A Fluid-Diffusion-Hybrid Limiting Approximation for Priority Systems with Fast and Slow Customers

Lun Yu
Department of Industrial Engineering and Management Science, Evanston, IL, 60208, lunyu2014@u.northwestern.edu

Seyed Iravani
Department of Industrial Engineering and Management Science, Evanston, IL, 60208, s-iravani@northwestern.edu

Ohad Perry
Department of Industrial Engineering and Management Science, Evanston, IL, 60208, ohad.perry@northwestern.edu

We consider a large service system with two customer classes that are distinguished by their urgency and service requirements. In particular, one of the customer classes is considered urgent, and is therefore prioritized over the other class; further, the average service time of customers from the urgent class is significantly larger than that of the non-urgent class. We therefore refer to the urgent class as "slow," and to the non-urgent class as "fast." Due to the complexity and intractability of the system's dynamics, our goal is to develop and analyze an asymptotic approximation, which captures the prevalent fact that, in practice, customers from both classes are likely to experience delays before entering service. However, under existing many-server limiting regimes, only two of the following options can be captured in the limit: (i) either the customers from the prioritized (slow) customer class do not wait at all, or (ii) the fast-class customers do not receive any service. We therefore propose a novel *Fluid-Diffusion Hybrid* (FDH) many-server asymptotic regime, under which the queue of the slow class behaves like a diffusion limit, while the queue of the fast class evolves as a (random) fluid limit that is driven by the diffusion process. That FDH limit is achieved by assuming that the service rate of the fast class scales with the system's size, while the service rate of the slow class is kept fixed. Numerical examples demonstrate that our FDH limit is accurate when the difference between the service rates of the two classes is sufficiently large. We then employ the FDH approximation to study the costs and benefits of de-pooling the service pool, by reserving a small number of servers for the fast class. We prove that, in some cases, a two-pool structure is the asymptotically optimal system design.

1. Introduction

We consider a large-scale service system that handles two classes of customers with substantially different service requirements: a class of "urgent" (or "guaranteed") customers, that should be served quickly, and a class of "non-urgent" (or "best effort") customers, that can be delayed for relatively long time periods. Due to the practical relevance, variants of such systems were studied extensively in the literature in various settings and application domains. For example, in healthcare settings, "urgent" may refer to high-acuity patients that should be prioritized over lower-acuity (non-urgent) patients. In economic models, "urgent" may refer to customers who pay a premium in order to receive service within a *guaranteed* time period, and are thus prioritized over "non-urgent" customers, who receive only the remaining service capacity (which is not allocated to the guaranteed customers), and can therefore experience long delays.

Our aim in this paper is to capture the following ubiquitous phenomenon: Despite the fact that the urgent customers are prioritized over the non-urgent ones, they may nevertheless experience delay. As we elaborate below, this phenomenon presents modeling and analytical challenges, since it cannot be captured by standard many-server asymptotic regimes. Further, delays for both customer classes can co-exist in the asymptotic approximation only if the high-priority customers require longer services than the low-priority customers. We therefore consider systems in which the average service time of the urgent class is substantially longer than that of the non-urgent one, and we refer to the former as the "slow class," and to the latter as the "fast class." We will also refer to the slow- and fast-class customers as "slow customers" and "fast customers," respectively.

1.1. Motivation

The main motivation for this work comes from the observation that the setting just described applies in several important systems, in which the customers who receive high priority also require long service times. For example, contact centers employing "blending" of inbound calls with other types of jobs, such as outbound calls or emails, are prevalent in practice, see Gans et al. (2003), Pang and Perry (2014). While service-level constraints for inbound calls require that they be replied to relatively quickly (often within several seconds), the other type of jobs can be delayed for long time periods (hours or even days) before being processed. Further, the average duration of an inbound call is typically several minutes long, while email replies may follow a generic template, and require only several seconds to process. Similarly, the average duration of outbound calls is often short, because those calls are not initiated by the customers, who may not be interested in having a conversation with the agent.

Other important cases to which our setting applies are healthcare systems that treat patients with different levels of severity. In such cases, the level of severity is typically positively correlated with the length of the treatment, as well as the prioritization of the different patient types. For example, emergency rooms (ER)[1] in the US employ a five-level Emergency Severity Index (ESI) to rank the acuity of patients during the triage stage Gilboy et al. (2012). Patients granted ESI-1 are in need for an immediate, life-saving treatment, while ESI-2 patients require treatment "as soon as possible" due to risk of deterioration. Patients with ESI levels 3-5 (the particular ESI level of those patients differ by the amount of resources the triage nurse estimates they will need) can wait until a bed is available in the ER. Since ESI-1 patients constitute approximately 1-3% of all ER patients, Eitel et al. (2003), and since large ERs typically reserve resources (beds) for those patients, one can consider the ER as a two-class service system with our modeling characteristics, with ESI-2 patients being the "slow-class customers" (as those patients require long treatment times), and the lower-acuity patients with ESI 3-5 being the "fast-class customers." Indeed, ESI-2 patients are prioritized over the lower-acuity patients, and thus experience relatively short waiting times, whereas the waiting times for the ESI 3–5 patients are long (can be measured in hours) relative to the waiting times of ESI-2 patients, and relative to their own treatment times; see Song et al. (2015). A similar characteristic can be found in Inpatient Units (IPs) that treat Observation patients in addition to the Inpatients, since the former type of patients has lower priority during bed assignments, and shorter average treatment times than the latter patient type.

An immediate question for the above examples is whether it is beneficial to split the service pool into two separate pools—one that is dedicated to the slow (urgent) class, and the other that can serve both classes. Specifically, a fundamental implication of many-server asymptotic analysis is that pooling reduces waiting times of all customers, due to associated economies of scales Whitt (1992). However, in the multi-class setting, significant improvements, in terms of waiting times, can be achieved for the fast class with only minor impacts on the slow class. We therefore study a two-server-pool system as well, and show that de-pooling may be asymptotically optimal, in our proposed asymptotic regime, when abandonment, holding and staffing costs are incurred.

1.2. Modeling and Analytical Approaches

To repeat, the examples discussed above all share the two features that we aim to model, namely, (i) the service requirements of the prioritized (urgent) class is substantially longer than that of the lower-priority class, and (ii) a non-negligible proportion of the customers from either class experiences delays in queue before entering service. Unfortunately, exact analysis of the system is intractable, even if it evolves as a continuous-time Markov chain (CTMC), as one must keep track of the number of customers from each class in service and in queue, so that the minimal Markov representation of the system is four-dimensional in the single-pool case, and five-dimensional in the two-pool case. Furthermore, little insight can be obtained from numerical computations of the system's steady-state, or from simulations that aim to approximate steady-state performance metrics, and it is therefore natural to resort to a Many-Server Heavy-Traffic (MSHT) approximation. However, under existing MSHT limiting regimes, only the following four scenarios are possible in the limit:

(I) The system is **underloaded**, in which case both classes are served, and neither class experiences any delay.

(II) The system is **critically loaded**, in which case both classes are served, and the urgent (prioritized) class experiences negligible delays.

(III) The system is **overloaded**, but there is sufficient service capacity to serve the urgent class alone. In this case, the urgent class experiences negligible delays, and a significant proportion of the non-urgent class abandons the queue.

(IV) The system is **overloaded**, and there is at most a negligible service capacity left for the non-urgent class. In this case, the urgent class may experience non-negligible delays, and most non-urgent customers abandon the queue.

(See §2 for a more rigorous discussion on the four scenarios above.) Therefore, in order to capture our desired dynamics, we propose a new MSHT regime for a system with sufficient service capacity to handle all customers (unlike in scenario (IV) above), such that both customer classes experience non-negligible delays asymptotically (unlike scenarios (I)–(III)). We achieve this by considering a properly-scaled sequence of queueing systems in which, in addition to the arrival rates and the number of servers, the service rate of the fast class is accelerated appropriately. Under that scaling, the queue of the fast class converges to a (random) fluid limit, whose dynamics are governed by the resulting diffusion limit of the slow class. We therefore refer to this limiting approximation as a *Fluid-Diffusion Hybrid* (FDH).

As usual, a limiting approximation for an intractable stochastic system is useful because, in addition to providing quantitative estimations for key performance measures, it also provides qualitative insights that are not available otherwise. Here, we demonstrate this by employing the FDH limit to consider the impacts of *de-pooling*, namely, of splitting the service pool to two separate pools—one that handles both classes, and the other that is dedicated to the fast class. Since the fast class requires short service times, the size of the dedicated pool is an order of magnitude smaller than that of the shared pool. Motivated by the ER setting, in which a relatively small pool of beds that are dedicated to patients who have low priority in the general ER is referred to as "fast track," we refer to the dedicated pool by the same name. A schematic representation of the single- and the two-pool system is depicted in Figure 1, which clarifies why the single-pool system is often referred to as the V-model (or V-system), while the two-pool system is known as an N-model.

To summarize, the contribution of this paper is threefold:

(1) We propose a new asymptotic regime for a two-class many-server queueing system (the V-model) in which the service rates of the two classes are significantly different. In that new FDH

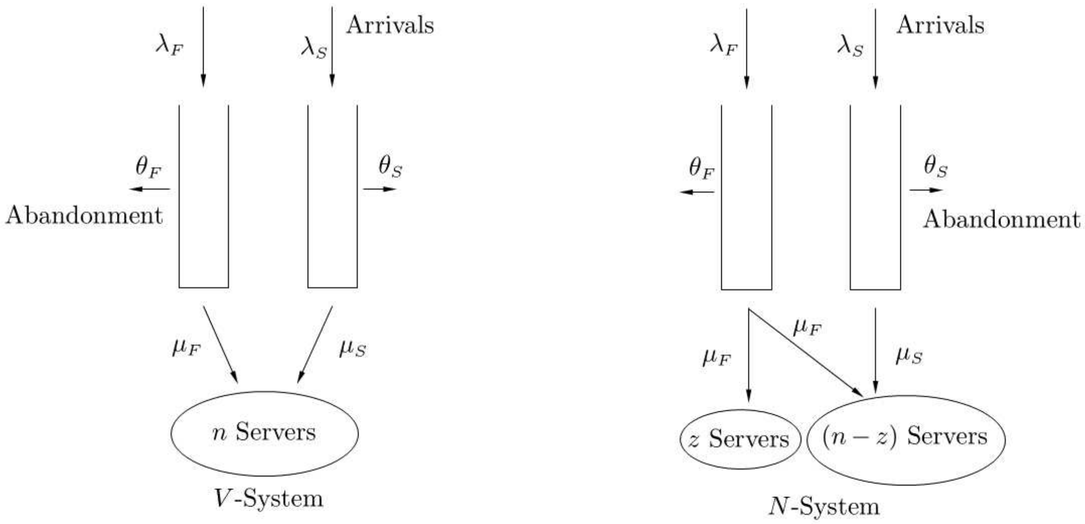

Figure 1 A single-pool "V-system" (left), and a two-pool "N-system" with a fast track (right).

limit (i) both classes, including the high-priority class, have a non-negligible probability of waiting for service, and (ii) both classes, including the low-priority class, receive service.

(2) We employ the FDH limit to study the potential benefits of de-pooling (the N-model). We show that a small number of dedicated servers, that is an order of magnitude smaller than the total number of servers, can substantially reduce the overall congestion in the system by reducing the delay of the fast-class customers at the expense of a *negligible* increase in the delay of the slow class. Our analysis thus confirms existing evidence, that having a small fast track can reduce overall waiting times for patients in the ER; see e.g., Sanchez et al. (2006) and Cooke et al. (2002).

(3) Finally, we demonstrate how the FDH limit can be used to optimize the system's topology when a holding and staffing cost is incurred. In particular, we prove important structural results for the optimal system-design problem in the FDH limit, and prove that the FDH-optimal system topology is asymptotically optimal in an appropriate sense (see Proposition 1 in Section 6).

1.3. Conventions About Notation

All random variables and processes are defined on a probability space $(\Omega, \mathcal{F}, P)$. We let $\mathbb{Z}_+ := \{1, 2, \dots\}$ denote the positive integers, $\mathbb{R}$ denote the real numbers, and $\mathbb{R}^k$, $k > 1$, denote all the k-dimensional vectors with components in $\mathbb{R}$. We use e to denote the identity function, $e(t) = t$.

The indicator function of a set A, denoted by 1_A, is the function that equal to 1 on A and to 0 otherwise. We denote by $D^k := D([0,\infty),\mathbb{R}^k)$ the space of $\mathbb{R}^k$-valued right-continuous functions with limits from the left, endowed with Skorohod J_1 topology; see, e.g., Whitt (2002), and write D for D^1.

In the subspace $C^k \subset D^k$ of $\mathbb{R}^k$-valued continuous functions, the J_1 topology reduces to the topology of uniform convergence over compact intervals, which is induced by the uniform metric

$$\|x\|_t := \sup_{0\le s\le t} \|x(s)\| := \sup_{0\le s\le t} \max_{1\le i\le k} |x_i(t)|, \quad x=(x_1,\ldots,x_k)\in C^k;$$

note that we have used $\|\cdot\|$ to denote the maximum norm in $\mathbb{R}^k$. We use $\Rightarrow$ to denote weak convergence (convergence in distribution).

For a sequence of positive real numbers $\{a^n : n\in\mathbb{Z}_+\}$ and a sequence of real numbers $\{b^n : n\in\mathbb{Z}_+\}$, we write (i) $b^n = o(a^n)$ if $|b^n/a^n| \to 0$ as $n\to\infty$; (ii) $b^n = O(a^n)$ if $|b^n/a^n|$ is bounded from above; (iii) $b^n = \Theta(a^n)$ if $|b^n/a^n|$ is bounded from above and from below by strictly positive numbers, namely, if $m \le |b^n/a^n| \le M$ for some $0 < m < M < \infty$ and for all n.

For a sequence of random variables $\{y^n : n\in\mathbb{Z}_+\}$ and a sequence of positive real numbers $\{a^n : n\in\mathbb{Z}_+\}$, we write (i) $y^n = o_P(a^n)$ if $\|y^n\|/a^n \Rightarrow 0$ as $n\to\infty$; (ii) $y^n = O_P(a^n)$ if $\{\|y^n\|/a^n : n\in\mathbb{Z}_+\}$ is a tight sequence in $\mathbb{R}$; and (iii) $y^n = \Theta_P(a^n)$ if y^n is $O_P(a^n)$, but **not** $o_P(a^n)$. Finally, for a sequence of stochastic processes $\{Y^n : n\in\mathbb{Z}_+\}$ and a sequence of positive real numbers $\{a^n : n\in\mathbb{Z}_+\}$, we write (i) $Y^n = o_P(a^n)$ if for any $t\ge 0$, $\|Y^n\|_t/a^n \Rightarrow 0$ as $n\to\infty$; (ii) $Y^n = O_P(a^n)$ if for any $t\ge 0$, $\{\|Y^n\|_t/a^n : n\in\mathbb{Z}_+\}$ is a tight sequence in $\mathbb{R}$; and (iii) $Y^n = \Theta_P(a^n)$ if Y^n is $O_P(a^n)$, but **not** $o_P(a^n)$.

Organization. The rest of the paper is organized as follows: We provide background on relevant many-server heavy-traffic asymptotics, and expand on the theoretical need to develop the FDH regime in Section 2. A review of related literature is presented in Section 3. Sections 4 and 5 are dedicated to analyzing the "V-system" and the "N-system," respectively. In Section 6 we consider the V and N models under a cost structure, and establish the asymptotically optimal system design. We present numerical examples in Section 7, and summarize in Section 8.

The paper includes an appendix. Appendix A is dedicated to generalizations to the setting of the main paper. In particular, we consider more general scaling regimes, as well as systems with time-varying arrival processes and systems with non-exponential service time distributions. Appendices B–E are devoted to the proofs of the results in the main paper.

2. Background on MSHT Asymptotics and Relevant Insights

In this section, we provide background information on heavy traffic limiting approximations and relevant insights for the FDH regime. Since we are interested in systems with many agents (or servers), we focus on the MSHT limiting regime, which is achieved by considering a sequence of queueing systems in which the number of servers increases to infinity, and the traffic intensity is scaled appropriately so that non-trivial limits are achieved.

Existing MSHT Limiting Regimes. In their seminal paper, Halfin and Whitt (1981) classified three heavy-traffic regimes, which were later named in Garnett et al. (2002) as Quality-Driven (QD), Quality-and-Efficiency Driven (QED), and Efficiency-Driven (ED) regimes. Under the QD regime, an arrival will—with probability converging to 1—find an idle agent, and will therefore not be delayed in queue. Thus, a pool of servers operating under the QD regime is asymptotically equivalent to an infinite-server queue, as in Iglehart (1965). In contrast, under the ED regime, an arrival—with probability converging to 1—will need to wait in a queue to be served. Under the QED regime, which was first identified in Halfin and Whitt (1981), and is therefore also called the *Halfin-Whitt regime*, the probability that an arrival will find all servers busy is, asymptotically, strictly between 0 and 1, even though most servers are working at any given time. More specifically, at most order $\sqrt{n}$ servers can be idle as $n \to \infty$, where n is the number of servers in the pool. In this regime, the queue is of order $\sqrt{n}$ so that waiting times, as well as the proportion of abandonment, are decreasing to 0 at rate $1/\sqrt{n}$. For a single class and single-pool system with no abandonment, it was shown in Halfin and Whitt (1981) that the QED regime is achieved via the square-root staffing rule, stipulating that

$$\lim_{n\to\infty} \sqrt{n}(1-\rho_n) = \beta,$$

for some $\beta > 0$, where $\rho_n < 1$ is the traffic intensity (arrival rate divided by the total service rate of the pool). This result was extended in Garnett et al. (2002) to include abandonment, in which case $\rho_n \geq 1$ (and $\beta \leq 0$) is allowed.

For a single-class, single-pool system with abandonment, we can therefore distinguish between the three different regimes according to the value of β: For $\beta = +\infty$, $\beta \in \mathbb{R}$, or $\beta = -\infty$, the system operates in, respectively, the QD, QED or ED asymptotic regime. Further, abandonment and waiting times are asymptotically negligible in all cases, unless $\liminf_{n\to\infty} \rho^n > 1$ (and in particular, when $\rho^n \to \rho > 1$ as $n \to \infty$), in which case the proportion of abandonment is asymptotically non-negligible, and waiting times are of the same order as service times. Note that this latter case corresponds to having a genuinely overloaded system, because the traffic intensity is bounded away from its critical value 1. The queue process is then well-approximated by a fluid limit; see Whitt (2004).

A fourth MSHT regime, named *non-degenerate slowdown* (NDS), was proposed in Atar (2012). The NDS regime is of “ED-type,” because arrivals are delayed in queue with a probability converging to 1, but, unlike previous ED approximations, waiting times are of the same order as the service times while simultaneously, the abandonment proportion is negligible. In particular, the proportion of abandonment is of order $1/\sqrt{n}$, as in the QED regime. The NDS asymptotic regime is achieved by scaling the number of agents, as well as the service rate of each individual agent, by $\sqrt{n}$.

In practice, engineering consideration is required in order to determine which regime is an appropriate approximation for a given system. If most customers enter service immediately upon arrival, then the QD approximation is appropriate. If a non-negligible proportion (which is not too close to 1) of the arrivals is delayed in queue, but waiting times of delayed customers are short relative to their average service times, then the QED regime is an appropriate asymptotic approximation. On the other hand, if almost all arrivals are delayed in queue, then the ED approximation should be employed. (The exact type of ED approximation can be chosen based on the proportion of abandonment, e.g., NDS when abandonment is negligible, and a fluid approximation when abandonment is substantial.)

Traffic Intensity	Slow Class	Fast Class
$\sqrt{n}(1-\rho_S^n-\rho_F^n)\to+\infty$	QD	QD
$1-\rho_S^n-\rho_F^n=O(n^{-1/2})$	QD	QED
$\rho_S+\rho_F>1$ and $\rho_S<1$	QD	ED
$\rho_S+\rho_F>1$ and $\rho_S\geq 1$	QED or ED	No Service

Table 1 Summary of existing MSHT regimes for two-class priority systems

Relevant Insights. With the insights obtained from the single-class single-pool setting, we can explain why the four scenarios in §1.2 are the only possible ones. Consider a sequence of single server-pool systems indexed by the number of servers n, and for $i=S,F$, let λ_i^n, μ_i and θ_i denote the arrival rate, service rate and abandonment rate of the class-i customers, respectively, in system n. (S and F are mnemonic for “fast” and “slow.”) Assume that $\lambda_i^n/n\to\lambda_i$ as $n\to\infty$, but that the service and abandonment rates are kept fixed along the sequence. Note that abandonment keeps the two queues stable even if the total arrival rate to the system is higher than its total processing rate. Let $\rho_i^n:=\lambda_i^n/(n\mu_i)$ denote the traffic intensity of class i and $\rho_i:=\lambda_i/\mu_i$ denote the limit, so that $\rho_i^n/n\to\rho_i$ as $n\to\infty$, for $i=S,\ F$.

If $\rho_S+\rho_F<1$, then the system operates in the QD regime, and neither class experiences any waiting, asymptotically. The same continues to hold if $\rho_S+\rho_F=1$, but $1-\rho_S^n-\rho_F^n$ converges to 0 at a slower rate than $\sqrt{n}$, namely, if $\sqrt{n}(1-\rho_S^n-\rho_F^n)\to\infty$ as $n\to\infty$; see Iglehart (1965) and the discussion in the introduction of Halfin and Whitt (1981). These two cases correspond to scenario (I). Scenario (II) arises when $\rho_S+\rho_F=1$, but $1-\rho_S^n-\rho_F^n$ converges to 0 at rate $\sqrt{n}$ or faster, while Scenario (III) arises when $\rho_S+\rho_F>1$, but $\rho_S<1$. In this case, the delay in queue of the slow class is negligible asymptotically with respect to the delay of the fast class; see, e.g., Theorem 3 and the discussion following it in Maglaras et al. (2017). Finally, if $\rho_S\geq 1$, then, asymptotically, there is no service capacity left to handle the low-priority (fast) class, so that the proportion of fast customers that are served is negligible, and practically all those customers leave the system via abandonment, as in scenario (IV). Table 1 summarizes the four scenarios.

