Applications of Fluid Models in Service Operations Management

Noa Zychlinski
Faculty of Industrial Engineering and Management,
Technion – Israel Institute of Technology, Haifa 3200003, Israel
noazy@technion.ac.il

The service sector is an indisputable and fundamental pillar of today's business world, encompassing nearly 80 percent of the workforce in the United States. Service operations management has been a fertile research area addressing strategic, tactical and operational challenges related to service systems. Such problems usually have a complex, dynamic stochastic nature, often leading to models that are both analytically and computationally complicated. In such cases, fluid deterministic models that approximate the dynamics of stochastic/queueing systems can yield accurate and tractable optimization formulations. These formulations enable the construction of intuitive, insightful policies that are implementable in practice, even in time-varying systems. This paper focuses on the applicative aspects of fluid models in addressing various problems in service and healthcare operations management. We review the literature on fluid model applications, discuss the situations in which fluid models are less adequate as well as the implementation of a fluid-based policy into a stochastic discrete system. Lastly, we identify future research opportunities and challenges that have yet to be addressed.

Key words: queueing systems, service/healthcare operations management, assymptotic analysis, Functional Strong Law of Large Numbers

1. Introduction

Service systems constitute an integral part of the world's economy and people's daily lives. Such systems are usually characterized by complex, stochastic, dynamic networks that involve multiple server pools and customer classes. Addressing operational management problems can, therefore, be highly complicated and lead to intractable models. In such cases, fluid frameworks can yield accurate manageable approximations usable for optimizing system performance, allowing us to glean structural operational insights and develop easy-to-implement policies. In fluid models, entities that go through the system are animated as continuous deterministic fluid. These approximations allow the system's dynamics to be captured by a set of differential equations that are intuitive and much more analysis-convenient than their stochastic counterparts.

Additionally, many-server time-varying fluid models, which approximate under, over and critically loaded stochastic systems, provide an excellent fit to the average transient behavior of time-dependent stochastic systems. There is ample literature justifying the assertion that fluid models

accurately approximate their underlying stochastic systems under the "right" operational-limiting regime [96, 97, 120, 133, 83, 135, 110, 89, 92]. In this paper, however, we focus on the applications and potential of fluid models in addressing operational management problems in service systems. That is, we look at how fluid models can be used, in practice, to model, control and optimize the performance of such systems.

Since fluid deterministic models allow an accurate, simple yet realistic perception of stochastic discrete systems, they have been successfully used to model different types of service systems. These models cover the early applications for post offices and social security offices [108, 130], transportation and ride-sharing services [128, 30, 109], call centers [56, 7, 1] and healthcare systems [139, 54, 13, 73].

The very basic deterministic fluid model (Figure 1) refers to the system as a black box having a single input and a single output: an arrival rate function $\lambda(t)$ and a departure rate function $\delta(t)$, $t \geq 0$.

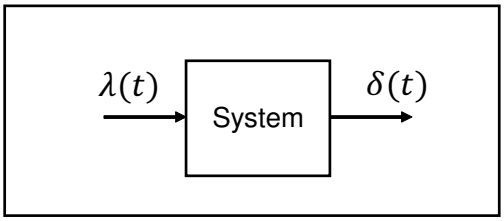

Figure 1 **A black box illustration with a single input and a single output.**

In this model, the stochastic arrival and departure processes are approximated by continuous deterministic functions in which customers are modeled as fluid flow. Let $q(t)$ denote the fluid content in the system at time t. The rate of change in the number of customers in the system at time t is given by the following differential equation:

$$\dot{q}(t) = \lambda(t) - \delta(t)$$

or equivalently,

$$q(t) = q(0) + \int_0^t [\lambda(s) - \delta(s)]\mathrm{d}s, \quad t \geq 0,$$

where $q(0)$ is the initial number of customers in the system at $t = 0$.

As we demonstrate below, the departure rate $\delta(t)$ is constructed based on the system's characteristics: its configuration (e.g., a single station or a complex network of stations), the service distributions, the number of servers in each station/pool, and whether the system is time-varying or

stationary. Moreover, the explicit formulation for q needs to assure that $q(t) \geq 0$, for every $t \geq 0$. In equilibrium, the number of customers in the fluid model remains constant and, therefore, $\dot{q}(t) = 0$.

Note that while most fluid models are continuous in time, there are several papers that consider discrete-time fluid models (e.g., [135, 142])

The remainder of the paper is organized as follows. We conclude this section by briefly overviewing some early applications of fluid models. In Section 2, we present the two main viewpoints of fluid applications: a tool for stability establishment and models in their own right. Thereafter, we discuss the two most common asymptotic regimes studied in the literature: conventional and many-server heavy-traffic. In Section 3, we demonstrate the construction of time-varying fluid models. Sections 4–8 review the literature on applications of fluid models in the area of service operations and management. We chose to organize these sections according to their *main* motivating application; specifically, call center operations: staffing scheduling and routing problems, healthcare operations management, matching queues, organ transplantation, ride-sharing systems, delay-informed queue access, customer acquisition and retention. In Section 9, we discuss the cases where fluid models might not be effective and ways to translate a fluid-based policy back into the original stochastic system. Lastly, in Section 10, we provide concluding comments and suggest opportunities for future research.

1.1. Early Applications of Fluid Models

One of the first applications of the fluid approach was suggested by Oliver and Samuel [108] for analyzing mail sorting procedures in post offices in order to reduce delivery delays. The optimal processing rates at each station were found by optimizing a deterministic model of network flow. The suggested solution was implemented in a large post office in the United States and reduced letter delays by 25%.

Vandergraft [130] utilized a fluid approach for analyzing time-varying networks in a district office of the social security administration. Staffing levels at each station need to match the variability of daily claim arrival rates. The network flow was modeled by a set of differential equations describing the rate of change in the number of customers at each station at any time. The equation set was then solved numerically, and performance measurements (such as productivity, resource utilization and waiting time) were evaluated.

Another early application of fluid models was suggested by Porteus [116, 117, 118] for analyzing the operations of the National Cranberry Cooperative. The main two problems that required improvement were extremely long queues of trucks waiting to unload the fruit they carried and high overtime costs due to the heavy burden of work. Using deterministic flow analysis of the

process network (berries flow continuously and at a constant rate throughout the process) together with inventory build-up models, the bottlenecks were identified and improvement alternatives were suggested (e.g., seasonal starting times and staffing levels). Although the deterministic fluid model neglects the stochasticity of the system, the analysis and suggestions were shown to improve the system performance significantly.

Several early applications of fluid models were offered in the context of airport terminal operations, such as analyzing arrival immigration control, check-in counters, and departure lounge capacity [126, 111, 88, 128, 48]. Horonjeff et al. [72] developed a fluid deterministic model for analyzing baggage claim facilities in terminals. They considered two flows – passengers and baggage – arriving at the service facility. The two flows are merged into a single departure flow after each passenger finds and takes their bags. For earlier fluid applications we refer readers to Mandelbaum and Zeltyn [101] who surveyed a service engineering course, taught at the Technion–Israel Institute of Technology. Influenced by Hall [64], the course teaches fluid models before stochastic models and by doing so emphasizes science-based applications.

2. Ways of Viewing Fluid Applications

One basic requirement from an approximating model is that it accurately describe the dynamics and performance of the system it approximates. There is a broad stream of research that includes the development of fluid models for different settings in order to evaluate and optimize performance. In that context, there are two ways of understanding fluid applications. The first [43] sees fluid models as a tool for establishing stability of their underlying stochastic systems. The second viewpoint [97, 91] states that fluid models are stand-alone, often first-order models that provide most of the insight needed. In what follows, we elaborate on each of these viewpoints.

Dai [42] proved that a Stochastic Processing Networks (SPN) is stable (i.e., its ambient Markov chain is positive recurrent), if the fluid limit of the SPN is stable; that is, the corresponding fluid limit model reaches zero and stays there regardless of the system's initial state. The importance of this result lies in its application simplicity when compared to working directly within/on the stochastic model. Practical usage of this result and more applications are discussed in Dai and Harrison [43].

As fluid models have become more common in operations research literature, more and more papers do not rigorously develop the fluid models as limits for corresponding stochastic systems, but rather posit the fluid model as is. As Liu and Whitt [91] pointed out: "It is important to recognize that the fluid model can be considered directly as a legitimate model in its own right". For example, Hall [65] used a fluid approach to model and analyze patient flow through a healthcare system so as to reduce delays and improve healthcare delivery.

Newell [105] and Hall [65], who established fluid approximating models for service systems, based their work on the analogy to transportation systems. They distinguished between two types of queues/variations: deterministic (predictable) and stochastic. The former can describe known variations throughout the day or day of the week, while the latter is caused by random variations around averages. The fluid model is a deterministic one and, thus, captures the first type of queues/variations (i.e., the first-order mean of the system's dynamics). Neglecting the second type is justified when the variations around the averages are relatively small compared to the averages. For instance, when the arrivals follow a Poisson process with rate $\Lambda(t)$, the coefficient of variance, $1/\sqrt{\Lambda(t)}$, decreases as $\Lambda(t)$ increases. In other words, the variations relative to the mean become smaller. This phenomenon allows fluid deterministic models to be accurate approximations of the system's average behavior. Refining the fluid models to capture the stochastic variations around the average is achievable through diffusion models.

2.1. Limiting Regimes

The above two ways of understanding fluid applications are also related to the different limiting regimes under which fluid models are derived. Indeed, fluid models are limits of corresponding stochastic systems, which are established by considering a sequence of properly scaled stochastic systems and using the Functional Law of Large Numbers. Two main asymptotic regimes are studied in the literature: the conventional heavy traffic regime and the many-server heavy traffic regime.

