OPERATIONS RESEARCH
Vol. 61, No. 2, March–April 2013, pp. 328–343
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.1120.1151

Staffing and Control of Instant Messaging Contact Centers

Jun Luo, Jiheng Zhang
Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology, Hong Kong S.A.R., China {jluolawren@ust.hk, j.zhang@ust.hk}

In addition to traditional call centers, many companies have started building a new kind of customer contact center, in which agents communicate with customers via instant messaging (IM) over the Internet rather than phone calls. A distinctive feature of the service centers based on IM is that one agent can serve multiple customers in parallel. We choose to model such a center as a server pool consisting of many limited processor-sharing servers. We characterize the underlying stochastic processes by establishing a fluid approximation in the many-server heavy-traffic regime. The limiting behavior of the stochastic processes is shown to involve a stochastic averaging principle, and the fluid approximation provides insights into the optimal staffing and control for such service centers.

Subject classifications: many-server queues; limited processor sharing; fluid models; staffing and control.
Area of review: Stochastic Models.
History: Received March 2012; revision received September 2012; accepted December 2012.

1. Introduction

Communicating with customers has become an indispensable part of modern business. Call centers have traditionally played an important role in communication. With the development of technology, instant messaging (IM) over the Internet has become a favored way of communicating in many situations. More and more companies are building IM-based customer contact centers to supplement their traditional call centers. For example, some online stores offer real-time chat so that customers can ask sales representatives for more information about the listed products. On Dell's online store website there is a link leading to "24/7 live sales help." The option "Chat with us" is listed first, together with other options such as "Call us." Some companies such as Hewlett-Packard (HP) even perform remote diagnostics and troubleshooting as part of after-sale service via IM. Communicating via IM has the advantage of efficient information exchange (imagine a sales representative sending a link to the webpage of their products rather than simply describing the products over the phone), but it is not as convenient as a phone call because it is difficult for customers to access the service on the go. In general, an IM conversation may take a longer time than a phone conversation service since the former requires both the user and the agent to read and type to communicate; see Shae et al. (2007). Nevertheless, IM serves as a good alternative channel for communicating with customers. In some industries, such as the online retail industry, this new mode of communication is rapidly gaining popularity. This motivates the study of models for IM-based customer contact centers to better manage such services.

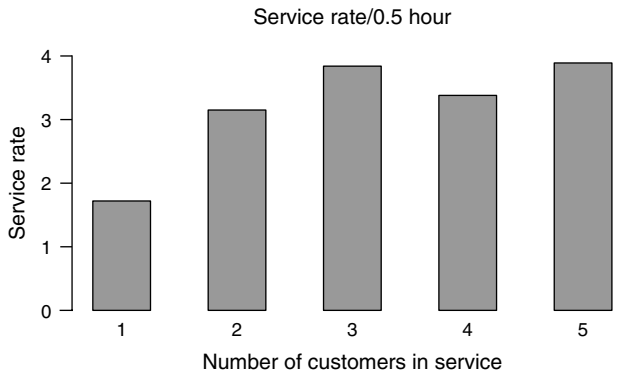
IM-based service centers have some unique features not shared by call centers. An agent (sales representative, technician) at a traditional call center can talk to only one customer at a time, but an agent who is providing service via IM can chat with multiple customers simultaneously. During an IM conversation, customers can be processed in a round-robin fashion—an agent responds to one request and then immediately shifts to another outstanding one from the customers he is serving. Such a system is best modeled using the processor-sharing protocol, where an agent can distribute his attention simultaneously to all customers in service. This modeling method first appeared in computer science, as described in Ritchie and Thompson (1974) and Kleinrock (1976). In many computer systems, a central processing unit handles all active jobs in parallel (a technique known as parallel processing). It should be pointed out that the protocol is an approximation of the actual situation, but the macroscopic model promises to reveal how the system performs in response to changes in the key parameters, which can be identified from historical data.

For this study, we obtained data from a company that is operating a large IM-based service center. The data were recorded using a standard timestamp approach that keeps track of when a customer contacted the center and when a customer service case was opened and closed, etc. From this data set, it emerged that it was hard to identify the required service time of a customer. However, the processor-sharing protocol enables the calculation of the rate γ_k at which an agent can complete cases when there are k customers being served simultaneously. Figure 1 illustrates how the service rate γ_k varies with k. In the data, there are agents serving more than 5 (up to 13) customers

simultaneously. However, we have decided to filter them out, because such "chats" contribute less than 0.1% of the total records, providing too few records to obtain a reliable estimate. One reason an agent is allowed to serve multiple customers is that customers need time to process the information the agent sends to them and to type out their next requests. During this time, the agent would be idling if he is not handling any other customers. Arranging for one agent to serve multiple customers helps reduce such idling, thus making better use of the agents' time. That is why γ_k exhibits an increasing trend when k is small. However, as the number of parallel jobs increases, an agent may become less efficient because of his limited capacity and cognitive issues caused by switching among too many different customers. The pattern of γ_k for large ks is thus uncertain. For this reason, some IM-based service centers enforce a limit on the number of customers an agent is allowed to serve at a time. Thus, our model uses the limited processor-sharing (LPS) protocol for each agent. If all agents have reached the limit, an arriving customer will have to wait to be served. With this background, we develop and evaluate a macroscopic model for an IM-based service center. The model is basically a many-server queue with each server operating under the LPS protocol with state-dependent service rate. We assume that all customers are homogeneous in terms of their requirements, and all servers in the server pool are homogeneous in the sense that they operate at the same state-dependent service rates (see Figure 1).

The remainder of this paper is organized as follows. Section 2 provides a mathematical description of our IM contact center model and presents the concept of asymptotic optimality in the proposed heavy-traffic asymptotic regime. Section 3 introduces a fluid model and uses it to approximate the stochastic model. Based on the approximation, §4 proposes optimal staffing and control policies for IM contact centers. In §5, extensive numerical experiments illustrating both the approximation and the optimality are reported. Section 6 concludes with some remarks and directions for future research. Finally, technical proofs are collected in two appendices, which are part of the electronic companion for the paper (available as supplemental material at http://dx.doi.org/10.1287/opre.1120.1151; proofs in the fluid analysis are in §EC1, and proofs in the stochastic analysis are in §EC2).

Figure 1. Varying service rate.



1.1. Service Quality, Staffing, and Control

In measuring the service qualities in traditional call centers, we often focus on quantities related to the waiting time, such as the expected waiting time or the probability that the waiting time exceeds a certain threshold. This makes sense because once a customer manages to get hold of an agent on the phone, he will exclusively have the service of the agent. Thus, the length of the actual service time (i.e., the duration of the phone call) depends only on the nature of the particular customer's request. However, with Web-chat services the customer is likely to share an agent with other customers. An agent will normally become less responsive the more customers he has to handle simultaneously. The service time (the period from the start of conversation to when the case is closed) of a customer with a particular service request varies with the number of other customers the agent is also serving in parallel. This is similar to the situation in computer systems where the time it takes for a computer processor to complete a task such as opening the Web browser varies with the number of other tasks the processor is also asked to perform in parallel.

In order to capture this distinctive feature of customer experience in an IM-based service center, we choose a holding-cost function on the system status as the indicator for service quality. We call the agents who are serving k customers "level k agents." The status of the system is defined as the number of customers waiting in queue, and the number of agents in each level (a rigorous description will be given in the modeling part). The rationale behind using a holding cost is Little's law, which relates customer's sojourn (waiting and actual service) time to the total number of customers in the system. This type of holding-cost function is not new in the literature. For instance, George and Harrison (2001) defined the holding cost as a function of the number of jobs in the system and used it to describe the congestion in a single-server queueing system. In this study, we allow the holding-cost function to be more general than a linear function of the total number of customers in the system. As will be shown in the mathematical model, the cost is allowed to be a general function of the queue length and the number of customers being served by level k agents for each k. This is to reflect different responsiveness that customers may sense while being served by agents of different levels.

Simply increasing staffing will of course reduce the sojourn time of a customer. However, staffing costs are a major part of the operating expenses of such service centers. On the other hand, reducing staffing may lead to poor service quality and the loss of goodwill, or even a direct loss of revenue. For example, customers may be so annoyed by the slow response caused by inadequate staffing that they end up not buying anything from the online retailer. In this paper, the framework allows the arrival process to

be both stationary and time varying. However, for the time-varying arrivals, we still use a stationary staffing policy. This is applicable to cases where staffing cannot be changed frequently or cannot be arranged to achieve a desired time-varying staffing. In a more general sense, optimal time-varying staffing as proposed by Feldman et al. (2008) is an interesting future direction.

In addition to staffing decisions, control decisions are also important in operating such a system. There are basically two types of controls: admission control and routing control. In this study, we do not reject customers. Admission control determines whether to admit a customer into *service* immediately upon arrival. This control is implemented by setting a control threshold K, which is the maximum number of customers an agent can serve simultaneously. If all agents are serving K customers each, then an arriving customer will have to wait in queue. Otherwise, the arrival is admitted into service. Routing control, on the other hand, is a lot more complicated because there are many ways of assigning arriving customers to agents. A routing control policy must be specified in order to operate the system. The design of optimal routing control policy alone is a very interesting research direction (see Tezcan 2011 for study of the optimal routing policy in steady state). In this study, we adopt the simplest and possibly the most widely used routing policy. Each new arrival is assigned to one of the agents with the "lightest load" at that time. An agent is said to have the lightest load if he is handling the least number of customers compared to all other agents. If more than one agent has the lightest load, one is chosen randomly to serve the arriving customer. We show in this paper that even this simple policy gives rise to some complicated issues in studying the system in transient. A larger control threshold K would help reduce customers' waiting time before being served, but it is doubtful whether this strategy is optimal. This study sets out to model the underlying stochastic processes in order to generate some insights into the joint staffing and admission control decisions involved in managing such service centers efficiently.

1.2. An Asymptotic Framework

Balancing staffing costs and service quality will be formulated in this study as a discrete optimization problem (see (12)). To solve this problem in a stochastic environment, we translate it into a continuous and deterministic problem by examining the system in a meaningful limiting regime.

Like call centers, IM-based service centers also employ a large number of agents to handle heavy demand. This motivates the study of models in the many-server heavy-traffic regime proposed in the call center study, which will be formulated in detail in §2.1. The basic idea is to put the stochastic system in a regime where the demand (the arrival rate) increases and the service capacity (the number of servers) also grows to balance out the demand. The service rate of each individual server remains the same. This heavy-traffic formulation is useful in applications involving humans such as operations in call centers and patient-flow management in a hospital, because the management can only increase the number of servers rather than making each individual server work faster to accommodate large demand.