2.1. A Singular Perturbation Approach

The discussion above shows that a different MSHT approach is required in order to have an asymptotic approximation for the system under which customers from either class are delayed, but most customers (from either class) are eventually served. Since we want the probability that a slow customer is delayed in queue to be strictly positive, we should assume that $\rho_S \geq 1$, but then only a negligible service capacity can be allocated to the fast class. One might try to circumvent this problem by exploiting the fact that the fast class requires short service times, and take $1/\mu_F = 0$. This perturbation approach can be effective in some cases, as in Whitt (2005), but it is easy to see that it trivializes the problem in our setting. Indeed, if the fast class is served instantaneously, then a single dedicated server for that class would suffice to ensure that no queueing of fast-class customers ever occurs. Asymptotically, the system is then equivalent to the single-class $M/M/n+M$ (Erlang-A) queue, serving the slow class only. Further, prioritizing the fast customers in this case does not impact the service quality of the slow customers at all. Therefore, such an approximation has no useful implication for the practical settings we consider.

Instead, we propose a *singular perturbation* approach, in which the service time of the fast class approaches 0 (equivalently, the service rate increases without bound), but remains strictly positive along the sequence of systems. We achieve our modeling goals by letting the service rate of the fast class increase with n at an order $O(\sqrt{n})$, while maintaining the service rate of the slow class fixed. Under an appropriate spatial scaling, the queue of the fast class converges to a diffusion process, and the queue of the slow class to a fluid limit whose dynamics are governed by those of the diffusion limit.

3. Literature Review

The V and N models have both been studied extensively in the conventional heavy traffic setting, e.g., see Whitt (1971), Bell and Williams (2001), Ghamami and Ward (2013) (with customer abandonment), and Harrison (1998), as well as in the MSHT setting, which is our focus here; see, e.g., Atar et al. (2010), Harrison and Zeevi (2004), Atar et al. (2004) and Gurvich et al. (2008), for works related to V-systems, and Tezcan and Dai (2010) for an N-system. Also related are the papers Gurvich and Perry (2012), which considers overflow from a main pool of agents to a second pool (or pools), and Perry and Whitt (2009, 2011, 2015), which consider an automatic control designed to transform an X-model (with two-way sharing) into an N-model. Unlike our FDH limit, the limits in all these works (and also in other works considering heavy traffic approximations for queueing systems) are either fluid or diffusion processes. Further, the numbers of servers in the two pools in the N-systems are of the same order, whereas the fast track in our N-system is an order of magnitude smaller than in the main pool.

Our work relates to the literature on service systems that handle two types of customers: *guaranteed* and *best effort*. The service quality for the former customer class (in terms of delay times in queue or in terms of service rates) is guaranteed, whereas for the latter class, the allocation of service capacity is based on availability; see Afeche (2013), Maglaras and Zeevi (2004), Maglaras and Zeevi (2005) and references therein.

Assuming that customers are strategic and seek to maximize their utility, Maglaras et al. (2017) shows that firms providing a service to a market consisting of several customer classes should offer a menu of delays and costs in order to maximize their profits. In particular, optimal market segmentation might require that low-priority classes are delayed in queue, even when such delays can be eliminated due to having sufficient service capacity. In this case, the optimal staffing is to have the high-priority class operate in the QD regime, and the low priority in the ED regime. Here we do not consider a customer choice model, but it is intuitively clear that having the low-priority class operate in the QED (instead of the QD) regime might be optimal in some cases; the FDH approximation can be used to study such cases when the service times of guaranteed and best-effort customers are substantially different. (Note that longer service times can be offered as part of a delay, service-time and cost menu.) We also refer to Nazerzadeh and Randhawa (2018), which considers a related problem in the single-server setting, and Gurvich et al. (2018), which compares the priority schemes of revenue-maximizing firms to those of a social planner.

Another closely related paper is Ata and Van Mieghem (2009), which considers a queueing system in which an "express class" is served by a fast service pool, and a "standard class" is served by a slow service pool. The problem considered in this paper is whether letting the fast servers process customers from the standard class is beneficial, namely, whether the system should operate two independent dedicated service pools, or an N-system with a shared service pool and a second pool dedicated to the slow class.

Perturbation and Singular-Perturbation Techniques. Perturbation of a (possibly stochastic) dynamical system is an analytical method in which a "small" parameter or process ε is replaced by 0 (0 may be the zeroth function, depending on the setting). If the limit point $\varepsilon = 0$ differs in important ways from the approach to the limit as $\varepsilon \to 0$, then a singular perturbation technique is required, in which ε (which is fixed for the given system) is taken to 0 in a suitable way, so as to achieve a meaningful limiting approximation; see, e.g., Hinch (1991).

Whitt (2005) considers the heavy-traffic limit for the $G/H_2^*/n/m$ queue, in which the service-time distribution H_2^* is exponential with mean $1/\nu$ with some probability p, and has point mass at 0 with probability $1-p$. Thus, the system with the H_2^* service-time distribution can be considered as a perturbation for a system with an hyperexponential service-time distribution H_2 (a mixture of two exponentials) in which the service time is, with probability $1-p$, small relative to ν. This perturbation technique was shown to be useful for developing closed-form expressions for performance measures for the $M/G/n$ model in Whitt (1983). Maglaras and Zeevi (2004) employs a perturbation approach, in which the service rates of different customer classes are perturbed about a single value in order to develop a diffusion limit that approximates the intractable diffusion limit of the original system with arbitrary service rates.

Singular perturbation techniques have been used extensively in the study of stochastic systems. An example for a fluid limit of a queueing model can be found in (Perry and Whitt 2016, §6), where one of the control parameters is replaced by 0 in certain states of the system. The resulting singularly perturbed dynamical system is then amenable to qualitative long-run analysis that is intractable for the original fluid limit. Perhaps the most prevalent technique is the *method of time scales*, under which a small and "fast" process is replaced by its local stationary behavior; see, e.g., Yin and Zhang (2005), Yin and Zhang (2012) and Khasminskii and Yin (2005). In the queueing literature, we mention pointwise stationary approximations, as in Bassamboo et al. (2009) and Whitt (1991), and stochastic averaging principles, as in Hunt and Kurtz (1994) and Coffman Jr et al. (1995). We refer to Gurvich and Perry (2012) and Perry and Whitt (2012) for detailed discussions and literature reviews; see also Wu et al. (2018) and Moyal and Perry (2017). However, we emphasize that our singular-perturbation approach here is different than in any of the aforementioned papers, since our diffusion process evolves in the same time scale as the fluid process, so that no separation of time scales occurs.

Finally, scaling of service times was proposed in Atar (2012) to develop the NDS regime. See Atar and Gurvich (2014) for an application of the NDS regime in multi-class multi-pool systems. However, unlike our setting, the number of agents in the NDS regime scales in the same order as the service rates, and the service times of all customer classes scale in the same fashion. More importantly, the NDS regime was developed so as to have the service time and delay in queue of a typical customer decay at the same order $n^{1/2}$; in particular, both are comparable to each other. In the FDH regime, however, the service time of a fast customer decays at rate $n^{-1/2}$, whereas the average delay is bounded away from 0 as $n \to \infty$. Thus, the fast customers experience delays that are an order of magnitude larger than their service times, and so the corresponding queue does not operate in the NDS regime.

4. The FDH Limit for the V-System

We consider a single pool of many statistically-homogeneous agents that handle two customer classes, as depicted in the left panel of Figure 1. The service times of class-i customers are assumed to be Independent and Identically-Distributed (IID) exponential random variables with mean $1/\mu_i$, $i = S$ or $i = F$, and to satisfy $1/\mu_F \ll 1/\mu_S$; see Assumption 1 below. We refer to class-S and to class-F customers as "slow" and "fast," respectively.

We let the arrival process of class-i customers follow a Poisson process with rate λ_i. A class-i customer that is not routed to an agent immediately upon arrival is placed in an infinite buffer (there are two buffers, one for each class), and waits for his turn to be served. We assume that each class-i customer has a finite patience time that is exponentially distributed with mean $1/\theta_i$, and will abandon the queue if his delay in queue exceeds his patience time. All random variables are assumed to be independent from each other, as well as from the two independent Poisson arrival processes.

Agents are non-idling, namely, an agent does not idle if a customer is waiting in either queue, and give strict priority to the slow class. For tractability, we assume that the routing policy is preemptive, so that a slow customer never waits in queue if there are fast customers in service. A fast customer who is replaced by a slow customer is put back at the head of his designated queue, and resumes his service at a later time. As we explain in Section 7.3, the difference between the queueing dynamics under the preemptive and the non-preemptive priority policies diminishes as the size of the system increases, so that our results are meaningful also if the non-preemptive priority policy is employed.

4.1. The FDH Scaling

The FDH approximation is obtained in a MSHT limiting regime for a sequence of systems indexed by the number of servers n, as n increases without bound. We append with a superscript n the arrival, service, and abandonment rates, as well as the stochastic processes corresponding to system n. We let λ_S^n and λ_F^n increases proportionally to n, so that neither one is asymptotically negligible, but take the abandonment rates of both classes, and the service rate of the *slow class*, to be fixed along the sequence. The aforementioned singular-perturbation technique corresponds to letting the service rate of the fast class scale with n so as to achieve a non-trivial limit. It will become clear (see the discussion below Theorem 1) that, since we consider the slow class to be operating in the QED regime, μ_F^n must increases at a rate $\sqrt{n}$. We formalize our MSHT scaling in the following assumption.

Let r_F^n denote the scaled offered load of the fast class; in particular,

$$r_F^n := R_F^n/\sqrt{n} \quad \text{where } R_F^n := \lambda_F^n/\mu_F^n. \tag{1}$$

ASSUMPTION 1 **(FDH scaling)**. *For $\beta \in \mathbb{R}$ and $\theta_S > 0$, the following holds for the slow class.*

$$\lim_{n\to\infty} (n - \lambda_S^n)/\sqrt{n} = \beta, \quad \mu_S^n = 1, \quad \textit{and} \quad \theta_S^n = \theta_S \quad \textit{for all } n \geq 1.$$

For strictly positive real numbers λ_F, r_F and θ_F, the following holds for the fast class

$$\lim_{n\to\infty} \lambda_F^n/n = \lambda_F, \quad \lim_{n\to\infty} r_F^n = r_F, \quad \textit{and} \quad \theta_F^n = \theta_F \quad \textit{for all } n \geq 1.$$

We remark that the assumption $\mu_S^n = 1$ is taken without loss of generality, because we can also measure time in terms of the expected service-time of the slow class.

Let $X_i^n(t)$ and $Q_i^n(t)$ denote the number of class-i customers in the system and in queue at time t, respectively, and let $X^n(t) := (X_S^n(t), X_F^n(t))$ and $Q^n(t) := (Q_S^n(t), Q_F^n(t))$. Note that X^n is a CTMC, but that Q^n is not Markov. The FDH-scaled processes are defined via

$$\widetilde{X}^n := \left(\widetilde{X}_S^n, \widetilde{X}_F^n\right) = \left(\frac{X_S^n - n}{\sqrt{\lambda_S^n}}, \frac{X_F^n}{\lambda_F^n}\right) \quad \text{and} \quad \widetilde{Q}^n := (\widetilde{Q}_S^n, \widetilde{Q}_F^n) = \left(\frac{Q_S^n}{\sqrt{\lambda_S^n}}, \frac{Q_F^n}{\lambda_F^n}\right). \tag{2}$$

Notice that the processes corresponding to the slow class, X_S^n and Q_S^n, are diffusion-scaled, whereas the processes corresponding to the fast class, X_F^n and Q_F^n, are fluid-scaled.

4.2. The FDH Limit

The FDH limit of $\widetilde{X}^n$ in (2) depends on having the sequence of initial conditions $\widetilde{X}^n(0)$ converge in $\mathbb{R}^2$. We therefore must guarantee that the initial conditions in the limit and the pre-limit are "legitimate" as in the following assumption.

ASSUMPTION 2 **(initial condition for the V-system)**. $Q_S^n(0) = (X_S^n(0) - n)^+$ *and* $Q_F^n(0) \geq 0$ *for all $n \geq 1$.*

Both Assumptions 1 and 2 are assumed to hold *throughout this section.*

Below is the main result for the V-system—the FDH limit for the scaled sequence $\{\widetilde{X}^n : n \geq 1\}$. This limit is characterized via a stochastic differential equation (SDE) whose solution is a fluid-diffusion hybrid, and we thus refer to that SDE as a *Hybrid Stochastic Differential Equation* (HSDE).

THEOREM 1 **(FDH limit for the V-system)**. *If $\widetilde{X}^n(0) \Rightarrow X(0)$ in $\mathbb{R}^2$, then $\widetilde{X}^n \Rightarrow X$ in D^2, where $X := (X_S, X_F)$ is the unique solution to the following HSDE with initial condition $X(0)$*

$$dX_S(t) = (-\beta + X_S(t)^- - \theta_S X_S(t)^+)dt + \sqrt{2}dB(t), \tag{3}$$

$$dX_F(t) = (1 - r_F^{-1} X_S(t)^- - \theta_F X_F(t))dt + dI(t) \quad \text{and} \quad X_F(t) \geq 0, \tag{4}$$

where B is a standard Brownian motion, and I is the unique non-decreasing process satisfying

$$I(0) = 0 \quad \text{and} \quad \int_0^t 1_{\{X_F(s)>0\}} dI(s) = 0, \quad \text{for all } t \geq 0. \tag{5}$$

Observe that the expression characterizing the process X_S in (3) does not involve X_F; it is the piecewise Ornstein-Uhlenbeck (OU) process that was shown in Garnett et al. (2002) to arise as the limit for the Erlang-A model operating in the QED regime. However, X_F and X_S are **dependent processes**, as is clear from (4). (Observe that X_S and $X_F(0)$ are the only sources of randomness in the equations for X_F in (4) and (5).) From the fact that X_S is the Garnett diffusion, it follows that the number of agents working with fast customers is $O_P(\sqrt{n})$ in the pre-limit. This explains why the service rate of the fast class must scale at a rate $\sqrt{n}$. Further, due to the fluid scaling of $\widetilde{X}^n_F$, the limit process X_F is therefore reflected at 0, and its non-negativity is preserved by the *regulator process* I in (5). It is also easy to see that X_F is bounded w.p.1 by $\max\{X_F(0), \theta_F^{-1}\}$, and in fact, one can show that if $X_F(0) > \theta_F^{-1}$, then X_F will decrease towards $[0, \theta_F^{-1})$ and will be absorbed in this interval.

It is easy to see that the limit process X in Theorem 1 also characterizes the FDH limit of $\{\widetilde{Q}^n : n \geq 1\}$: For each $n \geq 1$, we have $Q^n_S = (X^n_S - n)^+$ and $Q^n_S + Q^n_F = (X^n_S + X^n_F - n)^+$. Therefore, Theorem 1 and the continuous mapping theorem imply that $Q := (Q_S, Q_F) := (X^+_S, X_F)$ is the

FDH limit of $\{\widetilde{Q}^n : n \geq 1\}$. (Notice that $\widetilde{X}_F^n$ and $\widetilde{Q}_F^n$ both converge weakly to the same limit X_F due to the fact that the number-in-service process of the fast class is $O_P(\sqrt{n})$.)

Now consider the pre-limit *cumulative idleness process*

$$I^n(t) = \int_0^t (n - X_S^n(s) - X_F^n(s))^+ ds,$$

and its scaled version $\widetilde{I}^n = I^n/R_F^n$. Note that the integrand in the above expression represents the number of idle agents at time s. Since idleness is non-decreasing and "accumulates" only when the queue of the fast class is empty, we have

$$\int_0^t 1_{\{\widetilde{Q}_F^n(s)>0\}} d\widetilde{I}^n(s) = 0, \quad \text{for all } t \geq 0,$$

which is analogous to (4), due to the aforementioned asymptotic equivalence of Q_F and X_F. Indeed, we can prove that I is the FDH limit of $\widetilde{I}^n$. We summarize in the following corollary to Theorem 1.

Corollary 1. *If $\widetilde{X}^n(0) \Rightarrow X(0)$ in $\mathbb{R}^2$, then $(\widetilde{X}^n, \widetilde{Q}^n, \widetilde{I}^n) \Rightarrow (X, Q, I)$ in D^5 as $n \to \infty$, where (X_S, X_F, I) is characterized in* (3)–(5).

4.3. FDH Approximation for Limiting Distributions

Due to the abandonment, the process X^n, which is clearly an irreducible CTMC, is positive recurrent for each $n \geq 1$, and thus ergodic; in particular, it possesses a unique stationary distribution which is also its limiting distribution. One expects to have the FDH-scaled sequence of stationary distributions converge weakly as $n \to \infty$, to the stationary distribution of the FDH limit, but such a result is not guaranteed to hold in general. We note that the (marginal) stationary distributions of the processes $\widetilde{X}_S^n$, $n \geq 1$, have been shown to converge to the stationary distribution of the limiting Garnett diffusion in (Garnett et al. 2002, Appendix C). Here, however, we must prove the result for the sequence of *joint stationary distributions* of the processes $(\widetilde{X}_S^n, \widetilde{X}_F^n)$.

For a sequence of ergodic CTMCs that converges to a fluid limit, it is typical to have the corresponding sequence of stationary distributions converge to a stationary point of the fluid limit.

(A point x^* is stationary, if $X_F(t) = x^*$ for all $t \geq 0$, whenever $X_F(0) = x^*$.) However, the fluid part of the FDH limit X_F clearly keeps oscillating indefinitely, and therefore cannot possess a stationary point. Nevertheless, X_F is a *stochastic fluid limit*, and its driving diffusion process X_S *does possess a stationary distribution*, as was just mentioned. We use this latter fact to show that the FDH limit X is regenerative with a finite expected cycle length, thus possessing a unique limiting distribution. We then show that this limiting distribution is the weak limit of the stationary distributions of $\{\widetilde{X}^n : n \geq 1\}$ as $n \to \infty$.

Let $(X^n(\infty), Q^n(\infty))$ denote an $\mathbb{R}^4$ random variable having the limiting distribution of (X^n, Q^n), and define the FDH-scaled random variables

$$\widetilde{X}^n(\infty) := (\widetilde{X}^n_S(\infty), \widetilde{X}^n_F(\infty)) = \left(\frac{X^n_S(\infty) - n}{\sqrt{\lambda^n_S}}, \frac{X^n_F(\infty)}{\lambda^n_F} \right), \text{ and}$$
$$\widetilde{Q}^n(\infty) := \left(\widetilde{Q}^n_S(\infty), \widetilde{Q}^n_F(\infty)\right) = \left(\frac{Q^n_S(\infty)}{\sqrt{\lambda^n_S}}, \frac{Q^n_S(\infty)}{\lambda^n_F} \right).$$

Theorem 2. *The following hold:*

1. *The FDH process* (X, Q) *possesses a unique stationary distribution, which is also the limiting distribution, namely,* $(X(t), Q(t)) \Rightarrow (X(\infty), Q(\infty))$ *in* $\mathbb{R}^4$ *as* $t \to \infty$, *with*
$$Q(\infty) := (Q_S(\infty), Q_F(\infty)) = (X_S(\infty)^+, X_F(\infty)).$$
2. $(\widetilde{X}^n(\infty), \widetilde{Q}^n(\infty)) \Rightarrow (X(\infty), Q(\infty))$ *in* $\mathbb{R}^4$ *as* $n \to \infty$. *In particular,*
$$\lim_{n\to\infty} \lim_{t\to\infty} E[f(\widetilde{X}^n(t), \widetilde{Q}^n(t))] = \lim_{t\to\infty} \lim_{n\to\infty} E[f(\widetilde{X}^n(t), \widetilde{Q}^n(t))] = E[f(X(\infty), Q(\infty))],$$
for any bounded and continuous function $f : \mathbb{R}^4 \to \mathbb{R}$.
3. $\{\widetilde{Q}^n_i(\infty) : n \geq 1\}$ *is Uniformly Integrable (UI) for* $i = S, F$, *so that*
$$\lim_{n\to\infty} E[\widetilde{Q}^n_i(\infty)] = E[Q_i(\infty)].$$

The random variables $X_S(\infty)$ and $Q_S(\infty)$ are the steady-state distributions of the Garnett number-in-system and queue process, respectively; see part (2) of Theorem 2* in Garnett et al. (2002). Note that, just like the process X_F, the corresponding limiting distribution $X_F(\infty)$ has support on $[0, \theta_F^{-1})$, with a positive probability mass on state 0. Indeed, X_F^n can be bounded from above, in sample-path stochastic order, by an infinite server queue having service rate θ_F, giving the upper bound of the support of the limiting process X_F (see the proof of Theorem 4). Further, it follows from (4) that, if $X_S(t) < -r_F$, then X_F is strictly decreasing at time t. Since X_S is an ergodic diffusion process, it almost-surely experiences excursions below $-r_F$ for sufficiently long time intervals so as to allow the (bounded) process X_F to empty, and then remain at state 0 until X_S experiences an excursion in the set $[-r_F, \infty)$, which causes X_F to increase.

It is also worth noting that, since $X_F(\infty)$ has a positive probability mass at 0, the probability that a fast-class customer does not need to wait is positive asymptotically (as $n \to \infty$). Thus, even though the fast class is highly congested, and has fluid queue building up over much of the time, it does not strictly operate in the ED regime, as defined in Garnett et al. (2002); see Table 1 in this reference.

Approximating Performance Measures. Due to Theorem 2, we can use the limiting distribution of the FDH limit, as well as the expected values of the limiting FDH queues, to approximate key performance measures for the pre-limit stochastic system. For $i = S, F$ and for n large, we consider the following measures: the probability of delay in queue $P(W_i^n > 0)$; the average waiting time of delayed customers (including the waiting of the customers who eventually abandon the queue, but excluding customers who are not delayed) $E[W_i^n | W_i^n > 0]$; and the probability of abandonment $P(Ab_i^n)$.

The approximation of $P(W_S^n > 0)$ is straightforward: Since the event $\{W_S^n > 0\}$ is equivalent to the event $\{X_S^n \geq n\}$, both events have the same probability. Since $X_S(\infty)$ is a continuous random variable, the event $\{X_S(\infty) = 0\}$ has probability 0, and so we can approximate the limiting probability that the slow customers are delayed by $P(X_S(\infty) > 0) = P(Q_S(\infty) > 0)$.