Conventional heavy traffic. Under this regime, the time is scaled up (which is equivalent to scaling up the arrival and service rates), while the number of servers is held fixed [132, Section 5].

Under this scaling regime, service times are instantaneous and almost all arriving customers have to wait for service. This brings up questions of how to prioritize customers when servers free up [67, 123, 100]. Moreover, since service times are negligible under this scaling, such fluid models can support general service-time distributions. In other words, the time in the system is the time in the queue.

Consider a single-server queueing system where the arrival and service processes are renewal processes with finite mean $1/\lambda$ and $1/\mu$, respectively. The arrival process $\{A(t),\, t \geq 0\}$ and the service process $\{S(t),\, t \geq 0\}$ are scaled so that

$$\bar{A}^n(t) = A(nt)/n \to \lambda t, \quad \bar{S}^n(t) = S(nt)/n \to \mu t \quad \text{u.o.c, as } n \to \infty.$$

Let $Q(t)$ denote the number of customers in the stochastic system at time t. Space is scaled down by considering the fluid-scaled process $\bar{Q}^n(t) := Q^n(nt)/n$. Then, $\bar{Q}^n(t) \to q(t)$ as $n \to \infty$ where

$q(t),\ t \geq 0$, denotes the total fluid content in the system at time t. Given the initial condition, $q(0)$, we have (see Section 6.3 in Chen and Yao [39, Section 6]):

$$q(t) = [q(0) + (\lambda - \mu)t]^+ .$$

That is, if $\lambda < \mu$, the fluid content will eventually reach zero and stay there. If, however, $\lambda > \mu$, the fluid content will grow at rate $\lambda - \mu$. If $\lambda = \mu$, the fluid content will stay constant at the initial level.

Other fluid models were developed under conventional heavy traffic. For example, Mandelbaum and Massey [95] analyzed the $M_t/M_t/1$ queue with time-varying arrival and service processes. The key insight in this paper is that the fluid model determines the operational regime: Positive implies over-loaded; zero requires information from the reflection-term in order to decide if the system is underloaded or critically loaded. The main challenge arises from the need to analyze regimes that are changing in time. The time-varying traffic intensity is a way of decoding which regime prevails. Whitt [138] developed heavy-traffic limits under conventional scaling for the $G_t/GI/N$ queue with a periodic arrival process. Fluid models under conventional heavy traffic were also developed for queueing networks [132, 39]. In general, due to the type of scaling, conventional heavy traffic approximations are less effective in underloaded situations [8]. For further technical background on stochastic-process limits and fluid approximations, we refer readers to Whitt [132, Section 5] and Chen and Yao [39, Section 6].

Many-server heavy-traffic regime. Under this regime, the arrival rates and the number of servers are scaled up to infinity, while the service rates are held fixed. As such, the many-server approximation is adequate for large systems. Such approximations might also work well for single-digit staffing [29, 139] because of the fast rate of convergence [81]. The many-server heavy-traffic regime is useful for analyzing transient and time-varying systems. Moreover, as opposed to conventional heavy traffic, service rates are not scaled, hence service times and their distribution are significant.

Consider the $M_t/M/N$ system having Poisson non-homogeneous arrivals with rate $\lambda(t)$, exponential service times with average rate μ and N servers. Next, we introduce a sequence of stochastic systems indexed by $\eta > 0$, $\eta \to \infty$. In the η-th system, the number of servers is scaled to ηN, and the arrival rate is scaled: $A^\eta = \{\eta A(t),\ t \geq 0\}$, so that

$$\{A^\eta(t)/\eta,\, t \geq 0\} \to \left\{ \int_0^t \lambda(u)du,\, t \geq 0 \right\} \quad u.o.c.\ as\ \ \eta \to \infty.$$

The space is scaled down by considering the fluid-scaled process $Q^\eta(t) := Q^\eta(t)/\eta$. Then, $Q^\eta(\cdot) \to q(\cdot)$ as $\eta \to \infty$, where $q(\cdot)$ is the unique solution to the following differential equation:

$$q(t) = q(0) + \int_0^t [\lambda(s) - \mu\,(q(s) \wedge N)]\mathrm{d}s,$$

where $q(t) \wedge N = \min(q(t), N)$ denotes the number of occupied servers at time t.

Mandelbaum et al. [96, 97] established the foundations of fluid approximations for modeling time-varying queueing systems with abandonment and retrials, where inter-arrival times are exponentially distributed, hence the number of arrivals in an interval is Poisson; service, abandonment, and retry rates are exponential. These works allow the analysis of time-varying systems within specific regimes or operations, rather than forcing researchers to use simulation or piecewise constant stationary analysis. Different variations and extensions were developed under the many-server heavy-traffic regime: general distributions and abandonments [134, 135] and systems with multiple customer classes and multiple service pools [124]. More general many-server heavy-traffic fluid models were developed by [89, 90, 91, 60, 139, 92, 143, 54]. Within the framework of Resource-Driven Activity Networks (RANs), Carmeli et al. [31] developed a fluid model for closed networks in the many-server heavy-traffic regime.

Fluid models for long service times. The analysis of fluid models for systems with long service times must be distinguished from the analysis of systems with short service times. When service times are short, departures from the queue and departures from the system are very close and can be considered the same. When service times are long, however, there is a delay between these two departures. Specifically, let $\bar{A}(t)$ denote the cumulative arrivals to the system in the deterministic fluid approximation. Additionally, let $\bar{D}_q(t), \bar{D}_s(t)$ denote the cumulative departures from the queue and from the system, respectively. In the deterministic fluid approximation, we have $\bar{D}_s^{-1}(n) = \bar{D}_q^{-1}(n) + 1/\mu$. In systems with long service times compared to waiting times, this delay is not negligible. Thus, Hall [64] suggested the following construction of both deterministic accumulated departures:

$$\bar{D}_s(t + 1/\mu) = \bar{D}_q(t); \qquad \bar{D}_q(t) = \min\left(\bar{A}(t), \bar{D}_s(t) + N\right).$$

The first condition describes the delay caused by the service time: A customer who departs the queue and begins service at time t will depart the system at time $t + 1/\mu$. The second condition states that as long as there is no queue, the departure rate from the queue equals the arrival rate, since service starts immediately upon arrival. When the queue begins to form, all N servers work and the total number of customers that departed from the queue equals the total number that departed from the system plus the N customers in service. The vertical distance at time t between the cumulative arrivals and the cumulative departures from the queue (from the system) represents the fluid content in the queue (system) at time t. Hall's suggested method applies also

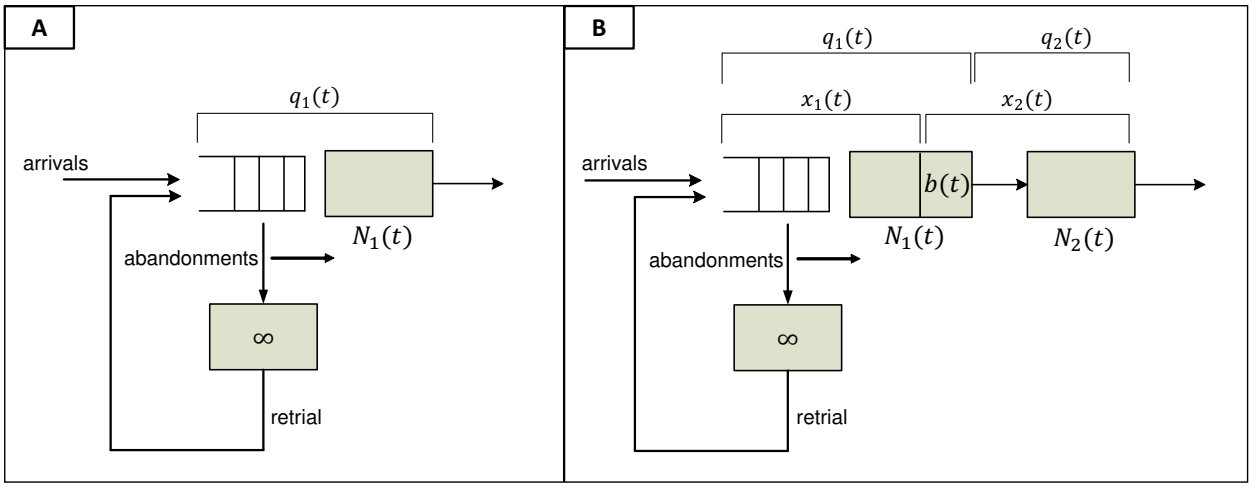

Figure 2 **Two models illustrating abandonments, retrials, and blocking.**

to many-server systems where service times do not become negligible in the limit. Long general service times in many-server queues were addressed by Carmeli et al. [31] under a transient finite horizon.

3. Constructing Time-Varying Fluid Models for Queueing Systems

The fluid model for a queueing network is characterized by a set of differential equations (one for each station) that describes the rate of change in the number of customers in each station at any time. These models can be built directly from systems/data (rather than as approximations). We demonstrate the construction of these equations using the two examples illustrated in Figure 2. We start with Model A, which was developed in Mandelbaum et al. [97] for a time-varying multiple-server queue with memoryless service times, abandonments and retrials. That is, some customers waiting in the queue give up, i.e., abandon; some retry later if the service is important to them. Specifically, we have service Station 1, and a retrial pool with infinite capacity for all customers who abandon the queue and may potentially retry. The model is characterized by the following (deterministic) parameters:

1. Exogenous arrival rate $\lambda(t)$, $t \geq 0$, to Station 1.
2. Service rate $\mu_1(t) > 0$ at Station 1 at time t.
3. Retry rate $\mu_\infty(t) > 0$ in the retrial pool at time t.
4. Number of servers $N_1(t)$ in Station 1 at time t.
5. Abandonment rate $\theta(t)$ in the waiting room before Station 1 at time t.
6. Probability of no retry $\psi(t)$ at time t.