The limit obtained in the heavy-traffic regime serves as an approximation to the original stochastic process. The optimal solution for the continuous and deterministic optimization problem provides an approximately optimal solution for the original problem in the asymptotic sense. Roughly speaking, the difference between the optimal value for the original problem and that for the limiting problem vanishes as the size of the original system approaches infinity.

1.3. Literature Review

To the best of our knowledge, this is the first study to approach the problem in this manner. Tezcan (2011) has completed a parallel study using the same model, but with a different focus. Whereas our study emphasizes both the transient behavior and the steady-state behavior of the stochastic system under a fixed routing policy, Tezcan (2011) studied an optimal routing policy in the steady state. In fact, they showed that under certain assumptions, the optimal routing policy coincides with the one chosen here.

There is a vast literature on call centers, providing the foundation and inspiration for the research reported here. The survey paper of Gans et al. (2003) provided a tutorial on how call centers function and a survey of academic research devoted to the management of their operations. The basic idea in their study is to model a voice call center as a multiserver queue and model each agent as a server serving only one customer at a time. The optimization problem formulated in this study is, however, based on the framework proposed in Borst et al. (2004). It is worth pointing out that Whitt (2006) showed that fluid models can be quite useful in approximating the performance measures of multi-server queues. In a study of an extension of the multiserver queue, where there are multiple customer classes and a single-server pool, Atar et al. (2010, 2011) showed that fluid approximations can be useful in designing optimal control policies for the operations. The work of Mandelbaum et al. (1998) and Puhalskii (2008) provided a nice theoretical framework for the study of many-server queues (see Mandelbaum et al. 1998 for a general network of many-server queues) with exponentially distributed service times. Their works helped build the foundation for some of the methodologies in this study.

The methodology in this study involves averaging principles. Only a handful of studies in the queueing literature involve averaging principles (see Whitt 2002 for a review). Some notable works include Coffman et al. (1995), which studied the diffusion limit of a two-queue polling

model with asymptotically negligible switchover times and Coffman et al. (1998), which studied the same subject but with nonnegligible switchover times. Recently, Perry and Whitt, Perry and Whitt, Perry and Whitt, Perry and Whitt (2011a, b, c; 2012) studied an extension of such a principle, which they named a "stochastic averaging principle" in Perry and Whitt (2011c), to obtain both the fluid and diffusion limits for an overloaded X-model of multiserver queues proposed in Perry and Whitt (2009). Their approximations also led to useful insights about the asymptotic optimal control of the system. Some of their methodologies were based on the one developed by Hunt and Kurtz (1994), who exploited martingales and random measures. The work of Hunt and Kurtz (1994) considered large loss networks with a large family of control policies, building on a fundamental theory of Kurtz (1992). Although based on different models, Hunt and Kurtz (1994), Perry and Whitt (2012) have inspired some of the methods adopted in this study to deal with a very similar stochastic averaging principle involved in this model.

The LPS protocol is a key feature of the study. Zhang and Zwart (2008) and Zhang et al. (2009, 2011) studied extensively models with a single LPS server. In their studies, both fluid and diffusion limits were established to approximate the transient behavior of the underlying stochastic processes. They have also studied the steady-state limits of the system and validated the interchange of heavy-traffic and steady-state limits. This justifies the use of the steady-state of the diffusion limit, which is tractable, to approximate the steady state of the original system. Closed-form formulae were provided in Zhang and Zwart (2008) to reveal how the performance measures depend on the system parameters. Recently, Gupta and Zhang (2011) also studied a single LPS server with a state-dependent service rate. Their model is closely related to the model studied here, where the service rate of each LPS server also depends on the state (the number of customers in service).

2. Model Formulation and Asymptotic Framework

2.1. The Stochastic Model

Consider a sequence of stochastic systems indexed by n. In the nth system, there are N^n agents, which are modeled as a server pool with N^n homogeneous LPS servers. Each agent can process multiple customers simultaneously. Let K be the maximum number of customers each agent can handle at any time, which is called control threshold throughout this paper. The state of the server pool can then be described using a $(K+1)$-dimensional vector $Z^n(t) = (Z_0^n(t), Z_1^n(t), \ldots, Z_K^n(t)) \in \mathbb{N}^{K+1}$. For each $k \in \{0, 1, \ldots, K\}$, $Z_k^n(t)$ denotes the number of agents who are serving k customers at time t. We call them "level k agents." Note that $Z_0^n(t)$ is the number of idling agents, and we have

$$\sum_{k=0}^{K} Z_k^n(t) = N^n, \quad t \geqslant 0. \tag{1}$$

When all agents are each serving K customers, i.e., $Z_K^n(t) = N^n$, an arriving customer must wait in a buffer. We assume that waiting customers are served based on the first-come-first-served principle. Let $Q^n(t)$ denote the number of customers who are waiting for service at time t. In what follows, we assume that no customer waits in queue if there is an agent who is serving fewer than K customers, i.e.,

$$Q^n(t)(N^n - Z_K^n(t)) = 0, \quad t \geqslant 0. \tag{2}$$

Customers arrive to the nth system according to a general nonhomogeneous Poisson process $\Lambda^n(t)$ with intensity function $\lambda^n(t)$. As mentioned above, if all agents are serving K customers, then an arrival has to wait. Otherwise, the arriving customer is assigned to one of the agents who has the "lightest load" at the time. If there are multiple agents with the same "lightest load," one is chosen randomly to serve the arrival. Mathematically, we introduce the index process

$$i_*^n(t) = \min\{0 \leqslant k \leqslant K\colon Z_k^n(t) > 0\} \tag{3}$$

to identify the lightest load at time t. The process $i_*^n(t)$ serves as the indicator of how arrivals should be assigned to agents. For example, if $i_*^n(t_-) = 0$, then an arrival at time t is assigned to an idling agent, yielding $Z_0^n(t) = Z_0^n(t_-) - 1$ and $Z_1^n(t) = Z_1^n(t_-) + 1$. If $i_*^n(t_-) = K$, then an arrival at time t joins the queue, incrementing the queue size by 1.

The data we have collected indicates that any realistic model must allow an agent's service speed to vary depending on how many customers he is serving simultaneously. Let γ_k denote the service rate of a level k agent. In this paper, we assume that the service times are exponentially distributed. Let $S_k^n(t), k = 1, \ldots, K$ be independent Poisson processes with rate 1. Then the total number of customers who have been served by level k agents by time t is

$$D_k^n(t) = S_k^n\left(\gamma_k \int_0^t Z_k^n(s)\, ds\right). \tag{4}$$

With the arrival, assignment, and service processes thus defined, the following stochastic dynamic equations describe the evolution of the nth system.

$$Z_0^n(t) = Z_0^n(0) - \int_0^t \mathbf{1}_{\{i_*^n(s-)=0\}}\, d\Lambda^n(s) + D_1^n(t), \tag{5}$$

$$\begin{aligned} Z_k^n(t) = Z_k^n(0) &+ \int_0^t \mathbf{1}_{\{i_*^n(s-)=k-1\}}\, d\Lambda^n(s) \\ &- \int_0^t \mathbf{1}_{\{i_*^n(s-)=k\}}\, d\Lambda^n(s) - D_k^n(t) \\ &+ \int_0^t \mathbf{1}_{\{Q^n(s-)=0\}}\, dD_{k+1}^n(s), \quad 0 < k < K, \end{aligned} \tag{6}$$

$$\begin{aligned} Z_K^n(t) = Z_K^n(0) &+ \int_0^t \mathbf{1}_{\{i_*^n(s-)=K-1\}}\, d\Lambda^n(s) \\ &- \int_0^t \mathbf{1}_{\{Q^n(s-)=0\}}\, dD_K^n(s), \end{aligned} \tag{7}$$

$$\begin{aligned} Q^n(t) = Q^n(0) &+ \int_0^t \mathbf{1}_{\{i_*^n(s-)=K\}}\, d\Lambda^n(s) \\ &- \int_0^t \mathbf{1}_{\{Q^n(s-)>0\}}\, dD_K^n(s). \end{aligned} \tag{8}$$

Note that the indicator function $\mathbf{1}_{\{Q^n(s-)=0\}}$ in (6) is only effective when $k = K-1$. When we study the evolution of $Z^n_{K-1}(\cdot)$ at time epoch s, the level K may happen to be $Z^n_K(s-) = N^n$. Suppose there is a service completion from level K agents at time s, the status of the system immediately after the service completion depends on whether there are customers in the queue. If the queue is not empty, then a customer in queue is immediately served upon a service completion. So $Z^n_{K-1}(s)$ stays at 0 and $Z^n_K(s)$ stays at N^n. Only when there are no customers waiting in the queue, does $Z^n_{K-1}(s)$ increase by 1 and $Z^n_K(s)$ decrease by 1.

2.2. The Heavy-Traffic Regime and Fluid Scaling

For the sequence of systems indexed by n, let the arrival rate and the number of agents grow in proportion to n as n increases to infinity, while keeping the service rate $\{\gamma_k, k = 0, 1, \ldots, K\}$ fixed. Therefore, we assume the following heavy-traffic assumption throughout this paper.

Assumption 1 (Heavy Traffic). *The arrival rate and the number of agents of the nth system satisfy the condition that $\bar{\lambda}^n(\cdot)$ is bounded and*

$$\bar{\lambda}^n(t) = \frac{\lambda^n(t)}{n} \to \lambda(t), \tag{9}$$

$$\bar{N}^n = \frac{N^n}{n} \to N, \tag{10}$$

as $n \to \infty$, for some function $\lambda(t)$ and $N > 0$.

The fluid scaling for the processes Λ^n, Q^n, and Z^n can be defined as

$$\bar{\Lambda}^n(t) = \frac{\Lambda^n(t)}{n}, \qquad \bar{Q}^n(t) = \frac{Q^n(t)}{n},$$
$$\bar{Z}^n_k(t) = \frac{Z^n_k(t)}{n}, \quad k = 0, \ldots, K. \tag{11}$$

The relevant heavy-traffic regime is essentially the many-server heavy-traffic regime studied in the call center literature, but there are quite interesting limiting dynamics that are different from the call center models. In the heavy-traffic regime the size of the system grows in proportion to n, and the fluid scaling (11) divides all the quantities by n, making the processes "smooth" at the limit. However, the stochastic dynamics Equations (5)–(8) heavily rely on the index process $i^n_*(t)$. Note that $i^n_*(t)$ does not scale like $Z^n(t)$. It jumps on the fixed discrete grid $\{0, 1, \ldots, K\}$ for all n. Therefore, unlike traditional stochastic processes that are smoothed out in the heavy-traffic limit, $i^n_*(t)$ is a jump process that instead oscillates infinitely often in heavy traffic. In fact, the index process fluctuates more and more frequently as n increases. Figure 2 depicts one simulated sample path to show the evolution of the index process in comparison with the process $Z^n(t)$ when $\lambda^n = 400$, $N^n = 200$, and $K = 6$. In fact, the process $i^n_*(t)$ fluctuates so frequently that only lines consisting of dense dots are apparent at this level of resolution. The oscillation of $i^n_*(t)$ brings complexity in applying traditional fluid approximations to the stochastic system, and motivates seeking another approach involving the stochastic averaging principle to understand the dynamics of the system.

