ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Int. J. Electron. Commun. (AEÜ) ▮▮▮ (▮▮▮) ▮▮▮–▮▮▮

International Journal AEÜ of Electronics and Communications

www.elsevier.de/aeue

A multi-class fluid model for a contact center with skill-based routing

Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, 500 West 120th Street, New York, NY 10027-6699, USA

Abstract

A multi-class deterministic fluid model is proposed to describe and improve the performance of a customer contact center with skill-based routing. The fluid model can be regarded as an approximation for a stochastic queueing system with multiple customer classes and multiple server groups, with customer abandonment and non-exponential service-time and time-to-abandon distributions. The fluid model is attractive to provide a rough analysis of large systems, with high arrival rate and many servers. Even though the fluid model evolves deterministically, the service-time distributions and time-to-abandon distributions beyond their means play a critical role. The fluid model can be used for staffing, routing and system design, because it is possible to formulate tractable optimization problems.

Keywords: Deterministic fluid models; Multi-server queues with abandonment; Contact centers; Call centers; Skill-based routing

1. Introduction

It is a pleasure to contribute to this special issue in honor of Paul Kuehn on his 65th birthday. The time our paths first crossed 28 years ago was a turning point in my research career. Even though I had received a doctorate in engineering, I had not yet developed the perspective of an engineer. I was doing academic research, writing research papers motivated by research papers. Being vaguely aware that something was missing, I left academia in 1977 and joined Bell Labs, the research branch of AT&T, and began to see engineering in action. There were many impressive people at Bell Labs, and among them, Paul stood out. Paul was the epitome of an engineer: His talks evoked organization, clarity and technical depth.

In those early Bell Labs days, my research was primarily aimed at developing new methods for approximately analyzing complex non-Markovian multi-class queueing networks, which serve as models of communication networks, computer systems and manufacturing facilities. That effort led to a software package called *The Queueing Network Analyzer* (QNA) [1–3], and it led to related theory and applications [4–10]. As acknowledged in [2], there were important precedents; notable among them was Kuehn [11].

In 2002, I left AT&T and joined the IEOR Department of Columbia University. My main research focus has shifted from communication networks to service systems, specifically to *customer contact centers*. A contact center is a collection of resources providing an interface between a service provider and its customers. The classical contact center is a telephone call center, containing service representatives (agents) who talk to customers over the telephone. In modern call centers, agents are supported by elaborate information-and-communication-technology equipment, such as an interactive voice response unit, an automatic call distributor (ACD), a personal computer and assorted databases. The operational efficiency has been improved through voice over IP.

With the rapid growth of e-commerce, contact is often made via e-mail or the Internet instead of by telephone. There often are many types of service requests, requiring different service skills, such as knowledge of different languages or technical information, and the agents differ in

E-mail address: ww2040@columbia.edu.

1434-8411/$ - see front matter © 2005 Elsevier GmbH. All rights reserved.
doi:10.1016/j.aeue.2005.11.005

their ability to respond to these requests. The ACD is able to route calls to different agents through skill-based routing (SBR), but there remains an opportunity for better design and control, including routing and staffing. Since contact centers play a vital role throughout the service sector, and since the service sector is a growing part of the economy, there is great potential for new technological contributions in this area. For background on contact centers, see [12].

Even though contact centers are quite different from the Internet and web server farms, many of the same modelling techniques used to analyze the performance of computer systems and communication networks apply. Indeed, to a large extent, modern contact centers can be regarded as special kinds of computer systems and communication networks. But differences in detail lead to different modelling approaches. Clearly, there is a difference in time scale between the transmission time of a data packet and the duration of a service contact: in a contact center the relevant time scale is minutes or seconds, corresponding to the duration of a service contact or the waiting time before service can begin. In many ways, a contact center is more like a classical circuit-switched telephone network, because the items to be processed are again calls and it is natural for the queues to have multiple servers (the agents). That connection suggests considering stochastic network models such as in [13–15], but the network structure plays less of a role here (e.g. each customer needs only a single server) and delay in providing service is an important consideration.

In this paper we propose a new model to study and improve the design and performance of customer contact centers. It is a multi-class deterministic fluid model, which arises as the limit of a multi-class many-server queueing system as both the arrival rate and the number of servers increase. The fluid model is appealing in contrast to fluid models associated with single-server queues and networks of such queues, e.g. see Section 5.3.1 of [10], because the performance descriptions depend on important model probability distributions beyond their means. The reason these distributions beyond their means play a critical role is that the state of the fluid model does not just consist of the numbers of customers of each class in queue and in service at each service group at each time, but in addition contains the length of time each customer has been in queue or in service. *The system state is given an important extra time dimension*, which we are not accustomed to seeing in fluid models. (What we do here should be contrasted with the customary state in Markovian many-server queueing models [13,16–18]. Precedents for adding time to the state exist in the analysis of non-Markovian many-server queues, e.g. [19,20].) In detail, the multi-class fluid model introduced here is quite different from QNA in [2], but it is similar in spirit. We introduce an approximation that greatly reduces the complexity, and yet still captures important system dynamics. In both cases, there is an attempt to treat non-Markovian models and to capture the impact of probability distributions beyond their means.

