This article was downloaded by: [152.1.60.159] On: 03 December 2020, At: 09:50
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Operations Research

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

To Pool or Not to Pool: Queueing Design for Large-Scale Service Systems

Ping Cao, Shuangchi He, Junfei Huang, Yunan Liu

To cite this article:
Ping Cao, Shuangchi He, Junfei Huang, Yunan Liu (2020) To Pool or Not to Pool: Queueing Design for Large-Scale Service Systems. Operations Research

Published online in Articles in Advance 03 Dec 2020

. https://doi.org/10.1287/opre.2019.1976

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2020, INFORMS

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

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

OPERATIONS RESEARCH
Articles in Advance, pp. 1–20
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

To Pool or Not to Pool: Queueing Design for Large-Scale Service Systems

Ping Cao,[a] Shuangchi He,[b] Junfei Huang,[c] Yunan Liu[d]

[a] School of Management, University of Science and Technology of China, 230026 Hefei, China; [b] Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 117576; [c] Department of Decision Sciences and Managerial Economics, CUHK Business School, Chinese University of Hong Kong, Shatin, Hong Kong; [d] Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina 27695

Contact: pcao@ustc.edu.cn, https://orcid.org/0000-0001-7801-9754 (PC); heshuangchi@nus.edu.sg, https://orcid.org/0000-0003-4107-3946 (SH); junfeih@cuhk.edu.hk, https://orcid.org/0000-0002-3764-354X (JH); yliu48@ncsu.edu, https://orcid.org/0000-0001-9961-2610 (YL)

Received: June 25, 2018
Revised: September 11, 2019; November 15, 2019
Accepted: November 21, 2019
Published Online in Articles in Advance: December 3, 2020

OR/MS Subject Classifications: queues: approximations, balking and reneging, limit theorems; probability: stochastic model applications
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2019.1976

Copyright: © 2020 INFORMS

Abstract. There are two basic queue structures commonly adopted in service systems: the pooled structure, where waiting customers are organized into a single queue served by a group of servers, and the dedicated structure, where each server has her own queue. Although the pooled structure, known to minimize the servers' idle time, is widely used in large-scale service systems, this study reveals that the dedicated structure, along with the join-the-shortest-queue routing policy, could be more advantageous for improving certain performance measures, such as the probability of a customer's waiting time being within a delay target. The servers' additional idleness resulting from the dedicated structure will be negligible when the system scale is large. Using a fluid model substantiated by asymptotic analysis, we provide a performance comparison between the two structures for a moderately overloaded queueing system with customer abandonment. We intend to help service system designers answer the following question: To reach a specified service-level target, which queue structure will be more cost effective? Aside from structure design, our results are of practical value for performance analysis and staffing deployment.

Funding: The work of P. Cao was supported in part by the National Natural Science Foundation of China [Grants 71771202 and 71520107002]. The work of S. He was supported in part by the Singapore Ministry of Education Academic Research Fund [Grant MOE2017-T2-1-012]. The work of J. Huang was supported in part by the Hong Kong Research Grants Council [Project 14502815].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2019.1976.

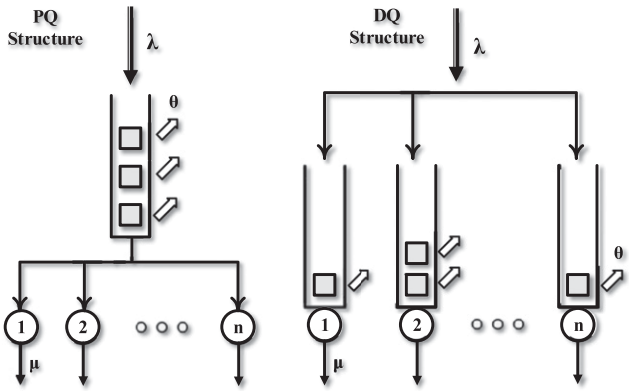
Keywords: dedicated queue • join-the-shortest-queue • power-of-*d* • join-idle-queues • customer abandonment • overloaded queue • many-server heavy-traffic limit

1. Introduction

In designing service systems, it is a common practice to organize customers with similar service requirements into a single queue served by a group of servers. This pooled queue (PQ) structure (see the left panel of Figure 1) is deemed highly efficient because each customer can be served by any available server, which will minimize the servers' idleness, thus reducing customers' waiting times (see, e.g., chap. 5.1 in Kleinrock (1976)). Such a structure is prevalent in call centers, emergency departments, outpatient clinics, etc.

In certain service systems, customers are assigned to specific servers upon arrival—that is, each server has her own queue for waiting customers. Such a dedicated queue (DQ) structure (see the right panel of Figure 1) is prevalent in supermarkets, immigration checkpoints, toll collection stations, etc. Clearly, the DQ structure is less efficient in terms of capacity utilization because servers could be idle while there are customers waiting in other servers' queues. However, pooling queues may not always improve a service system's performance, possible reasons for which include heterogeneous service requirements, customer reaction, and increased service times and costs (Rothkopf and Rech 1987). In particular, combining multiple streams of customers whose service time distributions are significantly different will prolong the mean customer waiting time (Smith and Whitt 1981, Mandelbaum and Reiman 1998, Whitt 1999). A service system's queue structure may also have a psychological influence on the servers' productivity. Recent empirical findings reveal that with their own queues, servers tend to work faster in certain service systems such as emergency departments and supermarkets (Song et al. 2015, Wang and Zhou 2018). It could thus be advantageous to adopt the DQ structure if the resulting improvement in service productivity outweighs the loss of capacity utilization.

Figure 1. Service Systems with a Pooled Queue (Left) and with Dedicated Queues (Right)



In this paper, we compare the two queue structures for designing large-scale service systems. Because customer abandonment is ubiquitous in service systems (e.g., corresponding to callers' hanging up in call centers or patients' leaving without being seen in emergency departments), we assume that each customer has a patience time to model this phenomenon. When the waiting time exceeds the patience time, a customer will leave the system without being served. In service operations, a *service level* is typically defined as the percentage of customers served within a given delay. Service contracts often specify performance targets for service levels, such as "at least 80% of calls should be answered within 20 seconds" (which is widely known as the "80/20 rule") and "at least 70% of patients should be seen by physicians within 30 minutes." By means of asymptotic analysis, we demonstrate that when the system is *moderately overloaded* (see condition (1)), it could achieve a *lower* probability of delay or a *higher* service level under the DQ structure by using the *join-the-shortest-queue* (JSQ) routing policy, whereas the induced loss of capacity utilization is negligible. Hence, to meet a certain service-level objective, a large-scale service system may need *fewer* servers under the DQ structure than under the PQ structure. In this sense, the DQ–JSQ design could be more efficient when staffing costs are expensive. Such a phenomenon is somewhat surprising at first glance, as it is a common belief that resource pooling would generally reduce staffing expenses. By converting the commonly used pooled structure into the dedicated structure, the resulting cost reduction could be even greater if the latter design may improve the servers' productivity. For example, the DQ structure may allow agents in a call center to access their own waiting customers' background information while serving other customers. As pointed out by Song et al. (2015), the more certain ownership of customers can help agents improve their service rates without compromising the quality of service.

To illustrate a queueing system's performance under the two queue structures, let us consider a Markovian model that has a Poisson arrival process and exponentially distributed service and patience times. The system has $n = 100$ servers, the customer arrival rate is $\lambda = 60$ per minute, the mean service time is $1/\mu = 2.0$ minutes, and the mean patience time is $1/\theta = 4.0$ minutes. The system is overloaded with traffic intensity $\rho = 1.2$. In Table 1, we report some performance statistics under the two structures obtained by simulation. Under the PQ structure, the probability of delay (denoted by P(De) in the table) is nearly 100%, and thus almost all customers have to wait before being served. The JSQ policy is used under the DQ structure. Because the system is moderately overloaded with $\rho = 1.2$, there are many queues having no waiting customers in the steady state. When a server who has no waiting customers completes a service, the next incoming customer will immediately enter service without needing to wait. Therefore, we may expect the probability of delay to be much smaller than 100% under the DQ–JSQ design: by simulation, about one-half of customers join idle servers and receive service immediately. Similarly, if we consider the percentage of customers whose waiting times are within 20 seconds (dubbed the "20-second service level"), we may expect a higher service level under the DQ–JSQ design as well: in Table 1, the service levels are 37.0% and 55.4% under the PQ and DQ structures, respectively. The simulation results show that the probabilities of customer abandonment (denoted by P(Ab) in the table) are nearly identical under the two queue structures. Hence, the servers' idleness induced by the DQ structure is negligible when the JSQ policy is used.

When queues are dedicated, the mean potential waiting time (PWT) of a delayed customer is about $1/\mu = 2.0$ minutes, greater than the mean PWT under the PQ structure (about 44 seconds). That is to say,

Table 1. Performance of a Markovian Queueing System Under the PQ and DQ Structures

Design	P(De) (%)	20-second service level (%)	P(Ab) (%)	Mean PWT (in seconds)
PQ	99.6	37.0	16.7	44
DQ–JSQ	52.5	55.4	17.2	61

Notes. The arrival rate is $\lambda = 60$ customers per minute, the mean service time is $1/\mu = 2.0$ minutes, and the mean patience time is $1/\theta = 4.0$ minutes. The JSQ routing policy is used under the DQ structure.

although about one-half of customers may join idle servers upon arrival under the JSQ policy, almost all other customers will join servers with exactly one customer, having to wait until their servers complete the current services. The benefit of no delay for some customers under the DQ–JSQ design is gained at the price of making others wait longer. Therefore, the DQ structure may not be appropriate for systems where the service discipline must be globally first-come, first-served (FCFS). Instead, the DQ structure could be more relevant to service systems where customers are relatively patient and queues are completely or partially invisible, such as service-oriented call centers and emergency departments (Mandelbaum et al. 2001, Armony et al. 2015). Nevertheless, the JSQ policy may help to address this "fairness" issue by placing each customer in one of the best queue positions upon arrival (see, e.g., Raz et al. (2005) for how the JSQ policy may improve service fairness under the DQ structure without customer abandonment).

Because exact analysis is generally difficult, we exploit asymptotic analysis by proving a fluid limit for the queue length process under the DQ–JSQ design. Compared with the fluid model for the PQ structure (see, e.g., Whitt (2004)), our fluid model has two distinct features: First, the fluid queue length process must be multidimensional to track the DQ–JSQ fluid model's state, whereas the PQ fluid model is one dimensional. Second, the DQ–JSQ fluid model can be used to characterize and approximate the distribution of customer waiting times, whereas the PQ fluid model can only be used to estimate the mean waiting time. In particular, the DQ–JSQ fluid model allows us to derive approximate formulas for service levels; by contrast, more refined diffusion models are required to estimate service levels under the PQ structure.

The influence of the queue structure on the system's performance could be distinct in different operational regimes. When a many-server system is underloaded (i.e., the traffic intensity is strictly less than 1), few customers will experience delay under the PQ structure. The DQ–JSQ system will be asymptotically equivalent to the PQ system as the number of servers increases. When a many-server system is critically loaded (i.e., the traffic intensity is about 1), we expect the mean PWT and the probability of abandonment to be close to 0 under both designs. Although the probability of delay is expected to be close to 0 in the DQ–JSQ system (see Eschenfeldt and Gamarnik (2018) for the analysis of DQ–JSQ systems without abandonment), it will be strictly between 0 and 1 in the PQ system (Halfin and Whitt 1981, Garnett et al. 2002). The asymptotic analysis in this paper reveals essential differences in performance under the two queue structures when the system is moderately overloaded (i.e., condition (1) holds). As a rigorous basis for the observations from Table 1, Theorem 5 in Section 5.2 points out that when the traffic intensity satisfies

$$1 < \rho < 1 + \frac{\theta}{\mu}, \tag{1}$$

where θ/μ is the ratio of the abandonment rate to the service rate, there will be a proportion of servers having no waiting customers under the DQ–JSQ design. Because such a server will be idle upon service completion, the probability of delay turns out to be strictly between 0 and 1 under the DQ–JSQ design, as opposed to being close to 1 under the PQ structure. That is to say, the DQ–JSQ design allows an overloaded system that satisfies condition (1) to achieve comparable delay performance to a critically loaded system under the PQ structure (see Remark 1 in Section 4.1 and Remark 3 in Section 5.1 for more discussion). If the traffic intensity satisfies $\rho \geq 1 + \theta/\mu$, the probability of delay will also approach 1 in the DQ–JSQ system, whereas some other performance measures still remain distinct under the two queue structures (see Section EC.5 of the e-companion for the performance comparison of the two queueing designs for all $\rho > 1$).

1.1. Our Contributions

First, this study provides theoretical support for the DQ–JSQ design in large-scale service systems. We develop a fluid model for many-server queueing systems with customer abandonment, and we prove a limit theorem for both process-level and steady-state convergence to justify the fluid model. Second, using the fluid model for the DQ–JSQ system, we obtain approximate formulas not only for mean performance measures but also for service levels. By comparing the system performance under the two queue structures, we provide new insights into queueing design: when there are many servers in the system, the loss of capacity utilization induced by the DQ structure will become negligible by employing the JSQ policy, which strives to balance workload across servers upon the arrival of each customer. Therefore, the DQ structure could be beneficial even when servers are identical and customers are homogeneous. Such results complement the predominant insights in the literature that the DQ structure could be advantageous when customers are heterogeneous (Smith and Whitt 1981), when it has a positive influence on the servers' productivity (Shunko et al. 2018), or when jockeying between queues is allowed (Rothkopf and Rech 1987). Third, we solve a staffing problem subject to a service-level constraint and obtain an asymptotically optimal solution using fluid analysis. Even with a lower staffing level, the optimal DQ–JSQ design may still be less sensitive than the optimal PQ design to forecasting

errors in the customer arrival rate. Although the PQ structure is prevalent in large-scale service systems, this study would suggest considering the DQ structure as well, especially when staffing costs are expensive and fairness may not be a serious concern. Aside from the above-mentioned contributions, we also demonstrate by numerical experiments that serving as low-overhead alternatives to the JSQ policy, the *power-of-d* policy and the *join-idle-queues* (JIQ) policy may achieve comparable performance under the DQ structure.

