Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Lun Yu, Seyed Iravani, Ohad Perry

To cite this article:
Lun Yu, Seyed Iravani, Ohad Perry (2022) A Fluid-Diffusion-Hybrid Limiting Approximation for Priority Systems with Fast and Slow Customers. Operations Research 70(4):2579-2596. https://doi.org/10.1287/opre.2021.2154

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 70, No. 4, July–August 2022, pp. 2579–2596
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

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

Lun Yu,[a] **Seyed Iravani,**[a] **Ohad Perry**[a]

[a] Department of Industrial Engineering and Management Science, Northwestern University, Evanston, Illinois 60208

Contact: lunyu2014@u.northwestern.edu, https://orcid.org/0000-0003-0044-2514 (LY); s-iravani@northwestern.edu (SI); ohad.perry@northwestern.edu, https://orcid.org/0000-0002-4584-3015 (OP)

Received: May 29, 2019
Revised: April 9, 2020; October 19, 2020; February 2, 2021
Accepted: April 6, 2021
Published Online in Articles in Advance: November 15, 2021

Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2021.2154

Abstract. 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 nonurgent class. We therefore refer to the urgent class as "slow," and to the nonurgent 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, whereas 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, whereas 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.

Funding: O. Perry and L. Yu were partially supported by the National Science Foundation [Grant CMMI 1763100].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2021.2154.

Keywords: priority queue • many-server heavy-traffic • fluid-diffusion hybrid • queueing network

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 "nonurgent" (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 (nonurgent) 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 "nonurgent" 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 nonurgent ones, they may nevertheless experience delay. As we elaborate, this phenomenon presents modeling and analytical challenges, because it cannot be captured by standard many-server asymptotic regimes. Further, delays for both customer classes can coexist 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 nonurgent 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). Although 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, whereas 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 United States 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, whereas 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. Because ESI-1 patients constitute approximately 1%–3% of all ER patients, Eitel et al. (2003), and because 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, because 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 nonnegligible 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 nonurgent class abandons the queue.

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

(See Section 2 for a more rigorous discussion on the four scenarios.) 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)), such that both customer classes experience nonnegligible 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. Because 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), whereas 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 limit (i) both classes, including the high-priority class, have a nonnegligible probability of waiting for service, and (ii) both classes, including the low-priority class, receive service.

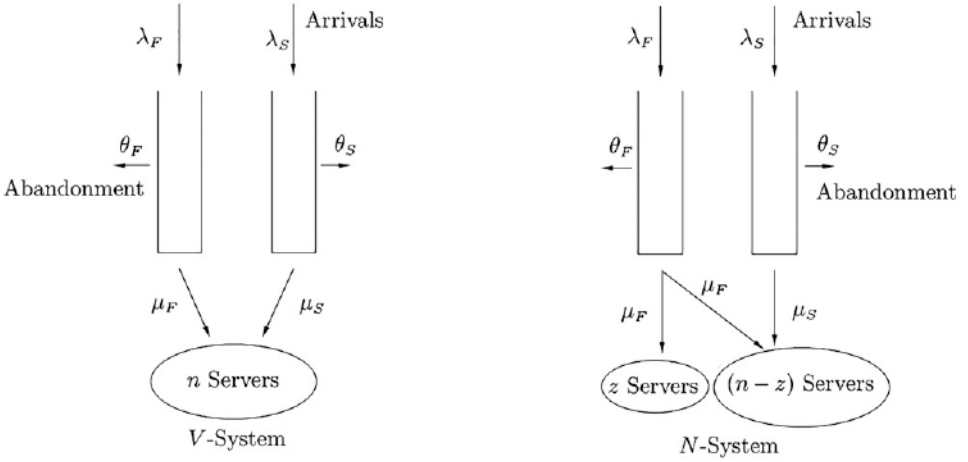
(2) We employ the FDH limit to study the potential benefits of de-pooling (the N-model). We show that a small number of dedicated servers, that is an order of magnitude smaller than the total number of servers, can substantially reduce the overall congestion in the system by reducing the delay of the 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., Cooke et al. 2002, Sanchez et al. 2006).