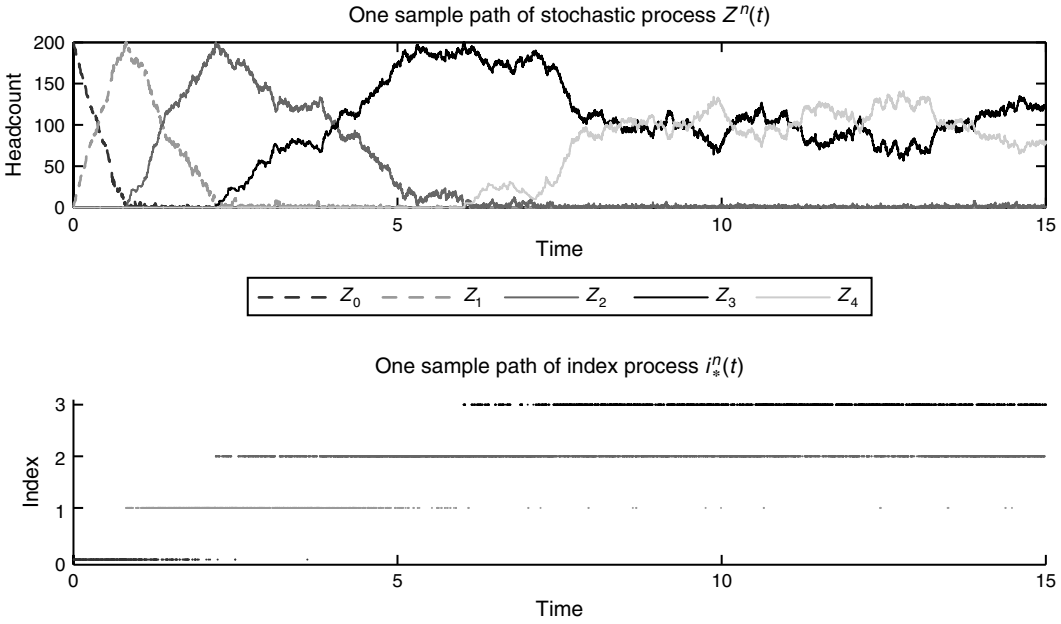
2.3. A Cost Function and Asymptotic Optimality

The system manager for such a service center confronts a joint problem of staffing and control. The manager needs to balance his choice of staffing level N^n and control threshold K, because an agent works at varying speeds depending on the number of customers he is handling. Consider a general holding-cost function of the system status. Let $h(z, q)$ denote the cost per unit of time the system is in state (z, q). Let c denote the cost of employing an agent per unit of time. For staffing level N^n and control level k, consider the normalized average cost over the time horizon $[0, T]$:

$$\bar{C}^n_T(\bar{N}^n, K) = c\bar{N}^n + \frac{1}{T}\mathbb{E}\left[\int_0^T h(\bar{Z}^n(s), \bar{Q}^n(s))\, ds\right]. \tag{12}$$

The rationale of considering this type of cost function is to take the service quality into consideration. Traditionally, only holding cost for the queue is considered in many models arising from call center applications, such as Atar et al. (2010) and Bassamboo et al. (2006). In a call center, the service quality depends only on the queue or waiting time in the queue, because a customer's service time depends entirely on the nature of his requirement. However, for IM-based service centers, the actual service time of a customer is significantly affected by the number of other customers sharing the agent during the service period. A longer actual service time means a customer senses worse responsiveness during his service. As pointed in Shae et al. (2007), both time in queue and total service duration are important measures for the quality of IM services. Intuitively, customers being served by a level 2 agent should feel better "responsiveness" than being served by a level 5 agent. This intuition is behind the idea of using a general function h, which can assign different "waiting costs" to customers in different stages. For example, we can set the function to be $h(z, q) = w_0 q + \sum_{k=1}^K k w_k z_k$, where $w_0, \ldots, w_K$ are weights. The condition on the cost function is stated in Assumption 2. Suppose we specialize the cost function to be $h(z, q) = \lambda^{-1}(q + \sum_{k=1}^K k z_k)$; then, Little's law gives an intuitive explanation in the stationary case. The holding cost is then essentially counted into the customers' average sojourn time, which is the metric used by Tezcan (2011). Putting the holding cost and the staffing cost together is in the same spirit as the cost function in Bassamboo et al. (2006). The objective is to optimize the trade-off between personnel cost and the holding cost. To arrive at a the general formulation for the cost function and also provide a solution to asymptotically optimize the total average cost, define (as in Bassamboo et al. 2006) the *asymptotic optimality* in both finite-horizon and infinite-horizon cases.

Figure 2. One sample path of the stochastic process $Z^n(t)$ and the index process $i^n_*(t)$ for system n with $\lambda^n = 400$, $N^n = 200$, $K = 6$, and $\gamma = (1, 1.6, 1.8, 2.2, 2.3, 2.4)$.



Definition 1. A sequence of staffing level $\{\bar{N}^n_*\}$ and a control threshold K_* are said to be asymptotically optimal for the finite horizon $[0, T]$ if for any other sequence of staffing levels $\{N^n\}$ and control threshold K,

$$\limsup_{n\to\infty} \bar{C}^n_T(\bar{N}^n_*, K_*) \leqslant \liminf_{n\to\infty} \bar{C}^n_T(\bar{N}^n, K). \tag{13}$$

A sequence of staffing level $\{\bar{N}^n_*\}$ and a control threshold K_* are said to be asymptotically optimal for the infinite horizon $[0, \infty]$ if for any other sequence of staffing levels $\{N^n\}$ and control threshold K,

$$\limsup_{T\to\infty} \limsup_{n\to\infty} \bar{C}^n_T(\bar{N}^n_*, K_*) \leqslant \liminf_{T\to\infty} \liminf_{n\to\infty} \bar{C}^n_T(\bar{N}^n, K). \tag{14}$$

We now propose a fluid model to approximate the performance of the stochastic system in the heavy-traffic regime, and then connect the asymptotic optimization problem to the fluid model.

3. Fluid Approximations in the Heavy-Traffic Regime

Such a complicated system is not amenable to exact analysis, so approximations to the original stochastic system are essential. The idea of approximating the complicated underlying stochastic processes is to use a fluid model, which is analogous to the original stochastic model with all the randomness removed by replacing the stochastic processes with their rate functions. However, due to the involvement of the index process, the reduction for such models is quite difficult. In this section, we first consider a fluid model, which can be formulated using a set of ordinary differential equations (ODEs). After justifying its validity, we will show that the solution to the ODEs approximates the fluid scaled stochastic processes in the heavy-traffic regime.

3.1. A Fluid Model

Let

$$I(z) = \min\{0 \leqslant k \leqslant K\colon z_k > 0\} \tag{15}$$

denote the smallest index of z's nonzero component. In fact, the stochastic index process defined in (3) can be written as $i^n_*(t) = I(\bar{Z}^n(t))$. The simulation presented in Figure 2 shows that the stochastic process for all levels seems to be "smoothed out," despite the fact that the index process cannot be. An appropriate fluid model must characterize how customer arrivals are allocated to the server pool, which is based on the index process. For this purpose, we introduce the mapping $f\colon [0, N]^{K+1} \times \mathbb{R}_+ \to [0, 1]^{K+1}$,

$$f(z, \lambda) = \big(f_0(z, \lambda), f_1(z, \lambda), \ldots, f_K(z, \lambda)\big),$$

where each component $f_k(z, \lambda)$ is formally defined as

$$f_k(z, \lambda) = \begin{cases} \dfrac{\gamma_{k+1} z_{k+1}}{\lambda} \wedge 1, & k = I(z) - 1, \\ \left(1 - \dfrac{\gamma_k z_k}{\lambda}\right)^+, & k = I(z), \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

Intuitively, $f_k(z, \lambda)$ indicates the fraction of the arrival stream that is injected into level k, whereas the current state

of the "fluid" server pool is z and the arrival rate is λ. We will show the connection to the stochastic model in §3.2 when we analyze the underlying stochastic processes.

It is clear that the fluid model lives in the space

$$\mathbb{S}=\left\{(z_0,\dots,z_K,q)\in\mathbb{R}_+^{K+2}\colon \sum_{k=0}^{K} z_k=N \text{ and } q(N-z_K)=0\right\}. \quad (17)$$

An ODE of the form

$$(z'(t),q'(t))=\Psi(t,z(t),q(t)) \quad (18)$$

can then be used to define the fluid model. For clearer presentation, divide space $\mathbb{S}$ into two subspaces $\mathbb{S}=\mathbb{S}_+\cup\mathbb{S}_0$, where

$$\mathbb{S}_+=\{(z,q)\in\mathbb{S}\colon q>0\},\qquad \mathbb{S}_0=\{(z,q)\in\mathbb{S}\colon q=0\}.$$

In space $\mathbb{S}_0$, ODE (18) takes the form

$$z_0'(t)=-f_0(z(t),\lambda(t))\lambda(t)+\gamma_1 z_1(t), \quad (19)$$

$$z_k'(t)=f_{k-1}(z(t),\lambda(t))\lambda(t)-f_k(z(t),\lambda(t))\lambda(t) -\gamma_k z_k(t)+\gamma_{k+1}z_{k+1}(t),\quad 0<k<K, \quad (20)$$

$$z_K'(t)=f_{K-1}(z(t),\lambda(t))\lambda(t)-\gamma_K z_K(t), \quad (21)$$

$$q'(t)=f_K(z(t),\lambda(t))\lambda(t), \quad (22)$$

and in the space $\mathbb{S}_+$, the ODE (18) takes the form

$$z_k'(0)=0,\quad 0\leqslant k\leqslant K, \quad (23)$$

$$q'(t)=\lambda(t)-\gamma_K N. \quad (24)$$

The transitions between $\mathbb{S}_0$ and $\mathbb{S}_+$ occurs at the critical point $(z,q)=(0,\dots,0,N,0)$, where all the agents reach the threshold K. Whether the solution to the ODE will stay in $\mathbb{S}_0$ or transit to $\mathbb{S}_+$ depends on whether or not $\lambda(t)\leqslant\gamma_K N$. Despite the complicated form of (19)–(24), the following theorem shows that ODE (18) is well defined.

