MANAGEMENT SCIENCE
Vol. 54, No. 8, August 2008, pp. 1513–1527
ISSN 0025-1909 | EISSN 1526-5501 | 08 | 5408 | 1513

informs

DOI 10.1287/mnsc.1080.0868

Fluid Models for Overloaded Multiclass Many-Server Queueing Systems with First-Come, First-Served Routing

Rishi Talreja, Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027
{rt2146@columbia.edu, ww2040@columbia.edu}

Motivated by models of tenant assignment in public housing, we study approximating deterministic fluid models for overloaded queueing systems having multiple customer classes (classes of tenants) and multiple service pools (housing authorities), each with many servers (housing units). Customer abandonment acts to keep the system stable, yielding a proper steady-state description. Motivated by fairness considerations, we assume that customers are selected for service by newly available servers on a first-come, first-served (FCFS) basis from all classes the corresponding service pools are allowed to serve. In this context, it is challenging to determine stationary routing flow rates between customer classes and service pools. Given those routing flow rates, each single fluid queue can be analyzed separately using previously established methods. Our ability to determine the routing flow rates depends on the structure of the network routing graph. We obtain the desired routing flow rates in three cases: when the routing graph is (i) a tree (sparsely connected), (ii) complete bipartite (fully connected), and (iii) an appropriate combination of the previous two cases. Other cases remain unsolved. In the last two solved cases, the routing flow rates are actually not uniquely determined by the fluid model, but become so once we make stochastic assumptions about the queueing models that the fluid model approximates.

Key words: deterministic fluid models; service networks; many-server queues; customer abandonment; overloaded queues; tenant assignment in public housing
History: Accepted by Michael Fu, stochastic models and simulation; received March 20, 2007. This paper was with the authors 1 week for 1 revision. Published online in *Articles in Advance* June 20, 2008.

1. Introduction

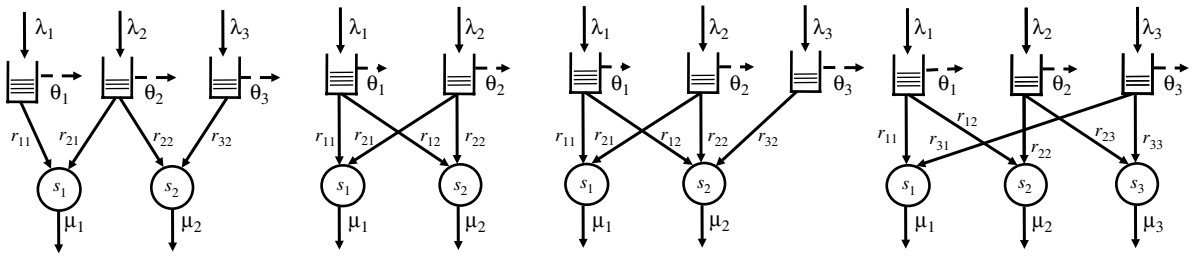
In this paper, we investigate deterministic fluid approximations for overloaded queueing systems having multiple customer classes and multiple service pools, each with many servers. Each customer class has a fixed subset of service pools where it can be served. Customer abandonment acts to keep the system stable, yielding a proper steady-state description (e.g., queue lengths and waiting times). We consider the case in which customers are selected for service by newly available servers on a first-come, first-served (FCFS) basis from all classes that the corresponding service pools are allowed to serve.

We have in mind (overloaded) queueing systems as depicted in Figure 1. Customers from class i arrive at rate λ_i and abandon the queue at rate θ_i. In the long run, class i customers are served by service pool j at rate $r_{i,j}$. Customers complete service from pool j at rate μ_j. The natural direct queueing model consistent with that description has independent Poisson arrival processes, independent and identically distributed (i.i.d.) exponential times to abandon, and multiple pools of servers, each with i.i.d. exponential service times. However, here we focus on an approximating deterministic fluid model, where we regard all flows as being deterministic. In the queueing model, there are s_j servers at service pool j, each serving customers at rate η_j. Because the system is overloaded, these servers are busy almost all of the time, so that the overall service rate at pool j in the fluid model is $\mu_j = s_j\eta_j$. We are interested in the steady-state behavior, where the fluid flow rates are constant.

1.1. Motivation

This problem was suggested to us by E. H. Kaplan (2006), who had previously introduced both network queueing models and associated fluid models to study tenant assignment policies in public housing (Kaplan 1984, 1985, 1986, 1988; Caldentey and Kaplan 2002). New housing applicants (prospective tenants) are the customers, whereas different housing authorities form the service pools, with the individual housing units playing the role of servers. Applicants fall into one of several different classes characterized by the housing authorities to which the applicant is eligible to apply. Unfortunately, these systems are often overloaded. Of course, that points to a need

Figure 1 Possible Network Models



Note. The first three pictures depict W, X, and $X+1$ models, whereas the last picture depicts a more complicated model that we are unable to analyze.

for additional public housing, but the supply may inevitably lag behind demand. With the prevailing overloaded systems, it may be helpful to have good models that make it possible to analyze the consequence of alternative policies. We might be able to investigate how much tenant abandonment probabilities and waiting times can be reduced by different actions. Indeed, the results of this paper can be applied for that purpose. For other recent analyses of public housing, see Johnson (2001, 2003) and references cited there.

Kaplan (2006) also suggested another application: the adoption of children. The customers are prospective parents seeking to adopt children, whereas the service pools are different adoption programs, with service rates determined by the arrival rates of children to be placed for adoption. The possible routing is determined by parental preferences (country of origin, gender, age, etc.) and parental participation in more than one waiting list (agency program, private lawyer, etc.). In addition, potential parents may renege for many reasons.

1.2. Routing Flow Rates

As soon as we introduce a multiclass, multi-server-pool fluid model of the kind above, we discover, as Kaplan did, that the routing flow rates $r_{i,j}$ are critical. We might directly specify these routing flow rates as part of the system design, but it is natural to ask what the routing flow rates would be for various scheduling rules. In particular, it is natural to consider the FCFS scheduling rule from fairness considerations. Kaplan specifically asked how the routing flow rates could be determined with FCFS scheduling.

Given any feasible set of routing rates yielding an overloaded fluid model (with all queues positive), we can analyze the queues for each customer class separately, using the methods for analyzing a single overloaded fluid queue with abandonment in Whitt (2006a). A network fluid model was analyzed this way in Whitt (2006b), but there the routing flow rates were exogenously specified, as part of the model construction. Here, in contrast, we want the routing flow rates to arise endogenously as a consequence of applying the FCFS assignment rule.

1.3. A Hierarchy of Complexity

We are able to determine the desired routing flow rates as well as the equilibrium behavior for a large subclass of the FCFS fluid models, but not for all possible fluid models. We identify a hierarchy of complexity among these systems based on the structure of the routing graph of the system. The easiest models to analyze are sparsely connected models, those with tree routing graphs, such as the W model, appearing first in Figure 1. (We follow the notation for routing topologies in service networks given in Garnett and Mandelbaum 2000.) Next in line are the fully connected models, those with complete bipartite routing graphs, such as the X model, shown second in Figure 1 above. Then, there are hybrid connected models, models that consist of a fully connected component with one or more sparsely connected components attached. The so-called $X+1$ model belongs to this class. Finally, there are models that do not fall into any of these three categories. We are not yet able to analyze those models. One example from each of the four cases appears in Figure 1.

Our analysis can be summed up as follows. In the sparsely connected case, if global FCFS is possible, flow rates can be obtained by using Equations (5)–(8). In the fully connected case, they can be determined by using Equation (16). In the hybrid case, the same equations can be used, but in a stepwise manner, first treating the sparsely connected components separately, and then treating the remaining fully connected component as above, but with appropriate new parameters.

1.4. Associated Fluid Limits

We use the fluid-model description to generate approximations for the corresponding queueing system. To make a clear connection between the queueing system and the fluid model, it is helpful to think of the fluid model as the limit of the queueing model as the arrival rate and the number of servers at each service-pool increase. Indeed, we conjecture that such a many-server heavy-traffic limit exists and is described by our fluid model, and hope to prove this in future work. Examples of such single-queue

fluid limits in this overloaded regime are contained in Whitt (2004, 2006a). Although we do not establish such a limit here, we are able to answer practical questions concerning the corresponding queueing system, assuming that the limit does hold. The fluid model can also be considered as a direct model or a direct approximation of the queueing model, without considering the limit, but the limit is helpful.

1.5. The Importance of Associated Stochastic Queueing Models

In fact, we discover what we believe is an important new phenomenon associated with fluid models, for which the limit plays an essential role. We show that the convergence of a sequence of queueing systems to the fluid model limit is itself important for determining unique routing flow rates $r_{i,j}$ in the fluid model. In other words, if we assume that the fluid model is a limit of appropriate queueing systems, satisfying general stochastic assumptions (to be specified), then we can identify unique routing flow rates $r_{i,j}$ in the fluid model. However, if we do not make that assumption, then the rates are not uniquely determined. This phenomenon occurs for fully connected models and hybrid models having a fully connected component.

1.6. Call-Center Models

The class of models we are considering is also relevant to the design and management of telephone call centers and related customer contact centers; see Gans et al. (2003). In the call-center literature, such systems are said to have *skill-based routing*. In models with skill-based routing, the routing policy used is crucial to the analysis of the model. Well managed call centers are usually not overloaded, but are more likely to be when the call center is service oriented as opposed to revenue-generating. We may also be interested in overloaded call centers in order to understand how they perform under exceptional circumstances, such as unanticipated high load or in the face of some system failures. Thus, overloaded models often prove to be useful in the call-center context as well.

