of a larger order of magnitude than the stochastic fluctuation of the system.

We start by providing an overview of how we address the two challenges listed Section 2 to derive the optimal scheduling and staffing rules jointly. When facing demand uncertainty, solving (2) analytically is more challenging than the case with deterministic arrival rates. This is because we now face two sources of randomness: One is the parameter uncertainty, the other is the stochasticity of the queue, i.e., random interarrival, service, and patience times. In this section, we again take a heavy-traffic asymptotic approach where we send the arrival rate parameter λ to infinity and quantify how the optimal staffing rule scales with λ. Under the assumption that $\alpha_i > 1/2$, we employ a stochastic-fluid approximation where we suppress the stochastic fluctuation of the queues and focus on parameter uncertainty only (Harrison and Zeevi 2005). In our setting, the stochastic-fluid optimal staffing problem is a special case of the single-period multi-product inventory problem with demand substitution (Netessine and Rudi 2003). Based on the stochastic-fluid staffing solution, it also becomes easier to develop good scheduling policies. We then show that the staffing and scheduling rules derived from the stochastic fluid problem achieve an $O(\lambda^{1-\alpha_2})$ optimality gap.

The key intuition behind our development is that when parameter uncertainty dominates system stochasticity, optimally hedging against parameter uncertainty is more important. Indeed, with high probability, the system with realized arrival rate is no longer in the QED regime. In these cases, any "fluid-optimal" scheduling policy is "good enough". We will make these intuitions more precise in the subsequent development.

4.1. Stochastic-Fluid Optimization Problem

For our model, the rate of customer abandonment can be expressed as

$$\theta \mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F)].$$

By rate conservation, the rate of customer abandonment can also be approximated by

$$\mathbb{E}\left[((\Lambda_1 - n_1\mu)^+ + (\Lambda_2 - n_2\mu)^+ - n_F\mu_F)^+\right].$$

Thus, we can approximate the steady-state queue length by

$$\frac{1}{\theta}\mathbb{E}\left[((\Lambda_1 - n_1\mu)^+ + (\Lambda_2 - n_2\mu)^+ - n_F\mu_F)^+\right].$$

This allows us to approximate (2) by the following stochastic-fluid optimization problem

$$\begin{aligned}\min_{\tilde{n}_1\geq 0,\tilde{n}_2\geq 0,\tilde{n}_F\geq 0} \tilde{\Pi}(\tilde{n}_1,\tilde{n}_2,\tilde{n}_F) :=& c(\tilde{n}_1+\tilde{n}_2) + c_F\tilde{n}_F \\ &+ (h/\theta + a)\mathbb{E}\left[((\Lambda_1 - \tilde{n}_1\mu)^+ + (\Lambda_2 - \tilde{n}_2\mu)^+ - \tilde{n}_F\mu_F)^+\right].\end{aligned} \tag{9}$$

For (9), we relax the integer requirement on n_1, n_2, n_F and only require them to be non-negative. We denote its optimal solution as $(\tilde{n}_1^*, \tilde{n}_2^*, \tilde{n}_F^*)$ and the optimal value as $\tilde{\Pi}^*$.

The optimization problem (9) can be viewed as a special case of the single-period multi-product inventory management problem with demand substitution, i.e., demand Λ_i is best met by dedicated resources, but may also be met by flexible resources if there is a shortfall of dedicated resources. This class of inventory management problems has been studied in the literature in much more general forms (Ernst and Kouvelis 1999, Rajaram and Tang 2001, Van Mieghem and Rudi 2002, Netessine and Rudi 2003, Bassamboo et al. 2010a). Restricting it to our special setting allows us to derive more analytical insights.

To simplify the notation, we define $c_P := h/\theta + a$, i.e., the performance cost. Let q_i denote the solution of the following equation

$$\mathbb{P}(Y_i > q_i) = \frac{c}{c_P\mu}.$$

We first study the case where $\alpha_1 = \alpha_2 = \alpha$. If $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) > \frac{c_F}{c_P\mu_F}$, let $r_1, r_2 \in \mathbb{R}$, and $r_F > 0$ denote the solution to the following system of equations:

$$\begin{aligned}\mathbb{P}(Y_1 > r_1, Y_1 - r_1 + (Y_2 - r_2)^+ > r_F) &= \frac{c}{c_P\mu}, \\ \mathbb{P}((Y_1 - r_1)^+ + (Y_2 - r_2)^+ > r_F) &= \frac{c_F}{c_P\mu_F}.\end{aligned}$$

The next lemma characterizes the optimal solution to (9) when $\alpha_1 = \alpha_2$.

LEMMA 4. *Suppose* $\alpha_1 = \alpha_2 = \alpha$.
If $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) \leq \frac{c_F}{c_P\mu_F}$, $\tilde{n}_i^* = (p_i\lambda + q_i\lambda^\alpha)/\mu$ *for* $i = 1, 2$, *and* $\tilde{n}_F^* = 0$.
If $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) > \frac{c_F}{c_P\mu_F}$, $\tilde{n}_i^* = (p_i\lambda + r_i\lambda^\alpha)/\mu$ *for* $i = 1, 2$, *and* $\tilde{n}_F^* = r_F\lambda^\alpha/\mu_F$.

Lemma 4 reveals that the optimal solution to (9) has a very neat structure. The optimal number of dedicated servers involves a baseline level to meet the mean demand, $p_i\lambda/\mu$, and an uncertainty hedging of order λ^α. We also note that the size of the flexible pool is $O(\lambda^\alpha)$, indicating that the flexible pool is mostly used for uncertainty hedging.

We next consider the case where $\alpha_1 > \alpha_2$. In this case, we do not have explicit expressions for $\tilde{n}_i^*$'s and $\tilde{n}_F^*$ as in Lemma 4. However, (9) can still be solved numerically very efficiently, as it

is a convex optimization problem. In addition, we can derive structural insights into the optimal staffing levels. When $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) > \frac{c_F}{c_P \mu_F}$, define $l \in \mathbb{R}, l_F > 0$ to be the solution to the following system of equations:

$$\mathbb{P}(Y_2 > l + l_F \text{ or } \{Y_1 > q_1, Y_2 > l\}) = \frac{c}{c_P \mu},$$
$$\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > l + l_F) = \frac{c_F}{c_P \mu_F}.$$

The next lemma characterizes the optimal solution to (9) when $\alpha_1 > \alpha_2$.

LEMMA 5. *Suppose* $\alpha_1 > \alpha_2$.
If $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) \le \frac{c_F}{c_P \mu_F}$, $\tilde{n}_i^* = (p_i \lambda + q_i \lambda^{\alpha_i})/\mu + o(\lambda^{\alpha_i})$ *for* $i = 1, 2$ *and* $\tilde{n}_F^* = o(\lambda^{\alpha_2})$.
If $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) > \frac{c_F}{c_P \mu_F}$, $\tilde{n}_1^* = (p_1 \lambda + q_1 \lambda^{\alpha_1})/\mu + o(\lambda^{\alpha_1})$, $\tilde{n}_2^* = (p_2 \lambda + l \lambda^{\alpha_2})/\mu + o(\lambda^{\alpha_2})$ *and* $\tilde{n}_F^* = l_F \lambda^{\alpha_2}/\mu_F + o(\lambda^{\alpha_2})$.

We note from Lemma 5 that the optimal size of the dedicated pool again contains a baseline level to meet the mean demand and an uncertainty hedging. The size of the flexible pool is $O(\lambda^{\alpha_2})$. As $\alpha_2 < \alpha_1$, the hedging functionality of the flexible pool is targeted for the less uncertain class.

We conduct numerical sensitivity analysis. For illustration, consider $p_1 = p_2 = 1$, $\alpha_1 = \alpha_2 = \alpha$, $Y_1 = Z_1$, and $Y_2 = \rho Z_1 + \sqrt{1 - \rho^2} Z_2$, where Z_1 and Z_2 are independent standard Normal random variables. In this case, $\text{Cor}(Y_1, Y_2) = \rho$. Due to the symmetry of the two classes, we have $q_1^* = q_2^* := q^*$, $\tilde{n}_i^* = \lambda/\mu + q^* \lambda^{\alpha}/\mu$, and $\tilde{n}_F^* = q_F^* \lambda^{\alpha}/\mu_F$. Figures 4 and 5 show how q^* and q_F^* vary with ρ for different values of c_F or μ_F. We note that as ρ increases, q_F^* decreases while q^* increases. This is because when the demand of the two classes are highly positively correlated, there is not much room for load-balancing. We also observe in Figure 4 that for a fixed value of ρ, the higher the cost of flexible servers, the smaller the value of q_F^*. Similarly, the less efficient the flexible servers, the smaller the value of q_F^* (see Figure 5).

4.2. Asymptotically Optimal Staffing and Scheduling Rules

In this section, we quantify the quality of the staffing rule derived from the stochastic-fluid approximation (9). We also develop a corresponding scheduling rule.

Consider a sequence of systems indexed by λ. The superscript λ is used to denote the quantities related to the λ-th system. For example, $\tilde{n}^{\lambda,*}, \tilde{n}_F^{\lambda,*}$ is the stochastic-fluid optimal solution when $\Lambda_i = p_i \lambda + \lambda^{\alpha_i} Y_i$ for Class i, $i = 1, 2$.

Our proposed staffing rule for the λ-th system is $(\lceil \tilde{n}_1^{\lambda,*} \rceil, \lceil \tilde{n}_2^{\lambda,*} \rceil, \lfloor \tilde{n}_F^{\lambda,*} \rfloor)$. We next introduce a scheduling policy. Given a realization of the arrival rate $\Lambda = \gamma := (\gamma_1, \gamma_2)$, let $\delta(\gamma) \in [0, 1]$ be a solution of

$$((\gamma_1 - n_1^{\lambda}\mu)^+ + (\gamma_2 - n_2^{\lambda}\mu)^+ - n_F^{\lambda}\mu_F)^+ = (\gamma_1 - n_1^{\lambda}\mu - \delta n_F^{\lambda}\mu_F)^+ + (\gamma_2 - n_2^{\lambda}\mu - (1-\delta) n_F^{\lambda}\mu_F)^+. \quad (10)$$

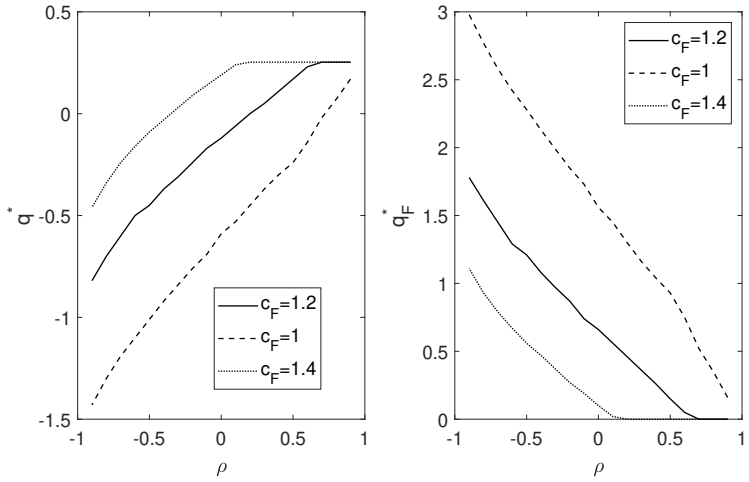

Figure 4 **How q^* and q_F^* vary with ρ when $\mu_F = 0.9$ and $c_F \in \{1, 1.2, 1.4\}$**

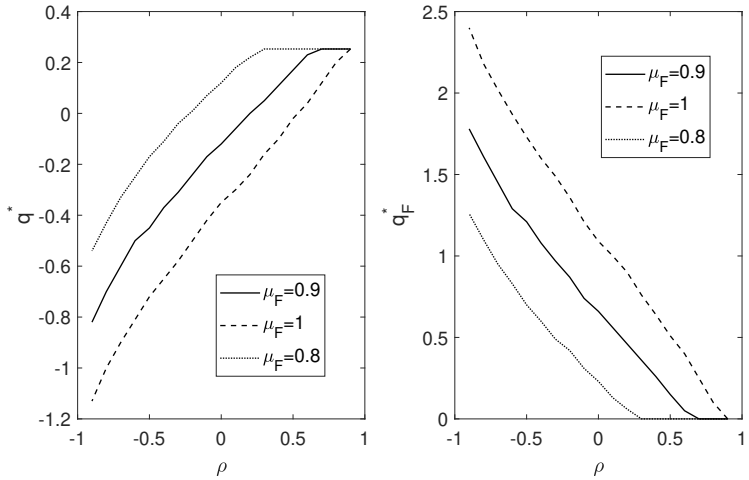

Figure 5 **How q^* and q_F^* vary with ρ when $c_F = 1.2$ and $\mu_F \in \{0.8, 0.9, 1\}$**

Note that the solution to (10) may not be unique. When (10) has multiple optimal solutions, we can set $\delta(\gamma)$ to be any one of them. For a fixed $\delta(\gamma)$, the scheduling policy $\tilde{\nu}^\lambda$ allocates $\lfloor \delta(\gamma) n_F^\lambda \rfloor$ flexible servers to Class 1 and the remaining $\lceil (1-\delta(\gamma)) n_F^\lambda \rceil$ flexible servers to Class 2. When assigning customers to servers, the dedicated servers are prioritized over the flexible servers. That is, upon each realization of the arrival rates $\Lambda = \gamma$, the policy $\tilde{\nu}^\lambda$ turns the M-model into two independent inverted-V models. For each inverted-V model, we follow the fastest-server-first policy.

To quantify the optimality gap of the stochastic-fluid based policies, we first quantify the difference between Π^λ defined in (2) and $\tilde{\Pi}^\lambda$ defined in (9).

LEMMA 6. *For $\theta > 0$, $\alpha_1 \geq \alpha_2 > 1/2$, and $n_i^\lambda + n_F^\lambda = \Theta(\lambda)$, for any scheduling policy ν^λ,*

$$\tilde{\Pi}^\lambda(n_1^\lambda, n_2^\lambda, n_F^\lambda) \leq \Pi^\lambda(n_1^\lambda, n_2^\lambda, n_F^\lambda; \nu^\lambda).$$

For policy $\tilde{\nu}^\lambda$, we also have

$$\Pi^\lambda(n_1^\lambda, n_2^\lambda, n_F^\lambda; \tilde{\nu}^\lambda) \le \tilde{\Pi}^\lambda(n_1^\lambda, n_2^\lambda, n_F^\lambda) + O(\lambda^{1-\alpha_2}).$$

Based on Lemma 6, we have the following optimality gap quantification.

THEOREM 4. *For $\alpha_1 \ge \alpha_2 > 1/2$ and $\theta > 0$,*

$$\Pi^\lambda(\lceil \tilde{n}_1^{\lambda,*} \rceil, \lceil \tilde{n}_2^{\lambda,*} \rceil, \lfloor \tilde{n}_F^{\lambda,*} \rfloor; \tilde{\nu}^\lambda) = \Pi^{\lambda,*} + O(\lambda^{1-\alpha_2}).$$

Theorem 4 indicates that the staffing rule based on the stochastic-fluid approximation together with the scheduling policy $\tilde{v}^\lambda$ is asymptotically optimal, i.e., it achieves an $o(\sqrt{\lambda})$ optimality gap. In addition, we note from Theorem 4 that the optimality gap of our proposed staffing and scheduling rule is determined by the smaller α_i. This is expected as the size of flexible pool, $\tilde{n}_F^{\lambda,*}$, is determined by the smaller α_i (Lemma 5).

When $\tilde{n}_F^{\lambda,*} > 0$, comparing to the case where no flexible server is available, i.e., $n_F^\lambda \equiv 0$, we have

$$\min_{\tilde{n}_1^\lambda \ge 0, \tilde{n}_2^\lambda \ge 0} \tilde{\Pi}^\lambda(\tilde{n}_1^\lambda, \tilde{n}_2^\lambda, 0) = \tilde{\Pi}^{\lambda,*} + \Theta(\lambda^{\alpha_2}).$$

Then, Theorem 4 indicates that in this case, having access to flexible servers can lead to an $\Theta(\lambda^{\alpha_2})$ cost-saving. This is different from the case without demand uncertainty (Section 3), where flexible servers only lead to an $\Theta(\sqrt{\lambda})$ cost-saving.

We conclude this section with some remarks about good scheduling policies when facing a high level of uncertainty in demand. Our proposed scheduling policy $\tilde{\nu}^\lambda$ is quite simple but is sufficient for achieving a good performance. This is because for most realized arrival rates, the system is no longer in the critically loaded regime. Thus, any fluid-optimal scheduling policy will achieve a similar optimality gap. To reinforce this point, consider another scheduling policy $\tilde{\nu}_R^\lambda$ defined as follows. Similar to $\tilde{v}^\lambda$, for a realized arrival rate $\Lambda = \gamma$, we allocate $\lfloor \delta(\gamma) n_F^\lambda \rfloor$ flexible servers to Class 1 and the remaining $\lceil (1-\delta(\gamma)) n_F^\lambda \rceil$ flexible servers to Class 2. However, unlike $\tilde{v}^\lambda$, when assigning customers to servers, $\tilde{\nu}_R^\lambda$ prioritizes the slower flexible servers over the faster dedicated servers, i.e., the policy turns the M-model into two independent slowest-server-first inverted-V models. Following similar lines of argument as the proof of Theorem 4, one can show that this policy also achieves an $O(\lambda^{1-\alpha_2})$ optimality gap.

Although the scheduling policy $\tilde{\nu}^\lambda$ is asymptotically optimal, it can be improved further. We next introduce a simple improved version of $\tilde{\nu}^\lambda$, which we denote as $\tilde{\nu}_I^\lambda$. For a realized arrival rate, $\Lambda = \gamma$, we again follow the same server allocation rule as $\tilde{\nu}^\lambda$, and when assigning customers to servers, we prioritize the dedicated servers. However, under $\tilde{\nu}_I^\lambda$, the $\lfloor \delta(\gamma) n_F^\lambda \rfloor$ flexible servers 'assigned' to Class 1 only give priority to Class 1 customers, and the remaining $\lceil (1-\delta(\gamma)) n_F^\lambda \rceil$

flexible servers 'assigned' to Class 2 only give priority to Class 2. For example, when one of the $\lfloor \delta(\gamma) n_F^\lambda \rfloor$ flexible servers assigned to Class 1 becomes available and there is no Class 1 customer waiting, the flexible server can then serve a Class 2 customer waiting in queue. It is easy to see that $\tilde{\nu}_I^\lambda$ is also asymptotically optimal. Indeed, following similar coupling arguments as those in Appendix B, one can show that $\tilde{\nu}_I^\lambda$ leads to a smaller steady-state average queue length than $\tilde{\nu}^\lambda$ for any given arrival rate realization.

In Section 5.2, we conduct some numerical experiments demonstrating the pre-limit performance of $\tilde{\nu}^\lambda$, $\tilde{\nu}_R^\lambda$, and $\tilde{\nu}_I^\lambda$ introduced above (see Table 6).

5. Numerical Experiments

In this section, we demonstrate the pre-limit performance of our proposed staffing and scheduling rules using simulation experiments.

5.1. Deterministic Arrival Rates

Based on the result in Theorem 3, we set the staffing levels

$$(\hat{n}^\lambda, \hat{n}_F^\lambda) = (\lceil R^\lambda + \beta^* \sqrt{R^\lambda} \rceil, \lfloor \beta_F^* \sqrt{R^\lambda} \rfloor). \tag{11}$$

In the first numerical experiment, we consider the case with no abandonment, i.e., $\theta = 0$. We set $h = c = 1$, $\mu = 1, \mu_F = 0.85$, and vary the values of λ and c_F. In Table 3, we compare the staffing rule (11) to the optimal staffing levels $(n^{\lambda,*}, n_F^{\lambda,*})$ (solved by exhaustive search using simulation). We observe that the staffing levels suggested by the diffusion optimization problem is almost identical to the optimal staffing levels. In most cases, the difference between the two is less than or equal to 1, and the largest difference is 3. Table 3 also reports $\Pi^{\lambda,*}$ and the optimality gaps, i.e., $\Pi^\lambda(\hat{n}^\lambda, \hat{n}_F^\lambda) - \Pi^{\lambda,*}$. As expected, the optimality gaps are extremely small, even for systems as small as $\lambda = 25$.

Table 4 reports the results of a similar experiment when there is abandonment. In this example, we set $h = a = 8$, $c = 1$, $\mu = 1$, and $\mu_F = \theta = 0.85$. We observe again that the prescription (11) works very well for all system sizes. Specifically, the optimality gap across all cases are less than 0.1.

5.2. Random Arrival Rates

In this section, we study the pre-limit performance of the stochastic-fluid based staffing and scheduling rules when the arrival rates are random. For simplicity of illustration, we consider a symmetric system where $p_1 = p_2 = 1$ and $\alpha_1 = \alpha_2 = \alpha$. In this case, $\tilde{n}_1^{\lambda,*} = \tilde{n}_2^{\lambda,*} := \tilde{n}^{\lambda,*}$. Based on the result in Theorem 4, we set the staffing level

$$(\hat{n}^\lambda, \hat{n}_F^\lambda) = (\lceil \tilde{n}^{\lambda,*} \rceil, \lfloor \tilde{n}_F^{\lambda,*} \rfloor), \tag{12}$$

c_F	$(\hat{n}^\lambda, \hat{n}_F^\lambda)$	$(n^{\lambda,*}, n_F^{\lambda,*})$	$\Pi^{\lambda,*}$	Gap
	$\lambda = 25$			
1	(27,10)	(26,11)	65.91	0.17
1.2	(28,7)	(28,7)	67.76	0
1.4	(29,5)	(30,4)	69.12	0.05
	$\lambda = 100$			
1	(103,20)	(102,22)	230.94	0.08
1.2	(106,15)	(106,15)	234.79	0
1.4	(108,11)	(108,10)	237.27	0.19
	$\lambda = 400$			
1	(406,40)	(405,42)	861.42	0.16
1.2	(412,30)	(413,27)	868.71	0.25
1.4	(416,22)	(416,21)	873.85	0.01

Table 3 **Performance of $(\hat{n}^\lambda, \hat{n}_F^\lambda)$ for systems with different scales, λ's. ($\mu = 1, \mu_F = 0.85, \theta = 0, h = 8, c = 1$)**

c_F	$(\hat{n}^\lambda, \hat{n}_F^\lambda)$	$(n^{\lambda,*}, n_F^{\lambda,*})$	$\Pi^{\lambda,*}$	Gap
	$\lambda = 25$			
1	(26,11)	(25,13)	65.95	0.02
1.2	(28,7)	(28,7)	67.94	0
1.4	(29,5)	(29,5)	69.26	0
	$\lambda = 100$			
1	(101,23)	(101,23)	231.29	0
1.2	(105,15)	(105,15)	235.12	0
1.4	(107,11)	(108,10)	237.72	0.03
	$\lambda = 400$			
1	(402,46)	(402,46)	862.01	0
1.2	(410,30)	(410,30)	869.50	0
1.4	(414,22)	(415,21)	874.60	0.05

Table 4 **Performance of $(\hat{n}^\lambda, \hat{n}_F^\lambda)$ for systems with different scales, λ's. ($\mu = 1, \mu_F = 0.85 = \theta, c = 1, h = a = 8$)**

where $\hat{n}_1^\lambda = \hat{n}_2^\lambda := \hat{n}^\lambda$.