As queueing design is much more complex in practice, it is not our intention to assert that one queue structure would be superior to the other. Instead, we suggest system designers consider different queue structures so that their service systems may achieve performance objectives in a more efficient manner.

1.2. Organization of the Rest of the Paper

Section 2 is devoted to the review of related literature. We describe the Markovian queueing system under the DQ–JSQ design in Section 3. In Section 4, we introduce and analyze the fluid model, which is justified by a limit theorem. We compare the system's performance under the two queue structures in Section 5 and compare optimal staffing levels subject to a service-level constraint in Section 6. We evaluate the performance of the power-of-d policy and the JIQ policy in Section 7. The paper is concluded in Section 8. We leave the proofs of all theorems and propositions, along with additional numerical examples, in the e-companion.

2. Related Literature

We sketch related studies in the literature to position our work. Both the literature on the JSQ routing policy and the literature on asymptotic analysis of many-server queueing systems are well established. It is not our intention to be exhaustive.

2.1. Exact and Asymptotic Analysis of the JSQ Policy

The optimality of the JSQ policy was investigated by Winston (1977), Weber (1978), and Whitt (1986). It is well known that when the service time distribution has a nondecreasing hazard rate, the JSQ policy will minimize each customer's individual expected waiting time and the system's long-run average waiting time. The asymptotic analysis of this policy in the conventional heavy-traffic regime was studied by Foschini and Salz (1978), Reiman (1984), Zhang and Wang (1989), and Zhang et al. (1995). These papers demonstrate that by adopting the JSQ policy, a queueing system with multiple servers can achieve complete resource pooling in heavy traffic, thus becoming asymptotically equivalent to a single-server system having the same service capacity. A recent study by Eschenfeldt and Gamarnik (2018) is the most relevant to our work. They established a multidimensional diffusion limit for the Markovian queueing model without abandonment when the JSQ policy is used in the Halfin–Whitt regime. The probability of delay is proved to be close to 0 in this critically loaded regime. Their paper proposes a novel representation of the queue length process by counting servers according to their respective customer numbers. This approach enables tractable asymptotic analysis of the many-server system. Following their work, Braverman (2020) proved the steady-state convergence using Stein's method, Banerjee and Mukherjee (2019) established the steady-state tail asymptotics of the diffusion limit, and Mukherjee et al. (2016) analyzed a class of load balancing policies including the JSQ policy as a special case. Considering the Markovian queueing model in the nondegenerate slowdown (NDS) regime, Gupta and Walton (2019) obtained a diffusion limit for the customer-count process when the JSQ policy is used. By diffusion analysis, they demonstrated that in the NDS regime, the mean sojourn time in the DQ–JSQ system is at most 14% longer than that in the PQ system. They also proved that with much less communication overhead, the policy that prioritizes idle servers first and then servers with one customer is asymptotically equivalent to the JSQ policy.

To prove the limit theorem, we also adopt the queue length representation introduced by Eschenfeldt and Gamarnik (2018). Our work is distinguished from other recent studies in the following aspects: First, as an important phenomenon in service systems, customer abandonment is investigated in our queueing models, whereas this feature is not considered in the previous studies. Second, we focus on overloaded systems; they studied systems in the Halfin–Whitt regime or the NDS regime, both of which are critically loaded regimes. The presence of customer abandonment is essential to stabilizing an overloaded system. Third, from the methodological perspective, performance analysis in this paper relies on fluid approximations, whereas the aforementioned studies are mainly concerned with diffusion approximations.

2.2. Asymptotic Analysis of Many-Server PQ Systems with Abandonment

One may refer to Ward (2012) for a comprehensive survey on the asymptotic analysis of many-server PQ systems. From the long list of brilliant papers on this topic, we will name a few that are closely related to our work. By analyzing a diffusion limit for the $\mathrm{M/M}/n+\mathrm{M}$ system with many servers, Garnett et al. (2002) and Whitt (2004) extended the framework proposed by Halfin and Whitt (1981) to the Markovian model with customer abandonment. Whitt (2004) provided useful performance formulas for overloaded systems,

some of which are used in this paper for comparison. The asymptotic analysis of many-server systems with customer abandonment has been extended to general patience time distributions (Zeltyn and Mandelbaum 2005, Liu and Whitt 2014, Huang et al. 2017, Liu 2018) and to general service time distributions (Whitt 2006, Dai et al. 2010, Kang and Ramanan 2012, Aras et al. 2018, He 2020) under various assumptions and in different regimes.

2.3. Psychological and Behavioral Influence of Queue Structures

By revealing the psychological and behavioral influence of queue structures, recent empirical and methodological studies have added a new dimension to queueing design for service systems. Jouini et al. (2008) reported performance improvement in the call center of a French telecommunications company, after the queueing configuration had been converted from a single queue served by all agents into multiple queues served by independent agent clusters. The authors argued that the new configuration had motivated agents to improve their service rates and quality. They also proposed queueing models to justify their findings, assuming each agent cluster to have its own stream of customer arrivals. Unlike in our study, the JSQ routing policy is not considered in their paper. Song et al. (2015) examined a set of patient flow data from an emergency department. When patients were sent to physicians under the DQ structure, the mean length of stay was reduced by 17% and the mean waiting time was reduced by 9% compared with the statistics under the PQ structure. This study attributes the performance improvement to physicians' more definite ownership of patients under the DQ structure, which enables them to proactively retrieve patient information, manage assessment and treatment, and optimize the patient flow, thus accelerating their service rates. Wang and Zhou (2018) analyzed a set of transaction data from a supermarket, along with the associated queue information collected from video recordings. Under the DQ structure, cashiers in this supermarket were 10.7% faster than working under the PQ structure. Social loafing is believed to be the main cause of lower productivity when queues are pooled. Shunko et al. (2018) designed behavioral experiments to study the impact of the queue structure on service rates. In their experiments, pooling queues also slowed down the servers. To better understand the psychological influence of queue structures, Armony et al. (2018) proposed a game-theoretic model in a Markovian queueing system. In their setting, the DQ structure will reduce the expected work-in-process when the servers have discretion over their service rates and exhibit high degrees of workload aversion or low degrees of busyness aversion. In contrast to these studies, our paper demonstrates that the DQ structure may still help improve service levels in large-scale service systems even if the service rates remain unchanged.

The influence of queue structures on customers' behavior was investigated by Sunar et al. (2018). As in our paper, their study assumes identical servers and homogeneous customers. In their setting, customers must decide whether to join the system based on queue length information upon arrival. Because the JSQ policy also requires all queues to be observable, our paper conveys a similar message as theirs: the DQ structure may enable us to exploit the queue length of each server (as opposed to the total number of customers under the PQ structure) so that certain performance measures will be improved. (In Section 7, we will study the performance of routing policies that exploit partial queue length information under the DQ structure.) Although both papers are concerned with delay-sensitive customers, they remain distinct in the following aspects: First, we consider customer abandonment, whereas they considered balking, under the two queue structures. Second, our study focuses on how the DQ–JSQ design may render the loss of capacity utilization negligible and improve service levels; they focused on how customers' rational balking decisions may lead to a shorter mean sojourn time and a greater social welfare under the DQ structure.

3. The DQ–JSQ System

Consider a queueing system with n parallel servers. Each server has her own queue for waiting customers who are served on the FCFS basis. Customers arrive according to a Poisson process with rate λ, and service times are independent and exponentially distributed with mean $1/\mu$. The system's traffic intensity is $\rho := \lambda/(n\mu)$. Each customer has a random patience time. When the waiting time exceeds the patience time, a customer will abandon the system without being served. Patience times are independent and exponentially distributed with mean $1/\theta$. The sequences of interarrival, service, and patience times are mutually independent. We assume that customers follow the JSQ policy, joining a server that has the fewest customers upon arrival. Jockeying between queues is not permitted. Such a queueing system is referred to as the *DQ–JSQ system*.

We follow the representation proposed by Eschenfeldt and Gamarnik (2018) to describe the system's dynamics. Let $Q_i(t)$ be the number of servers who have at least i customers at time t, either waiting or being served. Then,

$$0 \le Q_{i+1}(t) \le Q_i(t) \le n \quad \text{for } i \in \mathbb{N}_0 \text{ and } t \ge 0, \qquad (2)$$

where $\mathbb{N}_0$ is the set of nonnegative integers. In particular, $Q_0(t) = n$ for $t \ge 0$, and $Q_1(t)$ is the number of

busy servers at time t. Let $\mathbf{Q}(t) := (Q_i(t) : i \in \mathbb{N})$, where $\mathbb{N}$ is the set of positive integers. We refer to $\mathbf{Q}$ as the *augmented queue length process*.

Let $A(t)$ be the number of customer arrivals by time t. We must track the numbers of arrivals at different queues to describe the dynamics of the augmented queue length process. To this end, we use $U_i(t)$ to denote the cumulative number of customers by time t whose servers had at least i customers right before their arrivals. If someone joins a server having i customers, all servers must have at least i customers by the JSQ policy. Therefore, $U_i(t)$ can be represented by

$$U_i(t) := \int_0^t \mathbb{1}_{\{Q_i(u-)=n\}}\,\mathrm{d}A(u) \quad \text{for } i \in \mathbb{N}_0. \tag{3}$$

In particular,

$$U_0(t) = A(t) \quad \text{for } t \geq 0. \tag{4}$$

By this definition, we have

$$\int_0^\infty \mathbb{1}_{\{Q_i(t-)<n\}}\,\mathrm{d}U_i(t) = 0 \quad \text{for } i \in \mathbb{N}_0. \tag{5}$$

Then by (2),

$$\Delta U_{i+1}(t) \leq \Delta U_i(t) \quad \text{for } i \in \mathbb{N}_0 \text{ and } t \geq 0, \tag{6}$$

where $\Delta U_i(t) := U_i(t+) - U_i(t-)$ is the increment at time t. We define the *augmented arrival process* $\mathbf{U}$ by $\mathbf{U}(t) := (U_i(t) : i \in \mathbb{N}_0)$.

Let $\{S_i, F_i : i \in \mathbb{N}\}$ be a set of independent Poisson processes with rate 1. Because $Q_i(u) - Q_{i+1}(u)$ is the number of servers having exactly i customers at time u, the cumulative number of service completions by time t from servers having i customers can be represented by

$$D_i(t) := S_i\left(\mu \int_0^t (Q_i(u) - Q_{i+1}(u))\,\mathrm{d}u\right) \quad \text{for } i \in \mathbb{N}. \tag{7}$$

Similarly, the cumulative number of abandoning customers by time t from servers having i customers can be represented by

$$G_i(t) := F_i\left(\theta(i-1) \int_0^t (Q_i(u) - Q_{i+1}(u))\,\mathrm{d}u\right) \quad \text{for } i \in \mathbb{N}. \tag{8}$$

The *augmented departure process* $\mathbf{D}$ is given by $\mathbf{D}(t) := (D_i(t) : i \in \mathbb{N})$, and the *augmented abandonment process* $\mathbf{G}$ is given by $\mathbf{G}(t) := (G_i(t) : i \in \mathbb{N})$.

By (3), $U_{i-1}(t) - U_i(t)$ is the cumulative number of customers who joined servers having exactly $i-1$ customers by time t. Under the JSQ policy, $Q_i(t)$ may increase only when a customer joins a server having $i-1$ customers. Hence,

$$Q_i(t) = Q_i(0) + U_{i-1}(t) - U_i(t) - D_i(t) - G_i(t) \quad \text{for } i \in \mathbb{N}. \tag{9}$$

Let $X(t)$ be the total number of customers in the system at time t, with $X(0) < \infty$ by convention. Because $Q_i(t) - Q_{i+1}(t)$ is the number of servers having i customers at time t, we may write $X(t) := \sum_{i=1}^\infty i(Q_i(t) - Q_{i+1}(t))$, which is equivalent to $X(t) := \sum_{i=1}^\infty Q_i(t)$. By (4) and (7)–(9),

$$X(t) = X(0) + A(t) - S\left(\mu \int_0^t Q_1(u)\,\mathrm{d}u\right) - F\left(\theta \int_0^t (X(u) - Q_1(u))\,\mathrm{d}u\right), \tag{10}$$

where S and F are two independent Poisson processes with rate 1.

4. Fluid Approximations

We investigate a fluid model for the DQ–JSQ system. This model is presented in Section 4.1, and the limiting state is studied in Section 4.2. We establish a limit theorem in Section 4.3 to justify the fluid model.