A related paper in the call-center context, with references to many other recent related papers, is Gurvich and Whitt (2007). That paper establishes many-server heavy-traffic limits for the same type of queueing model, except routing is performed using a fixed-queue ratio (FQR) rule instead of FCFS. There it is observed that the same methods apply equally well to an analogous fixed-waiting ratio (FWR) rule, which reduces to FCFS as a special case. The motivation for the FQR and FWR routing policies is to perform service-level differentiation among the customer classes. That work differs from the present study in several respects: It focuses on the quality-and-efficiency-driven (QED) many-server heavy-traffic limiting regime, as in Halfin and Whitt (1981), rather than the efficiency-driven (ED) or overloaded many-server heavy-traffic limiting regime. It also restricts attention to Markovian models, although we also consider non-Markovian models. As in Whitt (2006a), the steady-state performance measures in this paper depend on general time-to-abandon cumulative distribution functions beyond their means.

1.7. Organization

We begin in §2 by introducing the queueing model we are approximating, and defining the classes of routing graphs we are able to analyze. Then, in §3, we introduce the corresponding fluid model. The fluid model is intended to arise as a limit of appropriately scaled queueing models as the arrival rates and number of servers increase. In this section we conjecture the existence and form of this limit. In §4, we go on to discuss stationary dynamics of the fluid model. It is here that we specify the important routing flow rates that must be determined and present a system of equations that they obey. Transient dynamics for our model are discussed in the e-companion[1] to this paper. In §§5–8, we discuss the fluid and queueing models in more detail for the special classes of routing graphs. In §5, we focus on the sparsely connected case. In §6, we focus on the fully connected case. In §7, we give an example to show that we must take into account the stochastic structure of the converging queueing models in order to compute the routing flow rates in the fully connected case. In §8, we show that it is easy to extend the analysis of the fully connected case to analyze the hybrid case. In §9, we show that our approximations for the routing flow rates are effective by making comparisons with simulations. Finally, in §10, we draw conclusions and discuss future work. Additional supporting material appears in the e-companion.

2. The Queueing Model

In our model there are sets $\mathscr{C} \equiv \{1, \ldots, n\}$ of customer classes and $\mathscr{S} \equiv \{1, \ldots, m\}$ of service pools ($\equiv$ here denotes equality by definition). We assume that there is a queue with unlimited capacity for each customer class. Arriving customers who cannot enter service immediately go to the end of the queue for their class, to be served thereafter in order of arrival. Throughout this paper, we use the FCFS routing policy.

Definition 1. We say that customers are routed according to the *FCFS routing policy* if, when a server becomes free and there are customers waiting from more than one customer class eligible for service by that server, the customer who entered the system first from the eligible classes is assigned to the freed server.

[1] An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

The definition leaves unspecified the way customers are assigned upon arrival to servers if there are idle servers in more than one eligible service pool. We are assuming that case will occur infrequently, and eventually not at all in the approximating fluid model.

2.1. Stochastic Model Elements

Figure 1 suggests a Markovian queueing model, but we actually consider more general queueing models. We define arrival processes $A_i \equiv \{A_i(t),\ t \geq 0\}$ with rates λ_i for each class i; i.e., A_i is a stochastic point process such that

$$\frac{A_i(t)}{t} \to \lambda_i \quad \text{as } t \to \infty \quad \text{with probability 1.} \tag{1}$$

We let $\lambda \equiv \{\lambda_i,\ i \in \mathscr{C}\}$.

Each class-i customer may elect to abandon the queue prior to starting service. We assume that these abandonment decisions for different customers are mutually independent, and independent of the arrival processes and service times. A class-i customer will abandon within time t of entering the system with probability $F_i(t)$, if the customer has not entered service by that time. We assume that F_i is a continuous cumulative distribution function (cdf), strictly increasing on its support, with $F_i(0) = 0$, mean $1/\theta_i$, and probability density function (pdf) f_i. Let $F \equiv \{F_i,\ i \in \mathscr{C}\}$.

We let the service-time cdf depend only on the service pool. We assume that the service times are mutually independent random variables independent of the arrival processes and the abandonment decisions. Let pool j have s_j servers, each with continuous service-time cdf G_j having mean $1/\eta_j$ and pdf g_j. Let $G \equiv \{G_j,\ j \in \mathscr{S}\}$, $\eta \equiv \{\eta_j,\ j \in \mathscr{S}\}$, and $s \equiv \{s_j,\ j \in \mathscr{S}\}$. The maximum possible total service rate for pool j is $\mu_j \equiv s_j\eta_j$ for all $j \in \mathscr{S}$.

2.2. The Routing Graphs

We assume that each class-i customer can be served by any server in a subset $S(i) \subseteq \mathscr{S}$ of the service pools. Let $C(j)$ denote the set of customer classes that can be served by service pool j; i.e., $C(j) \equiv \{i \in \mathscr{C}\colon j \in S(i)\}$, for all $j \in \mathscr{S}$. We characterize the allowed routing between customer classes and service pools by a *routing graph*. The nodes of the routing graph $\{1, \ldots, n-1, n, n+1, \ldots, n+m\}$ correspond to the customer classes in $\mathscr{C}$ and the service pools in $\mathscr{S}$. There is an arc $(i, n+j)$ in the graph if class i can be served by service pool j. Thus, the set of arcs in the routing graph is a subset of $\{1, 2, \ldots, n\} \times \{n+1, n+2, \ldots, n+m\}$. The routing graph is *bipartite* because all arcs connect customer classes to service pools.

We now characterize the classes of models we will be analyzing. We start by assuming that the bipartite graph is connected; i.e., it is not possible to decompose the bipartite graph into two unconnected components. If we could do so, then we could analyze each component separately. The first class of routing graphs we consider are acyclic connected bipartite graphs or trees. In these models, there are precisely $n + m - 1$ arcs connecting the n customer classes to the m service pools, without it being possible to decompose the system into two separate subsystems.

Definition 2. We say the model is *sparsely connected* if its routing graph is a tree.

The prototype of a sparsely connected model is the W model (see Figure 1).

The second class has routing graphs that are complete bipartite. In these models, each customer class can be served by any service pool.

Definition 3. We say a model is *fully connected* if its routing graph is complete.

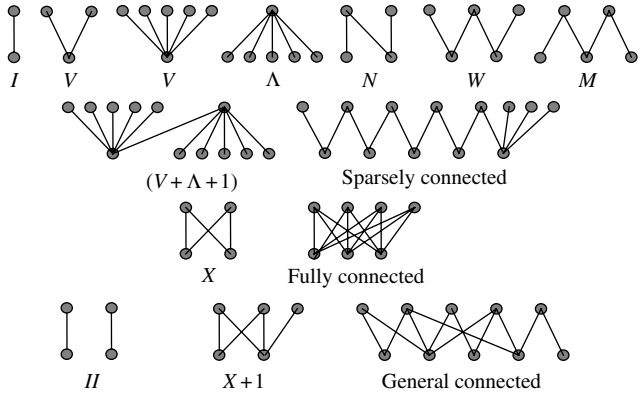
The prototype of a fully connected model is the X model.

We will also deal with models that are combinations of the sparsely and fully connected cases.

Definition 4. We say a model is a *hybrid* if its routing graph can be decomposed into one complete graph component and one or more tree component, such that the tree components are disjoint and each of them shares exactly one node with the complete component.

The $X+1$ model in Figure 1 is a hybrid model. Figure 2 shows a broader range of routing graphs. The routing graphs in the first row correspond to sparsely connected models, the I, V, Λ, N, W, and M models, respectively. The second row contains more complicated sparsely connected models. The third row contains two fully connected models, one being the X model and the other being a 4×3 fully connected model. The final row contains examples of three routing graphs that do not belong to the first two classes mentioned above. The first of these is the double-I model, which is disconnected. The second is the $X+1$ model, which is a hybrid model because it can be

Figure 2 Candidate Routing Graphs



Note. The first two rows contains sparsely connected routing graphs, the third row contains fully connected routing graphs, and the fourth row contains other routing graphs.

decomposed into an X model and an I model. We will discuss this model in more detail in §8. Finally, there is an example of a connected 5×5 model that does not belong to any of the three classes defined. Naturally we would like to treat models with such routing graphs, but we are currently unable to do so.

3. The Fluid Model

We are interested in analyzing deterministic fluid models corresponding to these queueing models. In this section, we discuss how these fluid models can arise as limits of sequences of appropriately scaled queueing models. We also define the notion of global FCFS for fluid models.

3.1. Fluid Limit

The fluid model is intended to arise as the limit of a sequence of queueing models indexed by r, where in model r we have arrival processes for each $i \in \mathcal{C}$ defined by $A_i^r(t) \equiv A_i(rt)$, $t \geq 0$, and each pool $j \in \mathcal{S}$ has rs_j servers. We regard r as a positive integer and let $r \to \infty$. The service times and abandonment times are left unchanged, independent of r.

Our first conjecture is a stochastic-process limit, as in Billingsley (1999), Chen and Yao (2001), and Whitt (2002). To express it, let $\Rightarrow$ denote convergence in distribution; let $D \equiv D([0,\infty), \mathbb{R})$ be the function space of all right-continuous real-valued functions on the interval $[0,\infty)$ with limits from the left everywhere in $(0,\infty)$, endowed with the usual Skorohod (J_1) topology; and let e be the identity function in D, i.e., $e(t) \equiv t$, $t \geq 0$. Let $D_2 \equiv D([0,\infty)^2, \mathbb{R})$ be the associated two-parameter function space (see Whitt 2006a). We will also consider product spaces, which are always understood to be endowed with the product topology; e.g., $D^n \equiv D([0,\infty), \mathbb{R})^n$.