Let $c = 1, c_F = 1.2, h = a = 8, \mu = 1, \mu_F = 0.9$, and $\theta = 0.5$. In addition, let $Y_1 = Z_1, Y_2 = \rho Z_1 + \sqrt{1-\rho^2} Z_2$, where Z_1 and Z_2 are independent standard Normal random variables, and $\rho = 0.5$. In this case, $r_1 = r_2 = 1.22$ and $r_F = 0.70$ in Lemma 4. Thus,

$$\hat{n}^\lambda = \lceil \lambda + 1.22\lambda^\alpha \rceil \text{ and } \hat{n}_F^\lambda = \lfloor 0.70\lambda^\alpha / 0.9 \rfloor.$$

Next, to define $\tilde{\nu}^\lambda$, we need to specify $\delta(\gamma)$ (the results of Section 4 are valid for any choice that satisfies (10)). In our experiments we choose $\delta(\gamma)$ to strike a balance between the capacity of the two classes, i.e. $\gamma_1 - \hat{n}_1^\lambda \mu - \delta \hat{n}_F^\lambda \mu_F$ versus $\gamma_2 - \hat{n}_2^\lambda \mu - (1-\delta)\hat{n}_F^\lambda \mu_F$. For example, if $\hat{n}^\lambda = 34$, $\hat{n}_F^\lambda = 5$ and $\gamma = (33, 35)$, $\tilde{\nu}^\lambda$ allocates 1 flexible server to Class 1 and 4 to Class 2.

Table 5 reports the performance of our proposed staffing and scheduling rules for different values of α and λ. The optimality gap in Theorem 4 cannot be computed numerically, because the optimal scheduling policy for (2) is unknown. However, because by Lemma 6, the optimality gap satisfies

$$\Pi^\lambda(\hat{n}^\lambda, \hat{n}_F^\lambda; \tilde{\nu}^\lambda) - \Pi^{\lambda,*} \leq \Pi^\lambda(\hat{n}^\lambda, \hat{n}_F^\lambda; \tilde{\nu}^\lambda) - \tilde{\Pi}^{\lambda,*},$$

where $\tilde{\Pi}^{\lambda,*}$ is the optimal value of (9), we use $\Pi^{\lambda}(\hat{n}^{\lambda},\hat{n}_F^{\lambda};\tilde{\nu}^{\lambda})-\tilde{\Pi}^{\lambda,*}$ as an approximation of the optimality gap. Note that this approximation is larger than the actual optimality gap. We refer to it as "Approx. Gap" in Table 5.

We observe that for a fixed value of λ, the gap decreases as α increases. For example, when $\lambda=25$, as α increases from 0.6 to 1, the gap decreases from 9.6 to 4.4. This agrees with the results in Theorem 4, i.e., the optimality gap is $O(\lambda^{1-\alpha})$, which decreases as α increases. For a fixed value of α, the ratio between the gap and $\tilde{\Pi}^{\lambda,*}$ decreases as λ increase. For example, when $\alpha=0.8$, as λ increases from 25 to 100, the gap decreases from 5.5% of $\tilde{\Pi}^{\lambda,*}$ to 2% of $\tilde{\Pi}^{\lambda,*}$.

	$\lambda=25$		$\lambda=50$		$\lambda=100$		$\lambda=200$	
α	$\tilde{\Pi}^{\lambda,*}$	Approx. Gap	$\tilde{\Pi}^{\lambda,*}$	Approx. Gap	$\tilde{\Pi}^{\lambda,*}$	Approx. Gap	$\tilde{\Pi}^{\lambda,*}$	Approx. Gap
0.6	78.4	9.6	143.0	12.2	265.2	14.7	498.9	18.6
0.8	104.1	5.7	194.1	6.7	363.9	7.2	685.3	7.9
1	152.9	4.4	305.8	4.6	611.7	4.5	1223.3	4.2

Table 5 **Performance of $(\hat{n}^{\lambda},\hat{n}_F^{\lambda};\tilde{\nu}^{\lambda})$ for systems with different values of λ and α.**
($c=1, c_F=1.2, h=a=8, \mu=1, \mu_F=0.9$, $\theta=0.5$, $\rho=0.5$)

We next compare the pre-limit performance of three asymptotically optimal scheduling policies: $\tilde{\nu}^{\lambda}$, $\tilde{\nu}_R^{\lambda}$, and $\tilde{\nu}_I^{\lambda}$ introduced in Section 4.2. We observe in Table 6 that

$$\Pi^{\lambda}(\hat{n}^{\lambda},\hat{n}_F^{\lambda};\tilde{\nu}_I^{\lambda})<\Pi^{\lambda}(\hat{n}^{\lambda},\hat{n}_F^{\lambda};\tilde{\nu}^{\lambda})<\Pi^{\lambda}(\hat{n}^{\lambda},\hat{n}_F^{\lambda};\tilde{\nu}_R^{\lambda}).$$

The performance gaps between $\tilde{\nu}_R^{\lambda}$ and $\tilde{\nu}_I^{\lambda}$ are small in all cases. This demonstrates that using as crude a policy as $\tilde{\nu}_R^{\lambda}$ still leads to good performances.

	$\lambda=25$			$\lambda=50$			$\lambda=100$			$\lambda=200$		
α	$\tilde{\nu}_I^{\lambda}$	$\tilde{\nu}^{\lambda}$	$\tilde{\nu}_R^{\lambda}$	$\tilde{\nu}_I^{\lambda}$	$\tilde{\nu}^{\lambda}$	$\tilde{\nu}_R^{\lambda}$	$\tilde{\nu}_I^{\lambda}$	$\tilde{\nu}^{\lambda}$	$\tilde{\nu}_R^{\lambda}$	$\tilde{\nu}_I^{\lambda}$	$\tilde{\nu}^{\lambda}$	$\tilde{\nu}_R^{\lambda}$
0.6	86.3	88.0	88.3	153.0	155.2	155.5	276.8	279.9	280.3	513.6	517.5	518.1
0.8	108.3	109.8	110.0	199.5	200.8	201.0	369.2	371.1	371.4	691.2	693.2	693.5
1	156.2	157.3	157.5	309.6	310.4	310.6	614.3	616.2	616.4	1226.6	1227.5	1227.7

Table 6 **The cost under other scheduling policies $\nu\in\{\tilde{\nu}_I^{\lambda},\tilde{\nu}^{\lambda},\tilde{\nu}_R^{\lambda}\}$ for different values of λ and α.**
($c=1, c_F=1.2, h=a=8, \mu=1, \mu_F=0.9$, $\theta=0.5$, $\rho=0.5$)

6. Concluding Remarks

In this paper, we study the joint optimal staffing and scheduling problem for a two-class queue with both dedicated and flexible servers. We quantify how the cost of flexibility affects the optimal size of the flexible pool. We conclude the paper with some remarks for future research.

Non-preemption For the deterministic arrival rate setting, our scheduling policy $\nu^{\lambda,*}$ is preemptive. For example, we allow a customer in service with the flexible pool to be transferred to the dedicated pool if a dedicated server becomes available. If we restrict ourselves to non-preemptive policies, one may be tempted to define a non-preemptive version of the policy, and prove that it performs asymptotically as well as the preemptive version in the many-server heavy-traffic regime. Unfortunately, this asymptotic result is unlikely to hold in our case. This is because the size of the flexible pool, $O(\sqrt{\lambda})$, is not large enough to cause instantaneous changes in X^λ in the limit, which indicates that the non-preemptive version of the policy may not be able to closely 'track' the preemptive policy (see Atar (2005) for a similar argument).

For the random arrival rate setting, our scheduling policy $\tilde{\nu}^\lambda$ is preemptive, but this is not needed to achieve the optimality gap in Theorem 4. Indeed, a simple coupling argument can show that a non-preemptive fastest-server-first scheduling policy outperforms the preemptive slowest-server-first scheduling policy $\tilde{\nu}^\lambda_R$, and the latter is still asymptotically optimal.

Multiple customer classes When there are k customer classes, servers can potentially have 2^k-1 different skill sets, i.e., each of the non-empty subsets of $\{1,\cdots,k\}$. In this case, we need to specify the optimal size of each potential server pool as well as the corresponding scheduling policy. As k increases, the number of possible system configurations can become very large, posing substantial analytical challenges.

When facing demand uncertainty, we can still approximate the optimal staffing problem with a multi-product inventory management problem with demand substitution. Bassamboo et al. (2010a) study such inventory networks when the 'staffing' costs are affine (or convex) in the degree of flexibility. Let $S_1, S_2 \subseteq \{1,\cdots,k\}$, and let n_{S_i} denote the number of servers with skill set S_i. We also write $|S_i|$ for the cardinality of set S_i. Bassamboo et al. (2010a) finds that if it is optimal to set $n_{S_1}, n_{S_2} > 0$ with $S_1 \subsetneq S_2$, then $|S_1| = |S_2| - 1$. This implies that if the optimal sizes of the dedicated pools are all positive, i.e., $n_{\{i\}} > 0$ for $i = 1, 2, \ldots, k$, then the only other server pools we need to consider are those with skill set $\{i, j\}$, $i \neq j$, $i, j = 1, \ldots, k$. This can help reduce the number of possible system configurations that one needs to consider.

Appendix A: Two important stochastic dominance results

In this section, we present two important stochastic dominance results that are useful for our subsequent analysis, e.g., the proofs of Theorem 1, Lemma 1, and Lemma 3. These results build on coupling arguments and can be of independent interest.

Let $\{Y(t) = (Y_1(t), Y_2(t)); t \geq 0\}$ and $\{\tilde{Y}(t) = (\tilde{Y}_1(t), \tilde{Y}_2(t)); t \geq 0\}$ be two positive recurrent birth-and-death processes. The birth (arrival) rates are λ for both Y_i and $\tilde{Y}_i$, $i = 1, 2$. Let $\zeta_i(y)$ be the death (departure) rate of Y_i when $Y(t) = y$. We also define $\zeta_\Sigma(y) = \zeta_1(y) + \zeta_2(y)$, $\zeta_M(y) = \zeta_1(y)1\{y_1 \geq y_2\} + \zeta_2(y)1\{y_1 < y_2\}$, and

$\zeta_m(y)=\zeta_1(y)1\{y_1\le y_2\}+\zeta_2(y)1\{y_1>y_2\}$. Similarly, let $\tilde{\zeta}_i(\tilde{y})$ be the death rate of $\tilde{Y}_i$, $i=1,2$, when $\tilde{Y}(t)=\tilde{y}$, $\tilde{\zeta}_\Sigma(\tilde{y})=\tilde{\zeta}_1(\tilde{y})+\tilde{\zeta}_2(\tilde{y})$, $\tilde{\zeta}_M(y)=\tilde{\zeta}_1(\tilde{y})1\{\tilde{y}_1\ge\tilde{y}_2\}+\tilde{\zeta}_2(\tilde{y})1\{\tilde{y}_1<\tilde{y}_2\}$, and $\tilde{\zeta}_m(y)=\tilde{\zeta}_1(\tilde{y})1\{\tilde{y}_1\le\tilde{y}_2\}+\tilde{\zeta}_2(\tilde{y})1\{\tilde{y}_1>\tilde{y}_2\}$.

The following two lemmas provide sufficient conditions to establish stochastic dominance between $Y(\infty)$ and $\tilde{Y}(\infty)$.

LEMMA 7. *For $\{Y(t);t\ge0\}$ and $\{\tilde{Y}(t);t\ge0\}$, suppose*

P1) $\zeta_\Sigma(y)\ge\tilde{\zeta}_\Sigma(\tilde{y})$ whenever $y_1+y_2=\tilde{y}_1+\tilde{y}_2$ and $y_1\vee y_2\le\tilde{y}_1\vee\tilde{y}_2$;

P2) $\zeta_M(y)\ge\tilde{\zeta}_M(\tilde{y})$ whenever $y_1\vee y_2=\tilde{y}_1\vee\tilde{y}_2$ and $y_1+y_2\le\tilde{y}_1+\tilde{y}_2$.

Then, $Y_1(\infty)+Y_2(\infty)\le_{st}\tilde{Y}_1(\infty)+\tilde{Y}_2(\infty)$ and $Y_1(\infty)\vee Y_2(\infty)\le_{st}\tilde{Y}_1(\infty)\vee\tilde{Y}_2(\infty)$.

Proof. We prove the lemma by constructing a coupling, under which

$$Y_1(t)+Y_2(t)\le\tilde{Y}_1(t)+\tilde{Y}_2(t)\text{ and }Y_1(t)\vee Y_2(t)\le\tilde{Y}_1(t)\vee\tilde{Y}_2(t)$$

for all $t\ge0$ path-by-path (Dong et al. 2015).

We start by introducing the coupling. Let $Y(0)=\tilde{Y}(0)=y_0$ for any fixed $y_0\in\mathbb{N}_0^2$. We denote the k-th potential transition time in both systems by t_k with $t_0=0$. In particular, for $Y(t_k)=y$ and $\tilde{Y}(t_k)=\tilde{y}$, let

$$\delta_M=\begin{cases}e_1 & y_1\ge y_2\\ e_2 & y_1<y_2\end{cases}\text{ and }\delta_m=\begin{cases}e_2 & y_1\ge y_2\\ e_1 & y_1<y_2\end{cases}.$$

Similarly let

$$\tilde{\delta}_M=\begin{cases}e_1 & \tilde{y}_1\ge\tilde{y}_2\\ e_2 & \tilde{y}_1<\tilde{y}_2\end{cases}\text{ and }\tilde{\delta}_m=\begin{cases}e_2 & \tilde{y}_1\ge\tilde{y}_2\\ e_1 & \tilde{y}_1<\tilde{y}_2\end{cases}.$$

We then generate $t_{k+1}-t_k$ from an exponential distribution with rate $\Lambda:=2\lambda+\zeta_\Sigma(y)\vee\tilde{\zeta}_\Sigma(\tilde{y})$. We also generate a random variable U uniformly distributed on $[0,1]$. We update the states of the two systems according to the following:

$$Y(t_{k+1})=Y(t_k)+\begin{cases}\delta_M & 0\le U\le\lambda/\Lambda\\ \delta_m & \lambda/\Lambda<U\le2\lambda/\Lambda\\ -\delta_M & 2\lambda/\Lambda<U\le(2\lambda+\zeta_M(y))/\Lambda\\ -\delta_m & (2\lambda+\zeta_M(y))/\Lambda<U\le(2\lambda+\zeta_\Sigma(y))/\Lambda\\ 0 & \text{Otherwise;}\end{cases}$$

and

$$\tilde{Y}(t_{k+1})=\tilde{Y}(t_k)+\begin{cases}\tilde{\delta}_M & 0\le U\le\lambda/\Lambda\\ \tilde{\delta}_m & \lambda/\Lambda<U\le2\lambda/\Lambda\\ -\tilde{\delta}_M & 2\lambda/\Lambda<U\le(2\lambda+\tilde{\zeta}_M(\tilde{y}))/\Lambda\\ -\tilde{\delta}_m & (2\lambda+\tilde{\zeta}_M(\tilde{y}))/\Lambda<U\le(2\lambda+\tilde{\zeta}_\Sigma(\tilde{y}))/\Lambda\\ 0 & \text{Otherwise.}\end{cases}$$

Now, let $S=\{k\in\mathbb{N}_0:Y_1(t_k)+Y_2(t_k)=\tilde{Y}_1(t_k)+\tilde{Y}_2(t_k)\}$, and let the elements of S, s_i, be ordered such that $0=s_0<s_1<\cdots$. We will prove by induction that

$$Y_1(t_k)\vee Y_2(t_k)\le\tilde{Y}_1(t_k)\vee\tilde{Y}_2(t_k)\text{ and }Y_1(t_k)+Y_2(t_k)\le\tilde{Y}_1(t_k)+\tilde{Y}_2(t_k)\text{ for }0\le k\le s_i\text{ for any }i\in\mathbb{N}.\quad(13)$$

For $i=0$, we have $Y_1(0)+Y_2(0)=\tilde{Y}_1(0)+\tilde{Y}_2(0)$ and $Y_1(0)\vee Y_2(0)=\tilde{Y}_1(0)\vee\tilde{Y}_2(0)$ by construction.

Suppose (13) **holds for some** i**,** $i \in \mathbb{N}_0$**.** We first note that for $k = s_i + 1$, if $s_i + 1 \in S$, we have $Y_1(t_k) + Y_2(t_k) = \tilde{Y}_1(t_k) + \tilde{Y}_2(t_k)$. If $s_i + 1 \notin S$, then by the coupling construction and P1, there must be a departure from Y but not $\tilde{Y}$. Consequently, $Y_1(t_k) + Y_2(t_k) < \tilde{Y}_1(t_k) + \tilde{Y}_2(t_k)$. This also implies that for $s_i + 1 < k < s_{i+1}$ (we set $s_{i+1} = \infty$ if s_i is the last element in S), $Y_1(t_k) + Y_2(t_k) < \tilde{Y}_1(t_k) + \tilde{Y}_2(t_k)$. We next note that for $s_i < k \le s_{i+1}$, by our coupling construction, if there is an arrival, it either joins the larger queue in both systems or the smaller queue in both systems. Thus, in this case $Y_1(t_k) \vee Y_2(t_k) \le \tilde{Y}_1(t_k) \vee \tilde{Y}_2(t_k)$. If there is a departure, then we further consider two cases.

Case 1. $Y_1(t_{k-1}) \vee Y_2(t_{k-1}) < \tilde{Y}_1(t_{k-1}) \vee \tilde{Y}_2(t_{k-1})$: since the difference between the two quantities changes by at most 1 at each epoch, we have $Y_1(t_k) \vee Y_2(t_k) \le \tilde{Y}_1(t_k) \vee \tilde{Y}_2(t_k)$.

Case 2. $Y_1(t_{k-1}) \vee Y_2(t_{k-1}) = \tilde{Y}_1(t_{k-1}) \vee \tilde{Y}_2(t_{k-1})$: by P2, if there is a departure from the larger queue in $\tilde{Y}$, there must be a departure from the larger queue in Y. Moreover, if $Y_1(t_{k-1}) = Y_2(t_{k-1})$, as $Y_1(t_{k-1}) + Y_2(t_{k-1}) \le \tilde{Y}_1(t_{k-1}) + \tilde{Y}_2(t_{k-1})$, we have $\tilde{Y}_1(t_{k-1}) = \tilde{Y}_2(t_{k-1})$. Thus, $Y_1(t_k) \vee Y_2(t_k) \le \tilde{Y}_1(t_k) \vee \tilde{Y}_2(t_k)$.

Above all, $Y_1(t) + Y_2(t) \le \tilde{Y}_1(t) + \tilde{Y}_2(t)$ and $Y_1(t) \vee Y_2(t) \le \tilde{Y}_1(t) \vee \tilde{Y}_2(t)$ for all $t \ge 0$ under our coupling construction. This further implies the stochastic dominance results for the stationary distributions. □

LEMMA 8. *For* $\{Y(t); t \ge 0\}$ *and* $\{\tilde{Y}(t); t \ge 0\}$*, suppose*

P1) $\zeta_\Sigma(y) \le \tilde{\zeta}_\Sigma(\tilde{y})$ *whenever* $y_1 + y_2 = \tilde{y}_1 + \tilde{y}_2$ *and* $y_1 \wedge y_2 \ge \tilde{y}_1 \wedge \tilde{y}_2$*;*

P2) $\zeta_m(y) \le \tilde{\zeta}_m(\tilde{y})$ *whenever* $y_1 \wedge y_2 = \tilde{y}_1 \wedge \tilde{y}_2$ *and* $y_1 + y_2 \ge \tilde{y}_1 + \tilde{y}_2$*.*

Then, $Y_1(\infty) + Y_2(\infty) \ge_{st} \tilde{Y}_1(\infty) + \tilde{Y}_2(\infty)$ *and* $Y_1(\infty) \wedge Y_2(\infty) \ge_{st} \tilde{Y}_1(\infty) \wedge \tilde{Y}_2(\infty)$*.*

Proof. The coupling construction follows a similar coupling idea as the proof of Lemma 7. We highlight the difference here for completeness.