(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, for example, Whitt (2002), and write D for D^1.

Figure 1. A Single-Pool "V-system" (Left), and a Two-Pool "N-system" with a Fast Track (Right)



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** $o_P(a^n)$.

1.3.1. 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.

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. Because 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 nontrivial limits are achieved.

2.1. 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 $\rho_n \ge 1$ (and $\beta \le 0$) 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 nonnegligible, 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 *nondegenerate 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 nonnegligible 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.)

2.2. Relevant Insights

With the insights obtained from the single-class single-pool setting, we can explain why the four scenarios in Section 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, whereas 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, for example, 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.

Table 1. Summary of Existing MSHT Regimes for Two-Class Priority Systems

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

2.3. 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. Because 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; for example, 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, for example, 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. (2019), 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.

3.1. 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, for example, 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, section 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, for example, 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 et al. (1995). We refer to Gurvich and Perry (2012) and Perry and Whitt (2013) 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, because 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 multiclass multipool 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. 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 nonidling, 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 nonpreemptive priority policies diminishes as the size of the system increases, so that our results are meaningful also if the nonpreemptive 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 nontrivial limit. It will become clear (see the discussion below Theorem 1) that, because 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.*

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

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

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

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

$$\tilde{X}^n := \left(\tilde{X}_S^n, \tilde{X}_F^n\right) = \left(\frac{X_S^n - n}{\sqrt{\lambda_S^n}}, \frac{X_F^n}{\lambda_F^n}\right) \quad \text{and}$$

$$\tilde{Q}^n := (\tilde{Q}_S^n, \tilde{Q}_F^n) = \left(\frac{Q_S^n}{\sqrt{\lambda_S^n}}, \frac{Q_F^n}{\lambda_F^n}\right). \tag{2}$$

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 $\tilde{X}^n$ in (2) depends on having the sequence of initial conditions $\tilde{X}^n(0)$ converge in $\mathbb{R}^2$. We therefore must guarantee that the initial conditions in the limit and the prelimit 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 $\{\tilde{X}^n : n \geq 1\}$. 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 $\tilde{X}^n(0) \Rightarrow X(0)$ in $\mathbb{R}^2$, then $\tilde{X}^n \Rightarrow X$ in D^2, where $X := (X_S, X_F)$ is the unique solution to the following HSDE with initial condition* X(0)

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

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

where B is a standard Brownian motion, and I is the unique nondecreasing process satisfying

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

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 prelimit. This explains why the service rate of the fast class must scale at a rate $\sqrt{n}$. Further, due to the fluid scaling of $\tilde{X}_F^n$, the limit process X_F is therefore reflected at 0, and its nonnegativity 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 toward $[0, \theta_F^{-1})$ 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 $\{\tilde{Q}^n : n \geq 1\}$: For each $n \geq 1$, we have $Q_S^n = (X_S^n - n)^+$ and $Q_S^n + Q_F^n = (X_S^n + X_F^n - n)^+$. Therefore, Theorem 1 and the continuous mapping theorem imply that $Q := (Q_S, Q_F) := (X_S^+, X_F)$ is the FDH limit of $\{\tilde{Q}^n : n \geq 1\}$. (Notice that $\tilde{X}_F^n$ and $\tilde{Q}_F^n$ 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 prelimit *cumulative idleness process*

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

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

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

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 $\tilde{I}^n$. We summarize in the following corollary to Theorem 1.

Corollary 1. *If $\tilde{X}^n(0) \Rightarrow X(0)$ in $\mathbb{R}^2$, then $(\tilde{X}^n, \tilde{Q}^n, \tilde{I}^n) \Rightarrow (X, Q, I)$ 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 $\tilde{X}_S^n$, $n \geq 1$, 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 $(\tilde{X}_S^n, \tilde{X}_F^n)$.