THEOREM 1 (EXISTENCE AND UNIQUENESS). *Assume that $\lambda(t)$ is a continuous function of t. There exists a unique solution to the* ODE (18) *specified by* (19)–(24)*, with the initial condition $(z(0),q(0))\in\mathbb{S}$.*

The proof of this theorem is available in the e-companion, §EC1. This result justifies the existence and uniqueness of the solution to the fluid model, thus providing a foundation for the rest of the study in this paper. To make the result more applicable, it would be helpful to extend it to a case with a more general arrival process. Suppose the arrival rate $\lambda(t)$ is a piecewise-continuous function. Let $0<t_1<t_2<\cdots$ be the jump points of $\lambda(t)$. Solving ODE (18) in the time interval $[0,t_1]$ gives a unique solution. Considering t_1 as the initial time point, the ODE can be then studied in the next time interval $[t_1,t_2]$. Iteratively, we can thus show the existence and uniqueness of the solution to the ODE over the entire time horizon.

COROLLARY 1. *Assume that $\lambda(t)$ is a piecewise-continuous function of t. There then exists a unique solution to* ODE (18) *with initial condition $(z(0),q(0))\in\mathbb{S}$.*

3.2. Stochastic Analysis

It is now necessary to show that the well-defined fluid model serves as an approximation for the fluid-scaled stochastic processes in the heavy-traffic regime. Let $\mathbb{D}([0,T],\mathbb{R}^{K+2})$ be the space of all $\mathbb{R}^{K+2}$-valued functions on $[0,T]$, which are right continuous with left limits.

THEOREM 2 (FWLLN). *Under Assumption* 1*, if the initial states converge in distribution to some constants, i.e.,*

$$(Z^n(0)/n,Q^n(0)/n)\Longrightarrow(z(0),q(0)),\quad \text{as } n\to\infty, \quad (25)$$

for some $(z(0),q(0))\in\mathbb{S}$, then the fluid-scaled process $(\bar{Z}^n,\bar{Q}^n)$ converges in distribution to the fluid model solution (z,q) in Theorem 1*, i.e., in the space $\mathbb{D}([0,T],\mathbb{R}^{K+2})$ equipped with uniform topology,*

$$(\bar{Z}^n(t),\bar{Q}^n(t))\Longrightarrow(z(t),q(t)),\quad \text{as } n\to\infty, \quad (26)$$

where (z,q) is the solution to the ODE (18) *with initial condition $(z(0),q(0))$.*

For the solution (z,q), define the associated *fluid cost* as

$$C_T(N,K)=cN+\frac{1}{T}\int_0^T h(z(s),q(s))\,ds. \quad (27)$$

Based on Theorem 2, it can be shown that the expected cost will converge to the fluid cost. We require some additional assumption on the holding-cost function.

ASSUMPTION 2. *The holding-cost function h is a nondecreasing continuous function with respect to each component. In addition, we assume there exist an α, A, and C such that*

$$h(2Ne,q)\leqslant A\exp(\alpha q/2)\quad \text{for all } q>C, \quad (28)$$

where e is the $(K+1)$-dimensional vector with each component being 1*. In other words, we assume that the "tail" of the holding cost in queue does not grow faster than all exponential functions.*

COROLLARY 2. *Under the same condition as Theorem* 2*, if the holding-cost function h satisfies Assumption* 2 *and*

$$\sup_n \mathbb{E}[\exp(\alpha\bar{Q}^n(0))]<\infty, \quad (29)$$

then

$$\bar{C}_T^n(\bar{N}^n,K)\to C_T(N,K),\quad \text{as } n\to\infty. \quad (30)$$

The proof of this result is available in the e-companion, §EC2.

REMARK 1. Assumption 2 is in fact quite general. All polynomial functions clearly satisfy it. The condition (29) is required mainly for technical reason. Alternatively, we may assume that initially no customers wait in queue, which is a reasonable assumption in this application.

The rest of this section will be devoted to establishing Theorem 2. The essential connection between the fluid and the stochastic models lies in the index process i_*^n and the function f defined in (16). Consider a small interval $[t, t+\delta]$. The number of arrivals in that interval who are assigned to a level k agent is $\int_t^{t+\delta} \mathbf{1}_{\{i_*^n(s-)=k\}}\, d\bar{\Lambda}^n(s)$, and the amount of fluid injected into z_k is $f_k(z(t), \lambda(t))\lambda(t)\delta$. Informally, the basic principle behind the convergence result in Theorem 2 is that

$$\lim_{\delta\to 0}\lim_{n\to\infty}\frac{1}{\delta}\int_t^{t+\delta}\mathbf{1}_{\{i_*^n(s-)=k\}}\, d\bar{\Lambda}^n(s) = f_k(z(t),\lambda(t))\lambda(t). \quad (31)$$

The interplay here between the $i_*^n(t)$ and $\bar{Z}^n(t)$ in this model is quite interesting. The process $\bar{Z}^n(t)$ evolves slowly and determines the transition rates for $i_*^n(t)$, whereas the process $i_*^n(t)$ evolves quickly and its "steady state" determines the evolution of $\bar{Z}^n(t)$. To see this intuitively, replace $\bar{\Lambda}^n(t)$ by λt in the above, yielding

$$\frac{1}{\delta}\int_t^{t+\delta}\mathbf{1}_{\{i_*^n(s-)=k\}}\lambda\, ds = \frac{1}{n\delta}\int_0^{n\delta}\mathbf{1}_{\{i_*^n(t+(s-)/n)=k\}}\lambda\left(t+\frac{s-}{n}\right) ds.$$

When n becomes large, what determines that the above integral is actually the "steady state" of the process $i_*^n(t+\frac{\cdot}{n})$. The above is just an informal illustration of the stochastic averaging principle involved in the model. The coexistence of two different time scales requires an untraditional method to analyze the stochastic model in the limiting regime. One idea is to use the stochastic averaging principle to prove the convergence (31). However, because $i_*^n(t)$ depends on a multidimensional Markov process $(\bar{Z}^n(t), \bar{Q}^n(t))$, a direct analysis using the stochastic averaging principle may be complicated. Instead, we propose to use an approach involving random measures and martingale representation. The approach was initiated by Hunt and Kurtz (1994), and has been adopted by Perry and Whitt (2012).

We now provide the proof for Theorem 2. Define the random measure ν^n by

$$\nu^n([0,t]\times A) = \int_0^t \mathbf{1}_{\{Z^n(s-)\in A\}}\, ds, \quad (32)$$

for any $t>0$ and subset $A\subset \bar{\mathbb{Z}}_+^K$, where $\bar{\mathbb{Z}}_+ = \mathbb{Z}_+\cup\{+\infty\}$. This is a common approach to compactify the space. The interested reader may refer to Kurtz (1992) and Perry and Whitt (2012) for detailed discussions. Consider the space $\mathbb{M}$ of all measures ν on the product space $[0,\infty)\times\bar{\mathbb{Z}}_+^K$ satisfying $\nu([0,t]\times\bar{\mathbb{Z}}_+^K) = t$ for all $t>0$. Endowing $\mathbb{M}$ with the Prohorov metric as in (1.1) of Kurtz (1992), then $\mathbb{M}$ inherits the compactness because $\bar{\mathbb{Z}}_+^K$ is compact. This will provide convenience for the proofs later on. Let $\mathcal{A}_k = \{z\in\bar{\mathbb{Z}}^K\colon z_k>0 \text{ and } z_j=0, j<k\}$. The indicator function of the index process can then be written as

$$\mathbf{1}_{\{i_*^n(t-)=k\}} = \mathbf{1}_{\{Z^n(t-)\in\mathcal{A}_k\}}.$$

Define martingales related to the arrival and service processes

$$\bar{M}_a^n(t) = \bar{\Lambda}^n(t) - \int_0^t \bar{\lambda}^n(s)\, ds, \quad (33)$$

$$\bar{M}_k^n(t) = \frac{1}{n}\left(S_k^n\left(\gamma_k\int_0^t Z_k^n(s)\, ds\right) - \gamma_k\int_0^t Z_k^n(s)\, ds\right), \quad k=1,\dots,K. \quad (34)$$

Using the random measure ν^n and the above-introduced martingales, the fluid-scaled stochastic dynamic Equations (5)–(8) can be written as

$$\bar{Z}_0^n(t) = \bar{Z}_0^n(0) - \int_0^t \mathbf{1}_{\{Z^n(s-)\in\mathcal{A}_0\}}\, d\bar{M}_a^n(s) + \bar{M}_1^n(t) - \int_{[0,t]\times\mathcal{A}_0}\bar{\lambda}^n(s)\nu^n(ds\times dy) + \gamma_1\int_0^t \bar{Z}_1^n(s)\, ds, \quad (35)$$

$$\begin{aligned}\bar{Z}_k^n(t) = \bar{Z}_k^n(0) &+ \int_0^t \mathbf{1}_{\{Z^n(s-)\in\mathcal{A}_{k-1}\}}\, d\bar{M}_a^n(s) - \int_0^t \mathbf{1}_{\{Z^n(s-)\in\mathcal{A}_k\}}\, d\bar{M}_a^n(s) \\ &- \bar{M}_k^n(t) + \int_0^t \mathbf{1}_{\{\bar{Q}^n(s-)=0\}}\, d\bar{M}_{k+1}^n(s) \\ &+ \int_{[0,t]\times\mathcal{A}_{k-1}}\bar{\lambda}^n(s)\nu^n(ds\times dy) - \int_{[0,t]\times\mathcal{A}_k}\bar{\lambda}^n(s)\nu^n(ds\times dy) \\ &- \gamma_k\int_0^t \bar{Z}_k^n(s)\, ds + \gamma_{k+1}\int_0^t \mathbf{1}_{\{\bar{Q}^n(s-)=0\}}\bar{Z}_{k+1}^n(s)\, ds, \quad 0<k<K, \end{aligned} \quad (36)$$

$$\begin{aligned}\bar{Z}_K^n(t) = \bar{Z}_K^n(0) &+ \int_0^t \mathbf{1}_{\{Z^n(s-)\in\mathcal{A}_{K-1}\}}\, d\bar{M}_a^n(s) - \int_0^t \mathbf{1}_{\{\bar{Q}^n(s-)=0\}}\, d\bar{M}_K^n(s) \\ &+ \int_{[0,t]\times\mathcal{A}_{K-1}}\bar{\lambda}^n(s)\nu^n(ds\times dy) - \gamma_K\int_0^t \mathbf{1}_{\{\bar{Q}^n(s-)=0\}}\bar{Z}_K^n(s)\, ds, \end{aligned} \quad (37)$$

$$\begin{aligned}\bar{Q}^n(t) = \bar{Q}^n(0) &+ \int_0^t \mathbf{1}_{\{Z^n(s-)\in\mathcal{A}_K\}}\, d\bar{M}_a^n(s) - \int_0^t \mathbf{1}_{\{\bar{Q}^n(s-)>0\}}\, d\bar{M}_K^n(s) \\ &+ \int_{[0,t]\times\mathcal{A}_K}\bar{\lambda}^n(s)\nu^n(ds\times dy) - \gamma_K\int_0^t \mathbf{1}_{\{\bar{Q}^n(s-)>0\}}\bar{Z}_K^n(s)\, ds. \end{aligned} \quad (38)$$

The following lemma establishes that the above stochastic processes are relatively compact and gives some preliminary characterization of the limit.