Let $Y(0) = \tilde{Y}(0) = y_0$ for any fixed $y_0 \in \mathbb{N}_0^2$. We denote the k-th potential transition time in both systems by t_k with $t_0 = 0$. In particular, for $Y(t_k) = y$ and $\tilde{Y}(t_k) = \tilde{y}$, we generate $t_{k+1} - t_k$ from an exponential distribution with rate $\Lambda := 2\lambda + \zeta_\Sigma(y) \vee \tilde{\zeta}_\Sigma(\tilde{y})$. We also generate a random variable U uniformly distributed on $[0, 1]$ and update the states of the two systems according to the following:

$$Y(t_{k+1}) = Y(t_k) + \begin{cases} \delta_M & 0 \le U \le \lambda/\Lambda \\ \delta_m & \lambda/\Lambda < U \le 2\lambda/\Lambda \\ -\delta_m & 2\lambda/\Lambda < U \le (2\lambda + \zeta_m(y))/\Lambda \\ -\delta_M & (2\lambda + \zeta_m(y))/\Lambda < U \le (2\lambda + \zeta_\Sigma(y))/\Lambda \\ 0 & \text{Otherwise;} \end{cases}$$

and

$$\tilde{Y}(t_{k+1}) = \tilde{Y}(t_k) + \begin{cases} \tilde{\delta}_M & 0 \le U \le \lambda/\Lambda \\ \tilde{\delta}_m & \lambda/\Lambda < U \le 2\lambda/\Lambda \\ -\tilde{\delta}_m & 2\lambda/\Lambda < U \le (2\lambda + \tilde{\zeta}_m(\tilde{y}))/\Lambda \\ -\tilde{\delta}_M & (2\lambda + \tilde{\zeta}_m(\tilde{y}))/\Lambda < U \le (2\lambda + \tilde{\zeta}_\Sigma(\tilde{y}))/\Lambda \\ 0 & \text{Otherwise.} \end{cases}$$

We next prove by contradiction that

$$Y_1(t_k) \wedge Y_2(t_k) \ge \tilde{Y}_1(t_k) \wedge \tilde{Y}_2(t_k) \text{ and } Y_1(t_k) + Y_2(t_k) \ge \tilde{Y}_1(t_k) + \tilde{Y}_2(t_k) \text{ for all } k \ge 0. \tag{14}$$

Let $k > 0$ be the minimal index such that either (i) $Y_1(t_k) \wedge Y_2(t_k) < \tilde{Y}_1(t_k) \wedge \tilde{Y}_2(t_k)$ or (ii) $Y_1(t_k) + Y_2(t_k) < \tilde{Y}_1(t_k) + \tilde{Y}_2(t_k)$, assuming the existence of such k.

In Scenario (i), $Y_1(t_{k-1}) \wedge Y_2(t_{k-1}) = \tilde{Y}_1(t_{k-1}) \wedge \tilde{Y}_2(t_{k-1})$ and $Y_1(t_{k-1}) + Y_2(t_{k-1}) \geq \tilde{Y}_1(t_{k-1}) + \tilde{Y}_2(t_{k-1})$. If there is an arrival event at time t_k, then based on our coupling construction, this is an arrival to both systems, and the arrival is either to the smaller queue in both systems or to the large queue in both. If $Y_1(t_{k-1}) = Y_2(t_{k-1})$ so that $Y_1 \wedge Y_2$ does not increase at t_k, then as $Y_1(t_{k-1}) + Y_2(t_{k-1}) \geq \tilde{Y}_1(t_{k-1}) + \tilde{Y}_2(t_{k-1})$, $\tilde{Y}_1(t_{k-1}) = \tilde{Y}_2(t_{k-1})$, and so $\tilde{Y}_1 \wedge \tilde{Y}_2$ does not increase either. Hence, in this case $Y_1(t_k) \wedge Y_2(t_k) = \tilde{Y}_1(t_k) \wedge \tilde{Y}_2(t_k)$. Suppose instead there is a departure event at t_k. There must be a departure from the smaller component in $\tilde{Y}$. However, by P2 and our coupling construction, there must be a departure from the smaller component in Y as well. Hence, we again have that $Y_1(t_k) \wedge Y_2(t_k) = \tilde{Y}_1(t_k) \wedge \tilde{Y}_2(t_k)$. Thus, Scenario (i) is not feasible.

In Scenario (ii), $Y_1(t_{k-1}) \wedge Y_2(t_{k-1}) \geq \tilde{Y}_1(t_{k-1}) \wedge \tilde{Y}_2(t_{k-1})$ and $Y_1(t_{k-1}) + Y_2(t_{k-1}) = \tilde{Y}_1(t_{k-1}) + \tilde{Y}_2(t_{k-1})$. Since arrivals coincide in both systems, there must be a departure from Y at t_k. However, by P1 and our coupling construction, there must be a departure from $\tilde{Y}$ as well. This rules out Scenario (ii).

Combining the analysis for the two scenarios, there is a contradiction. Thus, (14) holds, which further implies the stochastic dominance results for the stationary distributions. □

Appendix B: Application of the stochastic dominance results

B.1. Proofs of Lemma 1 and Lemma 3

In this section, we apply Lemma 7 to compare two system configurations. Lemmas 1 and 3 then follow as corollaries to this comparison.

Fix policy $\nu^{\lambda,*}$ for X^λ, which has n^λ servers in each dedicated server pool and n_F^λ flexible servers. Consider two auxiliary queueing systems $\tilde{X}^\lambda$ and $\check{X}^\lambda$ based on X^λ. $\tilde{X}^\lambda$ has no flexible servers. Each dedicated pool of $\tilde{X}^\lambda$ has n^λ servers that can work at rate μ and $n_F^\lambda/2$ servers that can work at rate μ_F. When assigning customers to servers, the rate-μ servers are prioritized. On the other hand, $\check{X}^\lambda$ does not have any dedicated servers. Instead, it has $2n^\lambda + n_F^\lambda$ flexible servers, among which $2n^\lambda$ servers can work at rate μ and n_F^λ servers can work at rate μ_F. When assigning customers to servers, we again prioritize the faster servers.

LEMMA 9. *Suppose $\theta \leq \mu_F$. For $\tilde{X}^\lambda$, $\check{X}^\lambda$ and X^λ, if $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$,*

$$\check{X}_1^\lambda(\infty) + \check{X}_2^\lambda(\infty) \leq_{st} X_1^\lambda(\infty) + X_2^\lambda(\infty) \leq_{st} \tilde{X}_1^\lambda(\infty) + \tilde{X}_2^\lambda(\infty);$$

$$\check{X}_1^\lambda(\infty) \vee \check{X}_2^\lambda(\infty) \leq_{st} X_1^\lambda(\infty) \vee X_2^\lambda(\infty) \leq_{st} \tilde{X}_1^\lambda(\infty) \vee \tilde{X}_2^\lambda(\infty);$$

$$\left(\check{X}_1^\lambda(\infty) + \check{X}_2^\lambda(\infty) - 2n^\lambda - n_F^\lambda\right)^+ \leq_{st} Q_\Sigma^\lambda(\infty) \leq_{st} \left(\tilde{X}_1^\lambda(\infty) - n^\lambda - n_F^\lambda/2\right)^+ + \left(\tilde{X}_2^\lambda(\infty) - n^\lambda - n_F^\lambda/2\right)^+.$$

Proof of Lemma 9. Because all three processes are two-dimensional birth-and-death processes with common arrival rate λ, we can apply Lemma 7. To simplify the notation, we omit the superscript λ. Set $Y = X$ and $\tilde{Y} = \tilde{X}$. Then the death rates take the form:

$$\zeta_1(y_1, y_2) = \begin{cases} \mu(y_1 \wedge n) + \mu_F((y_1 - n)^+ \wedge n_F) + \theta(y_1 - n - n_F)^+ & y_1 \geq y_2 \\ \mu(y_1 \wedge n) + \mu_F((y_1 - n)^+ \wedge (n_F - (y_2 - n)^+)^+) + \theta((y_1 - n)^+ - (n_F - (y_2 - n)^+)^+) & y_1 < y_2 \end{cases}$$

$$\zeta_2(y_1, y_2) = \begin{cases} \mu(y_2 \wedge n) + \mu_F((y_2 - n)^+ \wedge (n_F - (y_1 - n)^+)^+) + \theta((y_2 - n)^+ - (n_F - (y_1 - n)^+)^+) & y_1 \geq y_2 \\ \mu(y_2 \wedge n) + \mu_F((y_2 - n)^+ \wedge n_F) + \theta(y_2 - n - n_F)^+ & y_1 < y_2 \end{cases}$$

$$\tilde{\zeta}_1(y_1, y_2) = \mu(y_1 \wedge n) + \mu_F(n_F/2 \wedge (y_1 - n)^+) + \theta(y_1 - n - n_F/2)^+$$

$$\tilde{\zeta}_2(y_1, y_2) = \mu(y_2 \wedge n) + \mu_F(n_F/2 \wedge (y_2 - n)^+) + \theta(y_2 - n - n_F/2)^+$$

Since $\mu \geq \mu_F \geq \theta$, it is straightforward to verify that P1 and P2 in Lemma 7 hold. Thus, from the proof of Lemma 7, we can construct a coupling such that

$$Y_1(t) + Y_2(t) \leq \tilde{Y}_1(t) + \tilde{Y}_2(t) \text{ and } Y_1(t) \vee Y_2(t) \leq \tilde{Y}_1(t) \vee \tilde{Y}_2(t)$$

for $t \geq 0$ path-by-path. In addition,

$$\begin{aligned} Q_\Sigma(t) &= \left((X_1(t)-n)^+ + (X_2(t)-n)^+ - n_F\right)^+ \\ &\leq (X_1(t) - (n+n_F/2))^+ + (X_2(t) - (n+n_F/2))^+ \\ &\leq (\tilde{X}_1(t) - (n+n_F/2))^+ + (\tilde{X}_2(t) - (n+n_F/2))^+ \end{aligned}$$

As $\tilde{Y}$ is positive recurrent for $(n, n_F) \in \Omega^\lambda(\theta)$, so is Y. Sending t to infinity for the coupled processes, we have the stochastic dominance results in stationarity. The stochastic dominance results for X over $\check{X}$ follow similarly. $\square$

For Lemma 1, we note that under the policy $\nu^{\lambda,*}$, for $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(0)$,

$$X_1^\lambda(\infty) + X_2^\lambda(\infty) \leq_{st} \tilde{X}_1^\lambda(\infty) + \tilde{X}_2^\lambda(\infty)$$

by Lemma 9. Then, the stability of $\tilde{X}_1^\lambda$ and $\tilde{X}_2^\lambda$ implies the stability of $(X_1^\lambda, X_2^\lambda)$.

For Lemma 3, we have $\mu = \mu_F$. In this case,

$$Q_\Sigma^\lambda(\infty; 0, 2n^\lambda + n_F^\lambda) \stackrel{d}{=} \left(\check{X}_1^\lambda(\infty) + \check{X}_2^\lambda(\infty) - 2n^\lambda - n_F^\lambda\right)^+.$$

Then, by Lemma 9, under the policy $\nu^{\lambda,*}$, we have

$$Q_\Sigma^\lambda(\infty; 0, 2n^\lambda + n_F^\lambda) \leq_{st} Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda).$$

B.2. Proof of Theorem 1

We apply Lemma 7 to prove Theorem 1. To simplify notation, we omit the superscript λ. Consider $Y(t) = X(t; n, n_F; \nu^*)$ and $\tilde{Y}(t) = X(t; n, n_F; \nu)$. We will first verify that P1 and P2 in Lemma 7 hold.

For P1, $y_1 + y_2 = \tilde{y}_1 + \tilde{y}_2$ and $y_1 \vee y_2 \leq \tilde{y}_1 \vee \tilde{y}_2$. Since $\mu \geq \mu_F \geq \theta$,

$$\tilde{\zeta}_\Sigma(\tilde{y}) \leq \zeta_\Sigma(\tilde{y}) \leq \zeta_\Sigma(y).$$

For P2, $y_1 \vee y_2 = \tilde{y}_1 \vee \tilde{y}_2$ and $y_1 + y_2 \leq \tilde{y}_1 + \tilde{y}_2$. Without loss of generality, suppose $y_1 \geq y_2$ and $y_1 = \tilde{y}_1 \geq \tilde{y}_2$. Then,

$$\tilde{\zeta}_M(\tilde{y}) \leq \zeta_M(\tilde{y}) = \mu(y_1 \wedge n) + \mu_F(n_F \wedge (y_1 - n)^+) + \theta(y_1 - n - n_F)^+ = \zeta_M(y).$$

The positive recurrence of Y is established in Lemma 1. For $\tilde{Y}$, if it is not positive recurrent, we define $\tilde{Y}_i(\infty) = \infty$. Then by Lemma 7,

$$X_1^\lambda(\infty; n, n_F; \nu^*) + X_2^\lambda(\infty; n, n_F; \nu^*) \leq_{st} X_1^\lambda(\infty; n, n_F; \nu) + X_2^\lambda(\infty; n, n_F; \nu)$$

and

$$X_1^\lambda(\infty; n, n_F; \nu^*) \vee X_2^\lambda(\infty; n, n_F; \nu^*) \leq_{st} X_1^\lambda(\infty; n, n_F; \nu) \vee X_2^\lambda(\infty; n, n_F; \nu).$$

Lastly, for the queue length, consider the function $f : \mathbb{N}_0^2 \to \mathbb{N}_0$ defined by $f(y_1, y_2) = ((y_1 - n)^+ + (y_2 - n)^+ - n_F)^+$. Note that if $y_1 + y_2 \leq \tilde{y}_1 + \tilde{y}_2$ and $y_1 \vee y_2 \leq \tilde{y}_1 \vee \tilde{y}_2$, then $f(y) \leq f(\tilde{y})$. Therefore,

$$\begin{aligned} Q_\Sigma^\lambda(\infty; n, n_F, \nu^*) &= \left((X_1^\lambda(\infty; n, n_F; \nu^*) - n)^+ + (X_2^\lambda(\infty; n, n_F; \nu^*) - n)^+ - n_F\right)^+ \\ &\leq_{st} \left((X_1^\lambda(\infty; n, n_F; \nu) - n)^+ + (X_2^\lambda(\infty; n, n_F; \nu) - n)^+ - n_F\right)^+ \leq_{st} Q_\Sigma^\lambda(\infty; n, n_F; \nu). \end{aligned}$$

$\square$

B.3. Optimal scheduling rule when $\theta \geq \mu_F = \mu$

Define $\phi^{\lambda,*}$ by

$$Z_i^\lambda(t) = \min\{n^\lambda, X_i^\lambda(t)\} \text{ for } i = 1, 2; \tag{15}$$

and if $X_1^\lambda(t) \leq X_2^\lambda(t)$,

$$Z_{F1}^\lambda(t) = \min\{n_F^\lambda, (X_1^\lambda(t) - n^\lambda)^+\}, \quad Z_{F2}^\lambda(t) = \min\{n_F^\lambda - Z_{F1}^\lambda(t), (X_2^\lambda(t) - n^\lambda)^+\}; \tag{16}$$

otherwise,

$$Z_{F1}^\lambda(t) = \min\{n_F^\lambda - Z_{F2}^\lambda(t), (X_1^\lambda(t) - n^\lambda)^+\}, \quad Z_{F2}^\lambda(t) = \min\{n_F^\lambda, (X_2^\lambda(t) - n^\lambda)^+\}. \tag{17}$$

That is, the flexible pool gives priority to the class with fewer customers in the system. The next theorem show that $\phi^{\lambda,*}$ is optimal when $\theta \geq \mu_F = \mu$.

THEOREM 5. *Suppose $\theta \geq \mu = \mu_F$. For any deterministic Markovian scheduling policy ν^λ,*

$$\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \nu^\lambda)] \geq \mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \phi^{\lambda,*})],$$

which implies that $\Pi^\lambda(n^\lambda, n_F^\lambda; \nu^\lambda) \geq \Pi^\lambda(n^\lambda, n_F^\lambda; \phi^{\lambda,*})$.

Proof of Theorem 5. The proof of Theorem 5 uses a coupling construction similar to that of Theorem 1, but does so by considering a 'dual' problem where we maximize the number of busy servers. In particular, the key observation is that

$$\theta\mathbb{E}[Q_\Sigma^\lambda(\infty)] = 2\lambda - \mu\mathbb{E}[Z_1^\lambda(\infty) + Z_2^\lambda(\infty)] - \mu_F\mathbb{E}[Z_F^\lambda(\infty)] \tag{18}$$

so that minimizing $\mathbb{E}[Q_\Sigma^\lambda(\infty)]$ is equivalent to maximizing

$$\mu\mathbb{E}[Z_1^\lambda(\infty) + Z_2^\lambda(\infty)] + \mu_F\mathbb{E}[Z_F^\lambda(\infty)].$$

This may be accomplished by keeping $X_1^\lambda + X_2^\lambda$ and $X_1^\lambda \wedge X_2^\lambda$ both large. Based on this observation, we shall prove Theorem 5 using Lemma 8.

To simplify the notation, we drop the superscript λ. Let $Y(t) = X(t; n, n_F; \phi^*)$ and $\tilde{Y}(t) = X(t; n, n_F; \nu)$. We next verify P1 and P2 of Lemma 8.

For P1, $y_1 + y_2 = \tilde{y}_1 + \tilde{y}_2$ and $y_1 \wedge y_2 \geq \tilde{y}_1 \wedge \tilde{y}_2$. In this case, we have

$$\zeta_\Sigma(y) \leq \zeta_\Sigma(\tilde{y}) \leq \tilde{\zeta}_\Sigma(\tilde{y}).$$

For P2, $y_1 \wedge y_2 = \tilde{y}_1 \wedge \tilde{y}_2$ and $y_1 + y_2 \geq \tilde{y}_1 + \tilde{y}_2$. Without loss of generality, suppose $y_1 \leq y_2$ and $y_1 = \tilde{y}_1 \leq \tilde{y}_2$. Then,

$$\tilde{\zeta}_m(\tilde{y}) = \tilde{\zeta}_1(\tilde{y}) \geq \mu(y_1 \wedge (n + n_F)) + \theta(y_1 - n - n_F)^+ = \zeta_m(y)$$

From Lemma 8, we can construct a coupling, under which

$$Y_1(t) + Y_2(t) \geq \tilde{Y}_1(t) + \tilde{Y}_2(t) \text{ and } Y_1(t) \wedge Y_2(t) \geq \tilde{Y}_1(t) \wedge \tilde{Y}_2(t).$$

This further implies that

$$\mu(Z_1(t) + Z_2(t)) + \mu_F Z_F(t) \geq \mu(\tilde{Z}_1(t) + \tilde{Z}_2(t)) + \mu_F \tilde{Z}_F(t).$$

As $\theta > 0$, both Y and $\tilde{Y}$ are positive recurrent. Thus,

$$\mu(Z_1(\infty) + Z_2(\infty)) + \mu_F Z_F(\infty) \geq_{st} \mu(\tilde{Z}_1(\infty) + \tilde{Z}_2(\infty)) + \mu_F \tilde{Z}_F(\infty),$$

This completes the proof due to (18). □

REMARK 2. It is hard to extend the results in Theorem 5 to the case where $\mu > \mu_F$. This is because when $\mu > \mu_F$, P1 in Lemma 8 no longer holds. For example, consider $n = n_F = 1$, $y = (1,1)$ and $\tilde{y} = (0,2)$. In this case, $\zeta_\Sigma(y) = 2\mu > \mu + \mu_F \geq \tilde{\zeta}_\Sigma(\tilde{y})$.

Appendix C: Proofs of the Results in Section 3.2

C.1. Proof of Lemma 2.

Note that $\mathbb{E}[Q_\Sigma^\lambda(\infty; \lfloor R^\lambda + \sqrt{R^\lambda} \rfloor, 0)] = O(\sqrt{\lambda})$ (Garnett et al. 2002). Thus,

$$\Pi^{\lambda,*} \leq \Pi^\lambda(\lfloor R^\lambda + \sqrt{R^\lambda} \rfloor, 0) = 2cR^\lambda + O(\sqrt{\lambda}).$$

To prove $\Pi^{\lambda,*} = 2cR^\lambda + O(\sqrt{\lambda})$, it suffices to prove that $n^{\lambda,*} = R^\lambda + O(\sqrt{\lambda})$ and $n_F^{\lambda,*} = O(\sqrt{\lambda})$.

We first prove $\limsup_{\lambda\to\infty} \frac{n^{\lambda,*} - R^\lambda}{\sqrt{\lambda}} < \infty$. Suppose by contradiction that there exists a subsequence $\{\lambda_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} \lambda_k = \infty$ and $\lim_{k\to\infty}(n^{\lambda_k,*} - R_k)/\sqrt{\lambda_k} = \infty$, where $R_k = \lambda_k/\mu$. Then,

$$\frac{\Pi^{\lambda_k}(n^{\lambda_k,*}, n_F^{\lambda_k,*}) - 2cR_k}{\sqrt{\lambda_k}} \geq \frac{c(2n^{\lambda_k,*} + n_F^{\lambda_k,*} - 2R_k)}{\sqrt{\lambda_k}} \geq \frac{2c(n^{\lambda_k,*} - R_k)}{\sqrt{\lambda_k}} \to \infty,$$

contradicting that $\Pi^{\lambda,*} \leq 2cR^\lambda + O(\sqrt{\lambda})$.

We next prove that $\liminf_{\lambda\to\infty} \frac{n^{\lambda,*} - R^\lambda}{\sqrt{\lambda}} > -\infty$ and $\limsup_{\lambda\to\infty} \frac{n_F^{\lambda,*}}{\sqrt{\lambda}} < \infty$.

Consider the case where $\theta = 0$. Note that for stability, $2n^{\lambda,*}\mu + n_F^{\lambda,*}\mu_F > 2\lambda$. To prove $\limsup_{\lambda\to\infty} \frac{n_F^{\lambda,*}}{\sqrt{\lambda}} < \infty$, we suppose for contradiction that there exists a subsequence $\{\lambda_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} \lambda_k = \infty$ and $\lim_{k\to\infty} n_F^{\lambda_k,*}/\sqrt{\lambda_k} = \infty$. Note that $2n^{\lambda_k,*} > 2\lambda_k/\mu - n_F^{\lambda_k,*}\mu_F/\mu$. Then,

$$\frac{\Pi^{\lambda_k}(n^{\lambda_k,*}, n_F^{\lambda_k,*}) - 2cR_k}{\sqrt{\lambda_k}} \geq \frac{c(2n^{\lambda_k,*} - 2R_k) + c_F n_F^{\lambda_k,*}}{\sqrt{\lambda_k}} \geq \frac{n_F^{\lambda_k,*}(c_F - c\mu_F/\mu)}{\sqrt{\lambda_k}} \to \infty,$$

contradicting that $\Pi^{\lambda,*} \leq 2cR^\lambda + O(\sqrt{\lambda})$. Since $2n^{\lambda_k,*} > 2\lambda_k/\mu - n_F^{\lambda_k,*}\mu_F/\mu$, this also shows that $\liminf_{\lambda\to\infty} \frac{n^{\lambda,*} - R^\lambda}{\sqrt{\lambda}} > -\infty$.