Let the functions q_1, q_∞ denote the number of customers in Station 1 and in the retrial pool, respectively, according to the fluid model. Given initial conditions, $q_1(0), q_\infty(0)$, these functions are the unique solution of the following non-linear differential equations:

$$\begin{aligned}\dot{q}_1(t) &= \lambda(t) + \mu_\infty(t)q_\infty(t) - \mu_1(t)(q_1(t) \wedge N_1(t)) - \theta(t)\left(q_1(t) - N_1(t)\right)^+; \\ \dot{q}_\infty(t) &= \theta(t)(1-\psi(t))(q_1(t) - N_1(t))^+ - \mu_\infty(t)q_\infty(t),\end{aligned} \tag{1}$$

where $(x \wedge y) = \min(x, y)$ and $x^+ = \max(x, 0)$. Each equation is basically the total input minus the total output from each station at time t.

The development in [97] is based on the many-server heavy-traffic regime where the arrivals and the number of servers are scaled up to infinity, while the distribution of service duration (or service rate) is held fixed.

Next, we move to Model B in Figure 2, which presents a more complicated version of Model A with another service station, Station 2, located after Station 1. Station 2 has N_2 servers; if a customer completes service in Station 1 and there are no available servers in Station 2, the customer is blocked while occupying a server in Station 1. Following [143], we define $x_1(t)$ as the number of customers that arrived at Station 1 up to time t and have not completed their service there; $b(t)$ is the number of blocked customers in Station 1 at time t. Lastly, $x_2(t)$ is the number of customers that have completed service in Station 1 but have not completed service in Station 2 up to time t. We thus have

$$b(t) = \left(x_2(t) - N_2(t)\right)^+.$$

Given initial conditions, $x_1(0), x_2(0)$, the functions x_1, x_2, b are the unique solution of:

$$\begin{aligned}\dot{x}_1(t) &= \lambda(t) + \mu_\infty(t)q_\infty(t) - \mu_1(t)\left(x_1(t) \wedge (N_1(t) - b(t))\right) \\ &\quad - \theta(t)\left(x_1(t) + b(t) - N_1(t)\right)^+; \\ \dot{x}_2(t) &= \mu_1(t)\left(x_1(t) \wedge (N_1(t) - b(t))\right) - \mu_2(t)(x_2(t) \wedge N_2(t)); \\ \dot{q}_\infty(t) &= \theta(t)(1-\psi(t))\left(x_1(t) + b(t) - N_1(t)\right)^+ - \mu_\infty(t)q_\infty(t).\end{aligned} \tag{2}$$

Here, $(x_1(t) \wedge (N_1(t) - b(t)))$ is the number of customers in service in Station 1 at time t (i.e., the total number of customers occupying a server in Station 1 excluding the number of blocked customers).

The total number of customers in each station at any time t is given by

$$\begin{aligned}q_1(t) &= x_1(t) + b(t); \\ q_2(t) &= x_2(t) \wedge N_2(t).\end{aligned} \tag{3}$$

Note how low the barriers are for applications. The equations can easily be solved numerically even in a spreadsheet via discretization of time (or using any high-level software such as MATLAB of Python).

Next we use simulation to evaluate the performance of the fluid model. We plot how the average number of customers in each system evolves over time for systems that start empty. This provides a copious amount of details about the system dynamics. Figure 3 compares the total number of customers in each station over time according to the fluid model (Equations (2)–(3)) and according to a simulation model of the corresponding stochastic system. In the latter, the arrivals were samples from a non-homogeneous Poisson process; service times were samples from exponential distributions. We consider two system sizes. In the first (top plots), $N_1 = 30$, $N_2 = 50$ and the arrival rate is $\lambda(t) = 0.3t$, $0 \leq t \leq 80$; the second system (middle plots) is half the size of the first (i.e., $N_1 = 30$, $N_2 = 50$ and the arrival rate is $\lambda(t) = 0.15t$, $0 \leq t \leq 80$); the third system (bottom plots) has $N_1 = 8$, $N_2 = 17$ and the arrival rate is $\lambda(t) = 0.08t$, $0 \leq t \leq 80$. The simulation results are averages that were estimated based on 10 (left plots) and 150 (right plots) independent replications.

In sum, the fluid model accurately describes the average flow of customers in the stochastic system. As expected, the fluid approximation becomes more accurate as the number of replications increases, and as the system grows in size. Nonetheless, even when the system is relatively small ($N_1 = 8,\ N_2 = 13$), the fluid model describes the average behavior of the system well.

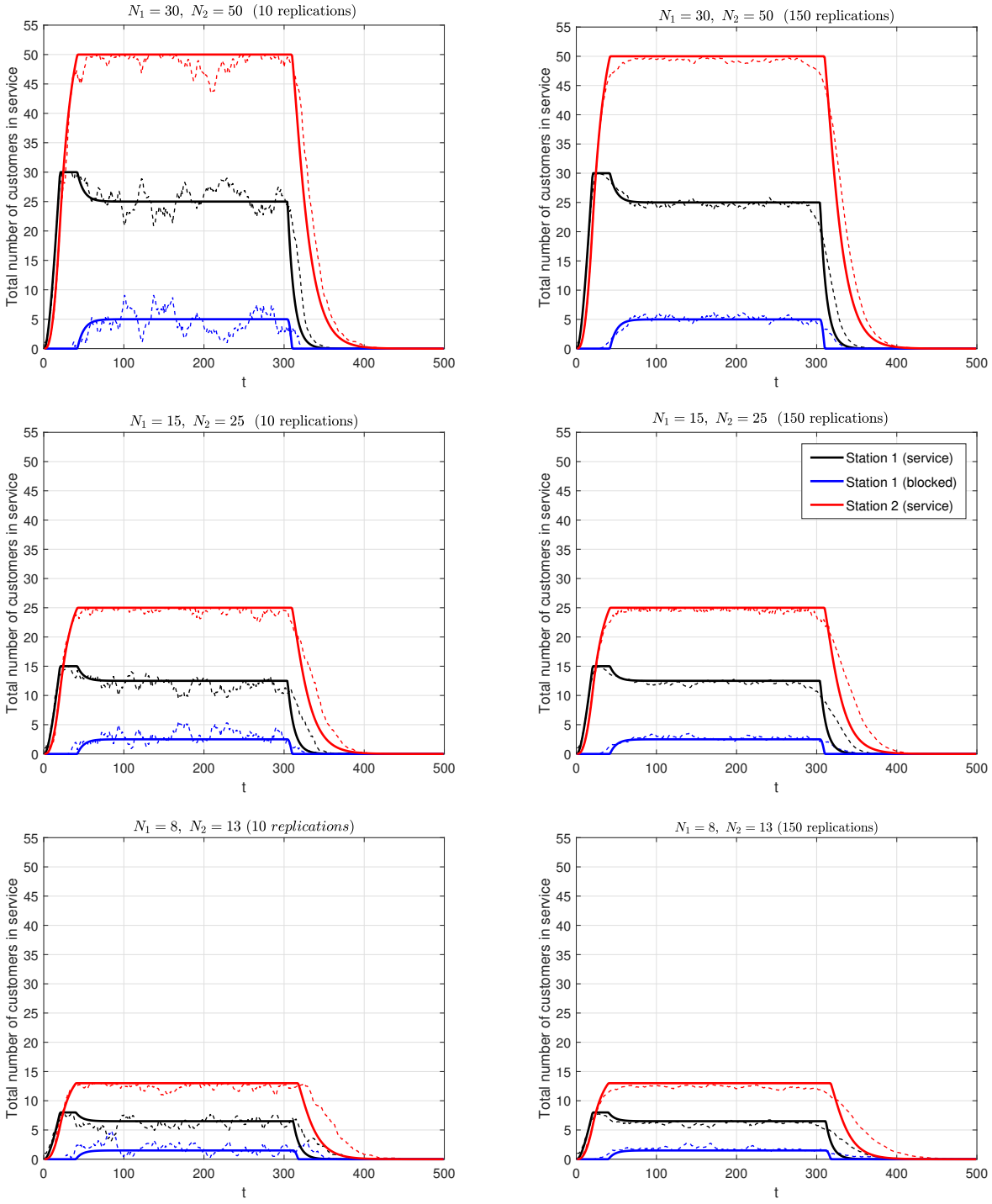

Figure 3 **Total number of customers in each station – fluid model vs. simulation. The simulation results are averages that were estimated based on 10 (left plots) and 150 (right plots) independent replications. The parameters are $\mu_1 = 1/8$, $\mu_2 = 1/16$, in the top plots, $\lambda(t) = 0.3t$, $0 \le t \le 80$, In the middle plots, $\lambda(t) = 0.15t$, $0 \le t \le 80$ and in the bottom plots, $\lambda(t) = 0.8t$, $0 \le t \le 80$.**

Fluid Model Applications

In the following sections we review the main applications of fluid models that are used to address various problems in service operations management. We chose to organize these problems according to their *main* motivating application, though we note that some problems are relevant to more than one application. We start with applications that are motivated by call center operations, which stimulated the development of fluid models; specifically, staffing, scheduling and routing problems (Section 4) – all of which constitute a major part of the service operations literature. Then, we move on to applications in healthcare operations managements (Section 5). Section 6 covers applications related to matching queues and, specifically, organ transplant and ridesharing systems. Then, we discuss fluid-based models of delay-informed queue access (Section 7). Lastly, we review the applications related to customer acquisition and retention (Section 8).