The approximation of $P(W_F^n > 0)$ is more intricate, although it too can be approximated by the probability that the corresponding queue is strictly positive, namely, by $P(Q_F(\infty) > 0)$. The intricacy here is that $\widetilde{Q}_F^n(\infty) \Rightarrow Q_F(\infty)$ in $\mathbb{R}$ does not directly imply that $P(Q_F^n(\infty) > 0) \to P(Q_F(\infty) >$

0) as $n \to \infty$, because $Q_F(\infty) \equiv X_F(\infty)$ has a probability mass at 0; hence, the cumulative distribution function (cdf) of $X_F(\infty)$ is discontinuous at state 0. (Recall that weak convergence is defined to hold in continuity points of the limit cdf.) Nevertheless, we claim that $P(Q_F(\infty)=0)$ approximates $P(Q_F^n(\infty)=0)$ for large n, so that $P(Q_F(\infty)>0)$ also approximates $P(Q_F^n(\infty)>0)$. To see why, note that $\{Q_F(t)=0\}$ implies that $\{X_S(t) \le -r_F\}$, because Q_F is bounded from below by 0 and is strictly increasing whenever $X_S > -r_F$. Specifically, idleness appears in the system (so that Q_F is fixed at 0) immediately once Q_F reaches state 0 and $X_S < -r_F$, whereas Q_F begins to increase immediately when X_S crosses $-r_F$ from below. Therefore, in the limit, either the fluid queue of the fast class is strictly positive, and waiting times are positive, or the queue is empty, in which case there is idleness, and so no waiting.

The approximation for the expected waiting of delayed customers builds on the equality $E[W_i^n] = E[Q_i^n(\infty)]/\lambda_i^n$ which holds by virtue of Little's law, from which it follows that

$$E[W_i^n|W_i^n>0] = E[W_i^n]/P(W_i^n>0) = (\lambda_i^n)^{-1}E[Q_i^n(\infty)]/P(Q_i^n(\infty)>0).$$

Finally, we define the abandonment rate from queue i to be $\theta_i E[Q_i^n(\infty)]$.

To summarize, we have the approximations

$$P(W_S^n>0) \approx P(Q_S(\infty)>0), \quad E[W_S^n|W_S^n>0] \approx \frac{(\lambda_S^n)^{-1/2}E[Q_S(\infty)]}{P(Q_S(\infty)>0)}, \quad P(Ab_S^n) \approx \theta_S \frac{E[Q_S(\infty)]}{\sqrt{\lambda_S^n}}; \tag{6}$$

$$P(W_F^n>0) \approx P(Q_F(\infty)>0), \quad E[W_F^n|W_F^n>0] \approx \frac{E[Q_F(\infty)]}{P(Q_F(\infty)>0)}, \quad P(Ab_F^n) \approx \theta_F E[Q_F(\infty)]. \tag{7}$$

4.4. An Example

We now demonstrate the effectiveness of the FDH approximation by comparing its predictions to simulation of a stochastic system. The system we consider has $n=50$ servers that are fed by two independent Poisson processes having arrival rates $\lambda_S^n = 46$ and $\lambda_F^n = 15$. The service rates are $\mu_S^n = 1$ and $\mu_F^n = 5$, and the abandonment rates are $\theta_S = 0.1$ and $\theta_F = 0.3$. Note that the traffic intensity of the slow class $(\lambda_S^n/(n\mu_S^n) = 0.92)$ is close to 1, and that the service rate of the fast class is 5 times larger than that of the slow class. For the computation of the FDH approximation, we take

$$\beta = (n - \lambda_S^n)/\sqrt{\lambda_S^n} \quad \text{and} \quad r_F = \lambda_F^n/(\mu_F^n \sqrt{\lambda_S^n}). \tag{8}$$

The computation of the FDH limit is carried out numerically by generating 400 independent sample paths via the Euler scheme, as in Asmussen and Glynn (2007, Chapter X.3), using step size 0.002. To estimate the stationary performance measures of the stochastic system we averaged 400 independent simulation runs, each was ran for $1,000$ time units, and considered after a warm-up period of 100 time units. The results, given in Table 2, show that the FDH approximation is accurate for the four performance measures, and for each customer class. In particular, the relative errors of the limiting approximations for the expected queue lengths $E[Q_i^n(\infty)]$ and waiting times $E[W_i^n | W_i^n > 0]$, $i = 1, 2$, are less than 3%.

	Slow ($i = S$)		Fast ($i = F$)	
	simulation	FDH	simulation	FDH
$E[Q_i^n(\infty)]$	3.42 (0.03)	3.28 (0.02)	14.06 (0.07)	13.57 (0.07)
$E[W_i^n \| W_i^n > 0]$	0.18 (9e-4)	0.18 (8e-4)	1.28 (0.005)	1.29 (0.005)
$P(W_i^n > 0)$	0.41 (0.001)	0.39 (0.001)	0.73 (0.001)	0.70 (0.001)
$P(Ab_i^n)$	0.01 (6e-5)	0.01 (5e-5)	0.28 (0.001)	0.27 (0.001)

Table 2 Comparison of performance measures for a stochastic system and its FDH approximation. The "simulation" columns give the results for the stochastic system, and the "FDH" columns show the results for the FDH approximation. Standard errors are presented in parentheses.

It is useful to contrast the simulation results with existing many-server asymptotic approximations. Specifically, recall from Section 1.2 that, under existing MSHT limiting regimes, one of the following three scenarios must hold asymptotically: (I) neither class experiences any delay; (II) both classes are served, and all the delay is experienced by the lower-priority class; (III) the slow class experience delay, in which case the fast class receives no service, asymptotically. Clearly, none of these three scenarios is consistent with the simulation results presented in Table 2, as the slow class has a significant delay (0.18 time units) while most of the customers (72%) of the fast class are served.

To demonstrate that the limiting distribution of the FDH approximates well the limiting distribution of the stochastic system (beyond the means), we compared the (marginal) limiting cdf's of the two simulated queues to the corresponding FDH distributions. The results are depicted in Figure 2.

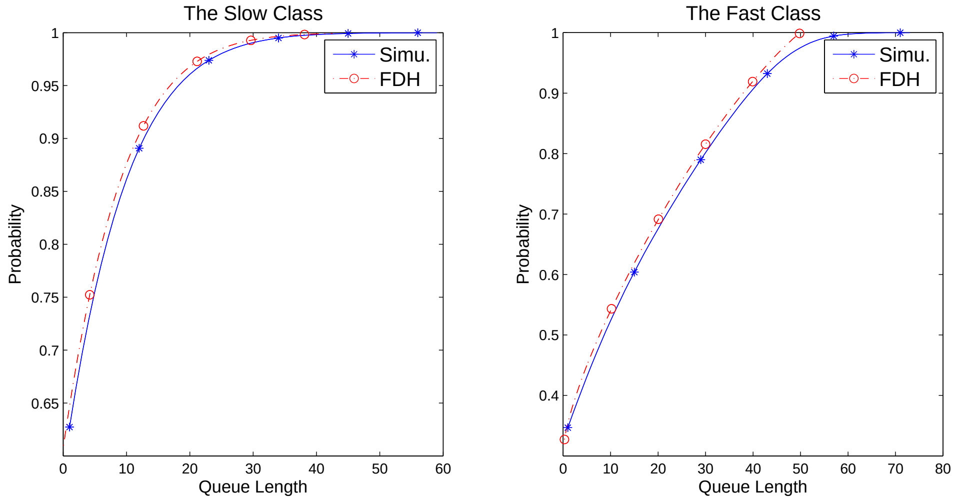

Figure 2 The marginal cdf's computed from simulations of $Q_S^n(\infty)$ and $\widetilde{Q}_F^n(\infty)$ (solid, starred line) and the corresponding cdf's computed for the FDH approximation $(\sqrt{\lambda_S^n}Q_S(\infty), \lambda_F^n Q_F(\infty))$ (dashed, circled line).

5. The FDH Limit for the N-System

We now consider the FDH approximation for the N-system. For comparison purposes, we think of the single pool of the V-system as being split into two distinct pools: a "regular track" which, as before, serves both classes with strict priority to the slow class, and a "fast track," which is dedicated to serving the fast class. Such a system design often makes sense, because it provides some of the benefits of pooling while requiring only part of the agents to be cross-trained. As we show below, the N-system design is especially useful in our setting, since a small number, that is asymptotically negligible, of dedicated agents can dramatically decrease the waiting times of the fast class, while maintaining good service levels for the slow class.

The benefits of having a fast track are especially pronounced when the dedicated pool is cheaper to operate, which is often the case in practice. For example, in the hospital setting, residents can replace physicians in the ER's fast track, and the required nurse-to-patient ratio in observation units is lower than in general inpatient units. In the contact-center setting, agents that handle inbound calls (slow customers) and emails, may receive higher pay and be more costly to train, than agents that only respond to emails.

The Setting. We assume that the arrival processes, patience and service times are as in Section 4. We further assume that the service time distribution of the fast class is the same in both pools, namely, the service times are class-dependent, and are not pool-dependent. As before, the slow customers receive preemptive priority over the fast customers in the regular track. However, an interrupted service due to preemption can be resumed in the fast track. We let z^n denote the number of servers in the fast track in system n, and assume that

$$\lim_{n\to\infty} z^n/R_F^n = z, \text{ for some } z \in [0,1],$$

so that z is the limiting capacity of the fast track. In particular, the case $z=0$, corresponding to having no fast track, will be seen shortly to agree with the corresponding limit for the single-pool V model. On the other hand, when $z=1$, all the fast customers are served in the fast track. Note that the number of servers assigned to the fast track is $z^n = O(\sqrt{n})$ and in particular, $z^n/\sqrt{n} \to r_F z$ as $n\to\infty$ by Assumption 2.

We let $X^{z,n} := (X_S^{z,n}, X_F^{z,n})$ denote the number-in-system process, $Q^{z,n} := (Q_S^{z,n}, Q_F^{z,n})$ denote the queue-length process, and $I^{z,n}(t) := \int_0^t (n - X_S^{z,n}(s) - X_F^{z,n}(s))^+ ds$ denote the cumulative idleness process for a given z in system n (so that the fast track size in that nth system is z^n). The FDH scaling is as follows

$$\widetilde{X}^{z,n} := (\widetilde{X}_S^{z,n}, \widetilde{X}_F^{z,n}) := \left(\frac{X_S^{z,n} - (n - z^n)}{\sqrt{\lambda_S^n}}, \frac{X_F^{z,n}}{\lambda_F^n} \right);$$

note that we center the process $X_S^{z,n}$ about the number of servers in the regular track $n - z^n$.

Let $\widetilde{Q}^{z,n}$ and $\widetilde{I}^{z,n}$ be the FDH-scaled versions of the processes just defined, as in (2). We make the following assumption in order to avoid a jump at time 0 in the limiting process.

Assumption 3 **(Initial condition for the N-system)**. *$Q_S^{z,n}(0) = (X_S^{z,n}(0) - (n - z^n))^+$ and $Q_F^{z,n}(0) \geq 0$ for all $n \geq 1$.*

The following theorem provides the FDH limit for the N-system as the solution to an HSDE.

Theorem 3 **(FDH limit for the N-system)**. *If $\widetilde{X}^{z,n}(0) \Rightarrow X^z(0)$ in $\mathbb{R}^4$ and, in addition, Assumptions 1 and 3 hold, then $\left(\widetilde{X}^{z,n}(t), \widetilde{Q}^{z,n}(t), \widetilde{I}^n(t)\right) \Rightarrow \left(X^z(t), Q^z(t), I^z(t)\right)$ in D^5 as $n \to \infty$, where the component process of X^z is the unique solutions to the HSDE*

$$dX_S^z(t) = \left(-\beta + r_F z + X_S^z(t)^- - \theta_S X_S^z(t)^+\right) dt + \sqrt{2} dB(t), \tag{9}$$

$$dX_F^z(t) = (1 - z - r_F^{-1} X_S^z(t)^- - \theta_F X_F^z(t))dt + dI^z(t) \quad \text{and} \quad X_F^z(t) \geq 0, \tag{10}$$

where B is a standard Brownian motion, $Q^z := ((X_S^z)^+, X_F^z)$, and I^z is the unique non-decreasing process satisfying

$$I^z(0) = 0 \quad \text{and} \quad \int_0^t 1_{\{X_F^z(s) > 0\}} dI^z(s) = 0, \quad \text{for all } t \geq 0. \tag{11}$$

Observe the similarity between the FDH limit for the N-system and for the V-system in Theorem 1. In particular, (9) becomes (3) if we replace $\beta - r_F z$ by β, while (10) becomes (4) if we scale both sides by $1 - z$. Thus, $(X_S^z, (1-z)^{-1} X_F^z)$ is the FDH limit for a sequence of V-systems, in which the number of servers in the nth system is reduced by z^n, while the fast-class arrival in the nth system is reduced by $\mu_F^n z^n$.

It is useful to consider the two extreme values of z, $z = 0$ and $z = 1$, to see how the FDH limit X^z depends on z: (i) When $z = 0$, the N-system reduces to the V-system; indeed, the expressions in (9) and (10) reduce to the expressions in (3) and (4), respectively. Therefore, the FDH limit for the single-pool model is a special case of the FDH limit for the N-system. (ii) When $z = 1$, (10) implies that $X_F^z(\infty)$ is identically zero. In this case, both classes have asymptotically negligible delay, implying that only a relatively negligible proportion of the arrivals abandon asymptotically. Compared to the V-system, in which a non-negligible portion of fast-class customers abandon the system, we conclude that a fast track can significantly increase the throughput of the system; a numerical example is presented in Appendix A.4.

We note that having no fluid queue for the fast class may not be desirable, because, in this case, the delay of the fast class may not be sufficiently larger than the delay of the slow class, which should receive high priority. Given the imposed priority, this implicitly means that too much of the service resources are taken from the high-priority class in order to reduce delays for the low-priority class. There are therefore clear tradeoffs that must be taken into account when deciding whether a fast-track should be operated, and what its size should be. We formalize this problem under a cost structure in Section 6.

5.1. FDH Approximation for the Limiting Distribution

Similar to Theorem 2, we can show that the limiting distribution of the FDH limit exists and is also the limit of the sequence of stationary versions of the processes, $(\widetilde{X}^{z,n}, \widetilde{Q}^{z,n})$, which we denote by $(\widetilde{X}^{z,n}(\infty), \widetilde{Q}^{z,n}(\infty))$, respectively.

Theorem 4. *For each $z \in [0,1]$, the following hold:*

1. *The FDH process (X^z, Q^z) possesses a limiting distribution $(X^z(\infty), Q^z(\infty))$, namely, $(X^z(t), Q^z(t)) \Rightarrow (X^z(\infty), Q^z(\infty))$ as $t \to \infty$ in $\mathbb{R}^4$, where*

$$Q^z(\infty) := (Q^z_S(\infty), Q^z_F(\infty) = (X^z_S(\infty)^+, X^z_F(\infty)). \qquad (12)$$

2. *$(\widetilde{X}^{z,n}(\infty), \widetilde{Q}^{z,n}(\infty)) \Rightarrow (X^z(\infty), Q^z(\infty))$ in $\mathbb{R}^4$ as $n \to \infty$. In particular,*

$$\lim_{n\to\infty} \lim_{t\to\infty} E[f(\widetilde{X}^{z,n}(t), \widetilde{Q}^{z,n}(t))] = \lim_{t\to\infty} \lim_{n\to\infty} E[f(\widetilde{X}^{z,n}(t), \widetilde{Q}^{z,n}(t))] = E[f(X^z(\infty), Q^z(\infty))],$$

for any bounded and continuous function $f : \mathbb{R}^4 \to \mathbb{R}$.

3. *For $i = S, F$ the sequences $\{(\widetilde{Q}^{z,n}_i(\infty)) : n \geq 1\}$ are UI, so that*

$$\lim_{n\to\infty} E[\widetilde{Q}^{z,n}_i(\infty)] = E[Q^z_i(\infty)].$$

Analogously to (6) and (7), Theorem 4 allows us to employ the limiting distribution of FDH limit to approximate key performance measures for each class when $z < 1$. When $z = 1$, there is sufficient service capacity in the fast track to ensure that the fast queue is not overloaded under fluid scaling, namely, $Q^z_F(\infty) = 0$ w.p.1, so that more refined asymptotic analysis is required in order to approximate the queue of the fast class. As before, the established UI can be used to approximate performance measures corresponding to the limiting distributions of the queues. In Section 6 below we use it to optimize expected costs.

6. Employing the FDH Limit to Optimize System Design

It is often the case that a fast track is considered because the slow customers must receive strict priority in the regular pool over the fast customers. The fast track is then used in order to "circumvent" this policy constraint, by having a small pool that is dedicated to the low-priority customers. On the other hand, the fast track is taking resources away from the regular pool, and so introduces a non-trivial cost-benefit tradeoff. Indeed, in a private communication with the management of a large hospital in Chicago, we were told that a fast track is operated in order to attract low-acuity patients, since those patients provide large revenues, but require simple (and thus, cheaper) treatments. In a different hospital, we were told that the fast track was recently eliminated, in order to deter low-acuity patients from arriving to the ER.

We now demonstrate how the FDH approximation can be employed to optimize (asymptotically) systems' design when holding, abandonment, and staffing costs are incurred. Specifically, we consider an N-system, and employ the FDH limit to establish the size of the fast track that asymptotically minimizes the incurred cost (where we recall that $z = 0$ corresponds to having no fast track).

For $i = S, F$, let a^n_i denote the cost incurred per abandoning class-i customer, and h^n_i denote the rate at which holding costs are incurred in system n. Let d^n_R and d^n_F be the per-server cost in the regular track and the fast track, respectively. For a system with z^n fast track servers and $n - z^n$ regular-track servers, the cost has the form of:

$$\sum_{i=S,F} (h^n_i E[Q^{z,n}_i(\infty)] + a^n_i \theta^n_i E[Q^{z,n}_i(\infty)]) + d^n_R(n - z^n) + d^n_F z^n.$$

Let $d^n := d_F^n - d_R^n$ and $c_i^n := h_i^n + \theta_i a_i^n$. Since the term nd_R^n has no impact on the optimal solution, we consider the objective function

$$C^n(z^n) := \sum_{i=S,F} c_i^n E[Q_i^{z,n}(\infty)] + d^n z^n. \tag{13}$$

Minimizing $C^n(\cdot)$ is clearly prohibitive because the stationary distribution of the system is hard to compute for any given value of z^n. However, an asymptotically optimal system design can be efficiently computed by utilizing the FDH limit, as we show below. The interesting (non-trivial) case to consider is when the total costs of queueing for both classes are proportional, implying that the cost incurred due to queueing of the slow class is significantly higher than the cost incurred by the fast class. Indeed, unlike low-acuity patients, the condition of high-acuity patients may deteriorate if they do not receive treatment in a timely manner. Similarly, there is typically more flexibility regarding when to process outbound work in contact centers than there is regarding inbound customers, who expect to receive service quickly. Since the fast queue is $O_P(n)$ while the slow queue is $O_P(\sqrt{n})$ in the FDH scaling, we therefore assume that $c_F^n/c_S^n = O(n^{-1/2})$. We further assume that the staffing costs corresponding to agents working only with the fast class are lower than those corresponding to the slow class. Formally,

Assumption 4. *$c_S^n = c_S$, $c_F^n = c_F/\mu_F^n$, and $d_F^n - d_R^n = d \le 0$.*

Then by virtue of Assumption 4 and Theorem 4(3), we have that

$$C(z) := \lim_{n\to\infty} n^{-1/2} C^n(R_F^n z) = c_S E[Q_S^z(\infty)] + c_F r_F E[Q_F^z(\infty)] + d r_F z, \tag{14}$$

where we utilized the fact that $n^{-1/2}\sqrt{\lambda_S^n} \to 1$ as $n \to \infty$. Let

$$z^* := \arg\min_{z\in[0,1]} C(z). \tag{15}$$

For z^* to be well-defined, we need the following lemma.

Lemma 1. *$z \mapsto E[Q_i^z(\infty)]$ is continuous in $[0,1]$ for $i = S, F$.*

The value of z^* can be numerically computed using grid search; it is relevant for the pre-limit stochastic system since it asymptotically minimizes the operating cost (under the control we consider), as we prove next.

Consider a sequence of systems with a corresponding sequence of fast tracks $\{z^n : n \geq 1\}$. To avoid having redundant service capacity in the fast track, which is clearly sub-optimal, we assume that

$$\limsup_{n\to\infty} \frac{z^n}{R_F^n} \leq 1. \tag{16}$$

For $x \in \mathbb{R}$, let $\lfloor x \rfloor$ denote the largest integer that is smaller than or equal to x.

Proposition 1. *$z^{n*} := \lfloor R_F^n z^* \rfloor$ asymptotically minimizes $C^n(z^n)$, in the sense that*

$$\limsup_{n\to\infty} \frac{1}{\sqrt{n}}(C^n(z^{n*}) - C^n(z^n)) \leq 0$$

for any sequence $\{z^n : n \geq 1\}$ that satisfies (16).

6.1. Structural Results

We can say more about the limiting cost function $C(\cdot)$ in (14) and z^* if we impose more assumptions on the system's parameters. First, we require that $\theta_S < \mu_S$. This condition tends to hold in service systems, as reviewed in Gans et al. (2003) (which mentions that the rate of abandonment rate of customers tends to be about half that of their service rate). This also suggests our second requirement, that $\theta_S < \theta_F$ (because $\mu_S^n \ll \mu_F^n$). Finally, consistent with the imposed priority rule, we assume that the $c\mu$-type condition $c_S^n \mu_S^n > c_F^n \mu_F^n$ holds. (Loosely speaking, this condition suggests that delaying a slow-class customer is more costly than delaying a fast-class customer, even after incorporating their service times.) Due to Assumption 4, this $c\mu$ condition is equivalent to the assumption that $c_S > c_F$. We summarize these three conditions in the following formal assumption, which is assumed to hold throughout this section, in addition to Assumptions 1, 3 and 4.

Assumption 5. *$\theta_S < \mu_S$, $\theta_S < \theta_F$ and $c_S > c_F$.*

Under this extra assumption, we can prove important structural results for the limiting cost function $C(\cdot)$ in (14).

Proposition 2. *$C:[0,1]\to\mathbb{R}$ is strictly convex. Hence, there exists a unique minimizer z^* to (15).*

Together with the continuity of $C(\cdot)$, Proposition 2 implies that a simple binary search can efficiently find the global minimizer z^*.

Quantifying the Tradeoffs of Having a Fast-Track. Even though a fast track reduces the waiting time of the fast class and increases the throughput of the system, it increases the delays of slow class, and thus the overall delay cost. Specifically, let

$$C_q(z) := c_S E[Q_S^z(\infty)] + c_F r_F E[Q_F^z(\infty)],$$

and note that $C(z) = C_q(z) + dr_F z$. The second term $dr_F z$ corresponds to the fast-track staffing cost, whereas $C_q(\cdot)$ is the cost corresponding to the queues (holding and abandonment costs), and thus the delays. Since the fast-track staffing cost is smaller than the staffing cost of the main pool, the following proposition demonstrates that there is a clear tradeoff in operating a fast-track, as it increases the overall queueing cost.