The multi-class fluid model introduced here extends a corresponding single-class fluid model introduced in Whitt [21]. That single-class fluid model already has had some applications: to study the impact of uncertain model parameters [22], to study the impact on aggregate system performance of delay announcements [23], and to study outsourcing strategies [24]. Alternative fluid models to develop new approaches to contact centers have been proposed by Bassamboo, Harrison and Zeevi [25–27].

Here we start in Section 2 by specifying a reference SBR queueing model. Next in Section 3 we introduce the proposed fluid model, concentrating on describing the equilibrium or steady-state behavior. In Section 4 we consider optimization problems for system design, which specify staffing and partially specify routing. In Section 5 we discuss issues arising in the implementation of fluid-model results in actual contact centers, in particular, accounting for stochastic fluctuations and producing associated routing strategies in the actual contact center. Finally, in Section 6 we draw conclusions.

2. An SBR queueing model

In this section we define a multi-class queueing model of an SBR contact center, which we denote by $(G/GI+GI)^m/s^n$. This queueing model helps put the fluid model we introduce in the next section in perspective, because we can view the fluid model as an approximation of it.

In the queueing model there are m customer classes and n service groups, with s_j servers in service group j, $1 \leqslant j \leqslant n$, and thus a total of $s = s_1 + \cdots + s_n$ servers. The individual customer classes and service groups are homogeneous: Customers from each customer class are assumed to have common characteristics, and servers in each server group are assumed to have common characteristics. (The priority skill matrix in [28] is one way to relax that requirement.) The servers have skills, specifying which customer classes they can serve. For example, service group 3 might be able to serve customer classes 1 and 4. There is a queue associated with each customer class, where arriving customers of that class wait if they do not enter service immediately upon arrival. There also is a queue for each service group, where idle servers of that service group wait if they are not assigned to serve waiting customers immediately upon service completion. These various queues might be virtual. For examples of queueing models of contact centers, see [12,28] and references therein.

The system is operated by making decisions at two transition epochs: (1) at the epochs of *customer arrivals*, and (2) at the epochs of *service completions*. First, upon each arrival of a class-i customer, we consider whether we should assign that customer to an idle server in one of the service groups that can serve class i, if one is available, or we put the customer in the class-i queue to wait. Second, upon each service completion by a server from server group j, we consider whether we should assign that server to one of the waiting customers in the customer classes that server can serve or we put the server in the queue of idle class-j servers. We assume that some non-preemptive, non-anticipating policy is used to assign customers to idle servers, defined to address both those two decisions. For both queues – of waiting customers and idle servers – we must have some *queue discipline* for deciding which is to be assigned next. One natural candidate is first-in first-out (FIFO), but others are possible. For example, we might assign the server that has the largest proportion of idle time during the last half hour.

Now we define the *stochastic model elements*. Customer class i has arrivals according to a *general stationary point process* $A_i \equiv \{A_i(t) : t \geqslant 0\}$ with *arrival-rate* λ_i; that is, we assume that $A_i(t)/t \to \lambda_i > 0$ as $t \to \infty$ with probability one (w.p.1). It is natural to assume that the arrival processes are mutually independent Poisson processes, but we do not require it.

Each class-i customer who is required to wait before starting service *balks* (leaves immediately upon arrival) with probability β_i, and elects to wait with probability $1 - \beta_i$, independently of the current state and history. Each class-i customer that cannot enter service immediately and does not balk may subsequently *abandon* (leave after joining the queue, before starting service). Successive times to abandon of class-i customers are *independent and identically distributed* (i.i.d.) random variables with a *cumulative distribution function* (cdf) F_i. That is natural with invisible queues (when waiting customers cannot see the state of the system, as is typical for contact centers without delay announcements).

We assume that customers do not abandon after they start service. The service times of class-i customers may depend on the service group where they are served. Successive service times of class-i customers by servers from service group j are i.i.d. random variables with a cdf $G_{i,j}$. We assume that the balking decisions, the times to abandon and the service times are all mutually independent random variables, independent of the system history. (We assume that balking and abandonment do not influence future arrivals.)

Let T_i be a generic time to abandon for a class-i customer, and let $S_{i,j}$ be a generic service time of a class-i customer served by a server from service group-j. Then our assumptions above imply that $F_i(t) \equiv P(T_i \leqslant t)$ and $G_{i,j}(t) \equiv P(S_{i,j} \leqslant t)$ for $t \geqslant 0$. We assume that these random variables have finite means: $m_{a,i} \equiv E[T_i]$ and $m_{s,i,j} \equiv E[S_{i,j}]$.

As advertised at the outset, an important goal is to capture the impact of probability distributions beyond their means. It is significant that we do not assume that the service-time cdf's $G_{i,j}$ and the time-to-abandon cdf's F_i are exponential. Statistical analysis of telephone-holding-time data has shown that the probability distributions of both service times and times to abandon often are not nearly exponential [29,30].

In running the SBR contact center, there are two decisions to make: staffing and routing. The staffing is the choice of the numbers s_j, for $1 \leqslant j \leqslant n$, while the routing specifies the assignment of servers to customers. There also is a larger design question, specifying the customer classes to be served, perhaps including the arrival rates λ_i, and the service groups to provide the service, perhaps including the service-time cdf's $G_{i,j}$. Our approximate fluid model in the next section is intended to focus on the higher-level issues such as design, as opposed to determining the optimal routing strategy for each individual service interaction (call).