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 $\{\tilde{X}^n : n \geq 1\}$ 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

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

$$\tilde{Q}^n(\infty) := \left(\tilde{Q}^n_S(\infty), \tilde{Q}^n_F(\infty)\right) = \left(\frac{Q^n_S(\infty)}{\sqrt{\lambda^n_S}}, \frac{Q^n_S(\infty)}{\lambda^n_F}\right).$$

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

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

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

3. $\{\tilde{Q}^n_i(\infty) : n \geq 1\}$ *is Uniformly Integrable (UI) for* $i = S, F$, *so that*

$$\lim_{n\to\infty} E[\tilde{Q}^n_i(\infty)] = E[Q_i(\infty)].$$

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^n_F 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. Because 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, because $X_F(\infty)$ 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.

4.3.1. 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 prelimit stochastic system. For $i = S, F$ and for n large, we consider the following measures: the probability of delay in queue $P(W^n_i > 0)$; 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) $E[W^n_i | W^n_i > 0]$; and the probability of abandonment $P(Ab^n_i)$.

The approximation of $P(W^n_S > 0)$ is straightforward: Because the event $\{W^n_S > 0\}$ is equivalent to the event $\{X^n_S \geq n\}$, both events have the same probability. Because $X_S(\infty)$ 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 $P(W^n_F > 0)$ 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 $\tilde{Q}^n_F(\infty) \Rightarrow Q_F(\infty)$ in $\mathbb{R}$ does not directly imply that $P(Q^n_F(\infty) > 0) \to P(Q_F(\infty) > 0)$ 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 $P(Q^n_F(\infty) = 0)$ for large n, so that $P(Q_F(\infty) > 0)$ also approximates $P(Q^n_F(\infty) > 0)$. To see why, note that $\{Q_F(t) = 0\}$ implies that $\{X_S(t) \leq -r_F\}$, 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

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

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

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 five 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 run 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%.

It is useful to contrast the simulation results with existing many-server asymptotic approximations.

Table 2. Comparison of Performance Measures for a Stochastic System and Its FDH Approximation

	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)

Notes. 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.

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.

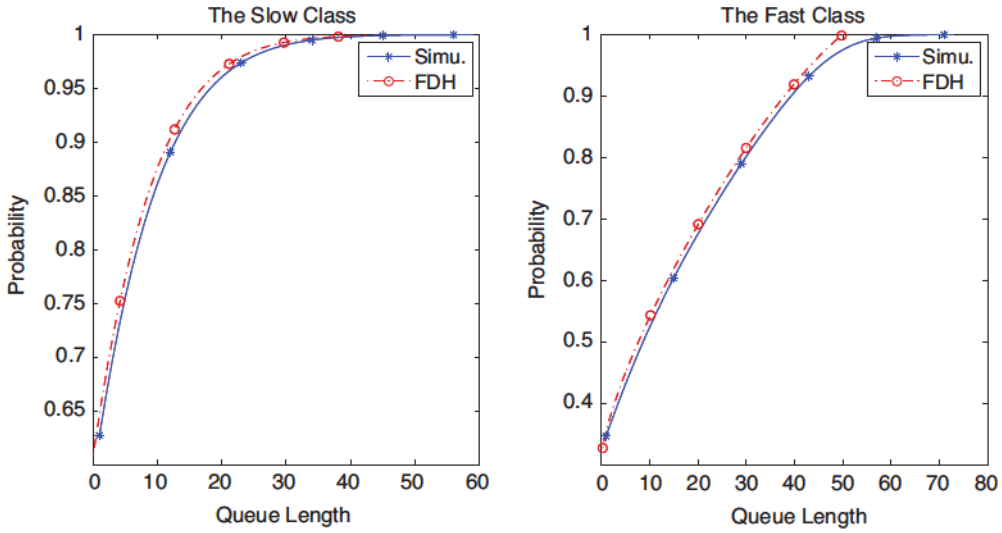
To demonstrate that the limiting distribution of the FDH approximates well the limiting distribution of the stochastic system (beyond the means), we compared the (marginal) limiting cdf's of the two simulated queues to the corresponding FDH distributions. The results are depicted in Figure 2.