We now turn to the case where $\theta > 0$. We first note that $\theta\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda)] \geq 2\lambda - 2n^\lambda\mu - n_F^\lambda\mu_F$. To prove $\limsup_{\lambda\to\infty} \frac{n_F^{\lambda,*}}{\sqrt{\lambda}} < \infty$, suppose for contradiction that there exists a subsequence $\{\lambda_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} \lambda_k = \infty$ and $\lim_{k\to\infty} n_F^{\lambda_k,*}/\sqrt{\lambda_k} = \infty$. Note that

$$\frac{\Pi^{\lambda_k}(n^{\lambda_k,*}, n_F^{\lambda_k,*}) - 2cR_k}{\sqrt{\lambda_k}} \geq \frac{2c(n^{\lambda_k,*} - R_k) + c_F n_F^{\lambda_k,*}}{\sqrt{\lambda_k}}. \tag{19}$$

Since the LHS of (19) must be bounded, say by $2cC$ for some constant $C > 0$, we have

$$n^{\lambda_k,*} - R_k \leq C\sqrt{\lambda_k} - \frac{c_F}{2c} n_F^{\lambda_k,*}.$$

Therefore,

$$\lambda_k - n^{\lambda_k,*}\mu \geq \frac{c_F\mu}{2c} n_F^{\lambda_k,*} - C\mu\sqrt{\lambda_k} \geq \frac{1}{2} n_F^{\lambda_k,*}\mu_F - C\mu\sqrt{\lambda_k}.$$

Next, for $h/\theta + a = c_F/\mu_F + \delta = c/\mu + \epsilon$ satisfying $0 < \delta < \epsilon$,

$$\begin{aligned}
\frac{\Pi^{\lambda_k}(n^{\lambda_k,*}, n_F^{\lambda_k,*}) - 2cR_k}{\sqrt{\lambda_k}} &= \frac{(h/\theta + a)\theta\mathbb{E}[Q_\Sigma^\lambda(\infty; n^{\lambda_k,*}, n_F^{\lambda_k,*})] + 2c(n^{\lambda_k,*} - R_k) + c_F n_F^{\lambda_k,*}}{\sqrt{\lambda_k}} \\
&\geq \frac{(h/\theta + a)(2\lambda_k - 2n^{\lambda_k,*}\mu - n_F^{\lambda_k,*}\mu_F) + 2c(n^{\lambda_k,*} - R_k) + c_F n_F^{\lambda_k,*}}{\sqrt{\lambda_k}} \\
&= \frac{2\epsilon(\lambda_k - n^{\lambda_k,*}\mu) - \delta n_F^{\lambda_k,*}\mu_F}{\sqrt{\lambda_k}} \\
&\geq \frac{(\epsilon - \delta) n_F^{\lambda_k,*}\mu_F - 2\epsilon C\mu\sqrt{\lambda_k}}{\sqrt{\lambda_k}} \\
&\to \infty
\end{aligned}$$

as $k \to \infty$. This contradicts that $\Pi^{\lambda,*} \le 2cR^\lambda + O(\sqrt{\lambda})$, and so $n_F^\lambda = O(\sqrt{\lambda})$.

To prove that $\liminf_{\lambda\to\infty} \frac{n^{\lambda,*}-R^\lambda}{\sqrt{\lambda}} > -\infty$, assume for contradiction that there exists a subsequence $\{\lambda_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty}\lambda_k = \infty$ and $\lim_{k\to\infty}(n^{\lambda_k,*}-R_k)/\sqrt{\lambda_k} = -\infty$. Then for $n_F^{\lambda_k,*} = O(\sqrt{\lambda_k})$

$$\begin{aligned}\frac{\Pi^{\lambda_k}(n^{\lambda_k,*}, n_F^{\lambda_k,*}) - 2cR_k}{\sqrt{\lambda_k}} &\ge \frac{2\epsilon(\lambda_k - n^{\lambda_k,*}\mu) - \delta n_F^{\lambda_k,*}\mu_F}{\sqrt{\lambda_k}} \\ &= \frac{-2\epsilon\mu(n^{\lambda_k,*}-R_k) - \delta n_F^{\lambda_k,*}\mu_F}{\sqrt{\lambda_k}} \\ &\to \infty.\end{aligned}$$

This is a contradiction. □

C.2. Some auxiliary lemmas

Before we prove Theorem 2, we first present three auxiliary lemmas.

LEMMA 10. *Let $M^\lambda = \{M^\lambda(t) : t \ge 0\}$ be a sequence of ergodic Markov chains taking values in $\mathbb{R}^m$, and $h : \mathbb{R}^m \to \mathbb{R}^n$ be a measurable function. Suppose*

1. *$h(M^\lambda(t)) \Rightarrow R(t)$ in D^n if $h(M^\lambda(0)) \Rightarrow R(0)$ as $\lambda \to \infty$, where R is a continuous ergodic process with a unique stationary distribution $R(\infty)$;*

2. *$\{h(M^\lambda(\infty)) : \lambda \ge 1\}$ is tight.*

Then, $h(M^\lambda(\infty)) \Rightarrow R(\infty)$ as $\lambda \to \infty$.

Proof. The proof follows similar lines of argument as Gamarnik and Zeevi (2006). As $\{(h(M^\lambda(\infty)) : \lambda \ge 1\}$ is tight, every subsequence has a convergent further subsequence. Let Y be a weak limit of $\{(h(M^\lambda(\infty)) : \lambda \ge 1\}$, i.e., there exists a sequence $\{\lambda_k : k \in \mathbb{N}\}$, such that $h(M^{\lambda_k}(\infty)) \Rightarrow Y$ as $k \to \infty$.

Now for each k, set $M^{\lambda_k}(0) \stackrel{d}{=} M^{\lambda_k}(\infty)$. Then, we have $M^{\lambda_k}(t) \stackrel{d}{=} M^{\lambda_k}(\infty)$ for any $t \ge 0$. This implies that $h(M^{\lambda_k}(0)) \Rightarrow Y$, which further implies that $h(M^{\lambda_k}(t)) \Rightarrow R(t)$ in D^n as $k \to \infty$. As $R(0) \stackrel{d}{=} Y$, $R(t) \stackrel{d}{=} Y$. Furthermore, as $R(t) \Rightarrow R(\infty)$ as $t \to \infty$, $Y \stackrel{d}{=} R(t) \stackrel{d}{=} R(\infty)$. Therefore, every weak limit of $\{(h(M^\lambda(\infty)) : \lambda \ge 1\}$ follows the same distribution as $R(\infty)$. This indicates that $h(M^\lambda(\infty)) \Rightarrow R(\infty)$ as $\lambda \to \infty$. □

Let $\tilde{X}_1^\lambda(\cdot)$ denote the number of customers in a system with arrival rate λ, n^λ rate-μ servers and $n_F^\lambda/2$ rate-μ_F servers.

LEMMA 11. *If either (i) $\theta = 0$ and $\lambda < n^\lambda\mu + \frac{n_F^\lambda}{2}\mu_F = \lambda + \Theta(\sqrt{\lambda})$ or (ii) $\theta > 0$, $n^\lambda\mu = \lambda + O(\sqrt{\lambda})$, and $n_F^\lambda = O(\sqrt{\lambda})$,*

$$\sup_{\lambda>1} \mathbb{E}\left[\left(\frac{(\tilde{X}_1^\lambda(\infty) - n^\lambda - n_F^\lambda/2)^+}{\sqrt{\lambda}}\right)^2\right] < \infty.$$

Proof. Let $C^\lambda = n^\lambda + n_F^\lambda/2$. We first note that $\tilde{X}_1^\lambda(\cdot)$ is a positive-recurrent birth-death process. Let π^λ denote its stationary distribution. In Case (i), for $k \ge C^\lambda$, we have

$$\pi^\lambda(k) = \pi^\lambda(C^\lambda)\left(\frac{\lambda}{n^\lambda\mu + \frac{n_F^\lambda}{2}\mu_F}\right)^{k-C^\lambda}.$$

This implies that $(\tilde{X}_1^\lambda(\infty) - C^\lambda)^+$ is stochastically bounded by a geometric random variable with probability of success

$$1 - \frac{\lambda}{n^\lambda\mu + n_F^\lambda\mu_F/2} = \Theta(1/\sqrt{\lambda}).$$

Thus, $\mathbb{E}\left[\left(\tilde{X}_1^\lambda(\infty)-C^\lambda)^+\right)^2\right]=O(\lambda)$.

In Case (ii), choose $l^\lambda\geq 0$ such that $l^\lambda=O(\sqrt{\lambda})$ and $n^\lambda\mu+\frac{n_F^\lambda}{2}\mu_F+l^\lambda\theta=\lambda+\Theta(\sqrt{\lambda})$, and note that it suffices to prove that

$$\sup_{\lambda>1}\mathbb{E}\left[\left(\frac{(\tilde{X}_1^\lambda(\infty)-n^\lambda-n_F^\lambda/2-l^\lambda)^+}{\sqrt{\lambda}}\right)^2\right]<\infty.$$

Let $D^\lambda=n^\lambda+n_F^\lambda/2+l^\lambda$. For $k\geq D^\lambda$, we have

$$\pi^\lambda(k)=\pi^\lambda(D^\lambda)\prod_{j=1}^{k-D^\lambda}\left(\frac{\lambda}{n^\lambda\mu+\frac{n_F^\lambda}{2}\mu_F+l^\lambda\theta+j\theta}\right)\leq\pi^\lambda(D^\lambda)\left(\frac{\lambda}{n^\lambda\mu+\frac{n_F^\lambda}{2}\mu_F+l^\lambda\theta}\right)^{k-D^\lambda}.$$

Thus $(\tilde{X}_1^\lambda(\infty)-D^\lambda)^+$ is stochastically bounded by a geometric random variable with probability of success

$$1-\frac{\lambda}{n^\lambda\mu+n_F^\lambda\mu_F/2+l^\lambda\theta}=\Theta(1/\sqrt{\lambda}),$$

Thus, $\mathbb{E}\left[\left(\tilde{X}_1^\lambda(\infty)-D^\lambda)^+\right)^2\right]=O(\lambda)$. $\square$

LEMMA 12. *For $(n^\lambda,n_F^\lambda)\in\Omega^\lambda(\theta)$, $2n^\lambda+n_F^\lambda=2R^\lambda+\gamma\sqrt{R^\lambda}+o(\sqrt{R^\lambda})$, and $n_F^\lambda=O(\sqrt{\lambda})$, we have*

$$\sup_{\lambda>1}\mathbb{E}\left[\frac{(X_i^\lambda(\infty)-n^\lambda)^-}{\sqrt{\lambda}}\right]<\infty.$$

Proof. We prove the lemma for $i=1$ only; the case $i=2$ is similar. Let $\pi_1^\lambda(k)=\mathbb{P}(X_1^\lambda(\infty)=k)$. Then for $0\leq k<n^\lambda$, we have $\lambda\pi_1^\lambda(k)=(k+1)\mu\pi_1^\lambda(k+1)$. This implies that

$$\begin{aligned}\mathbb{E}[(X_1^\lambda(\infty)-n^\lambda)^-]&=\pi_1^\lambda(n^\lambda)\sum_{k=0}^{n^\lambda}(n^\lambda-k)\frac{\mu^{n^\lambda-k}n^\lambda!}{\lambda^{n^\lambda-k}k!}\\&\leq\frac{1}{\sum_{k=0}^{n^\lambda}\frac{\mu^{n^\lambda-k}n^\lambda!}{\lambda^{n^\lambda-k}k!}}\cdot\sum_{k=0}^{n^\lambda}(n^\lambda-k)\frac{\mu^{n^\lambda-k}n^\lambda!}{\lambda^{n^\lambda-k}k!}\\&=\frac{1}{\sum_{k=0}^{n^\lambda}\frac{\mu^{n^\lambda-k}n^\lambda!}{\lambda^{n^\lambda-k}k!}}\cdot\frac{1}{\pi_c^\lambda(n^\lambda)}\cdot\pi_c^\lambda(n^\lambda)\sum_{k=0}^{n^\lambda}(n^\lambda-k)\frac{\mu^{n^\lambda-k}n^\lambda!}{\lambda^{n^\lambda-k}k!}\\&=\frac{1}{\sum_{k=0}^{n^\lambda}\pi_c^\lambda(k)}\cdot\mathbb{E}[(X_c^\lambda(\infty)-n^\lambda)^-]\end{aligned}$$

where X_c^λ denotes the number-in-system process of an $M/M/(n^\lambda+n_F^\lambda)+M$ queue with arrival rate λ and service rate μ, and abandonment rate $\theta\geq 0$, and π_c^λ denotes the stationary distribution of X_c^λ. As $\mathbb{E}[(X_c^\lambda(\infty)-n^\lambda-n_F^\lambda)^-]=O(\sqrt{\lambda})$ (Garnett et al. 2002) and $n_F^\lambda=O(\sqrt{\lambda})$, $\mathbb{E}[(X_c^\lambda(\infty)-n^\lambda)^-]=O(\sqrt{\lambda})$. We also note that $\sum_{k=0}^{n^\lambda}\pi_c^\lambda(k)=\mathbb{P}(X_c^\lambda\leq n^\lambda)=\Theta(1)$. Thus, $\mathbb{E}[(X_1^\lambda(\infty)-n^\lambda)^-]=O(\sqrt{\lambda})$. $\square$

C.3. Proof of Theorem 2

Define, for $i=1,2$, the fluid-scale processes

$$\bar{Z}_i^\lambda(t)=\frac{Z_i^\lambda(t)}{n^\lambda},\quad\bar{A}_i^\lambda(t)=\frac{A_i(\lambda t)}{n^\lambda},\quad\text{and}\quad\bar{S}_i^\lambda(t)=\frac{S_i(n^\lambda\mu t)}{n^\lambda}.$$

We also define, for $i=1,2$,

$$\bar{G}_i^\lambda(t)=\frac{G_i(\theta\sqrt{\lambda}t)}{\sqrt{\lambda}}\text{ and }\bar{S}_{F_i}^\lambda(t)=\frac{S_{F_i}(\mu_F\sqrt{\lambda}t)}{\sqrt{\lambda}}.$$

Define, for $i=1,2$, the diffusion-scale processes

$$\hat{A}_i^\lambda(t)=\frac{A_i(\lambda t)-\lambda t}{\sqrt{\lambda}} \text{ and } \hat{S}_i^\lambda(t)=\frac{S_i(n^\lambda\mu t)-n^\lambda\mu t}{\sqrt{\lambda}}.$$

We first note that because

$$X_i^\lambda(t)=X_i^\lambda(0)+A_i(\lambda t)-G_i\left(\theta\int_0^t Q_i^\lambda(s)\,ds\right)-S_i\left(\mu\int_0^t Z_i^\lambda(s)\,ds\right)-S_{Fi}\left(\mu_F\int_0^t Z_{Fi}^\lambda(s)\,ds\right),$$

we have that

$$\hat{X}_i^\lambda(t)=\hat{X}_i^\lambda(0)+\hat{Y}_i^\lambda(t)+F_i(\hat{X}^\lambda)(t),$$

where

$$\hat{Y}_i^\lambda(t)=\hat{A}_i^\lambda(t)-\hat{S}_i^\lambda\left(\int_0^t\bar{Z}_i^\lambda(s)\,ds\right)-\left(\frac{S_{Fi}\left(\mu_F\int_0^t Z_{Fi}^\lambda(s)\,ds\right)}{\sqrt{\lambda}}-\mu_F\int_0^t f_i(\hat{X}^\lambda(s))\,ds\right)$$
$$-\left(\frac{G_i\left(\theta\int_0^t Q_i^\lambda(s)\,ds\right)}{\sqrt{\lambda}}-\theta\int_0^t(\hat{X}^\lambda(s)^+-f_i(\hat{X}^\lambda(s))\,ds\right)+\frac{\lambda-n^\lambda\mu}{\sqrt{\lambda}}t$$

and

$$F_i(\hat{X}^\lambda)(t)=\mu\int_0^t\hat{X}_i^\lambda(s)^-\,ds-(\mu_F-\theta)\int_0^t f_i(\hat{X}^\lambda(s))\,ds-\theta\int_0^t\hat{X}_i^\lambda(s)^+\,ds. \tag{20}$$

The proof of Theorem 2 is then divided into six steps.

Step 1. Establish the convergence of the fluid-scale number-in-service processes $\bar{Z}_i^\lambda$.

LEMMA 13. *For $(n^\lambda,n_F^\lambda)\in\Omega^\lambda(\theta)$, suppose $n^\lambda=R^\lambda+\beta\sqrt{R^\lambda}+o(\sqrt{R^\lambda})$ and $n_F^\lambda=\beta_F\sqrt{R^\lambda}+o(\sqrt{R^\lambda})$. If $\bar{Z}_i^\lambda(0)\to 1$, $i=1,2,$, then $\bar{Z}_i^\lambda\Rightarrow I$ in D as $\lambda\to\infty$.*

Proof of Lemma 13. For any fixed $\epsilon>0$ and $T>0$, we shall prove that

$$\lim_{\lambda\to\infty}\mathbb{P}\left(\inf_{0\le t\le T}\bar{Z}_1^\lambda(t)<1-\epsilon\right)\to 0.$$

Define $\bar{\tau}_1^\lambda=\inf\{0\le t\le T:\bar{Z}_1^\lambda(t)<1-\epsilon\}$ and $\bar{\tau}_2^\lambda=\sup\{0\le t<\bar{\tau}_1^\lambda:\bar{Z}_1^\lambda(t)>1-\epsilon/2\}$. Let $\bar{E}^\lambda$ be the event that $\bar{\tau}_1^\lambda$ and $\bar{\tau}_2^\lambda$ are well-defined, i.e. $\bar{\tau}_i^\lambda\le T$. The initial condition $\bar{Z}_i^\lambda(0)\to 1$ implies that $\left\{\inf_{0\le t\le T}\bar{Z}_1^\lambda(t)<1-\epsilon\right\}\subseteq\bar{E}^\lambda$ for λ sufficiently large.

As $\bar{Z}_1^\lambda(t)<1$ for $t\in[\bar{\tau}_2^\lambda,\bar{\tau}_1^\lambda]$, all Class 1 arrivals on $[\bar{\tau}_2^\lambda,\bar{\tau}_1^\lambda]$ join the dedicated server pool immediately on arrival. Moreover there are no abandonments from Class 1. Thus,

$$(\bar{A}_1^\lambda(\bar{\tau}_1^\lambda)-\bar{A}_1^\lambda(\bar{\tau}_2^\lambda-))-\left(\bar{S}_1^\lambda\left(\int_0^{\bar{\tau}_1^\lambda}\bar{Z}_1(s)ds\right)-\bar{S}_1^\lambda\left(\int_0^{\bar{\tau}_2^\lambda}\bar{Z}_1(s)ds\right)\right)=\bar{Z}_1^\lambda(\bar{\tau}_1^\lambda)-\bar{Z}_1^\lambda(\bar{\tau}_2^\lambda-)\le-\epsilon/2.$$

This further implies that

$$\mathbb{P}(\bar{E}^\lambda)\le\mathbb{P}\left(\inf_{\substack{0\le s\le t\le T\\0\le u\le s}}(\bar{A}_1^\lambda(t)-\bar{A}_1^\lambda(s))-(\bar{S}_1^\lambda(u+t-s)-\bar{S}_1^\lambda(u))\le-\epsilon/2\right)\to 0,$$

where the convergence follows from the fact that, by the functional strong law of large numbers (FSLLN) for Poisson processes, $(\bar{A}_1^\lambda,\bar{S}_1^\lambda)\Rightarrow(\mu\chi,\mu\chi)$ as $\lambda\to\infty$. The analysis for $\bar{Z}_2^\lambda$ follows similarly. □

We note from Lemma 13 that in the fluid scale, the dedicated servers are busy all the time.

Step 2. Establish proper limits for the diffusion-scale processes $\hat{Y}_i^\lambda$.

LEMMA 14. *For $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$, suppose $n^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ and $n_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$. If $\bar{Z}_i^\lambda(0) \to 1$, $i = 1, 2$, then*

$$\left(\hat{Y}_1^\lambda, \hat{Y}_2^\lambda\right) \Rightarrow (\sqrt{2}B_1 - \beta\sqrt{\mu}\chi, \sqrt{2}B_2 - \beta\sqrt{\mu}\chi) \text{ in } D^2 \text{ as } \lambda \to \infty,$$

where B_1 and B_2 are independent Brownian motions.

Proof of Lemma 14. Recall that

$$\hat{Y}_i^\lambda(t) = \hat{A}_i^\lambda(t) - \hat{S}_i^\lambda\left(\int_0^t \bar{Z}_i^\lambda(s)\,ds\right) - \left(\frac{S_{Fi}\left(\mu_F\int_0^t Z_{Fi}^\lambda(s)\,ds\right)}{\sqrt{\lambda}} - \mu_F\int_0^t f_i(\hat{X}^\lambda(s))\,ds\right)$$
$$- \left(\frac{G_i\left(\theta\int_0^t Q_i^\lambda(s)\,ds\right)}{\sqrt{\lambda}} - \theta\int_0^t (\hat{X}^\lambda(s)^+ - f_i(\hat{X}^\lambda(s))\,ds\right) + \frac{\lambda - n^\lambda\mu}{\sqrt{\lambda}}t.$$

We shall analyze the five components of $\hat{Y}_i$ in sequence.

First, by the functional central limit theorem (FCLT) for Poisson processes, $\hat{A}_i^\lambda \Rightarrow B_i$ in D as $\lambda \to \infty$, where B_i is a Brownian motion.

Second, as $\int_0^\cdot \bar{Z}_i^\lambda(s)\,ds \Rightarrow \chi$ (Lemma 13), by a random time change, the FCLT for Poisson processes, and the continuous mapping theorem (Chapter 13 of Whitt (2002)), we have $\hat{S}_i^\lambda\left(\int_0^\cdot \bar{Z}_i^\lambda(s)\,ds\right) \Rightarrow \tilde{B}_i$ in D as $\lambda \to \infty$, where $\tilde{B}_i$ is a Brownian motion and is independent of B_i.