We start by applying (1) to obtain a corresponding stochastic-process limit. The ordinary strong law of large numbers (SLLN) in (1) is actually equivalent to the stronger functional strong law of large numbers (FSLLN); e.g., see Theorem 5.10 of Chen and Yao (2001) and Theorem 5.3.2 of Whitt (2002) and the cited material in the Internet supplement to that book. We will state our conjecture in the weaker form $\Rightarrow$ (even though we think the stronger FSLLN actually does hold). We remark that convergence in distribution $\Rightarrow$ to a deterministic limit is equivalent to convergence in probability. We also remark that there is no equivalence between the weak LLN (WLLN) and its functional counterpart, the FWLLN. We obtain the desired FWLLN through the implications: SLLN $\Rightarrow$ FSLLN $\Rightarrow$ FWLLN. Hence, starting from (1), we deduce that a scaled version of $A \equiv \{A_i,\ i \in \mathcal{C}\}$ satisfies the FWLLN

$$\bar{A}^r \equiv \frac{A^r(\cdot)}{r} \equiv \frac{A(r\cdot)}{r} \ \Rightarrow\ \lambda e(\cdot) \quad \text{in } D^n \text{ as } r \to \infty. \tag{2}$$

We describe the system content in model r using the following stochastic processes:

- $B^r \equiv \{B_{i,j}^r,\ i \in \mathcal{C}, j \in \mathcal{J}\}$, where $B_{i,j}^r(t,y)$ is the number of class i customers in service in pool j at time t that have been in service for time less than or equal to y.
- $Q^r \equiv \{Q_i^r,\ i \in \mathcal{C}\}$, where $Q_i^r(t,y)$ is the number of class i customers in queue at time t that have been in queue for time less than or equal to y.

Then, form the scaled processes $\bar{B}^r \equiv \{\bar{B}^r(t,y),\ t \geq 0, y \geq 0\}$ and $\bar{Q}^r \equiv \{\bar{Q}^r(t,y),\ t \geq 0, y \geq 0\}$ by

$$\bar{B}^r(t,y) \equiv \frac{B^r(t,y)}{r} \quad \text{and} \quad \bar{Q}^r(t,y) \equiv \frac{Q^r(t,y)}{r} \quad \text{for } t \geq 0,\ y \geq 0,\ \text{and } r \geq 0.$$

Also, to further describe the dynamics of the system, we will need to keep track of how customers are routed and virtual waiting times at the queues. For model r, let

- $R^r \equiv \{R_{i,j}^r,\ i \in \mathcal{C}, j \in \mathcal{S}\}$, where $R_{i,j}^r(t)$ is the number of class i customers routed to pool j by time t.
- $W^r \equiv \{W_i^r,\ i \in \mathcal{C}\}$, where $W_i^r(t)$ is the virtual waiting time of a class i customer at time t.

Then, just as for B^r and Q^r, form the scaled processes $\bar{R}^r \equiv \{\bar{R}^r(t),\ t \geq 0\}$ by

$$\bar{R}^r(t) \equiv \frac{R^r(t)}{r}, \quad t \geq 0,\ r \geq 0.$$

We do not have to scale the virtual waiting times W^r because our fluid scaling does not involve time scaling of the content stochastic process.

A discussion of the transient dynamics of the limiting fluid model is given in the e-companion. Here, we will only be interested in stationary dynamics of the model. We assume that the limiting models are initially empty. However, we believe this assumption can be generalized to appropriate convergence of the initial conditions.

Conjecture 1 (Fluid Limit). *Consider an initially empty queueing model of the type described in* §2 *and construct a sequence of fluid-scaled versions of this model indexed by positive integers r as indicated above. Then,*

$$(\bar{B}^r, \bar{Q}^r, \bar{R}^r, W^r) \ \Rightarrow\ (B, Q, R, W) \tag{3}$$

in $D_2^{nm+n} \times D^{nm+n}$ as $r \to \infty$, where B and Q are continuous deterministic functions in (t,y) and R and W are continuous deterministic functions in t, consistent with the description of the transient dynamics given in §1 *of the e-companion.*

The deterministic functions $B(t,y)$, $Q(t,y)$, $R(t)$, and $W(t)$ serve as fluid approximations for the scaled stochastic processes $\bar{B}^r(t,y)$, $\bar{Q}^r(t,y)$, $\bar{R}^r(t)$, and the

nonscaled process $W^r(t)$ so that

$$\begin{aligned} B^r(t,y) &\approx rB(t,y), \qquad Q^r(t,y) \approx rQ(t,y), \\ R^r(t) &\approx rR(t), \qquad W^r(t) \approx W(t) \end{aligned} \tag{4}$$

for $t \geq 0$, $y \geq 0$, and r large.

3.2. Global FCFS

When going from the queueing model to the fluid model, we would expect that the FCFS routing policy of the limiting queueing models would imply some notion of FCFS for the fluid model. In fact, in the fluid model we can hope to achieve a stronger notion of FCFS scheduling: global FCFS. In the queueing model, global FCFS means that all customers enter service in order of arrival. For queueing models in which servers do not sit idle when there is a waiting customer they can serve, global FCFS is only possible for fully connected models. However, we can hope to achieve global FCFS in fluid models with more general structure. Global FCFS holds in the fluid model if when one atom of fluid arrives to the system before a second atom of fluid, it starts service before the second atom of fluid. Formally, we have the following:

Definition 5. We say that *global FCFS* holds in the fluid model if, for all $0 \leq t_1 < t_2$ and for all $i, j \in \mathscr{C}$, we have

$$t_1 + W_i(t_1) < t_2 + W_j(t_2).$$

A key (intuitive) observation for our analysis is that this global FCFS condition for the fluid model is actually equivalent to the wait times $W_i(t)$ not depending on i.

Lemma 1. *For the fluid model described above, with its evolution described by continuous functions, the specified global FCFS condition is equivalent to the existence of a function* $W\colon [0,\infty) \to \mathbb{R}$ *such that* $W_i(t) = W(t)$ *for all* $i \in \mathscr{C}$ *and* $t \geq 0$.

Proof. The implication of global FCFS from the existence of a single waiting-time function is immediate from the FCFS property of each individual queue. Assuming that global FCFS holds, we will show that unequal waiting times leads to a contradiction. Thus, suppose $W_i(t) > W_j(t) + \epsilon$ for some $i, j \in \mathscr{C}$, $t \geq 0$, and $\epsilon > 0$. By the definition of $W_i(t)$, class i fluid that enters the system at time t, enters service at time $t + W_i(t)$. By the continuity of W_j, we must have

$$\lim_{\delta \downarrow 0} W_j(t+\delta) + \delta = W_j(t),$$

so that for $0 < \eta < \epsilon$ there exists δ small enough so that $W_j(t+\delta) + \delta - W_j(t) < \eta < \epsilon$. Therefore, class j customers entering the system at time $t + \delta$, after a class i customer entering at time t, enters service at time

$$t + \delta + W_j(t+\delta) < t + W_j(t) + \eta < t + W_j(t) + \epsilon < t + W_i(t),$$

contradicting the FCFS assumption. Therefore, we must have $W_i(t) = W_j(t)$ for all $i \neq j$. □

We will want to find stationary routing rates yielding global FCFS. However, it is not guaranteed that we can achieve global FCFS in the limiting fluid model, even when FCFS scheduling is used in the corresponding prelimit queueing models. When global FCFS cannot be achieved, the waiting times of different classes cannot all be identical. We believe for that to happen in a connected model, one or more flows must be zero, which effectively changes the network. In particular, it implies that some arcs can be removed, so that the network actually can be decomposed into two or more subnetworks, with global FCFS prevailing in each. In this paper, we do not fully analyze what happens when global FCFS fails to occur, because we do not determine which arcs need to be removed, but we provide insight. Our analysis helps identify appropriate decompositions.

4. Stationary Dynamics of the Overloaded Fluid Model

We now describe the stationary behavior of the fluid model. Except for the routing flow rates, the story is the same as in Whitt (2006a), so we are primarily interested in the routing flow rates. We assume our fluid models are overloaded in stationarity.

Definition 6. A fluid model is said to be *overloaded in stationarity* if in stationarity all queues are nonempty, which in turn implies that all the service pools are working at full capacity.

This overloaded property can be determined from the stationary equations of the FCFS fluid model, to be developed below. A necessary condition for this property is $\sum_{i=1}^{n} \lambda_i > \sum_{j=1}^{m} \mu_j$.