4.1. A Fluid Model for the DQ–JSQ System

Because the exact analysis of the DQ–JSQ system is difficult, we rely on an approximate model to investigate the system's performance. By the functional law of large numbers, the normalized augmented queue length process is expected be close to a deterministic process when n is large. We would thus seek a deterministic process $\bar{\mathbf{Q}}$ such that $\mathbf{Q}(t)/n \approx \bar{\mathbf{Q}}(t)$ for $t \geq 0$, where $\bar{\mathbf{Q}}(t) := (\bar{Q}_i(t) : i \in \mathbb{N})$. In other words, $\bar{Q}_i(t)$ is used to approximate the proportion of servers having at least i customers at time t. We refer to $\bar{\mathbf{Q}}$ as the *fluid queue length process*. To represent this process, we use a deterministic process $\bar{\mathbf{U}}$ to approximate the normalized augmented arrival process, with $\bar{\mathbf{U}}(t) := (\bar{U}_i(t) : i \in \mathbb{N}_0)$. We expect the pair $(\bar{\mathbf{Q}}, \bar{\mathbf{U}})$ to satisfy the following system of dynamical equations corresponding to (2)–(9) for the DQ–JSQ system:

$$\begin{cases} \bar{Q}_i(t) = \bar{Q}_i(0) + \bar{U}_{i-1}(t) - \bar{U}_i(t) - (\mu + \theta(i-1)) \\ \qquad\qquad \times \int_0^t (\bar{Q}_i(u) - \bar{Q}_{i+1}(u))\,\mathrm{d}u, & (11) \\ \bar{U}_0(t) = \rho\mu t, & (12) \\ \bar{U}_i(0) = 0, & (13) \\ \int_0^\infty \mathbb{1}_{\{\bar{Q}_i(t-)<1\}}\,\mathrm{d}\bar{U}_i(t) = 0, & (14) \\ 0 \leq \bar{Q}_{i+1}(t) \leq \bar{Q}_i(t) \leq 1, & (15) \\ 0 \leq \bar{U}_i'(t) \leq \bar{U}_{i-1}'(t), & (16) \end{cases}$$

for $i \in \mathbb{N}$, where $f'(t)$ denotes the derivative at $t \geq 0$ (if it exists) of a real-valued function f defined on $[0, \infty)$. The dynamical system given by (11)–(16) characterizes the *fluid model* for the DQ–JSQ system. The pair $(\bar{\mathbf{Q}}, \bar{\mathbf{U}})$ is said to be a *fluid solution* if it is a solution to (11)–(16) for $t \geq 0$ almost everywhere. (The existence and uniqueness of a fluid solution under a certain

condition is stated in Theorem 1.) Let $\bar{X}(t) := \sum_{i=1}^{\infty} \bar{Q}_i(t)$ be the total fluid content at time t. By (11) and (12),

$$\bar{X}(t) = \bar{X}(0) + \rho\mu t - \int_0^t \big(\mu\bar{Q}_1(u) + \theta\big(\bar{X}(u) - \bar{Q}_1(u)\big)\big)\,\mathrm{d}u. \tag{17}$$

To study the behavior of the fluid model, we write

$$q := \frac{\mu(\rho-1)^+}{\theta}, \quad \bar{q} := \lfloor q \rfloor + 1, \quad r := q - \lfloor q \rfloor, \tag{18}$$

where $\lfloor q \rfloor$ is the greatest integer that is less than or equal to q. Let us consider the DQ–JSQ system with n servers and traffic intensity $\rho > 1$. By conservation of flow, the total abandonment rate from the queueing system should satisfy

$$\theta \times \text{the mean number of waiting customers} \approx \lambda - n\mu = n(\rho-1)\mu,$$

which implies that q can be used to approximate the mean number of waiting customers in each queue. Because most queues differ by at most one customer under the JSQ policy, they have either $\lfloor q \rfloor$ or $\bar{q}$ waiting customers. By the fact that $q = r\bar{q} + (1-r)\lfloor q \rfloor$, the fraction of servers having $\bar{q}$ waiting customers is about r.

Remark 1. When condition (1) holds, we obtain $0 < q < 1$ by (18), in which case the fraction of servers having no waiting customers is about $1 - r > 0$. Upon service completion, such a server will be idle so that the next incoming customer will enter service without delay. Then, the probability of delay should be strictly less than 1 under this moderate overloading condition.

It is worth mentioning that when θ is close to 0, the traffic intensity should be close to 1 in order for condition (1) to hold. Because customers' patience times are relatively long in this case, there will be a proportion of servers having no waiting customers only if the traffic intensity is just "slightly" greater than 1 (i.e., the system is in a critically loaded regime rather than in an overloaded regime). We provide numerical examples when θ is small in Section EC.10 of the e-companion.

Let $\mathbb{S} := \{(x_i : i \in \mathbb{N}) : 0 \le x_i \le 1 \text{ and } x_{i+1} \le x_i \text{ for } i \in \mathbb{N}\}$ and $\mathbb{S}_N := \{(x_i : i \in \mathbb{N}) \in \mathbb{S} : x_i = 0 \text{ for } i \ge N\}$ for $N \in \mathbb{N}$. For technical convenience, we impose an initial condition on the fluid model:

$$\bar{\mathbf{Q}}(0) \in \mathbb{S}_N \quad \text{for some } N > \bar{q} + 1. \tag{19}$$

Because $\bar{Q}_i(0)$ approximates the proportion of servers having at least i customers at time 0, then roughly speaking, condition (19) corresponds to the assumption that the initial queue lengths of all servers in the DQ–JSQ system are bounded by N. With this condition, we can show that $\bar{\mathbf{Q}}(t) \in \mathbb{S}_N$ for all $t \ge 0$ (see Theorem 1), so that the fluid queue length process is essentially finite dimensional.

Theorem 1. *Assume that condition (19) holds. Then, there exists a unique solution $(\bar{\mathbf{Q}}, \bar{\mathbf{U}})$ to (11)–(16) for $t \ge 0$ almost everywhere. This solution has the following properties:* (i) *both $\bar{Q}_i$ and $\bar{U}_i$ are Lipschitz continuous for $i \in \mathbb{N}$, and* (ii) *$\bar{\mathbf{Q}}(t) \in \mathbb{S}_N$ for $t \ge 0$.*

Please refer to Section EC.1 for the proof of Theorem 1. If the initial condition of the fluid model is not far from the invariant state (see Theorem 2 in Section 4.2), the fluid solution may have a closed-form expression as specified in the following proposition, the proof of which is given in Section EC.8.

Proposition 1. *Assume that $\rho > 1$. If $\bar{Q}_i(0) = 1$ for $i \le \bar{q}$ and $\bar{Q}_i(0) = 0$ for $i \ge \bar{q} + 2$, the fluid solution $(\bar{\mathbf{Q}}, \bar{\mathbf{U}})$ is given by*

$$\bar{Q}_i(t) = \begin{cases} 1 & 1 \le i \le \bar{q}, \\ r + (\bar{Q}_i(0) - r)\mathrm{e}^{-\theta t} & i = \bar{q} + 1, \\ 0 & i \ge \bar{q} + 2, \end{cases}$$

$$\bar{U}_i(t) = \begin{cases} \rho\mu t & 0 \le i \le \bar{q} - 1, \\ (\mu + \theta\bar{q})rt + \theta^{-1}(\mu + \theta(\bar{q} - 1)) \\ \qquad \times (\bar{Q}_{\bar{q}+1}(0) - r)(1 - \mathrm{e}^{-\theta t}), & i = \bar{q}, \\ 0 & i \ge \bar{q} + 1. \end{cases}$$

4.2. The Limiting State

Let $\mathbf{q}^* := (q_i^* : i \in \mathbb{N})$ and $\mathbf{u}^*(t) = (u_i^*(t) : i \in \mathbb{N}_0)$ for $t \ge 0$, where

$$q_i^* = \begin{cases} 1 & 1 \le i \le \bar{q}, \\ r & i = \bar{q} + 1, \\ 0 & i \ge \bar{q} + 2, \end{cases} \quad \text{and} \quad u_i^*(t) = \begin{cases} \rho\mu t & 0 \le i \le \bar{q} - 1, \\ (\mu + \theta\bar{q})rt & i = \bar{q}, \\ 0 & i \ge \bar{q} + 1. \end{cases}$$

With $\bar{\mathbf{Q}}(0) = \mathbf{q}^*$ and $\rho > 1$, the fluid solution is given by $\bar{\mathbf{Q}}(t) = \mathbf{q}^*$ and $\bar{\mathbf{U}}(t) = \mathbf{u}^*(t)$ for $t \ge 0$. In other words, $\mathbf{q}^*$ is an *invariant state* of the fluid model. The next theorem states that as the unique invariant state, $\mathbf{q}^*$ is also the *limiting state* of the fluid model as t increases.

Theorem 2. *Assume that $\rho > 1$ and that condition (19) holds. Then, $\mathbf{q}^*$ is the unique invariant state in $\mathbb{S}_N$. Moreover, there exists some $c > 0$ such that*

$$|\bar{Q}_i(t) - q_i^*| \le c \cdot \mathrm{e}^{-\theta t} \quad \textit{for } i = 1, \ldots, N-1 \textit{ and } t \ge 0. \tag{20}$$

By this theorem, $\lim_{t\to\infty} \bar{\mathbf{Q}}(t) = \mathbf{q}^*$ in $\mathbb{S}_N$. Please refer to Section EC.2 for the proof.

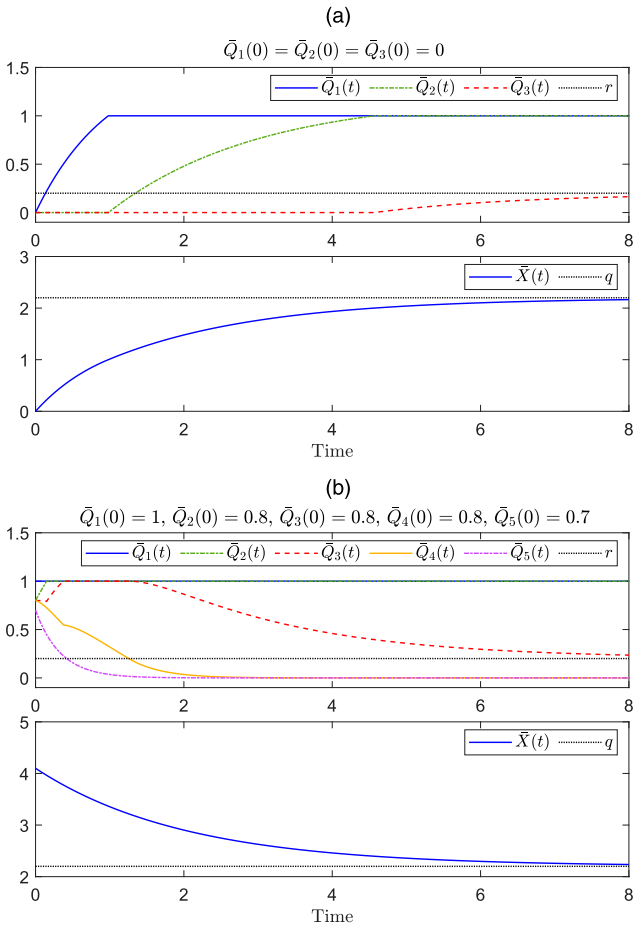
Let us illustrate the transient and limiting behavior of the fluid model by a numerical example.

Example 1. Take $\rho = 1.6$, $\mu = 1$, and $\theta = 0.5$. This set of parameters produces $q = 1.2$, $\bar{q} = 2$, and $r = 0.2$. We consider two different initial conditions: (i) $\bar{\mathbf{Q}}(0) = \mathbf{0}$ (i.e., the system is initially empty), and (ii) $\bar{\mathbf{Q}}(0) = (1, 0.8, 0.8, 0.8, 0.7, 0, 0, \ldots)$.

In case (i), we plot the fluid queue length process on the top panel of Figure 2. The system is initially empty. By (14), $\bar{U}_i'(0) = 0$ for $i \geq 1$, making $\bar{Q}_1$ increase while $\bar{Q}_2$ and $\bar{Q}_3$ remain at 0. At time $T_1 = 0.981$, $\bar{Q}_1$ reaches 1.0 and stays there ($q_1^* = 1.0$). Correspondingly, $\bar{U}_1'$ is positive. Because $\bar{Q}_2$ is strictly below 1.0, $\bar{U}_i'$ must be 0 for $i \geq 2$. As a result, $\bar{Q}_2$ begins to increase while $\bar{Q}_3$ remains at 0. At time $T_2 = 4.564$, $\bar{Q}_2$ reaches 1.0 and stays there ($q_2^* = 1.0$). After T_2, $\bar{Q}_3$ begins to increase and will eventually converge to its invariant value $q_3^* = r = 0.2$. Moreover, $\bar{Q}_i(t) = 0$ for all $t \geq 0$ and $i \geq 4$.

In case (ii), the fluid queue length process exhibits more complex dynamics. For instance, the path of $\bar{Q}_3$ is not monotone, first increasing and then decreasing. Nevertheless, the fluid queue length process will converge to the same invariant state regardless of the initial condition. Because $\bar{Q}_1(t) = 1.0$ for all $t \geq 0$, the total fluid content $\bar{X}$ is an exponential function as the solution to (17). This fact is consistent with the bottom plot of Figure 2.

Figure 2. (Color online) Two Sample Paths of the Fluid Queue Length Process



4.3. A Limit Theorem

To justify the fluid model by asymptotic analysis, we consider a sequence of DQ–JSQ systems indexed by the number of servers n. In each system, the initial augmented queue length, the arrival process, the sequence of service times, and the sequence of patience times are all mutually independent. These systems have the same service time distribution with mean $1/\mu$ and the same patience time distribution with mean $1/\theta$. Let λ^n be the arrival rate of the nth system. To establish the asymptotic framework, we assume that these arrival rates satisfy

$$\lim_{n\to\infty} \frac{\lambda^n}{n\mu} = \rho > 1. \tag{21}$$

By convention, we add a superscript n to some processes introduced in Section 3 to denote the corresponding processes in the nth DQ–JSQ system, such as $\mathbf{Q}^n$, $\mathbf{U}^n$, and X^n. Their fluid-scaled versions are given by

$$\bar{\mathbf{Q}}^n(t) := \frac{1}{n}\mathbf{Q}^n(t), \quad \bar{\mathbf{U}}^n(t) := \frac{1}{n}\mathbf{U}^n(t), \quad \bar{X}^n(t) := \frac{1}{n}X^n(t).$$

Both $\bar{\mathbf{Q}}^n(t)$ and $\bar{\mathbf{U}}^n(t)$ are random vectors taking values in $\mathbb{R}^\infty$. To study related continuity and convergence results, we define a metric d on $\mathbb{R}^\infty$ by

$$d(\mathbf{y}, \mathbf{z}) := \sum_{i=1}^{\infty} 2^{-i}(|y_i - z_i| \wedge 1),$$

where $\mathbf{y} := (y_i : i \in \mathbb{N}) \in \mathbb{R}^\infty$ and $\mathbf{z} := (z_i : i \in \mathbb{N}) \in \mathbb{R}^\infty$. Let $\mathbb{D}$ and $\mathbb{D}^\infty$ be the respective spaces of real-valued and $\mathbb{R}^\infty$-valued functions defined on $[0, \infty)$ that are right continuous on $[0, \infty)$ and have left limits on $(0, \infty)$. The space $\mathbb{D}$ is endowed with the Skorohod metric σ (see, e.g., section 12 in Billingsley (1999)), and the space $\mathbb{D}^\infty$ is endowed with a metric σ_∞ defined by

$$\sigma_\infty(\mathbf{Y}, \mathbf{Z}) := \sum_{i=1}^{\infty} 2^{-i}(\sigma(Y_i, Z_i) \wedge 1),$$

where $\mathbf{Y} := (Y_i : i \in \mathbb{N}) \in \mathbb{D}^\infty$ and $\mathbf{Z} := (Z_i : i \in \mathbb{N}) \in \mathbb{D}^\infty$.