Remark 1 *Various applications of diffusion models in service operations management exist in the literature. For example, there are staffing [63, 131, 57, 137, 21], scheduling [129, 100, 68, 74, 93], routing [9, 62, 12, 18, 99] as well as joint optimization [16, 61, 127, 23, 58, 10] problems. To remain focused on fluid model applications, we do not cover these valuable works and their significant contribution to the field.*

4. Call Center Operations: Staffing, Scheduling and Routing Problems

Call centers have been a fertile ground for research in the operations management area [56, 7]. Most operational challenges faced by call centers boil down to staffing, scheduling and routing problems – all tightly intertwined. These dependencies impose additional challenges beyond those that arise when addressing each problem separately. Exact analysis of such stochastic settings is analytically/computationally complex and often intractable. The goal then becomes to find good, structurally simple solutions that are both near optimal and asymptotically optimal as the system's scale increases. Fluid approximations are desirable in that sense because they often enable the formulation of tractable optimization problems [136]. In what follows, we review the relevant approaches for each problem.

4.1. Staffing and Capacity Planning

Setting staffing requirements is essential when designing and managing service systems. The staffing decision specifies the number of agents needed during each staffing interval over the day. In many realistic cases (e.g., many-server queues, nonstationary and/or general distributions and abandonments), deriving insightful practical staffing decisions is possible only through heavy-traffic approximations. One common approach for setting staffing requirements is based on the *offered*

load approximation (i.e., an infinite-server queue). The offered-load represents the average amount of work being processed by resources, under the assumption of an infinite number of resources. The known square-root safety staffing formula is obtained by setting $N = R + \beta\sqrt{R}$, where N is the required number of servers and β is a parameter reflecting the quality of service in terms of delay and congestion [131]. The applicability of fluid models in setting staffing requirements under an uncertain arrival rate and agent absenteeism was discussed in [135, 137].

The above staffing problems focus on a single homogeneous server pool. In multi-skill settings, however, servers have different skills, and customers have different service requirements. Harrison and Zeevi [69] used stochastic fluid models to study a staffing problem of different agent pools in a setting with different customer classes, each with an instantaneous arrival rate that is time dependent and varies stochastically. By considering the trade-off between staffing costs and abandonment penalties, and using stochastic fluid models, the staffing problem is reduced to a multidimensional newsvendor problem. Following this work, Bassamboo, Randhawa and Zeevi [25] studied a capacity sizing problem in a multiple-server service system where customers may renege while waiting for service. The paper focused on the impact of uncertainty (caused by arrival rate prediction) on capacity planning. The authors identified two regimes, each requiring a different solution approach. The first is an uncertainty-dominated regime, where uncertainty effects dominate stochastic fluctuations. In this case, the stochastic fluctuations can be ignored and a suitable newsvendor problem (such as in Harrison and Zeevi [69]) can be used. The second regime the authors identified is a variability-dominated one, where stochastic fluctuations dominate uncertainty effects. In this case, the square-root approach that incorporates a capacity safety buffer is more adequate. The "right" regime is decided upon according to a threshold policy for the coefficient of variation.

Other versions of staffing problems were addressed using fluid models. Aguir et al. [4, 3] used a fluid model to study the effect of retrials on the performance of call centers. Using numerical analysis they demonstrated how misleading it is to consider retrials as first-time calls, and how this practice affects forecasting and staffing decisions. Ren and Zhou [119] used a fluid approximation model to study the operations in call centers that are outsourced to other companies. Staffing levels and exertion effort are determined in order to improve service quality (i.e., number of calls that are served and resolved). Gurvich, Luedtke and Tezcan [59] studied the staffing of call centers with uncertain demand forecasts. Yom-Tov and Mandelbaum [139] addressed staffing questions for the Erlang-R model with re-entrant customers.

Staffing flexible (contractors) and/or fixed (full-time) agents in service systems was studied by Dong and Ibrahim [52]. Using fluid and stochastic fluid models, the optimal staffing policy to minimize operating costs under varying customer demand patterns was derived.

In a long-term planning horizon, managers need to set the system capacity. This usually involves hardware decisions, for example, setting the required number of agent positions in a call center. Such decisions involve demand forecasting and trade-off considering the proper balance between operating costs and service quality (e.g., waiting times). Jennings, Massey and McCalla [82] used fluid approximations to optimize the number of leased private lines in a telecommunication setting to maximize profit. Because of very long service times (measured in years), their analysis is transient; indeed, the system does not reach steady state within the observation period. Bassamboo and Randhawa [24] studied the accuracy of fluid models for capacity sizing of queueing systems where customers abandon according to a general patience distribution. They derived prescriptions that are asymptotically optimal for large customer arrival rates, and showed that as the customer arrival rate increases, the optimality gap of the prescription remains bounded.

4.2. Scheduling

A scheduling problem determines how to assign waiting customers to a server/server pool upon completion of a service. The $c\mu$ rule (Cox and Smith [41]) for scheduling a multiclass single server queue when holding costs are linear was shown to be optimal.

Atar, Giat and Shimkin [17] derived the asymptotic optimality of an index-based policy, known as the $c\mu/\theta$ rule, for many-server queues with abandonment in the overloaded regime. Zychlinski, Chan and Dong [142] studied the scheduling of queues with different resource requirements. They derived the idle-avoid $c\mu/m$ rule, where m is the number of servers each class requires. For general multi-class multiserver queues, where exact analysis becomes prohibitively tedious, the asymptotic optimality of the rule is established in the many-server regime.

Dobson, Tezcan and Tilson [50] used a fluid model to study the scheduling of new customers and customers that are already in the system. The service process includes three steps conducted by two resources where service might be interrupted by other customers in the system. They show how the interruptions effect the optimal prioritization and the system's throughput. Dong and Ibrahim [53] studied the shortest-remaining-processing-time (SRPT) scheduling policy in multiserver queues with abandonment. They showed that from among all scheduling disciplines, the SRPT discipline maximizes, asymptotically, the system throughput.

4.3. Routing

A routing problem in a multi-skill setting is also known as skills-based routing [56]. Since exact analysis of stochastic routing problems is challenging and complex, fluid approximation models are a useful tool to derive insightful implementable policies. Motivated by tenant assignment models in public housing, Talreja and Whitt [124] used fluid models to determine the stationary routing rates between customer classes and service pools in overloaded queueing systems. The routing flow rates

are derived for different network routing graphs. By analyzing a fluid approximation model, Perry and Whitt [114, 115] proposed a fixed-queue-ratio (FQR) routing policy for a two-class and two-server pool (X-model) to address unexpected overload. The proposed FQR policy has thresholds that automatically detect class overload.

Mandelbaum and Momčilović [98] used a fluid model to study personalized queues where information at the level of individual customers/servers affects system dynamics. To understand the benefits from personalized customer information, the authors compared the personalized least-patience first (LPF) routing policy with FCFS. They found that LPF is significantly better when the overloaded durations are comparable to (im)patience times.

Routing and Staffing. Staffing is closely related to routing when agents have different skills. Bassamboo, Harrison and Zeevi [23] addressed both problems in a call center model with multiple customer classes and multiple server pools. The model has instantaneous arrival rates that vary both temporally and stochastically. To minimize the sum of staffing costs and abandonment penalties, a new asymptotic parameter regime is developed. In this regime, service rates and abandonment rates are accelerated in a linear manner, while arrival rates grow super-linearly. The key feature of this two-scale parameter regime is that the dynamic control problem becomes tractable. As a solution, the authors suggested a linear-programming based method for staffing and routing, which is asymptotically optimal.

5. Healthcare Operations Management

Healthcare systems tend to have complex dynamic processes, which are difficult to analyze directly. In such cases, fluid models can help capture and analyze decision makers' key trade-offs. Fluid models have been used to analytically derive operational insights regarding system dynamics and performance. Such insights are usually unattainable via exact analysis of the underlying stochastic systems.

Mills, Argon and Ziya [103] used a fluid model for setting triage priorities among patients in a mass casualty event (MSE), by considering resource limitations and time-dependent survival probabilities. In the same context, Cohen, Mandelbaum and Zychlinski [40] used a fluid model to address resource allocation problems in hospitals between different treatment stations during MSEs. Fluid models have been used to analyze systems with slowdowns [51] and speedups [35] that can occur in hospitals as a result of high congestion.

Dai and Shi [45] used fluid control combined with single-pool approximation to study the overflow phenomenon in hospitals that occurs when, due to excessive waiting times, patients have the option of being assigned to a non-primary bed. A queueing model for inpatient wards including

the need for a physician's approval to discharge patients was studied in Dong and Perry [54] using a nonstationary fluid model. Motivated by the bed blocking problem within hospital networks, Zychlinski et al. [144, 143, 145] developed time-varying fluid models for networks with blocking. In [144] a fluid-based approach was used to determine the optimal number of beds in a long-term care facility by incorporating blocking costs incurred when there are not enough beds. Chan, Huang and Sarhangian [33] used a fluid model to study the reassignment of ED nurses to different services at the beginning of shifts. By analyzing a fluid control problem, the authors minimized transient holding costs over a finite horizon, and showed that an appropriate "translation" of the solution to the fluid control problem is asymptotically optimal for the original stochastic system. Recently, Chan et al. [34] extended the model to incorporate two different phases of care: treatment and boarding, and two types of nursing staff.