We now define quantities of interest for the stationary fluid model. Let $B_{i,j}(y)$ be the amount of class i fluid in service pool j that has been in service for time less than or equal to y, and $Q_i(y)$ the amount of class i fluid in queue that has been in queue for time less than or equal to y. We assume that the functions $B_{i,j}(y)$ and $Q_i(y)$ have densities $b_{i,j}$ and q_i, so that

$$B_{i,j}(y) = \int_0^y b_{i,j}(u)\,du \quad \text{and} \quad Q_i(y) = \int_0^y q_i(u)\,du$$
$$\text{for all } i \in \mathscr{C},\ j \in \mathscr{S},\ \text{and } y \geq 0.$$

Also, let $B_{i,j} \equiv B_{i,j}(\infty)$ be the total amount of class i fluid in pool j and let $Q_i \equiv Q_i(\infty)$ be the total amount of class i fluid in queue for $i \in \mathscr{C}$, $j \in \mathscr{S}$. Furthermore, define the following quantities:

$\alpha_i \equiv$ the rate at which class i fluid abandons the queue,

$\nu_i \equiv$ the total rate at which class i fluid enters service,

$r_{i,j} \equiv$ the rate at which class i fluid enters service at pool j (*routing flow rates*),
$\sigma_{i,j} \equiv$ the rate at which class i fluid is served by pool j,
$p_{i,j} \equiv$ the proportion of class i fluid served by pool j,
$w \equiv$ the common waiting time for the system

for all $i \in \mathscr{C}$, $j \in \mathscr{S}$. So far, we do not know if these quantities exist. It turns out that if we assume $r_{i,j}$ exists for $i \in \mathscr{C}$, $j \in \mathscr{S}$, and we assume the conjectures of Whitt (2006a) concerning the single-customer-class single-service-pool case, then we can analyze each individual queue and service pool in our model separately using Theorem 3.1 of Whitt (2006a), implying existence of the rest of the quantities defined above.

Conjecture 2. *The quantities* $r_{i,j} \equiv \lim_{t\to\infty} R_{i,j}(t)/t$ *exist as proper limits for all* $i \in \mathscr{C}$, $j \in \mathscr{S}$.

From now on, we assume Conjectures 1 *and* 2 *here as well as Conjectures* 2.1 *and* 2.2 *of Whitt* (2006a).

Lemma 2. *If the fluid model is overloaded in stationarity, then all rates defined above exist. Furthermore, the fluid model has a unique steady state where each individual queue* $i \in \mathscr{C}$ *and each individual service pool* $j \in \mathscr{S}$ *satisfies the relations of Theorem* 3.1(b) *of Whitt* (2006a) *with* ρ *there set to* λ_i/ν_i.

Proof. By the existence of the quantities $r_{i,j}$, $i \in \mathscr{C}$, $j \in \mathscr{S}$, $\nu_i = \sum_{j\in C(i)} r_{i,j}$ must exist for each $i \in \mathscr{C}$. For each individual queue $i \in \mathscr{C}$, our fluid limit Q_i in (3) is constructed exactly like the fluid limit of the queueing process in Whitt (2006a) except, for each individual queue, customers enter the queue at rate $r\lambda_i$ and leave the queue at the rate $r\nu_i$. By the overloaded in stationarity assumption, each queue is overloaded so that $\rho = \lambda_i/\nu_i > 1$ and the relations of Theorem 3.1(b) of Whitt (2006a) involving the queueing process hold with $\rho = \lambda_i/\nu_i$. Similarly, each $B_{i,j}$, $i \in \mathscr{C}$, $j \in \mathscr{S}$, in (3) can be analyzed separately. □

Using this lemma, we can now characterize our stationary performance measures in terms of the rates $r_{i,j}$, $i \in \mathscr{C}$, $j \in \mathscr{S}$, and the common waiting time w. As in Whitt (2006a), not all the model elements from the original queueing model remain relevant in the fluid model. For stationary behavior of the fluid processes (except for $b_{i,j}(x)$, which is not crucial), the relevant model elements are (λ, μ, F). In particular, the arrival processes A appear only in the fluid model via the rates λ, and the service-time cdfs G and the quantities s appear only via $\mu_j = s_j\eta_j$, $j \in \mathscr{S}$, but the time-to-abandon cdfs F_i beyond their means $1/\theta_i$ remain relevant. For a cdf F_i, let $F_i^c \equiv 1 - F$ be the associated complementary cdf.

Theorem 1. *Given a fluid model described by* (λ, G, F, s), *it is overloaded in stationarity and global FCFS holds if and only if for* $i \in \mathscr{C}$, $j \in \mathscr{S}$, *there exist numbers* $r'_{i,j} \geq 0$, $\nu'_i > 0$, *and* $w' > 0$ *satisfying the following three sets of equations. First, we have the* n global FCFS equations

$$\lambda_i F_i^c(w') = \nu'_i \quad \textit{for all } i \in \mathscr{C}. \tag{5}$$

Second, we have the single flow-conservation equation

$$\sum_{i=1}^{n} \nu'_i = \sum_{j=1}^{m} \mu_j. \tag{6}$$

Third, we have the $n + m$ flow-rate equations

$$\sum_{i\in C(j)} r'_{i,j} = \mu_j \quad \textit{for all } j \in \mathscr{S}, \tag{7}$$

$$\sum_{j\in S(i)} r'_{i,j} = \nu'_i \quad \textit{for all } i \in \mathscr{C}. \tag{8}$$

Given that these equations have a solution, w' *and* ν'_i *will always be uniquely determined. Also, there will always exist* $r'_{i,j}$, $i \in \mathscr{C}$, $j \in \mathscr{S}$, *satisfying the equations, but they will not necessarily be uniquely determined and nonnegative. If a solution exists for which these quantities are in fact nonnegative, then* $\nu_i = \nu'_i$, $w = w'$ *for each* $i \in \mathscr{C}$, *and*

$$\alpha_i = \lambda_i - \nu_i, \tag{9}$$

$$\sigma_{i,j} = r_{i,j}, \qquad p_{i,j} = \frac{r_{i,j}}{\sum_{k\in S(i)} r_{i,k}}, \tag{10}$$

$$b_{i,j}(x) = \frac{r_{i,j}}{s_j} G_j^c(x), \quad x \geq 0, \tag{11}$$

$$q_i(x) = \begin{cases} \lambda_i F_i^c(x) & 0 \leq x \leq w, \\ 0 & x > w, \end{cases} \tag{12}$$

$$B_{i,j} = \frac{r_{i,j}}{\mu_j}, \tag{13}$$

$$Q_i = \int_0^\infty q_i(x)\,dx = \lambda_i \int_0^w F_i^c(x)\,dx \tag{14}$$

for all $i \in \mathscr{C}$ *and* $j \in \mathscr{S}$. *If, in addition, the quantities* $r'_{i,j}$, $i \in \mathscr{C}$, $j \in \mathscr{S}$ *are uniquely determined, then we also have* $r_{i,j} = r'_{i,j}$, $i \in \mathscr{C}$, $j \in \mathscr{S}$.

Proof. Assume that the fluid model is overloaded in stationarity and admits global FCFS. Then the equations are satisfied by the quantities $r'_{i,j} = r_{i,j}$, $\nu'_i = \nu_i$, and $w' = w$, the common wait time at all the queues, which exists by Lemma 1: Equation (5) holds by Lemma 2 and (3.9) of Whitt (2006a), (6) holds simply by conservation of flow, and (7)–(8) hold by local balance of flow at each queue and each service pool.

Conversely, suppose the equations have a solution $r'_{i,j} \geq 0$, $\nu'_i > 0$, $w' > 0$ for $i \in \mathscr{C}$, $j \in \mathscr{S}$. Then, it is possible to route fluid from the queues to the service pools in such a way that wait times at all queues will be w'. This implies that the fluid model admits global FCFS. Since $w' > 0$, the model is overloaded in stationarity. Furthermore, combining (5) with (3.9) of Whitt (2006a), for each $i \in \mathscr{C}$ we will have $\nu'_i = \nu_i$.

We now argue that when the flow equations have a solution, w' and ν'_i are uniquely determined. We can combine Equations (5) and (6) to obtain the single *waiting-time equation*

$$\sum_{i=1}^{n} \lambda_i F_i^c(w') = \sum_{j=1}^{m} \mu_j. \tag{15}$$

Consider the left side of this equation as a function of w'. Notice that this function is continuous and strictly decreasing when it is positive because we assumed that all the abandonment cdfs F_i are continuous and strictly increasing on their support. When $w' = 0$, by the overloaded in stationarity assumption, we have $\sum_{i=1}^{n} \lambda_i > \sum_{j=1}^{m} \mu_j$. Also, as $w' \to \infty$, the left side approaches zero. Therefore, by the intermediate value theorem, there must exists some $w' > 0$ such that the equation holds.

If a solution to (5)–(8) exists for which $r'_{i,j} \geq 0$, $i \in \mathscr{C}$, $j \in \mathscr{S}$, then by Lemma 2 we must have (9)–(14) for the true $r_{i,j}$ given by Conjecture 2. The $r'_{i,j}$ do not appear in these equations unless they are uniquely determined by the flow-rate equations. □

The focus of the rest of the paper is to compute the flow rates $r_{i,j}$, $i \in \mathscr{C}$, $j \in \mathscr{S}$ and, thus determine stationary dynamics of the fluid model. We are able to do this in the three cases:

1. The model is sparsely connected.
2. The model is fully connected with nonlattice service-time distributions with no mass at zero.
3. The model is hybrid connected with nonlattice service-time distributions with no mass at zero.

In the next section, we show that our system of stationary fluid equations has a unique solution in the sparsely connected case (but not necessarily nonnegative!). In §§6 and 7, we show that in other cases we *must* make further assumptions on the stochastic elements of the queueing model to determine stationary behavior of the corresponding fluid model. The flow rates $r_{i,j}$ are not uniquely determined by the fluid model alone.

5. The Sparsely Connected Case

We now show that we can uniquely determine the stationary dynamics of the fluid model when the model is sparsely connected.