A *functional weak law of large numbers* for the sequence of DQ–JSQ systems is presented in the next theorem, where we also establish the convergence of the normalized steady-state augmented queue lengths to the invariant state of the fluid model. Please refer to Section EC.3 for the proof.

Theorem 3.

(i) *Assume that* $\bar{\mathbf{Q}}^n(0) \Rightarrow \bar{\mathbf{Q}}(0)$ *in* $\mathbb{R}^\infty$ *as* $n \to \infty$ *for some* $\bar{\mathbf{Q}}(0)$ *that satisfies condition* (19). *If condition* (21)

holds, then $(\bar{\mathbf{Q}}^n, \bar{\mathbf{U}}^n) \Rightarrow (\bar{\mathbf{Q}}, \bar{\mathbf{U}})$ *in* $\mathbb{D}^\infty$ *as* $n \to \infty$, *where* $(\bar{\mathbf{Q}}, \bar{\mathbf{U}})$ *is the fluid solution* (*i.e., the unique solution to* (11)–(16) *for* $t \geq 0$ *almost everywhere*).

(ii) *Assume that condition* (21) *holds. For each* $n \in \mathbb{N}$, *there exists an* $\mathbb{R}^\infty$*-valued random vector* $\bar{\mathbf{Q}}^n(\infty)$ *such that* $\bar{\mathbf{Q}}^n(t) \Rightarrow \bar{\mathbf{Q}}^n(\infty)$ *in* $\mathbb{R}^\infty$ *as* $t \to \infty$. *Moreover,* $\bar{\mathbf{Q}}^n(\infty) \Rightarrow \mathbf{q}^*$ *in* $\mathbb{R}^\infty$ *as* $n \to \infty$.

Let $X^n(\infty)$ be the steady-state number of customers in the system. Then, the fluid-scaled version satisfies $\bar{X}^n(\infty) = \sum_{i=1}^\infty \bar{Q}_i^n(\infty)$. Part (ii) of Theorem 3 implies that $\bar{X}^n(\infty) \Rightarrow q + 1$ as $n \to \infty$. The number of customers waiting in the queues should thus be about nq in the steady state.

It is worth mentioning that one may also establish a diffusion limit for the augmented queue length process, which may serve as a refined approximate model. We will briefly discuss that as a future research topic in Section 8, where a numerical example is given to illustrate the distribution of augmented queue lengths in the DQ–JSQ system.

Remark 2. Because a server could be idle while there are customers waiting in other servers' queues, the DQ structure is less efficient in terms of capacity utilization than the PQ structure. However, by part (ii) of Theorem 3, $\bar{Q}_1^n(\infty) \Rightarrow q_1^* = 1$ as $n \to \infty$—that is, when the number of servers is large, almost all of them should be busy in the steady state. Therefore, the system's workload will be well balanced across servers by the JSQ policy, and the loss of capacity utilization induced by the DQ structure will be *negligible* when n is large (also see Eschenfeldt and Gamarnik (2018), Banerjee and Mukherjee (2019), Gupta and Walton (2019), and Braverman (2020) for DQ–JSQ systems without customer abandonment).

5. Performance Analysis of the DQ–JSQ and PQ Designs

We compare the system's performance under the two queue structures in this section. In service systems such as call centers, quality of service is usually measured in terms of service level (i.e., the percentage of customers served within a given delay target), the percentage of abandoning customers, mean waiting time, etc. (see, e.g., chap. 2 in Koole (2013) for a detailed discussion). Among these measures, service level is not only the most common performance indicator specified in service contracts for call centers (Milner and Olsen 2008, Baron and Milner 2009), but also a crucial metric for quality of care in hospital inpatient wards (Shi et al. 2016) and emergency departments (Ding et al. 2019, He et al. 2019). As a sign of customer dissatisfaction, abandonment should be maintained at a low level. The probability of customer abandonment is thus of great concern to service system managers. Whether a customer will abandon the system depends on both the patience time and the PWT. Statistics of waiting times are also taken into account in our analysis. Aside from the mean waiting time, we consider the variance of waiting times as well, because it is often used to measure "unfairness" in queueing design (Kingman 1962, Avi-Itzhak and Levy 2004). By the comparison of these performance measures, we would help service system designers better understand the two queue structures so that they may choose a more efficient design according to their needs.

We establish a limit for the steady-state PWTs under the DQ–JSQ design in Section 5.1 and compare the aforementioned performance measures under the two designs in Section 5.2.

5.1. Potential Waiting Time in the Steady State

Let W^n be the steady-state PWT in the nth DQ–JSQ system (i.e., an arbitrary customer's waiting time for service in the steady state if his patience time were infinite). The PWT quantifies the effort that a customer has to expend to receive service. The steady-state *actual waiting time* (AWT) is given by $V^n := W^n \wedge R$, where R stands for a generic patience time independent of W^n.

To characterize W^n, let us consider a pure death process on the state space $\mathbb{N}_0$, where the death rate of state i is $\mu + \theta(i-1)$ for $i \in \mathbb{N}$. Let $\{T_a(i) : i \in \mathbb{N}_0\}$ be a sequence of independent random variables, where $T_a(0) = 0$ and $T_a(i)$ is the time to absorption of the pure death process starting from state $i \in \mathbb{N}$. Clearly, $T_a(i)$ can be written as the sum of i independent exponential random variables—that is, $T_a(i) := \sum_{k=0}^{i-1} \xi_{i,k}$, where $\xi_{i,k}$ is exponentially distributed with mean $1/(\mu + k\theta)$. The Laplace transform of $T_a(i)$ is given by

$$\mathbb{E}[\mathrm{e}^{-sT_a(i)}] = \prod_{k=0}^{i-1} \frac{\mu + k\theta}{s + \mu + k\theta} \quad \text{for } i \in \mathbb{N}, \tag{22}$$

and the complementary cumulative distribution function of $T_a(i)$ is given by

$$\mathbb{P}(T_a(i) > x) = \sum_{j=1}^{i} \mathrm{e}^{-(\mu+(j-1)\theta)x} \prod_{k=1,k\neq j}^{i} \frac{\mu + (k-1)\theta}{(k-j)\theta} \quad \text{for } x \geq 0 \tag{23}$$

(see, e.g., Amari and Misra 1997). If a customer joins a server having i customers (either waiting or being served), the PWT will have the same distribution as $T_a(i)$. By part (ii) of Theorem 3, we obtain the following limit of the steady-state PWTs.

Theorem 4. *Let* χ *be a Bernoulli random variable with* $p := \mathbb{P}(\chi = 1) = r(\mu + \theta\bar{q})/(\rho\mu)$. *Assume that* χ *is independent of* $T_a(\bar{q})$ *and* $T_a(\lfloor q \rfloor)$ *and that condition* (21) *holds. Then,* $W^n \Rightarrow W$ *as* $n \to \infty$, *where* $W := \chi \cdot T_a(\bar{q}) + (1-\chi) \cdot T_a(\lfloor q \rfloor)$.

Please refer to Section EC.4 for the proof of Theorem 4. By this theorem, we may use a mixture of $T_a(\bar{q})$ and $T_a(\lfloor q \rfloor)$ to approximate the PWT. In the steady state, a customer will join a server having $\bar{q}$ customers (including one being served and $\lfloor q \rfloor$ waiting) with a probability of about p or join a server having $\lfloor q \rfloor$ customers with a probability of about $1-p$. We may thus estimate an arbitrary service level (i.e., the probability of W^n not exceeding a given threshold $x \geq 0$) by

$$\mathbb{P}(W^n \leq x) \approx p \cdot \mathbb{P}\big(T_a(\bar{q}) \leq x\big) + (1-p) \cdot \mathbb{P}\big(T_a(\lfloor q \rfloor) \leq x\big).$$

The approximate fraction p is interpreted as follows: as we discussed in Section 4.1, with $\rho > 1$, almost all servers in the DQ–JSQ system have either $\bar{q}+1$ or $\bar{q}$ customers (including one customer being served), with the fraction of servers that have $\bar{q}+1$ customers being about r. By conservation of flow, the arrival rate to servers having $\bar{q}$ customers should be equal to the sum of the service completion rate from servers having $\bar{q}+1$ customers and the abandonment rate from their queues—that is, $\lambda^n p \approx nr(\mu + \theta\bar{q})$—by which, together with (21), we obtain $p = r(\mu+\theta\bar{q})/(\rho\mu)$.

In the preceding discussion, we estimate the fraction of servers with a specific number of customers using fluid approximations; then by analyzing individual queues, we obtain an approximate distribution for the PWT. By combining a deterministic fluid analysis with a stochastic queueing analysis, we obtain a distributional performance measure for the DQ–JSQ system. By contrast, using fluid approximations under the PQ structure, we can only obtain mean performance measures such as the mean PWT and the mean queue length (Whitt 2006, Liu and Whitt 2012). We have to rely on diffusion approximations to estimate a distributional performance measure under the PQ structure (Garnett et al. 2002, Whitt 2004).

Remark 3. Consider a DQ–JSQ system that satisfies the moderate overloading condition (1). It follows from (18) that $0 < q < 1$, $\lfloor q \rfloor = 0$, $\bar{q} = 1$, and $r = q$. By Theorem 4 and the fact that $T_a(0) = 0$, the fraction of customers getting into service upon arrival is about $1-p$. The probability of delay is thus about $p = r(\mu+\theta)/(\rho\mu) = 1-(1-r)/\rho$, strictly less than 1, as we discussed in Section 4.1. To achieve such performance under the PQ structure, the queueing system must be critically loaded ($\rho \approx 1$), requiring more servers than under the DQ structure (Halfin and Whitt 1981, Garnett et al. 2002).

5.2. Performance Comparison of Moderately Overloaded Systems

We focus on the moderately overloaded regime and compare the sequence of DQ–JSQ systems with a sequence of PQ systems, the nth of which is an $M/M/n+M$ system having the same arrival rate, mean service time, and mean patience time as the nth DQ–JSQ system. To differentiate corresponding performance measures, we add subscripts D and P to respective quantities for the DQ–JSQ and PQ systems. The next theorem provides performance formulas in the moderately overloaded regime under the two queue structures. Some formulas for the PQ system are taken from Whitt (2004).

Theorem 5. *Assume that condition (21) holds with $1 < \rho < 1 + \theta/\mu$. Then, the performance of the nth DQ–JSQ system and that of the $M/M/n+M$ system have the following asymptotic relationships:*

(i) *For the mean fluid-scaled numbers of customers in the system,*

$$\lim_{n\to\infty} \mathbb{E}\big[\bar{X}^n_{\mathrm{D}}(\infty)\big] = q+1 = \lim_{n\to\infty} \mathbb{E}\big[\bar{X}^n_{\mathrm{P}}(\infty)\big],$$

where $0 < q < 1$.

(ii) *For the probabilities of customer abandonment,*

$$\lim_{n\to\infty} P^n_{\mathrm{D}}(\mathrm{Ab}) = \frac{\rho-1}{\rho} = \lim_{n\to\infty} P^n_{\mathrm{P}}(\mathrm{Ab}).$$

(iii) *For the mean AWTs,*

$$\lim_{n\to\infty} \mathbb{E}\big[V^n_{\mathrm{D}}\big] = \frac{q}{\rho\mu} = \lim_{n\to\infty} \mathbb{E}\big[V^n_{\mathrm{P}}\big].$$

(iv) *For the probabilities of delay,*

$$\lim_{n\to\infty} P^n_{\mathrm{D}}(\mathrm{De}) = \frac{r(\mu+\theta)}{\rho\mu} < 1 = \lim_{n\to\infty} P^n_{\mathrm{P}}(\mathrm{De}).$$

(v) *For the mean PWTs of delayed customers,*

$$\lim_{n\to\infty} \mathbb{E}\big[W^n_{\mathrm{D}} | W^n_{\mathrm{D}} > 0\big] = \frac{1}{\mu} > w = \lim_{n\to\infty} \mathbb{E}\big[W^n_{\mathrm{P}} | W^n_{\mathrm{P}} > 0\big],$$

where $w := \ln(\rho)/\theta$.

(vi) *For the mean PWTs,*

$$\lim_{n\to\infty} \mathbb{E}\big[W^n_{\mathrm{D}}\big] = \frac{r(\mu+\theta)}{\rho\mu^2} > w = \lim_{n\to\infty} \mathbb{E}\big[W^n_{\mathrm{P}}\big].$$

(vii) *For the mean AWTs of served customers,*

$$\lim_{n\to\infty} \mathbb{E}\big[V^n_{\mathrm{D}} | W^n_{\mathrm{D}} \leq R\big] = \frac{r}{\mu+\theta} < w = \lim_{n\to\infty} \mathbb{E}\big[V^n_{\mathrm{P}} | W^n_{\mathrm{P}} \leq R\big].$$

(viii) *For the mean AWTs of abandoning customers,*

$$\lim_{n\to\infty} \mathbb{E}\big[V^n_{\mathrm{D}} | W^n_{\mathrm{D}} > R\big] = \frac{1}{\mu+\theta} > -\frac{\mu w}{\theta q} + \frac{1}{\theta} = \lim_{n\to\infty} \mathbb{E}\big[V^n_{\mathrm{P}} | W^n_{\mathrm{P}} > R\big].$$

(ix) *For the variances of PWTs,*

$$\lim_{n\to\infty} \mathrm{Var}(W_{\mathrm{D}}^n) = \frac{r(\mu+\theta)}{\rho\mu^3}\left(2 - \frac{r(\mu+\theta)}{\rho\mu}\right) > 0 = \lim_{n\to\infty} \mathrm{Var}(W_{\mathrm{P}}^n).$$

The proof of Theorem 5 is given in Section EC.5, where we also provide performance comparison for all $\rho > 1$. Before we discuss the comparison results, let us examine the approximate formulas by a numerical example.