3. An approximating fluid model

In this section we introduce a fluid model approximating the $(G/GI+GI)^m/s^n$ SBR queueing model from the last section. The fluid model arises by scaling up the arrival rates and the numbers of servers, while holding the balking probabilities β_i, time-to-abandon cdf's F_i and service-time cdf's $G_{i,j}$ fixed, but we do not establish any limits here.

We can do the scaling by introducing a family of models indexed by a scaling parameter η, and then let $\eta \to \infty$. We let the arrival rates and number of servers be functions of η, and then let

$$\frac{\lambda_i(\eta)}{\eta} \to \lambda_i \quad \text{and} \quad \frac{s_j(\eta)}{\eta} \to s_j \quad \text{as } \eta \to \infty. \tag{1}$$

We then scale the customer number-in-service and queue-length processes by dividing by η, converting individual customers into "atoms of fluid" in the limit. Thus $\lambda_i(\eta) \approx \eta\lambda_i$ is the arrival rate of customers in the queueing model η, but λ_i becomes the arrival rate of class-i fluid in the limit after scaling. Similarly, $s_j(\eta) \approx \eta s_j$ is the number of class-j servers in queueing system η, while s_j is the class-j fluid service capacity in the limiting fluid model obtained after scaling.

For the single-class ($m = 1$), single-service-group ($n = 1$) case (where routing is not an issue), the fluid model has been shown to be asymptotically correct in the regime (1) in [21]. (So far, the asymptotic correctness has only been verified for the Markovian $M/M/s+M$ special case [18,31] and a discrete-time analog of the general $G_t(n)/GI/s+GI$ fluid model, allowing both time-dependent and state-dependent arrivals [21], but since the time increments can be arbitrarily short, that discrete-time setting suffices for practical purposes.) The steady-state behavior of the single-class fluid model has been shown to yield accurate approximations for the corresponding queueing systems with 100 servers in overloaded scenarios through comparisons with exact numerical results obtained from numerical algorithms and simulations [21,31,32]. At the same time, the fluid model provides great simplification that makes it possible to investigate other more complicated questions. For additional discussion of fluid models, see [17,27,33–35].

A major complication arising when we go from the single-class fluid model in [21] to the multi-class fluid model here is the routing. However, recent work indicates that it is possible to treat the routing in a relatively "broad-stroke" way; e.g.

see [28]. With that in mind, we treat the routing in a very elementary way. That leaves open the question of how best to do the routing in practice. In Section 5 we discuss how that might be done, but that remains to be more fully explored.

Here is how we handle routing (assign class-i fluid to class-j service groups) in the fluid model: At the outset, we allocate a proportion $r_{i,j}$ of all class-i fluid to service group j. Thus, $\sum_{j=1}^{n} r_{i,j} = 1$ for all i. Mathematically, these proportions can be regarded as probabilities, but we are not explicitly assuming that Markovian routing is being employed. We regard these proportions $r_{i,j}$ as decision variables, to be specified. They may be limited by the available skills of the servers in the service groups.

Thus, our fluid model is characterized by the parameter six-tuple $(\lambda, \beta, \mathbf{F}, \mathbf{r}, \mathbf{G}, \mathbf{s})$, where $\lambda \equiv (\lambda_1, \ldots, \lambda_m)$ and $\beta \equiv (\beta_1, \ldots, \beta_m)$ are m-tuples of numbers, $\mathbf{F} \equiv (F_1, \ldots, F_m)$ is an m-tuple of cdf's, $\mathbf{r} \equiv (r_{i,j} : 1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n)$ is an $m \times n$ matrix of numbers, $\mathbf{G} \equiv (G_{i,j} : 1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n)$ is an $m \times n$ matrix of cdf's, and $\mathbf{s} \equiv (s_1, \ldots, s_n)$ is an n-tuple of positive integers. The matrix $\mathbf{r}$ can be regarded as the transition matrix of a discrete-time Markov chain, because the rows are probability vectors. The way the parameters simplify going from the queueing model to the fluid model provides insight. Note that the arrival processes A_i enter in only through their rates λ_i, but the full cdf's F_i and $G_{i,j}$ remain relevant in the description of the fluid model.

We now describe how the fluid model evolves over time. Class-i fluid arrives at rate λ_i and, a priori, we know that a proportion $r_{i,j}$ of it will be served at service group j, so we can think of the queue for class i partitioned into n parts, depending upon the ultimate destination. For some classes i, the class-i fluid can enter service immediately upon arrival, but for other i the fluid must wait in queue. Of the class-i fluid that has to wait before starting service, a proportion $F_i(t)$ abandons by time t after it arrives if it has not started service by that time. It will turn out that all class-i fluid that is served will enter service at a fixed deterministic time w_i.