Third, by the FSLLN for Poisson processes, $\bar{S}_{Fi}^\lambda \Rightarrow \mu_F\chi$ as $\lambda \to \infty$. Next, we rewrite

$$\frac{\int_0^t Z_{Fi}^\lambda(s)\,ds}{\sqrt{\lambda}} = \int_0^t f_i^\lambda(\hat{X}_1^\lambda(s), \hat{X}_2^\lambda(s))\,ds,$$

where

$$f_1^\lambda(x_1, x_2) = \begin{cases} x_1^+ \wedge \frac{n_F^\lambda}{\sqrt{\lambda}}, & x_1 \geq x_2, \\ x_1^+ \wedge \left(\frac{n_F^\lambda}{\sqrt{\lambda}} - x_2^+\right)^+, & x_1 < x_2; \end{cases} \quad \text{and} \quad f_2^\lambda(x_1, x_2) = \begin{cases} x_2^+ \wedge \left(\frac{n_F^\lambda}{\sqrt{\lambda}} - x_1^+\right)^+, & x_1 \geq x_2, \\ x_2^+ \wedge \frac{n_F^\lambda}{\sqrt{\lambda}}, & x_1 < x_2. \end{cases}$$

Then, as $f^\lambda \to f$ as $\lambda \to \infty$,

$$\begin{aligned} &\frac{S_{Fi}\left(\mu_F\int_0^t Z_{Fi}^\lambda(s)\,ds\right)}{\sqrt{\lambda}} - \mu_F\int_0^t f_i(\hat{X}^\lambda(s))\,ds \\ =&\bar{S}_{Fi}^\lambda\left(\frac{1}{\sqrt{\lambda}}\int_0^t Z_{Fi}^\lambda(s)\,ds\right) - \mu_F\int_0^t f_i^\lambda(\hat{X}^\lambda(s))\,ds + \mu_F\int_0^t f_i^\lambda(\hat{X}^\lambda(s)) - f_i(\hat{X}^\lambda(s))\,ds \\ \Rightarrow&0 \text{ as } \lambda \to \infty. \end{aligned}$$

Fourth, by the FSLLN for Poisson processes, $\bar{G}_i^\lambda \to \theta\chi$ as $\lambda \to \infty$. Then, because

$$\frac{\theta\int_0^t Q_i^\lambda(s)\,ds}{\sqrt{\lambda}} = \frac{\theta\int_0^t ((X_i^\lambda(s) - n^\lambda)^+ - Z_{Fi}^\lambda(s))\,ds}{\sqrt{\lambda}} = \theta\int_0^t (\hat{X}_i^\lambda(s)^+ - f_i^\lambda(\hat{X}_1^\lambda(s), \hat{X}_2^\lambda(s))\,ds,$$

$$\begin{aligned} &\frac{G_i\left(\theta\int_0^t Q_i^\lambda(s)\,ds\right)}{\sqrt{\lambda}} - \theta\int_0^t (\hat{X}^\lambda(s)^+ - f_i(\hat{X}^\lambda(s))\,ds \\ =&\bar{G}_i^\lambda\left(\frac{1}{\sqrt{\lambda}}\int_0^t Q_i^\lambda(s)\,ds\right) - \frac{\theta}{\sqrt{\lambda}}\int_0^t Q_i^\lambda(s)\,ds + \theta\int_0^t (f_i(\hat{X}^\lambda(s) - f_i^\lambda(\hat{X}^\lambda(s))\,ds \\ \Rightarrow&0 \text{ as } \lambda \to \infty. \end{aligned}$$

Fifth, under the assumption of the lemma, $(\lambda - n^\lambda\mu)/\sqrt{\lambda} \to -\beta\sqrt{\mu}$ as $\lambda \to \infty$.

Finally, putting the five parts together, we have the result. $\square$

Step 3. Establish the C-tightness of the $\{\hat{X}^\lambda : \lambda \geq 1\}$.

LEMMA 15. *For $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$, suppose $n^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ and $n_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$. If $\hat{X}^\lambda(0) \Rightarrow \hat{X}(0)$ as $\lambda \to \infty$, $\left\{\hat{X}^\lambda : \lambda \geq 1\right\}$ is C-tight in $[0,T]$ for all $T > 0$.*

Proof of Lemma 15. Following the C-tightness definition in Chen and Zhang (2000), we will prove that for any fixed $\epsilon, \gamma > 0$, there exist $\delta > 0$ and $\lambda_0 > 0$ such that for all $\lambda \geq \lambda_0$,

$$\mathbb{P}\left(\sup_{\substack{0\leq s<t\leq T\\ |s-t|<\delta}} |\hat{X}_i^\lambda(t) - \hat{X}_i^\lambda(s)| \geq \epsilon\right) \leq \gamma$$

for $i = 1, 2$. Consider the representation

$$\begin{aligned}\hat{X}_i^\lambda(t) =& \hat{X}_i^\lambda(0) + \hat{A}_i^\lambda(t) - \hat{S}_i^\lambda\left(\int_0^t \bar{Z}_i^\lambda(s)\,ds\right) + \frac{\lambda - n^\lambda\mu}{\sqrt{\lambda}}t + \mu\int_0^t \hat{X}_i^\lambda(s)^-\,ds \\ &- \frac{S_{Fi}\left(\mu_F\int_0^t Z_{Fi}^\lambda(s)\,ds\right)}{\sqrt{\lambda}} - \frac{G_i(\theta\int_0^t Q_i^\lambda(s)\,ds)}{\sqrt{\lambda}}.\end{aligned}$$

First, as

$$\hat{X}_i^\lambda(0) + \hat{A}_i^\lambda(t) - \hat{S}_i^\lambda\left(\int_0^t \bar{Z}_i^\lambda(s)\,ds\right) + \frac{\lambda - n^\lambda\mu}{\sqrt{\lambda}}t \Rightarrow \hat{X}_i(0) + \sqrt{2}B_i(t) - \beta\sqrt{\mu}t \text{ in } D \text{ as } \lambda \to \infty$$

and $\sqrt{2}B_i(t) - \beta\sqrt{\mu}t$ is continuous, $\{\hat{X}_i^\lambda(0) + \hat{A}_i^\lambda(t) - \hat{S}_i^\lambda\left(\int_0^t \bar{Z}_i^\lambda(s)\,ds\right) + (\lambda - n^\lambda\mu)/\sqrt{\lambda} : \lambda \geq 1\}$ is C-tight (Lemma 4.2 of Chen and Zhang (2000)).

Second, for $0 \leq s \leq t \leq T$,

$$\begin{aligned}&\frac{1}{\sqrt{\lambda}}\left(S_{Fi}\left(\mu_F\int_0^t Z_{Fi}^\lambda(u)\,du\right) - S_{Fi}\left(\mu_F\int_0^s Z_{Fi}^\lambda(u)\,du\right)\right) \\ \leq& \bar{S}_{Fi}^\lambda\left(\frac{\int_0^s Z_{Fi}^\lambda(u)\,du}{\sqrt{\lambda}} + \frac{n_F^\lambda(t-s)}{\sqrt{\lambda}}\right) - \bar{S}_{Fi}^\lambda\left(\frac{\int_0^s Z_{Fi}^\lambda(u)\,du}{\sqrt{\lambda}}\right).\end{aligned}$$

Then, the C-tightness of $\{\frac{1}{\sqrt{\lambda}}S_{Fi}\left(\mu_F\int_0^\cdot Z_{Fi}^\lambda(s)\,ds\right)\}$ follows from the fact that $n_F^\lambda/\sqrt{\lambda} \to \beta_F/\sqrt{\mu} < \infty$ and $\bar{S}_{Fi}^\lambda \Rightarrow \mu_F\chi$ in D as $\lambda \to \infty$.

Third, for $0 \leq s \leq t \leq T$, we note that

$$\begin{aligned}&\frac{1}{\sqrt{\lambda}}\left(G_i\left(\theta\int_0^t Q_i^\lambda(u)\,du\right) - G_i\left(\theta\int_0^s Q_i^\lambda(u)\,du\right)\right) \\ \leq& \bar{G}_i^\lambda\left(\frac{\int_0^s Q_i^\lambda(u)\,du + (t-s)\sup_{0\leq v\leq T} Q_i^\lambda(v)}{\sqrt{\lambda}}\right) - \bar{G}_i^\lambda\left(\frac{\int_0^s Q_i^\lambda(u)\,du}{\sqrt{\lambda}}\right).\end{aligned}$$

Then, to prove that $\{\frac{1}{\sqrt{\lambda}}G_i(\theta\int_0^\cdot Q_i^\lambda(s)\,ds)\}$ is C-tight, it suffices to prove that for any $\gamma > 0$, there exists $K, \lambda_0 > 0$, such that $\mathbb{P}\left(\sup_{0\leq v\leq T} Q_i^\lambda(v)/\sqrt{\lambda} \geq K\right) \leq \gamma/2$ for every $\lambda > \lambda_0$. Furthermore, since $n_F^\lambda = O(\sqrt{\lambda})$, it is sufficient to prove that $\mathbb{P}\left(\sup_{0\leq v\leq T} \hat{X}_i^\lambda(v) \geq K\right) \leq \gamma/2$, which follows from Lemma 9.

Fourth, we prove that $\{\mu\int_0^\cdot \hat{X}_i^\lambda(s)^-\,ds : \lambda \geq 1\}$ is C-tight. For $0 \leq s \leq t \leq T$, we first note that

$$\mu\int_0^t \hat{X}_i^\lambda(u)^-\,du - \mu\int_0^s \hat{X}_i^\lambda(u)^-\,du \leq \mu(t-s)\sup_{0\leq u\leq T} \hat{X}_i^\lambda(u)^-.$$

Next from Lemma 12 we have that for any $\gamma > 0$, there exists $K > 0$ and $\lambda_0 > 0$ such that for all $\lambda > \lambda_0$,

$$\mathbb{P}\left(\sup_{0\le u\le T} \hat{X}_i^\lambda(u)^- > K\right) \le \gamma.$$

Thus, $\{\mu \int_0^\cdot \hat{X}_i^\lambda(s)^- \, ds : \lambda \ge 1\}$ is C-tight.

Putting the four parts together, we have the C-tightness of $\left\{\hat{X}^\lambda : \lambda \ge 1\right\}$. $\square$

Lemma 15 implies that any subsequence of $\hat{X}^\lambda$ has a weakly convergent further subsequence and the limit is continuous almost surely (Proposition 4.1 in Chen and Zhang (2000)).

Step 4. Establish that F is continuous at almost all limit points of $\hat{X}^\lambda$.

LEMMA 16. *For $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$, suppose $n^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ and $n_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$. The mapping $F : D^2 \to D^2$ defined in (20) is continuous at almost all limit points of $\hat{X}^\lambda$.*

Proof of Lemma 16. From the C-tightness of $\{\hat{X}^\lambda : \lambda \ge 1\}$, almost all sub-sequential limits of $\hat{X}^\lambda$ are continuous. Thus, it suffices to prove continuity of F under the uniform topology. We denote by $\hat{X}$ a generic sub-sequential limit of $\hat{X}^\lambda$.

Fix $T > 0$. For $X \in D^2$, define $\|X\|_T = \sup_{0\le t\le T} \max(|X_1(t)|, |X_2(t)|)$. Now, fix $\epsilon > 0$. Consider $X, Y \in D^2$ with X continuous and $\|X - Y\|_T < \epsilon/2$.

For $0 \le t \le T$,

$$\left|\int_0^t X_i(s)^- \, ds - \int_0^t Y_i(s)^- \, ds\right| < \epsilon t/2 \le \epsilon T/2.$$

Similarly,

$$\left|\int_0^t X_i(s)^+ \, ds - \int_0^t Y_i(s)^+ \, ds\right| < \epsilon t/2 \le \epsilon T/2.$$

Next, for f_i, when $|X_1(t) - X_2(t)| \ge \epsilon$, if $X_1(t) > X_2(t)$, then $Y_1(t) > Y_2(t)$, and if $X_1(t) < X_2(t)$, then $Y_1(t) < Y_2(t)$. In this case, we have $|f_i(X(t)) - f_i(Y(t))| \le \epsilon/2$. If, instead, $|X_1(t) - X_2(t)| < \epsilon$, $|f_i(X(t)) - f_i(Y(t))| \le \beta_F/\sqrt{\mu}$. Putting the two cases together, we have

$$\begin{aligned}
&\left|\int_0^t f_i(X(s))\, ds - \int_0^t f_i(Y(s))\, ds\right| \\
\le &\frac{\epsilon}{2}\int_0^t 1\{|X_1(s) - X_2(s)| \ge \epsilon\}\, ds + \frac{\beta_F}{\sqrt{\mu}}\int_0^t 1\{|X_1(s) - X_2(s)| < \epsilon\}\, ds \\
\le &\frac{\epsilon T}{2} + \frac{\beta_F}{\sqrt{\mu}}\int_0^T 1\{|X_1(s) - X_2(s)| < \epsilon\}\, ds.
\end{aligned}$$

Above all,

$$\begin{aligned}
|F_i(X)(t) - F_i(Y)(t)| \le &\mu\left|\int_0^t X_i(s)^- \, ds - \int_0^t Y_i(s)^- \, ds\right| + \theta\left|\int_0^t X_i(s)^+ \, ds - \int_0^t Y_i(s)^+ \, ds\right| \\
&+ (\mu_F - \theta)\left|\int_0^t f_i(X(s))\, ds - \int_0^t f_i(Y(s))\, ds\right| \\
\le &\frac{\epsilon\mu T}{2} + \frac{\epsilon\theta T}{2} + \frac{\epsilon(\mu_F - \theta)T}{2} + \frac{\beta_F}{\sqrt{\mu}}(\mu_F - \theta)\int_0^T 1\{|X_1(t) - X_2(t)| < \epsilon\}\, dt \\
\to &\frac{\beta_F}{\sqrt{\mu}}(\mu_F - \theta)\int_0^T 1\{X_1(t) = X_2(t)\}\, dt \text{ as } \epsilon \downarrow 0.
\end{aligned}$$

This implies that to prove continuity of F at $\hat{X}$, it suffices to prove that $\mathbb{P}\left(\int_0^T 1\{\hat{X}_1(t)=\hat{X}_2(t)\}\,dt=0\right)=1$. Note that $\hat{X}^\lambda \Rightarrow \hat{X}$ implies that $\hat{X}$ takes the form

$$\hat{X}_i(t)=\hat{X}_i(0)+\sqrt{2}B_i(t)-\beta\sqrt{\mu}t+\mu\int_0^t \hat{X}_i(s)^-\,ds+\theta\int_0^t \hat{X}_i(s)^+\,ds-L_i(t),$$

where $L_i(t)$ is a weak limit of $\{(\mu_F-\theta)\int_0^t f_i(\hat{X}^\lambda(s))\,ds\}$. We also note that $L_i(t)$ is monotone increasing and bounded by $(\mu_F-\theta)\beta_F t/\sqrt{\mu}$. Thus, L_i has finite total variation. Meanwhile, since $\hat{X}$ is continuous, $\left\|\hat{X}\right\|_T<\infty$. As $\int_0^t \hat{X}_i(s)^-\,ds\le\int_0^T \hat{X}_i(s)^-\,ds<\infty$, $\mu\int_0^t \hat{X}_i(s)^-\,ds$ has finite total variation as well. Similarly, $\theta\int_0^t \hat{X}_i(s)^+\,ds$ has finite total variation as well. It then follows that $\hat{X}(t)$ is the sum of a Brownian motion and other terms of finite total variation. Therefore $\hat{X}$ spends almost surely zero time on $\{\hat{X}_1(s)=\hat{X}_2(s)\}$ (Turner 2000). $\square$

Step 5. Establish that $\hat{X}$ is suitably well-posed.

The following lemma follows directly from Proposition 5.3.10 in Karatzas and Shreve (1998).

LEMMA 17. *The diffusion equation*

$$\hat{X}_i(t)=\hat{X}_i(0)+\sqrt{2}B_i(t)-\beta\sqrt{\mu}t+\mu\int_0^t \hat{X}_i(s)^-\,ds-(\mu_F-\theta)\int_0^t f_i(\hat{X}(s))\,ds-\theta\int_0^t \hat{X}_i(s)^+\,ds$$

has a unique (weak) solution.

Steps 1-5 together establish the process level convergence of $\hat{X}^\lambda$, i.e.,

$$\hat{X}^\lambda \Rightarrow \hat{X} \text{ in } D^2 \text{ as } \lambda\to\infty.$$

We also note that

$$\hat{Q}^\lambda_\Sigma(t)=\left(\hat{X}^\lambda_1(t)^+ + \hat{X}^\lambda_2(t)^+ - n^\lambda_F/\sqrt{\lambda}\right)^+ = \left(\hat{X}^\lambda_1(t)^+ + \hat{X}^\lambda_2(t)^+ - \beta_F/\sqrt{\mu}\right)^+ + g^\lambda(\hat{X}^\lambda_1(t),\hat{X}^\lambda_2(t))$$

where

$$\begin{aligned}\left|g^\lambda(\hat{X}^\lambda_1(t),\hat{X}^\lambda_2(t))\right| &= \left|\left(\hat{X}^\lambda_1(t)^+ + \hat{X}^\lambda_2(t)^+ - n^\lambda_F/\sqrt{\lambda}\right)^+ - \left(\hat{X}^\lambda_1(t)^+ + \hat{X}^\lambda_2(t)^+ - \beta_F/\sqrt{\mu}\right)^+\right| \\ &\le |n^\lambda_F/\sqrt{\lambda}-\beta_F/\sqrt{\mu}|\to 0 \text{ as } \lambda\to\infty.\end{aligned}$$

This implies that $\hat{Q}^\lambda_\Sigma \Rightarrow \left(\hat{X}^+_1+\hat{X}^+_2-\beta_F/\sqrt{\mu}\right)^+$ in D as $\lambda\to\infty$.

Step 6. Establish the appropriate interchange of limits and uniform integrability results.

LEMMA 18. *For $(\beta,\beta_F)\in\hat{\Omega}(\theta)$, the diffusion process $\hat{X}$ is positive recurrent.*

Proof of Lemma 18. We will show that the function $V(x_1,x_2)=\frac{1}{2}(x_1^2+x_2^2)$ is a Lyapunov function. The generator G of $\hat{X}$ applied to V is given by

$$GV(x)=\sum_{i=1}^2 x_i\left(-\beta\sqrt{\mu}+\mu x_i^- - \theta x_i^+ - (\mu_F-\theta)f_i(x)\right)$$

for $x\in\mathbb{R}^2$.

We first consider the case $\theta>0$. Because f_i is bounded (by $\beta_F/\sqrt{\mu}$), we have that $-\beta\sqrt{\mu}+\mu x_i^- - \theta x_i^+ - (\mu_F-\theta)f_i(x)\le -1$ for all $x_i>0$ large enough, and $-\beta\sqrt{\mu}+\mu x_i^- - \theta x_i^+ - (\mu_F-\theta)f_i(x)\ge 1$ for all $-x_i>0$ large enough. It follows that $GV(x)\le -1$ for all $|x|$ large enough.

Suppose instead $\theta = 0$. If $\beta > 0$, $-\beta\sqrt{\mu} + \mu x_i^- - \mu_F f_i(x) \le -\beta\sqrt{\mu} < 0$ for all $x_i > 0$, and $-\beta\sqrt{\mu} + \mu x_i^- - \mu_F f_i(x) \ge 1$ for all $-x_i > 0$ large enough. Thus we may suppose $\beta \le 0$.

Suppose first both x_i are non-negative, with $x_1 \ge x_2 \ge 0$ (the case $x_2 > x_1 \ge 0$ is similar). Then, if $x_1 \ge \beta_F/\sqrt{\mu}$,

$$GV(x) = x_1(-\beta\sqrt{\mu} - \mu_F\beta_F/\sqrt{\mu}) - x_2\beta\sqrt{\mu} \le \frac{-x_1}{\sqrt{\mu}}(2\beta\mu + \beta_F\mu_F) \le -1$$

for x_1 large enough, since $2\beta\mu + \beta_F\mu_F > 0$.

Next, suppose exactly one x_i is non-negative, with $x_1 \ge 0 > x_2$ (the case $x_2 \ge 0 > x_1$ is similar). We have, if $x_1 \ge \beta_F/\sqrt{\mu}$,

$$GV(x) = x_1(-\beta\sqrt{\mu} - \mu_F f_i(x)) - \mu x_2^2 - \beta\sqrt{\mu}x_2 \le -\frac{x_1}{\sqrt{\mu}}(\beta\mu + \beta_F\mu_F) - \mu x_2^2 \le -\frac{x_1}{\sqrt{\mu}}(2\beta\mu + \beta_F\mu_F) - \mu x_2^2 \le -1$$

for $|x|$ large enough, since $2\beta\mu + \beta_F\mu_F > 0$. If instead $0 \le x_1 < \beta_F/\sqrt{\mu}$, we have that $x_1(-\beta\sqrt{\mu} - \mu_F f_i(x))$ is bounded, so again $GV(x) \le -1$ for $|x|$ large enough.

Finally, suppose $x_i < 0$ for $i = 1, 2$. We have

$$GV(x) = \sum_{i=1}^{2} x_i(-\beta\sqrt{\mu} - \mu x_i) \le -1$$

for $|x|$ large enough. This completes the proof. □

Lemma 18 implies that $\hat{X}(\infty)$ is well defined.