Theorem 2 (Sparsely Connected Models). *For sparsely connected fluid models that are overloaded in stationarity, the system of stationary fluid equations in* (5)–(8) *has a unique solution for w', ν'_i, and $r'_{i,j}$ with $w' > 0$ and $\nu'_i > 0$ for all $i \in \mathscr{C}$. If, in addition, $r'_{i,j} \geq 0$ for all $i \in \mathscr{C}$, $j \in \mathscr{S}$, then the corresponding fluid model has a unique stationary solution satisfying global FCFS, where $w = w'$, $\nu_i = \nu'_i$, and $r_{i,j} = r'_{i,j}$ for $i \in \mathscr{C}$, $j \in \mathscr{S}$.*

Proof. As indicated in the proof of Theorem 1, for models that are overloaded in stationarity, we can first use (15) to uniquely solve for w', and thus ν'_i, $i \in \mathscr{C}$. We are then left with the linear system (7)–(8). In the sparsely connected case, the routing graph is a tree, so that as in Proposition A.2 of Atar (2005) we can solve for the flow rates uniquely as follows. Because the routing graph is a tree, it contains at least one leaf. If the leaf is a customer class $i \in \mathscr{C}$ and its adjacent edge is (i, j), then let $r'_{ij} = \nu'_i$. If the leaf is a service pool $j \in \mathscr{S}$ and its adjacent edge is (j, i), then let $r'_{ij} = \mu_j$. Then strip off the leaf and its incident edge and recurse. Finally, the last claim follows from the last sentence of Theorem 1. □

It may happen that one or more of the unique routing rates $r'_{i,j}$ is negative. In that case, by Theorem 1 we can conclude that it is not possible to achieve a solution that is overloaded in stationarity and exhibits global FCFS. We do not carefully analyze the case when negative flows occur. It is our experience that when there is a single negative flow, then we can describe the network behavior by replacing this flow by 0, which is equivalent to removing the arc from the routing graph. If the model is sparsely connected, this action always produces two separate sparsely connected models, which then can be analyzed in precisely the same way.

When there are multiple negative flows, the situation is more complicated. We have found examples where more than one of these should be set equal to 0, but also examples where only one should be set equal to 0. We do not yet have a systematic way to determine which case prevails and, thus, what actually happens in the network. Right now, we would rely on simulation to confirm that we have the correct reduced network with global FCFS in each separate sparsely connected component. It is our experience that, in all cases, the actual behavior can be captured by such reductions; it only remains to be determined which reduction is the correct one.

The negative flow rates also tell us that, if we want to achieve global FCFS, then we need to redesign the network. One approach is to eliminate one or more arcs, by fiat as a redesign decision, by setting routing flow rates equal to 0, giving us two separate sparsely connected components. Then, we may possibly achieve global FCFS in each component, but way may need to make further reductions. Indeed, by this recursive network reduction, we will necessarily arrive at a collection of sparsely connected subnetworks, each of which exhibits global FCFS, provided that all the subnetworks are overloaded. The analysis above provides guidance for the network redesign.

6. The Fully Connected Case

In the fully connected case, customers in each class can be served by servers in any pool. In this case, after solving (15), our system of stationary flow-rate equations (7)–(8) has $nm + 1$ unknowns and $n + m$

equations. When either $n = 1$ or $m = 1$, the model is also sparsely connected, so we can apply the method of the previous section. When $n > 1$ and $m > 1$, we have $nm + 1 > n + m$, so the analysis in the previous section will not go through; we have extra degrees of freedom. To resolve the additional degrees of freedom, we exploit stochastic properties of the queueing model in addition to the fluid model. In fact, we will show in the next section that we *must* consider stochastic elements of the queueing model.

Assuming that the fluid model is indeed overloaded in stationarity, all queue lengths are strictly positive, so that all service pools are processing at maximum rate. Therefore, for large r and t, after time t, all service pools in models with an index greater than r are almost always busy (but never all busy with probability 1). We will carry out an analysis of the queueing model based on the assumption that all servers are in fact *always* busy. This significantly simplifies analysis.

Even though we scale the queueing models indexed by r in order to relate them to the limiting fluid model, we will want to look at the departure processes of the queueing model in the time scale of individual departures. That means that we must dilate time by a factor of r. (For a previous asymptotic analysis with such time dilation, see Whitt 1984.) Let $D_j^r(t)$ be the number of departures from service pool j by time t, assuming that all servers are always busy. Then, we consider the processes $\widetilde{D}_j^r(t) \equiv D_j^r(t/r)$, as $r \to \infty$. We first discuss the case of exponential service-time distributions, and then generalize.

6.1. Exponential Service Times

Let each service pool contain servers having a common exponential service-time distribution. Because we assume that all servers are always busy, pool j in model r of the sequence of fluid scaled models has a Poisson departure process with intensity $r\mu_j = r\eta_j s_j$. Therefore, the dilated departure processes $\widetilde{D}_j^r(t)$ are Poisson processes with intensity $\mu_j = \eta_j s_j$.

Lemma 3 (Busy Exponential Service Pools). *For pools of permanently busy servers, each with an exponential service distribution with mean* $1/\eta_j$, *the dilated process* $\widetilde{D}_j^r(t)$ *is a Poisson process with intensity* μ_j *for each* $r \geq 0$.

Now consider the customer who has been waiting in the queue the longest. Because we have a fully connected model, that customer can be served by any service pool. The customer will enter service in the service pool where a server frees up first. Because all residual service times are exponential, the probability that the customer ends up being served at a pool $j \in \mathcal{S}$ is $\mu_j / \sum_{k=1}^m \mu_k$. By the memoryless property, the pools where customers get served are mutually independent random variables. Therefore, by the SLLN, the proportion of class-i customers that get served at pool j must be $\mu_j / \sum_{k=1}^m \mu_k$.

Now, let w be the solution to the waiting-time equation (15) for our system. Then, the rate that fluid leaves queue $i \in \mathcal{C}$ for service is $\nu_i = \lambda_i F_i^c(w)$ and we have

$$r_{i,j} = \frac{\mu_j}{\sum_{k=1}^m \mu_k} \nu_i \quad \text{for all } i \in \mathcal{C},\ j \in \mathcal{S}. \tag{16}$$

Notice that this solution (16) satisfies our system of stationary fluid equations (7)–(8) with $r_{i,j} > 0$ for all i and j.

6.2. General Service Times

We now argue that the result in the previous section extends to the more general case of i.i.d. nonlattice service times with no point mass at zero, in the limit as $r \to \infty$. We do this by showing that the stationary departure process of each service pool $j \in \mathcal{S}$ in the time scale of individual departures is *asymptotically* Poisson with rate μ_j.

In model r, the departure process D_j^r, $j \in \mathcal{S}$, can be represented as the superposition of rs_j departure processes $D_{j,k}$, with one for each of the servers in the service pool:

$$D_j^r(t) = \sum_{k=1}^{rs_j} D_{j,k}(t).$$

In general, for different values of r, the sample paths of the departure processes will differ, but the distributions of the departure processes will not. Because we have assumed that the servers are permanently busy, each of these individual departure processes is a (possibly delayed) renewal process with inter-renewal-time cdf G_j for all r. Because the distribution of G_j has been assumed to be nonlattice, each of these renewal processes approaches the associated equilibrium renewal process as time t increases, by Theorem V.4.3 (4.6) of Asmussen (2003, p. 155). Hence, in stationarity, the rs_j departure processes are i.i.d. stationary point processes (with stationary increments), each having rate η_j. Hence, we can apply the classical limit theorem for sums of i.i.d. point processes with stationary increments; see Chapter 5 of Khintchine (1960) and Proposition 9.2.VI of Daley and Vere-Jones (1988). Our assumption of no point mass at zero implies that we have unit jumps in the individual departure processes $D_{j,k}$. This gives us

Lemma 4 (Busy General Service Pools). *For pools of permanently busy servers, each with nonlattice service-time cdf* G_j *having no point mass at zero and with mean* $1/\eta_j$, *we have*

$$\widetilde{D}_j^r(\cdot) \equiv \sum_{k=1}^{rs_j} D_{j,k}(\cdot/r) \ \Rightarrow\ \Pi_j(\cdot) \quad \text{as } r \to \infty,$$

where Π_j *is a Poisson process with intensity* μ_j.

Because in the time scale of individual departures the departure process of each service pool $j \in \mathcal{S}$ is

asymptotically Poisson with intensity μ_j, our analysis from §6.1 applies, giving us

THEOREM 3. *If the service-time cdfs G_j are nonlattice with no point mass at zero, and if the servers are permanently busy, then the routing flow rates in the fluid model are as given in* (16).

Overall, this reasoning must be regarded as heuristic, because we have assumed that the servers are always busy. However, simulation experiments confirm that this reasoning does indeed produce the correct fluid flow rates. We will see in the next section that the stochastic conditions on the service-time distributions cannot be eliminated altogether.

7. Dependence Upon the Stochastic Models

In this section we demonstrate that the behavior of the fully connected fluid model depends on the stochastic structure of the sequence of prelimit queueing models that converge to the fluid model. For that purpose, we present a sequence of deterministic queueing models indexed by r that would converge to a symmetric X fluid model described by (16) as $r \to \infty$ if they satisfied the stochastic assumptions of Theorem 3, but which actually admit an asymmetric solution. Note that the deterministic queueing models in our example are simply special cases of the more general stochastic queueing models we are studying.