The fluid model can describe the evolution of performance (flow through the system, queue lengths and times spent) over time as a function of the initial conditions (and the model elements), as discussed for the single-class, single-service-group case in [21]. Indeed, a discrete-time analog of the fluid model is introduced in Section 6 of [21], which makes it possible to numerically calculate the time-dependent performance of a fluid model with time-dependent and state-dependent arrivals, as a function of the initial conditions. However, here we will only consider the stationary fluid model in steady state. We intend to discuss the time-dependent fluid model elsewhere.

We start by describing the offered loads (requested service times), noting that the amount of work depends on the routing, since the required service time depends on the routing. However, we now take the routing as specified by the proportions $r_{i,j}$. The (i, j) – *offered load* is the arrival rate times the mean service time, i.e. $L_{i,j} = \lambda_i r_{i,j} m_{s,i,j}$. The (i, j) offered load represents the type-j service capacity needed to serve class-i customers, provided that we can ignore stochastic fluctuations, which is precisely what the fluid model does.

The model behavior is much more interesting if some classes are not served immediately. Then balking and abandonment play an important role. We now partition the set $\mathscr{C}$ of customer classes into two subsets: $\mathscr{I}$ and $\mathscr{C} - \mathscr{I}$. The customers in classes $\mathscr{I}$ get served immediately upon arrival, while the customers in the remaining classes will have to wait before starting service. We can explore the different partitions separately. We now assume that one specific partition has been selected, with $\mathscr{C} - \mathscr{I}$ non-empty. That will have implications on what happens for the (i, j) pairs and for the system as a whole.

Let $\mathscr{S} \equiv \{1, 2, \ldots, n\}$ be the set of all server groups and let $\mathscr{S}_i$ be the set of server groups that are allowed to serve customers of class i, $1 \leqslant i \leqslant m$. Let $\mathscr{C} \equiv \{1, 2, \ldots, m\}$ be the set of all customer classes and let $\mathscr{C}_j$ be the customer classes that can be served by server group j, $1 \leqslant j \leqslant n$. We require that $i \in \mathscr{C}_j$ if and only if $j \in \mathscr{S}_i$.

We assume that all the customers in classes belonging to $\mathscr{I}$ can be served immediately, while the remainder cannot. That leads to two different sets of constraints

$$\sum_{i \in \mathscr{I} \cap \mathscr{C}_j} \lambda_i r_{i,j} m_{s,i,j} = s_{j,I} < s_j, \tag{2}$$

$$\sum_{i \in (\mathscr{C} - \mathscr{I}) \cap \mathscr{C}_j} \lambda_i r_{i,j} (1 - \beta_i) m_{s,i,j} > s'_j \equiv s_j - s_{j,I} \tag{3}$$

for all j, where $s_{j,I}$ is defined in (2). In (3) we have assumed that the offered load after balking exceeds the available capacity, after deleting the committed capacity for those classes to be served immediately. We thus allow balking and abandonment for classes i in $\mathscr{C} - \mathscr{I}$ to reduce the load faced by the servers, enabling the servers to meet the requirements. We are thus thinking of our overall SBR contact center as a queueing system with customer balking and abandonment operating in the so-called *efficiency-driven* (*ED*) *heavy-traffic regime* [21,31,34,36,37].

We also assume for each waiting class i (in $\mathscr{C} - \mathscr{I}$) that all class-$i$ fluid that is served enters service at a fixed positive time w_i. We regard these waiting times as decision variables, along with the routing proportions $r_{i,j}$ and the capacities s_j, but these waiting times w_i must satisfy equations, just as for the one-dimensional fluid model in [21], namely,

$$\sum_{i \in (\mathscr{C} - \mathscr{I}) \cap \mathscr{C}_j} L_{i,j} (1 - \beta_i) F_i^{\mathrm{c}}(w_i) = s'_j \tag{4}$$

for all j, $1 \leqslant j \leqslant n$, where s'_j is the reduced service-group-j capacity defined in (3) and $F_i^{\mathrm{c}}(t) \equiv 1 - F_i(t)$ is the complementary cdf (ccdf) associated with the cdf F_i. Eq. (4) says that the total reduced offered load at each server group after balking and abandonment should coincide with the available capacity there.

The steady-state behavior of the fluid model is determined by the systems of equations in (2) and (4), where the variables s_j, $r_{i,j}$ and w_i are allowed to vary. Here is what happens for class-i fluid that must wait before beginning service: A proportion β_i of all arriving class-i fluid balks. All class-i fluid that is served waits precisely w_i before entering service. A proportion $r_{i,j}$ of all class-i input is allocated to server group j. Let $P(A_i)$ be the proportion of class-i fluid that abandons; let $P(S_i)$ be the proportion of class-i fluid that is served; and let $P(S_{i,j})$ be the proportion of all class-i fluid that is served by service group j. Then $P(A_i) = (1-\beta_i)F_i(w_i)$, $P(S_i) = (1-\beta_i)F_i^{\mathrm{c}}(w_i)$ and $P(S_{i,j}) = P(S_i)r_{i,j}$.

The *density* of class-i fluid that has been waiting for time t is

$$q_i(t) = \lambda_i(1-\beta_i)F_i^{\mathrm{c}}(t), \quad 0 \leqslant t \leqslant w_i, \tag{5}$$

with $q_i(t) = 0$ for all $t > w_i$. The *queue content* for class i is

$$Q_i = \lambda_i(1-\beta_i)\int_0^{w_i} F_i^{\mathrm{c}}(t)\,\mathrm{d}t. \tag{6}$$

Even though these queue-content descriptions are deterministic functions, the time-to-abandon cdf's F_i play a prominent role in the description. Under assumption (3), the abandonment cdf's determine the final steady-state performance via the critical system of equations (4), and they influence the queue content via (6).