LEMMA 19. *Suppose* $n^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ *and* $n_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$, *with* $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$ *and* $(\beta, \beta_F) \in \hat{\Omega}(\theta)$. *Then,*

$$\hat{Q}_\Sigma^\lambda(\infty) \Rightarrow \left(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu}\right)^+ \text{ as } \lambda \to \infty$$

and

$$\mathbb{E}[\hat{Q}_\Sigma^\lambda(\infty)] \to \mathbb{E}\left[\left(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu}\right)^+\right] \text{ as } \lambda \to \infty.$$

Proof of Lemma 19. Note that

$$\begin{aligned}
&\sup_{\lambda>1} \mathbb{E}[(\hat{X}_i^\lambda(\infty)^+)^2] \\
&= \sup_{\lambda>1} \mathbb{E}\left[\left(\frac{(X_i^\lambda(\infty) - n^\lambda)^+}{\sqrt{\lambda}}\right)^2\right] \\
&\le \sup_{\lambda>1} \mathbb{E}\left[\left(\frac{\sum_{j=1}^2 (X_j^\lambda(\infty) - n^\lambda)^+}{\sqrt{\lambda}}\right)^2\right] \\
&\le \sup_{\lambda>1} \mathbb{E}\left[\left(\frac{\sum_{j=1}^2 \left((X_j^\lambda(\infty) - n^\lambda - n_F^\lambda/2)^+ + n_F^\lambda/2\right)}{\sqrt{\lambda}}\right)^2\right] \\
&\le \sup_{\lambda>1} 4\mathbb{E}\left[\left(\frac{(\tilde{X}_1^\lambda(\infty) - n^\lambda - n_F^\lambda/2)^+ + n_F^\lambda/2}{\sqrt{\lambda}}\right)^2\right] \text{ by Lemma 9 and Cauchy-Schwarz Inequality} \\
&< \infty \text{ by Lemma 11.}
\end{aligned} \tag{21}$$

In addition,

$$\sup_{\lambda>1} \mathbb{E}[\hat{X}_i^\lambda(\infty)^-] = \sup_{\lambda>1} \mathbb{E}\left[\frac{(X_i^\lambda(\infty) - n^\lambda)^-}{\sqrt{\lambda}}\right] < \infty$$

by Lemma 12. Then we have $\sup_{\lambda>1}\mathbb{E}[|\hat{X}_i^\lambda(\infty)|]<\infty$, i.e., $\{\hat{X}^\lambda(\infty):\lambda>1\}$ is tight. Thus, $\hat{X}^\lambda(\infty)\Rightarrow\hat{X}(\infty)$ as $\lambda\to\infty$ by Lemma 10. By the continuous mapping and converging together theorems, we have $\hat{Q}_\Sigma^\lambda(\infty)\Rightarrow\left(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu}\right)^+$ as $\lambda\to\infty$.

Next, the bound in (21) also implies that $\{\hat{X}_i^\lambda(\infty)^+:\lambda>1\}$ is uniformly integrable. As $\hat{Q}_\Sigma^\lambda(\infty)\le\hat{X}_1^\lambda(\infty)^+ + \hat{X}_2^\lambda(\infty)^+$, $\{\hat{Q}_\Sigma^\lambda(\infty):\lambda>1\}$ is also uniformly integrable. Thus,

$$\mathbb{E}[\hat{Q}_\Sigma^\lambda(\infty)]\to\mathbb{E}\left[\left(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu}\right)^+\right]\text{ as }\lambda\to\infty.$$

□

This concludes the proof of Theorem 2.

C.4. Proof of Theorem 3.

We first prove the 'only if' part. Let (n^λ, n_F^λ) be asymptotically optimal, and suppose for contradiction that it is not of the form stated in the theorem. That is, there exists $\epsilon>0$ and a subsequence, which we index again by λ for convenience, satisfying

$$\min_{(a,b)\in\arg\min_{\beta,\beta_F}\hat{V}_p(\beta,\beta_F)}\frac{\left|n^\lambda - R^\lambda - a\sqrt{R^\lambda}\right| + \left|n_F^\lambda - b\sqrt{R^\lambda}\right|}{\sqrt{R^\lambda}}>\epsilon$$

for each λ. This subsequence is asymptotically optimal, and so it follows from the proof of Lemma 2 that

$$n_F^\lambda = b_\lambda\sqrt{R^\lambda}+o(\sqrt{R^\lambda})\text{ and }n^\lambda = R^\lambda + a_\lambda\sqrt{R^\lambda}+o(\sqrt{R^\lambda})$$

for some bounded sequences $\{a_\lambda\}$ and $\{b_\lambda\}$. Then, there exist finite constants $(a,b)\notin\arg\min_{\beta,\beta_F}\hat{V}_p(\beta,\beta_F)$ and a subsequence indexed by λ', such that

$$a_{\lambda'}\to a\text{ and }b_{\lambda'}\to b\text{ as }\lambda'\to\infty.$$

For the ease of notation, we re-index this subsequence by λ. As $(a,b)\notin\arg\min_{\beta,\beta_F}\hat{V}_p(\beta,\beta_F)$, there exists (β,β_F) such that $\hat{V}_p(\beta,\beta_F)<\hat{V}_p(a,b)$. Define

$$\bar{n}_F^\lambda = \beta_F\sqrt{R^\lambda}+o(\sqrt{R^\lambda})\text{ and }\bar{n}^\lambda = R^\lambda+\beta\sqrt{R^\lambda}+o(\sqrt{R^\lambda}).$$

Then,

$$\begin{aligned}
&\limsup_{\lambda\to\infty}\frac{\Pi^\lambda(n^\lambda,n_F^\lambda)-\Pi^{\lambda,*}}{\sqrt{\lambda}}\\
\ge&\limsup_{\lambda\to\infty}\frac{\Pi^\lambda(n^\lambda,n_F^\lambda)-\Pi^\lambda(\bar{n}^\lambda,\bar{n}_F^\lambda)}{\sqrt{\lambda}}\\
=&\limsup_{\lambda\to\infty}\frac{2c(n^\lambda-R^\lambda)+c_F n_F^\lambda+(h+a\theta)\mathbb{E}[Q_\Sigma^\lambda(\infty;n^\lambda,n_F^\lambda)]-(2c(\bar{n}^\lambda-R^\lambda)+c_F\bar{n}_F^\lambda+(h+a\theta)\mathbb{E}[Q_\Sigma^\lambda(\infty;\bar{n}^\lambda,\bar{n}_F^\lambda)])}{\sqrt{\lambda}}\\
=&\hat{V}_p(a,b)-\hat{V}_p(\beta,\beta_F)>0
\end{aligned}$$

where the last equality follows from Theorem 2, contradicting asymptotic optimality.

It remains to prove the 'if' part. From the proof of the 'only if' part, the sequence of optimal staffing levels $(n^{\lambda,*},n_F^{\lambda,*})$ satisfy

$$n_F^{\lambda,*}=d_\lambda\sqrt{R^\lambda}+o(\sqrt{R^\lambda})\text{ and }n^{\lambda,*}=R^\lambda+c_\lambda\sqrt{R^\lambda}+o(\sqrt{R^\lambda})$$

for some $(c_\lambda, d_\lambda) \in \arg\min_{\beta,\beta_F} \hat{V}_p(\beta,\beta_F)$. Next, consider any sequence

$$n_F^\lambda = b_\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda}) \text{ and } n^\lambda = R^\lambda + a_\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$$

where $(a_\lambda, b_\lambda) \in \arg\min_{\beta,\beta_F} \hat{V}_p(\beta,\beta_F)$. Then,

$$\begin{aligned}
&\limsup_{\lambda\to\infty} \frac{\Pi^\lambda(n^\lambda, n_F^\lambda) - \Pi^{\lambda,*}}{\sqrt{\lambda}} \\
=&\limsup_{\lambda\to\infty} \frac{\Pi^\lambda(n^\lambda, n_F^\lambda) - \Pi^\lambda(n^{\lambda,*}, n_F^{\lambda,*})}{\sqrt{\lambda}} \\
=&\limsup_{\lambda\to\infty} \frac{2c(n^\lambda - R^\lambda) + c_F n_F^\lambda + (h+a\theta)\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda)] - \left(2c(n^{\lambda,*} - R^\lambda) + c_F n_F^{\lambda,*} + (h+a\theta)\mathbb{E}[Q_\Sigma^\lambda(\infty; n^{\lambda,*}, n_F^{\lambda,*})]\right)}{\sqrt{\lambda}} \qquad (22) \\
=&\hat{V}_p^* - \hat{V}_p^* = 0,
\end{aligned}$$

where $\hat{V}_p^* = \min_{\beta,\beta_F} \hat{V}_p(\beta,\beta_F)$. To see (22), note that by Theorem 2, for any $(a,b) \in \arg\min_{\beta,\beta_F} \hat{V}_p(\beta,\beta_F)$,

$$\frac{2ca\sqrt{R^\lambda} + c_F b\sqrt{R^\lambda} + o(\sqrt{R^\lambda})}{\sqrt{\lambda}} + (h+a\theta)\mathbb{E}[\hat{Q}_\Sigma^\lambda(\infty; a, b)] = \hat{V}_p^* + o(1).$$

Then, (22) follows because $\arg\min_{\beta,\beta_F} \hat{V}_p(\beta,\beta_F)$ is finite under Assumption 1. □

Appendix D: Proofs of the Results in Section 4

For $x, y \in \mathbb{R}$ and $z \geq 0$, define

$$K_\lambda(x,y,z) = \tilde{\Pi}^\lambda\left(\frac{p_1\lambda + x\lambda^{\alpha_1}}{\mu}, \frac{p_2\lambda + y\lambda^{\alpha_2}}{\mu}, \frac{z\lambda^{\alpha_2}}{\mu_F}\right).$$

D.1. Proof of Lemma 4.

In this case,

$$K_\lambda(x,y,z) = c(p_1+p_2)R^\lambda + \lambda^\alpha\left(\frac{c}{\mu}x + \frac{c}{\mu}y + \frac{c_F}{\mu_F}z\right) + c_P\lambda^\alpha\mathbb{E}\left[\left((Y_1-x)^+ + (Y_2-y)^+ - z\right)^+\right].$$

In the first case, note that $K_\lambda(x,y,z)$ is convex and

$$\nabla K_\lambda(q_1,q_2,0) = \lambda^\alpha\left(0, 0, \frac{c_F}{\mu_F} - c_P\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2)\right).$$

As $\frac{c_F}{\mu_F} - c_P\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) \geq 0$, $(q_1, q_2, 0)$ is optimal.

In the second case, we have

$$\nabla K_\lambda(r_1, r_2, r_F) = (0,0,0).$$

The optimality of (r_1, r_2, r_F) follows due to the convexity of $K_\lambda(x,y,z)$. □

D.2. Proof of Lemma 5.

In this case,

$$\begin{aligned}
K_\lambda(x,y,z) =& c(p_1+p_2)R^\lambda + \lambda^{\alpha_1}\frac{c}{\mu}x + \lambda^{\alpha_2}\left(\frac{c}{\mu}y + \frac{c_F}{\mu_F}z\right) \\
&+ c_P\lambda^{\alpha_1}\mathbb{E}\left[\left((Y_1-x)^+ + \lambda^{\alpha_2-\alpha_1}(Y_2-y)^+ - \lambda^{\alpha_2-\alpha_1}z\right)^+\right].
\end{aligned}$$

Let $(x_\lambda^*, y_\lambda^*, z_\lambda^*)$ be the minimizer of K_λ.

We first show that $x^*_\lambda = q_1 + o(1)$. Note that $K^*_\lambda := \tilde{\Pi}^{\lambda,*} \le K_\lambda(0,0,0) = c(p_1+p_2)R^\lambda + O(\lambda^{\alpha_1})$. Since $K_\lambda(x,y,z) \ge c(p_1+p_2)R^\lambda + \lambda^{\alpha_1}\frac{c}{\mu}x$, we have that $x^{*+}_\lambda = O(1)$.

Now suppose for contradiction that there exists a subsequence, indexed by λ_k, such that either i) $x^*_{\lambda_k} \to -\infty$ or ii) $x^*_{\lambda_k} \to C \in \mathbb{R}\backslash\{q_1\}$. Note that

$$\begin{aligned} K_\lambda(x,y,z) &\ge c(p_1+p_2)R^\lambda + \lambda^{\alpha_1}\frac{c}{\mu}x + \lambda^{\alpha_2}\frac{c_F}{\mu_F}z + c_P\lambda^{\alpha_1}\mathbb{E}[(Y_1-x)^+ - \lambda^{\alpha_2-\alpha_1}z)^+] \\ &= c(p_1+p_2)R^\lambda + \lambda^{\alpha_1}\frac{c}{\mu}x + \lambda^{\alpha_2}\frac{c_F}{\mu_F}z + c_P\lambda^{\alpha_1}\mathbb{E}[(Y_1-x-\lambda^{\alpha_2-\alpha_1}z)^+] \\ &= c(p_1+p_2)R^\lambda + \lambda^{\alpha_1}\frac{c}{\mu}(x+\lambda^{\alpha_2-\alpha_1}z) + c_P\lambda^{\alpha_1}\mathbb{E}[(Y_1-x-\lambda^{\alpha_2-\alpha_1}z)^+] + \lambda^{\alpha_2}\left(\frac{c_F}{\mu_F}-\frac{c}{\mu}\right)z. \quad (23) \end{aligned}$$

First suppose that $x^*_{\lambda_k} \to -\infty$. Since $c_P > c_F/\mu_F > c/\mu$ and $K^*_\lambda \le c(p_1+p_2)R^\lambda + O(\lambda^{\alpha_1})$, it follows that $(x^*_\lambda + \lambda^{\alpha_2-\alpha_1}z^*_\lambda)^- = O(1)$. This in turn implies that $\lambda^{\alpha_2-\alpha_1}z^*_\lambda \to \infty$, so that $\lambda^{\alpha_2}\left(\frac{c_F}{\mu_F}-\frac{c}{\mu}\right)z^*_\lambda$ grows to infinity faster than $O(\lambda^{\alpha_1})$. Then, the second and third terms of the last equation (23) will be $O(\lambda^{\alpha_1})$, while the last is of a larger order. This contradicts that $(x^*_\lambda, y^*_\lambda, z^*_\lambda)$ is optimal.

Consider the second case $x^*_\lambda \to C \in \mathbb{R}\backslash\{q_1\}$. Note that

$$K_\lambda(q_1,0,0) = c(p_1+p_2)R^\lambda + \lambda^{\alpha_1}f(q_1)$$

where $f(x) = \frac{c}{\mu}x + c_P\mathbb{E}[(Y_1-x)^+]$. From (23), we have that

$$K_\lambda(x^*_\lambda, y^*_\lambda, z^*_\lambda) \ge c(p_1+p_2)R^\lambda + \lambda^{\alpha_1}f(x^*_\lambda + \lambda^{\alpha_2-\alpha_1}z^*_\lambda) + \lambda^{\alpha_2}\left(\frac{c_F}{\mu_F}-\frac{c}{\mu}\right)z^*_\lambda.$$

Since q_1 is uniquely optimal for f and $x^*_\lambda \to C$, we must have $\lambda^{\alpha_2-\alpha_1}z^*_\lambda \to q_1 - C > 0$. But then $\lambda^{\alpha_2}\left(\frac{c_F}{\mu_F}-\frac{c}{\mu}\right)z^*_\lambda \ne o(\lambda^{\alpha_1})$, contradicting optimality.

This completes the proof that $x^*_\lambda = q_1 + o(1)$. Next, we prove that y^*_λ and z^*_λ are of the appropriate form. We first show they are $O(1)$. Consider the partial derivatives

$$\frac{\partial K_\lambda}{\partial y} = \lambda^{\alpha_2}\left(\frac{c}{\mu} - c_P\mathbb{P}[Y_2 > y, \lambda^{\alpha_1-\alpha_2}(Y_1-x)^+ + Y_2 - y > z]\right)$$

and

$$\frac{\partial K_\lambda}{\partial z} = \lambda^{\alpha_2}\left(\frac{c_F}{\mu_F} - c_P\mathbb{P}[\lambda^{\alpha_1-\alpha_2}(Y_1-x)^+ + (Y_2-y)^+ > z]\right).$$

By optimality, we have

$$0 < \frac{c}{c_P\mu} = \mathbb{P}[Y_2 > y^*_\lambda, \lambda^{\alpha_1-\alpha_2}(Y_1-x^*_\lambda)^+ + Y_2 - y^*_\lambda > z^*_\lambda] \le \mathbb{P}[Y_2 > y^*_\lambda]$$

which implies that $y^{*+}_\lambda = O(1)$. If $y^{*-}_\lambda \ne O(1)$, then there is a subsequence (re-indexed by λ) such that $y^*_\lambda \to -\infty$, which implies that

$$\begin{aligned} 1 > \frac{c}{c_P\mu} = \mathbb{P}[Y_2 > y^*_\lambda, \lambda^{\alpha_1-\alpha_2}(Y_1-x^*_\lambda)^+ + Y_2 - y^*_\lambda > z^*_\lambda] &= \mathbb{P}[\lambda^{\alpha_1-\alpha_2}(Y_1-x^*_\lambda)^+ + Y_2 - y^*_\lambda > z^*_\lambda] + o(1) \\ &\ge \mathbb{P}[Y_2 > y^*_\lambda + z^*_\lambda] + o(1). \end{aligned}$$

This in turn implies $z^*_\lambda \to \infty$, and in particular $z^*_\lambda > 0$. But then

$$\mathbb{P}[\lambda^{\alpha_1-\alpha_2}(Y_1-x^*_\lambda)^+ + (Y_2-y^*_\lambda)^+ > z^*_\lambda] = \mathbb{P}[Y_2 > y^*_\lambda, \lambda^{\alpha_1-\alpha_2}(Y_1-x^*_\lambda)^+ + Y_2 - y^*_\lambda > z^*_\lambda] + o(1) = \frac{c}{c_P\mu} + o(1) < \frac{c_F}{c_P\mu_F}$$

so that $\frac{\partial K_\lambda}{\partial z}(x^*_\lambda, y^*_\lambda, z^*_\lambda) > 0$, contradicting optimality. Hence, $y^*_\lambda = O(1)$.

Next, we show $z^*_\lambda = O(1)$. If not, we can obtain a subsequence indexed again by λ such that $z^*_\lambda \to \infty$, and in particular $z^*_\lambda > 0$. Since $y^*_\lambda = O(1)$ and $x^*_\lambda = q_1 + o(1)$, we have

$$\mathbb{P}[\lambda^{\alpha_1-\alpha_2}(Y_1 - x^*_\lambda)^+ + (Y_2 - y^*_\lambda)^+ > z^*_\lambda] = \mathbb{P}[\lambda^{\alpha_1-\alpha_2}(Y_1 - q_1)^+ > z^*_\lambda] + o(1) \le \mathbb{P}[Y_1 > q_1] + o(1) = \frac{c}{c_P\mu} + o(1)$$

so that $\frac{\partial K_\lambda}{\partial z}(x^*_\lambda, y^*_\lambda, z^*_\lambda) > 0$ contradicting optimality. Thus $z^*_\lambda = O(1)$.

Finally, we show that y^*_λ and z^*_λ have the right asymptotics. Suppose $\tilde{n}^*_2$ and $\tilde{n}^*_F$ are not of the specified form. First suppose $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) > \frac{c_F}{c_P\mu_F}$. Then, there is a subsequence re-indexed by λ such that $x^*_\lambda \to q_1$, $y^*_\lambda \to D \in \mathbb{R}$ and $z^*_\lambda \to E \ge 0$, where either $D \ne l$ or $E \ne l_F$. Note that

$$\begin{aligned}\mathbb{P}[Y_2 > y^*_\lambda, \lambda^{\alpha_1-\alpha_2}(Y_1 - x^*_\lambda)^+ + Y_2 - y^*_\lambda > z^*_\lambda] &= \mathbb{P}[Y_2 > D, Y_1 > q_1 \text{ or } Y_2 > D + E] + o(1)\\ &= \mathbb{P}[Y_2 > D + E \text{ or } (Y_2 > D, Y_1 > q_1)] + o(1)\end{aligned}$$

and

$$\mathbb{P}[\lambda^{\alpha_1-\alpha_2}(Y_1 - x^*_\lambda)^+ + (Y_2 - y^*_\lambda)^+ > z^*_\lambda] = \mathbb{P}[Y_1 > q_1 \text{ or } Y_2 > D + E] + o(1).$$

By optimality, we must have either (i) the first probability is $\frac{c}{c_P\mu}$ and the second is $\frac{c_F}{c_P\mu_F}$ or (ii) the first probability is $\frac{c}{c_P\mu}$, the second is $\le \frac{c_F}{c_P\mu_F}$ and $E = 0$. Case (i) is ruled out by the uniqueness of l and l_F. If (ii), then $D > q_2$ in order for the second probability to be $\le \frac{c_F}{c_P\mu_F}$, but $D = q_2$ is necessary for the first probability to be $\frac{c}{c_P\mu}$. This is a contradiction and thus $y^*_\lambda = l + o(1)$ and $z^*_\lambda = l_F + o(1)$.

We now turn to the other case $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) < \frac{c_F}{c_P\mu_F}$. Using the previous notation, we can again obtain a subsequence such that either $D \ne q_2$ or $E > 0$. The case $E = 0$ and $D \ne q_2$ can be ruled out by our previous discussion, so suppose $E > 0$. By optimality, we must have that

$$\mathbb{P}[Y_2 > D + E \text{ or } (Y_2 > D, Y_1 > q_1)] = \frac{c}{c_P\mu}$$

and

$$\mathbb{P}[Y_1 > q_1 \text{ or } Y_2 > D + E] = \frac{c_F}{c_P\mu_F}.$$

The second equation ensures $D + E < q_2$. Then, the first probability is at least $\mathbb{P}(Y_2 > D + E) > \mathbb{P}(Y_2 > q_2) = \frac{c}{c_P\mu}$, a contradiction. This completes the proof. □

D.3. Two auxiliary lemmas

Note that for any given arrival rate realization, under the scheduling policy $\tilde{\nu}^\lambda$, the two-class queue can be decomposed into two independent single-class queues with two types of servers: the high-priority rate-μ servers and the low-priority rate-μ_F servers. In this section, we study the single-class queue with arrival rate γ, n high-priority rate-μ servers, n_F low-priority rate-μ_F servers, and abandonment rate $\theta = \mu_F$. To simplify the notation, we denote by X the steady-state number of customers in the system, Q the steady-state number of customers waiting in queue, Z the steady-state number of customers in service with rate-μ servers, and Z_F the steady-state number of customers in service with rate-μ_F servers.