Example 2. We consider a queueing system with mean service time $1/\mu = 1.0$, mean patience time $1/\theta = 2.0$, and traffic intensity $\rho = 1.2$. The number of servers is taken to be $n = 400$, 100, or 25. We summarize both simulation results (with 95% confidence intervals) and fluid approximations under the DQ–JSQ design in Table 2. As implied by Theorem 5, fluid approximations turn out to be more and more accurate as n increases. We also provide exact performance measures for the PQ design, by which the comparison results established in Theorem 5 are confirmed.

By adopting the JSQ policy, the loss of capacity utilization induced by the DQ structure will vanish as n increases. As implied by parts (i)–(iii) of Theorem 5, the fluid-scaled number of customers, the probability of customer abandonment, and the mean AWT will be approximately equal under the two designs.

As a special service level, the probability of delay is a widely used metric for quality of service. Part (iv) of Theorem 5 states that the limit of the probability of delay is strictly less than 1 under the DQ–JSQ design, as opposed to being equal to 1 under the PQ design. Although almost all servers of the DQ–JSQ system are busy in the steady state, one or several idle servers may frequently appear for short periods, allowing a proportion of customers to enter service without delay. By contrast, idle servers barely appear in the PQ system, and as a result, the probability of delay must be close to 1 under the PQ structure (Garnett et al. 2002, Whitt 2004).

Part (v) of Theorem 5 reveals the price for the smaller probability of delay in the DQ–JSQ system: because almost all delayed customers join servers having one customer (who is being served), the mean PWT of them are close to $1/\mu$, greater than that in the PQ system. The benefit of no delay for a proportion of customers in the DQ–JSQ system is gained at the expense of making others wait longer.

We may obtain the approximate mean PWTs in part (vi) of Theorem 5 using parts (iv) and (v). When n is large, the mean PWT of all customers is also longer in the DQ–JSQ system. With a proportion of customers entering service without delay, the mean AWT of served customers is shorter in the DQ–JSQ system, as stated in part (vii) of Theorem 5. Because the mean AWTs are approximately equal under the two designs, the mean AWT of abandoning customers must be longer in the DQ–JSQ system, as stated in part (viii) of Theorem 5.

The variance of PWTs in part (ix) of Theorem 5 is a measure of unfairness in queueing design. In the PQ system, the steady-state PWT converges in distribution to the constant w as n increases (Whitt 2004). Accordingly, the variance of PWTs converges to 0. Because every customer needs to wait almost the same amount of time for service, the PQ design is able to achieve "absolute" fairness when the system has many servers. By contrast, the DQ structure is intrinsically unfair: whereas some customers may enter service without delay, others may have to wait until their servers complete the current services. The JSQ policy could mitigate the unfairness induced by the

Table 2. Performance Comparison Between the DQ–JSQ and PQ Designs

	$n = 400$			$n = 100$			$n = 25$		
	DQ–JSQ		PQ	DQ–JSQ		PQ	DQ–JSQ		PQ
Parameter	Sim.	App.	Exact	Sim.	App.	Exact	Sim.	App.	Exact
$P(\mathrm{De})$	0.508 ± 0.003	0.500	1.000	0.525 ± 0.003	0.500	0.997	0.577 ± 0.003	0.500	0.928
$P(\mathrm{Ab})$	0.168 ± 0.002	0.167	0.167	0.174 ± 0.002	0.167	0.167	0.195 ± 0.002	0.167	0.174
$\mathbb{E}[X(\infty)]$	561.2 ± 0.4	560.0	560.0	141.0 ± 0.2	140.0	140.0	35.84 ± 0.06	35.00	35.23
$\mathbb{E}[W]$	0.506 ± 0.004	0.500	0.366	0.520 ± 0.005	0.500	0.370	0.587 ± 0.004	0.500	0.400
$\mathbb{E}[W\|W > 0]$	0.996 ± 0.007	1.000	0.354	0.995 ± 0.007	1.000	0.371	1.011 ± 0.005	1.000	0.431
$\mathbb{E}[V]$	0.338 ± 0.003	0.333	0.333	0.350 ± 0.003	0.333	0.334	0.390 ± 0.003	0.333	0.349
$\mathbb{E}[V\|W < R]$	0.272 ± 0.003	0.267	0.363	0.284 ± 0.003	0.267	0.360	0.324 ± 0.003	0.267	0.366
$\mathbb{E}[V\|W > R]$	0.663 ± 0.008	0.667	0.183	0.666 ± 0.008	0.667	0.202	0.672 ± 0.006	0.667	0.265
$\mathrm{Var}(W)$	0.748 ± 0.014	0.750	0.005	0.762 ± 0.015	0.750	0.020	0.845 ± 0.013	0.750	0.070

Notes. The Markovian queueing system has $1/\mu = 1.0$; $1/\theta = 2.0$; $\rho = 1.2$; and $n = 400, 100,$ *and* 25 under the two queue structures. Both simulation results (with 95% confidence intervals) and fluid approximations are provided for the DQ–JSQ design; exact results are provided for the PQ design.

DQ structure by placing each customer in one of the best queue positions upon arrival. However, it cannot achieve the fairness provided by the PQ design: by Theorem 4 and part (ix) of Theorem 5, the steady-state PWT in the DQ–JSQ system converges in distribution to a random variable with a positive variance, as opposed to a constant as in the PQ system.

Remark 4. By part (ix) of Theorem 5, we may write the limiting variance of PWTs in the DQ–JSQ system as

$$\lim_{n\to\infty} \mathrm{Var}\big(W_{\mathrm{D}}^n\big) = \frac{p(2-p)}{\mu^2},$$

where $p = r(\mu+\theta)/(\rho\mu)$. This limiting variance is increasing in p for $0 < p < 1$. By (18), we further have $p = (1-1/\rho)(1+\mu/\theta)$, which is increasing in ρ—as a result, so is the limiting variance of PWTs. By part (iv) of Theorem 5, p is identical to the limiting probability of delay. When the traffic intensity goes from 1 to $1+\theta/\mu$, the probability of delay will increase from 0 to 1, and correspondingly, the limiting variance of PWTs will increase from 0 to $1/\mu^2$. In addition, p is decreasing in θ and so is the limiting variance of PWTs. In other words, if the moderate overloading condition holds with the traffic intensity being fixed, the variance of PWTs will be reduced when customers are less patient.

We take the variance of PWTs as the measure of unfairness mainly because of its simplicity. In the literature, there are other measures of unfairness that may be relevant (see Avi-Itzhak et al. (2008) for a survey). For example, using a measure introduced in their earlier paper (Raz et al. 2004), Raz et al. (2005) demonstrated that either combining dedicated queues or adopting the JSQ policy may improve service fairness in queueing systems without customer abandonment. Although such measures may capture more subtle aspects in queueing design, it could be difficult to analyze them for the DQ–JSQ system in the many-server setting. As suggested by Milner and Olsen (2008), including terms about the second moment (or the variance) of waiting times in service contracts may mitigate the fairness issue. We would thus employ this simple statistic as the measure of unfairness for analytical purposes.

Example 3. To better illustrate delay performance under the two designs, we simulate the system in Example 2 with $n = 100$ for 4.0×10^5 units of time, taking the first 1.0×10^5 units of time as the burn-in period before the steady state. The simulation results of the probability density functions of PWTs, AWTs, and AWTs of abandoning customers are plotted in Figure 3. Under the DQ–JSQ design, a customer's PWT could be either 0 or another customer's remaining service time. The probability density function of that appears to be an exponential function with a point mass at 0 (which is truncated in the figure). By contrast, the steady-state PWT exhibits relatively low variability under the PQ design, concentrating on a much narrower range. The distribution appears to be Gaussian (see formula (3.16) in Whitt 2004). Because PWTs have a smaller variance in the PQ system, AWTs and AWTs of abandoning customers also have smaller variances there. Moreover, almost all abandoning customers in the PQ system have relatively short patience times. By contrast, customers with much longer patience times may still abandon the DQ–JSQ system, because the remaining service times of their servers could be long. This phenomenon is evident from the bottom panel of Figure 3.

6. Staffing Under the DQ–JSQ and PQ Designs

In this section, we compare the two queue structures from a system designer's perspective, considering a staffing problem for meeting a service-level objective. More specifically, we solve the optimal staffing problem in Section 6.1 and compare the two designs for different threshold probabilities in Section 6.2.

6.1. Optimal Staffing Subject to a Service-Level Constraint

We would like to find the minimum number of servers for a Markovian queueing system so that the probability of the steady-state PWT exceeding a given delay target is below a specified threshold—that is, we would solve the following problem:

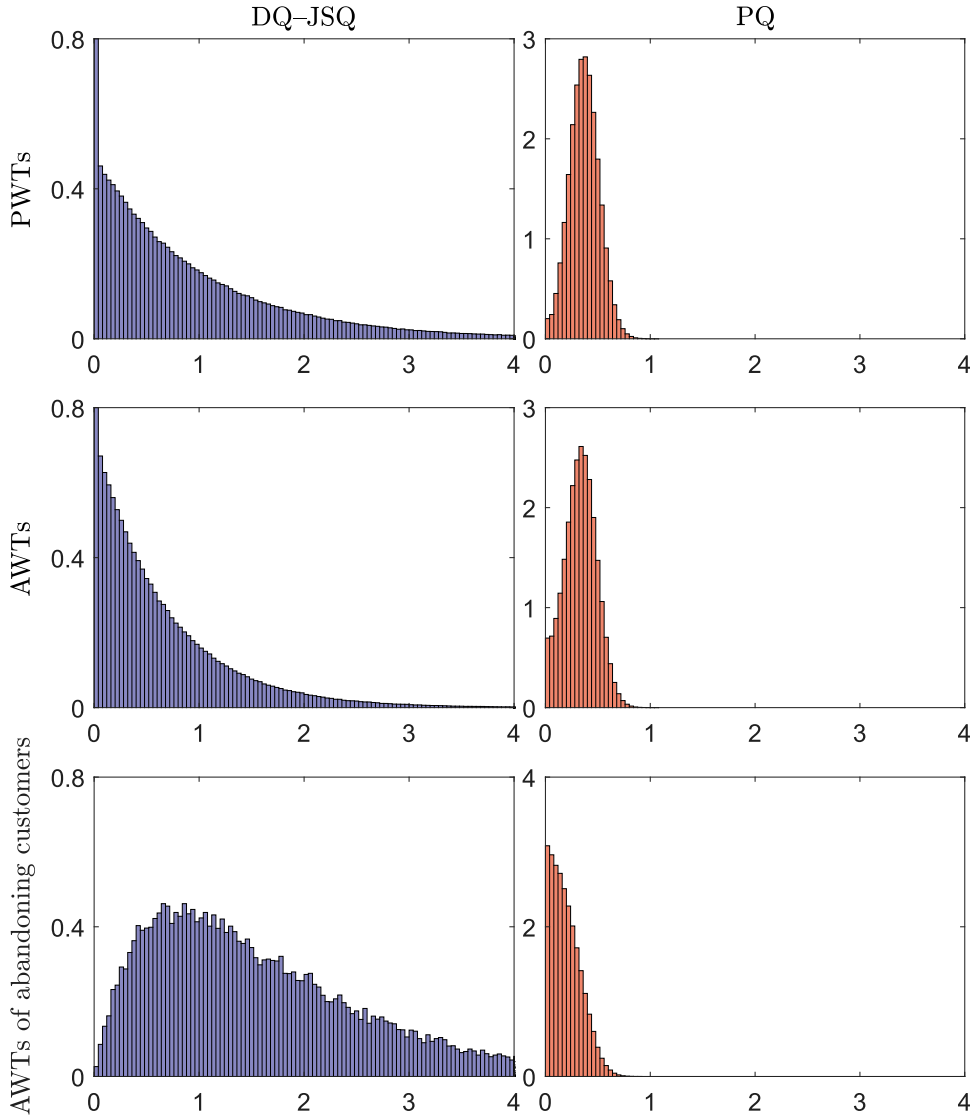
$$\hat{n}(\lambda,T,\alpha) := \min\big\{n\in\mathbb{N} : \mathbb{P}\big(W_\lambda^n > T\big) \le \alpha\big\} \quad \text{for } \lambda>0,\ T\ge 0,\ \text{and } 0<\alpha<1, \tag{24}$$

where W_λ^n is the steady-state PWT when the arrival rate is λ and the system has n servers. In particular, $\hat{n}(\lambda,0,\alpha)$ is the minimum staffing level that keeps the probability of delay not exceeding α. As in the previous section, we add subscripts D and P to staffing levels under the DQ–JSQ and PQ designs, respectively. Because it is difficult to obtain the closed-form solution to (24), we rely on asymptotic analysis to obtain an approximate solution. The optimal staffing level under the DQ–JSQ design is characterized in the next theorem, the proof of which is given in Section EC.6.

Theorem 6. *The optimal staffing level under the DQ–JSQ design follows*

$$\hat{n}_{\mathrm{D}}(\lambda,T,\alpha) = \frac{\mu+\theta(\kappa(T,\alpha)+r_0(T,\alpha))}{\big(\mu+\theta\kappa(T,\alpha)\big)\big(\mu+\theta(\kappa(T,\alpha)+1)\big)}\cdot\lambda + o(\lambda) \quad \textit{for } \lambda>0,\ T\ge 0,\ \textit{and } 0<\alpha<1, \tag{25}$$

Figure 3. (Color online) Estimates of the Probability Density Functions of PWTs, AWTs, and Abandoning Customers' AWTs



where $\kappa(T,\alpha) := \max\{i \in \mathbb{N}_0 : \mathbb{P}(T_a(i) > T) \le \alpha\}$ *and*

$$r_0(T,\alpha) := \frac{\mathbb{P}(T_a(\kappa(T,\alpha)+1) > T) - \alpha}{\mathbb{P}(T_a(\kappa(T,\alpha)+1) > T) - \mathbb{P}(T_a(\kappa(T,\alpha)) > T)}.$$

In particular,

$$\hat{n}_{\mathrm{D}}(\lambda,0,\alpha) = \left(1 - \alpha \cdot \frac{\theta}{\mu+\theta}\right) \cdot \frac{\lambda}{\mu} + o(\lambda). \tag{26}$$

For comparison, the optimal staffing level under the PQ design follows

$$\hat{n}_{\mathrm{P}}(\lambda,T,\alpha) = \begin{cases} \mathrm{e}^{-\theta T} \cdot \frac{\lambda}{\mu} + \sqrt{\theta \mathrm{e}^{-\theta T}} \cdot \bar{\Phi}^{-1}(\alpha) \\ \qquad \cdot \sqrt{\frac{\lambda}{\mu}} + o\left(\sqrt{\lambda}\right), & T > 0, \\ \frac{\lambda}{\mu} + \beta \cdot \sqrt{\frac{\lambda}{\mu}} + o\left(\sqrt{\lambda}\right), & T = 0, \end{cases} \tag{27}$$

where the expression for $T > 0$ is obtained by theorem 4.4 in Mandelbaum and Zeltyn (2009) and that for $T = 0$ is obtained by theorem 4 in Garnett et al. (2002). In this formula, $\bar{\Phi}^{-1}$ is the inverse survival function of the standard normal distribution, and β is the unique solution to

$$\alpha = \frac{\sqrt{\theta^{-1}\mu}h(-\beta)}{\sqrt{\theta^{-1}\mu}h(-\beta) + h\left(\beta\sqrt{\theta^{-1}\mu}\right)},$$

where h is the hazard rate function of the standard normal distribution.