The service-time cdf's $G_{i,j}$ enter in so far only via their means $m_{s,i,j}$. In steady state, the servers are all always busy. Class-i fluid is processed from service group j at rate $1/m_{s,i,j}$. Class-i fluid enters and leaves service group j at a total rate of $P(S_{i,j})/m_{s,i,j}$. The entire system is kept in steady state by having the class-i arrival rate λ_i balanced by the class-i balking, abandonment and service rates – $\lambda_i\beta_i$, $\lambda_i P(A_i)$ and $\lambda_i P(S_i)$, respectively: $\lambda_i = \lambda_i\beta_i + \lambda_i P(A_i) + \lambda_i P(S_i)$.

We can also describe the density of the fluid that is in service. For classes $i \in (\mathscr{C} - \mathscr{I}) \cap \mathscr{C}_j$, the density of class-$i$ fluid that has been in service at service group j for time t, and thus has been in the system for time $w_i + t$ is

$$b_{i,j}(t) = \lambda_i(1-\beta_i)F_i^{\mathrm{c}}(w_i)r_{i,j}G_{i,j}^{\mathrm{c}}(t), \quad t \geqslant 0. \tag{7}$$

For classes $i \in \mathscr{I} \cap \mathscr{C}_j$, the density of class-$i$ fluid that has been in service at service group j (and thus also in the system) for time t is $b_{i,j}(t) = \lambda_i r_{i,j} G_{i,j}^{\mathrm{c}}(t)$, $t \geqslant 0$. The total fluid content that has been in the service at service group j for time t is $b_j(t) = \sum_{i\in\mathscr{C}_j} b_{i,j}(t)$, $t \geqslant 0$.

While the density of fluid content is deterministic, we interpret the experience of individual customers or "atoms of fluid" as stochastic, regarding these as i.i.d. (The strong law of large numbers is acting behind the scenes to convert the individual independent actions into an overall system deterministic behavior.) For $i \in \mathscr{C} - \mathscr{I}$, each "class-$i$ customer" abandons before time t with probability $F_i(t)$, provided that $0 < t < w_i$. With probability $F_i^{\mathrm{c}}(t)$, the customer remains in the system after time t. However, any customer that has not abandoned by time w_i enters service at that time. Thus, as stated above, the waiting time (before entering service) is precisely w_i for all class-i customers that do enter service. The expected or average waiting time for *all* class-i fluid is

$$E[W_i] = P(S_i)w_i + (1-\beta_i)\int_0^{w_i} x\,\mathrm{d}F_i(x) = \frac{Q_i}{\lambda_i} \tag{8}$$

as shown in [21]. (The last relation can be viewed as a consequence of Little's law, $L = \lambda W$). The mean $E[W_i]$ is of course less than or equal to the waiting time w_i of the class-i fluid that is served. We regard W_i as a random variable because we interpret the experience of individual customers (atoms of fluid) is random.

4. Costs and benefits

We first consider the special case in which all customers are served immediately upon arrival. Then afterwards we consider the more interesting remaining cases.

4.1. The case of no waiting

When all customers are served immediately upon arrival, we can serve all fluid without any congestion if the number of servers in each service group exactly matches the offered load at that service group, i.e. if $\sum_{i=1}^{m} L_{i,j} = s_j$ for all j, $1 \leqslant j \leqslant n$.

In this case, there is no balking or abandonment, so the balking probabilities β_i and the abandonment cdf's F_i play no role. In this context, we can use the fluid model to design the system, i.e. to choose the numbers of servers s_j and the scheduled routing $r_{i,j}$ in order to meet specified demand, specified in terms of the arrival rates λ_i and the mean service times $m_{s,i,j}$. To do so, we formulate an optimization problem.

Let $v_{i,j}(x)$ be the rate value is accrued from serving a quantity x of class-i fluid by service group j and let $c_j(y)$ be the cost rate of providing a quantity y of capacity for service group j. The *optimization problem* is to maximize the net reward rate $R \equiv R(\mathbf{r}, \mathbf{s})$, where

$$R(\mathbf{r}, \mathbf{s}) \equiv \sum_{i=1}^{m}\sum_{j=1}^{n} v_{i,j}(\lambda_i r_{i,j}) - \sum_{j=1}^{n} c_j(s_j) \tag{9}$$

over all allowed $r_{i,j}$ and s_j, *subject to the constraints* that

$$\begin{aligned} &\sum_{i\in\mathscr{C}_j} \lambda_i r_{i,j} m_{s,i,j} = s_j \quad \text{for all } j, \\ &\sum_{j\in\mathscr{S}_i} r_{i,j} = 1, \qquad \sum_{j\in\mathscr{S}-\mathscr{S}_i} r_{i,j} = 0 \quad \text{for all } i, \end{aligned} \tag{10}$$

where $s_j \geqslant 0$ and $r_{i,j} \geqslant 0$ for all i and j.