Lemma 1. *Under Assumption* 1, *if* (25) *holds, then the sequence* $\{(\bar{Z}^n, \bar{Q}^n), \nu^n\}_{n\in\mathbb{N}}$ *is relatively compact in the*

space $\mathbb{D}([0,T],\mathbb{R}^{K+2})\times\mathbb{M}$*, and the limit of any convergent subsequence satisfies*

$$z_0(t)=z_0(t)-\int_{[0,t]\times\mathscr{A}_0}\lambda(s)\nu(dy\times ds)+\gamma_1\int_0^t z_1(s)\,ds, \tag{39}$$

$$z_k(t)=z_k(0)+\int_{[0,t]\times\mathscr{A}_{k-1}}\lambda(s)\nu(dy\times ds)-\int_{[0,t]\times\mathscr{A}_k}\lambda(s)\nu(dy\times ds)-\gamma_k\int_0^t z_k(s)\,ds+\gamma_{k+1}\int_0^t \mathbf{1}_{\{q(s)=0\}}z_{k+1}(s)\,ds,\quad 0<k<K, \tag{40}$$

$$z_K(t)=z_K(0)+\int_{[0,t]\times\mathscr{A}_{K-1}}\lambda(s)\nu(dy\times ds)-\gamma_K\int_0^t \mathbf{1}_{\{q(s)=0\}}z_K(s)\,ds, \tag{41}$$

$$q(t)=q(0)+\int_{[0,t]\times\mathscr{A}_K}\lambda(s)\nu(dy\times ds)-\gamma_K\int_0^t \mathbf{1}_{\{q(s)>0\}}z_K(s)\,ds. \tag{42}$$

The proofs of Lemma 1 and Lemma 3 are available in the e-companion, §EC2. To further study the limit in the above lemma, we need to characterize the limiting measure ν. The following lemma is taken from Kurtz (1992), which states that the measure ν has the product form.

Lemma 2 (Kurtz 1992). *Let* $\{(z,q),\nu\}$ *be the limit of a convergent subsequence of the processes* $\{(\bar{Z}^n,\bar{Q}^n),\nu^n\}_{n\in\mathbb{N}}$. *Then for all measurable subsets* Γ *of* $[0,T]$ *and* A *of* $\bar{\mathbb{Z}}_+^{K+1}$

$$\nu(\Gamma\times A)=\int_\Gamma \pi_s(A)\,ds, \tag{43}$$

where π_s *is a probability measure on* $\bar{\mathbb{Z}}_+^{K+1}$ *for all* $s\geqslant 0$.

This is a very useful result. It says that the measure ν on the product space $[0,t]\times\bar{Z}_+^K$ can be separated in product form. In other words,

$$\int_{[0,t]\times A}\lambda(s)\nu(dy\times ds)=\int_0^t \pi_s(A)\lambda(s)\,ds. \tag{44}$$

To characterize the probability measure π_s, we introduce the Markov process $m_{x(s)}$ on $\bar{\mathbb{Z}}_+^{K+1}$, where $x=(z,q,\lambda)$ (recall that λ is the limit in Assumption 1). In other words, for each s, $m_{x(s)}$ is a Markov process whose transition rate depends on $x(s)$. For $j=0,1,\ldots,K$, let $e_j=(0,\ldots,0,1,0,\ldots,0)$ be a $(K+1)$-dimensional vector with its $(j+1)$th component being 1 and all the rest being 0. We first define

$$e_j m_{x(s)}=\infty,\quad \text{for all } j \text{ such that } z_j(s)>0.$$

The Markov process $m_{x(s)}$ evolves with the following transition rate when $0\leqslant j<K$,

$$m_{x(s)}\to\begin{cases} m_{x(s)}-e_{j+1}+e_j, & \text{at rate } \gamma_{j+1}z_{j+1}(s),\\ m_{x(s)}-e_j+e_{j+1}, & \text{at rate } \lambda(s)\mathbf{1}_{\{m_{x(s)}\in\mathscr{A}_j\}},\end{cases} \tag{45}$$

and

$$m_{x(s)}\to m_{x(s)}-e_K+e_{K-1},\quad \text{at rate } \mathbf{1}_{\{q(s)=0\}}\gamma_K N. \tag{46}$$

Suppose $I(z(s))=k$ for some $k=0,\ldots,K$; then it is clear that the states in $S_<=\{y\in\bar{Z}_+^{K+1}\colon y_j>0 \text{ for any } j<k-1\}$ are transient for the Markov process, because the rates $\gamma_{j+1}z_{j+1}(s)=0$ for all $j<k-1$. The states in $S_<$ cannot be accessible from the states out of $S_<$, so the Markov process is reducible. In fact, the probability that the Markov process returns to any state in $S_<$ once having left it is 0. Suppose $\max\{k\colon z_k(s)>0\}=k^*$ for some $k^*\geqslant k$; then the states in $S_>=\{y\in\bar{Z}_+^{K+1}\colon y_j>0 \text{ for any } j>k^*\}$ are also transient because $\mathbf{1}_{\{m_{x(s)}\in\mathscr{A}_j\}}=0$ for any $j>k$. It is also clear that $\lim_{t\to\infty}\mathbb{P}(e_j m_{x(s)}(t)=\infty\mid e_j m_{x(s)}(0)=x)=1$ for any initial state x and $k\leqslant j\leqslant k^*$. Therefore, the only interesting component is the kth component, i.e., $e_{k-1}m_{x(s)}$. Let $\tau_<$ be the first time the Markov process exits $S_<$; then the component $e_{k-1}m_{x(s)}$ evolves as a birth–death process with birth rate $\gamma_k z_k(s)$ and death rate $\lambda(s)$. Denote by $\mathbb{P}_\infty$ the steady-state probability of the Markov process $m_{x(s)}$. We are only concerned with calculating $\mathbb{P}_\infty(m_{x(s)}\in\mathscr{A}_j)$ for $0\leqslant j\leqslant K$ for the purposes of this discussion. So, essentially we only need to focus on the evolution of the kth and the $(k+1)$th components of $m_{x(s)}$. Because $e_{k-1}m_{x(s)}$ is a birth–death process, the stationary distribution of $m_{x(s)}$ satisfies

$$\mathbb{P}_\infty(m_{x(s)}\in\mathscr{A}_j)=0,\quad j<k-1, \tag{47}$$

$$\mathbb{P}_\infty(m_{x(s)}\in\mathscr{A}_{k-1})=\frac{\gamma_k z_k(s)}{\lambda(s)}\wedge 1. \tag{48}$$

Because the $(k+1)$th component of $m_{x(s)}$ is defined to be infinity, we have

$$\mathbb{P}_\infty(m_{x(s)}\in\mathscr{A}_k)=\left(1-\frac{\gamma_k z_k(s)}{\lambda(s)}\right)^+, \tag{49}$$

$$\mathbb{P}_\infty(m_{x(s)}\in\mathscr{A}_j)\quad=0,\quad j>k+1. \tag{50}$$

The following lemma helps to connect the above-defined Markov process with the probability π_s in Lemma 2.

Lemma 3. *If* $I(z(s))=k$ *for some* $0\leqslant k\leqslant K$, *then for all bounded function* $g\colon Z_+^{K+1}\to\mathbb{R}$,

$$\int_{\bar{\mathbb{Z}}^{K+1}}\left\{\sum_{j=0}^{k\wedge(K-1)}[g(y-e_j+e_{j+1})-g(y)]\mathbf{1}_{\{y\in\mathscr{A}_j\}}\lambda(s)+\sum_{j=k}^{K}[g(y-e_j+e_{j-1})-g(y)]\mathbf{1}_{\{q(s)=0\}}\gamma_j z_j(s)\right\}\cdot\pi_s(dy)=0, \tag{51}$$

where π_s *is defined in Lemma* 2.

With the above preparation, we are now ready to present the proof of the main result.

Proof of Theorem 2. It now remains to show that the limit $((z, q), \nu)$ satisfies the ODE (18). According to (39)–(42), and (44), we need only show that

$$\pi_s(\mathcal{A}_k) = f_k(z(s), \lambda(s)), \tag{52}$$

where $f_k(z(s), \lambda(s))$ is defined as in (16). Suppose that $I(z(s)) = 0$, then the Markov process degenerates to $e_0 m_{x(s)} = \infty$. Therefore, $\mathbb{P}_\infty(m_{x(s)} \in \mathcal{A}_0) = 1$, which is consistent with (16). Suppose $I(z(s)) = k$ for some $1 \leqslant k \leqslant K - 1$, then $q(s) = 0$. According to (51) with $\mathbf{1}_{\{q(s)=0\}}$ being just 1, it follows in Ethier and Kurtz (1986, Proposition 4.9.2) that π_s is the stationary distribution for the Markov process $m_{x(s)}$. Then (52) follows from (47)–(50). Suppose $I(z(s)) = K$; then there are two cases. The first is the case where $q(s) = 0$. In this case, the situation is the same as that discussed above. The second case, where $q(s) > 0$, is actually quite easy. According to (45) and (46), the rate $m_{x(s)} \to m_{x(s)} - e_{j+1} + e_j$ is 0 for all $0 \leqslant j < K$. Therefore, all the states in $\{y \in \bar{Z}_+^{K+1}\colon y_j > 0 \text{ for any } j < K\}$ are transient. It is clear in this case that $\pi_s(\mathcal{A}_k) = 0$ for all $k < K$; thus, $\pi_s(\mathcal{A}_K) = 1$. Plugging π_s into (39)–(42) and separating the expressions into the two cases, depending on whether or not $q(s) > 0$, yields the ODEs (19)–(21) or (23)–(24), respectively.

4. Asymptotic Optimal Staffing and Control Policies

4.1. An Asymptotically Optimal Policy When the Planning Horizon Is Finite

Let us first develop a connection between the asymptotic optimization problem proposed in §2.3 and the fluid model. Let (z, q) denote the solution to the fluid model characterized by ODE (18). Due to the tractability of the deterministic process (z, q), what can be done in general is to numerically solve the optimization problem:

$$\begin{aligned} &\text{minimize} \quad C_T(N, K) \\ &\text{subject to} \quad N > 0, \ K \in \mathbb{N}. \end{aligned} \tag{53}$$

Under additional assumptions, we will show later that there are closed-form solutions when we consider the infinite-horizon ($T \to \infty$) problem. Let (N_*, K_*) be an optimal solution to (53). Because the fluid model serves as a reasonable approximation for this complicated system, one would expect that the optimal solution based on the fluid model might suggest an asymptotically optimal solution for the stochastic problem.