Hu, Chan and Dong [73] used fluid approximation to study the scheduling of proactive services when less critical patients might deteriorate if their treatment is delayed. In their analysis, they distinguished between long-run average performance and the transient performance. A fluid model was used by Armony and Yom-Tov [13] to study the trade-off between infection and mortality risks when deciding when to discharge hematology patients from hospital.

Yom-Tov and Mandelabaum [139] used a fluid approximation to study the Erlang-R model, motivated by healthcare systems, in which patients go through a repetitive service process. A fluid approximation was used by Chan et al. [36] to study service systems with returns in the context of hospital readmission prevention programs. In their system, there is a cost associated with the return probability of departing customers. The decision maker can determine this probability based on the system's congestion. Zychlinski [141] used fluid approximation to study the scheduling of a hybrid healthcare system that provides three service channels: in-person, virtual and supplementary in-person service for virtual patients that require a follow-up in-person visit.

6. Matching Queues: Organ Transplant and Ride-Sharing Systems

Matching problems arise in different applications, such as organ transplants, ride-sharing systems and public housing, when there is a need to allocate different types of resources to different types of customers. Because of the complex stochastic nature of such problems, fluid deterministic approximations become helpful in designing effective allocation policies.

In organ transplant problems, transplant candidates and organ donors need to be matched, in order to optimize factors such as expected quality-adjusted life expectancy of transplant candidates (quality-adjusted life-year – QALY), post-transplant survival probability, matching quality, number of patient deaths, and number of wasted organs (see David and Yechiali [47]).

Zenios, Chertow and Wein [140] considered the trade-off between clinical efficiency and equity when designing a dynamic kidney allocation policy. The clinical efficiency is captured by QALY, while the equity is captured by the transplant likelihood of patient type, and the differences in mean waiting times across them. By using a fluid model and an approximate analysis of the optimal control problem, they developed a dynamic index policy that is effective and implementable.

Later on, [5] used a fluid model to study liver transplants and analyzed the trade-off between medical urgency and efficiency over a finite time horizon. The medical urgency is captured by the number of patient deaths while waiting for a transplant (NPDWT). Patients waiting in a queue may abandon the system or switch between classes (when deteriorating or improving). When considering only the NPDWT criterion (without the clinical efficiency), the current United Network for Organ Sharing (UNOS) policy is optimal. On the other hand, considering only the efficiency criterion yields a dynamic index policy that prioritizes patients on the waiting list according to their potential marginal benefit from transplantation. The work in [5] was extended in [70] who studied the allocation problem under fairness constraints. The optimal policy of their proposed fluid model is a dynamic priority rule. More recently, Ata, Ding and Zenios [14] developed a fluid model to study a kidney allocation problem that considers strategic patients when choosing to accept or decline an offered organ.

In ride-sharing systems, arriving customers need to be matched with available drivers. Özkan and Ward [109] used a fluid model to develop dynamic matching for a real-time ride-sharing system. They proposed a randomized policy, which is based on the solution of a continuous linear program that incorporates the time-varying different arrival rates of customers/drivers in different city areas, and the time customers/drivers are willing to wait. The policy was shown to be asymptotically optimal under fluid scaling. Braverman et al. [30] studied the control of an empty-car routing problem in a closed queueing network, where the objective is to optimize utility functions such as the availability of empty cars when a passenger arrives. They proposed an efficient fluid-based optimal routing policy for empty vehicles in a large market regime.

Ding, McCormick and Nagarajan [49] used a fluid model to study a one-sided bipartite matching problem with match-dependent rewards. That is, a resource is allocated to the customer with the highest score, which is calculated as the sum of a customer's waiting time and matching score.

A two-sided controlled problem arises when, in addition to customers, resources also wait in a queue. Afeche, Caldentey and Gupta [2] studied the design of a matching topology in a multiclass multiserver queueing system under a first come first served–assign longest idle server (FCFS–ALIS) service discipline. The trade-off they considered is between minimizing customers' waiting time and maximizing customer–server matching rewards. The risk of losing a waiting customer/resource

while waiting for a match is considered by Aveklouris et al. [20], who developed a fluid-based two-sided matching policy under generally distributed customer patience.

7. Delay-Informed Queue Access

Many service providers share delay information with their customers. Sharing such information was shown to be beneficial. It increases customer satisfaction by reducing undesirable uncertainty, it enables customers to have increased control over their wait, and it provides customers with a sense of progress while they are waiting [87, 104, 76]. Delay announcements are then used by potential customers in order to decide whether to join the system or to balk. Balking customers can be observed or non-observed by the system operator [80]. Time-varying arrivals and abandonments while waiting impose additional challenges on delay prediction. A substantial body of research addresses the challenges involved in effectively managing the provision of delay announcements. Since large systems are the main focus in this context and direct analysis is prohibitively complex, deterministic fluid models become a natural effective tool within the many-server heavy-traffic framework.

Armony, Shimkin and Whitt [11] investigated the effect of delay announcements when the abandonments due to the announcements are non-exponential. The authors suggested approximations of the steady-state performance with delay announcements. One such approximation is based on the equilibrium delay in a deterministic fluid model, where the expected steady-state delay coincides with the delay announcement. The fluid model provides useful insights in an overloaded regime, when delay announcements, according to the authors, are important. Based on the deterministic fluid model, conditions for customer response are derived to guarantee the existence and uniqueness of that equilibrium. Moreover, in the fluid model, the LES (Last customer to have Entered Service) delay coincides with the FD (Fixed Delay) announcement at equilibrium. The authors also showed that when the abandonment response to the announcements is not smooth, the fluid model may not be accurate. In such cases, a further refinement through diffusion approximation, as in [75], is required.

In a series of papers, Ibrahim and Whitt [77, 78, 79] studied the asymptotic accuracy of real-time delay predictors in service systems with a single class of customers. Using a many-server heavy-traffic framework, they developed and examined different predictors that are based on either the queue length or the history of delays for queueing systems with time-varying arrivals, abandonments and general distributional assumptions. In [79], for example, two predictors that exploit a deterministic fluid approximation for a many-server queueing model are suggested for a system with a time-varying arrival rate, a time-varying number of servers and customer abandonments.

In a different series of papers, Pender et al. [112, 113] and Novitzky et al. [106, 107] used deterministic fluid models to study the effect of providing customers with delayed information. They showed how delayed information can cause oscillations or asynchronous behavior and produce unwanted system dynamics. With the goal of evenly distributing the workload among queues, they analyzed the kind of information that should be provided to arriving customers.

8. Customer Acquisition and Retention

Fluid models have been used to set customer acquisition strategies and revenue maximization. Ata et al. [15] used fluid models to analyze the expected profit of hospice care. They proposed an alternative reimbursement policy for the US Medicare system and determined the recruiting rates of short- and long-stay patients to maximize profitability of the hospice. Afeche, Araghi and Baron [1] utilized a fluid model for setting customer acquisition investments and bottleneck capacity allocation to maximize the profit of service firms. Specifically, they addressed the questions of how much to spend on customer acquisition, how much capacity to deploy, and how to allocate capacity and adjust service access quality levels in line with different customer types. Recently, Furman, Diamant and Kristal [55] focused on the trade-off between acquisition and retention efforts when customers are sensitive to service quality. The model, which is a multi-class queueing network with new and returning customers, time-dependent arrivals, and abandonment, is approximated by a fluid model. The model is used to determine optimal stationary staffing levels for new and returning customers.

Savin et al. [121] used a fluid approximation model to study capacity allocation decisions in a rental-equipment system with two classes of customers. Upon a customer's arrival, the system's controller decides whether to admit the customer for service (given that there is available capacity) or to reject the arrival. Based on the fluid model, they developed heuristic capacity allocation policies and obtained closed-form expressions for the heuristic's control parameters. Akan and Ata [6] considered a continuous-time stochastic fluid model of network revenue management. They showed that when the volume and capacity are sufficiently large, the revenue management systems can be approximated by fluid models and that the optimal bid-price processes are martingales in the stochastic fluid models. Dai, Kleywegt and Xiao [44] studied airline network revenue management problems with customer cancellations and no-shows. Their goal was to optimize booking policies in an independent-demand model and a choice-based one. Using a fluid model, they derived a policy that is asymptotically optimal when the arrival rates become high and the seat capacities become large.

9. Translation of the Fluid Solution Back to the Stochastic System

Fluid models are deterministic continuous approximations of underlying discrete stochastic systems. Under the limiting regime where the approximation is valid, the fluid model can be extremely accurate. Specifically, conventional heavy-traffic regime accurately approximates heavily loaded systems (i.e., when traffic intensity approaches 1); many-server heavy-traffic regime accurately approximates large systems with high arrival rates and many servers. When these conditions are not met, the fluid approximation can be inaccurate. This may lead to poorly performing polices, and even instability when implemented straightforwardly. Moreover, since fluid models are continuous, a translation mechanism is needed to reinstate the fluid solution into the original discrete stochastic system. The translation mechanism needs to guarantee stability and good performance of the stochastic system; it also needs to be asymptotically optimal in the relevant scaling regime.