If the functions $v_{i,j}$ and c_j appearing in the objective function (9) are linear, then the optimization is a *linear program*, but we think it may be important to consider nonlinear rewards and costs. By introducing a sequence of linear approximations for the nonlinear objective function, it may be possible to develop an effective iterative optimization algorithm.

For the many applications with few customer classes and few service groups, it may be possible to essentially evaluate the performance of all cases, by performing a search, by performing evaluations over successive finite grids. For example, with two classes and two service groups, we have four proportions $r_{i,j}$ to define, but $r_{1,1} = 1 - r_{1,2}$ and $r_{2,2} = 1 - r_{2,1}$. We thus can let $r_{1,2} = j_1/k$ and $r_{2,1} = j_2/k$, and consider alternative vectors (j_1, j_2), with $0 \leqslant j_i \leqslant k$, $i = 1, 2$.

4.2. Some classes are not served immediately

The model behavior is much more interesting if some classes are not served immediately. Then balking and abandonment play an important role. In the fluid model specified above there are four sets of decision variables: The fluid steady-state depends on (1) the partition of the set $\mathscr{C}$ of customer classes into the subset $\mathscr{I}$ that will be served immediately and the complement $\mathscr{C} - \mathscr{I}$ that will have to wait, (2) the service-group capacities s_j, (3) the routing proportions $r_{i,j}$ and (4) the waiting times w_i for each class $i \in \mathscr{C} - \mathscr{I}$. These decision variables are collectively required to satisfy the systems of equations in (2) and (4). (We also require condition (3) to ensure that all service groups in $\mathscr{C} - \mathscr{I}$ operate in the overloaded regime.) But there typically are many immediate-service subsets $\mathscr{I}$ and sets of these variables s_j, $r_{i,j}$ and w_i that will yield a valid steady-state for the fluid model. To discriminate between them, we can impose costs and benefits, similar to those in (9). We consider how we might do that in this section.

Suppose that a reward rate (positive value) $v(i, j, t)$ is earned per unit of fluid per unit time for serving class-i fluid by service group j after these customers have waited time t. This reward is presumably decreasing in the waiting time t. Suppose that a cost rate $c^b(i)$ is incurred per unit of fluid per unit of time for having class-i fluid balk. (Under assumption (3), that cannot be controlled.) Suppose that a cost rate $c^a(i, t)$ is incurred per unit of fluid per unit time for having class-i fluid abandon after having waited time t. This cost is presumably increasing in the time t. Suppose that there is a holding cost rate $c^h(i, x)$ incurred per unit time for having x units of class-i fluid waiting in queue. This cost rate is presumably increasing in the level x.

Then the *total reward rate per unit time*, as a function of the decision variables $\mathscr{I}$, $\mathbf{s} \equiv (s_j)$, $\mathbf{r} \equiv (r_{i,j})$ and $\mathbf{w} \equiv (w_i : i \in \mathscr{C} - \mathscr{I})$ is

$$\begin{aligned} R \equiv R(\mathscr{I}, \mathbf{s}, \mathbf{r}, \mathbf{w}) = &\sum_{i \in \mathscr{I}} \left(\lambda_i \sum_{j \in \mathscr{S}_i} r_{i,j} v(i, j, 0) \right) \\ &+ \sum_{i \in \mathscr{C} - \mathscr{I}} \left(\lambda_i (1 - \beta_i) F_i^{\mathrm{c}}(w_i) \sum_{j \in \mathscr{S}_i} r_{i,j} v(i, j, w_i) \right. \\ &- \lambda_i \beta_i c_i^b - \lambda_i (1 - \beta_i) \int_0^{w_i} c^a(i, t) \, \mathrm{d}F_i(t) \\ &\left. - c^h(i, Q_i) \right), \end{aligned} \tag{11}$$

where Q_i is given in (6). Given values of the decision variables $\mathscr{I}$, $\mathbf{r} \equiv (r_{i,j})$, $\mathbf{s} \equiv (s_j)$ and $\mathbf{w} \equiv (w_i)$, we can calculate the total reward and its components. It is also natural to consider the *optimization problem*

$$\text{maximize} \quad R(\mathscr{I}, \mathbf{r}, \mathbf{w}, \mathbf{s}) \tag{12}$$

over the decision variables $\mathscr{I}$, $\mathbf{r} \equiv (r_{i,j})$, $\mathbf{s} \equiv (s_j)$ and $\mathbf{w} \equiv (w_i)$, *subject to the constraints*

$$\begin{aligned} &\sum_{i \in \mathscr{I} \cap \mathscr{C}_j} \lambda_i r_{i,j} m_{s,i,j} \equiv s_{j,I} < s_j, \\ &\sum_{i \in (\mathscr{C} - \mathscr{I}) \cap \mathscr{C}_j} L_{i,j} (1 - \beta_i) F_i^{\mathrm{c}}(w_i) = s_j' \equiv s_j - s_{j,I}, \\ &\sum_{j \in \mathscr{S}_i} r_{i,j} = 1, \qquad \sum_{j \in \mathscr{S} - \mathscr{S}_i} r_{i,j} = 0 \end{aligned} \tag{13}$$

with $s_j \geqslant 0$ and $r_{i,j} \geqslant 0$ for all i and j, assuming condition (3). The first two constraints in (13) are just (2) and (4).