Theorem 3. *If $\bar{N}_*^n \to N_*$ as $n \to \infty$, then the sequence of staffing level $\{\bar{N}_*^n\}$ and the control threshold K_* are asymptotically optimal.*

Proof. Pick any sequence of staffing levels $\bar{N}^n$ and a control threshold K. For any convergent subsequence $\{\bar{N}^{n_l}\}$, suppose that $\bar{N}^{n_l} \to N_s$ as $n_l \to \infty$. It follows from Corollary 2 and optimization problem (53) that

$$\begin{aligned} \lim_{n_l \to \infty} \bar{C}_T^{n_l}(\bar{N}^{n_l}, K) &= C_T(N_s, K) \\ &\geqslant C_T(N_*, K_*) = \lim_{n \to \infty} \bar{C}_T^n(\bar{N}_*^n, K_*). \end{aligned}$$

Because the above inequality holds for any convergent subsequence of $\{\bar{N}^n\}$ and any control level K, the sequence $\{\bar{N}_*^n\}$ and K_* satisfy the definition of asymptotic optimality in Definition 1.

Theorem 3 prescribes a numerical approach to asymptotically solve the joint staffing and control problem for managing an IM-based service center. Due to the time-varying arrival rate and the complexity of the underlying model, the objective function in (53), which involves the solution to a set of ODEs, is extremely complicated. Nevertheless, the solution to the ODEs are tractable in the sense that optimization problem (53) can be solved numerically. The solution we provide is particularly helpful when the arrival rate varies and staffing cannot be adjusted as quickly. We illustrate through a numerical example in §5.3 that is contrary to the common sense, it is not always optimal to set the control threshold to be the level where agents achieve the greatest efficiency.

4.2. An Asymptotically Optimal Policy When the Planning Horizon Is Infinite

Consider now the stationary case where the arrival rate $\lambda(\cdot)$ is constant. In this case, the outcome of interest is the long-run average cost over an infinite-time horizon, i.e., $\lim_{T \to \infty} \bar{C}_T^n(\bar{N}^n, K)$. An additional assumption in this case is the monotonicity of the state-dependent service rate,

$$0 < \gamma_1 < \gamma_2 < \cdots. \tag{54}$$

Interested readers are referred to Tezcan (2011) for the supporting logic of this assumption. Mathematically, the sequence $\{\gamma_k\}$ is allowed to increase to infinity or to be bounded. For stability reasons, we must also require that

$$N > \lambda / \sup_k \gamma_k. \tag{55}$$

As a consequence of (54) and (55), there exist some k' such that

$$\gamma_{k'} N \leqslant \lambda < \gamma_{k'+1} N. \tag{56}$$

Define $\tilde{z}(N)$ to be the point where

$$\tilde{z}_k(N) = \begin{cases} 0, & k < k', \\ \dfrac{\gamma_{k'+1} N - \lambda}{\gamma_{k'+1} - \gamma_{k'}}, & k = k', \\ \dfrac{\lambda - \gamma_{k'} N}{\gamma_{k'+1} - \gamma_{k'}}, & k = k' + 1, \\ 0, & k > k' + 1. \end{cases} \tag{57}$$

PROPOSITION 1. *Assume that* $\lambda(t) \equiv \lambda$ *and* (54)–(55) *hold. For any control threshold* $K > k'$, *the point* $(\tilde{z}(N), 0)$ *with* $\tilde{z}(N)$ *defined by* (57) *is an invariant point of the fluid model. For any fluid model solution* (z, q) *with* $(z(0), q(0)) \in \mathbb{S}$,

$$(z(t), q(t)) \to (\tilde{z}(N), 0), \quad \text{as } t \to \infty.$$

The proof of this proposition is available in the e-companion, §EC1. Based on this proposition, it is easy to see that the fluid cost

$$\lim_{T\to\infty} C_T(N, K) \to cN + h(\tilde{z}(N), 0) \triangleq C(N). \tag{58}$$

Let N_* denote an optimal solution to the problem

$$\begin{aligned} &\text{minimize} \quad C(N) \\ &\text{subject to} \quad N > 0. \end{aligned} \tag{59}$$

THEOREM 4. *If* $\bar{N}^n_* \to N_*$ *as* $n \to \infty$, *then the sequence of staffing level* $\{\bar{N}^n_*\}$ *and any control threshold* $K_* > k'$ *are asymptotically optimal for the long-run average cost on infinite time horizon.*

PROOF. The proof of this result is similar to the one for Theorem 3 in invoking Corollary 2 and checking the requirement of asymptotic optimality in Definition 1. In addition, the result of Proposition 1 is also needed, due to the infinite horizon.

For staffing levels $\bar{N}^n$ and control threshold K, let $\{\bar{N}^{n_l}\}$ be a convergent subsequence such that $\bar{N}^{n_l} \to N_s$ as $n_l \to \infty$. By Corollary 2 and optimization problem (59),

$$\begin{aligned} \lim_{T\to\infty}\lim_{n_l\to\infty} \bar{C}^{n_l}_T(\bar{N}^{n_l}, K) &= \lim_{T\to\infty} C_T(N_s, K) \\ &= C(N_s) \geqslant C(N_*) \\ &= \lim_{T\to\infty} C_T(N_*, K_*) \\ &= \lim_{T\to\infty}\lim_{n\to\infty} \bar{C}^n_T(\bar{N}^n_*, K_*). \end{aligned}$$

Because the above inequality holds for any convergent subsequence of $\{\bar{N}^n\}$ and any control level K, the sequence $\{\bar{N}^n_*\}$ and K_* satisfy the definition of asymptotic optimality in Definition 1.

The optimization problem (59) in some cases can be solved explicitly. For example, when the holding cost is a linear function $h(z, 0) = h\sum_{k=0}^{K} kz_k$ where h is a positive constant. In this case, the objective function (59) is a piecewise-linear function in N. For each interval, $(\lambda/\gamma_{k+1}, \lambda/\gamma_k]$, $C(N)$ takes a linear form. Because the optimal value for a linear programming always occurs at the boundary, optimization problem (59) becomes

$$\begin{aligned} &\text{minimize} \quad \lambda c\frac{1}{\gamma_k} + \lambda h\frac{k}{\gamma_k} \\ &\text{subject to} \quad k \in \{1, 2, \ldots, K\}. \end{aligned} \tag{60}$$

Then the optimal level where all agents should be is simply $k' = \arg\min_k(\lambda/\gamma_k)(c + hk)$, and the best staffing level is $N^* = \lambda/\gamma_{k'}$. A numerical example will be presented in §5.4. We just point out here that despite the monotonicity of γ_i, the graph of function $C(N)$ (e.g., Figure 7(a)) may still zigzag quite irregularly, rather than being convex as the total cost function does in many ostensibly similar applications. It is thus quite important to quantitatively calculate which level is best and choose the appropriate staffing to reach that level. We also demonstrate a numerical example in §5.4 where the holding-cost function h is not linear. It is interesting to see that in this case, the steady state of the system may be somewhere between two levels rather than focusing on one level.

5. Numerical Experiments

In this section, we present some of the numerical experiments we have carried out on the IM-based service center model. The main purpose is to confirm our understanding of how the stochastic process works, and test the approximations obtained from the asymptotic analysis. We also illustrate through some examples the importance of using quantitative insights to guide the design and operation of such service centers.

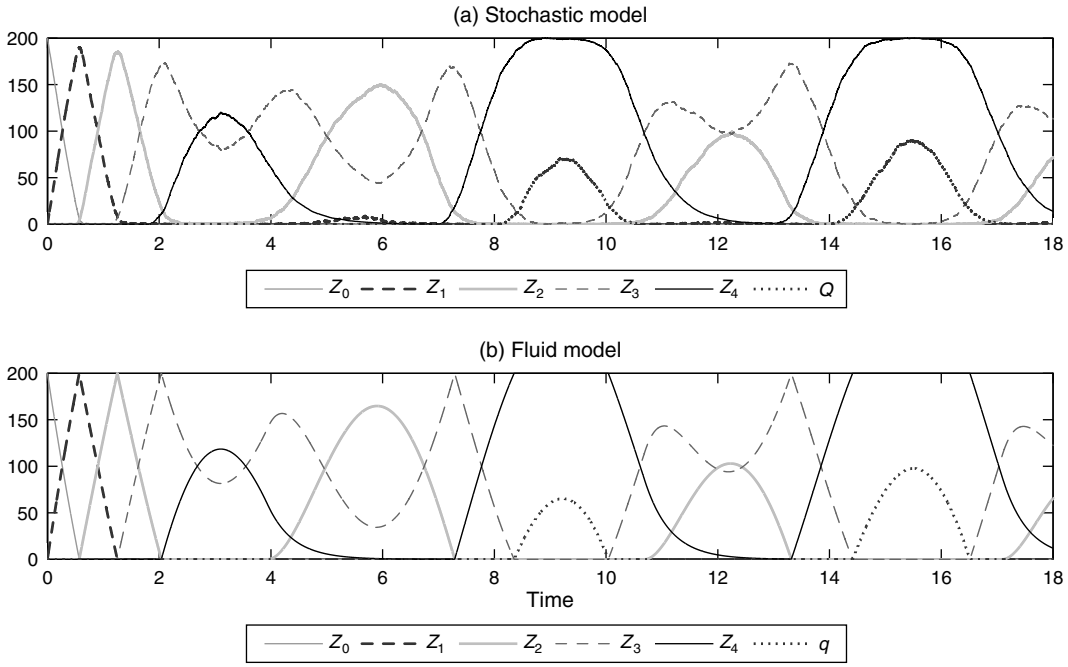
5.1. Validity of the Transient Approximation

In order to demonstrate the fluid approximation, we simulate a system with number of agents $N^{200} = 200$. Each agent can serve at most $K = 4$ customers, and the service rate $\gamma = (1, 1.6, 1.8, 2.2)$. Because time-varying arrivals are allowed in the analysis, we set the arrival rate to be $\lambda^{200}(t) = 200\lambda(t)$, where $\lambda(t) = 2 + 1\sin(t)$. Figure 3 gives an overview of the system's evolution with time. The upper graph depicts an aggregation of 30 simulated sample paths, and the lower graph draws the trajectory of the fluid model obtained by solving the ODE (18). Thirty sample paths are aggregated to reduce the stochastic fluctuation, which the fluid model cannot capture. To obtain a better idea how close the fluid approximation is, the fluid model solution and the aggregate of 30 simulated sample paths are overlaid in Figure 4. Systems of three different sizes $N^n = n$, $n = 50, 100, 200$ are simulated. The corresponding arrival rates are scaled accordingly: $\lambda^n(t) = n\lambda(t)$, $n = 50, 100, 200$. For comparison purposes, the fluid-scaled sample paths, i.e., $n^{-1}Z^n(t)$ and $n^{-1}Q^n(t)$, are plotted. To save space, only level 2 and queue are shown in the figure; the comparisons for the other levels are similar. The approximation becomes more accurate for larger systems.