To illustrate the advantages of the DQ–JSQ design in staffing, let us consider the special case of $T = 0$ and $0 < \alpha < 1$. By comparing (26) with (27), the number of servers reduced by adopting the DQ–JSQ design follows

$$\hat{n}_{\mathrm{P}}(\lambda,0,\alpha) - \hat{n}_{\mathrm{D}}(\lambda,0,\alpha) = \alpha \cdot \frac{\theta}{\mu+\theta} \cdot \frac{\lambda}{\mu} + o(\lambda).$$

The fraction of reduced servers is given by

$$\frac{\hat{n}_{\mathrm{P}}(\lambda,0,\alpha)-\hat{n}_{\mathrm{D}}(\lambda,0,\alpha)}{\hat{n}_{\mathrm{P}}(\lambda,0,\alpha)}=\alpha\cdot\frac{\theta}{\mu+\theta}+o(1). \tag{28}$$

Indeed, we may reduce the staffing level because more customers will abandon the system under the DQ–JSQ design: when the number of servers is close to the optimal level, the probability of delay in the DQ–JSQ system will be about α. In this case, the system must operate in the moderately overloaded regime (i.e., condition (1) holds), and almost all customers who cannot be served immediately have to wait until their respective servers complete the current services. Because the service times and the patience times are exponentially distributed and mutually independent, the probability of such a customer's PWT exceeding his patience time is $\theta/(\mu+\theta)$. Therefore, the fraction of customers abandoning the DQ–JSQ system will be about $\alpha\cdot\theta/(\mu+\theta)$. This implies that in order for the probability of delay to be about α, the number of servers in the DQ–JSQ system should be about $(1-\alpha\cdot\theta/(\mu+\theta))\cdot\lambda/\mu$, which is consistent with (26). By contrast, the PQ system must operate in the critically loaded regime, as long as the probability of delay is required to be strictly between 0 and 1 (Garnett et al. 2002). With the number of servers being about λ/μ, the customers' waiting times are generally short, and only a very small fraction of customers will abandon the PQ system. Hence, by allowing more customers to abandon the system, the DQ–JSQ design needs fewer servers than the PQ design to maintain a given probability of delay. In addition, when customers are less patient, the DQ–JSQ design may help us save more servers because there is a greater proportion of waiting customers abandoning the system. This insight can be confirmed by (28), where the fraction of reduced servers is approximately proportional to both α and the ratio $\theta/(\mu+\theta)$. The percentage of reduction will tend to α as θ increases.

When $T=0$, it follows from (26) that the optimal staffing level under the DQ–JSQ design will change with α at rate

$$\frac{\partial\hat{n}_{\mathrm{D}}(\lambda,0,\alpha)}{\partial\alpha}\approx-\frac{\theta}{\mu+\theta}\cdot\frac{\lambda}{\mu},$$

which is proportional to both the offered load λ/μ and the ratio $\theta/(\mu+\theta)$. In order to reduce the probability of delay by $\delta>0$, we should add about $\theta/(\mu+\theta)\cdot(\lambda/\mu)\cdot\delta$ servers. Clearly, more servers should be added when customers are less patient.

In the next numerical example, we assess the approximate optimal staffing level under the DQ–JSQ design and compare that with the optimal staffing level under the PQ design.

Example 4. We consider a queueing system with arrival rate $\lambda=400$, mean service time $1/\mu=1.0$, and mean patience time $1/\theta=2.0$. We would find the minimum numbers of servers to meet various service-level objectives with delay targets $T=1/6, 0.5, 1.0$, and 1.5 and threshold probabilities $\alpha=0.2, 0.5$, and 0.8, respectively. Both exact results and approximate results are summarized in Table 3. We obtain approximate optimal staffing levels under the DQ–JSQ design by (25) without the $o(\lambda)$ term, those under the PQ design by (27) without the $o(\sqrt{\lambda})$ term, and all exact optimal staffing levels using exhaustive search by simulation. Below each approximate staffing level, we provide the corresponding probability of the steady-state PWT exceeding the delay target T (with 95% confidence intervals), which is computed by simulation.

Table 3. Optimal Staffing Levels for Various Service-Level Objectives

		$\alpha=0.2$		$\alpha=0.5$		$\alpha=0.8$	
Case	Result type	DQ–JSQ	PQ	DQ–JSQ	PQ	DQ–JSQ	PQ
$T=1/6$	Exact	372	380	322	368	273	357
	App.	369	380	322	369	274	357
		0.215 ± 0.005	0.196 ± 0.005	0.499 ± 0.006	0.495 ± 0.006	0.795 ± 0.004	0.800 ± 0.005
$T=0.5$	Exact	359	323	292	312	219	302
	App.	357	323	291	312	219	302
		0.205 ± 0.005	0.201 ± 0.005	0.501 ± 0.006	0.492 ± 0.006	0.799 ± 0.005	0.796 ± 0.005
$T=1.0$	Exact	329	252	237	241	167	233
	App.	328	252	237	243	167	234
		0.202 ± 0.005	0.199 ± 0.005	0.498 ± 0.005	0.454 ± 0.006	0.799 ± 0.004	0.786 ± 0.005
$T=1.5$	Exact	281	197	192	190	128	181
	App.	281	198	192	189	129	181
		0.199 ± 0.004	0.183 ± 0.005	0.494 ± 0.005	0.504 ± 0.006	0.794 ± 0.005	0.797 ± 0.005

Notes. The Markovian queueing system has $\lambda=400$, $1/\mu=1.0$, and $1/\theta=2.0$ under the two queue structures. Exact results obtained by simulation are compared with approximations. The simulation results of corresponding performance (with 95% confidence intervals) are provided below the approximate optimal staffing levels.

With the largest staffing error being about 0.8%, the system's delay performance is generally satisfactory under the approximate optimal staffing levels given by (25). In particular, the case of $T = 1/6$ and $\alpha = 0.2$ corresponds to a scenario similar to the 80/20 rule in call centers: if the mean service time is about 2 minutes, then the delay target of 20 seconds is about 1/6 of the mean service time. In this case, we can reduce about 2.1% of servers under the DQ–JSQ design compared with the commonly used PQ design. Such a reduction is considerable when staffing costs are expensive.

Remark 5. As discussed in Remark 3 and justified by Theorem 5, in order for the probability of delay to be about $\alpha \in (0,1)$, the system can be staffed either in the moderately overloaded regime under the DQ–JSQ design or in the critically loaded regime under the PQ design. In the latter case, a small increase in the arrival rate may result in a significant change in delay performance. If the arrival rate is estimated to be λ_0, the optimal staffing level under the PQ design follows $\hat{n}_{\mathrm{P}}(\lambda_0, 0, \alpha) = \lambda_0/\mu + o(\lambda_0)$. Under such a staffing level, the system could be overloaded if the actual arrival rate is $\lambda = \lambda_0(1+\varepsilon)$ for some $\varepsilon > 0$. In this case, the probability of delay will be close to 1 even if ε is not large. Based on the estimated arrival rate, the optimal staffing level under the DQ–JSQ design satisfies $\hat{n}_{\mathrm{D}}(\lambda_0, 0, \alpha) = (\mu + \theta(1-\alpha))/(\mu(\mu+\theta)) \cdot \lambda_0 + o(\lambda_0)$. By part (iv) of Theorem 5, with the actual arrival rate $\lambda = \lambda_0(1+\varepsilon)$, the probability of delay will be about

$$P_{\mathrm{D}}(\mathrm{De}) \approx \left(\frac{\alpha}{1+\varepsilon} + \frac{\varepsilon(\mu+\theta)}{(1+\varepsilon)\theta}\right) \wedge 1,$$

which is strictly below 1 for $\varepsilon < \theta(1-\alpha)/\mu$. Hence, although the staffing level is lower, the delay performance under the DQ–JSQ design is less sensitive to estimation error in the arrival rate. In general, if a service-level objective requires the system to be staffed in the critically loaded regime under the PQ design, it could be more advantageous to adopt the DQ–JSQ design instead, not only because it may reduce the staffing level but also because it may mitigate the sensitivity issue that is intrinsic to the PQ structure.

Example 5. To illustrate the sensitivity issue in queueing design, we consider a system with mean service time $1/\mu = 1.0$, mean patience time $1/\theta = 2.0$, and estimated arrival rate $\lambda_0 = 100$. We would determine the minimum number of servers so that the probability of delay will not exceed $\alpha = 0.5$. By (26) and (27), the staffing level is set to $n = 84$ under the DQ–JSQ design and $n = 102$ under the PQ design. Then, we increase the customer arrival rate to $\lambda = \lambda_0(1+\varepsilon)$ with forecasting errors $\varepsilon = 0, 0.05, 0.1$, and 0.15. In Table 4, we report the simulation results of the probabilities of delay (with 95% confidence intervals) under the DQ–JSQ design and compare them with the exact results under the PQ design. The corresponding fluid approximations under the DQ–JSQ design are also listed in the same table. As we discussed earlier, the probability of delay appears to be less sensitive to a slight increase in the arrival rate under the DQ–JSQ design.

6.2. Boundary Function for Threshold Probabilities

To understand how the service-level objective may affect the queueing design, let us fix the delay target and compare the two designs as the threshold probability varies. To this end, we let

$$\begin{aligned}\hat{\alpha}(T) := {} & \mathbb{P}(T_a(m(T)+1) > T) - \psi(T) \\ & \times (\mathbb{P}(T_a(m(T)+1) > T) - \mathbb{P}(T_a(m(T)) > T)) \\ & \text{for } T \geq 0,\end{aligned} \tag{29}$$

where $m(T) := \lfloor \mu(\mathrm{e}^{\theta T} - 1)/\theta \rfloor$ and

$$\psi(T) := \frac{1}{\theta}\big(\mu + \theta m(T)\big)\big(\mu + \theta(m(T)+1)\big) \times \left(\frac{\mathrm{e}^{-\theta T}}{\mu} - \frac{1}{\mu + \theta(m(T)+1)}\right).$$

As implied by the next theorem, the function $\hat{\alpha}$ specifies the *boundary* for threshold probabilities, above which the DQ–JSQ design may require fewer servers than the PQ design.

Theorem 7. *For $T \geq 0$, $\lim_{\lambda\to\infty} \hat{n}_{\mathrm{D}}(\lambda, T, \alpha)/\hat{n}_{\mathrm{P}}(\lambda, T, \alpha) \leq 1$ if and only if $\alpha \geq \hat{\alpha}(T)$. Moreover, $\hat{\alpha}$ is continuous and strictly increasing on $[0, \infty)$ with*

$$\hat{\alpha}(0) = 0 \quad \textit{and} \quad \hat{\alpha}(\infty) := \lim_{T\to\infty} \hat{\alpha}(T) = \frac{\gamma(\mu/\theta, \mu/\theta)}{\Gamma(\mu/\theta)},$$

Table 4. Probabilities of Delay Under the PQ and DQ–JSQ Designs

Design	$\lambda = 100.0$ ($\varepsilon = 0$)	$\lambda = 105.0$ ($\varepsilon = 0.05$)	$\lambda = 110.0$ ($\varepsilon = 0.1$)	$\lambda = 115.0$ ($\varepsilon = 0.15$)
PQ ($n = 102$)	0.503	0.727	0.887	0.966
DQ–JSQ ($n = 84$)	0.490 ± 0.004	0.594 ± 0.004	0.694 ± 0.004	0.782 ± 0.005
App.	0.480	0.600	0.709	0.809

Notes. The Markovian queueing system has $1/\mu = 1.0$, $1/\theta = 2.0$, and $\lambda_0 = 100$ under the two queue structures. Both simulation results (with 95% confidence intervals) and fluid approximations are provided for the DQ–JSQ design; exact results are provided for the PQ design.

where Γ *is the gamma function (i.e.,* $\Gamma(s) := \int_0^\infty t^{s-1}e^{-t}\,dt$ *for* $s > 0$*) and* γ *is the lower incomplete gamma function (i.e.,* $\gamma(s,x) := \int_0^x t^{s-1}e^{-t}\,dt$ *for* $s > 0$ *and* $x \geq 0$*).*

Please refer to Section EC.7 for the proof of Theorem 7. We illustrate this boundary function in Figure 4 for $1/\mu = 1.0$ and $1/\theta = 2.0$. With a sufficiently large λ, the DQ–JSQ design requires fewer servers when (T, α) is above the curve (i.e., in the unshaded region), whereas the PQ design requires fewer servers when (T, α) is below the curve (i.e., in the shaded region). In particular, the fact that $\hat{\alpha}(0) = 0$ confirms that the DQ–JSQ design is more efficient in achieving any given probability of delay $\alpha > 0$. Because the boundary function is continuous, with a small delay target T (e.g., in terms of seconds in typical call centers), the DQ–JSQ design will also be more efficient as long as α is not too small. If the threshold probability satisfies $\alpha \geq \hat{\alpha}(\infty)$, the monotonicity of $\hat{\alpha}$ implies that the DQ–JSQ design must be more efficient, regardless of the value of T.