5. Implementation

We briefly discuss two issues in relating the fluid model to actual service systems: (1) coping with stochastic fluctuations, and (2) routing consistent with the fluid model.

5.1. Stochastic fluctuations

Since the fluid model ignores all uncertainty and stochastic fluctuations, it is natural to consider some adjustments to take account of the stochastic fluctuations. That might not be necessary, because balking and abandonment should act to prevent overload. But as a means to address stochastic fluctuations directly, we suggest augmenting the staffing by a *square-root safety factor* [28,34,36,38,39].

Given that we have found the desired staffing vector $\mathbf{s} = (s_1, \ldots, s_n)$, we let the total staff be $\tilde{s} \equiv s + \gamma\sqrt{s}$, where $s \equiv s_1 + \cdots + s_n$ and γ is a *quality-of-service* (*QoS*) *parameter*, with higher γ yielding higher quality of service. Then, as in Eqs. (5.1) and (5.2) of [28], we allocate the spare capacity to the service groups proportionally according to their square roots; i.e. we let

$$\tilde{s}_j = s_j + x\sqrt{s_j}, \quad 1 \leqslant j \leqslant n, \tag{14}$$

where

$$x = \frac{\tilde{s} - s}{\sum_{j=1}^{n} \sqrt{s_j}} = \frac{\gamma\sqrt{s}}{\sum_{j=1}^{n} \sqrt{s_j}}. \tag{15}$$

The QoS parameter γ can be chosen assuming the load in entire system has a normal distribution with mean and variance s.

5.2. Routing

The fluid model is consistent with many different routing schemes, but the performance of the approximation may well depend upon the method used. For example, the routing could be Markovian with routing probabilities coinciding with the proportions $r_{i,j}$. That Markovian routing is evidently asymptotically correct in the heavy-traffic regime (1), but it might not perform so well in practice. One natural improvement is to perform a form of generalized round robin, that deterministically allocated class-i customers to class-j service groups in the right proportions. Such a generalized round-robin scheme eliminates the randomness associated with Markovian routing. But neither of these two routing schemes responds flexibly to the actual state of the system.

As an alternative routing scheme that responds flexibly to the system state, we propose a *dynamic priority scheme* based on a *tracking index*. Let $\hat{r}_{i,j}(t)$ be the proportion of class-i customers that have been routed to service group j during the last time interval of length t (among those that have been routed to some service group during that time). We can then construct a dynamic priority index, such as

$$p_{i,j}(t) \equiv \frac{r_{i,j}}{\hat{r}_{i,j}(t)}. \tag{16}$$

With the dynamic priority index in (16), a new arriving class-i customer at some time would select a free server from among the eligible service groups with free servers, with the service group chosen being the one having the highest priority index $p_{i,j}(t)$ at that arrival instant (among all eligible service groups). A server from service group j who becomes free at some time would select a customer to serve from one of the customer-class queues, with the customer class chosen being the one having the highest priority index $p_{i,j}(t)$ at that arrival instant (among all eligible customer classes). The idea, of course, is that the dynamic priority scheme should assign customers to servers in a way that will produce the desired proportions in the long run, but at the same time, avoid unnecessary server idleness when there are customers requiring service. That is, with the dynamic priority scheme, we hope to ensure that we obtain the available economies of scale from sharing among service groups.

We have only illustrated one way in which a routing policy can be generated from the proportions chosen in the fluid model. It remains to examine alternative routing schemes, to see if they are roughly consistent with the fluid model and if they perform desirably.

6. Conclusions

We have introduced a multi-class deterministic fluid model of an SBR contact center having m customer classes and n service groups, which can be used to study problems of design and control. An important realistic feature is the use of balking and abandonment to ensure stable model behavior, where the net input is balanced by the net output. Another important realistic feature is the use of non-exponential service-time and time-to-abandon distributions. The key to successfully treating the resulting complex non-Markovian model is to: first, consider an idealized view of a large system (with high arrival rate and many servers), which is formalized by the asymptotic regime (1) and, second, to augment the system state by including the time in service and the time in queue.

In this short space we have only been able to present a framework that can be used for further analysis, but we believe that there is much that can be done within that framework. Moreover, within that framework, we have only described the steady-state behavior of a stationary fluid model. As indicated in [21], the framework can also be used to describe time-dependent behavior, of both a non-stationary model (having time-dependent input) and a stationary model with different initial conditions. Further work on time-dependent behavior and supporting stochastic-process limits is in progress.

Acknowledgments

The author was supported by US NSF Grant DMI-0457095.

References

[1] Segal M, Whitt W. A queueing network analyzer for manufacturing. In: Bonatti M, editor. Teletraffic science for new cost-effective systems, networks and services. Proceedings of ITC 12. Amsterdam: North-Holland; 1989. p. 1146–52.

[2] Whitt W. The queueing network analyzer. Bell Syst Tech J 1983;62:2779–815.

[3] Whitt W. Performance of the queueing network analyzer. Bell Syst Tech J 1983;62:2817–43.