5.2. Validity of the Steady-State Approximation

In this section, we study the approximation for the steady state of the system using the invariant state of the fluid model in Proposition 1. Consider an example where $K = 6$ with the service rate $\gamma = (1, 1.6, 1.8, 2.2, 2.3, 2.4)$. Choose the system size to be $N^{200} = 200$ with the arrival rate fixed

Figure 3. Simulated stochastic model and the fluid model.



to be a constant $\lambda^n = 390$. Figure 5 depicts the aggregate of 30 simulated sample paths over a relatively long time horizon. It shows that the system "stabilizes" in the state where about 62.5% of the agents are in level 3 and the rest in level 4. With this set of parameters, we can easily calculate by (57) that the invariant point is $\tilde{z} = (0, 0, 0, 5/8, 3/8, 0, 0)$ and $\tilde{q} = 0$.

It is worth pointing out that the approximation using the fluid invariant performs well not only for systems with exponential service times, but also for systems with general service times. We simulate the system with three different service-time distributions—exponential(1), Erlang$(2, 0.5)$ and log-normal $(1, 4)$—which all have mean 1. The control threshold is set at $K = 6$, with service rate $\gamma = (1, 1.6, 1.8, 2.2, 2.3, 2.4)$. The system size is $N^{200} = 200$, and the arrival rate is $\lambda^{200} = 390$. We ran simulation experiments for 16 independent replications with the three service-time distributions over a relatively long time horizon $[0, 10^4]$. Table 1 reports both the estimates and the 95% confidence intervals. The "Approximation" column is calculated based on the invariant state (57) for the fluid model, with sojourn time being calculated via Little's law.

Figure 4. Comparisons of the simulated stochastic model and the fluid model for systems of different sizes.

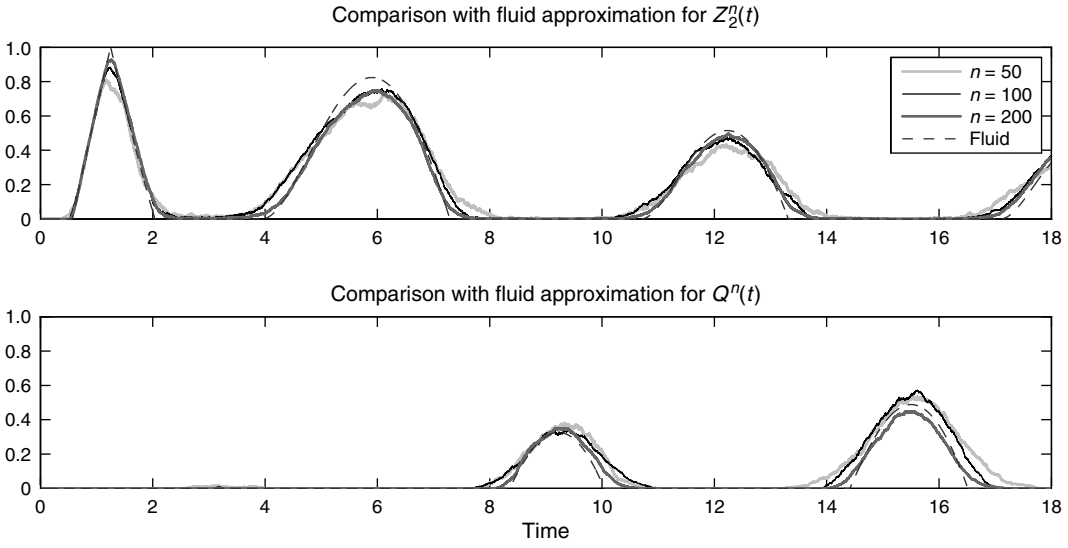
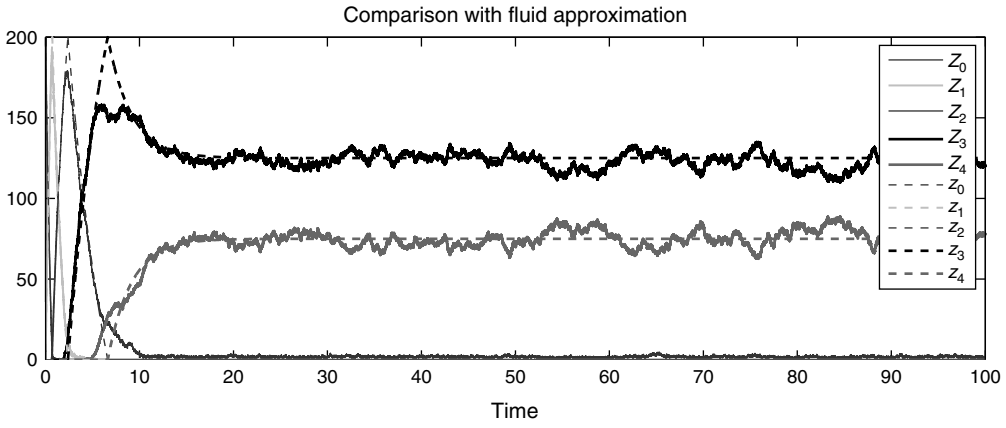


Figure 5. Simulated long-term behaviors of the stochastic model and the fluid model.



5.3. Optimal Staffing and Control with Time-Varying Arrivals on a Finite Horizon

Consider now a staffing and control problem with time-varying arrivals. Assume that the service rate $\gamma = (2.0, 3.0, 2.7, 3.2)$. Thus, the most efficient level is level 4, where an agent achieves maximum service speed. However, it may not always be optimal to set the control threshold at 4, as demonstrated in the following numerical study. In fact, one can use the fluid approximation and Theorem 3 to serve as a quantitative guide.

We now illustrate the usefulness of the quantitative insights through a concrete example. Suppose the service center needs to cater for the time-varying demand depicted in Figure 6(a). We use a scaled log-normal density function, $\lambda(t) = 2.5 + 0.76(0.02t\sqrt{2\pi})^{-1}\exp(-(1/2)\log^2(0.02t))$, to mimic the unimodal shape of the arrival rate over a planning horizon of length 100. Assume that the holding-cost function is $h(z, q) = 1 \times (\sum_k kz_k + q)$ (in other words, the holding cost is linear and the rate is equal to 1) and the

Table 1. Comparison of fluid approximations with simulation estimates of steady-state performance measures with general service-time distribution.

Performance	Exponential	Erlang-2	LN(1, 4)	Approximation
Level 0	0.0004	0.0003	0.0005	0
	±0.0008	±0.0007	±0.0010	—
Level 1	0.0088	0.0084	0.0102	0
	±0.0032	±0.0024	±0.0059	—
Level 2	1.7325	1.7174	1.7553	0
	±0.0201	±0.0154	±0.0310	—
Level 3	122.2821	122.2991	122.3772	125
	±0.3716	±0.2532	±0.4488	—
Level 4	75.9753	75.9740	75.8561	75
	±0.3837	±0.2649	±0.4683	—
Level 5	0.0010	0.0006	0.0007	0
	±0.0007	±0.0003	±0.0006	—
Level 6	0	0	0	0
Sojourn time	1.7287	1.7287	1.7283	1.7308
	±0.0007	±0.0004	±0.0011	—

Figure 6. The arrival rate, and the fluid cost associated with different control thresholds and number of agents.

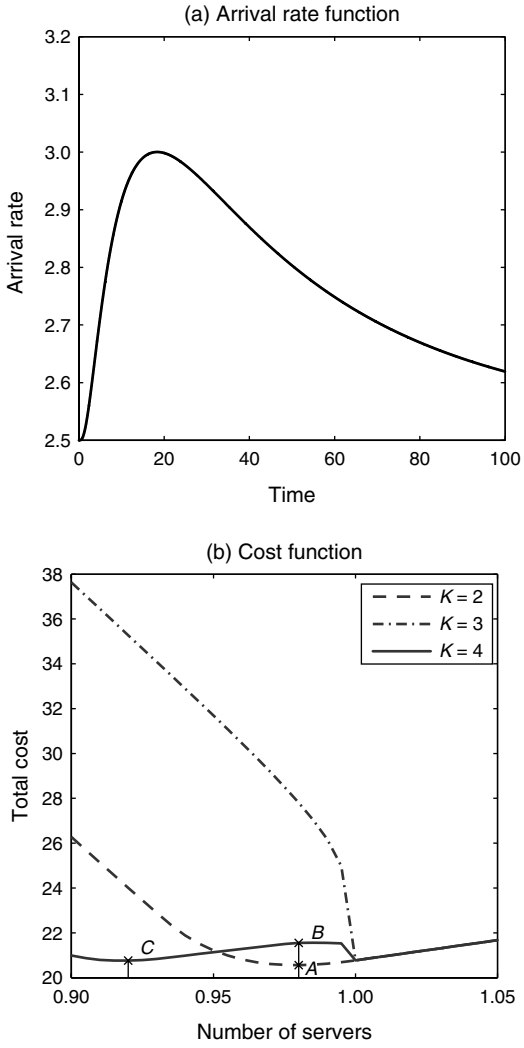


Table 2. Comparison between the expected cost and the fluid cost.

Staffing and control	Fluid cost	Expected cost	95% C.I.
A $(N^n = 196, K = 2)$	4,112.4	4,129.4	[4,112.6, 4,146.1]
B $(N^n = 196, K = 4)$	4,310.1	4,270.2	[4,252.9, 4,287.5]
C $(N^n = 183, K = 4)$	4,152.1	4,146.6	[4,131.0, 4,162.1]

staffing cost is $c = 19$. We plot the "fluid" cost for different control thresholds $K = 2, 3, 4$ in Figure 6(b). Clearly, the cost varies depending on the control threshold. For different Ks, the minimum occurs at different staffing levels. For $K = 4$, the minimum occurs at $N = 0.915$. However, the minimum for $K = 2$ occurs at $N = 0.98$. In this example, the optimal solution for problem (53) is $(N_*, K_*) = (0.98, 2)$, which corresponds to point A on the graph. This emphasizes the importance of making a joint decision. Even if a service center chooses the correct staffing level, but the wrong control threshold (e.g., $K = 4$), then it will experience a significantly higher cost at point B. Similarly, if a corrected threshold is chosen ($K = 2$ in this example), a wrong staffing decision would make the cost at some other points on the red dotted line, which is higher than the optimal.