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, the *N*-system design is especially useful in our setting, because 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)

Figure 2. (Color online) The Marginal cdf's Computed from Simulations of $Q_S^n(\infty)$ and $Q_F^n(\infty)$ (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)



and emails, may receive higher pay and be more costly to train, than agents that only respond to emails.

5.1. 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. In addition, fast customers are always routed to the fast track when both pools have idle servers. 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

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

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

Let $\tilde{Q}^{z,n}$ and $\tilde{I}^{z,n}$ 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* $\tilde{X}^{z,n}(0) \Rightarrow X^z(0)$ *in* $\mathbb{R}^4$ *and, in addition,* Assumptions 1 *and* 3 *hold, then* $(\tilde{X}^{z,n}(t), \tilde{Q}^{z,n}(t), \tilde{I}^n(t)) \Rightarrow (X^z(t), Q^z(t), I^z(t))$ *in* D^5 *as* $n \to \infty$, *where the component process of* X^z *is the unique solutions to the HSDE*

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

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

where B *is a standard Brownian motion,* $Q^z := ((X_S^z)^+, X_F^z)$, *and* I^z *is the unique nondecreasing 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, whereas (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, whereas 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 with the V-system, in which a nonnegligible portion of fast-class customers abandon the system, we conclude that a fast track can significantly increase the throughput of the system;

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.

5.2. 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, $(\tilde{X}^{z,n}, \tilde{Q}^{z,n})$, which we denote by $(\tilde{X}^{z,n}(\infty), \tilde{Q}^{z,n}(\infty))$, 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*

$$Q^z(\infty) := (Q_S^z(\infty), Q_F^z(\infty) = (X_S^z(\infty)^+, X_F^z(\infty)). \tag{12}$$

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

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

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

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

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

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, $Q_F^z(\infty)=0$ 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, 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 nontrivial 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, because 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_i^n denote the cost incurred per abandoning class-i customer, and h_i^n denote the rate at which holding costs are incurred in system n. Let d_R^n and d_F^n 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:

$$\sum_{i=S,F} \left(h_i^n E[Q_i^{z,n}(\infty)] + a_i^n \theta_i^n E[Q_i^{z,n}(\infty)]\right) + d_R^n(n-z^n) + d_F^n z^n.$$

Let $d^n := d_F^n - d_R^n$ and $c_i^n := h_i^n + \theta_i a_i^n$. Because 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. The interesting (nontrivial) 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. Because 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

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

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 prelimit stochastic system because 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 $\{z^n : n \ge 1\}$. To avoid having redundant service capacity in the fast track, which is clearly suboptimal, we assume that

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

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*

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

for any sequence $\{z^n : n \ge 1\}$ *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^*.

6.1.1. 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 $C(z) = C_q(z) + d r_F z$. The second term $d r_F z$ 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. Because 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 Section 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 Section 7.2, which demonstrates the robustness of the FDH approximation. Finally, in Section 7.3, we explain why the dynamics under the nonpreemptive 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

Because 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 $E[\tilde{Q}_i^n(\infty)]$ 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, because 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 $\Theta_P(n)$, whereas the stochastic fluctuations are $\Theta_P(\sqrt{n})$. For n small, the two orders are indistinguishable.)