Our fluid model will have two customer classes and two server pools. Routing is possible from each customer class to each pool, so that we may have $r_{i,j} > 0$ for $i = 1, 2$ and $j = 1, 2$. Let the fluid arrival rates be $\lambda_1 = \lambda_2 = 2$ and the fluid service rates be $\mu_1 = \mu_2 = 2$. This makes the model symmetric in i and in j. Because the total input rate is equal to the total output rate, we are actually in the QED heavy-traffic regime (for which this result is also of interest!), but we can easily put ourselves in the overloaded ED regime by adding deterministic abandonment.

Under the previous minimal stochastic assumptions for the service times in §6, we have the symmetric solution to this fluid model and the rates become $r_{1,1} = r_{1,2} = r_{2,1} = r_{2,2} = 1$. However, we can actually construct prelimit models achieving

$$r_{1,1} = r_{2,2} = 1 + \epsilon \quad \text{and} \quad r_{1,2} = r_{2,1} = 1 - \epsilon \quad \text{for any } \epsilon,\ -1 \le \epsilon \le 1, \tag{17}$$

which is an asymmetric solution for $\epsilon \neq 0$. Therefore, *the extra degrees of freedom in the system of linear equations in Theorem* 1 *can actually be exploited. And the resolution depends on extra structure beyond the limiting fluid model itself*. So the argument we gave in §6, starting from the stochastic model is actually necessary.

We now proceed to construct a sequence of queueing models indexed by r for which (17) holds with $\epsilon = 1$. From our construction, it will be evident that constructions can be made for any other ϵ in the specified range as well. In our queueing models, let all the service times be deterministic of length 1. In model r, let each service pool have $2r$ servers. Thus, the maximum possible long-run average total service rate is $2r$ for each service pool in model r, but we will scale by dividing by r, so that the total average fluid service rate from each service pool $j = 1, 2$ becomes $\mu_j = 2$.

Similarly, we will let the arrival rate in model r for each customer class be $2r$, making the fluid arrival rate $\lambda_i = 2$ after dividing by r, again as already stipulated for the fluid model. We will choose a special deterministic batch arrival process. We let arrivals occur in batches of size 2 with intervals between successive batches deterministic with length $1/r$ to ensure that the long-run average arrival rate is $2r$. But we make the two arrival processes for the two customer classes be asynchronous. In particular, in model r we let class-1 arrivals occur at times $(k+1)/r$, $k \ge 0$, but we let class 2 arrivals occur at times $(2k+1)/(2r)$, $k \ge 0$. As stated above, a batch of two arrivals comes at each arrival epoch. Note that the arrival epochs for batch arrivals in the two arrival processes alternate.

We also consider special initial conditions. We suppose that all servers are initially busy at time 0, but the two queues are initially empty. Moreover, we suppose that the $2r$ servers in service pool 1 free up two at a time at the times

$$\delta_r, \frac{1}{r} + \delta_r, \frac{2}{r} + \delta_r, \ldots, \frac{r-1}{r} + \delta_r, \tag{18}$$

whereas the $2r$ servers in service pool 2 free up two at a time at the times

$$\frac{1}{2r} + \delta_r, \frac{3}{2r} + \delta_r, \frac{5}{2r} + \delta_r, \ldots, \frac{2r-1}{2r} + \delta_r \tag{19}$$

for some constant δ_r with $0 < \delta_r < 1/(2r)$ for all r. For example, we could have $\delta_r = 1/(4r)$.

With these definitions, all customers from class 1 are served by servers from service pool 1, whereas all customers from class 2 are served by servers from service pool 2. The servers free up in pairs from alternating service pools. Following the requirement of the FCFS service discipline, these pairs of servers will serve the two customers that have been waiting the longest, but the customers that have been waiting the longest will also alternate from customer class to customer class, in a way that guarantees that servers from pool 1 always select customers from class 1, whereas servers from pool 2 always select customers from class 2. Moreover, in model r, all customers spend exactly time $w_r = \delta_r$ in queue so in the fluid limit we have $w = 0$.

Note that both batch arrivals and batch service can be relaxed by making small perturbations of the arrival times and service times. We thus could have an example with arrivals and service both occurring one at a time, but we would get the same assignments as above. Even without the batches, the conditions for the theorems in §6 are not satisfied. In particular, the example exploits lattice structure in the service-time distribution. For this model, the servers are always busy, but they do not eventually behave as the superposition of i.i.d. stationary point processes.

In summary, we constructed a sequence of special prelimit queueing models, which converges to the initial fluid model as $r \to \infty$, but for which the routing flow rates are not the unique solution we get for queueing models satisfying the assumptions of Theorem 3. We have thus proven that extra degrees of freedom in the routing flow rates cannot be resolved from the fluid model itself. They must be resolved by stochastic properties of the prelimit queueing models indexed by r.

8. The Hybrid Case

We now observe that we can combine the methods in §§5 and 6 to analyze hybrid models, as defined in §2. We analyze hybrid models by first treating the sparsely connected components, obtaining the unique solution for the flow rates of each. If these flow rates are nonnegative for each sparsely connected component, we subtract the committed flow from the connecting arc to the fully connected component. After we have done that for all sparsely connected components, we solve for the flows in the remaining fully connected model. If we get any negative flow rates when analyzing the sparsely connected components, we can conclude that the model is not overloaded in stationarity with global FCFS, by Theorem 1. For the final analysis of the fully connected component, we assume that the relevant service-time distributions are nonlattice without an atom at zero, so that we get the unique solution in (16); again, otherwise the model is not overloaded in stationarity with global FCFS.

We illustrate this technique with the $X+1$ model with exponential service as shown in Figure 1. The idea is that the allocation for class 3 can be easily determined. Then after removing that predetermined flow, we can solve the remaining X model as in the previous section. Note that the waiting time w can be determined using the conservation of flow equation. Then, we must have $r_{32} = \nu_3 \equiv \lambda_3 F_3^c(w)$. Given that $r_{32} = \nu_3$, the pool 2 service rate available for classes 1 and 2 becomes $\mu_3 - \nu_3$. Therefore, reasoning as for the X model, we get

$$r_{11} = \frac{\mu_1\nu_1}{\mu_1 + (\mu_2 - \nu_3)}, \quad r_{12} = \frac{(\mu_2 - \nu_3)\nu_1}{\mu_1 + (\mu_2 - \nu_3)},$$

$$r_{21} = \frac{\mu_1\nu_2}{\mu_1 + (\mu_2 - \nu_3)}, \quad \text{and} \quad r_{22} = \frac{(\mu_2 - \nu_3)\nu_2}{\mu_1 + (\mu_2 - \nu_3)}.$$

Note that this solution satisfies the steady-state equations (7)–(8).

This approach can easily be extended to more general hybrid connected models. Figure 3 illustrates the technique applied to slightly more complicated models.

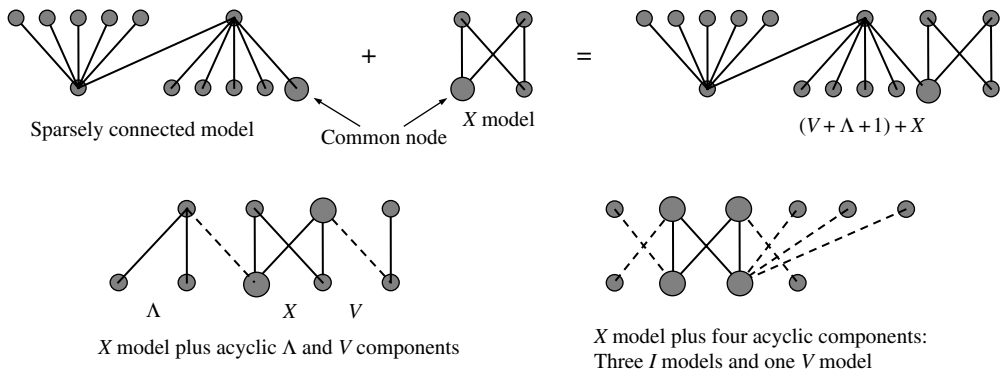
9. Simulation Results

In this section, we compare our fluid approximations to simulated values. First, we do this for a hybrid model, then for the X model with many different combinations of interarrival and service-time distributions, and finally for sparsely connected models yielding negative flows.

9.1. A Hybrid Model

Because computing the fluid approximation for a hybrid model requires computation of fluid approximations for both sparsely and fully connected components, presenting simulation results for a hybrid model helps validate our method for all three classes

Figure 3 More General Hybrid Connected Models



Note. The first hybrid contains one sparsely connected component connected to the fully connected component, the second contains two sparsely connected components, and the third contains four sparsely connected components.