Proposition 3. *$C_q:[0,1]\to\mathbb{R}_+$ is convex and strictly increasing.*

In ending we remark that, unlike the function C in the limit, the function C^n need not be convex for any given $n\in\mathbb{Z}_+$. For example, take $n=2$, $\lambda_S^n=10$, $\lambda_F^n=3$, $\mu_S^n=1$, $\mu_F^n=2$, $\theta_S=0.999$, $\theta_F=5$, $c_S^n=3$, $c_F^n=1$, and $d^n=0$. (Note that n is too small for the FDH approximation to be accurate). One can check that Assumption 5 is satisfied. We take $z_i^n=i$ for $i=0,1,2$ and let

$$\Delta := C^n(z_0^n) + C^n(z_2^n) - 2C^n(z_1^n).$$

A discrete event simulation with 400 replications reports $\Delta=-0.12$ with standard deviation 0.0003, suggesting that C^n is not convex in z^n.

7. Numerical Studies

We now present a numerical and simulation study in which we compare the FDH predications to simulations of the stochastic system it approximates. In particular, in §7.1 we demonstrate how the accuracy of the FDH approximation increases together with the size of the system. We perform a sensitivity analysis in §7.2, which demonstrates the robustness of the FDH approximation. Finally, in §7.3 we explain why the dynamics under the non-preemptive version of the strict-priority policy are asymptotically indistinguishable from the dynamics under the preemptive priority policy we considered. We support that explanation with simulation.

7.1. A Numerical Demonstration of the Convergence to the FDH Limit

Since the FDH approximation is obtained as a weak limit for stochastic systems with many servers, one expects its accuracy to improve as the size of the system increases. The following example shows that this is indeed the case, although the limit provides a good approximation also for a relatively small system, with only $n = 25$ agents. For the examples we consider, we take $\mu_S = 1$, $\beta = 0.5$, $r_F = 0.3$, $\theta_S = 0.1$, and $\theta_F = 0.3$ and vary the number of agents n, giving it the values in $\{25, 100, 400\}$. For each n we consider two values of the fast service rate, $\mu_F^n = \sqrt{n}$ and $\mu_F^n = 0.5\sqrt{n}$. We take these two values of μ_F^n because $\mu_F^n = \sqrt{n}$ is extremely large when $n = 400$, and $\mu_F^n = 0.5\sqrt{n}$ is quite small when $n = 25$. The values of λ_S^n and λ_F^n are chosen so as to satisfy (8). We compare the simulated values of $E[\widetilde{Q}_i^n(\infty)]$ and $P(W_i^n > 0)$, $i = S, F$, to their respective FDH approximations, where, for each of the six systems, we employ the same procedures as in the numerical example in Section 4.4 for the simulation of the stochastic system and the numerical solution for its FDH approximation. The results are shown in Table 3.

We observe that the accuracy of the approximations increases with n. The error is relatively large when $n = 25$ and $\mu_F^n = 0.5\sqrt{n} = 2.5$, as should be expected. Nevertheless, despite the lesser accuracy in this case, the limit still captures the key feature for which the FDH approximation is developed; in particular, the high-priority (slow) class operates in a QED-type fashion (its probability of delay is substantially larger than 0 and smaller than 1), while the low-priority (fast) class operates in an ED-type fashion. Note that, since the FDH approximation for the fast class is based on a fluid limit, the lesser accuracy for small systems is to be expected, because the stochastic fluctuations (which are not captured by the fluid approximation), are substantial relative to the “predictable dynamics" of the fluid limit. (Loosely speaking, the fluid limit captures dynamics that are $\Theta_P(n)$, while the stochastic fluctuations are $\Theta_P(\sqrt{n})$. For n small, the two orders are indistinguishable.)

		Discrete-Event Simulation			FDH
		$n=25$	$n=100$	$n=400$	
$\mu_F^n=\sqrt{n}$	$E[\widetilde{Q}_S^n(\infty)]$	0.62 (0.005)	0.61 (0.005)	0.60 (0.004)	0.59 (0.004)
	$P(W_S^n>0)$	0.47 (0.002)	0.46 (0.002)	0.45 (0.001)	0.44 (0.001)
	$E[\widetilde{Q}_F^n(\infty)]$	0.97 (0.005)	0.94 (0.005)	0.93 (0.005)	0.92 (0.005)
	$P(W_F^n>0)$	0.71 (0.001)	0.70 (0.001)	0.68 (0.001)	0.67 (0.001)
$\mu_F^n=0.5\sqrt{n}$	$E[\widetilde{Q}_S^n(\infty)]$	0.62 (0.005)	0.61 (0.004)	0.60 (0.004)	0.59 (0.004)
	$P(W_S^n>0)$	0.47 (0.002)	0.46 (0.001)	0.45 (0.001)	0.44 (0.001)
	$E[\widetilde{Q}_F^n(\infty)]$	1.03 (0.005)	0.96 (0.005)	0.93 (0.005)	0.92 (0.005)
	$P(W_F^n>0)$	0.74 (0.001)	0.71 (0.001)	0.70 (0.001)	0.67 (0.001)

Table 3 Comparison of the FDH predictions to simulation results for three components of a sequence of systems. Standard errors for the simulations are presented in parentheses.

We used the simulation experiments to approximate the cdf's of the stationary distributions of the fast-class queue in the three systems with $\mu_F^n=0.5\sqrt{n}$, and compare these cdf's to the corresponding cdf of the limiting distribution for the FDH approximation. The result, depicted in Figure 3, illustrates the weak convergence of the stationary distribution of the queue to the corresponding distribution of the FDH limit.

7.2. Sensitivity Analysis

Recall that the FDH limit was achieved by assuming that $1-\rho_S^n=O(n^{-1/2})$ (where $\rho_S^n:=\lambda_S^n/(n\mu_S)$ and $\mu_S=1$), and $\mu_F^n=O(\sqrt{n})$. Therefore, the FDH limit may not be a proper approximation when ρ_S^n is significantly smaller than 1, or when μ_F^n is not sufficiently larger than 1. To test how the values of ρ_S^n and μ_F^n affect the accuracy of the FDH approximation, we conduct a sensitivity analysis with three values of ρ_S^n and μ_F^n, for a total of nine different examples. We fix the number of agents to be $n=50$ and take the offered load to be equal to the service capacity, namely, $\lambda_S^n/\mu_S^n+\lambda_F^n/\mu_F^n=n$.

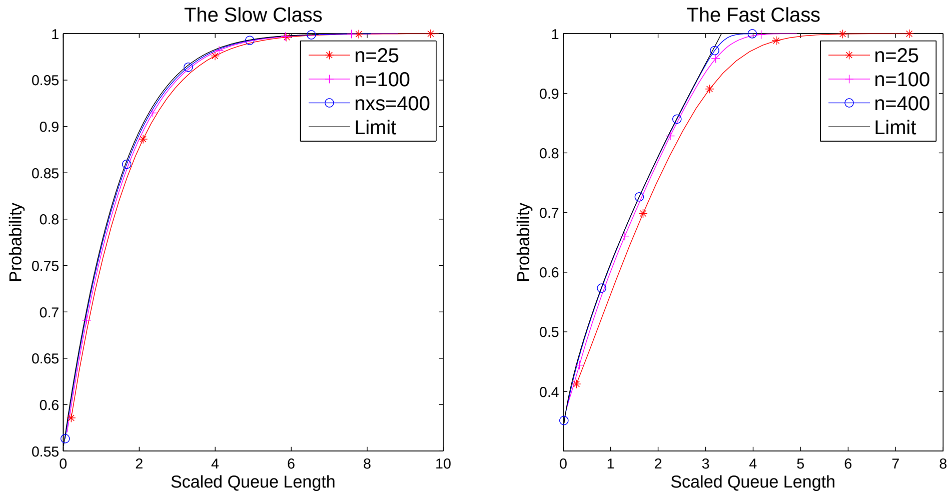

Figure 3 Empirical cdf's of $\widetilde{Q}_S^n(\infty)$ (left panel) and $\widetilde{Q}_F^n(\infty)$ (right panel) for $n \in \{25, 100, 400\}$, plotted together with the empirical cdf of $Q_S(\infty)$ (left panel) and $Q_F(\infty)$ (right panel).

The abandonment rates are fixed at $\theta_S = 0.1$ and $\theta_F = 0.3$. The results for the nine combinations are shown in Table 4. To facilitate the comparison between the different experiments, we show the expected values of the FDH-scaled queues.

The results in Table 4 make it clear that, as expected, the accuracy of the FDH approximation is sensitive to the value of μ_F^n. In particular, the FDH approximations for $E[\widetilde{Q}_F^n(\infty)]$ and $P(W_F^n > 0)$ have the largest errors when $\mu_F^n = 2$, while the error is significantly smaller for the larger two values of μ_F^n. (Note that the FDH approximation for $E[\widetilde{Q}_S^n(\infty)]$ and $P(W_S^n > 0)$ does not depend on μ_F^n.) Nevertheless, the FDH limit still exhibits the behavior and the main qualitative features it is designed to capture in this case.

On the other hand, the accuracy of the FDH approximation is not very sensitive with respect to ρ_S^n. For small ρ_S^n, i.e., $\rho_S^n = 0.75$, the results show that the slow class does not operate in the QED regime, because the probability of delay is close to 0. Of course, this is simply an indication that the traffic intensity of the slow class is too low for the QED regime to be an appropriate limiting approximation. In particular, an Erlang-A model with the same parameters λ_S, μ_S, θ_S and n as in this example is better approximated by the QD regime. Despite this, the FDH approximation is still a good quantitative approximation, especially for the larger values of μ_F^n, and it clearly captures the qualitative behavior of the simulated stochastic systems well.

		Discrete-Event Simulation			FDH
		$\mu_F^n=2$	$\mu_F^n=5$	$\mu_F^n=10$	
$\rho_S^n=0.75$	$E[\widetilde{Q}_S^n(\infty)]$	0.02 (0.002)	0.02 (0.002)	0.02 (0.002)	0.01 (0.002)
	$P(W_S^n>0)$	0.03 (0.001)	0.03 (0.001)	0.03 (0.001)	0.02 (0.001)
	$E[\widetilde{Q}_F^n(\infty)]$	0.47 (0.002)	0.44 (0.002)	0.43 (0.002)	0.41 (0.002)
	$P(W_F^n>0)$	0.77 (0.001)	0.77 (0.001)	0.77 (0.001)	0.76 (0.001)
$\rho_S^n=0.85$	$E[\widetilde{Q}_S^n(\infty)]$	0.14 (0.003)	0.14 (0.003)	0.14 (0.003)	0.12 (0.003)
	$P(W_S^n>0)$	0.18 (0.001)	0.18 (0.001)	0.18 (0.001)	0.16 (0.001)
	$E[\widetilde{Q}_F^n(\infty)]$	0.75 (0.003)	0.71 (0.003)	0.69 (0.003)	0.68 (0.003)
	$P(W_F^n>0)$	0.79 (0.001)	0.78 (0.001)	0.77 (0.001)	0.76 (0.001)
$\rho_S^n=0.95$	$E[\widetilde{Q}_S^n(\infty)]$	0.84 (0.006)	0.84 (0.006)	0.84 (0.006)	0.82 (0.005)
	$P(W_S^n>0)$	0.54 (0.001)	0.54 (0.001)	0.54 (0.001)	0.53 (0.001)
	$E[\widetilde{Q}_F^n(\infty)]$	1.38 (0.006)	1.31 (0.006)	1.29 (0.006)	1.27 (0.005)
	$P(W_F^n>0)$	0.83 (0.001)	0.81 (0.001)	0.79 (0.001)	0.78 (0.001)

Table 4 Sensitivity analysis for the accuracy of the FDH approximation. Standard error of the simulation experiments are presented in parentheses. The expected queue lengths are scaled according to the FDH scaling.

7.3. Non-Preemptive FDH Approximation

We now provide a *high-level* explanation as to why the queueing dynamics under the priority policy with no preemption are asymptotically (as $n\to\infty$) indistinguishable from the dynamics under the preemptive policy we analyzed. The explanation is given for the V-system, as similar arguments apply for the N-system.

Let $Z_F^n(t)$ and $Z_S^n(t)$ denote the number of agents at time t that are working with fast and slow customers, respectively, in system n. Now, the scaling of μ_F^n implies that $Z_F^n = O_P(\sqrt{n})$. Therefore, if a queue of slow customers is building up, then $O_P(\sqrt{n})$ fast customers are removed from service and added to their queue under the preemptive policy, a quantity that is negligible under the spatial fluid scaling of that queue. In particular, even if all the fast customers in service were removed and put back in their queue instantaneously, there would be no impact on the limiting queue $\widetilde{Q}_F$. Further, $Z^n_F = o_P(n)$ under either policy (indeed, $\widetilde{Q}_F = \widetilde{X}_F$), showing that the processes corresponding to the fast class are indistinguishable under the two policies in the FDH limit.

The reasoning as to why the processes corresponding to the slow class under the non-preemptive policy are unchanged asymptotically is more intricate, but again follows from the scaling of μ^n_F. Due to this scaling, the total output rate of fast customers from service is $\Theta_P(n)$ whenever $Z^n_F(t) = \Theta_P(\sqrt{n})$. This suggests that, if a queue of slow customers is starting to build up, the number of fast customers in service will drop to $o_P(\sqrt{n})$ in $o_P(1)$ time under the non-preemptive policy, because no new fast customers will be routed into service. In fact, the total service rate of all fast customers in service combined is always an order $\sqrt{n}$ larger than the order of the number of those customers. Specifically, if $Z^n_F(t) > 0$ and $Q^n_S(t) > 0$ for all $t \in [t^n_1, t^n_2]$, $0 \le t^n_1 < t^n_2 < \infty$, then Z^n_F behaves like a pure death process over this time interval, with death rates $k\mu^n_F = \Theta(k\sqrt{n})$, $k = 1, 2, \ldots$. It follows that for any $\epsilon > 0$, the sequence of events

$$B^n(\epsilon) := \{\{Z^n_F(t) > 0\} \cap \{Q^n_S(t) > 0\} : \; t \in [t^n_1, t^n_2], \; t^n_2 - t^n_1 > \epsilon\},$$

satisfies $P(B^n(\epsilon)) \to 0$ as $n \to \infty$, where P is the probability measure in the underlying probability space. In other words, having fast customers in service and slow customers in queue simultaneously over an interval is an asymptotically null event. (It is significant that the events $B^n(\epsilon)$ are defined in terms of the *unscaled* processes Z^n_F and Q^n_S.) In turn, whenever a queue of the slow class builds up in the limiting system, the number of fast customers in service drops to 0 instantaneously, so that all the service capacity is dedicated to the slow class, just like the case in which preemption is exercised.

We do not attempt to rigorously prove the asymptotic equivalence between the policies. Instead, we demonstrate that the dynamics of the queues are similar under both policies via simulation. Figure 4 plots two sample paths for the system considered in Section 4.4, with $n = 50$, $\lambda^n_S = 46$, $\lambda^n_F = 15$, $\mu^n_S = 1$, $\mu^n_F = 5$, $\theta_S = 0.1$ and $\theta_F = 0.3$. The two sample paths shown in the figure were generated by giving both the same arrival process of customers, with each customer having the same patience and service-time requirement. As can be seen, the two sample paths are in close

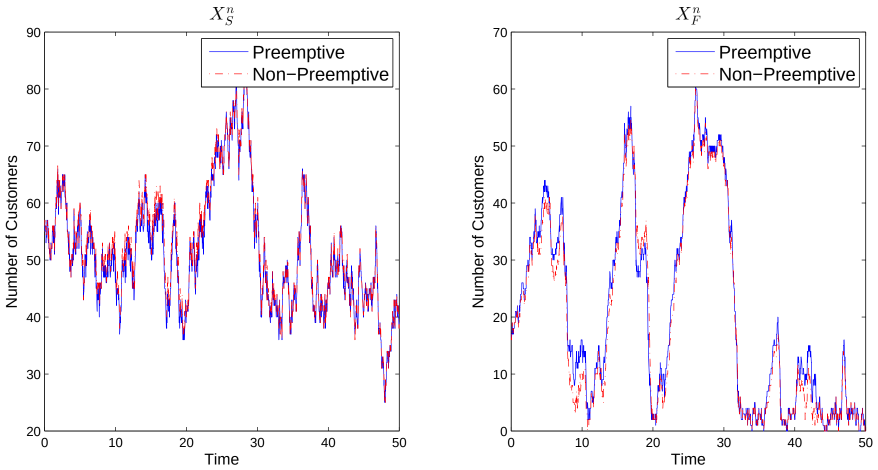

Figure 4 Sample path comparison of X_S^n (left) and X_F^n (right) in a system with $n = 50$ agents, operating under the preemptive and non-preemptive priority policy. The starred lines plot the sample paths under the preemptive policy, and the circled lines plot the sample paths under the non-preemptive policy.

agreement with each other. We also mention that the stationary performance measures are similar under the two policies. In particular, the values of $(E[X_S^n(\infty)],\, E[X_F^n(\infty)])$ are estimated to be $(49.1, 16.2)$ for the preemptive policy, and $(49.7, 14.6)$ for the non-preemptive policy, with standard errors smaller than 0.04.

8. Summary

In this paper, we proposed a fluid-diffusion hybrid process to approximate two-customer class many-server systems that operate under a priority policy. We assumed that the high-priority (“slow”) customers require substantially longer service times than the low-priority (“fast”) customers. The need to develop the FDH approximation stems from the fact that existing MSHT approximations cannot capture the setting in which both customer classes are delayed in queue with a non-negligible probability, and yet most customers, from either class, end up receiving service.

We first considered the V-system, in which the two customer classes are served by a single pool of agents, and then the N-system, in which one pool handles both customer classes (giving strict priority to the slow class), and the other pool, which we named “fast track,” is dedicated to the fast class. For both systems, we characterized the FDH limit, and proved that it possesses a limiting distribution, which is also the weak limit for the sequence of stationary distributions of the underlying sequence of systems. As we demonstrated via numerical examples, the FDH limit can be used to approximate key performance measures of the underlying stochastic system when the basic assumptions of the model hold. Sensitivity analysis demonstrated the robustness of the FDH approximation in that the main qualitative insights remain to hold even when it is questionable whether these assumptions are satisfied.

In Section 6 we demonstrated how the FDH limit can be employed to determine the asymptotically optimal system topology. In particular, we considered whether it is beneficial to split the server pool into two pools, and to determine the optimal size of the "fast-track" pool in the limit, assuming a linear holding and abandonment cost is incurred. One can employ the FDH regime and the framework we developed here in other optimization settings, such as in finding an asymptotically optimal control for either the one- or the two-pool system, when the priority policy is not enforced. Such implementations are currently under investigation.

Acknowledgments

We thank the AE and two anonymous referees for making many useful comments and suggestions which led to the results in Section 6.1 and to Appendix A. Ohad Perry and Lun Yu were partially supported by NSF grant CMMI 1763100.

Endnotes

1. We use ER instead of the now-common ED (for Emergency Department) to avoid confusion with the acronym for Efficiency Driven, which will be used repeatedly throughout the paper

Appendices:

The appendix is organized as follows: In Appendix A we consider three generalizations to the setting considered in the main paper: (i) generalized FDH scaling for the prelimit; (ii) implications of our analysis to systems with time-varying arrival processes; (iii) systems with general service-time distributions. Appendices B–E are devoted to the proofs of the results in the main paper.

Appendix A: Generalizations to the FDH Approximation

We consider several generalizations to the FDH regime. In Section A.1 we consider a general FDH scaling, assuming the service rate of the fast class scales like $\Theta(n^\alpha)$, for $1/2 \le \alpha < 1$; In Section A.2, we consider the FDH regime with time-varying arrival process; In Section A.3, we consider the FDH regime with general service time distributions. With each generalization, we show how we can apply the FDH approximation to the non-standard system settings.

A.1. Other Scalings of the Fast Class

In the FDH regime we assume the service rate of the fast class is scaled like $\sqrt{n}$, $\mu_F^n = O(\sqrt{n})$. This singular perturbation approach of the service times is a technical artifact that is taken to achieve a non-trivial limiting process with desirable characteristics, whenever the slow class operates in the QED regime (which is itself achieved by carefully balancing the arrival and service rates). Indeed, the assumption that $\mu_F^n = O(\sqrt{n})$ (which is carried out in Assumption 1 through $\lambda_F^n/n \to \lambda_F$ and $r_F^n \to r_F$ as $n \to \infty$) can be relaxed, as long as R_F^n keeps its $O(\sqrt{n})$ order. In particular, all results in the paper remain valid with the same proof, under the following generalized assumption:

Assumption 6 **(generalized FDH scaling)**. *For $\beta \in \mathbb{R}$ and $\theta_S > 0$, the following holds for the slow class.*

$$\lim_{n\to\infty} (n - \lambda_S^n)/\sqrt{n} = \beta, \quad \mu_S^n = 1 \quad \textit{and} \quad \theta_S^n = \theta_S \quad \textit{for all } n \geq 1.$$

For strictly positive real numbers λ_F, r_F and θ_F, the following holds for the fast class

$$\lim_{n\to\infty} \mu_F^n = \infty, \quad \lim_{n\to\infty} r_F^n = r_F, \quad \textit{and} \quad \theta_F^n = \theta_F \quad \textit{for all } n \geq 1.$$

Notice that $\lambda_F^n/n \to \lambda_F$ in Assumption 1 is replaced by the generalized condition $\mu_F^n \to \infty$ in Assumption 6. In particular, Assumption 6 holds if $\mu_F^n = O(n^\alpha)$ and $\lambda_F^n = O(n^{\alpha+1/2})$ for any $\alpha > 0$. Now we assume Assumption 6 instead of Assumption 1. Again scale the process (X^n, Q^n) by (2). Recall the limit process (X, Q) and the stationary distribution from Section 4, we have

Theorem 5. *As $n \to \infty$, (i) if $\widetilde{X}^n(0) \Rightarrow X(0)$, then $(\widetilde{X}^n, \widetilde{Q}^n) \Rightarrow (X, Q)$ in D^4; (ii) $(\widetilde{X}^n(\infty), \widetilde{Q}^n(\infty)) \Rightarrow (X(\infty), Q(\infty))$ in $\mathbb{R}^4$; and (iii) $E[\widetilde{Q}_i^n(\infty)] \to E[\widetilde{Q}_i^n(\infty)]$ for $i = S$, F.*

In other words, all results in Section 4 remain unchanged. We omit the detailed proof of Theorem 5, as it follows from the same arguments as of Theorem 1 and Theorem 2.