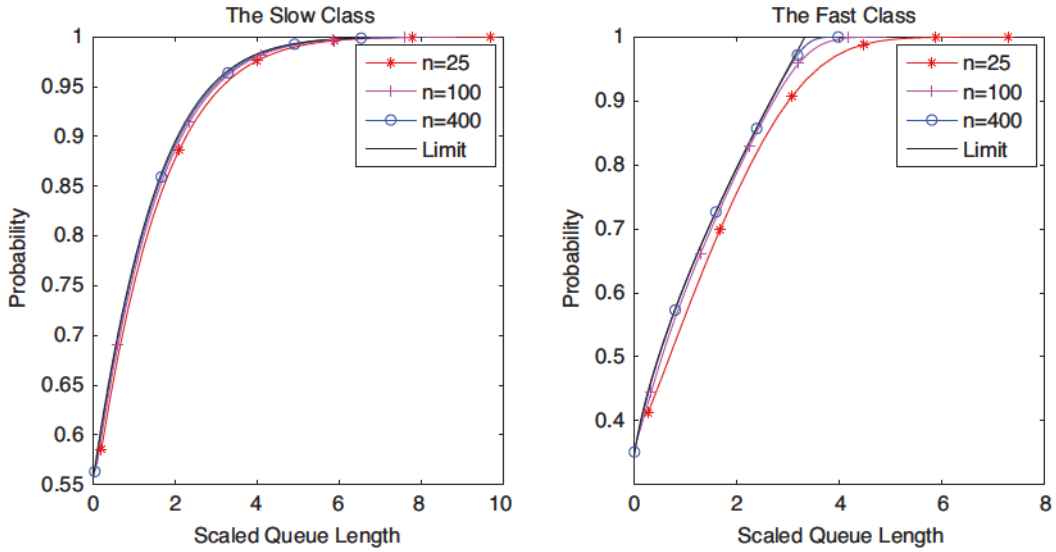
We used the simulation experiments to approximate the cdf's of the stationary distributions of the fast-class queue in the three systems with $\mu_F^n = 0.5\sqrt{n}$, and compare these cdf's to the corresponding cdf of the limiting distribution for the FDH approximation. The result, depicted in Figure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

Table 3. Comparison of the FDH Predictions to Simulation Results for Three Components of a Sequence of Systems

		Discrete-event simulation			
		$n = 25$	$n = 100$	$n = 400$	FDH
$\mu_F^n = \sqrt{n}$	$E[\tilde{Q}_S^n(\infty)]$	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)
	$E[\tilde{Q}_F^n(\infty)]$	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)
$\mu_F^n = 0.5\sqrt{n}$	$E[\tilde{Q}_S^n(\infty)]$	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)
	$E[\tilde{Q}_F^n(\infty)]$	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)

Note. Standard errors for the simulations are presented in parentheses.

Figure 3. (Color online) Empirical cdf's of $\tilde{Q}_S^n(\infty)$ (Left) and $\tilde{Q}_F^n(\infty)$ (Right) for $n \in \{25, 100, 400\}$, Plotted Together with the Empirical cdf of $Q_S(\infty)$ (Left) and $Q_F(\infty)$ (Right)



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, $\lambda_S^n/\mu_S^n + \lambda_F^n/\mu_F^n = n$. 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 $E[\tilde{Q}_F^n(\infty)]$ and $P(W_F^n > 0)$ have the largest errors when $\mu_F^n = 2$, whereas the error is significantly smaller for the larger two values of μ_F^n. (Note that the FDH approximation for $E[\tilde{Q}_S^n(\infty)]$ 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, that is, $\rho_S^n = 0.75$, 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.

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 $\tilde{Q}_F$. Further, $Z_F^n = o_P(n)$ under either policy (indeed, $\tilde{Q}_F = \tilde{X}_F$), showing that the processes corresponding to the fast class are indistinguishable under the two policies in the FDH limit.

Table 4. Sensitivity Analysis for the Accuracy of the FDH Approximation

		Discrete-event simulation			
		$\mu_F^n = 2$	$\mu_F^n = 5$	$\mu_F^n = 10$	FDH
$\rho_S^n = 0.75$	$E[\tilde{Q}_S^n(\infty)]$	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)
	$E[\tilde{Q}_F^n(\infty)]$	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)
$\rho_S^n = 0.85$	$E[\tilde{Q}_S^n(\infty)]$	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)
	$E[\tilde{Q}_F^n(\infty)]$	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)
$\rho_S^n = 0.95$	$E[\tilde{Q}_S^n(\infty)]$	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)
	$E[\tilde{Q}_F^n(\infty)]$	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)

Notes. Standard error of the simulation experiments are presented in parentheses. The expected queue lengths are scaled according to the FDH scaling.