LEMMA 20. *For the single-class queue with two types of servers and $\theta = \mu_F$, there are universal constants $K_1, K_2, K_3 > 0$ (i.e. not depending on n, n_F or γ), such that*

$$\theta\mathbb{E}[Q] \le (\gamma - n\mu - n_F\mu_F)^+ + K_1\sqrt{\gamma}\exp\left(-\frac{K_2}{\gamma}(\gamma - n\mu - n_F\mu_F)^2\right) + K_3.$$

Proof of Lemma 20. We start by showing that

$$\mathbb{E}[X] = \gamma/\theta - n\mu/\theta + n + (\mu-\theta)\mathbb{E}[(n-X)^+]/\theta. \tag{24}$$

Let $\xi(x)$ denote the death rate at state x. When $x > n$, $\xi(x) = x\theta + n(\mu-\theta)$; when $x \le n$, $\xi(x) = x\mu = x\theta + x(\mu-\theta)$. Equating the birth rate and the death rate in stationarity, we have

$$\gamma = \mathbb{E}[X\theta + n(\mu-\theta) - (n-X)^+(\mu-\theta)],$$

which implies (24).

First, consider the case where $\gamma \ge n\mu + n_F\mu_F$. We have

$$\begin{aligned}
\mathbb{E}[Q] &= \mathbb{E}[(X - n - n_F)^+] \\
&= \mathbb{E}[X] - n - n_F + \sum_{x=0}^{n+n_F-1} \mathbb{P}[X \le x] \\
&= \gamma/\theta - n\mu/\theta + n + (\mu-\theta)\mathbb{E}[(n-X)^+]/\theta - n - n_F + \sum_{x=0}^{n+n_F-1} \mathbb{P}[X \le x] \text{ by (24)} \\
&= (\gamma/\theta - n\mu/\theta - n_F) + \frac{\mu-\theta}{\theta}\sum_{x=0}^{n-1}\mathbb{P}[X \le x] + \sum_{x=0}^{n+n_F-1} \mathbb{P}[X \le x].
\end{aligned}$$

It suffices to bound $\sum_{x=0}^{n+n_F-1}\mathbb{P}[X \le x]$. Let $M_1 = \lfloor \gamma/\theta - n\mu/\theta + n \rfloor$, which is the mode of X (to see this, note that the death rate $\xi(x)$ is increasing in x, and that the birth rate γ is larger than the death rate for $x < M_1$, and is smaller than the death rate for $x > M_1$). We then have $M_1 \ge n + n_F$, and $\xi(M_1) = n\mu + \lfloor \gamma/\theta - n\mu/\theta \rfloor\theta \le \gamma$. For $0 < k \le M_1 - n + 1$, note that

$$\begin{aligned}
\mathbb{P}[X = M_1 - k] &= \mathbb{P}[X = M_1] \cdot \frac{\xi(M_1)(\xi(M_1)-\theta)\cdots(\xi(M_1)-(k-1)\theta)}{\gamma^k} \\
&\le \mathbb{P}[X = M_1] \cdot \frac{\gamma(\gamma-\theta)\cdots(\gamma-(k-1)\theta)}{\gamma^k} \\
&= \mathbb{P}[X = M_1] \cdot 1(1-\theta/\gamma)\cdots(1-(k-1)\theta/\gamma) \\
&\le \mathbb{P}[X = M_1] \cdot (1-(k-1)\theta/2\gamma)^k \\
&\le \mathbb{P}[X = M_1] \cdot \exp(-(k-1)k\theta/2\gamma) \text{ as } 1 - x \le \exp(-x).
\end{aligned}$$

Similarly, for $M_1 - n + 1 < k \le M_1$,

$$\begin{aligned}
&\mathbb{P}[X = M_1 - k] \\
=&\mathbb{P}[X = M_1] \cdot \frac{1}{\gamma^k}\xi(M_1)(\xi(M_1)-\theta)\cdots(\xi(M_1)-(M_1-n)\theta)\cdot(\xi(M_1)-(M_1-n)\theta-\mu) \\
&\cdots(\xi(M_1)-(M_1-n)\theta-(k-M_1+n-1)\mu) \\
\le&\mathbb{P}[X = M_1] \cdot \frac{\xi(M_1)(\xi(M_1)-\theta)\cdots(\xi(M_1)-(M_1-n)\theta)\cdot(\xi(M_1)-(M_1-n+1)\theta)\cdots(\xi(M_1)-(k-1)\theta)}{\gamma^k} \\
\le&\mathbb{P}[X = M_1] \cdot (1-(k-1)\theta/2\gamma)^k \\
\le&\mathbb{P}[X = M_1] \cdot \exp(-(k-1)k\theta/2\gamma).
\end{aligned}$$

Choose $A_1 > 0$ such that $A_1 k^2 \le k(k-1)\theta/2$ for all $k \ge 2$, then for any $1 < k \le M_1$

$$\mathbb{P}[X = M_1 - k] \le \mathbb{P}[X = M_1] \cdot \exp\bigl(-A_1 k^2/\gamma\bigr).$$

Next, we bound $\mathbb{P}[X = M_1]$. Note that for $k > 0$,

$$\mathbb{P}[X = M_1 + k] = \mathbb{P}[X = M_1] \cdot \frac{\gamma^k}{(\xi(M_1)+\theta)\cdots(\xi(M_1)+k\theta)}.$$

Then, because

$$\begin{aligned}
\frac{\gamma^{\lfloor\sqrt{\gamma}\rfloor}}{(\xi(M_1)+\theta)\cdots(\xi(M_1)+\lfloor\sqrt{\gamma}\rfloor\theta)} &\ge \frac{\gamma^{\lfloor\sqrt{\gamma}\rfloor}}{(\gamma+\theta)\cdots(\gamma+\lfloor\sqrt{\gamma}\rfloor\theta)} \\
&\ge \frac{\gamma^{\lfloor\sqrt{\gamma}\rfloor}}{(\gamma+\lfloor\sqrt{\gamma}\rfloor\theta)^{\lfloor\sqrt{\gamma}\rfloor}} \\
&= \left(1 - \frac{\lfloor\sqrt{\gamma}\rfloor\theta}{\gamma+\lfloor\sqrt{\gamma}\rfloor\theta}\right)^{\lfloor\sqrt{\gamma}\rfloor} \\
&\ge \left(1 - \frac{\lfloor\sqrt{\gamma}\rfloor\theta}{\gamma}\right)^{\lfloor\sqrt{\gamma}\rfloor} \to \exp(-\theta) \text{ as } \gamma \to \infty,
\end{aligned}$$

for γ large enough and $1 \le k \le \lfloor\sqrt{\gamma}\rfloor$,

$$\mathbb{P}[X = M_1 + k] \ge \mathbb{P}[X = M_1] \cdot \exp(-\theta)/2.$$

Thus, for γ large enough,

$$1 \ge \sum_{k=M_1}^{M_1+\lfloor\sqrt{\gamma}\rfloor} \mathbb{P}[X = k] \ge \mathbb{P}[X = M_1] \cdot (1 + \lfloor\sqrt{\gamma}\rfloor \exp(-\theta)/2)$$

which implies that

$$\mathbb{P}[X = M_1] \le \frac{1}{1+\lfloor\sqrt{\gamma}\rfloor\exp(-\theta)/2} \le \frac{B_1}{\sqrt{\gamma}},$$

where $B_1 = 2\exp(\theta)$.

Above all, we have proven that for γ large enough, when $1 < k \le M_1$,

$$\mathbb{P}[X = M_1 - k] \le B_1 \cdot \exp\bigl(-A_1 k^2/\gamma\bigr)/\sqrt{\gamma}.$$

Then, for $1 < k \le n + n_F$,

$$\begin{aligned}
\mathbb{P}[X \le n + n_F - k] &= \mathbb{P}[X \le M_1 - (M_1 - n - n_F + k)] \\
&= \sum_{j=M_1-n-n_F+k}^{M_1} \mathbb{P}[X = M_1 - j] \\
&\le \sum_{j=M_1-n-n_F+k}^{M_1} B_1 \exp\bigl(-A_1 j^2/\gamma\bigr)/\sqrt{\gamma} \\
&\le \int_{M_1-n-n_F+k-1}^{\infty} B_1 \exp\bigl(-A_1 j^2/\gamma\bigr)/\sqrt{\gamma}\, dj \\
&\le \frac{B_1\sqrt{2\pi}}{2\sqrt{A_1}} \exp\bigl(-A_1(M_1 - n - n_F + k - 1)^2/(2\gamma)\bigr) \text{ by Chernoff-Cramer bound.}
\end{aligned}$$

Choose $D_1 > 0$ such that $D_1 x^2 \le A_1(x-1)^2/2$ for all $x \ge 2$. Then, for γ large enough, and $2 \le k \le n + n_F$,

$$\mathbb{P}[X \le n + n_F - k] \le C_1 \exp\bigl(-D_1(M - n - n_F + k)^2/\gamma\bigr)$$

where $C_1 = \frac{B_1\sqrt{2\pi}}{2\sqrt{A_1}}$.

Finally, we have for γ large enough

$$\begin{aligned}
\sum_{x=0}^{n+n_F-1} \mathbb{P}[X \le x] &\le \int_0^{n+n_F} \mathbb{P}[X \le n+n_F-x]\,dx \\
&\le \int_2^{n+n_F} \mathbb{P}[X \le n+n_F-x]\,dx + 2 \\
&\le \int_0^{n+n_F} C_1 \exp\bigl(-D_1(M_1-n-n_F+x)^2/\gamma\bigr)\,dx + 2 \\
&= \int_{M_1-n-n_F}^{M_1} C_1 \exp\bigl(-D_1 x^2/\gamma\bigr)\,dx + 2 \\
&\le \frac{C_1\sqrt{2\pi\gamma}}{2\sqrt{D_1}} \exp\bigl(-D_1(M_1-n-n_F)^2/(2\gamma)\bigr) + 2.
\end{aligned}$$

Since $M_1 - n - n_F = \bigl\lfloor \frac{\gamma - n\mu - n_F\mu_F}{\theta} \bigr\rfloor$, this completes the proof for $\gamma > n\mu + n_F\mu_F$.

Next, consider the case where $n\mu \le \gamma < n\mu + n_F\mu_F$. Let $M_2 = n + \lfloor \frac{\gamma - n\mu}{\theta} \rfloor$ be the mode of X in this case. Note that

$$n + n_F - M_2 \ge \frac{n\mu + n_F\mu_F - \gamma}{\theta}.$$

Thus, for any $C > 0$,

$$\exp\bigl(-C\theta^2(n+n_F-M_2)^2\bigr) \le \exp\bigl(-C(n\mu+n_F\mu_F-\gamma)^2\bigr).$$

Next, note that

$$\mathbb{E}[Q] = \mathbb{E}[(X-n-n_F)^+] = \sum_{x=n+n_F}^{\infty} \mathbb{P}[X > x].$$

As $\xi(M_2) = n\mu + \lfloor \frac{\gamma-n\mu}{\theta} \rfloor \theta$, we have for $k > 0$,

$$\begin{aligned}
\mathbb{P}[X = M_2 + k] &= \mathbb{P}[X = M_2] \cdot \frac{\gamma^k}{(\xi(M_2)+\theta)\cdots(\xi(M_2)+k\theta)} \\
&\le \mathbb{P}[X = M_2] \cdot \frac{\gamma^k}{\gamma(\gamma+\theta)\cdots(\gamma+(k-1)\theta)} \\
&\le \mathbb{P}[X = M_2] \cdot \left(\frac{\gamma}{\gamma+(k-1)\theta/2}\right)^{k/2} \\
&= \mathbb{P}[X = M_2] \cdot \left(1 - \frac{(k-1)\theta/2}{\gamma+(k-1)\theta/2}\right)^{k/2} \\
&\le \mathbb{P}[X = M_2] \cdot \exp\left(-\frac{k(k-1)\theta/4}{\gamma+(k-1)\theta/2}\right) \\
&\le \mathbb{P}[X = M_2] \cdot \left(\exp\left(-\frac{k(k-1)\theta/4}{2\gamma}\right) + \exp\left(-\frac{k(k-1)\theta/4}{(k-1)\theta}\right)\right) \\
&= \mathbb{P}[X = M_2] \cdot \left(\exp\left(-\frac{k(k-1)\theta}{8\gamma}\right) + \exp\left(-k/4\right)\right).
\end{aligned}$$

The last inequality comes from the fact that $\frac{k(k-1)\theta/4}{\gamma+(k-1)\theta/2} \ge \frac{k(k-1)\theta/4}{2\gamma}$ if $\gamma \ge (k-1)\theta/2$ and $\frac{k(k-1)\theta/4}{\gamma+(k-1)\theta/2} \ge \frac{k(k-1)\theta/4}{(k-1)\theta}$ otherwise. Thus, for all $k > 0$,

$$\mathbb{P}[X \ge M_2 + k] = \sum_{j=k}^{\infty} \mathbb{P}[X = M_2 + j]$$

$$\leq \sum_{j=k}^{\infty} \mathbb{P}[X = M_2] \cdot \left(\exp\left(-\frac{j(j-1)\theta}{8\gamma} \right) + \exp\left(-j/4\right) \right)$$
$$\leq \mathbb{P}[X = M_2] \cdot \left(\int_{k-1}^{\infty} \exp\left(-\frac{j(j-1)\theta}{8\gamma} \right) dj + \frac{\exp(-k/4)}{1-\exp(-1/4)} \right).$$

Choose $B_2 > 0$ such that $B_2 x^2 \leq x(x-1)\theta/8$ for all $x \geq 2$, and choose $F_2 > 0$ such that $F_2 x^2 \leq B_2(x-1)^2/2$ for all $x \geq 3$. Also, choose $A_2 > 0$ such that $\mathbb{P}[X = M_2] \leq A_2/\sqrt{\gamma}$ (the existence of A_2 follows similarly to before.) Then, for all $k \geq 3$,

$$\begin{aligned}
\mathbb{P}[X \geq M_2 + k] &\leq \mathbb{P}[X = M_2] \cdot \left(\int_{k-1}^{\infty} \exp\left(-\frac{j(j-1)\theta}{8\gamma} \right) dj + \frac{\exp(-k/4)}{1-\exp(-1/4)} \right) \\
&\leq \frac{A_2}{\sqrt{\gamma}} \left(\int_{k-1}^{\infty} \exp\left(-B_2 j^2/\gamma\right) dj + \frac{\exp(-k/4)}{1-\exp(-1/4)} \right) \\
&\leq \frac{A_2\sqrt{2\pi}}{2\sqrt{B_2}} \cdot \exp\big(-B_2(k-1)^2/2\gamma\big) + C_2 \exp(-D_2 k)/\sqrt{\gamma} \\
&\leq E_2 \exp\big(-F_2 k^2/\gamma\big) + C_2 \exp(-D_2 k)/\sqrt{\gamma}
\end{aligned}$$

for some universal $C_2, D_2, E_2 > 0$. Finally,

$$\begin{aligned}
\sum_{x=n+n_F}^{\infty} \mathbb{P}[X > x] &= \int_{n+n_F-M_2}^{\infty} \mathbb{P}[X > M_2 + x]\, dx \\
&\leq 3 + \int_{n+n_F-M_2+3}^{\infty} E_2 \exp\big(-F_2 x^2/\gamma\big) + C_2 \exp(-D_2 x)/\sqrt{\gamma}\, dx \\
&\leq 3 + \int_{n+n_F-M_2}^{\infty} E_2 \exp\big(-F_2 x^2/\gamma\big) + C_2 \exp(-D_2 x)/\sqrt{\gamma}\, dx \\
&\leq 3 + L_2\sqrt{\gamma} \exp\big(-F_2(n+n_F-M_2)^2/2\gamma\big) + G_2 \exp(-D_2(n+n_F-M_2))/\sqrt{\gamma} \\
&\leq L_2\sqrt{\gamma} \exp\big(-F_2(n+n_F-M_2)^2/2\gamma\big) + 3 + G_2
\end{aligned}$$

for some universal $L_2, G_2 > 0$.

Lastly, consider the case where $0 < \gamma < n\mu$. This is very similar to the proof of the case $n\mu \leq \gamma < n\mu + n_F\mu_F$, but we include it here for completeness. Let $M_3 = \lfloor \gamma/\mu \rfloor$ be the mode of X in this case. Note that

$$n + n_F - M_3 \geq \frac{n\mu + n_F\mu - \gamma}{\mu} \geq \frac{n\mu + n_F\mu_F - \gamma}{\mu}.$$

Thus, for any $C > 0$,

$$\exp\big(-C\mu^2(n+n_F-M_3)^2\big) \leq \exp\big(-C(n\mu + n_F\mu_F - \gamma)^2\big).$$

Next, note that

$$\mathbb{E}[Q] = \mathbb{E}[(X-n-n_F)^+] = \sum_{x=n+n_F}^{\infty} \mathbb{P}[X > x].$$

For $\xi(M_3) = \lfloor \gamma/\mu \rfloor \mu$, we have for $0 < k \leq n - M_3$,

$$\mathbb{P}[X = M_3 + k] = \mathbb{P}[X = M_3] \cdot \frac{\gamma^k}{(\xi(M_3)+\mu)\cdots(\xi(M_3)+k\mu)},$$

and for $k > n - M_3$,

$$\begin{aligned}
&\mathbb{P}[X = M_3 + k] \\
=&\mathbb{P}[X = M_3] \cdot \frac{\gamma^k}{(\xi(M_3)+\mu)\cdots(\xi(M_3)+(n-M_3)\mu)(\xi(M_3)+(n-M_3)\mu+\theta)\cdots(\xi(M_3)+(n-M_3)\mu+(k-n+M_3)\theta)} \\
\leq&\mathbb{P}[X = M_3] \cdot \frac{\gamma^k}{(\xi(M_3)+\mu)\cdots(\xi(M_3)+k\mu)}.
\end{aligned}$$

Thus, for all $k > 0$, we have

$$\begin{aligned}
\mathbb{P}[X = M_3 + k] &\leq \mathbb{P}[X = M_3] \cdot \frac{\gamma^k}{(\xi(M_3)+\mu)\cdots(\xi(M_3)+k\mu)} \\
&\leq \mathbb{P}[X = M_3] \cdot \frac{\gamma^k}{\gamma(\gamma+\mu)\cdots(\gamma+(k-1)\mu)} \\
&\leq \mathbb{P}[X = M_3] \cdot \left(\frac{\gamma}{\gamma+(k-1)\mu/2}\right)^{k/2} \\
&= \mathbb{P}[X = M_3] \cdot \left(1 - \frac{(k-1)\mu/2}{\gamma+(k-1)\mu/2}\right)^{k/2} \\
&\leq \mathbb{P}[X = M_3] \cdot \exp\left(-\frac{k(k-1)\mu/4}{\gamma+(k-1)\mu/2}\right) \\
&\leq \mathbb{P}[X = M_3] \cdot \left(\exp\left(-\frac{k(k-1)\mu/4}{2\gamma}\right) + \exp\left(-\frac{k(k-1)\mu/4}{(k-1)\mu}\right)\right) \\
&= \mathbb{P}[X = M_3] \cdot \left(\exp\left(-\frac{k(k-1)\mu}{8\gamma}\right) + \exp\left(-k/4\right)\right).
\end{aligned}$$

The last inequality comes from the fact that $\frac{k(k-1)\mu/4}{\gamma+(k-1)\mu/2} \geq \frac{k(k-1)\mu/4}{2\gamma}$ if $\gamma \geq (k-1)\mu/2$ and $\frac{k(k-1)\mu/4}{\gamma+(k-1)\mu/2} \geq \frac{k(k-1)\mu/4}{(k-1)\mu}$ otherwise. Thus, for all $k > 0$,

$$\begin{aligned}
\mathbb{P}[X \geq M_3 + k] &= \sum_{j=k}^{\infty} \mathbb{P}[X = M_3 + j] \\
&\leq \sum_{j=k}^{\infty} \mathbb{P}[X = M_3] \cdot \left(\exp\left(-\frac{j(j-1)\mu}{8\gamma}\right) + \exp\left(-j/4\right)\right) \\
&\leq \mathbb{P}[X = M_3] \cdot \left(\int_{k-1}^{\infty} \exp\left(-\frac{j(j-1)\mu}{8\gamma}\right) dj + \frac{\exp(-k/4)}{1-\exp(-1/4)}\right).
\end{aligned}$$

Choose $B_3 > 0$ such that $B_3 x^2 \leq x(x-1)\mu/8$ for all $x \geq 2$, and choose $F_3 > 0$ such that $F_3 x^2 \leq B_3(x-1)^2/2$ for all $x \geq 3$. Also, choose $A_3 > 0$ such that $\mathbb{P}[X = M_3] \leq A_3/\sqrt{\gamma}$. Then, for all $k \geq 3$,

$$\begin{aligned}
\mathbb{P}[X \geq M_3 + k] &\leq \mathbb{P}[X = M_3] \cdot \left(\int_{k-1}^{\infty} \exp\left(-\frac{j(j-1)\mu}{8\gamma}\right) dj + \frac{\exp(-k/4)}{1-\exp(-1/4)}\right) \\
&\leq \frac{A_3}{\sqrt{\gamma}} \left(\int_{k-1}^{\infty} \exp\left(-B_3 j^2/\gamma\right) dj + \frac{\exp(-k/4)}{1-\exp(-1/4)}\right) \\
&\leq \frac{A_3\sqrt{2\pi}}{2\sqrt{B_3}} \cdot \exp\big(-B_3(k-1)^2/2\gamma\big) + C_3 \exp(-D_3 k)/\sqrt{\gamma} \\
&\leq E_3 \exp\big(-F_3 k^2/\gamma\big) + C_3 \exp(-D_3 k)/\sqrt{\gamma}
\end{aligned}$$

for some universal $C_3, D_3, E_3 > 0$. Finally,

$$\sum_{x=n+n_F}^{\infty} \mathbb{P}[X > x] = \int_{n+n_F-M_3}^{\infty} \mathbb{P}[X > M_3 + x]\, dx$$

$$
\begin{aligned}
&\leq 3+\int_{n+n_F-M_3+3}^{\infty} E_3\exp\bigl(-F_3x^2/\gamma\bigr)+C_3\exp(-D_3x)/\sqrt{\gamma}\,dx\\
&\leq 3+\int_{n+n_F-M_3}^{\infty} E_3\exp\bigl(-F_3x^2/\gamma\bigr)+C_3\exp(-D_3x)/\sqrt{\gamma}\,dx\\
&\leq 3+L_3\sqrt{\gamma}\exp\bigl(-F_3(n+n_F-M_3)^2/2\gamma\bigr)+G_3\exp(-D_3(n+n_F-M_3))/\sqrt{\gamma}\\
&\leq L_3\sqrt{\gamma}\exp\bigl(-F_3(n+n_F-M_3)^2/2\gamma\bigr)+3+G_3
\end{aligned}
$$

for some universal $L_3, G_3 > 0$. This completes the proof. □

Next, consider two single-class queues, A and B, with common arrival rate λ and abandonment rate θ. System A has m high priority rate-μ servers and m_F low priority rate-μ_F servers, while system B has m/r high priority rate-$r\mu$ servers and m_F/r low priority rate-$r\mu_F$ servers, for some $r > 1$. Let $Q^A(\infty)$ and $Q^B(\infty)$ denote the stationary queue lengths of the two systems respectively.