A.2. Time-Varying Arrivals

Service systems in practice are typically non-stationary because the arrival processes are time-varying (and, as a response, so are the staffing levels). Nevertheless, stationary analysis for such systems is still useful, because it provides guidance regarding how to staff (or control) so as to stabilize desirable performance measures. In this section, we consider two different prevalent methods—the Infinite-Server (IS) approximation and the Pointwise Stationary Approximation (PSA)—to determine the appropriate time-dependent staffing levels. We note that PSA works well when the arrival rates change slowly relative to the mean service times (as in typical contact centers), whereas the IS approximation is appropriate when this is not the case (as in healthcare systems which tend to have large service times). We refer to Whitt (2018) for a comprehensive review of time-varying queues, including the IS and PSA methods. Note that, in our setting, the relative changes in the arrival rates should naturally be compared to the service times of the slow class.

A.2.1. Piecewise Stationary Approximation As reviewed in Gans et al. (2003), a prevalent approach in practice to deal with time-varying systems is to divide the day into time segments, and treat the system as being stationary over each of those segments. Specifically, the arrival rates are approximated by piecewise-constant (step) functions, and stationary analysis is carried out over each interval over which the arrival rate is fixed. This approach is a special case of PSA, which prescribes treating the system as being stationary at each time t with a stationary distribution corresponding to the arrival rates at that time.

In our setting, we assume that, for $m \geq 2$, there are time points $0 = t_0 < t_1 < t_2 < \cdots < t_m = T$, such that values of $\lambda_S(t)$ and $\lambda_F(t)$ are fixed over each of the time intervals $[t_j, t_{j+1})$, $0 \leq j \leq m-1$, where T is the period of the arrival rate. We then treat the system as stationary in $[t_j, t_{j+1})$, and choose a staffing level so as to keep $E[Q_i(\infty)]$ fixed for both $i = S$ and $i = F$.

To demonstrate the effectiveness of PSA with piecewise constant rates, we consider a simulation example with arrival rate

$$\lambda_S(t) = 30(1 + 1_{\{t\in[0,50)\}}) \quad \text{and} \qquad \lambda_F(t) = \lambda_S(t)/3, \quad t \in [0, 150),$$

where $\lambda_i(150 + t) = \lambda_i(t)$, for each $t \in \mathbb{R}$. We take the other system parameters to be $\mu_S = 1$, $\mu_F = 5$, $\theta_S = 0.1$, and $\theta_F = 0.3$, and the number of servers $s(t)$ at time t to be

$$s(t) = 34 + 33 \times 1_{\{t\in[0,50)\}}, \quad t \in [0, 150], \quad \text{with } s(t + 150) = s(t) \quad \text{for} \quad t \geq 0.$$

With these parameters, the FDH approximation is computed separately for each of the two time intervals $[0, 50)$ and $[50, 150)$. Figure 5 compares the simulated average queue processes to their PSA approximation. As expected, the system is temporarily congested (overloaded) when the arrival rate and staffing levels drop at time 50, but nevertheless stabilizes quickly. We note that such “predictable spikes” due to changes in the arrival rate and staffing levels can be smoothed out by changing the staffing levels gradually, as is done in practice.

We next consider an optimal system design problem in the PSA setting with the same parameters as in the example above, and with cost parameters $c_S = 1$, $c_F = 1$, and $d = -0.8$. To this end, we solve for the optimal value z_j^* in (15), $j = 1, \ldots, m$, corresponding to the assumed stationary distribution associated with the time interval $[t_j, t_{j+1})$, where, in our numerical example, $t_1 = 50$ and $t_2 = T = 150$.

Solving the two stationary distributions corresponding to the two values of the arrival rates, we find that the optimal fast-track sizes are $z_1^{n*} = 3$, and $z_2^{n*} = 1$. In particular, the fast-track staffing function satisfies

$$z(t) = 1 + 2 \times 1_{\{t\in[0,50)\}}, \quad t \in [0, 150),$$

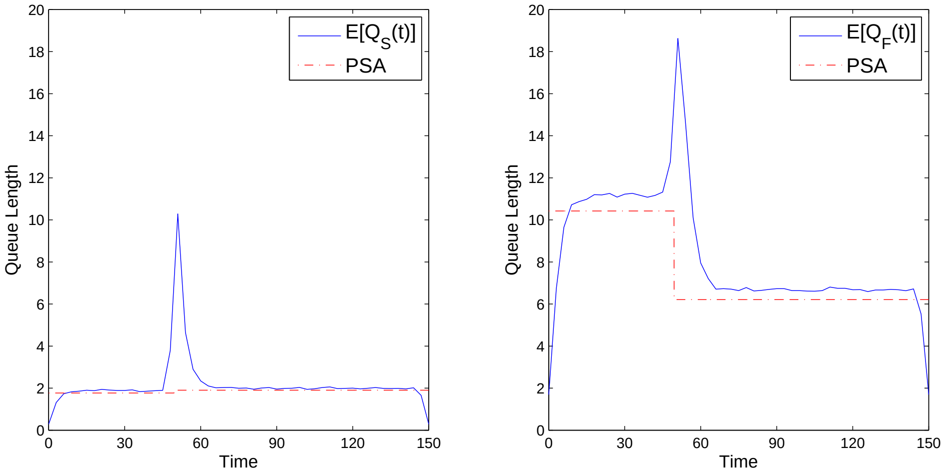

Figure 5 The simulated queue lengths and its piecewise stationary approximation.

and $z(t+150)=z(t)$ for $t\geq 150$.

To check the effectiveness of the system design, we conducted 100 independent simulation, each having length $300,000$, and averaged the results for both system designs, namely, single pool, and with a fast track. The resulting simulated total cost was $C=3.35$ for the system with the fast track (with standard error 0.003), and $C=3.87$ (standard error 0.004) for the single-pool design; in particular, operating a fast track provides a 13.4% reduction in cost relative to the single-pool design.

A.2.2. Infinite-Server Staffing Rule The QED limiting approximation is effective when the proportion of delayed customers (in the customer class operating under QED) is non-negligible, but is sufficiently smaller than 1. To achieve QED-type behavior in time-varying setting, one can take the number of agents $s(t)$ at time t be such that

$$s(t)=m(t)+\gamma\sqrt{m(t)}, \quad t\geq 0, \tag{17}$$

where $m(t)$ is the fluid limit of a corresponding IS queue having the same arrival rate function and service-time distribution as the system under consideration, and γ is a *quality-of-service* (QoS) constant chosen appropriately so as to maintain desirable performance measures. In our two-class setting, m takes the form of

$$m(t)=\sum_{i=S,F}\left(\int_0^t e^{\mu_i(t_1-t)}\lambda_i(t_1)dt_1+Q_i(0)e^{-\mu_i t}\right).$$

To demonstrate that the insights from our stationary analysis remain to hold in time-varying setting, we consider a numerical example. We take the arrival rates to be of the form

$$\lambda_i(t)=\lambda_i(1+\alpha\sin(At)), \quad i=S,F,$$

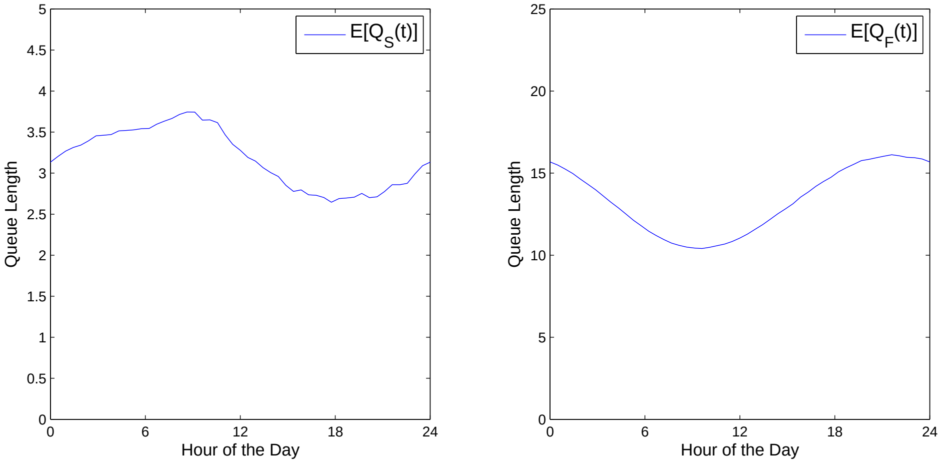

Figure 6 Time-varying queue lengths under the IS staffing rule.

where $\lambda_S = 9.2$, $\lambda_F = 3$, $A = 2\pi/24$ and $\alpha = 0.7$. The other parameters are $\mu_S = 0.2$, $\mu_F = 1$, $\theta_S = 0.02$ and $\theta_F = 0.06$. We further take $\gamma = 0.15$ and the staffing function $s(t)$ in (17).

Figure 6 plots the time-dependent queue processes, averaged over 100 independent simulation runs of periods 500 to 5000. Note that the time-dependent average queue of the slow class varies in the interval $(2.5, 4)$ and that of the fast class has values in $(10, 16)$, despite the fact that the maximum arrival rate is 5 times greater than the minimum arrival rate within a period. We further note that the probability that a slow-class customer is delayed is 0.31 (with standard deviation 0.001), indicating that its queue operates in a time-varying QED regime.

The example above demonstrates that the IS staffing rule can stabilize a system qualitatively in a time-varying FDH regime. In order to perform quantitative analysis, we propose a heuristic FDH approximation that builds on the limit for the stationary system. Specifically, we consider the stationary FDH limit with arrival rates $\bar{\lambda}_i := T^{-1}\int_0^T \lambda_i(s)ds$, $i = S, F$ (where T is the period) and staffing level (17) (after replacing $\lambda_i(t)$ by its average $\bar{\lambda}_i$).

Table 5 compares the heuristic FDH approximation to the simulation results for the time-varying setting just described. In the table, $E[Q_S]$ and $E[Q_F]$ are the simulated long-run average queue lengths, computed by averaging the queue-length processes from time 500 to 5000. To check how the accuracy of the heuristic dependence on the *predictable variability*, we consider four different values of α: $\alpha \in \{0, 0.3, 0.5, 0.7\}$. We observe that the heuristic averaging approach to compute the FDH approximation is effective.

A.3. Non-Exponential Service Time

We now provide a simulation experiment to demonstrate the robustness of the FDH limiting approximation to the assumption that service times are exponentially distributed. The parameters

	FDH	$\alpha = 0.0$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$
$E[Q_S]$	3.31 (0.04)	3.35 (0.02)	3.27 (0.02)	3.26 (0.02)	3.19 (0.02)
$E[Q_F]$	13.70 (0.1)	13.96 (0.07)	13.69 (0.06)	13.64 (0.06)	13.39 (0.06)

Table 5 Comparisons of the heuristic time-varying FDH to simulations for different values of α

we consider are the same as in Section 4.4, namely, $n = 50$, $\lambda_S^n = 46$, $\lambda_F^n = 15$, $\theta_S = 0.1$ and $\theta_F = 0.3$. We compare two service-time distributions to the exponential distribution—one that is more variable, and the other that is less variable, than the exponential distribution, whose coefficient of variations (CoV) equals 1. In particular, we fix the values of the service rates to be $\mu_S^n = 1$ and $\mu_F^n = 5$, and consider all the nine different combinations of the service time distributions, where each distribution is gamma with the specified mean and with CoV in $\{0.5,1,2\}$. Since the exponential distribution is a special case of the gamma, in which the CoV is equal to 1, the other gamma distributions we consider represent distributions with variability that is half or twice that of the exponential distribution. (Note that the mean and CoV completely determine the gamma distribution.)

Table 8 compares the FDH approximation to simulations results. Each simulation experiment consists of 100 independent simulation runs for 1000 time units with a warmup period of 100 time units. Since we consider extreme differences in the variability of the service times, and since the system is relatively small, the difference between the sizes of the slow-class queue in the different cases can be relatively large. However, the fast-class queue is much less sensitive to the service-time distribution, both due to its relative size (that is larger than that of the slow queue), and the fact that it behaves like a fluid model.

In ending we remark that one can replace the limiting approximation X_S, which is the Garnett diffusion limit for the $M/M/n+M$ queue, by the limiting approximation for the $G/G/n+G$ queue. This can be done in two steps using existing results: First, the service-time distribution is approximated with a phase-type distribution (this can always be done, because the class of phase-type distributions is dense in the family of probability distributions on $\mathbb{R}$). Second, for that specific phase-type distribution, X_S in the FDH limit is replaced with the diffusion limit for the $G/Ph/n+G$ queue in Dai et al. (2010).

A.4. A Numerical Comparison Between V-system and N-system.

We provide a numerical example comparing the performance metrics of systems with different fast-track sizes. We consider the same example as the one in Section 4.4 and take $n = 50$, $\lambda_S^n = 46$, $\lambda_F^n = 15$, $\mu_S^n = 1$, $\mu_F^n = 5$, $\theta_S = 0.1$ and $\theta_F = 0.3$. Table A.4 presents various performance metrics for

	$z^n = 0$ (V-system)		$z^n = 1$		$z^n = 2$	
	simulation	FDH	simulation	FDH	simulation	FDH
$E[Q_S^{z;n}(\infty)]$	3.40 (0.03)	3.25 (0.02)	4.77 (0.04)	4.67 (0.03)	6.65 (0.04)	6.59 (0.04)
$E[W_S^n \mid W_S^n > 0]$	0.18 (8e-4)	0.18 (8e-4)	0.21 (10e-4)	0.21 (10e-4)	0.25 (0.001)	0.25 (0.001)
$P(W_S^n > 0)$	0.41 (0.001)	0.39 (0.001)	0.50 (0.002)	0.48 (0.002)	0.58 (0.001)	0.57 (0.002)
$P(Ab_S^n)$	0.01 (5e-5)	0.01 (5e-5)	0.01 (7e-5)	0.01 (7e-5)	0.01 (9e-5)	0.01 (10e-5)
$E[Q_F^{z;n}(\infty)]$	13.96 (0.07)	13.51 (0.06)	10.73 (0.06)	10.36 (0.05)	6.48 (0.03)	5.81 (0.03)
$E[W_F^n \mid W_F^n > 0]$	1.28 (0.005)	1.28 (0.004)	1.00 (0.004)	0.98 (0.004)	0.62 (0.002)	0.55 (0.002)
$P(W_F^n > 0)$	0.73 (0.001)	0.70 (0.001)	0.72 (0.001)	0.71 (0.001)	0.70 (0.001)	0.71 (0.001)
$P(Ab_F^n)$	0.28 (0.001)	0.27 (0.001)	0.22 (0.001)	0.21 (0.001)	0.13 (6e-4)	0.12 (6e-4)

Table 6 Accuracy of approximation for different fast-track size. Numbers in parentheses are the standard error in the simulation.

systems with different fast-track sizes. In particular, the case $z^n = 0$ is the V-system considered in Section 4.4.

In this example, as the size of the fast-track grows, we see an increase in the queue of the slow class and a decrease in the queue of the fast class. This observation is expected since the fast track is essentially taking capacity away from the slow class. We also see that the approximation errors for $E[Q^{z,n}(\infty)]$ and $E[W_F^n|W_F^n > 0]$ are increasing in z^n. This too is to be expected, because the FDH approximation for the fast class is a (random) fluid limit that does not capture stochastic fluctuations of the fast-class queue, and its accuracy is thus decreasing as the queue decreases.

For the optimal system-design problem, we consider the parameters $c_S^n = 10$, $c_F^n = 1$, and $d^n = -15$. We have $c_S = 10$ and $c_F = 5$, so that Assumption 4 holds, implying that $C(z)$ is convex. For these parameters, we computed the value $z^* = 0.5$ so that $z^{n*} = 1$. From Table A.4, we see that $C^n(z^n)$ takes the values 47.0, 43.4, and 43.0 for $z^n = 0, 1, 2$, respectively. In particular, the cost for the N-system with one or two fast-track servers is approximately 8% lower than the cost for the V system.

A.5. The Non-preemptive Optimal System Design Problem

In this section, we provide additional numerical experiments to illustrate how we can use the FDH approximation to solve the optimal system design problem, when the scheduling policy is non-preemptive. We consider the same system as in Section 4.4 with $n = 50$, $\lambda_S^n = 46$, $\lambda_F^n = 15$, $\mu_S^n = 1$, $\mu_F^n = 5$, $\theta_S = 0.1$ and $\theta_F = 0.3$.

Let z_P^{n*} and z_N^{n*} be the optimal solutions corresponding to the preemptive and the non-preemptive policies, respectively. Notice that $R_F^n = \lambda_F^n/\mu_F^n = 3$, so that the possible value of z_i^{n*}, $i = N,\ P$, is in $\{0, 1, 2, 3\}$. Also let z^* be the optimal fast-track value for the FDH limit, and $z_{\mathrm{FDH}}^{n*} := R_F^n z^*$ be the (unscaled) value. We normalize the cost parameters relative to c_S by fixing $c_S = 1$ and vary the values of c_F and d.

The results for z_i^{n*}, $i = \mathrm{FDH},\ N,\ P$, are summarized in Table 7. Observe that the optimal size of the fast track under the preemptive and the non-preemptive policies are the same in all cases, except for the two cases $c_F = 0.8$ when $d = 0.8$ and $d = 2$. Even though this is a relatively small system the differences between the two systems are not large, as can also be seen by the numerical experiment Section 7.3, which compares the stationary queues under the two policies.

	$c_F = 0.4$			$c_F = 0.8$			$c_F = 1.2$			$c_F = 1.6$		
d	z_{FDH}^{n*}	z_P^{n*}	z_N^{n*}	z_{FDH}^{n*}	z_P^{n*}	z_N^{n*}	z_{FDH}^{n*}	z_P^{n*}	z_N^{n*}	z_{FDH}^{n*}	z_P^{n*}	z_N^{n*}
0.4	0	0	0	0	0	0	0	0	0	1.1	1	1
0.6	0	0	0	0	0	0	0.6	1	1	2.0	2	2
0.8	0	0	0	0.3	0	1	1.4	1	1	2.6	2	2
1.0	0.2	0	0	1.1	1	1	2.0	2	2	2.9	2	2
2.0	2.25	2	2	2.85	2	3	3	3	3	3	3	3

Table 7 Numerical Computations of z_i^{n*}, $i = N,\ P$, and its FDH approximation z_{FDH}^{n*}.

We also note that the FDH approximation tends to overestimate z^{n*}, because the fast class is approximated via a fluid limit, which is oblivious to the stochasticity of this class' queue. To see why, note that if the fast-track has sufficient service capacity to handle all of the fast class, then the respective queue will be null in the limit, unlike in the actual system, which will experience stochastic fluctuations.

Appendix B: Preliminaries for the Proofs

The remaining appendices are dedicated to the proofs of the results in the paper. In this section, we establish supporting results that are employed in the proofs of the main results in Section C. Additional technical results are further deferred to Section D and Section E.

We use the following notation, in addition to the notation introduced in the main paper. We write $x \vee y$ and $x \wedge y$ to denote $\max\{x, y\}$ and $\min\{x, y\}$, respectively. We denote the state space of the FDH limit processes X and X^z by $S := \mathbb{R} \times [0, \infty)$. We let $D_+ \subset D$ denote the space

		$\text{CoV}_F = 0.5$	$\text{CoV}_F = 1$	$\text{CoV}_F = 2$
$\text{CoV}_S = 0.5$	Slow	2.52 (0.02)	2.51 (0.02)	2.52 (0.02)
	Fast	13.08 (0.06)	12.79 (0.06)	12.02 (0.06)
$\text{CoV}_S = 1$	Slow	3.43 (0.03)	3.42 (0.03)	3.34 (0.02)
	Fast	14.27 (0.07)	13.97 (0.07)	13.01 (0.06)
$\text{CoV}_S = 2$	Slow	5.84 (0.06)	5.71 (0.05)	5.78 (0.06)
	Fast	15.72 (0.1)	15.42 (0.09)	14.72 (0.09)

Table 8 The Expected Queue Lengths $E[Q_i^n(\infty)]$, $i = S,\, F$.

of right-continuous functions with limits everywhere, having a nonnegative initial value, namely, $D_+ := \{w \in D : w(0) \geq 0\}$. Similarly, $C_+ := \{w \in C : w(0) \geq 0\}$. We use $\overset{\mathrm{d}}{=}$ to denote equality in distribution, and use $\leq_{st}$ to denote stochastic dominance under the usual stochastic order. In particular, for two random variables X_1 and X_2 we write $X_1 \leq_{st} X_2$ if $P(X_1 > x) \leq P(X_2 > x)$, for all $x \in \mathbb{R}$. Finally, $\|\cdot\|_{tv}$ denotes the total-variation norm; e.g., see (Asmussen 2008, A8).

We prove the convergence of the stochastic queueing processes to the FDH limit by representing these processes as a continuous mapping of their primitives. We therefore begin by introducing two continuous mappings in Section B.1, which we use in Section B.2 to characterize the stochastic systems.

B.1. Continuous Mappings

Let $\phi : D \to D$ be the map defined via $\phi(y) = x$, where $y, x \in D$ satisfy

$$x(t) = \int_0^t x(s)^- ds - \theta_S \int_0^t x(s)^+ ds + y(t). \tag{18}$$

Define the mapping $\eta : D_+ \to D^2$ via $\eta(y') = (q, \ell)$, where $y' \in D_+$ and $(q, \ell) \in D^2$ is the solution to a generalized Skorohod problem

$$\begin{gathered} q(t) = y'(t) - \theta_F \int_0^t q(s) ds + \ell(t); \\ \int_0^\infty 1_{\{q(t)>0\}} d\ell(t) = 0; \\ q(t) \geq 0,\ \ \ell(0) = 0,\ \text{ and } \ell(t) \text{ is non-decreasing.} \end{gathered} \tag{19}$$

The following lemma guarantees that the mappings above are well defined and continuous.