Harrison [66] suggested a general translation mechanism, the BIGSTEP approach for dynamic control of stochastic networks. Harrison [67] implemented the BIGSTEP approach to schedule a simple N-model with two stations and two classes of customers. In particular, a discrete-review control policy (i.e., system status is reviewed at fixed-length intervals and decisions are made for the next interval) was constructed and proved to be asymptotically optimal in the heavy traffic limit. This approach was extended in Maglaras [94] who proposed a class of discrete-review policies based on repeatedly solving fluid relaxation problems. Specifically, the status of the system is reviewed at discrete time points; then, the system operator formulates a processing plan for the next review period to best track the fluid control policy at that point. Within each period only customers that were present at the review time point are allowed to be processed. The implementation of a discrete-review policy requires the usage of safety stocks (i.e., thresholds on the number of customers of each class at the end of the period) to prevent undesirable idleness when some classes get depleted. These safety stocks are asymptotically negligible (i.e., under fluid scaling). Maglaras [94] also proved the stability of these discrete-review policies and established their fluid-scale asymptotic optimality.

Bertsimas, Gamarnik and Sethuraman [28] suggested an efficient asymptotically optimal algorithm for rounding an optimal fluid solution to the job-shop scheduling problem, where the fluid relaxation is solved once. The authors also provided an explicit convergence rate to optimality.

Other more specific translation mechanisms of fluid solutions to discrete near-optimal scheduling policies were suggested in [38, 19, 102, 46, 33].

10. Concluding Remarks and Future Opportunities

The stochasticity and complexity of service systems encourage researchers to look for useful effective approximations to gain insights and optimize system performance. Such approximations need to be accurate in describing the average performance of the corresponding stochastic systems; they

also need to be manageable in terms of analysis and derivation of simple yet insightful policies. Fluid models provide a useful framework for such approximations and have, therefore, been used to address a variety of problems in the area of service operations management.

We identify several future directions in which fluid models have the potential to capture and explore important phenomena in service and healthcare systems. Most existing network applications are based on common assumptions of independency between service stations, service times and system state as well as outcomes. In healthcare systems, the latter refers to readmissions, service times of readmitted patients and mortality. There is, however, a reasonable amount of empirical evidence that shows that dependency prevails in many service, healthcare and transportation systems [37, 84, 85, 122, 26, 27, 125, 32, 22]. For instance, transferring patients earlier than necessary to the next station or keeping patients blocked due to lack of availability in the next station is likely to affect their healthcare outcomes. Moreover, readmitted patients are likely to arrive in a worse condition than when first admitted that will necessitate longer hospitalization. Therefore, such dependencies need to be taken into account carefully. Another relevant example regards COVID controls: to maintain spacial distancing, downstream status (e.g., long queues) affected upstream decisions (e.g., strategic idling). Similar dependency also prevails in other service systems such as call centers, product development centers, white-collar work systems [71], and even in production systems. Capturing such dependencies through exact analysis of the stochastic network will likely lead to models that are intractable either analytically or computationally. Using fluid approximating models, in contrast, is more promising and can uncover the dependencies' effect on optimal decision making.

Another dependency that calls for further exploration is observed when longer service times in one station lead to shorter service times in the next station and vice versa. For example, in a healthcare setting (e.g., ICU and inpatient wards or inpatient wards and rehabilitation institutes), a shorter stay in one station can be compensated for by a longer stay in the next one [86]. On the other hand, a longer stay in one station may allow shortening the stay at the next station. Similar dependencies also prevail in maintenance-repair systems or systems with preventive maintenance. Developing fluid models that capture and optimize such a balance is both an interesting and important research direction.

Another research direction involves the embodying of dependencies between service times and an individual's state. Although there is a body of literature that focuses on developing fluid models for general service distributions, most existing applications are based on common assumptions of memoryless service times. Moreover, most fluid models take a "bird's eye view" and are thus unable to depict and control *individual* progression through the system. This calls

for the development of fluid models that capture both the individual's state and the system state. In a healthcare setting, the individual's state refers to the patient's health condition. Providing an underlying mechanism that monitors the individual's state in time throughout the network can promote the construction of fluid models in two dimensions: state and time. Such a mechanism can facilitate the analysis of more complex realistic systems and help address questions of control on a finer scale. For example, future fluid models should seek to answer questions such as what state determines that customers/patients should be transferred from one station to another, what state indicates that customers/patients should be returned/readmitted, and how do these decisions change under congestion and blocking that occurs when, due to congestion at the next station, available customers cannot be relocated. Other promising directions include the incorporation of dependencies between arrival times and system state or individual waiting/blocking time in one station and service times or individual state at the next station.

Declarations

Funding. Partial financial support was received from ISF Grant 277/21 and the Israel National Institute for Health Policy Research, grant 2021/160/R.

Competing interests. The author has no competing interests to declare that are relevant to the content of this article.

Acknowledgements

The author thanks the editor-in-chief, Michel Mandjes, and the anonymous editorial team for their constructive feedback and valuable suggestions that helped improve the paper. The author is very grateful to Avishai Mandelbaum and Petar Momčilović for their insightful comments and suggestions.

References

[1] Afèche, P., Araghi, M., Baron, O.: Customer acquisition, retention, and queueing-related service quality: Optimal advertising, staffing, and priorities for a call center. Manufacturing and Service Operations Management **19**(4), 674–691 (2017) 2, 19

[2] Afeche, P., Caldentey, R., Gupta, V.: On the optimal design of a bipartite matching queueing system. Operations Research (2021) 17

[3] Aguir, M., Akşin, O., Karaesmen, F., Dallery, Y.: On the interaction between retrials and sizing of call centers. European Journal of Operational Research **191**(2), 398–408 (2008) 13

[4] Aguir, S., Karaesmen, F., Akşin, O., Chauvet, F.: The impact of retrials on call center performance. OR Spectrum **26**(3), 353–376 (2004) 13

[5] Akan, M., Alagoz, O., Ata, B., Erenay, F., Said, A.: A broader view of designing the liver allocation system. Operations Research **60**(4), 757–770 (2012) 17

[6] Akan, M., Ata, B.: Bid-price controls for network revenue management: Martingale characterization of optimal bid prices. Mathematics of Operations Research **34**(4), 912–936 (2009) 19

[7] Aksin, Z., Armony, M., Mehrotra, V.: The modern call center: A multi-disciplinary perspective on operations management research. Production and Operations Management **16**(6), 665–688 (2007) 2, 12

[8] Altman, E., Jiménez, T., Koole, G.: On the comparison of queueing systems with their fluid limits. Probability in the Engineering and Informational Sciences **15**(2), 165 (2001) 6

[9] Armony, M.: Dynamic routing in large-scale service systems with heterogeneous servers. Queueing Systems **51**(3), 287–329 (2005) 12

[10] Armony, M., Mandelbaum, A.: Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. Operations Research **59**(1), 50–65 (2011) 12

[11] Armony, M., Shimkin, N., Whitt, W.: The impact of delay announcements in many-server queues with abandonment. Operations Research **57**(1), 66–81 (2009) 18

[12] Armony, M., Ward, A.: Fair dynamic routing in large-scale heterogeneous-server systems. Operations Research **58**(3), 624–637 (2010) 12

[13] Armony, M., Yom-Tov, G.: Hospitalization versus home care: Balancing mortality and infection risks for hematology patients. working paper (2021) 2, 16

[14] Ata, B., Ding, Y., Zenios, S.: An achievable-region-based approach for kidney allocation policy design with endogenous patient choice. Manufacturing & Service Operations Management **23**(1), 36–54 (2021) 17

[15] Ata, B., Killaly, B., Olsen, T., Parker, R.: On hospice operations under medicare reimbursement policies. Management Science **59**(5), 1027–1044 (2013) 19

[16] Atar, R.: Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. The Annals of Applied Probability **15**(4), 2606–2650 (2005) 12

[17] Atar, R., Giat, C., Shimkin, N.: The $c\mu/\theta$ rule for many-server queues with abandonment. Operations Research **58**(5), 1427–1439 (2010) 14

[18] Atar, R., Shaki, Y., Shwartz, A.: A blind policy for equalizing cumulative idleness. Queueing Systems **67**(4), 275–293 (2011) 12

[19] Atkins, D., Chen, H.: Performance evaluation of scheduling control of queueing networks: Fluid model heuristics. Queueing Systems **21**(3), 391–413 (1995) 20

[20] Aveklouris, A., DeValve, L., Ward, A., Wu, X.: Matching impatient and heterogeneous demand and supply. arXiv preprint arXiv:2102.02710 (2021) 18

[21] Baron, O., Milner, J.: Staffing to maximize profit for call centers with alternate service-level agreements. Operations Research **57**(3), 685–700 (2009) 12

[22] Bartel, A., Chan, C., Kim, S.H.: Should hospitals keep their patients longer? The role of inpatient care in reducing post-discharge mortality. Management Science **66**(6), 2326–2346 (2020) 21

[23] Bassamboo, A., Harrison, J., Zeevi, A.: Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. Queueing Systems **51**(3), 249–285 (2005) 12, 15

[24] Bassamboo, A., Randhawa, R.: On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. Operations Research **58**(5), 1398–1413 (2010) 14

[25] Bassamboo, A., Randhawa, R., Zeevi, A.: Capacity sizing under parameter uncertainty: Safety staffing principles revisited. Management Science **56**(10), 1668–1686 (2010) 13

[26] Batt, R., Terwiesch, C.: Doctors under load: An empirical study of state-dependent service times in emergency care. The Wharton School, the University of Pennsylvania, Philadelphia, PA **19104** (2012) 21

[27] Berry J., J.A., Tucker, A.: Hurry up and wait: Differential impacts of congestion, bottleneck pressure, and predictability on patient length of stay. Harvard Business School working paper series# 13-052 (2012) 21

[28] Bertsimas, D., Gamarnik, D., Sethuraman, J.: From fluid relaxations to practical algorithms for high-multiplicity job-shop scheduling: The holding cost objective. Operations Research **51**(5), 798–813 (2003) 20