The reasoning as to why the processes corresponding to the slow class under the nonpreemptive policy are unchanged asymptotically is more intricate, but again follows from the scaling of μ_F^n. Due to this scaling, the total output rate of fast customers from service is $\Theta_P(n)$ whenever $Z_F^n(t) = \Theta_P(\sqrt{n})$. 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 nonpreemptive 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 $Z_F^n(t) > 0$ and $Q_S^n(t) > 0$ for all $t \in [t_1^n, t_2^n]$, $0 \le t_1^n < t_2^n < \infty$, then Z_F^n behaves like a pure death process over this time interval, with death rates $k\mu_F^n = \Theta(k\sqrt{n})$, $k = 1, 2, \ldots$. It follows that for any $\epsilon > 0$, the sequence of events

$$B^n(\epsilon) := \{\{Z_F^n(t) > 0\} \cap \{Q_S^n(t) > 0\} : t \in [t_1^n, t_2^n],\ t_2^n - t_1^n > \epsilon\},$$

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_F^n and Q_S^n.) 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.

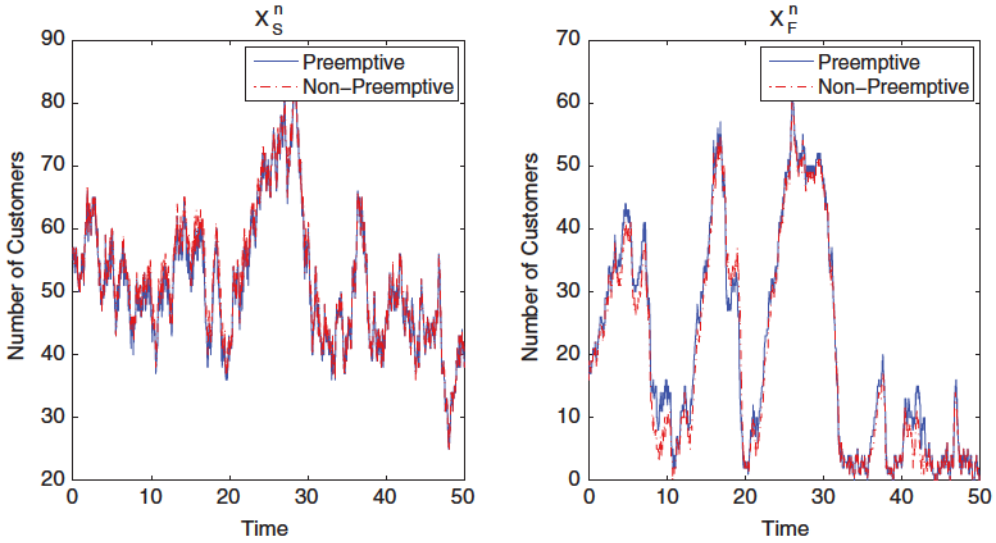
We do not attempt to rigorously prove the asymptotic equivalence between the policies. Instead, we demonstrate that the dynamics of the queues are similar under both policies via simulation. Figure 4 plots two sample paths for the system considered in Section 4.4, with $n = 50$, $\lambda_S^n = 46$, $\lambda_F^n = 15$, $\mu_S^n = 1$, $\mu_F^n = 5$, $\theta_S = 0.1$ 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 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 nonpreemptive 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 nonnegligible 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

Figure 4. (Color online) 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 Nonpreemptive Priority Policy



Note. The starred lines plot the sample paths under the preemptive policy, and the circled lines plot the sample paths under the nonpreemptive policy.

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

The authors thank the associate editor and two anonymous referees for making many useful comments and suggestions which led to the results in Section 6.1 and to Section EC.1.

Endnote

[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.

References

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

Asmussen S, Glynn PW (2007) *Stochastic Simulation: Algorithms and Analysis*, vol. 57 (Springer Science & Business Media, Berlin).

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

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

Atar R, Giat C, Shimkin N (2010) The $c\mu/\theta$ rule for many-server queues with abandonment. *Oper. Res.* 58(5):1427–1439.