Lemma 2 **(continuity)**. *There exists a unique solution x in D to* (18)*. Further, ϕ is continuous in D, and is Lipschitz continuous as a mapping from $C[0,T]$ into itself, for any $T>0$. Similarly, there exists a unique solution (q,ℓ) in D^2 to* (19)*; the map η is continuous in D_+, and is Lipschitz continuous as a mapping from $C_+[0,T]$ into $C^2[0,T]$, for any $T>0$.*

Proof. The statements regarding the solution x to (18), and the map ϕ follow from Theorem 4.1 Pang et al. (2007), by taking $b=0$ and $h(a)=-\theta_S a^+ - a^-$ in this theorem. The Lipschitz continuity of ϕ on $C[0,T]$ (with e^{cT} as a Lipschitz constant) follows from the proof of the cited theorem. The statements regarding (q,ℓ) and η follow from Theorem 7.3 in the same reference, by taking $b=0$, $y=-y'$, $\kappa=0$, $h(q)=-\theta_F q$ and $u(t)=-\ell(t)$. Once again, the Lipschitz continuity of η as a mapping from $C_+[0,T]$ follows from the proof of the cited theorem. □

We remark that the Lipchitz continuity of ϕ and η in the proofs of Theorems 4.1 and 7.3 in Pang et al. (2007) in shown to hold in the space D endowed with the uniform topology, and it therefore holds for the space C with the same topology. However, we endow D with the J_1 topology, which is why we only state the Lipschitz continuity when the domain (and the codomain) of ϕ and η is C.

B.2. Process Characterization

Consider the setting in Section 5, where system n operates with z^n fast-track servers. Notice that this setting reduces to the one in Section 4.2, by taking $z^n=0$ for all $n\geq 1$. To simplify the notation, we drop the superscript z in this section, and let (X^n,Q^n) denote the number-in-system and queue-length processes (as opposed to $(X^{z,n},Q^{z,n})$). Further, we assume that $\widetilde{X}^n(0)\Rightarrow X(0)$ in $\mathbb{R}^2$, as $n\to\infty$.

Recall that $Z^n:=(Z^n_S,Z^n_F)$ is the number-in-service process, i.e., $Z^n_i(t)$ is the number of class-i customers that are in service at time t. Clearly,

$$Q^n_S = X^n_S - Z^n_S \quad \text{and} \quad Q^n_F = X^n_F - Z^n_F. \tag{20}$$

Further, the (preemptive) priority policy to the slow class indicates that

$$Z^n_S = X^n_S \wedge (n-z^n) \quad \text{and} \quad Z^n_F = (n-Z^n_S)\wedge X^n_F. \tag{21}$$

Employing standard arguments, as reviewed Pang et al. (2007), we can represent X^n_i, $i=S,F$, as follows

$$X^n_i(t) = X^n_i(0) + A_i(\lambda^n_i t) - S_i\left(\mu^n_i\int_0^t Z^n_i(s)ds\right) - R_i\left(\theta^n_i\int_0^t Q^n_i(s)ds\right), \quad t\geq 0, \tag{22}$$

where $A_i(t)$, $S_i(t)$, and $R_i(t)$ are 6 independent unit-rate Poisson processes, which are also independent of $X^n(0)$. It follows immediately from (22) and (21) that X^n_S is the number-in-system process of an $M/M/(n-z^n)+M$ queue. (Compare, e.g., to Equation (97) in Pang et al. (2007).)

For $i = S, F$ and $t \geq 0$, we define the "martingale terms"

$$M_{iA}^n(t) := A(\lambda_i^n t) - \lambda_i^n t, \quad M_{iS}^n(t) := S\left(\mu_i^n \int_0^t Z_i^n(s)ds\right) - \mu_i^n \int_0^t Z_i^n(s)ds,$$
$$M_{iR}^n(t) := R_i\left(\theta_i \int_0^t Q_i^n(s)ds\right) - \theta_i \int_0^t Q_i^n(s)ds, \tag{23}$$

and

$$W_i^n(t) := \sum_{k=A,S,R} M_{ik}^n(t).$$

(For the fact that the above terms are indeed martingales with respect to an appropriate filtration see Pang et al. (2007); since this fact is irrelevant for our analysis, we do not dwell on this issue.) Subtracting the intensity from each of the Poisson processes in (22) and then adding those intensities back, and employing the martingale terms in (23) gives the "martingale representation"

$$X_i^n(t) = X_i^n(0) + \lambda_i^n t - \mu_i^n \int_0^t Z_i^n(s)ds - \theta_i \int_0^t Q_i^n(s)ds + W_i^n(t). \tag{24}$$

To represent $\widetilde{X}_S^n$ via the map ϕ, let

$$\tilde{z}^n := z^n/\sqrt{\lambda_S^n}, \quad \beta^n := (n - \lambda_S^n)/\sqrt{\lambda_S^n} \quad \text{and} \quad \widetilde{W}_S^n := (\lambda_S^n)^{-1/2} W_S^n,$$

so that $\tilde{z}^n \to r_F z$ and $\beta^n \to \beta$ as $n \to \infty$. It follows from (24) and Theorems 7.2 in Pang et al. (2007), that

$$\widetilde{X}_S^n = \phi\left(\widetilde{X}^n(0) - (\beta^n - \tilde{z}^n)e + \widetilde{W}_S^n\right). \tag{25}$$

Further, from the proof of Theorem 7.1 in Pang et al. (2007), we have

$$(\widetilde{W}_S^n, \widetilde{X}^n(0)) \Rightarrow (\sqrt{2}B, X(0)) \text{ in } D \times \mathbb{R}^2 \text{ as } n \to \infty, \tag{26}$$

where B is a standard Brownian motions that is independent to $X(0)$.

To represent $\widetilde{Q}_F^n$ via the map η, let

$$\widetilde{Z}_S^n := (Z_S^n - n)/\sqrt{\lambda_S^n} \quad \text{and} \quad \tilde{r}_F^n := R_F^n/\sqrt{\lambda_S^n},$$

so that $\tilde{r}_F^n \to r_F$ as $n \to \infty$. Scaling X_i^n in (24) when $i = F$ gives

$$\widetilde{X}_F^n(t) = \widetilde{X}_F^n(0) + \widetilde{W}_F^n(t) + (1 - \tilde{z}^n/\tilde{r}_F^n)t - (\tilde{r}_F^n)^{-1} \int_0^t \widetilde{X}_S^n(s)^- ds$$
$$+ \quad \mu_F^n \int_0^t \left(\widetilde{X}_F^n(s) + (\tilde{r}_F^n \mu_F^n)^{-1} \widetilde{Z}_S^n(s)\right)^- ds - \theta_F \int_0^t \widetilde{Q}_F^n(s)ds. \tag{27}$$

Since $I^n(t) := \int_0^t (n - Z_S^n(s) - Z_F^n(s))ds$ and $\widetilde{I}^n := I^n/R_F^n$, (21) gives

$$\widetilde{I}^n(t) = \mu_F^n \int_0^t \left(\widetilde{X}_F^n(s) + (\tilde{r}_F^n \mu_F^n)^{-1} \widetilde{Z}_S^n(s)\right)^- ds. \tag{28}$$

Let $\widetilde{W}^n_F := W^n_F/\lambda^n_F$. Using (20), (27), and (28), we have

$$\widetilde{Q}^n_F(t) = \left(\widetilde{W}^n_F(t) + \widetilde{X}^n_F(0) - Z^n_F(t)/\lambda^n_F + (1-\tilde{z}^n/\tilde{r}^n_F)t - (\tilde{r}^n_F)^{-1}\int_0^t \widetilde{X}^n_S(s)^- ds\right) - \theta_F \int_0^t \widetilde{Q}^n_F(s)ds + \widetilde{I}^n(t).$$

On the other hand, plugging (21) in (20) gives

$$\widetilde{Q}^n_F = \left(\widetilde{X}^n_F + (\tilde{r}^n_F \mu^n_F)^{-1}\widetilde{Z}^n_S\right)^+. \tag{29}$$

It follows from (28) and (29) that

$$\int_0^t \widetilde{Q}^n_F(s)d\widetilde{I}^n(s) = 0, \quad \text{for all } t \geq 0.$$

Since $\widetilde{I}^n(0) = 0$, $\widetilde{I}^n$ is non-decreasing, and $\widetilde{Q}^n_F \geq 0$, we conclude that

$$(\widetilde{Q}^n_F, \widetilde{I}^n) = \eta\left(\widetilde{W}^n_F + \widetilde{X}^n_F(0) - Z^n_F/\lambda^n_F + (1-\tilde{z}^n/\tilde{r}^n_F)e - (\tilde{r}^n_F)^{-1}\int_0^\cdot \widetilde{X}^n_S(s)ds\right). \tag{30}$$

Proposition 4. *$\widetilde{W}^n_F \Rightarrow 0e$ in D as $n \to \infty$.*

Proof. Using the functional central limit theorem for the Poisson process, e.g., (Pang et al. 2007, Theorem 8.1), we have

$$\left(\frac{A_F(n^2t) - n^2t}{n}, \frac{S_F(n^2t) - n^2t}{n}, \frac{R_F(n^2t) - n^2t}{n}\right) \Rightarrow (B_1, B_2, B_3) \text{ in } D^3, \text{ as } n \to \infty, \tag{31}$$

where (B_1, B_2, B_3) is a three-dimensional Brownian motion.

Define the functions $T^n_j \in D$, $j = A, S, R$, as follows

$$T^n_A(t) := \frac{\lambda^n_F}{n^2}t, \quad T^n_S(t) := \frac{\mu^n_F}{n^2}\int_0^t Z^n_F(s)ds \quad \text{and} \quad T^n_R(t) := \frac{\theta_F}{n^2}\int_0^t Q^n_F(s)ds, \quad t \geq 0.$$

By Assumption 1, $T^n_A \Rightarrow 0e$ in D as $n \to \infty$. To see that T^n_S and T^n_R also converge weakly to the zeroth function, note that

$$0 \leq \frac{\mu^n_F}{n^2}\int_0^t Z^n_F(s)ds \leq \frac{\mu^n_F}{n}t, \quad \text{and} \tag{32}$$

$$0 \leq \frac{\theta_F}{n^2}\int_0^t Q^n_F(s)ds \leq \frac{\theta_F}{n^2}\left(X^n_F(0)t + \int_0^t A^n_F(s)ds\right). \tag{33}$$

It follows from (32) and the scaling of μ^n_F in Assumption 1 that $T^n_S \Rightarrow 0e$ in D as $n \to \infty$. For T^n_R, we first employ the functional strong law of large number for the Poisson process to obtain that $A^n_F/n^2 \Rightarrow 0e$ in D. Therefore, the continuity of the integral (e.g., Theorem 11.5.1 in Whitt (2002)) implies that, as $n \to \infty$,

$$\frac{1}{n^2}\int_0^\cdot A^n_F(s)ds \Rightarrow 0e \text{ in } D.$$

Now, $\widetilde{X}_F^n(0) \Rightarrow X_F(0)$ implies that $X_F^n(0)/n^2 \Rightarrow 0e$ in $\mathbb{R}$, and by the continuity of addition at continuous limits, e.g., Theorem 4.1 in Whitt (1980),

$$\frac{1}{n^2}\left(\int_0^{\cdot} A_F^n(s)ds + X_F^n(0)\right) \Rightarrow 0e \text{ in } D.$$

Together with (33), this imply that $T_R^n \Rightarrow 0e$ in D.

For $j = A, S, R$ and M_{Fj}^n in (23), let $\widetilde{M}_{Fj}^n := M_{Fj}^n/\lambda_F^n$. Observe that $\widetilde{M}_{FA}^n$ is the composition of the process $(\lambda_F^n)^{-1}(A_F(n^2t) - n^2t)$ with the function T_A^n, which is nondecreasing. By the continuity of the composition map (Theorem 13.2.1 in Whitt (2002)) $\widetilde{M}_{FA}^n \Rightarrow 0e$, and similar arguments give that $\widetilde{M}_{FS}^n \Rightarrow 0e$ and $\widetilde{M}_{FR}^n \Rightarrow 0e$ in D as $n \to \infty$. Therefore,

$$(\widetilde{M}_{FA}^n, \widetilde{M}_{FS}^n, \widetilde{M}_{FR}^n) \Rightarrow (0e, 0e, 0e) \text{ in } D^3,$$

implying that $\widetilde{W}_F^n \Rightarrow 0e$ in D as $n \to \infty$, due to the continuity of addition when the limits are continuous. □

Appendix C: Proofs of the Results in the Main Text

The main idea of the proof of Theorem 3 is to represent the process X^z by (25) and (30), and then applying the continuous mapping theorem. The proof of Theorem 4 relies on Proposition 5, the proof of which requires a result from Down et al. (1995) and is further postponed to Section D.2.

To prove Theorems 3 and 4 we need the following result.

Lemma 3. *The HSDE in* (9)–(11) *possesses a unique solution* $X^z \in C^2$.

Proof. By (18) and (19), the HSDE in (9)–(11) is equivalent to

$$X_S^z = \phi\left(X_S^z(0) - (\beta - r_F z)e + \sqrt{2}B\right), \tag{34}$$

$$(X_F^z, I^z) = \eta\left(X_F^z(0) + e + \int_0^{\cdot} (r_F^{-1} X_S^z(s)) \wedge (-z) ds\right). \tag{35}$$

Therefore, the existence of a unique solution to the HSDE, as well as the continuity of X^z, follow from Lemma 2, noting that the arguments of ϕ in (34) and of η in (35) are elements in C w.p.1. □

In the following proof of Theorem 3 all the arrows signifying limits (strong or weak) are taken as $n \to \infty$, and we therefore omit a mention of this fact to streamline the writing.

Proof of Theorem 3. The limits $\tilde{z}^n \to r_F z$ and $\beta^n \to \beta$ in $\mathbb{R}$, together with (26), give

$$\left(\widetilde{W}_S^n + (\tilde{z}^n - \beta^n)e + \widetilde{X}_S^n(0), \widetilde{X}_F^{z,n}(0), \tilde{z}^n\right) \Rightarrow \left(\sqrt{2}B + (r_F z - \beta)e + X_S^z(0), X_F^z(0), r_F z\right),$$

in $D \times \mathbb{R}^2$. By Lemma 2, ϕ is a continuous map from D to D. By (25) and (34) the continuous mapping theorem gives $\widetilde{X}_S^{z,n} \Rightarrow X_S^z$ in D, which further implies the joint convergence

$$(\widetilde{X}_S^{z,n}, \widetilde{X}_F^{z,n}(0), \tilde{z}^n) \Rightarrow (X_S^z, X_F^z(0), r_F z) \text{ in } D \times \mathbb{R}^2.$$

Next, $\widetilde{Q}_S^{z,n} = (\widetilde{X}_S^{z,n})^+$ and $Q_S^z = (X_S^z)^+$, so that $\widetilde{Q}_S^{z,n} \Rightarrow Q_S^z$ in D and therefore

$$(\widetilde{X}_S^{z,n}, \widetilde{Q}_S^{z,n}, \widetilde{X}_F^{z,n}(0), \tilde{z}^n) \Rightarrow (X_S^z, Q_S^z, X_F^z(0), r_F z) \text{ in } D^2 \times \mathbb{R}^2. \tag{36}$$

This, the limit $\tilde{r}_F^n \to r_F$ and Lemma 4, give

$$(\widetilde{X}_S^{z,n}, \widetilde{Q}_S^{z,n}, \widetilde{W}_F^n, \widetilde{X}_F^{z,n}(0), \tilde{z}^n, \tilde{r}_F^n) \Rightarrow (X_S^z, Q_S^z, 0e, X_F^z(0), r_F z, r_F) \text{ in } D^3 \times \mathbb{R}^3.$$

In particular, we have the limit in D

$$\widetilde{X}_F^{z,n}(0) + (1 - \tilde{z}^n/\tilde{r}_F^n)e + (\tilde{r}_F^n)^{-1} \int_0^{\cdot} \widetilde{X}_S^{z,n}(s)ds + \widetilde{W}_F^n \Rightarrow X_F^z(0) + (1-z)e + r_F^{-1} \int_0^{\cdot} X_S^z(s)ds.$$

By Lemma 2, η is a continuous map from D_+ to D^2. By (30) and (35), the continuous mapping theorem gives $(\widetilde{X}_F^{z,n}, \widetilde{I}^{z,n}) \Rightarrow (X_F^z, I^z)$ in D^2, which further implies the joint convergence

$$(\widetilde{X}_S^{z,n}, \widetilde{Q}_S^{z,n}, \widetilde{Q}_F^{z,n}, \widetilde{I}^{z,n}) \Rightarrow (X_S^z, Q_S^z, Q_F^z, I^z) \text{ in } D^4$$

Finally, $\widetilde{Z}_S^n/\sqrt{n} \Rightarrow 0e$ in D and (21) imply that $Z_F^n/\lambda_F^n \Rightarrow 0e$ in D and the joint convergence

$$(\widetilde{X}_S^{z,n}, \widetilde{Q}_S^{z,n}, \widetilde{Q}_F^{z,n}, \widetilde{I}^{z,n}, Z_F^n/\lambda_F^n) \Rightarrow (X_S^z, Q_S^z, Q_F^z, I^z, 0e) \text{ in } D^5$$

Notice that $X_F^z = Q_F^z$ and $\widetilde{X}_F^n = \widetilde{Q}_F^n + Z_F^n/\lambda_F^n$, we have $\widetilde{X}_F^{z,n} \Rightarrow X_F^z$ in D and the joint convergence $(\widetilde{X}^{z,n}, \widetilde{Q}^{z,n}, \widetilde{I}^{z,n}) \Rightarrow (X^z, Q^z, I^z)$ in D^5. □

To prove Theorem 4 we express $\widetilde{Q}^{z,n}(\infty)$ in terms of $\widetilde{X}^{z,n}(\infty)$. Using (20) and (21), we have

$$\widetilde{Q}_S^{z,n} = (\widetilde{X}_S^{z,n})^+ \quad \text{and} \quad \widetilde{Q}_F^{z,n} = \left(\widetilde{X}_F^{z,n} - (\tilde{r}_F^n \mu_F^n)^{-1}(\widetilde{X}_S^{z,n})^- - (\tilde{r}_F^n \mu_F^n)^{-1} \tilde{z}^n\right)^+,$$

so that

$$\widetilde{Q}_S^{z,n}(\infty) = \widetilde{X}_S^{z,n}(\infty)^+ \quad \text{and} \quad \widetilde{Q}_F^{z,n}(\infty) = \left(\widetilde{X}_F^{z,n}(\infty) - (\tilde{r}_F^n \mu_F^n)^{-1} \widetilde{X}_S^{z,n}(\infty)^- - (\tilde{r}_F^n \mu_F^n)^{-1} \tilde{z}^n\right)^+. \tag{37}$$

We will also need the following proposition, whose proof appears in Appendix D. Let $\{P_t : t \geq 0\}$ denote the transition semigroup of X^z, namely, $P_t(x, A) = P(X^z(t) \in A | X^z(0) = x)$, for any $x \in S$ and Borel-measurable set $A \subset S$. Following standard terminology, we say that X^z is *exponentially ergodic* if there exists an invariant measure π, such that

$$\|P_t(x, \cdot) - \pi(\cdot)\|_{tv} \leq M(x)\gamma^t \tag{38}$$

for some $\gamma < 1$ and some finite $M(x)$, that depends only on the initial condition x.

Proposition 5 **(exponential ergodicity)**. *Fix $z \in [0,1]$, and let X^z be the solution to the HSDE* (9)–(11)*. Then there exist positive constants K_1, K_2, and $\gamma < 1$, and a random variable $X^z(\infty)$ in $\mathbb{R}^2$, such that the following holds for any $x = (x_S, x_F) \in S$ and any measurable function $f : \mathbb{R}^2 \to [1, \infty)$.*

$$\left|E[f(X^z(t)|X^z(0) = x)] - E[f(X^z(\infty))]\right| \le \sup_{y \in \mathbb{R}^2} \frac{|f(y)|}{\|y\| + 1}(K_1\|x\| + K_2)\gamma^t. \tag{39}$$

In particular, (38) *holds, so that X^z is exponentially ergodic.*

Proof of Theorem 4. The first statement of Theorem 4 follows immediately from Proposition 5. To prove the second statement, we first note that the sequence $\{(\widetilde{X}^{z,n}(\infty), \widetilde{Q}^{z,n}(\infty)) : n \ge 1\}$ is tight in $\mathbb{R}^4$, as will be proved below (note that this claimed tightness follows from the third statement of the theorem), so that any of its subsequences has a weakly converging sub-subsequence.

Consider $\widetilde{X}^{z,n}$ with an initial condition $\widetilde{X}^{z,n}(0) \stackrel{\mathrm{d}}{=} \widetilde{X}^{z,n}(\infty)$, and consider a subsequence of this sequence of initial conditions that converges to a random variable X_0^z in $\mathbb{R}^2$ as $n \to \infty$. By Theorem 3, $\{\widetilde{X}^{z,n} : n \ge 1\}$ converges weakly in D^2 to the solution X^z of the HSDE (9)–(11) with the initial condition $X^z(0) = X_0^z$. By Proposition 5, X^z has a unique stationary distribution $X^z(\infty)$. On the other hand, $\widetilde{X}^{z,n}(t)$ has the same law as $\widetilde{X}^{z,n}(0)$ for any $t \ge 0$, so that $X^z(t)$ also has the same law as X_0^z, implying that X_0^z must have the stationary distribution of the limit process X^z. The uniqueness of $X^z(\infty)$ implies that all converging subsequences have the same limit, and in turn, that the whole sequence $\widetilde{X}^{z,n}$ converges to this limit.

To prove the third statement of the theorem, recall that $\widetilde{X}_S^{z,n}$ is distributed as the number-in-system process of an $M/M/(n - z^n) + M$ queue with arrival rate λ_S^n, service rate 1, and abandonment rate θ_S (see (22) and the discussion following it), so that $\widetilde{X}_S^{z,n}(\infty)$ has the stationary distribution of this process. By Theorem 1 in Braverman and Dai (2017), there exists a real-valued random variable $X_S^z(\infty)$, such that

$$|\widetilde{X}_S^{z,n}(\infty)| \Rightarrow |X_S^z(\infty)| \quad \text{and} \quad E\left[|\widetilde{X}_S^{z,n}(\infty)|\right] \to E\left[|X_S^z(\infty)|\right] \quad \text{in } \mathbb{R} \text{ as } n \to \infty.$$

By Theorem 3.6 Billingsley (2009), $\{|\widetilde{X}_S^{z,n}(\infty)| : n \ge 1\}$ is UI. It follows from (37) that $E\left[|\widetilde{Q}_S^{z,n}(\infty)|\right] \le E\left[|\widetilde{X}_S^{z,n}(\infty)|\right] + \tilde{z}^n$, implying that both $\{\widetilde{Q}_S^{z,n}(\infty) : n \ge 1\}$ and $\{\widetilde{X}_S^{z,n}(\infty) : n \ge 1\}$ are UI.