[29] Borst, S., Mandelbaum, A., Reiman, M.: Dimensioning large call centers. Operations Research **52**(1), 17–34 (2004) 6

[30] Braverman, A., Dai, J., Liu, X., Ying, L.: Empty-car routing in ridesharing systems. Operations Research **67**(5), 1437–1452 (2019) 2, 17

[31] Carmeli, N., Mandelbaum, A., Yom-Tov, G.: Data-based resource-view of service networks: Performance analysis, delay prediction and asymptotics. Ph.D. thesis, Technion-Israel Institute of Technology (2020) 7, 8

[32] Chan, C., Farias, V., Escobar, G.: The impact of delays on service times in the intensive care unit. Management Science **63**(7), 2049–2072 (2017) 21

[33] Chan, C., Huang, M., Sarhangian, V.: Dynamic server assignment in multiclass queues with shifts, with applications to nurse staffing in emergency departments. Operations Research (2021) 16, 20

[34] Chan, C., Sarhangian, V., Talwai, P., G., K.: Utilizing partial flexibility to improve emergency department flow: Theory and implementation. Working paper (2022) 16

[35] Chan, C., Yom-Tov, G., Escobar, G.: When to use speedup: An examination of service systems with returns. Operations Research **62**(2), 462–482 (2014) 15

[36] Chan, T., Huang, S., Sarhangian, V.: Dynamic control of service systems with returns: Application to design of post-discharge hospital readmission prevention programs. Working paper (2022) 16

[37] Chen, C., Jia, Z., Varaiya, P.: Causes and cures of highway congestion. IEEE Control Systems Magazine **21**(6), 26–32 (2001) 21

[38] Chen, H., Yao, D.: Dynamic scheduling of a multiclass fluid network. Operations Research **41**(6), 1104–1115 (1993) 20

[39] Chen, H., Yao, D.: Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization. Springer Science & Business Media (2013) 6

[40] Cohen, I., Mandelbaum, A., Zychlinski, N.: Minimizing mortality in a mass casualty event: Fluid networks in support of modeling and staffing. IIE Transactions **46**(7), 728–741 (2014) 15

[41] Cox, D., Smith, W.: Queues. Methuen, London (1961) 14

[42] Dai, J.: On positive harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. The Annals of Applied Probability pp. 49–77 (1995) 4

[43] Dai, J., Harrison, J.: Processing Networks: Fluid Models and Stability. Cambridge University Press (2020) 4

[44] Dai, J., Kleywegt, A., Xiao, Y.: Network revenue management with cancellations and no-shows. Production and Operations Management **28**(2), 292–318 (2019) 19

[45] Dai, J., Shi, P.: Inpatient overflow: An approximate dynamic programming approach. Manufacturing & Service Operations Management **21**(4), 894–911 (2019) 15

[46] Dai, J., Weiss, G.: A fluid heuristic for minimizing makespan in job shops. Operations Research **50**(4), 692–707 (2002) 20

[47] David, I., Yechiali, U.: One-attribute sequential assignment match processes in discrete time. Operations Research **43**(5), 879–884 (1995) 16

[48] De Neufville, R., Odoni, A., Belobaba, P., Reynolds, T.: Airport systems: Planning, Design, and Management. McGraw-Hill Education (2013) 4

[49] Ding, Y., McCormick, S., Nagarajan, M.: A fluid model for one-sided bipartite matching queues with match-dependent rewards. Operations Research (2021) 17

[50] Dobson, G., Tezcan, T., Tilson, V.: Optimal workflow decisions for investigators in systems with interruptions. Management Science **59**(5), 1125–1141 (2013) 14

[51] Dong, J., Feldman, P., Yom-Tov, G.: Service systems with slowdowns: Potential failures and proposed solutions. Operations Research **63**(2), 305–324 (2015) 15

[52] Dong, J., Ibrahim, R.: Managing supply in the on-demand economy: Flexible workers, full-time employees, or both? Operations Research **68**(4), 1238–1264 (2020) 13

[53] Dong, J., Ibrahim, R.: SRPT scheduling discipline in many-server queues with impatient customers. Management Science **67**(12), 7708–7718 (2021) 14

[54] Dong, J., Perry, O.: Queueing models for patient-flow dynamics in inpatient wards. Operations Research **68**(1), 250–275 (2020) 2, 7, 16

[55] Furman, E., Diamant, A., Kristal, M.: Customer acquisition and retention: A fluid approach for staffing. Production and Operations Management **30**(11), 4236–4257 (2021) 19

[56] Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: Tutorial, review, and research prospects. Manufacturing & Service Operations Management **5**(2), 79–141 (2003) 2, 12, 14

[57] Garnett, O., Mandelbaum, A., Reiman, M.: Designing a call center with impatient customers. Manufacturing & Service Operations Management **4**(3), 208–227 (2002) 12

[58] Gurvich, I., Armony, M., Mandelbaum, A.: Service-level differentiation in call centers with fully flexible servers. Management Science **54**(2), 279–294 (2008) 12

[59] Gurvich, I., Luedtke, J., Tezcan, T.: Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. Management Science **56**(7), 1093–1115 (2010) 13

[60] Gurvich, I., Perry, O.: Overflow networks: Approximations and implications to call center outsourcing. Operations Research **60**(4), 996–1009 (2012) 7

[61] Gurvich, I., Whitt, W.: Queue-and-idleness-ratio controls in many-server service systems. Mathematics of Operations Research **34**(2), 363–396 (2009) 12

[62] Gurvich, I., Whitt, W.: Service-level differentiation in many-server service systems via queue-ratio routing. Operations Research **58**(2), 316–328 (2010) 12

[63] Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Operations Research **29**(3), 567–588 (1981) 12

[64] Hall, R.: Queueing Methods for Services and Manufacturing. Pearson College Division (1991) 4, 7

[65] Hall, R.: Patient Flow. AMC **10**, 12 (2013) 4, 5

[66] Harrison, J.: The BIGSTEP approach to flow management in stochastic processing networks. Stochastic Networks: Theory and Applications **4**, 147–186 (1996) 20

[67] Harrison, J.: Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete-review policies. Annals of Applied Probability pp. 822–848 (1998) 5, 20

[68] Harrison, J., Zeevi, A.: Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. Operations Research **52**(2), 243–257 (2004) 12

[69] Harrison, J., Zeevi, A.: A method for staffing large call centers based on stochastic fluid models. Manufacturing & Service Operations Management **7**(1), 20–36 (2005) 13

[70] Hasankhani, F., Khademi, A.: Is it time to include post-transplant survival in heart transplantation allocation rules? Production and Operations Management (2019) 17

[71] Hopp, W., Iravani, S., Yuen, G.: Operations systems with discretionary task completion. Management Science **53**(1), 61–77 (2007) 21

[72] Horonjeff, R., McKelvey, F., Sproule, W., Young, S.: Planning and Design of Airports. McGraw-Hill Education (2010) 4

[73] Hu, Y., Chan, C., Dong, J.: Optimal scheduling of proactive service with customer deterioration and improvement. Management Science (2020) 2, 16

[74] Huang, J., Carmeli, B., Mandelbaum, A.: Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. Operations Research **63**(4), 892–908 (2015) 12

[75] Huang, J., Mandelbaum, A., Zhang, H., Zhang, J.: Refined models for efficiency-driven queues with applications to delay announcements and staffing. Operations Research **65**(5), 1380–1397 (2017) 18

[76] Ibrahim, R.: Sharing delay information in service systems: A literature survey. Queueing Systems **89**(1), 49–79 (2018) 18

[77] Ibrahim, R., Whitt, W.: Real-time delay estimation based on delay history. Manufacturing & Service Operations Management **11**(3), 397–415 (2009) 18

[78] Ibrahim, R., Whitt, W.: Real-time delay estimation in overloaded multiserver queues with abandonments. Management Science **55**(10), 1729–1742 (2009) 18

[79] Ibrahim, R., Whitt, W.: Wait-time predictors for customer service systems with time-varying demand and capacity. Operations Research **59**(5), 1106–1118 (2011) 18

[80] Inoue, Y., Ravner, L., Mandjes, M.: Estimating customer impatience in a service system with unobserved balking. Stochasic Systems, to appear (2022) 18

[81] Janssen, A., Van Leeuwaarden, J., Zwart, B.: Refining square-root safety staffing by expanding Erlang C. Operations Research **59**(6), 1512–1522 (2011) 6

[82] Jennings, O., Massey, W., McCalla, C.: Optimal profit for leased lines services. In: Proceedings of the 15th International Teletraffic Congress – ITC, vol. 15, pp. 803–814 (1997) 14

[83] Jiménez, T., Koole, G.: Scaling and comparison of fluid limits of queues applied to call centers with time-varying parameters. OR Spectrum **26**(3), 413–422 (2004) 2

[84] Kc, D., Terwiesch, C.: Impact of workload on service time and patient safety: An econometric analysis of hospital operations. Management Science **55**(9), 1486–1498 (2009) 21

[85] Kc, D., Terwiesch, C.: An econometric analysis of patient flows in the cardiac intensive care unit. Manufacturing & Service Operations Management **14**(1), 50–65 (2012) 21

[86] Konetzka, R., Stuart, E., Werner, R.: The effect of integration of hospitals and post-acute care providers on medicare payment and patient outcomes. Journal of health economics **61**, 244–258 (2018) 21

[87] Leclerc, F., Schmitt, B., Dube, L.: Waiting time and decision making: Is time like money? Journal of Consumer Research **22**(1), 110–119 (1995) 18