Atar R, Gurvich I (2014) Scheduling parallel servers in the nondegenerate slowdown diffusion regime: Asymptotic optimality results. *Ann. Appl. Probab.* 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. *Ann. Appl. Probab.* 14(3):1084–1134.

Bassamboo A, Harrison JM, Zeevi A (2009) Pointwise stationary fluid models for stochastic processing networks. *Manufacturing Service Oper. 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. *Ann. Appl. Probab.* 11 (3):608–649.

Coffman E Jr, Puhalskii A, Reiman MI (1995) Polling systems with zero switchover times: A heavy-traffic averaging principle. *Ann. Appl. Probab.* 5(3):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. *Emerg. Med. J.* 19(1):28–30.

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

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

Garnett O, Mandelbaum A, Reiman MI (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. 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. *Math. Oper. Res.* 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, Rockville, MD).

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

Gurvich I, Lariviere MA, Ozkan C (2019) Coverage, coarseness, and classification: Determinants of social efficiency in priority queues. *Management Sci.* 65(3):1061–1075.

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

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

Harrison JM (1998) Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Ann. Appl. Probab.* 8(3):822–848.

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

Hinch EJ (1991) *Perturbation Methods* (Cambridge University Press, Cambridge, United Kingdom).

Hunt P, Kurtz T (1994) Large loss networks. *Stochastic Process. Appl.* 53(2):363–378.

Iglehart DL (1965) Limiting diffusion approximations for the many server queue and the repairman problem. *J. Appl. Probab.* 2(2): 429–441.

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

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

Maglaras C, Zeevi A (2004) Diffusion approximations for a multiclass Markovian service system with "guaranteed" and "best-effort" service levels. *Math. Oper. Res.* 29(4):786–813.

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

Moyal P, Perry O (2017) On the instability of matching queues. *Ann. Appl. Probab.* 27(6):3385–3434.

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

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

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

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

Perry O, Whitt W (2013) A fluid limit for an overloaded X model via a stochastic averaging principle. *Math. Oper. Res.* 38(2):294–349.

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

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

Sanchez M, Smally AJ, Grant RJ, Jacobs LM (2006) Effects of a fast-track area on emergency department performance. *J. Emerg. Med.* 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. Accessed October 8, 2021, 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. *Oper. Res.* 58(1):94–110.

Whitt W (1971) Weak convergence theorems for priority queues: preemptive-resume discipline. *J. Appl. Probab.* 8(1):74–94.

Whitt W (1983) Comparison conjectures about the $M/G/s$ queue. *Oper. Res. Lett.* 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 Sci.* 37(3):307–314.

Whitt W (1992) Understanding the efficiency of multi-server service systems. *Management Sci.* 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, Berlin).

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

Whitt W (2005) Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math. Oper. Res.* 30(1):1–27.

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

Yin GG, Zhang Q (2005) *Discrete-Time Markov Chains: Two-Time-Scale Methods and Applications*, vol. 55 (Springer Science & Business Media, Berlin).

Yin GG, Zhang Q (2012) *Continuous-Time Markov Chains and Applications: A Two-Time-Scale Approach*, vol. 37 (Springer Science & Business Media, Berlin).

Lun Yu is a post-doctoral research fellow in the Department of Industrial Engineering at Tsinghua University. His research focuses on developing asymptotic approximations to service and inventory systems that can be analyzed and optimized using applied probability, statistics, optimization, and machine learning.

Seyed Iravani is a professor of Industrial Engineering and Management Sciences at Northwestern University. His research focuses on the applications of stochastic processes, game theory, social networks, and queueing theory to the design and control of manufacturing, service operations systems, healthcare, supply chains, and non-profit systems focusing on improving their flexibility, coordination, and responsiveness. Professor Iravani has served on the editorial board of journals such as Operations Research, Management Science, Service Science, IIE Transactions, and Naval Research Logistics.

Ohad Perry is an associate professor in the Department of Industrial Engineering and Management Science at Northwestern University. His research focuses on applying methodologies from applied probability and dynamical-systems' control to approximate and analyze complex stochastic systems with applications to service and inventory systems.