Table 1 Simulation Results for Hybrid Model with Exponential and Uniform Service-Time Distributions

Route	Fluid approx.	*EXP*	*UNIFORM*
$p_{1,1}$	0.4667	0.4652 ± 0.0058	0.4623 ± 0.0083
$p_{1,2}$	0.4667	0.4682 ± 0.0084	0.4678 ± 0.0076
$p_{1,3}$	0.0667	0.0664 ± 0.0115	0.0698 ± 0.0107
$p_{2,3}$	0.4615	0.4617 ± 0.0134	0.4612 ± 0.0089
$p_{2,4}$	0.5385	0.5382 ± 0.0134	0.5387 ± 0.0089
$p_{3,3}$	0.2692	0.2691 ± 0.0124	0.2714 ± 0.0073
$p_{3,4}$	0.3141	0.3164 ± 0.0049	0.3164 ± 0.0052
$p_{3,5}$	0.4167	0.4144 ± 0.0141	0.4120 ± 0.0073
$p_{4,5}$	1	1	1

of routing graphs. Therefore, we will be working with a queueing model with the routing graph in the bottom left of Figure 3 and approximating its dynamics with the corresponding fluid model. We simulated the model with both exponential and uniform service-time distributions. For both cases, we simulated Poisson arrivals and exponential abandonment. We simulated 100,000 arrivals to the system, disregarding an initial transient 20% of the run. Labelling the nodes in the figure from left to right, the remaining model parameters we used are

$$\theta_i = 1 \quad \text{for } i = 1, \ldots, 4,$$
$$\lambda_1 = 10{,}000, \quad \lambda_2 = 4{,}000, \quad \lambda_3 = 8{,}000, \quad \lambda_4 = 6{,}000,$$
$$\eta_j = 2 \quad \text{for } j = 1, 2, 3, 4 \qquad \eta_5 = 4,$$
$$s_j = 1{,}000 \quad \text{for } j = 1, \ldots, 5.$$

We can think of this model as being model $r = 1{,}000$ in a sequence of models indexed by r, where $s_j = 1$, $\lambda_1 = 10$, $\lambda_2 = 4$, $\lambda_3 = 8$, and $\lambda_4 = 6$. We focus on the proportions $p_{i,j}$ given in (10). Table 1 gives our results. *EXP* refers to the model with service times exponential with mean $1/\eta_j$, whereas *UNIFORM* refers to the model with service times uniform in the interval $[1/(2\eta_j), 3/(2\eta_j)]$. The second column contains the fluid approximation values computed using the method described in §8. As indicated in §8, the dedicated flow rates for the sparsely connected components are computed first, and then the fully connected component is considered. All simulation results in this section are given as 95% confidence intervals. We see that for all routing proportions our fluid approximation gives results very close to the simulated values in both the exponential and uniform cases.

9.2. The X Model

We now focus on the X model in more detail. We work with the X model because it is the simplest nontrivial fully connected model. In light of §6, we should expect our fluid approximations to be close to accurate for arbitrary arrival processes satisfying the SLLN in (1) and for i.i.d. service times with nonlattice cdfs without atoms at 0. We see evidence that this is true by simulating the X model with mutually independent i.i.d. sequences of interarrival times and service times, for various combinations of interarrival-time and service-time distributions.

The parameters we used for our X model simulations are

$$\theta_1 = \theta_2 = \tfrac{1}{2}, \quad \lambda_1 = 2{,}000, \quad \lambda_2 = 3{,}000,$$
$$\eta_1 = 1, \quad \eta_2 = 2, \quad s_1 = s_2 = 1{,}000.$$

As in §6, the fluid approximation gives $p_{1,1} = p_{2,1} = 1 - p_{1,2} = 1 - p_{2,2} = 1/3$. Table 2 gives our simulation results. In this table, *EXP* and *UNIFORM* denote exponential and uniform distributions, respectively, as defined as in the previous section. In addition, we have the following:

- *GAMMA* denotes a gamma distribution with mean $1/\eta$ and squared coefficient of variation (SCV, variance divided by the square of the mean) 1/2 (equivalent to E_2, an Erlang of order 2),

Table 2 X Model Simulation Results for Various Combinations of Interarrival-Time and Service-Time Distributions

		Interarrival					
Service	Param.	*EXP*	*GAMMA*	*HYPEREXP*	*UNIFORM*	*TWOPOINT*	*CONSTANT*
EXP	p_{11}	0.3335 ± 0.0039	0.3320 ± 0.0063	0.3302 ± 0.0054	0.3328 ± 0.0055	0.3348 ± 0.0043	0.3330 ± 0.0041
	p_{21}	0.3320 ± 0.0048	0.3330 ± 0.0021	0.3334 ± 0.0049	0.3331 ± 0.0023	0.3334 ± 0.0049	0.3329 ± 0.0063
GAMMA	p_{11}	0.3344 ± 0.0056	0.3326 ± 0.0067	0.3333 ± 0.0054	0.3315 ± 0.0046	0.3315 ± 0.0055	0.3328 ± 0.0044
	p_{21}	0.3322 ± 0.0035	0.3335 ± 0.0045	0.3342 ± 0.0049	0.3346 ± 0.0046	0.3341 ± 0.0029	0.3333 ± 0.0027
HYPEREXP	p_{11}	0.3327 ± 0.0096	0.3334 ± 0.0066	0.3331 ± 0.0057	0.3312 ± 0.0046	0.3331 ± 0.0066	0.3355 ± 0.0072
	p_{21}	0.3334 ± 0.0058	0.3332 ± 0.0033	0.3335 ± 0.0057	0.3342 ± 0.0030	0.3332 ± 0.0061	0.3320 ± 0.0046
UNIFORM	p_{11}	0.3335 ± 0.0045	0.3334 ± 0.0048	0.3343 ± 0.0044	0.3342 ± 0.0047	0.3337 ± 0.0044	0.3337 ± 0.0037
	p_{21}	0.3328 ± 0.0027	0.3325 ± 0.0052	0.3327 ± 0.0044	0.3323 ± 0.0033	0.3339 ± 0.0033	0.3339 ± 0.0025
TWOPOINT	p_{11}	0.3335 ± 0.0066	0.3325 ± 0.0040	0.3341 ± 0.0062	0.3346 ± 0.0034	0.3311 ± 0.0059	0.3326 ± 0.0044
	p_{21}	0.3323 ± 0.0056	0.3333 ± 0.0031	0.3331 ± 0.0029	0.3322 ± 0.0026	0.3342 ± 0.0040	0.3329 ± 0.0030
CONSTANT	p_{11}	0.3343 ± 0.0056	0.3321 ± 0.0038	0.3328 ± 0.0022	0.3328 ± 0.0034	0.3333 ± 0.0034	0.3340 ± 0.0047
	p_{21}	0.3331 ± 0.0037	0.3346 ± 0.0022	0.3341 ± 0.0018	0.3337 ± 0.0020	0.3334 ± 0.0022	0.3328 ± 0.0032

• *HYPEREXP* denotes a hyperexponential distribution (a mixture of two exponentials with means $1/\xi_j$ and probability weights p_j) with mean $1/\eta \equiv (p_1/\xi_1) + (p_2/\xi_2)$, balanced means $(p_1/\xi_1) = (p_2/\xi_2)$ and SCV $= 4$,
• *TWOPOINT* refers to a two-point distribution attaching probability 1/2 to each of the two points $1/(2\eta)$ and $3/(2\eta)$,
• *CONSTANT* refers to the constant $1/\eta$.

Each box in the table contains 95% confidence intervals for $p_{1,1}$ and $p_{2,1}$. We exclude $p_{1,2}$ and $p_{2,2}$ because, necessarily, $p_{1,2} = 1 - p_{1,1}$ and $p_{2,2} = 1 - p_{2,1}$. For this example, we have no difficulties even with the lattice service distributions *TWOPOINT* and *CONSTANT* because the random, exponentially distributed abandonment serves to remove the deterministic regularity. We also ran the same simulations with batch arrivals of size 2 and obtained essentially the same results. This gives further evidence that the arrival distributions do not play a role in the fluid approximations. These results are presented in the e-companion.

9.3. Negative Flows

In this subsection, we consider examples of sparsely connected models in which the unique solution for the flow rates $r_{i,j}$ produces one or more negative values, so that global FCFS cannot be achieved for the model. First, we give a simple example with a single negative flow, and show that simulation matches the result obtained by decomposing the network into two separate sparsely connected networks by setting the negative flow equal to 0, with global FCFS in each component. Next, we consider a more complicated example with two negative flows. We give one example in which only one of these flows should be set equal to zero and another in which both should be set equal to zero. From the parameters, we can perhaps guess the right result, but we have yet to devise a systematic procedure for determining which negative flows should be set equal to zero. We have not yet shown that an initially positive flow never should be set equal to zero, but we have not experienced it.

9.3.1. One Negative Flow. Consider a W model with the following parameters: $\lambda_1 = \lambda_2 = 3$, $\lambda_3 = 10$, $\theta_1 = \theta_2 = \theta_3 = 2$, and $\mu_1 = \mu_2 = 1$. After solving the system of Equations (7) and (8), we find that $p_{2,2} = -2/3 < 0$, with all other flow rates positive. This indicates that global FCFS can not be achieved for this model. When simulating the model with Markovian parameters for $r = 1{,}000$, we find that

$$p_{1,1} = p_{2,1} = p_{3,2} = 1, \quad p_{2,2} = 0, \quad w_1 = 0.892 \pm 0.006,$$
$$w_2 = 0.893 \pm 0.006, \quad w_3 = 1.150 \pm 0.007.$$

Because $p_{22} = 0$, with all other proportions positive, the model actually decomposes into two models. The first model consists of customer classes 1 and 2 and service pool 1. The second model consists of customer class 3 and service pool 2. Because wait times for both queues are practically the same, we see that global FCFS is achieved in the first model in the decomposition. Because the second model only contains one queue, global FCFS is trivially achieved in the second model.

9.3.2. Two Negative Flows. We now simulate models for which the fluid flow-rate equations give two negative flow rates. In the first model, both flows in the simulated queueing model are actually close to zero. In the second, only one of the flows is actually close to zero. That shows that we have difficulty in deciding how to decompose the model, without performing simulations. But we do find that the fluid approximation closely matches simulation after we have found the proper decomposition. We show the fluid model description both before and after doing the decomposition.