LEMMA 21. *For systems A and B,*

$$Q^A(\infty) \leq_{st} Q^B(\infty).$$

Proof of Lemma 21. The proof follows the same lines of argument as the proof of Lemma 2 in Bassamboo et al. (2010b), and is only included for completeness. Let $X^A(t)$ and $X^B(t)$ denote the headcount processes, which are birth-death processes. Let $\alpha = m + m_F$ and $\beta = (m + m_F)/r$ be the number of servers in the two systems, and let $\xi^A(\cdot)$ and $\xi^B(\cdot)$ denote the death rates.

We first note that the two systems have the same birth rates. For $x \geq 0$, $\xi^A(\alpha + x) = \xi^B(\beta + x)$, and for $0 \leq x \leq \beta$, $\xi^A(\alpha - x) \geq \xi^B(\beta - x)$. Initialize $X^A(0) = \alpha$, $X^B(0) = \beta$. Couple the two systems such that (i) the arrivals to both systems coincide, and (ii) the departures in system B is a subset of the departures in system A. Then, for all $t \geq 0$, $X^A(t) - \alpha \leq X^B(t) - \beta$ and

$$Q^A(t) = (X^A(t) - \alpha)^+ \leq (X^B(t) - \beta)^+ = Q^B(t).$$

As the stationary distribution is well-defined, we have the stochastic dominance of the stationary distribution. □

D.4. Proof of Lemma 6

To simplify the notation, we drop the superscript λ and the '(∞)'. In particular, let X_i denote the stationary number of Class i customers in the system, Q_Σ denote the stationary total queue length, Z_i denote the stationary number of Class i customers served by the dedicated servers, and Z_{Fi} denote the stationary number of Class i customers served by the flexible servers.

We first prove the lower bound. *Consider the case where $\theta \leq \mu_F$.* Note that $Q_\Sigma \geq f(X_1, X_2) := ((X_1 - n_1)^+ + (X_2 - n_2)^+ - n_F)^+$. As f is convex, by Jensen's inequality

$$\mathbb{E}[Q_\Sigma|\Lambda = \gamma; \nu] \geq \mathbb{E}[f(X_1, X_2)|\Lambda = \gamma; \nu] \geq f(\mathbb{E}[X_1|\Lambda = \gamma; \nu], \mathbb{E}[X_2|\Lambda = \gamma; \nu]).$$

Thus,

$$\theta\mathbb{E}[Q_\Sigma|\Lambda = \gamma; \nu] \geq \bigl(\theta\mathbb{E}[X_1|\Lambda = \gamma; \nu] - \theta n_1)^+ + (\theta\mathbb{E}[X_2|\Lambda = \gamma; \nu] - \theta n_2)^+ - \theta n_F\bigr)^+.$$

Equating the arrival and departure rates in stationarity, we have

$$\gamma_i = \theta\mathbb{E}[X_i|\Lambda = \gamma; \nu] + (\mu - \theta)\mathbb{E}[Z_i|\Lambda = \gamma; \nu] + (\mu_F - \theta)\mathbb{E}[Z_{Fi}|\Lambda = \gamma; \nu].$$

Because $\mathbb{E}[Z_i|\Lambda=\gamma;\nu]\le n_i$ and $\mathbb{E}[Z_{F1}|\Lambda=\gamma;\nu]+\mathbb{E}[Z_{F2}|\Lambda=\gamma;\nu]\le n_F$, for some $\alpha=\alpha(\gamma)\in[0,1]$,

$$\gamma_1\le\theta\mathbb{E}[X_1|\Lambda=\gamma;\nu]+(\mu-\theta)n_1+\alpha(\mu_F-\theta)n_F$$

and

$$\gamma_2\le\theta\mathbb{E}[X_2|\Lambda=\gamma;\nu]+(\mu-\theta)n_2+(1-\alpha)(\mu_F-\theta)n_F.$$

Then,

$$\begin{aligned}\theta\mathbb{E}[Q_\Sigma|\Lambda=\gamma;\nu]&\ge\big(\theta\mathbb{E}[X_1|\Lambda=\gamma;\nu]-\theta n_1)^+ + (\theta\mathbb{E}[X_2|\Lambda=\gamma;\nu]-\theta n_2)^+-\theta n_F\big)^+\\&\ge\big((\gamma_1-\mu n_1-\alpha(\mu_F-\theta)n_F)^+ + (\gamma_2-\mu n_2-(1-\alpha)(\mu_F-\theta)n_F)^+-\theta n_F\big)^+\\&\ge\big((\gamma_1-\mu n_1)^+ + (\gamma_2-\mu n_2)^+-\mu_F n_F\big)^+,\end{aligned}$$

where the last inequality follows from the fact that $((a-c)^+ + (b-d)^+ - e)^+\ge(a^+ + b^+ - (c+d+e))^+$ for any $c,d,e\ge 0$.

Next, consider the case where $\theta>\mu_F$. Note that

$$\theta\mathbb{E}[Q_\Sigma|\Lambda=\gamma;\nu]=\gamma_1+\gamma_2-\mu\mathbb{E}[Z_1+Z_2|\Lambda=\gamma;\nu]-\mu_F\mathbb{E}[Z_{F1}+Z_{F1}|\Lambda=\gamma;\nu].$$

Consider an auxiliary system, $\tilde{X}$, with all parameters the same except that its abandonment rate is $\tilde{\theta}=\mu_F$. We next construct a scheduling policy ν' such that

$$\mathbb{E}[Z_1+Z_2|\Lambda=\gamma;\nu]=\mathbb{E}[\tilde{Z}_1+\tilde{Z}_2|\Lambda=\gamma;\nu']\text{ and }\mathbb{E}[Z_{F1}+Z_{F2}|\Lambda=\gamma;\nu]=\mathbb{E}[\tilde{Z}_{F1}+\tilde{Z}_{F2}|\Lambda=\gamma;\nu'].\tag{25}$$

The policy for ν' is constructed through a coupling that keeps $Z_i=\tilde{Z}_i$ and $Z_{Fi}=\tilde{Z}_{Fi}$ at all times. This can be achieved by assuming that arrivals and service completions in both systems coincide, and that the abandonments in the auxiliary system is a subset of the abandonments in the original system since $\tilde{\theta}<\theta$. From (25), we have

$$\theta\mathbb{E}[Q_\Sigma|\Lambda=\gamma;\nu]=\tilde{\theta}\mathbb{E}[\tilde{Q}_\Sigma|\Lambda=\gamma;\nu']\ge\big((\gamma_1-\mu n_1)^+ + (\gamma_2-\mu n_2)^+-\mu_F n_F\big)^+,$$

where the last inequality follows from our analysis of the case where $\theta\le\mu_F$.

We next prove the upper bound. *We first consider the case when $\theta=\mu_F$.* By Lemma 20, we have

$$\begin{aligned}&\theta\mathbb{E}[Q_\Sigma|\Lambda=\gamma;\tilde{\nu}]\\=&\theta\mathbb{E}[Q_1|\Lambda=\gamma;\tilde{\nu}]+\theta\mathbb{E}[Q_2|\Lambda=\gamma;\tilde{\nu}]\\\le&(\gamma_1-\mu n_1-\lfloor\delta n_F\rfloor\mu_F)^+ + K_1\sqrt{\gamma}_1\exp\bigg(-\frac{K_2}{\gamma_1}(\gamma_1-\mu n_1-\lfloor\delta n_F\rfloor\mu_F)^2\bigg)+K_3\\&+(\gamma_2-\mu n_2-\lceil(1-\delta)n_F\rceil\mu_F)^+ + K_1\sqrt{\gamma}_2\exp\bigg(-\frac{K_2}{\gamma_2}(\gamma_2-\mu n_2-\lceil(1-\delta)n_F\rceil\mu_F)^2\bigg)+K_3\\\le&(\gamma_1-\mu n_1-\delta n_F\mu_F)^+ + K_1\sqrt{\gamma}_1\exp\bigg(-\frac{K_2}{\gamma_1}(\gamma_1-\mu n_1-\lfloor\delta n_F\rfloor\mu_F)^2\bigg)+K_3+\mu_F\\&+(\gamma_2-\mu n_2-(1-\delta)n_F\mu_F)^+ + K_1\sqrt{\gamma}_2\exp\bigg(-\frac{K_2}{\gamma_2}(\gamma_2-\mu n_2-\lceil(1-\delta)n_F\rceil\mu_F)^2\bigg)+K_3+\mu_F\\=&((\gamma_1-\mu n_1)^+ + (\gamma_2-\mu n_2)^+-\mu_F n_F)^+ + 2(K_3+\mu_F)\\&+K_1\sqrt{\gamma}_1\exp\bigg(-\frac{K_2}{\gamma_1}(\gamma_1-\mu n_1-\lfloor\delta n_F\rfloor\mu_F)^2\bigg)+K_1\sqrt{\gamma}_2\exp\bigg(-\frac{K_2}{\gamma_2}(\gamma_2-\mu n_2-\lceil(1-\delta)n_F\rceil\mu_F)^2\bigg).\end{aligned}$$

The result then follows using the proof of Lemma 1 in Bassamboo et al. (2010b).

Next, consider the case when $\theta < \mu_F$. Let the original system be labeled I. We form an auxiliary system II with the same parameters, except that the abandonment rate is $\theta^{II} = \mu_F$ and the holding cost is $h^{II} = h\mu_F/\theta$. We write

$$\Pi^I(n_1, n_2, n_F; \tilde{\nu}) = c(n_1 + n_2) + c_F n_F + (a + h/\theta) A^I(n_1, n_2, n_F; \tilde{\nu})$$

where $A^I(n_1, n_2, n_F; \tilde{\nu}) = \theta \mathbb{E}[Q^I_\Sigma(n_1, n_2, n_F; \tilde{\nu})]$ is the stationary abandonment rate in system I. Similarly,

$$\Pi^{II}(n_1, n_2, n_F; \tilde{\nu}) = c(n_1 + n_2) + c_F n_F + (a + h/\theta) A^{II}(n_1, n_2, n_F; \tilde{\nu}).$$

We next show that

$$A^I(n_1, n_2, n_F; \tilde{\nu}) \le A^{II}(n_1, n_2, n_F; \tilde{\nu}). \tag{26}$$

Note that

$$A^i = \mathbb{E}_\Lambda[\Lambda - \mu Z^i - \mu_F Z^i_F]$$

where Z^i and Z^i_F are the stationary number of busy rate-μ servers and rate-μ_F servers respectively, in system i, $i = I, II$. Thus, we only need to show that $Z^I \ge_{st} Z^{II}$ and $Z^I_F \ge_{st} Z^{II}_F$. Based on the scheduling policy $\tilde{\nu}$, it suffices to verify the following: If X^I is the stationary headcount in a single-class queue with m high priority rate-μ servers, m_F low priority rate-μ_F servers, and abandonment rate θ, and X^{II} is the same but with abandonment rate μ_F, then $X^I \ge_{st} X^{II}$. This is true because the birth rates of the two corresponding processes are the same, while the death rate in II is higher than in I. This proves (26), which further implies that

$$\Pi^{\lambda,I}(n^\lambda_1, n^\lambda_2, n^\lambda_F; \tilde{\nu}^\lambda) \le \Pi^{\lambda,II}(n^\lambda_1, n^\lambda_2, n^\lambda_F; \tilde{\nu}^\lambda) \le \tilde{\Pi}^\lambda(n^\lambda_1, n^\lambda_2, n^\lambda_F) + O(\lambda^{1-\alpha_2}) \text{ as } \theta^{II} = \mu_F.$$

Lastly, consider the case where $\theta > \mu_F$. We form a new auxiliary system III with the same parameters as X, except that $\mu^{III}_F = \theta$, $\mu^{III} = \mu\theta/\mu_F$, $c^{III} = c\theta/\mu_F$, and $c^{III}_F = c_F\theta/\mu_F$. Then,

$$\begin{aligned}\Pi^{\lambda,I}(n^\lambda_1, n^\lambda_2, n^\lambda_F; \tilde{\nu}^\lambda) &\le \Pi^{\lambda,III}\left(\frac{\mu_F}{\theta} n^\lambda_1, \frac{\mu_F}{\theta} n^\lambda_2, \frac{\mu_F}{\theta} n^\lambda_F; \tilde{\nu}^\lambda\right) \text{ by Lemma 21} \\ &\le \tilde{\Pi}^\lambda(n^\lambda_1, n^\lambda_2, n^\lambda_F) + O(\lambda^{1-\alpha_2}) \text{ as } \mu^{III}_F = \theta.\end{aligned}$$

□

D.5. Proof of Theorem 4.

Let $(n^{\lambda,*}_1, n^{\lambda,*}_2, n^{\lambda,*}_F; \nu^{\lambda,*})$ be optimal for (2). We have

$$\begin{aligned}&\Pi^\lambda(\lceil \tilde{n}^{\lambda,*}_1 \rceil, \lceil \tilde{n}^{\lambda,*}_2 \rceil, \lfloor \tilde{n}^{\lambda,*}_F \rfloor; \tilde{\nu}^\lambda) \\ \le &\tilde{\Pi}^\lambda(\lceil \tilde{n}^{\lambda,*}_1 \rceil, \lceil \tilde{n}^{\lambda,*}_2 \rceil, \lfloor \tilde{n}^{\lambda,*}_F \rfloor) + O(\lambda^{1-\alpha_2}) \text{ by the upper bound in Lemma 6} \\ \le &\tilde{\Pi}^\lambda(\tilde{n}^{\lambda,*}_1, \tilde{n}^{\lambda,*}_2, \tilde{n}^{\lambda,*}_F) + 2c + c_P\mu_F + O(\lambda^{1-\alpha_2}) \\ \le &\tilde{\Pi}^\lambda(n^{\lambda,*}_1, n^{\lambda,*}_2, n^{\lambda,*}_F) + O(\lambda^{1-\alpha_2}) \\ \le &\Pi^\lambda(n^{\lambda,*}_1, n^{\lambda,*}_2, n^{\lambda,*}_F; \nu^{\lambda,*}) + O(\lambda^{1-\alpha_2}) \text{ by the lower bound in Lemma 6.}\end{aligned}$$

□

References

Andradóttir, Sigrún, Hayriye Ayhan, Douglas G Down. 2003. Dynamic server allocation for queueing networks with flexible servers. *Operations Research* **51**(6) 952–968.

Armony, Mor, Avishai Mandelbaum. 2011. Routing and Staffing in Large-Scale Service Systems: The Case of Homogeneous Impatient Customers and Heterogeneous Servers. *Operations Research* **59**(1) 50–65.

Atar, Rami. 2005. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability* **15**(4) 2606–2650.

Bassamboo, Achal, J. Michael Harrison, Assaf Zeevi. 2006. Design and Control of a Large Call Center: Asymptotic Analysis of an LP-Based Method. *Operations Research* **54**(3) 419–435.

Bassamboo, Achal, Ramandeep S. Randhawa, Jan A. Van Mieghem. 2012. A Little Flexibility Is All You Need: On the Asymptotic Value of Flexible Capacity in Parallel Queuing Systems. *Operations Research* **60**(6) 1423–1435.

Bassamboo, Achal, Ramandeep S. Randhawa, Jan A. Van Mieghem. 2010a. Optimal Flexibility Configurations in Newsvendor Networks: Going Beyond Chaining and Pairing. *Management Science* **56**(8) 1285–1303.

Bassamboo, Achal, Ramandeep S. Randhawa, Assaf Zeevi. 2010b. Capacity Sizing Under Parameter Uncertainty: Safety Staffing Principles Revisited. *Management Science* **56**(10) 1668–1686.

Bassamboo, Achal, Assaf Zeevi. 2009. On a data-driven method for staffing large call centers. *Operations Research* **57**(3) 714–726.

Bertsimas, Dimitris, Xuan Vinh Doan. 2010. Robust and data-driven approaches to call centers. *European Journal of Operational Research* **207**(2) 1072–1085.

Best, Thomas J, Burhaneddin Sandıkçı, Donald D Eisenstein, David O Meltzer. 2015. Managing hospital inpatient bed capacity through partitioning care into focused wings. *Manufacturing & Service Operations Management* **17**(2) 157–176.

Borst, Sem, Avi Mandelbaum, Martin I. Reiman. 2004. Dimensioning Large Call Centers. *Operations Research* **52**(1) 17–34.

Chen, Hong, Hanqin Zhang. 2000. Diffusion Approximations for Some Multiclass Queueing Networks with FIFO Service Disciplines. *Mathematics of Operations Research* **25**(4) 679–707.

Dai, J. G., Wuqin Lin. 2008. Asymptotic optimality of maximum pressure policies in stochastic processing networks. *The Annals of Applied Probability* **18**(6) 2239–2299.

Dai, J. G., Tolga Tezcan. 2011. State Space Collapse in Many-Server Diffusion Limits of Parallel Server Systems. *Mathematics of Operations Research* **36**(2) 271–320.

Dong, Jing, Pnina Feldman, Galit B. Yom-Tov. 2015. Service Systems with Slowdowns: Potential Failures and Proposed Solutions. *Operations Research* **63**(2) 305–324.

Ernst, Ricardo, Panagiotis Kouvelis. 1999. The Effects of Selling Packaged Goods on Inventory Decisions. *Management Science* **45**(8) 1142–1155.

Gamarnik, David, Assaf Zeevi. 2006. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *The Annals of Applied Probability* **16**(1) 56–90.

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

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a Call Center with Impatient Customers. *Manufacturing & Service Operations Management* **4**(3) 208–227.

Gurvich, Itai, James Luedtke, Tolga Tezcan. 2010. Staffing Call Centers with Uncertain Demand Forecasts: A Chance-Constrained Optimization Approach. *Management Science* **56**(7) 1093–1115.

Gurvich, Itay, Ward Whitt. 2009a. Queue-and-Idleness-Ratio Controls in Many-Server Service Systems. *Mathematics of Operations Research* **34**(2) 363–396.

Gurvich, Itay, Ward Whitt. 2009b. Service-Level Differentiation in Many-Server Service Systems via Queue-Ratio Routing. *Operations Research* **58**(2) 316–328.

Halfin, Shlomo, Ward Whitt. 1981. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research* **29**(3) 567–588.

Harrison, J. Michael, Assaf Zeevi. 2004. Dynamic Scheduling of a Multiclass Queue in the Halfin-Whitt Heavy Traffic Regime. *Operations Research* **52**(2) 243–257.

Harrison, J. Michael, Assaf Zeevi. 2005. A Method for Staffing Large Call Centers Based on Stochastic Fluid Models. *Manufacturing & Service Operations Management* **7**(1) 20–36.

Karatzas, Ioannis, Steven Shreve. 1998. *Brownian Motion and Stochastic Calculus*. 2nd ed. Graduate Texts in Mathematics, Springer-Verlag, New York.

Kim, Jeunghyun, Ramandeep S Randhawa, Amy R Ward. 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.

Koçağa, Yaşar Levent, Mor Armony, Amy R. Ward. 2015. Staffing Call Centers with Uncertain Arrival Rates and Co-sourcing. *Production and Operations Management* **24**(7) 1101–1117.

Mandelbaum, Avishai, Sergey Zeltyn. 2009. Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers. *Operations Research* **57**(5) 1189–1205.

Netessine, Serguei, Nils Rudi. 2003. Centralized and Competitive Inventory Models with Demand Substitution. *Operations Research* **51**(2) 329–335.

Pashler, Harold. 1994. Dual-task interference in simple tasks: data and theory. *Psychological bulletin* **116**(2) 220.

Rajaram, Kumar, Christopher S. Tang. 2001. The impact of product substitution on retail merchandising. *European Journal of Operational Research* **135**(3) 582–601.

Shi, Cong, Yehua Wei, Yuan Zhong. 2019. Process flexibility for multiperiod production systems. *Operations Research* **67**(5) 1300–1320.

Simchi-Levi, David, Yehua Wei. 2012. Understanding the Performance of the Long Chain and Sparse Designs in Process Flexibility. *Operations Research* **60**(5) 1125–1141.

Song, Hummy, Anita L. Tucker, Ryan Graue, Sarah Moravick, Julius J. Yang. 2019. Capacity Pooling in Hospitals: The Hidden Consequences of Off-Service Placement. *Management Science* **66**(9) 3825–3842.

Tezcan, Tolga, JG Dai. 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.

Tsitsiklis, John N., Kuang Xu. 2012. On the power of (even a little) resource pooling. *Stochastic Systems* **2**(1) 1–66.

Turner, Stephen R. E. 2000. A join the shorter queue model in heavy traffic. *Journal of Applied Probability* **37**(1) 212–223.

Van Mieghem, Jan A., Nils Rudi. 2002. Newsvendor Networks: Inventory Management and Capacity Investment with Discretionary Activities. *Manufacturing & Service Operations Management* **4**(4) 313–335.

Wallace, Rodney B., Ward Whitt. 2005. A Staffing Algorithm for Call Centers with Skill-Based Routing. *Manufacturing & Service Operations Management* **7**(4) 276–294.

Whitt, Ward. 2002. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer Series in Operations Research and Financial Engineering, Springer-Verlag, New York.

Whitt, Ward. 2006. Staffing a Call Center with Uncertain Arrival Rate and Absenteeism. *Production and Operations Management* **15**(1) 88–102.