Now, a simple coupling argument shows that the process $X_F^{z,n}$ is bounded from above, in the sample-path stochastic order sense (namely, sample-path wise and w.p.1), by the number-in-system process in an $M/M/\infty$ queue having arrival rate λ_F^n and service rate θ_F; see Lemma A.5 in Perry and Whitt (2012). Since a sample-path stochastic order between two processes implies that the corresponding stationary distributions are ordered accordingly in the usual stochastic-order sense (in $\mathbb{R}$), we conclude that $X_F^{z,n}(\infty) \le_{st} Y^n$, where Y^n is a random variables having the stationary distribution of the bounding $M/M/\infty$ queue, and in particular, Y^n is Poisson distributed with mean λ_F^n/θ_F. Then

$$E[Y^n]/\lambda_F^n = \theta_F^{-1}, \quad \text{and} \quad \lim_{n\to\infty} Y^n/\lambda_F^n = \theta_F^{-1} \ \ w.p.1,$$

so that $\{Y^n/\lambda_F^n : n \geq 1\}$ is UI. Since $0 \leq \widetilde{Q}_F^{z,n}(\infty) \leq \widetilde{X}_F^{z,n}(\infty) \leq_{st} Y^n/\lambda_F^n$, $\{\widetilde{X}_F^{z,n}(\infty): n \geq 1\}$ and $\{\widetilde{Q}_F^{z,n}(\infty): n \geq 1\}$ are also UI.

Finally, the tightness of each of the sequences of (univariate) random variables $\{\widetilde{X}_i^{n,z}(\infty): n \geq 1\}$ and $\{\widetilde{Q}_i^{n,z}(\infty): n \geq 1\}$, $i = S, F$, follows from the fact that the sequence converges for $i = S$, and is stochastically bounded by the infinite-server queue for $i = F$. It also follows from the fact that each of these sequences is UI. This implies the claimed tightness in $\mathbb{R}^4$; see, e.g., Lemma 5.2 in Pang et al. (2007). □

Proofs of Theorem 1, Corollary 1, and Theorem 2. Taking $z = 0$, Theorem 3 and Theorem 4 immediately give Theorem 1, Corollary 1, and Theorem 2. □

Proof of Proposition 1. Consider a converging subsequence $\{z^n : n \geq 1\}$ satisfying (16) (we keep the superscript n for notational convenience), so that $z^n/R_F^n \to z$ for some $z \in [0,1]$, as $n \to \infty$. Theorems 3 and 4 show that $E[\widetilde{Q}_i^{z,n}(\infty)] \to E[Q_i^z(\infty)]$ for $i = S, F$, implying that $n^{-1/2}C^n(z^n) \to C(z)$. In particular, $n^{-1/2}C^n(z^{n*}) \to C(z^*)$ as $n \to \infty$. The result follows from the fact that $C(z) \geq C(z^*)$. □

Appendix D: Remaining Proofs

It remains to prove Lemma 1 and Proposition 5. The proofs of these two result, which appear in §D.2 below, build on technical lemmas which we state and prove in §D.1.

D.1. Auxiliary Results

Recall that $S = \mathbb{R} \times [0, \infty)$ and that $\{P_t : t \geq 0\}$ denotes the transition semi-group of the solution X^z to the HSDE (9)–(11).

LEMMA 4. *For any compact set $A \subset S$, there exists a non-trivial measure ν_A on S such that*

$$\int_0^\infty e^{-t} P_t(x, \cdot) dt \geq \nu_A(\cdot), \ \textit{for all } x \in A. \tag{40}$$

Proof. Consider the point $a := (r_F, 0) \in S$. For any $x = (x_S, x_F) \in S$, denote by $\tau(x)$ the next recurrence time of a, i.e., $\tau(x) = \inf\{t \geq 0 | X^z(t) = a\}$, given $X^z(0) = x$. Let $F_{\tau(x)}$ denote the cdf of $\tau(x)$. For any $T > 0$ and measurable set $A' \subseteq S$, we have

$$\begin{aligned}\int_0^\infty e^{-t} P_t(x, A') dt &\geq \int_0^\infty e^{-t} P(X^z(t) \in A', \tau(x) \leq t | X^z(0) = x) dt \\ &= \int_0^\infty e^{-t} \int_0^t P(X^z(t) \in A' | \tau(x) = s) dF_{\tau(x)}(s) dt\end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty e^{-t}\int_0^t P_{t-s}(a,A')dF_{\tau(x)}(s)dt\\
&= \int_0^\infty e^{-s}\int_s^\infty e^{s-t}P_{t-s}(a,A')dtdF_{\tau(x)}(s)\\
&= \int_0^\infty e^{-s}dF_{\tau(x)}(s)\int_0^\infty e^{-t}P_t(a,A')dt\\
&\ge e^{-T}\inf_{x\in A}P(\tau(x)\le T)\int_0^\infty e^{-t}P_t(a,A')dt.
\end{aligned}
\tag{41}
$$

Since $P_t(a,\cdot)$ is a non-trivial measure on S for each $t\ge 0$, $\int_0^\infty e^{-t}P_t(a,\cdot)dt$ is also a non-trivial measure on S, and the statement will follow if we can find $T>0$ such that

$$\inf_{x\in A}P(\tau(x)\le T)>0. \tag{42}$$

Due to the compactness of A, there is a constant K such that $|x_S|+|x_F|+2r_F+\theta_F^{-1}\le K$ holds for any $x\in A$. For $x_S=X_S^z(0)$, let $\tau_M(x_S)$ be the first hitting time of $-r_F$ by the process X_S^z after an excursion of this process to the set $(-\infty,-K)$ that lasted at least K time units, i.e., given that $X_S^z(0)=x_S$,

$$
\begin{aligned}
\tau_M(x_S) &:= \inf\{t\ge \tau_E(x_S): X_S^z(t)=-r_F\}, \text{ where}\\
\tau_E(x_S) &:= \inf\{t\ge K: \sup_{u\in[t-K,t]} X_S^z(u)<-K\}.
\end{aligned}
\tag{43}
$$

We next show that $\tau(x)\le\tau_M(x_S)$ for any $x\in A$ by showing that $X^z(\tau_M(x_S))=a$. Since $X_S^z(\tau_M(x_S))=-r_F$ by (43), we need only to show that $X_F^z(\tau_M(x_S))=0$. Observe that (10) implies that $dX_F^z(t)/dt<-1_{\{X_F^z(t)>0\}}$ whenever $X_S^z(t)<-K<-2r_F$. This implies that $X_F^z(t)$ at time $t=\tau_E(x_S)-K$ is no larger than K, so that

$$X_F^z(\tau_E(x_S))\le[X_F^z(\tau_E(x_S)-K)-K]^+=0,$$

where the latter equality follows from the choice of K and the fact that X_F^z is bounded from below by 0 and from above by $\theta_F^{-1}\vee X_F^z(0)$. It follows from (43) that $X_S^z(s)\le -r_F$ for $s\in[\tau_E(x_S),\tau_M(x_S)]$, and since $X_F^z(t)$ is non-increasing at t whenever $X_S^z(t)\le -r_F$, we conclude that $X_F^z(\tau_M(x_S))=0$, so that $X^z(\tau_M(x_S))=a$. Now, $X^z(\tau_M(x_S))=a$ implies that $\tau(x)\le\tau_M(x_S)$, and in turn,

$$\inf_{x\in A}P(\tau(x)\le T)\ge\inf_{x\in A}P(\tau_M(x_S)\le T), \text{ for all } T\ge 0. \tag{44}$$

Next we establish a lower bound for $\inf_{x\in A}P(\tau_M(x_S)\le T)$. For $-K\le x_S\le x_S'$, we have $\tau_M(x_S)\le_{st}\tau_M(x_S')$, as any trajectory of X_S^z from x_S' to $-K$ passes through x_S. Using $|x_S|\le K$ for $x\in A$, we obtain

$$\inf_{x\in A}P(\tau_M(x_S)\le T)\ge P(\tau_M(K)\le T) \text{ for all } T\ge 0. \tag{45}$$

Now, X_S^z is a piecewise OU process, as in Dieker and Gao (2013), that is (locally) distributed like an OU process on $(-r_F z, \infty)$, which we denote by X^+, and like a different OU process on $(-\infty, -r_F z)$, which we denote by X^-. (The OU processes X^+ and X^- are characterized by the SDEs obtained by considering (10) for $t \in (-r_F z, \infty)$ and $t \in (-\infty, -r_F z)$, respectively.) It follows from Dieker and Gao (2013) that X_S^z is positive recurrent with $\mathbb{R}$ being its support. Further, the distribution of excursions of X_S^z to $(-\infty, -K)$ is the distribution of excursions of the OU process X^+ to this set, so that $P(\tau_M(K) < \infty) = 1$. Therefore, $P(\tau_M(K) \le T) > 0$ for some $T > 0$, which together with (44) and (45), implies (42). □

For $x = (x_S, x_F) \in S$ let

$$V(x) := u(x_S) + u(x_F) + 1, \tag{46}$$

where

$$u(y) := \left(1 - \frac{2}{\pi}\cos(\frac{\pi}{2}y)\right) 1_{\{|y| \le 1\}} + |y| 1_{\{|y|>1\}}, \quad y \in \mathbb{R}.$$

It is easily checked that u is a twice continuously differentiable convex function that is minimized at $y = 0$, with $u(0) = 1$, $u'(0) = 0$, and with u'' having a finite support; in particular, $|u''| < \pi/2$. Further, $|y| \le u(y) \le |y| + 1$, for all $y \in \mathbb{R}$. Therefore, $V \ge 1$ is also twice continuously differentiable, with $V(0) = 3$ and

$$\|x\| + 1 \le V(x) \le 2\|x\| + 3. \tag{47}$$

In particular, $V \ge 1$ and $V(x) \to \infty$ as $\|x\| \to \infty$. Let $\mathcal{A}^z$ be the extended generator of X^z; see Equations (12)–(13) in Down et al. (1995).

Lemma 5 **(drift condition)**. *The function V in* (46) *is in the domain of the extended generator $\mathcal{A}^z$. Further, there exist positive constants c and d, whose values do not depend on the value of z, and a compact set $A \subset S$, such that the following drift condition holds for all $z \in [0,1]$ and $x = (x_S, x_F) \in S$*

$$\mathcal{A}^z V(x) + cV(x) \le d 1_{\{x \in A\}}.$$

Proof. We first prove that V in (46) is in the domain of the extended generator $\mathcal{A}^z$. To this end, let

$$b_S^z(x_S) := -\beta + r_F z + x_S^- - \theta_S x_S^+, \quad b_F^z(x_S, x_F) := 1 - z - r_F^{-1} x_S^- - \theta_F x_F,$$
$$U(x) := b_S^z(x_S) u'(x_S) + b_F^z(x) u'(x_F) + u''(x_S), \text{ for } x = (x_S, x_F) \in S.$$

By Ito's formula

$$\begin{aligned} V(X^z(t)) &= V(X^z(0)) + \int_0^t \left[\frac{\partial V(X^z(s))}{\partial x_S} dX_S^z(s) + \frac{\partial V(X^z(s))}{\partial x_F} dX_F^z(s)\right] + \int_0^t \frac{\partial^2 V(X^z(s))}{\partial x_S^2} ds \\ &= V(X^z(0)) + \int_0^t U(X^z(s)) ds + \int_0^t u'(X_F^z(s)) dI^z(s), \quad t \ge 0. \end{aligned}$$

Since $u'(0)=0$, $u'(y)>0$ for $y>0$, and $\int_0^t 1_{\{X_F^z(s)>0\}}dI^z(s)=0$, it holds that $\int_0^t u'(X_F^z(s))dI^z(s)=0$ for all $t\geq 0$, implying that

$$V(X^z(t))=V(X^z(0))+\int_0^t U(X^z(s))ds. \tag{48}$$

It remains to show that, for any $x\in S$,

$$E\left[\,\|U(X^z(\cdot))\|_t\,\middle|X^z(0)=x\right]<\infty,\quad t\geq 0. \tag{49}$$

Observe that $|U(x)|\leq C(\|x\|+1)$, $x\in S$, for some finite constant $C>0$, and recall that

$$0\leq X_F^z(s)\leq X_F^z(0)\vee\theta_F^{-1}\quad\text{for all } s\geq 0.$$

It follows from (34) and the Lipschitz continuity of ϕ, that $\|X_S^z\|_t\leq C_t(\|B\|_t+|X_S^z(0)|)$ for some $C_t<\infty$, where B is a standard Brownian motion. Now,

$$E[\|B\|_t]\leq\frac{1}{2}(1+E[\|B\|_t^2])\leq\frac{1}{2}(1+4\|E[B^2]\|_t])=\frac{1}{2}(1+4t)<\infty,$$

where the first inequality following from the fact that $0\leq(1-a)^2$ for all $a\in\mathbb{R}$, and the second inequality follows from Doob's martingale inequality. We conclude that (49) holds, which together with (48), implies that V is in the domain of $\mathcal{A}^z$ and that $\mathcal{A}^zV=U$.

The proof that the drift condition holds is done by direct computation. Since $|u''(y)|\leq\pi/2$ for $y\in\mathbb{R}$, it suffices to prove that

$$b_S^z(x_S)u'(x_S)+b_F^z(x)u'(x_F)+c(|x_S|+|x_F|)+\pi/2\leq d1_{\{x\in A\}}. \tag{50}$$

Take $\theta_m=\min\{1,\theta_S,\theta_F\}$, and observe that $|u'(y)|\leq 1$ for all $y\in\mathbb{R}$ and $u'(y)=y/|y|$ for $|y|\geq 1$. Hence,

$$b_F^z(x_S,x_F)u'(x_F)\leq 1-\theta_Fx_F\leq 1-\theta_mx_F\quad\text{for all } x_F\geq 0$$

We next establish an upper bound for $b_S^z(x_S)u'(x_S)+\theta_m|x_S|$. For $x_S<-1$ we have $u'(x_S)=-1$, so that

$$b_S^z(x_S)u'(x_S)+\theta_m|x_S|\leq -b_S^z(x_S)-x_S\leq\beta+(1+\theta_S)(r_F+1).$$

For $x_S>1$ we have $u'(x_S)=1$, so that

$$b_S^z(x_S)u'(x_S)+\theta_m|x_S|\leq b_S^z(x_S)+\theta_Sx_S\leq\beta+r_F;$$

Finally, for $|x_S|\leq 1$, we have $|u'(x_S)|\leq 1$, which implies that

$$b_S^z(x_S)u'(x_S)+\theta_m|x_S|\leq|b_S^z(x_S)|+1\leq\beta+(1+\theta_S)(r_F+1)+1.$$

Then for $d_S := \beta + (1+\theta_S)(r_F+1)+1$, we have

$$b_S^z(x_S)u'(x_S) + \theta_m|x_S| \le d_S, \text{ for all } x_S \in \mathbb{R} \text{ and } z \in [0,1],$$

and

$$b_S^z(x_S)u'(x_S) + b_F^z(x)u'(x_F) + \theta_m(|x_S|+|x_F|) \le d_S+1, \text{ for all } x \in S.$$

Taking $c = \theta_m/2$, $d = d_S + 1 + \pi/2$, and $A = \{x \in S : |x_S| + |x_F| \le d/c\}$ gives (50). □

D.2. Proofs of Proposition 5 and Lemma 1

Proof of Proposition 5. Lemma 4 implies that any compact set in S is a petite set. By Lemma 5, Condition $\tilde{\mathscr{D}}$ in (Down et al. 1995, §5) holds for the process X^z with the "norm-like" Lyapunov function V in (46). We can thus employ Theorem 5.2(c) in Down et al. (1995) to conclude that there exist $K < \infty$ and $\gamma < 1$, and a random variable $X^z(\infty) \in \mathbb{R}^2$, such that, for any measurable function $f : \mathbb{R}^2 \to \mathbb{R}$ satisfying $|f(x)| \le \|x\| + 1 \le V(x)$,

$$\big|E[f(X^z(t)|X^z(0)=x)] - E[f(X^z(\infty))]\big| \le KV(x)\gamma^t \le K(2\|x\|+3)\gamma^t, \tag{51}$$

where the second inequality in (51) follows from (47). The stated claims follow by observing that (51) implies (39), which in turn implies (38) by taking $f(x) = 1_{\{x\in\cdot\}}$. □

In the proof of Lemma 1 below we will employ sample-path comparisons between solutions to the HSDE. To this end, Observe that X_F^z is a deterministic function of X_S^z, so that the solution $X^z = (X_S^z, X_F^z)$ to the HSDE is also a deterministic function of X_S^z, and in turn, of the initial condition $X^z(0)$ and the Brownian motion B in (9). Since there exists a unique strong solution X_S^z to the SDE part of the HSDE, see, e.g., Dieker and Gao (2013), this implies that there exists a unique strong solution X^z to the HSDE, namely, a solution X^z that is on the same probability space and is adapted to the filtration generated by $X^z(0)$ and B.

Proof of Lemma 1. We first prove that $E[\|X^z(\infty)\|]$ is uniformly bounded with respect to $z \in [0,1]$. By the definition of the extended generator $\mathcal{A}^z$ (see (48)) we have the following identity

$$E[V(X^z(t))|X^z(0)=x] = E\left[\int_0^t \mathcal{A}^z V(X^z(s))ds\big|X^z(0)=x\right] + V(x), \text{ for any } t \ge 0.$$

Dividing both sides by t and using the drift condition in Lemma 5 gives

$$t^{-1}E[V(X^z(t))|X^z(0)=x] + ct^{-1}E\left[\int_0^t V(X^z(s))ds\big|X^z(0)=x\right] \le d + t^{-1}V(x), \text{ for any } t > 0.$$

Recall that the values of c and d do not depend on z. Applying Proposition 5, we have

$$\lim_{t\to\infty} t^{-1}E[V(X^z(t))|X^z(0)=x] = 0, \quad \text{and} \quad \lim_{t\to\infty} t^{-1}E\left[\int_0^t V(X^z(s))ds\big|X^z(0)=x\right] = E[V(X^z(\infty))].$$

This implies that $cE[V(X^z(\infty))] \le d$, for any $z \in [0,1]$, which together with the bound $V(x) \ge \|x\| + 1$, give

$$E[\|X^z(\infty)\|] \le E[V(X^z(\infty))] \le d/c, \quad \text{for all } z \in [0,1]. \tag{52}$$

Our goal is to prove that, for any $z \in [0,1]$ and $M > 0$, there is an $\epsilon > 0$, such that

$$\|E[Q_i^w(\infty)] - E[Q_i^z(\infty)]\| \le M^{-1}, \quad i \in \{S, F\} \tag{53}$$

holds for all $w \in [0,1]$ in an ϵ-neighborhood of z, namely, for all $w \in [0,1]$ that satisfies $|w - z| \le \epsilon$. To this end, let $G = (G_S, G_F) : S \times [0,1] \times C \to C^2$, where

$$G_S(x,z,y) := \phi(x_S - (\beta - r_F z)e + \sqrt{2}y), \tag{54}$$

$$G_F(x,z,y) := \eta_1\left(x_F + (1-z)e - \int_0^{\cdot} r_F^{-1} G_S(x,z,y)^- ds\right), \tag{55}$$

where $x \in S$, $z \in [0,1]$, $y \in C$, and $\eta_1 : C_+ \to C$ is the map defined via $\eta_1(y) = q$, for q in (19), namely, η_1 is the projection of η to its first coordinate.

For $w \in [0,1]$ and a random variable $X^w(0) \in S$, let X^w be the solution to the HSDE (9)–(11) with fast track of size w, and let B^w be the corresponding standard Brownian motion. From (54), (55), and the proof of Lemma 3, $X^w = G(X^w(0), w, B^w)$. By Proposition 5, X^w has a unique stationary distribution $X^w(\infty)$, which does not depend on the initial state $X^w(0)$. In particular, we can take $X^w(0) \overset{\mathrm{d}}{=} X^w(\infty)$, so that $E[X_i^w(t)] = E[X_i^w(\infty)]$ for all $t \ge 0$ and $i \in \{S, F\}$. By (12), we have

$$Q^w = ((X_S^w)^+, X_F^w) \quad \text{and} \quad Q^w(\infty) = (X_S^w(\infty)^+, X_F^w(\infty)), \tag{56}$$

so that $E[Q_i^w(t)] = E[Q_i^w(\infty)]$ for all $t \ge 0$ and $i \in \{S, F\}$.

Next, fix $z \in [0,1]$ and let $X^{z,w} := G(X^w(0), z, B^w)$, so that $X^{z,w}$ is a solution to the HSDE (9)–(11) with the same initial condition and Brownian motion as in the solution X^w, but with fast-track of size z. Let $Q^{z,w} := ((X_S^{z,w})^+, X_F^{z,w})$, and $Q^{z,w}(\infty)$ denote a random variable with the corresponding limiting distribution, and observe that $Q^{z,w}(\infty) \overset{\mathrm{d}}{=} Q^z(\infty)$ for any $w \in [0,1]$; namely, it is independent of w. Take $f_S(x) := x_S^+$ and $f_F(x) := x_F$, respectively, in Proposition 5. Then there are constants $K_1 > 0$, $K_2 > 0$, and $0 < \gamma < 1$, such that

$$|E[Q_i^{z,w}(t)] - E[Q_i^{z,w}(\infty)]| \le \sup_{x \in \mathbb{R}^2} \frac{|f_i(x)|}{\|x\| + 1}(K_1 + K_2 E[\|X^{z,w}(0)\|])\gamma^t \le \sup_{x \in \mathbb{R}^2} \frac{|f_i(x)|}{\|x\| + 1}(K_1 + K_2 d/c)\gamma^t,$$

where the last inequality follows from (52) and the fact that $X^{z,w}(0) = X^w(0) \overset{\mathrm{d}}{=} X^w(\infty)$. Since $|f_i(x)| \le (1 + r_F)(\|x\| + 1)$ for $i \in \{S, F\}$ and $x \in \mathbb{R}^2$,

$$\left|E[Q_i^{z,w}(t)] - E[Q_i^{z,w}(\infty)]\right| \le (1 + r_F)(K_1 + K_2 d/c)\gamma^t, \text{ for all } t \ge 0.$$

Finally, fix $M > 0$ and take $t > 0$ such that $(1+r_F)(K_1+K_2d/c)\gamma^t < (2M)^{-1}$. We have

$$\begin{aligned}\big|E[Q_i^w(\infty)] - E[Q_i^{z,w}(\infty)]\big| &\le \big|E[Q_i^w(t) - Q_i^{z,w}(t)]\big| + \big|E[Q_i^{z,w}(t)] - E[Q_i^{z,w}(\infty)]\big| \\ &\le E[\|Q^w - Q^{z,w}\|_t] + \frac{1}{2M}.\end{aligned}$$

The inequality $|a^+ - b^+| \le |a-b|$, for all a, $b \in \mathbb{R}$ implies that

$$\big|(X_S^w(t))^+ - (X_S^{z,w}(t))^+\big| \le \big|X_S^w(t) - X_S^{z,w}(t)\big|,$$

for all $t \ge 0$. Together with $Q_F^w = X_F^w$ and $Q_F^{z,w} = X_F^{z,w}$, we have for $i \in \{S, F\}$,

$$\begin{aligned}\big|E[Q_i^w(\infty)] - E[Q_i^{z,w}(\infty)]\big| &\le E[\|X^w - X^{z,w}\|_t] + \frac{1}{2M} \\ &= E[\|G(X^w(0), w, B^w) - G(X^w(0), z, B^w)\|_t] + \frac{1}{2M}.\end{aligned}$$

Due to the Lipschitz continuity of η_1 and ϕ, there is a constant $c_t > 0$ such that

$$\|G(x,w,y) - G(x,z,y)\|_t \le c_t|w-z|, \text{ for all } x \in S \text{ and } y \in C. \tag{57}$$

Taking $\epsilon \in (0, \frac{1}{2Mc_t})$ gives

$$\big|E[Q_i^w(\infty)] - E[Q_i^{z,w}(\infty)]\big| \le c_t|w-z| + \frac{1}{2M} \le c_t\epsilon + \frac{1}{2M} \le \frac{1}{M},$$

for all w such that $|w-z| \le \epsilon$. Then (53) follows from the fact that $Q^{z,w}(\infty) \overset{\mathrm{d}}{=} Q^z(\infty)$ for all w, i.e., $Q^{z,w}(\infty)$ is independent of w. □

Appendix E: Proofs of Proposition 2 and Proposition 3.