[4] Fendick KW, Saksena VR, Whitt W. Dependence in packet queues. IEEE Trans Commun 1989;37:1173–83.

[5] Fendick KW, Whitt W. Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. Proc IEEE 1989;77:171–94.

[6] Sriram K, Whitt W. Characterizing superposition arrival processes in packet multiplexers for voice and data. IEEE J Sel Areas Commun 1986;SAC-4:833–46.

[7] Whitt W. Approximating a point process by a renewal process: the view through a queue, an indirect approach. Manage Sci 1981;27:619–36.

[8] Whitt W. Approximating a point process by a renewal process: two basic methods. Oper Res 1982;30:125–47.

[9] Whitt W. Variability functions for parametric-decomposition approximations of queueing networks. Manage Sci 1995;41:1704–15.

[10] Whitt W. Stochastic-process limits. Berlin: Springer; 2002.

[11] Kuehn PJ. Approximate analysis of general queueing networks by decomposition. IEEE Trans Commun 1979;27:113–26.

[12] Gans N, Koole G, Mandelbaum A. Telephone call centers: tutorial, review and research prospects. Manuf Serv Oper Manage 2003;5:79–141.

[13] Kelly FP. Blocking probabilities in large circuit-switched networks. Adv Appl Probab 1986;18:473–505.

[14] Ross KW. Multiservice loss models for broadband telecommunication networks. Berlin: Springer; 1995.

[15] Whitt W. Blocking when service is required from several facilities simultaneously. AT&T Tech J 1985;64:1807–56.

[16] Iglehart DL. Limit diffusion approximations for the many-server queue and the repairman problem. J Appl Probab 1965;2:429–41.

[17] Mandelbaum A, Massey WA, Reiman MI. Strong approximations for Markovian service networks. Queueing Syst 1998;30:149–201.

[18] Mandelbaum A, Pats G. State-dependent queues: approximations and applications. In: Kelly FP, Williams RJ, editors. Stochastic networks, IMA volumes in mathematics. Berlin: Springer; 1995. p. 239–82.

[19] Duffield NG, Whitt W. Control and recovery from rare congestion events in a large multi-server system. Queueing Syst 1997;26:69–104.

[20] Krichagina EV, Puhalskii AA. A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. Queueing Syst 1997;25:235–80.

[21] Whitt W. Fluid models for multi-server queues with abandonments. Oper Res, forthcoming. Available at http://columbia.edu/∼ww2040.

[22] Whitt W. Staffing a call center with uncertain arrival rate and absenteeism. Prod Oper Manage, forthcoming. Available at http://columbia.edu/∼ww2040.

[23] Armony M, Shimkin N, Whitt W. The impact of delay announcements in many-server queues with abandonments. Working Paper, 2005. Available at http://columbia.edu/∼ww2040.

[24] Ren ZJ, Zhou Y-P. Call center outsourcing: coordinated staffing level and service quality. Working Paper, 2004. University of Washington.

[25] Bassamboo A, Harrison JM, Zeevi A. Design and control of a large call center: asymptotic analysis of an LP-based method. Oper Res, forthcoming.

[26] Bassamboo A, Harrison JM, Zeevi A. Dynamic routing and admission control in high-volume service systems: asymptotic analysis via multi-scale fluid limits. Queueing Syst, forthcoming.

[27] Harrison JM, Zeevi A. A method for staffing large call centers based on stochastic fluid models. Manuf Serv Oper Manage 2005;7:20–36.

[28] Wallace RB, Whitt W. A staffing algorithm for call centers with skill-based routing. Manuf Serv Oper Manage 2005;7:276–294.

[29] Bolotin V. Telephone circuit holding-time distributions. In: Labetoulle J, Roberts JW, editors. Proceedings of the international teletraffic congress, ITC 14. Amsterdam: North-Holland; 1994. p. 125–34.

[30] Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L. Statistical analysis of a telephone call center: a queueing-science perspective. J Am Statist Assoc 2005;100:20–36.

[31] Whitt W. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. Manage Sci 2004;50:1449–61.

[32] Whitt W. Two fluid approximations for multi-server queues with abandonments. Oper Res Lett 2005;33:363–72.

[33] Altman E, Jiménez T, Koole G. On the comparison of queueing systems with their fluid limits. Probab Eng Inf Sci 2001;15:165–78.

[34] Garnett O, Mandelbaum A, Reiman MI. Designing a call center with impatient customers. Manuf Serv Oper Manage 2002;4:208–27.

[35] Jiménez T, Koole G. Scaling and comparison of fluid limits of queues applied to call centers with time-varying parameters. OR Spectrum 2004;26:413–22.

[36] Borst S, Mandelbaum A, Reiman MI. Dimensioning large call centers. Oper Res 2004;52:17–34.

[37] Zeltyn S, Mandelbaum A. Call centers with impatient customers: many-server asymptotics of the $M/M/s+G$ queue. Working Paper, 2005. Available at http://iew3.technion.ac.il/serveng/References.references.html.

[38] Halfin S, Whitt W. Heavy-traffic limits for queues with many exponential servers. Oper Res 1981;29:567–88.

[39] Whitt W. Understanding the efficiency of multi-server service systems. Manage Sci 1992;38:708–23.