[88] Lewin, D.: Queueuing at airport desks. In: Airport Forum, vol. 6 (1976) 4

[89] Liu, Y., Whitt, W.: Large-time asymptotics for the $G_t/M_t/s_t+GI_t$ many-server fluid queue with abandonment. Queueing systems **67**(2), 145–182 (2011) 2, 7

[90] Liu, Y., Whitt, W.: A network of time-varying many-server fluid queues with customer abandonment. Operations Research **59**(4), 835–846 (2011) 7

[91] Liu, Y., Whitt, W.: The $G_t/GI/s_t+GI$ many-server fluid queue. Queueing Systems **71**(4), 405–444 (2012) 4, 7

[92] Liu, Y., Whitt, W.: Many-server heavy-traffic limit for queues with time-varying parameters. Annals of Applied Probability **24**(1), 378–421 (2014) 2, 7

[93] Long, Z., Shimkin, N., Zhang, H., Zhang, J.: Dynamic scheduling of multiclass many-server queues with abandonment: The generalized cμ/h rule. Operations Research **68**(4), 1218–1230 (2020) 12

[94] Maglaras, C.: Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality. Annals of Applied Probability pp. 897–929 (2000) 20

[95] Mandelbaum, A., Massey, W.: Strong approximations for time-dependent queues. Mathematics of Operations Research **20**(1), 33–64 (1995) 6

[96] Mandelbaum, A., Massey, W., Reiman, M.: Strong approximations for Markovian service networks. Queueing Systems **30**(1-2), 149–201 (1998) 2, 7

[97] Mandelbaum, A., Massey, W., Reiman, M., Rider, B.: Time varying multiserver queues with abandonment and retrials. In: Proceedings of the 16th International Teletraffic Conference (1999) 2, 4, 7, 8, 9

[98] Mandelbaum, A., Momčilović, P.: Personalized queues: The customer view, via a fluid model of serving least-patient first. Queueing Systems **87**(1), 23–53 (2017) 15

[99] Mandelbaum, A., Momčilović, P., Tseytlin, Y.: On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. Management Science **58**(7), 1273–1291 (2012) 12

[100] Mandelbaum, A., Stolyar, A.: Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. Operations Research **52**(6), 836–855 (2004) 5, 12

[101] Mandelbaum, A., Zeltyn, S.: Service engineering: Data-based course development and teaching. INFORMS Transactions on Education **11**(1), 3–19 (2010) 4

[102] Meyn, S.: Stability and optimization of queueing networks and their fluid models. Lectures in applied mathematics-American Mathematical Society **33**, 175–200 (1997) 20

[103] Mills, A., Argon, N., Ziya, S.: Resource-based patient prioritization in mass-casualty incidents. Manufacturing & Service Operations Management **15**(3), 361–377 (2013) 15

[104] Munichor, N., Rafaeli, A.: Numbers or apologies? Customer reactions to telephone waiting time fillers. Journal of Applied Psychology **92**(2), 511 (2007) 18

[105] Newell, C.: Applications of queueing theory, vol. 4. Springer Science & Business Media (2013) 5

[106] Novitzky, S., Pender, J., Rand, R., Wesson, E.: Nonlinear dynamics in queueing theory: Determining the size of oscillations in queues with delay. SIAM Journal on Applied Dynamical Systems **18**(1), 279–311 (2019) 19

[107] Novitzky, S., Pender, J., Rand, R., Wesson, E.: Limiting the oscillations in queues with delayed information through a novel type of delay announcement. Queueing Systems **95**(3), 281–330 (2020) 19

[108] Oliver, R., Samuel, A.: Reducing letter delays in post offices. Operations Research **10**(6), 839–892 (1962) 2, 3

[109] Özkan, E., Ward, A.: Dynamic matching for real-time ride sharing. Stochastic Systems **10**(1), 29–70 (2020) 2, 17

[110] Pang, G., Whitt, W.: Heavy-traffic limits for many-server queues with service interruptions. Queueing Systems **61**(2), 167–202 (2009) 2

[111] Paullin, R., Horonjeff, R.: Sizing of departure lounges in airport buildings. Transportation Engineering Journal of ASCE **95**(2), 267–277 (1969) 4

[112] Pender, J., Rand, R., Wesson, E.: Queues with choice via delay differential equations. International Journal of Bifurcation and Chaos **27**(04), 1730,016 (2017) 19

[113] Pender, J., Rand, R., Wesson, E.: An analysis of queues with delayed information and time-varying arrival rates. Nonlinear Dynamics **91**(4), 2411–2427 (2018) 19

[114] Perry, O., Whitt, W.: Responding to unexpected overloads in large-scale service systems. Management Science **55**(8), 1353–1367 (2009) 15

[115] Perry, O., Whitt, W.: A fluid approximation for service systems responding to unexpected overloads. Operations Research **59**(5), 1159–1170 (2011) 15

[116] Porteus, E.: The case analysis section: National cranberry cooperative. Interfaces **19**(6), 29–39 (1989) 3

[117] Porteus, E.: Case analysis: Analyses of the national cranberry cooperative—1. tactical options. Interfaces **23**(4), 21–39 (1993) 3

[118] Porteus, E.: Case analysis: Analyses of the national cranberry cooperative—2. environmental changes and implementation. Interfaces **23**(6), 81–92 (1993) 3

[119] Ren, Z., Zhou, Y.P.: Call center outsourcing: Coordinating staffing level and service quality. Management Science **54**(2), 369–383 (2008) 13

[120] Ridley, A., Fu, M., Massey, W.: Fluid approximations for a priority call center with time-varying arrivals. In: Winter Simulation Conference, vol. 2, pp. 1817–1823 (2003) 2

[121] Savin, S., Cohen, M., Gans, N., Katalan, Z.: Capacity management in rental businesses with two customer bases. Operations Research **53**(4), 617–631 (2005) 19

[122] Staats, B., Gino, F.: Specialization and variety in repetitive tasks: Evidence from a japanese bank. Management Science **58**(6), 1141–1159 (2012) 21

[123] Stolyar, A.: Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. The Annals of Applied Probability **14**(1), 1–53 (2004) 5

[124] Talreja, R., Whitt, W.: Fluid models for overloaded multiclass many-server queueing systems with first-come, first-served routing. Management Science **54**(8), 1513–1527 (2008) 7, 14

[125] Tan, T., Netessine, S.: When does the devil make work? An empirical study of the impact of workload on worker productivity. Management Science **60**(6), 1574–1593 (2014) 21

[126] Tanner, J.: A queueing model for departure baggage handling at airports. Institute of Transportation and Traffic Engineering, University of California (1966) 4

[127] Tezcan, T., Dai, J.: Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. Operations Research **58**(1), 94–110 (2010) 12

[128] Tošić, V.: A review of airport passenger terminal operations analysis and modelling. Transportation Research Part A: Policy and Practice **26**(1), 3–26 (1992) 2, 4

[129] Van Mieghem, J.: Dynamic scheduling with convex delay costs: The generalized cμ rule. The Annals of Applied Probability pp. 809–833 (1995) 12

[130] Vandergraft, J.: A fluid flow model of networks of queues. Management Science **29**(10), 1198–1208 (1983) 2, 3

[131] Whitt, W.: Understanding the efficiency of multi-server service systems. Management Science **38**(5), 708–723 (1992) 12, 13

[132] Whitt, W.: Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues. Springer Science & Business Media (2002) 5, 6

[133] Whitt, W.: Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. Management Science **50**(10), 1449–1461 (2004) 2

[134] Whitt, W.: Two fluid approximations for multi-server queues with abandonments. Operations Research Letters **33**(4), 363–372 (2005) 7

[135] Whitt, W.: Fluid models for multiserver queues with abandonments. Operations Research **54**(1), 37–54 (2006) 2, 3, 7, 13

[136] Whitt, W.: A multi-class fluid model for a contact center with skill-based routing. AEU–International Journal of Electronics and Communications **60**(2), 95–102 (2006) 12

[137] Whitt, W.: Staffing a call center with uncertain arrival rate and absenteeism. Production and Operations Management **15**(1), 88–102 (2006) 12, 13

[138] Whitt, W.: Heavy-traffic limits for queues with periodic arrival processes. Operations Research Letters **42**(6-7), 458–461 (2014) 6

[139] Yom-Tov, G., Mandelbaum, A.: Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. Manufacturing & Service Operations Management **16**(2), 283–299 (2014) 2, 6, 7, 13, 16

[140] Zenios, S., Chertow, G., Wein, L.: Dynamic allocation of kidneys to candidates on the transplant waiting list. Operations Research **48**(4), 549–569 (2000) 17

[141] Zychlinski, N.: Managing queues with reentrant customers in support of hybrid healthcare. Working paper (2022) 16

[142] Zychlinski, N., Chan, C., Dong, J.: Managing queues with different resource requirements. Operations Research, forthcoming (2022) 3, 14

[143] Zychlinski, N., Mandelbaum, A., Momčilović, P.: Time-varying tandem queues with blocking: Modeling, analysis, and operational insights via fluid models with reflection. Queueing Systems **89**(1-2), 15–47 (2018) 7, 9, 16

[144] Zychlinski, N., Mandelbaum, A., Momčilović, P., Cohen, I.: Bed blocking in hospitals due to scarce capacity in geriatric institutions—cost minimization via fluid models. Manufacturing & Service Operations Management **22**(2), 396–411 (2020) 16

[145] Zychlinski, N., Momčilović, P., Mandelbaum, A.: Time-varying many-server finite-queues in tandem: Comparing blocking mechanisms via fluid models. Operations Research Letters **46**(5), 492–499 (2018) 16