With $\mu > 0$ and $T \geq 0$ being fixed, the boundary probability $\hat{\alpha}(T)$ will change when θ takes different values. In Figure 5, we plot $\hat{\alpha}(T)$ as a function of θ with $1/\mu = 1.0$. For each fixed T, the difference between $\hat{\alpha}(\infty)$ and $\hat{\alpha}(T)$ gets smaller and smaller as θ increases. We observe that it would be generally satisfactory to use $\hat{\alpha}(\infty)$ to approximate $\hat{\alpha}(T)$ when $T \geq 1.0$ and $\theta \geq 2.0$. Note that $m(T)$ grows exponentially with both T and θ. When either T or θ is large, it may become intractable to compute the exact value of $\hat{\alpha}(T)$ by (23) and (29). Because it is much simpler to evaluate the regularized incomplete gamma function, $\hat{\alpha}(\infty)$ would be an attractive approximation in this case.

By extensive numerical tests, we observed that $\hat{\alpha}(T)$ is strictly increasing in θ. However, it is difficult to prove such monotonicity given the complicated expression in (29). Instead, we may prove monotonicity and limit results for $\hat{\alpha}(\infty)$, as stated in the following proposition.

Proposition 2. *With a fixed* $\mu > 0$, $\hat{\alpha}(\infty)$ *is strictly increasing in* θ *on* $(0, \infty)$, *having limits*

$$\lim_{\theta \downarrow 0} \hat{\alpha}(\infty) = \frac{1}{2} \quad \textit{and} \quad \lim_{\theta \to \infty} \hat{\alpha}(\infty) = 1.$$

Proposition 2 may help us determine the queueing design when the delay target is *relatively long* (e.g., several times longer than the mean service time). Because $\hat{\alpha}(\infty) > 1/2$, the PQ design is likely to be more efficient if α is well below $1/2$, no matter how patient customers would be (such a phenomenon is also evident in Figure 4). If customers are highly impatient (i.e., θ is large), $\hat{\alpha}(\infty)$ will be closer to 1. In this case, the PQ design is also likely to be more efficient as long as α is not close to 1. Along with the earlier discussion, we may summarize that the DQ–JSQ design is likely to be more efficient when T is small and α is well above 0, whereas the PQ design is likely to be more efficient when T is large and α is well below $\hat{\alpha}(\infty)$.

7. The DQ Structure with Partial Information

The JSQ policy requires complete state information (i.e., the number of customers of each server) to make routing decisions. When communication overhead is a constraint, routing policies that are capable of exploiting partial queue length information may become more attractive. In this section, we evaluate the power-of-d policy and the JIQ policy under the DQ structure by numerical experiments. The asymptotic analysis of these two policies is beyond the scope of this study, and we would like to investigate such topics in the future.

Figure 4. (Color online) Boundary Function $\hat{\alpha}$

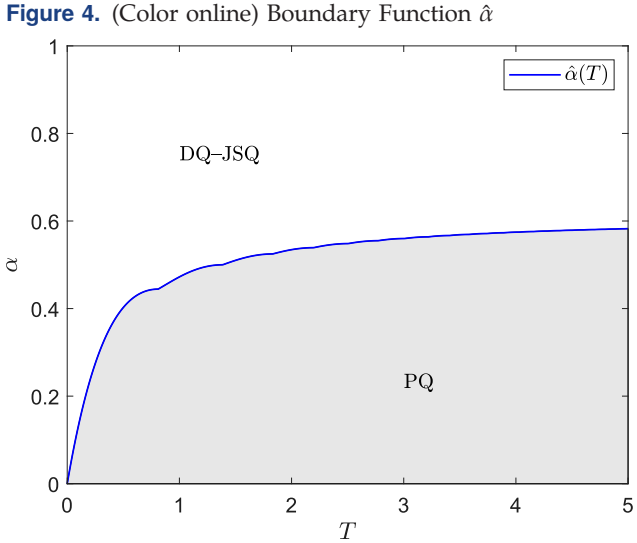
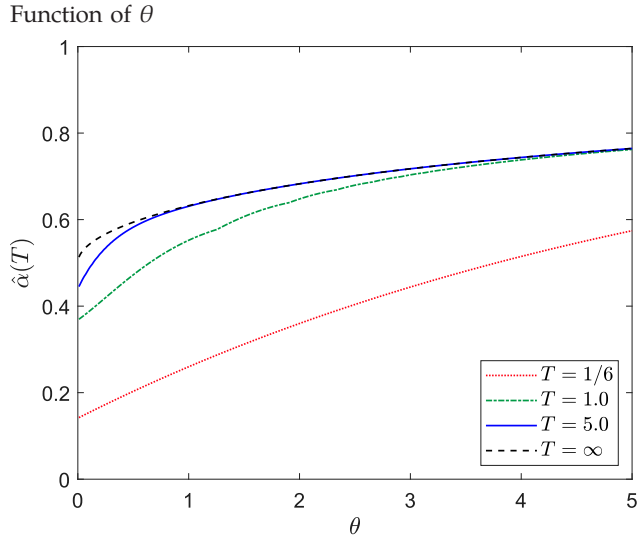


Figure 5. (Color online) Boundary Probability $\hat{\alpha}(T)$ as a Function of θ



Instead of comparing all queue lengths of the n servers, the power-of-d policy selects d servers at random and dispatches each customer to one of the d servers that has the fewest customers. This policy was analyzed by Mitzenmacher (2001), Chen and Ye (2012), Ying et al. (2017), and Mukherjee et al. (2020) under various settings. It is well known that the performance of the power-of-d policy could be close to that of the JSQ policy with d being much smaller than n (Mitzenmacher 2001, Chen and Ye 2012). The JIQ policy is another low-overhead alternative to the JSQ policy. If there are idle servers available upon a customer's arrival, the JIQ policy will dispatch the customer to an idle server; otherwise, the customer will be dispatched to a server selected at random. The JIQ policy may achieve comparable performance to the power-of-d policy with much lower communication overhead (see Lu et al. (2011) and Stolyar (2015), as well as the following discussion).

Example 6. We consider a queueing system with $1/\mu = 1.0$, $1/\theta = 2.0$, $\rho = 1.2$, and $n = 100$. In Figure 6, we plot the simulation results of several performance measures under the power-of-d policy (denoted by Pod in the figure) for d ranging from 1 to 30. These performance measures include the mean number of customers in the system, the mean PWT, the mean AWT, the probability of delay, the probability of abandonment, the mean number of idle servers (MNIS), and the probability of a customer joining a shortest queue (denoted by P(JSQ) in the figure).

The performance of the power-of-d policy converges to that of the JSQ policy as d increases. The power-of-d policy may achieve comparable performance to the JSQ policy for $d \geq 10$. The performance of the JIQ policy appears to be much better than the power-of-d policy when d is not large. This is because idle servers will frequently appear in the moderately overloaded regime and the probability of joining an

Figure 6. (Color online) Performance Comparison of the Power-of-d (Pod), JSQ, and JIQ Policies

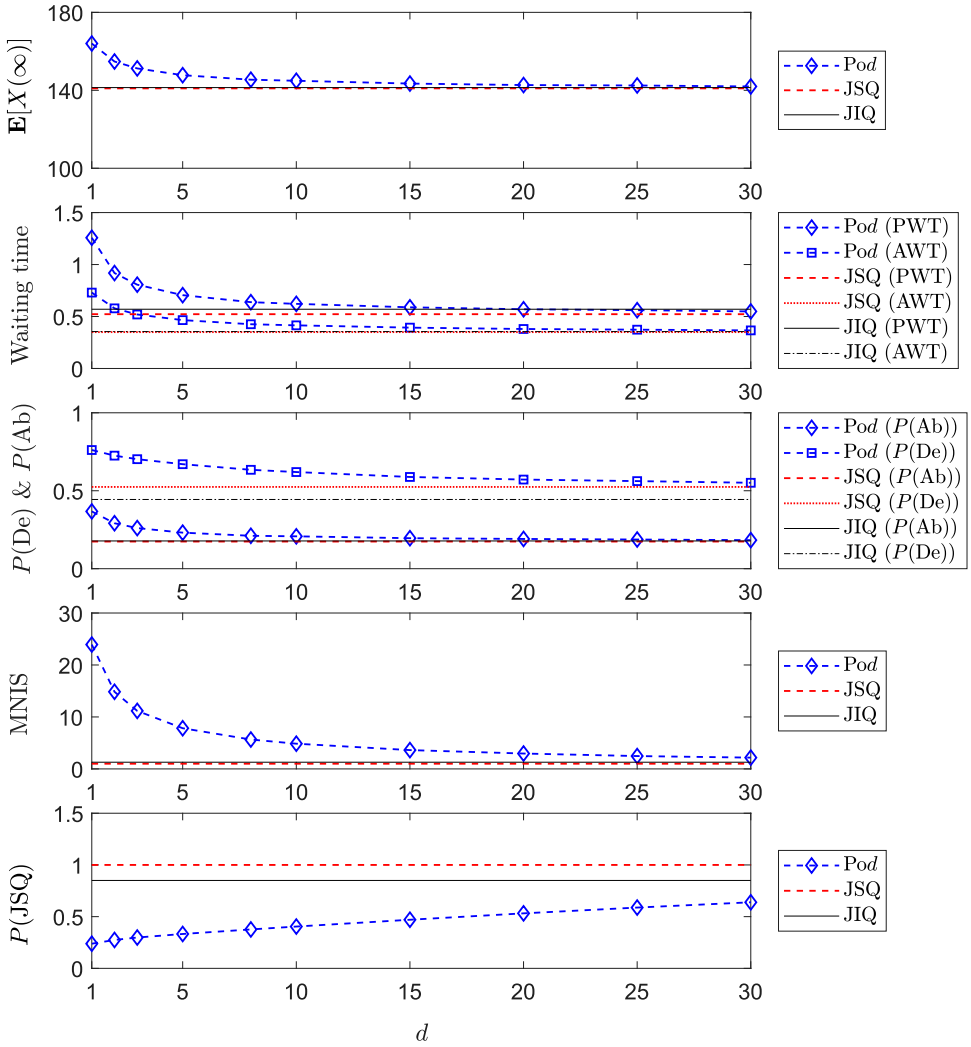


Table 5. Performance Comparison Between the JIQ and JSQ Policies

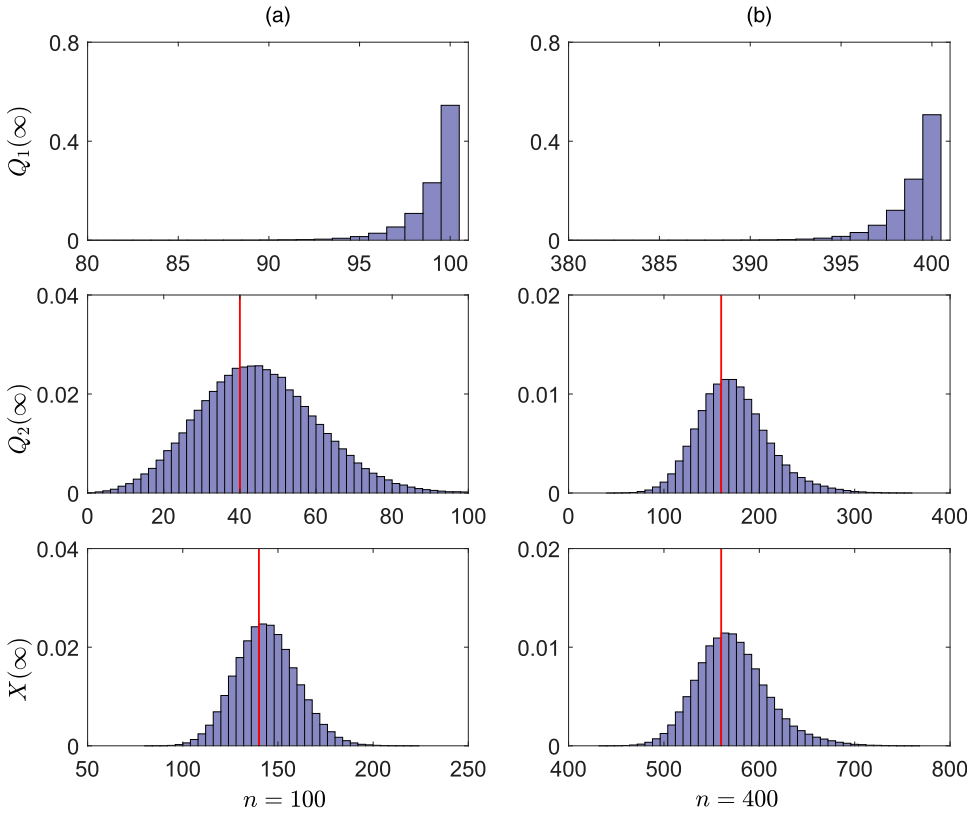
	$\rho = 1.02$			$\rho = 1.2$			$\rho = 1.5$		
	JIQ	JSQ		JIQ	JSQ		JIQ	JSQ	
Parameter	Sim.	Sim.	App.	Sim.	Sim.	App.	Sim.	Sim.	App.
P(De)	0.178 ± 0.002	0.188 ± 0.003	0.059	0.444 ± 0.003	0.525 ± 0.003	0.500	0.719 ± 0.003	0.940 ± 0.002	1.000
P(Ab)	0.064 ± 0.002	0.063 ± 0.002	0.020	0.178 ± 0.002	0.174 ± 0.002	0.167	0.336 ± 0.003	0.332 ± 0.003	0.333
$\mathbb{E}[X(\infty)]$	108.6 ± 0.2	108.4 ± 0.2	104.0	141.4 ± 0.2	141.0 ± 0.2	140.0	200.5 ± 0.2	200.2 ± 0.3	200.0
$\mathbb{E}[W]$	0.199 ± 0.004	0.189 ± 0.003	0.059	0.570 ± 0.005	0.520 ± 0.005	0.500	1.141 ± 0.007	1.025 ± 0.005	1.000
$\mathbb{E}[W\|W > 0]$	1.110 ± 0.012	1.005 ± 0.011	1.000	1.280 ± 0.008	0.995 ± 0.007	1.000	1.586 ± 0.007	1.089 ± 0.005	1.000
$\mathbb{E}[V]$	0.129 ± 0.002	0.126 ± 0.002	0.039	0.355 ± 0.003	0.350 ± 0.003	0.333	0.674 ± 0.004	0.670 ± 0.004	0.667
$\mathbb{E}[V\|W < R]$	0.089 ± 0.002	0.089 ± 0.002	0.027	0.268 ± 0.003	0.284 ± 0.003	0.267	0.585 ± 0.005	0.657 ± 0.004	0.667
$\mathbb{E}[V\|W > R]$	0.712 ± 0.014	0.674 ± 0.013	0.667	0.757 ± 0.008	0.666 ± 0.008	0.667	0.851 ± 0.007	0.699 ± 0.006	0.667
Var(W)	0.392 ± 0.011	0.346 ± 0.010	0.114	1.020 ± 0.019	0.762 ± 0.016	0.750	1.721 ± 0.029	1.116 ± 0.022	1.000
MNIS	4.496 ± 0.035	4.382 ± 0.035	—	1.293 ± 0.010	1.000 ± 0.010	—	0.395 ± 0.004	0.069 ± 0.002	—
P(JSQ)	0.971 ± 0.001	1.000 ± 0	1.000	0.850 ± 0.002	1.000 ± 0	1.000	0.582 ± 0.003	1.000 ± 0	1.000