"Super" N Model. Consider a model with parameters given in the left of Figure 4, again with exponential interarrival, abandonment, and service distributions.

Figure 4 "Super" *N* and "Super" *W* Model Simulation Parameters

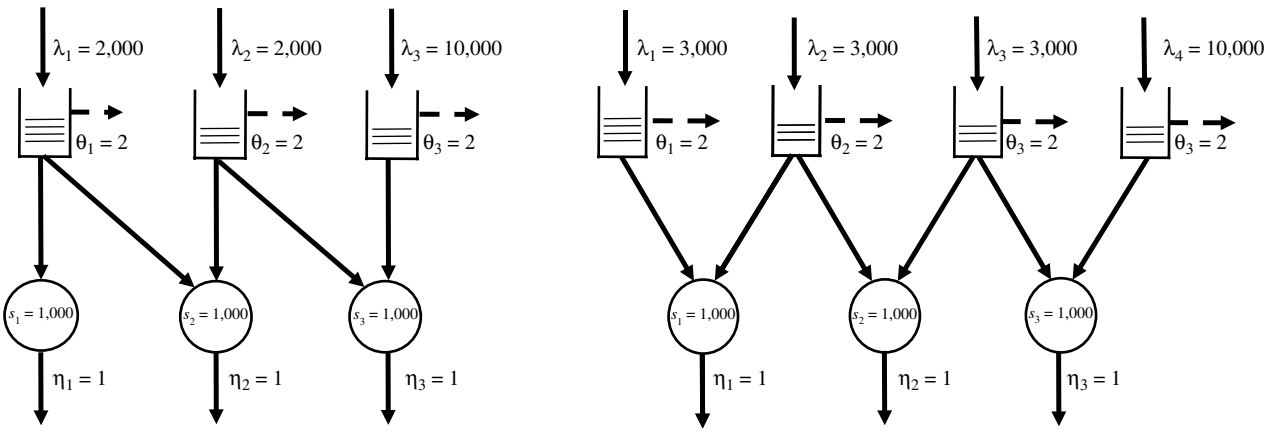


Table 3 Simulation Results for the "Super" *N* Model

	$p_{1,1}$	$p_{1,2}$	$p_{2,2}$	$p_{2,3}$	$p_{3,3}$	w_1	w_2	w_3
Fluid approx.	2.333	−1.333	3.667	−2.667	1	0.770	0.770	0.770
Removing (2, 3)	1	0	1	0	1	0.346	0.346	1.151
Removing (1, 2)	1	0	1	0	1	0.346	0.346	1.151
Simulation	0.976 ± 0.006	0.024 ± 0.006	1	0	1	0.333 ± 0.007	0.359 ± 0.006	1.151 ± 0.010

We refer to this model as a "super" N model. Our fluid approximation and simulation results are given in Table 3. Notice here we have $p_{12} < 0$ and $p_{23} < 0$ in the initial fluid approximation. For this example, our simulation gives $p_{12} \approx 0$ and $p_{23} \approx 0$ as well, and shows that the waiting times are quite different at the three queues. We clearly are not able to achieve global FCFS in this model. This result is consistent with decomposing the original model by either (i) removing the route from class 2 to pool 3 first or (ii) removing the route from class 1 to pool 2 first. In both cases, after the first decomposition we are left with the N model of the previous subsection and an I model, and in both cases we get the same approximation. In this example we could have simply eliminated the two negative flows at the outset. This is in contrast to the "super" W model discussed below.

"Super" W Model. Now consider a model with parameters given in the right of Figure 4, again fully Markovian. We refer to this model as a "super" W model. We give our fluid approximation and simulation results in Table 4. After solving the system of routing flow equations, we find that $p_{2,2} = -1/9$ and $p_{3,3} = -11/9$, with all other flow rates positive. Again, this indicates that global FCFS cannot be achieved for this model. Here we see that although the fluid rate equations give us a negative value for $p_{2,2}$, simulation tells us that this proportion is not actually zero. If we decompose the model by first removing the flow from class 3 to pool 3, then we get a symmetric W model and an I model (second row of Table 4), where each of these models can achieve global FCFS. Clearly, this would give us a good fluid approximation. However, if we instead decompose the model by removing the flow from class 2 to pool 2, we get a symmetric V model and an N model (third row of Table 4). But because simulation results tell us $p_{2,2} > 0$, this would not give us a good fluid approximation. In this example, we can anticipate in advance what will happen in the initial fluid calculation. Moreover, we can easily see how the network should be decomposed, but we have yet to specify an automatic procedure to determine the appropriate decomposition. This phenomenon tells us that further analysis is necessary in order to determine what actually happens in models for which the fluid rate equations yield multiple negative flow rates.

10. Conclusions

We have shown how to determine the fluid flow rates and, thus, also full steady-state performance descriptions, for a large subclass of overloaded multiclass fluid models with abandonment. We have identified a hierarchy of complexity in the models, which is captured by the routing graph. We obtain the desired routing flow rates in three cases: when the model is (i) sparsely connected, (ii) fully connected, and (iii) an appropriate combination of these two cases. Other cases remain unsolved. The analysis determines whether or not global FCFS is achievable; we obtain a complete description when it is (in the second case it always is). When it is not, the behavior is evidently captured by an appropriate decomposition in which one or more arcs with negative flows are removed from the model. In the last two solved cases, the routing flow rates are actually not uniquely determined in general, but become so once we make stochastic assumptions about prelimit queueing models converging to the fluid limit. But the results hold for quite general stochastic models, as a consequence of the classic limit theorem establishing convergence of the superposition of independent stationary point processes to a Poisson process.

There are many fascinating open problems. We have yet to determine the stationary behavior in the sparsely connected and hybrid cases when global FCFS fails. We have shown that this occurs when

Table 4 Simulation Results for the "Super" *W* Model

	$p_{1,1}$	$p_{2,1}$	$p_{2,2}$	$p_{3,2}$	$p_{3,3}$	$p_{4,3}$	w_1	w_2	w_3	w_4
Fluid approx.	1	1.111	−0.111	2.222	−1.222	1	0.923	0.923	0.923	0.923
Removing (3, 3)	1	0.500	0.500	1	0	1	0.752	0.752	0.752	1.151
Removing (2, 2)	1	1	0	2.167	−1.167	1	0.896	0.896	0.936	0.936
Simulation	1	0.500 ± 0.025	0.500 ± 0.025	1	0	1	0.754 ± 0.007	0.754 ± 0.00	0.754 ± 0.007	1.151 ± 0.008

the characterizing linear system yields negative flows. Evidently, the system behavior can be described by setting some of these negative flows to zero and analyzing the separate components of the resulting decomposed network, but we have yet to determine precisely what happens.

There are many routing graphs we have not yet been able to analyze. Moreover, it remains to prove Conjectures 1 and 2, stating that the steady-state fluid equations are asymptotically correct both as $r \to \infty$ and then as $t \to \infty$.

11. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

Acknowledgments

The authors thank Ed Kaplan for suggesting this problem. The authors were supported by National Science Foundation Grant DMI-0457095.

References

Asmussen, S. 2003. *Applied Probability and Queues*. Springer-Verlag, New York.

Atar, R. 2005. A diffusion model of scheduling control in queueing systems with many servers. *Ann. Appl. Probab.* **15**(1B) 820–852.

Billingsley, P. 1999. *Convergence of Probability Measures*. Wiley, New York.

Caldentey, R. A., E. H. Kaplan. 2002. A heavy-traffic approximation for queues with restricted customer-server matchings. Working paper, Yale University, New Haven, CT.

Chen, H., D. D. Yao. 2001. *Fundamentals of Queueing Networks*. Springer-Verlag, New York.

Daley, D. J., D. Vere-Jones. 1988. *An Introduction to the Theory of Point Processes*. Springer-Verlag, New York.

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

Garnett, O., A. Mandelbaum. 2000. An introduction to skills-based routing and its operational complexities. http://iew3.technion.ac.il/serveng/Lectures/SBR.pdf.

Gurvich, I., W. Whitt. 2007. Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing. Working paper, Columbia University, New York, http://www.columbia.edu/~ww2040.

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

Johnson, M. P. 2001. Tenant-based subsidized housing location planning under uncertainty. *Socio-Econom. Planning Sci.* **35**(3) 149–173.

Johnson, M. P. 2003. Single-period location models for subsidized housing: Tenant-based subsidies. *Ann. Oper. Res.* **123**(2) 105–124.

Kaplan, E. H. 1984. Managing the demand for public housing. Technical Report 183, MIT Operations Research Center, Cambridge, MA.

Kaplan, E. H. 1985. How PHA's fill their units. *J. Housing* **42**(1) 13–20.

Kaplan, E. H. 1986. Tenant assignment models. *Oper. Res.* **34**(6) 832–843.

Kaplan, E. H. 1988. A public housing queue with reneging and task-specific servers. *Decision Sci.* **19**(2) 383–391.

Kaplan, E. H. 2006. Personal communication.

Khintchine, A. Y. 1960. *Mathematical Methods in the Theory of Queueing*. Charles Griffin & Co. Limited, London.

Whitt, W. 1984. Departures from a queue with many busy servers. *Math. Oper. Res.* **9**(4) 534–544.

Whitt, W. 2002. *Stochastic-Process Limits*. Springer-Verlag, New York.

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

Whitt, W. 2006a. Fluid models with multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.

Whitt, W. 2006b. A multi-class fluid model for a contact center with skill-based routing. *Internat. J. Electronics Comm.* **60**(2) 95–102.