In the proof we repeatedly use the inequality $(a+b)^+ \le a^+ + b^+$ for a, $b \in \mathbb{R}$, and the following generalized Gronwall's inequality, see, e.g., Theorem 1.3.5 in Pachpatte (1997).

Lemma 6. *If $f \in C$ and $\theta \in \mathbb{R}_+$ satisfy $f(t) \ge -\theta\int_0^t f(s)^+ds$ for all $t \ge 0$, then $f \ge 0$.*

Consider the map G_S in (54). We first prove

Lemma 7. *The map $z \mapsto G_S(x + r_Fz, z, y) - r_Fz$ is non-decreasing and convex.*

Proof. Let $z_i \in [0,1]$ for $i = 1$, 2, 3 satisfy $z_1 \ge z_2$ and $2z_3 = z_1 + z_2$, and let

$$u_i := G_S(x + r_Fz_i, z_i, y) - r_Fz_i, \text{ for } i = 1, 2, 3.$$

It follows (18) and (54) that

$$\begin{aligned}u_i(t) + r_Fz_i &= \phi(x + r_Fz_i, z_i, y) \\ &= x_S + r_Fz_i - \beta t + \int_0^t (u_i(s) + r_Fz_i)^-ds - \theta_S\int_0^t (u_i(s) + r_Fz_i)^+ds + \sqrt{2}y(t).\end{aligned}$$

Using $a = a^+ - a^-$ for $a \in \mathbb{R}$, we have

$$u_i(t) = x_S - \beta t - \int_0^t u_i(s)ds + (1-\theta_S)\int_0^t (u_i(s) + r_F z_i)^+ ds + \sqrt{2}y(t). \tag{58}$$

Therefore,

$$\begin{aligned} u_1(t) - u_2(t) &= \int_0^t (u_2(s) - u_1(s))ds + (1-\theta_S)\int_0^t \left((u_1(s) + r_F z_1)^+ - (u_2(s) + r_F z_2)^+\right) ds \\ &\geq \int_0^t (u_2(s) - u_1(s))ds - (1-\theta_S)\int_0^t (u_2(s) + r_F z_2 - u_1(s) - r_F z_1)^+ ds \\ &\geq \int_0^t (u_2(s) - u_1(s))ds - \int_0^t (u_2(s) - u_1(s))^+ ds \\ &\geq -\int_0^t (u_1(s) - u_2(s))^+ ds, \end{aligned} \tag{59}$$

By Lemma 6 we have $u_1 \geq u_2$; hence, G_S is increasing in z as stated.

To prove the claimed convexity, it suffices to show that $u_1 + u_2 \geq 2u_3$, namely, that G_S is midpoint convex; see, e.g., (Roberts and Varberg 1974, pp. 220–221). First notice that

$$u_1^+ + u_2^+ - 2u_3^+ \geq -(2u_3 - u_1 - u_2)^+.$$

We can employ similar arguments as in (59) and conclude that

$$u_1(t) + u_2(t) - 2u_3(t) \geq -\int_0^t (u_1(s) + u_2(s) - 2u_3(s))^+ ds, \text{ for all } t \geq 0. \tag{60}$$

It then follows from Lemma 6 that $u_1 + u_2 \geq 2u_3$. □

Recall that $C_q(z) = C(z) - dr_F z$.

Lemma 8. *$C_q : [0,1] \to \mathbb{R}_+$ is strictly increasing and strictly convex.*

Proof. It follows from (9) and Theorem 4 that

$$0 = -\beta + r_F z + E[X_S^z(\infty)^-] - \theta_S E[X_S^z(\infty)^+].$$

By Assumption 5 we have $1 - \theta_S > 0$. Using $x = x^+ - x^-$ for $x \in \mathbb{R}$, we have

$$E[Q_S^z(\infty)] = E[X_S^z(\infty)^+] = (1-\theta_S)^{-1}(\beta + E[X_S^z(\infty)] - r_F z). \tag{61}$$

A straightforward computation gives

$$C_q(z) = (c_S - c_F)E[Q_S^z(\infty)] + \frac{c_F \beta}{1-\theta_S} + c_F E\left[\frac{X_S^z(\infty) - r_F z}{1-\theta_S} + r_F X_F^z(\infty)\right]. \tag{62}$$

We the initialize process X^z at $x_0^z := (-r_F z, 0) \in \mathbb{R}^2$. Using (54) we have $X_S^z = G_S(x_0^z, z, B)$. By Lemma 7 and (61), the map $z \mapsto X_S^z - r_F z$ is non-decreasing and convex. It then follows Theorem

4 that $E[X_S^z(\infty) - r_F z]$ is non-decreasing and convex in z. By (61). the map $z \mapsto Q_S^z(\infty)$ is also non-decreasing and convex. Let

$$Y^z(t) := (1-\theta_S(t))^{-1}(X_S^z(t) - r_F z) + r_F X_F^z(t), \quad t \geq 0,$$

and let $Y^z(\infty)$ be the stationary distribution of Y^z. By Theorem 4, $E[Y^z(t)] \to E[Y^z(\infty)]$ as $t \to \infty$. By Assumption 5 we have $c_S > c_F$. Therefore, it is sufficient to prove that (i) the map $z \mapsto Y^z$ is nondecreasing and convex, and (ii) the map $z \mapsto E[Q_S^z(\infty)]$ is strictly increasing and strictly convex.

To prove (i), we compute Y^z via (9) and (10):

$$Y^z(t) = \int_0^t \left[\frac{\theta_F - \theta_S}{1-\theta_S}(X_S^z(s) - r_F z) - \theta_F Y^z(s)\right] ds + \left(r_F - \frac{\beta}{1-\theta_S}\right) t + \frac{\sqrt{2} dB(t)}{1-\theta_S} + r_F dI^z(t). \quad (63)$$

As $X_F^z(0) = 0$ and X_F^z solves a generalized Skorohod problem, we can write

$$X_F^z(t) = \left[\int_0^t \left(1 - z - r_F^{-1} X_S^z(s)^- - \theta_F X_F^z(s)\right) ds\right] + I^z(t),$$

In other words, X_F^z is also the solution to a Skorohod problem (not generalized), where the regulator process I^z has the following representation:

$$I^z(t) = \sup_{s\in[0,t]} \left[\int_0^s \left(-1 + z + r_F^{-1} X_S^z(s_1)^- + \theta_F X_F^z(s_1)\right) ds_1\right]^+.$$

Combine with (9), we can compute

$$\begin{aligned} r_F I^z(t) = \sup_{s\in[0,t]} \Bigg[& \left(\frac{\beta}{1-\theta_S} - r_F\right) s + \frac{X_S^z(s) - r_F z - \sqrt{2}B(s)}{1-\theta_S} \\ & + \int_0^s \frac{\theta_F - \theta_S}{1-\theta_S}(X_S^z(s_1) - r_F z) ds_1 - \theta_F \int_0^s Y^z(s_1) ds_1 \Bigg]^+. \end{aligned} \quad (64)$$

To prove that Y^z is non-decreasing in z, take z_1, $z_2 \in [0,1]$ such that $z_1 \geq z_2$. It is straightforward to check that, for any a, $b \in C$ and $t \geq 0$,

$$\sup_{s\in[0,t]} [a(s)^+] - \sup_{s\in[0,t]} [b(s)^+] \geq - \sup_{s\in[0,t]} [(b(s) - a(s))^+]. \quad (65)$$

Using (65), the fact that $z \mapsto X_S^z - r_F z$ is nondecreasing and $\theta_F \geq \theta_S$ (Assumption 5) imply that

$$\begin{aligned} r_F I^{z_1}(t) - r_F I^{z_2}(t) \geq & - \sup_{s\in[0,t]} \Bigg[\frac{1}{1-\theta_S} \left(X_S^{z_2}(s) - X_S^{z_1}(s) - r_F(z_2 - z_1)\right) \\ & + \int_0^s \frac{\theta_F - \theta_S}{1-\theta_S} \left(X_S^{z_2}(s_1) - X_S^{z_1}(s_1) - r_F(z_2 - z_1)\right) ds_1 \\ & + \theta_F \int_0^s (Y^{z_2}(s_1) - Y^{z_1}(s_1)) ds_1 \Bigg]^+ \\ \geq & - \sup_{s\in[0,t]} \left[\theta_F \int_0^s (Y^{z_2}(s_1) - Y^{z_1}(s_1)) ds_1\right]^+ \\ \geq & - \theta_F \int_0^t (Y^{z_1}(s) - Y^{z_2}(s))^- ds. \end{aligned} \quad (66)$$

Plugging (66) in (63) and using the fact that $X_S^z - r_F z$ is nondecreasing in z give

$$\begin{aligned}Y^{z_1}(t)-Y^{z_2}(t) \geq & \theta_F\int_0^t (Y^{z_2}(s)-Y^{z_1}(s))\,ds - \theta_F\int_0^t (Y^{z_1}(s)-Y^{z_2}(s))^- ds\\ \geq & -\theta_F\int_0^t (Y^{z_1}(s)-Y^{z_2}(s))^+ ds.\end{aligned}$$

By Lemma 6, we can conclude that $Y^{z_1}(t)\geq Y^{z_2}(t)$ for all $t\geq 0$.

To prove that Y^z is convex in z let $z_i\in[0,1]$ for $i=1,2,3$ such that $z_3=(z_1+z_2)/2$. It is straightforward to check that, for any a, b, $c\in C$ and $t\geq 0$,

$$\sup_{s\in[0,t]}[a(s)^+] + \sup_{s\in[0,t]}[b(s)^+] - 2\sup_{s\in[0,t]}[c(s)^+] \geq -\sup_{s\in[0,t]}[2c(s)^+ - a(s)^+ - b(s)^+]$$

Similar to the computation of (66), using the fact that X_S^z is convex in z, we obtain

$$r_F(I^{z_1}(t)+I^{z_2}(t)-2I^{z_3}(t)) \geq -\theta_F\int_0^t (Y^{z_1}(s)+Y^{z_2}(s)-2Y^{z_3}(s))^- ds.$$

Plugging in (63), we conclude that

$$Y^{z_1}(t)+Y^{z_2}(t)-2Y^{z_3}(t) \geq -\theta_F\int_0^t (Y^{z_1}(s)+Y^{z_2}(s)-2Y^{z_3}(s))^+ ds.$$

By Lemma 6, we have $Y^{z_1}+Y^{z_2}\geq 2Y^{z_3}$, which finishes the proof of (i).

To prove (ii), we take z_1, $z_2\in[0,1]$ such that $z_1>z_2$ and define the interval $L:=[(z_2-z_1)r_F/4, 0]$. By (61) we have $Q_S^z(\infty)=X_S^z(\infty)^+$ and $(1-\theta_S)E[(X_S^z(\infty))^+] = E[X_S^z(\infty)-r_F z]+\beta$. By Lemma 7 we have $E[Q_S^{z_1}(\infty)]\geq E[Q_S^{z_2}(\infty)]$. To prove the latter inequality is strict, we recall that

$$X_S^{z_1}(t)-r_F z_1 \geq X_S^{z_2}(t)-r_F z_2 \text{ for all } t\geq 0.$$

Therefore, if $X_S^{z_2}(t)\in L$ for some $t\geq 0$, then $X_S^{z_1}(t)^+\geq X_S^{z_2}(t)^+ + (z_1-z_2)r_F/2$. Taking expectation, we obtain

$$E[X_S^{z_1}(t)^+]\geq E[X_S^{z_2}(t)^+] + P(X_S^{z_2}(t)\in L)(z_1-z_2)r_F/2.$$

for all $t\geq 0$. By Proposition 5 we have $E[X_S^{z_i}(t)]\to E[X_S^{z_i}(\infty)]$ and $P(X_S^{z_2}(t)\in L)\to P(X_S^{z_2}(\infty)\in L)$ as $t\to\infty$, implying that

$$E[X_S^{z_1}(\infty)^+]\geq E[X_S^{z_2}(\infty)^+] + P(X_S^{z_2}(\infty)\in L)(z_1-z_2)r_F/2, \tag{67}$$

As $X_S^{z_2}(\infty)$ has support on $\mathbb{R}$, we have $P(X_S^{z_2}(\infty)\in L)>0$. Therefore we have $E[X_S^{z_1}(\infty)^+] > E[X_S^{z_2}(\infty)^+]$, implying that $E[Q_S^{z_1}(\infty)] > E[Q_S^{z_2}(\infty)]$. Thus, $z\mapsto E[Q_S^z(\infty)]$ is strictly increasing.

To prove that $z\mapsto E[Q_S^z(\infty)]$ is also strictly convex, take $z_3=(z_1+z_2)/2$, and recall from statement (i) that $X_S^{z_1}+X_S^{z_2}\geq 2X_S^{z_3}$ and $X_S^{z_1}\geq X_S^{z_3}+(z_1-z_2)r_F/2$. Then

$$X_S^{z_1}(t)^+ + X_S^{z_2}(t)^+ \geq (X_S^{z_3}(t)+(z_1-z_2)r_F/2)^+ \geq (z_1-z_2)r_F/4 = 2X_S^{z_3}(t)^+ + (z_1-z_2)r_F/4,$$

hold for all $t \geq 0$ given $X_S^{z_3}(t) \in L$. In turn,

$$E[X_S^{z_1}(t)^+] + E[X_S^{z_2}(t)^+] \geq 2E[X_S^{z_3}(t)^+] + (z_1 - z_2) r_F / 4 \, P(X_S^{z_3}(t) \in L).$$

Taking $t \to \infty$ and using the fact that $P(X_S^{z_3}(\infty) \in L) > 0$, we have

$$E[Q_S^{z_1}(\infty)] + E[Q_S^{z_2}(\infty)] > 2E[Q_S^{z_3}(\infty)],$$

which finishes the proof of (ii). □

Proofs of Proposition 2 and Proposition 3. Proposition 2 and Proposition 3 follow immediately from Lemma 8. □

References

Afeche P (2013) Incentive-compatible revenue management in queueing systems: optimal strategic delay. *Manufacturing & Service Operations Management* 15(3):423–443.

Asmussen S (2008) *Applied probability and queues* (Springer Science & Business Media).

Asmussen S, Glynn PW (2007) *Stochastic simulation: algorithms and analysis*, volume 57 (Springer Science & Business Media).

Ata B, Van Mieghem JA (2009) The value of partial resource pooling: Should a service network be integrated or product-focused? *Management Science* 55(1):115–131.

Atar R (2012) A diffusion regime with nondegenerate slowdown. *Operations Research* 60(2):490–500.

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

Atar R, Gurvich I (2014) Scheduling parallel servers in the nondegenerate slowdown diffusion regime: Asymptotic optimality results. *The Annals of Applied Probability* 24(2):760–810.

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

Bassamboo A, Harrison JM, Zeevi A (2009) Pointwise stationary fluid models for stochastic processing networks. *Manufacturing & Service Operations Management* 11(1):70–89.

Bell SL, Williams RJ (2001) Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *The Annals of Applied Probability* 11(3):608–649.

Billingsley P (2009) *Convergence of probability measures* (John Wiley & Sons).

Braverman A, Dai J (2017) Stein's method for steady-state diffusion approximations of $M/Ph/n+M$ systems. *The Annals of Applied Probability* 27(1):550–581.

Coffman Jr E, Puhalskii A, Reiman MI (1995) Polling systems with zero switchover times: a heavy-traffic averaging principle. *The Annals of Applied Probability* 681–719.

Cooke M, Wilson S, Pearson S (2002) The effect of a separate stream for minor injuries on accident and emergency department waiting times. *Emergency Medicine Journal* 19(1):28–30.

Dai J, He S, Tezcan T, et al. (2010) Many-server diffusion limits for G/Ph/n+ GI queues. *The Annals of Applied Probability* 20(5):1854–1890.

Dieker AB, Gao X (2013) Positive recurrence of piecewise ornstein–uhlenbeck processes and common quadratic lyapunov functions. *The Annals of Applied Probability* 23(4):1291–1317.

Down D, Meyn SP, Tweedie RL (1995) Exponential and uniform ergodicity of markov processes. *The Annals of Probability* 1671–1691.

Eitel DR, Travers DA, Rosenau AM, Gilboy N, Wuerz RC (2003) The emergency severity index triage algorithm version 2 is reliable and valid. *Academic Emergency Medicine* 10(10):1070–1080.

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

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

Ghamami S, Ward AR (2013) Dynamic scheduling of a two-server parallel server system with complete resource pooling and reneging in heavy traffic: Asymptotic optimality of a two-threshold policy. *Mathematics of Operations Research* 38(4):761–824.

Gilboy N, Tanabe P, Travers D, Rosenau AM (2012) *Emergency Severity Index (ESI): a triage tool for emergency department care, Version 4* (Agency for healthcare research and quality).

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

Gurvich I, Lariviere MA, Ozkan C (2018) Coverage, coarseness, and classification: Determinants of social efficiency in priority queues. *Management Science* .

Gurvich I, Perry O (2012) Overflow networks: Approximations and implications to call center outsourcing. *Operations Research* 60(4):996–1009.

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

Harrison JM (1998) Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Annals of applied probability* 822–848.

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

Hinch EJ (1991) *Perturbation methods* (Cambridge university press).

Hunt P, Kurtz T (1994) Large loss networks. *Stochastic Processes and their Applications* 53(2):363–378.

Iglehart DL (1965) Limiting diffusion approximations for the many server queue and the repairman problem. *Journal of Applied Probability* 2(2):429–441.

Khasminskii R, Yin G (2005) Limit behavior of two-time-scale diffusions revisited. *Journal of Differential Equations* 212(1):85–113.

Maglaras C, Yao J, Zeevi A (2017) Optimal price and delay differentiation in large-scale queueing systems. *Management Science* 64(5):2427–2444.

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

Maglaras C, Zeevi A (2005) Pricing and design of differentiated services: Approximate analysis and structural insights. *Operations Research* 53(2):242–262.

Moyal P, Perry O (2017) On the instability of matching queues. *The Annals of Applied Probability* 27(6):3385–3434.

Nazerzadeh H, Randhawa RS (2018) Near-optimality of coarse service grades for customer differentiation in queueing systems. *Production and Operations Management* 27(3):578–595.

Pachpatte BG (1997) *Inequalities for differential and integral equations* (Elsevier).

Pang G, Perry O (2014) A logarithmic safety staffing rule for contact centers with call blending. *Management science* 61(1):73–91.

Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for markovian queues. *Probability Surveys* 4(193-267):7.

Perry O, Whitt W (2009) Responding to unexpected overloads in large-scale service systems. *Management Science* 55(8):1353–1367.

Perry O, Whitt W (2011) A fluid approximation for service systems responding to unexpected overloads. *Operations Research* 59(5):1159–1170.

Perry O, Whitt W (2012) A fluid limit for an overloaded X model via a stochastic averaging principle. *Mathematics of Operations Research* 38(2).

Perry O, Whitt W (2015) Achieving rapid recovery in an overload control for large-scale service systems. *INFORMS Journal on Computing* 27(3):491–506.

Perry O, Whitt W (2016) Chattering and congestion collapse in an overload switching control. *Stochastic Systems* 6(1):132–210.

Roberts A, Varberg D (1974) *Convex functions*. ISSN (Elsevier Science), ISBN 9780080873725, URL `https://books.google.com/books?id=cqyHkkCxVtcC`.

Sanchez M, Smally AJ, Grant RJ, Jacobs LM (2006) Effects of a fast-track area on emergency department performance. *The Journal of emergency medicine* 31(1):117–120.

Song H, Tucker AL, Murrell KL (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay URL `http://nrs.harvard.edu/urn-3:HUL.InstRepos:11591702`.

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

Whitt W (1971) Weak convergence theorems for priority queues: preemptive-resume discipline. *Journal of Applied Probability* 8(1):74–94.

Whitt W (1980) Some useful functions for functional limit theorems. *Mathematics of operations research* 5(1):67–85.

Whitt W (1983) Comparison conjectures about the $M/G/s$ queue. *Operations Research Letters* 2(5):203–209.

Whitt W (1991) The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science* 37(3):307–314.

Whitt W (1992) Understanding the efficiency of multi-server service systems. *Management Science* 38(5):708–723.

Whitt W (2002) *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues* (Springer Science & Business Media).

Whitt W (2004) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* 50(10):1449–1461.

Whitt W (2005) Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Mathematics of Operations Research* 30(1):1–27.

Whitt W (2018) Time-varying queues. *Queueing models and service management* 1(2).

Wu S, Zhang J, Zhang RQ (2018) Management of a shared-spectrum network in wireless communications. *Operations research* 66(4):1119–1135.

Yin GG, Zhang Q (2005) *Discrete-time Markov chains: two-time-scale methods and applications*, volume 55 (Springer Science & Business Media).

Yin GG, Zhang Q (2012) *Continuous-time Markov chains and applications: a two-time-scale approach*, volume 37 (Springer Science & Business Media).