Notes. The Markovian queueing system has $1/\mu = 1.0$; $1/\theta = 2.0$; $\rho = 1.02, 1.2,$ *and* 1.5; and $n = 100$ under the JIQ and JSQ policies. Simulation results (with 95% confidence intervals) are provided, along with fluid approximations for the JSQ policy.

idle server will be greater under the JIQ policy. The performance of the JIQ policy appears to be very close to that of the JSQ policy, with the probability of delay being even smaller. This is because customers may not join one of the shortest queues if there is no idle server. Those shortest queues will thus go empty more frequently, and more idle servers will appear. As a result, we may find the probability of delay even smaller than under the JSQ policy.

If servers are allowed to proactively send their states to the dispatcher, the communication overhead of the JIQ policy will be significantly lower than that of the power-of-d policy. Because a server needs to send a message to the dispatcher only when she completes a service and becomes idle, the average message exchange rate will not exceed one per customer under the JIQ policy. By contrast, the dispatcher needs to send d messages for requesting

Figure 7. (Color online) Estimates of the Steady-State Queue Length Distributions



queue lengths, and the d servers need to send d responses to the dispatcher under the power-of-d policy. The message exchange rate is $2d$ messages per customer. Because the JIQ policy could be more attractive as a low-overhead alternative, let us compare its performance with that of the JSQ policy more closely in the next example.

Example 7. We consider a queueing system with $1/\mu = 1.0$; $1/\theta = 2.0$; $n = 100$; and $\rho = 1.02, 1.2, 1.5$. Simulation results (with 95% confidence intervals) under the JIQ and JSQ policies are summarized in Table 5. With a lower communication overhead, the JIQ policy may achieve comparable performance in terms of the probability of abandonment, mean queue length, and mean waiting times. (The fluid model may not provide accurate approximations when the system operates in the critically loaded regime. In this table, fluid approximations for the JSQ policy appear less accurate with $\rho = 1.02$. Please refer to Section EC.10 of the e-companion for more discussion.) Because they may not join the shortest queues if no idle servers are available, customers are not well balanced across servers under the JIQ policy, having a greater variance of PWTs. In this sense, the fairness issue is more serious under the JIQ policy. As we discussed earlier, the JIQ policy renders a lower probability of delay, because more idle servers will be available for incoming customers.

8. Concluding Remarks

We conducted an asymptotic analysis of the DQ–JSQ system with customer abandonment, and we compared the system's performance with that of the PQ system in a moderately overloaded regime. We demonstrated that under the DQ–JSQ design, the queueing system may achieve a lower probability of delay or a higher service level, whereas the induced loss of capacity utilization is negligible. Therefore, the system may need a lower staffing level under the DQ–JSQ design to meet a certain service-level objective.

Although the fluid model may provide effective performance approximations for the DQ–JSQ system, such a model ignores stochastic fluctuations in the arrival, service completion, and abandonment processes. In Figure 7, we plot the estimates of the steady-state distribution of queue lengths in the DQ–JSQ system with $1/\mu = 1.0$; $1/\theta = 2.0$; $\rho = 1.2$; and $n = 100, 400$. Both $Q_2(\infty)$ and $X(\infty)$ appear to follow Gaussian distributions, with the mean values close to the corresponding fluid approximations (represented by solid vertical lines in the figure). To capture such distributions, we need to establish a refined model that takes into account the stochastic features of the queueing system. In the future, we may prove a diffusion limit for the DQ–JSQ system. Instead of approximating the queue length by $Q_i^n(t) = n\bar{Q}_i(t) + o(n)$, we expect that $Q_i^n(t) = n\bar{Q}_i(t) + \sqrt{n}\tilde{Q}_i(t) + o(\sqrt{n})$, where the diffusion limit $\tilde{Q}_i(t)$ characterizes stochastic fluctuations around the mean value, may provide more accurate approximations. This refined model may enable us to better estimate the distributions of queue lengths and PWTs, so that we may obtain more accurate solutions to staffing problems.

Acknowledgments

The authors thank the associate editor and two anonymous referees for many constructive comments and suggestions. This paper has been substantially improved by their work.

References

Amari SV, Misra RB (1997) Closed-form expressions for distribution of sum of exponential random variables. *IEEE Trans. Reliability* 46(4):519–522.

Aras AK, Chen X, Liu Y (2018) Many-server Gaussian limits for overloaded non-Markovian queues with customer abandonment. *Queueing Systems* 89(1–2):81–125.

Armony M, Roels G, Song H (2018) Pooling queues with strategic servers: The effects of customer ownership. Working paper, New York University, New York.

Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1):146–194.

Avi-Itzhak B, Levy H (2004) On measuring fairness in queues. *Adv. Appl. Probab.* 36(3):919–936.

Avi-Itzhak B, Levy H, Raz D (2008) Quantifying fairness in queuing systems. Principles, approaches, and applicability. *Probab. Engrg. Inform. Sci.* 22(4):495–517.

Banerjee S, Mukherjee D (2019) Join-the-shortest queue diffusion limit in Halfin–Whitt regime: Tail asymptotics and scaling of extrema. *Ann. Appl. Probab.* 29(2):1262–1309.

Baron O, Milner J (2009) Staffing to maximize profit for call centers with alternate service-level agreements. *Oper. Res.* 57(3):685–700.

Billingsley P (1999) *Convergence of Probability Measures*, 2nd ed. (John Wiley & Sons, New York).

Braverman A (2020) Steady-state analysis of the join-the-shortest-queue model in the Halfin–Whitt regime. *Math. Oper. Res.* 45(3): 1069–1103.

Chen H, Ye HQ (2012) Asymptotic optimality of balanced routing. *Oper. Res.* 60(1):163–179.

Dai JG, He S, Tezcan T (2010) Many-server diffusion limits for $G/Ph/n + GI$ queues. *Ann. Appl. Probab.* 20(5):1854–1890.

Ding Y, Park E, Nagarajan M, Grafstein E (2019) Patient prioritization in emergency department triage systems: An empirical study of Canadian Triage and Acuity Scale (CTAS). *Manufacturing Service Oper. Management* 21(4):723–741.

Eschenfeldt P, Gamarnik D (2018) Join the shortest queue with many servers. The heavy-traffic asymptotics. *Math. Oper. Res.* 43(3):867–886.

Foschini G, Salz J (1978) A basic dynamic routing problem and diffusion. *IEEE Trans. Comm.* 26(3):320–327.

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

Gupta V, Walton N (2019) Load balancing in the nondegenerate slowdown regime. *Oper. Res.* 67(1):281–294.

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

He S (2020) Diffusion approximation for efficiency-driven queues when customers are patient. *Oper. Res.* 68(4):1265–1284.

He S, Sim M, Zhang M (2019) Data-driven patient scheduling in emergency departments: A hybrid robust–stochastic approach. *Management Sci.* 65(9):4123–4140.

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

Jouini O, Dallery Y, Nait-Abdallah R (2008) Analysis of the impact of team-based organizations in call center management. *Management Sci.* 54(2):400–414.

Kang W, Ramanan K (2012) Asymptotic approximations for stationary distributions of many-server queues with abandonment. *Ann. Appl. Probab.* 22(2):477–521.

Kingman JFC (1962) The effect of queue discipline on waiting time variance. *Proc. Cambridge Philos. Soc.* 58(1):163–164.

Kleinrock L (1976) *Computer Applications, Volume II: Queueing Systems* (John Wiley & Sons, New York).

Koole G (2013) *Call Center Optimization* (MG books, Amsterdam).

Liu Y (2018) Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Oper. Res.* 66(2):514–534.

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

Liu Y, Whitt W (2014) Many-server heavy-traffic limit for queues with time-varying parameters. *Ann. Appl. Probab.* 24(1):378–421.

Lu Y, Xie Q, Kliot G, Geller A, Larus JR, Greenberg A (2011) Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation* 68(11):1056–1071.

Mandelbaum A, Reiman MI (1998) On pooling in queueing networks. *Management Sci.* 44(7):971–981.

Mandelbaum A, Zeltyn S (2009) Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.* 57(5):1189–1205.

Mandelbaum A, Sakov A, Zeltyn S (2001) Empirical analysis of a call center. Working paper, Technion–Israel Institute of Technology, Haifa.

Milner JM, Olsen TL (2008) Service-level agreements in call centers: Perils and prescriptions. *Management Sci.* 54(2):238–252.

Mitzenmacher M (2001) The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distributed Systems* 12(10):1094–1104.

Mukherjee D, Borst SC, van Leeuwaarden JSH, Whiting PA (2016) Universality of load balancing schemes on the diffusion scale. *J. Appl. Probab.* 53(4):1111–1124.

Mukherjee D, Borst SC, van Leeuwaarden JSH, Whiting PA (2020) Asymptotic optimality of power-of-*d* load balancing in large-scale systems. *Math. Oper. Res.*, ePub ahead of print January 20, https://doi.org/10.1287/moor.2019.1042.

Raz D, Avi-Itzhak B, Levy H (2005) Fair operation of multi-server and multi-queue systems. *SIGMETRICS Performance Evaluation Rev.* 33(1):382–383.

Raz D, Levy H, Avi-Itzhak B (2004) A resource-allocation queueing fairness measure. *SIGMETRICS Performance Evaluation Rev.* 32(1): 130–141.

Reiman MI (1984) Some diffusion approximations with state space collapse. Baccelli F, Fayolle G, eds. *Modelling and Performance Evaluation Methodology*, Lecture Notes in Control and Information Sciences, Vol. 60 (Springer, Berlin), 209–240.

Rothkopf MH, Rech P (1987) Perspectives on queues: Combining queues is not always beneficial. *Oper. Res.* 35(6):906–909.

Shi P, Chou MC, Dai JG, Ding D, Sim J (2016) Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Sci.* 62(1):1–28.

Shunko M, Niederhoff J, Rosokha Y (2018) Humans are not machines: The behavioral impact of queueing design on service time. *Management Sci.* 64(1):453–473.

Smith DR, Whitt W (1981) Resource sharing for efficiency in traffic systems. *Bell System Tech. J.* 60(1):39–55.

Song H, Tucker AL, Murrell KL (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Sci.* 61(12):3032–3053.

Stolyar AL (2015) Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems* 80(4):341–361.

Sunar N, Tu Y, Ziya S (2018) Pooled versus dedicated queues when customers are delay-sensitive. Working paper, University of North Carolina at Chapel Hill, Chapel Hill.

Wang J, Zhou YP (2018) Impact of queue configuration on service time: Evidence from a supermarket. *Management Sci.* 64(7):3055–3075.

Ward AR (2012) Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys Oper. Res. Management Sci.* 17(1):1–14.

Weber RR (1978) On the optimal assignment of customers to parallel servers. *J. Appl. Probab.* 15(2):406–413.

Whitt W (1986) Deciding which queue to join: Some counterexamples. *Oper. Res.* 34(1):55–62.

Whitt W (1999) Partitioning customers into service groups. *Management Sci.* 45(11):1579–1592.

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

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

Winston W (1977) Optimality of the shortest line discipline. *J. Appl. Probab.* 14(1):181–189.

Ying L, Srikant R, Kang X (2017) The power of slightly more than one sample in randomized load balancing. *Math. Oper. Res.* 42(3): 692–722.

Zeltyn S, Mandelbaum A (2005) Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems* 51(3–4):361–402.

Zhang H, Wang R (1989) Heavy traffic limit theorems for a queueing system in which customers join the shortest line. *Adv. Appl. Probab.* 21(2):451–469.

Zhang H, Hsu G-H, Wang R (1995) Heavy traffic limit theorems for a sequence of shortest queueing systems. *Queueing Systems* 21(1–2):217–238.

Ping Cao is an associate professor in the School of Management at University of Science and Technology of China. His research interests include dynamic control and optimization of stochastic systems, as well as their applications in operations management.

Shuangchi He is an associate professor in the Department of Industrial Systems Engineering and Management at the National University of Singapore. His research interests include stochastic modeling, analysis, and control as well as their applications in service, healthcare, and transportation systems. His work was awarded third place in the INFORMS Junior Faculty Interest Group Paper Competition in 2015.

Junfei Huang is an associate professor in the Department of Decision Sciences and Managerial Economics at the Chinese University of Hong Kong. His research interests are in asymptotic analysis and optimal control of queueing systems and their applications in manufacturing and services.

Yunan Liu is an associate professor in the Fitts Department of Industrial and Systems Engineering at North Carolina State University. His research interests include queueing theory; stochastic modeling; applied probability; simulation; online learning; and their applications in call centers, healthcare, transportation, blockchain, and manufacturing systems. His work was awarded first place in the INFORMS Junior Faculty Interest Group Paper Competition in 2016.