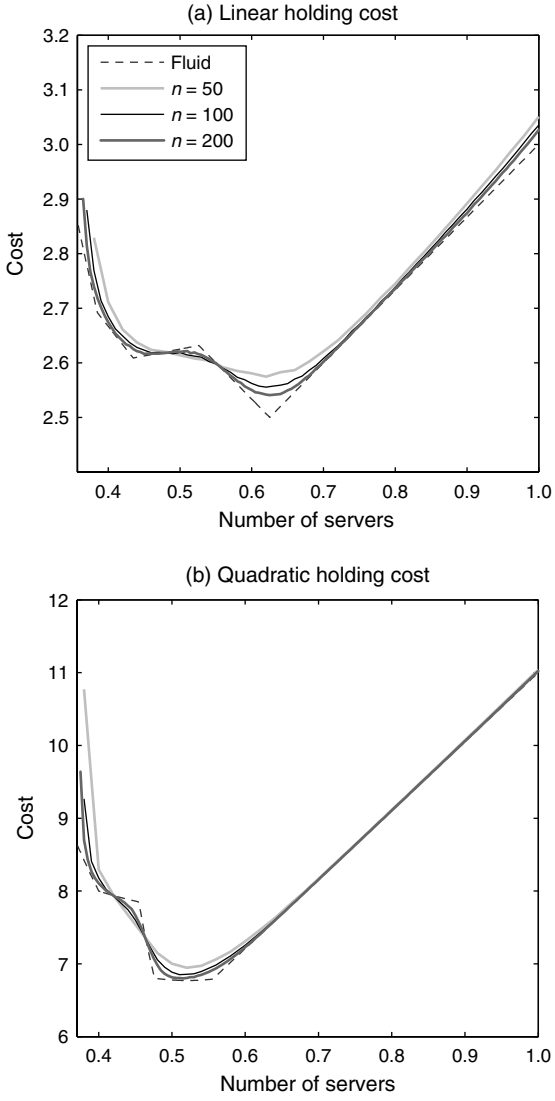
To illustrate how the optimal solution based on the fluid model helps with design and control, we simulate a stochastic system of scale $n = 200$. In other words, the system is fed with a nonhomogeneous Poisson process with rate $n\lambda(t)$. The staffing level corresponding to A and B in Figure 6 is $N^n = 200 \times 0.98 = 196$ and that for C is $N^n = 200 \times 0.915 = 183$. Table 2 summarizes the predicted costs based on the fluid model for different joint decisions A, B, and C on the graph, and compares them with simulations of the corresponding stochastic systems. Notice that choosing the correct staffing level but wrong control threshold (point B) would incur 3.4% more expected cost compared with the optimal choice (point A), quite consistent with the 4.8% increase predicted by the fluid cost. It is worth pointing out that this example also demonstrates that it may not always be optimal to set the control threshold at the most efficient level, i.e., where γ_k achieves its maximum.

5.4. Optimal Staffing and Control with Constant Arrivals on an Infinite Horizon

To illustrate the model for the case of a linear holding cost, set the arrival rate $\lambda = 1$, the staffing cost $c = 2$, and the holding-cost function to be $h(z, q) = 1 \times (\sum_k kz_k + q)$. Assume the service rate $\gamma = (1, 1.6, 1.9, 2.3, 2.6, 2.8)$, and the control threshold is set at $K = 6$. The fluid cost calculated using (58) is plotted in Figure 7(a) indicated by the dotted line.

To illustrate the case of nonlinear holding cost, set the arrival rate to be $\lambda = 1$, the staffing cost rate to be $c = 10$, and the holding-cost function to be $h(z, q) = (\sum_k kz_k + q)^2$.

Figure 7. Total cost function with linear and quadratic holding-cost functions.



The service rate is assumed to be $\gamma = (1, 1.8, 2.1, 2.2, 2.5, 2.7)$, and the control threshold is set at $K = 6$. Figure 7(b) plots the corresponding fluid cost.

In both examples, we also plot the expected cost estimated via simulation for systems of different scales $n = 50, 100, 200$. Both graphs show that the fluid approximation is suitable for staffing purposes, because the expected costs of the stochastic systems dips and peaks with the corresponding fluid cost. However, at some staffing levels, the approximation is not close enough to obtain an accurate performance evaluation, which is beyond the aim of this paper. For a more accurate performance evaluation, in particular for the turning points (where all agents are expected to be in the same level), more refined approximations such as a diffusion approximation are required.

It is also important to point out that the simulated cost functions exhibit "flat bottom" in both examples. This

suggests that the choice of staffing level can be quite robust. A relatively wide range of choices of the staffing level gives similar costs that are close to the optimum.

6. Conclusions and Future Work

In this paper, we study a new type of service centers where agents communicate with customers via instant messaging. A distinctive feature is that each agent can serve multiple customers simultaneously. This makes modeling and analysis more challenging than for traditional call centers. This study has shown that such a service center can be modeled as a pool of many homogeneous servers, each operating under the processor-sharing protocol. The number of customers an agent can serve at one time is limited, and the threshold is determined by a control policy. We provide an asymptotic analysis for the underlying process in the many-server heavy-traffic regime, which is widely used to study call centers. We obtain an approximation by using the stochastic averaging principle in the heavy-traffic analysis. The approximation helps to characterize the complicated queueing model using an ODE. Because the solution to the ODE is tractable, the approximation can then be applied to solve staffing and control problems. Our numerical experiments confirm that the approximation is reasonably good in both transient and steady-state studies. We also demonstrate via a few numerical examples how to use the approximation to guide the staffing and control of such service centers.

These results suggest quite a few interesting directions for future study. First, this study considers only a simple control policy that assigns arrivals to one of the agents with the lightest load. In fact, Tezcan (2011) has studied a more complicated control policy that skips some "inefficient" levels. Certainly, there are many interesting problems in the routing of arriving customers. Second, abandonment also happens in such service centers due to the impatient nature of human beings. Slow response from an agent handling too many customers may make a customer abandon during service. It would be interesting to allow a customer's abandonment rate to depend on the number of other customers being served by the same agent. Third, in this study we have assumed that service times are exponentially distributed, which facilitated the analysis. Future research might fruitfully apply the framework of measure-valued process to study the model with generally distributed service times. And finally, this study is limited to the fluid approximation, which relies on the insights of the functional law of large numbers. The functional central limit theorem might be applied to obtain a more refined approximation to the underlying stochastic processes.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.1120.1151.

Acknowledgments

This research was supported by the Hong Kong Research Grants Council [Grants 622110 and 624012].

References

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

Atar R, Giat C, Shimkin N (2011) On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost. *Queueing Systems* 67(2):127–144.

Bassamboo A, Harrison JM, Zeevi A (2006) Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.* 54(3):419–435.

Borst S, Mandelbaum A, Reiman MI (2004) Dimensioning large call centers. *Oper. Res.* 52(1):17–34.

Coffman EG Jr, Puhalskii AA, Reiman MI (1995) Polling systems with zero switchover times: A heavy-traffic averaging principle. *Ann. Appl. Probab.* 5(3):681–719.

Coffman EG Jr, Puhalskii AA, Reiman MI (1998) Polling systems in heavy traffic: A Bessel process limit. *Math. Oper. Res.* 23(2): 257–304.

Ethier SN, Kurtz TG (1986) *Markov Processes: Characterization and Convergence.* Wiley Series in Probability and Statistics (John Wiley & Sons, Inc., New York).

Feldman Z, Mandelbaum A, Massey WA, Whitt W (2008) Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* 54(2):324–338.

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

George JM, Harrison JM (2001) Dynamic control of a queue with adjustable service rate. *Oper. Res.* 49(5):720–731.

Gupta V, Zhang J (2011) Limited processor sharing queues with state dependent service rates. Technical report, The Hong Kong University of Science and Technology, Hong Kong.

Hunt PJ, Kurtz TG (1994) Large loss networks. *Stochastic Processes and Their Appl.* 53(2):363–378.

Kleinrock L (1976) Queueing systems. *Computer Applications*, Vol. 2 (Wiley-Interscience, New York).

Kurtz TG (1992) Averaging for martingale problems and stochastic approximation. *Applied Stochastic Analysis*. Lecture Notes in Control and Information Sciences, Vol. 177 (Springer, Berlin), 186–209.

Mandelbaum A, Massey WA, Reiman MI (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30(1–2):149–201.

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

Perry O, Whitt W (2011a) Diffusion approximation for an overloaded X model via an averaging principle. Working paper, Columbia University, New York.

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

Perry O, Whitt W (2011c) An ODE for an overloaded X model involving a stochastic averaging principle. *Stochastic Systems* 1:17–66.

Perry O, Whitt W (2012) A fluid limit for an overloaded X model via a stochastic averaging principle. *Math. Oper. Res.*, ePub ahead of print December 20, http://dx.doi.org/10.1287/moor.1120.0572.

Puhalskii A (2008) The $M_t/M_t/K_t+M_t$ queue in heavy traffic. http://arxiv.org/pdf/0807.4621.pdf.

Ritchie DM, Thompson K (1974) The UNIX time-sharing system. *J. ACM* 17(7):365–375.

Shae Z-Y, Garg D, Bhose R, Mukherjee R, Guven S, Pingali G (2007) Efficient Internet chat services for help desk agents. *IEEE Internat. Conf. Services Comput., SCC 2007* (IEEE, Piscataway, NJ), 589–596.

Tezcan T (2011) Design and control of customer service chat systems. Technical report, University of Rochester, Rochester, NY.

Whitt W (2002) *Stochastic-Process Limits.* Springer Series in Operations Research (Springer-Verlag, New York).

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

Zhang J, Zwart B (2008) Steady state approximations of limited processor sharing queues in heavy traffic. *Queueing Systems* 60(3–4):227–246.

Zhang J, Dai JG, Zwart B (2009) Law of large number limits of limited processor-sharing queues. *Math. Oper. Res.* 34(4):937–970.

Zhang J, Dai JG, Zwart B (2011) Diffusion limits of limited processor-sharing queues. *Ann. Appl. Probab.* 21(2):745–799.

Jun Luo is a Ph.D. student in the Department of Industrial Engineering and Logistics Management at the Hong Kong University of Science and Technology. His research interests include stochastic modeling and simulation, with their applications in service operations management and optimization via simulation.

Jiheng Zhang is an assistant professor in the Department of Industrial Engineering and Logistics Management at the Hong Kong University of Science and Technology. His research interests are in performance evaluation and optimal control via asymptotic analysis of queueing systems arising from applications in manufacturing and services.