



## Mathematics of Operations Research

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

### Diffusion-Based Staffing for Multitasking Service Systems with Many Servers

Jaap Storm, Wouter Berkelmans, René Bekker

To cite this article:

Jaap Storm, Wouter Berkelmans, René Bekker (2024) Diffusion-Based Staffing for Multitasking Service Systems with Many Servers. *Mathematics of Operations Research* 49(4):2684–2722. <https://doi.org/10.1287/moor.2021.0051>

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](mailto: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 © 2023, 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>

# Diffusion-Based Staffing for Multitasking Service Systems with Many Servers

Jaap Storm,<sup>a,b</sup> Wouter Berkelmans,<sup>c</sup> René Bekker<sup>b,\*</sup>

<sup>a</sup>Department of Mathematics and Computer Science, Eindhoven University of Technology, 5600 MB Eindhoven, Netherlands; <sup>b</sup>Department of Mathematics, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, Netherlands; <sup>c</sup>Stochastics Group at Centrum for Wiskunde and Informatica, 1098 XG Amsterdam, Netherlands

\*Corresponding author

Contact: [p.j.storm@tue.nl](mailto:p.j.storm@tue.nl) (JS); [wberkelmans@gmail.com](mailto:wberkelmans@gmail.com) (WB); [r.bekker@vu.nl](mailto:r.bekker@vu.nl),  <https://orcid.org/0000-0002-5769-3624> (RB)

Received: February 22, 2021

Revised: May 4, 2022; September 19, 2023

Accepted: November 19, 2023

Published Online in Advance:  
December 28, 2023

MSC2020 Subject Classifications: Primary:  
90B22; secondary: 60K25, 60F17, 60E15

<https://doi.org/10.1287/moor.2021.0051>

Copyright: © 2023 INFORMS

**Abstract.** We consider a many-server queue in which each server can serve multiple customers in parallel. Such multitasking phenomena occur in various applications areas (e.g., in hospitals and contact centers), although the impact of the number of customers who are simultaneously served on system efficiency may vary. We establish diffusion limits of the queueing process under the quality-and-efficiency-driven scaling and for different policies of assigning customers to servers depending on the number of customers they serve. We show that for a broad class of routing policies, including routing to the least busy server, the same one-dimensional diffusion process is obtained in the heavy-traffic limit. In case of assignment to the most busy server, there is no state-space collapse, and the diffusion limit involves a custom regulator mapping. Moreover, we also show that assigning customers to the least (most) busy server is optimal when the cumulative service rate per server is concave (convex), motivating the routing policies considered. Finally, we also derive diffusion limits in the nonheavy-traffic scaling regime and in the heavy-traffic scaling regime where customers can be reassigned during service.

**Funding:** The research of J. Storm is partly funded by the Netherlands Organization for Scientific Research (NWO) Gravitation project Networks [Grant 024.002.003].

**Keywords:** diffusion limits • routing policies • square-root staffing • multitasking effects • multitasking service systems

## 1. Introduction

Many service systems have employees who can manage multiple customers concurrently because a single customer does not require the constant attention of an employee during the service request. This allows the employee to multitask by devoting her or his attention to other service requests in the meantime, potentially increasing the employee's efficiency. An illustrative example from the healthcare industry is where multiple patients are treated by a single nurse in intensive care units, clinical wards, or emergency departments (Elkhuizen et al. [15], Hall [21], Kc [31], Véricourt and Jennings [54]). Another prominent example is that of modern contact centers, where customer service representatives communicate with multiple customers at once via a customer chat channel (Cui and Tezcan [10], Legros and Jouini [34], Luo and Zhang [36], Tezcan and Zhang [51]). In addition to these, multiprogramming functionalities in computer systems (Horváth et al. [25], Schroeder et al. [48]), parole officers, and social welfare agencies (Campello et al. [8]) are mentioned as motivating practical examples for service systems involved with multitasking.

In this paper, we focus on the analysis of *many-server* service systems with multitasking. These systems require an analysis different from that of classical many-server queueing systems (see, e.g., Gans et al. [17], Halfin and Whitt [20]) because of the multitasking property. In particular, the system performance is strongly influenced by the efficiency gained or lost because of multitasking employees. This *multitasking effect* is complex and likely depends on the field of application (Delasay et al. [13], Douglas et al. [14]). On the one hand, the service rate per customer is expected to decrease in the number of simultaneously handled customers because of the server dividing her or his attention. On the other hand, it is known that challenging an employee with more work can have a positive effect on the service rate per customer, thus increasing the employee's effectiveness (Wickens et al. [58]). The effects of workload on processing efficiency have been statistically investigated in various types of service systems, such as in clinical settings (Berry Jaeker and Tucker [3], Kc [31], Kc and Terwiesch [32]), restaurant chains (Tan and Netessine [50]), contact centers (Hasija et al. [24]), and job shops (Bertrand and Van Ooijen [4]). In fact, it is frequently hypothesized that multitasking has an inverted U-shaped effect on throughput (Berry Jaeker and Tucker [3], Kc [31]).

For the efficient management of service systems, two major decisions apply, namely the *staffing level* and the *routing policy*. The former concerns the number of servers necessary to meet a certain service-level target; the latter prescribes to which available server a customer is routed depending on the number of customers they serve. These decisions are closely related and should thus be considered jointly. In Tezcan and Zhang [51], fluid model solutions to these problems are presented. In particular, they find an asymptotically optimal policy for a specific case of the multitasking effect. For this policy, the staffing problem is addressed in Cui and Tezcan [10] using diffusion processes arising from heavy-traffic scaling limits.

The primary goal of this paper is to use heavy-traffic diffusion limits to analyze the performance of multitasking service systems for different variants of the multitasking effect. Compared with Cui and Tezcan [10], our analysis covers a broad class of intuitive and popular routing policies and thus, provides both a qualitative and quantitative understanding of the interplay between routing decisions, different multitasking effects, and preferred staffing levels; the latter can be derived from the scaling used for the diffusion limit. We consider this to be our primary contribution. Here, it should be noted that heavy traffic does not necessarily mean that employees are at the limit of workload that they can handle but rather, that the workload offered to an employee is just below a workload threshold that is set by, for example, a system manager. In addition to diffusion limits, we obtain explicit expressions for the stationary distribution of the limiting diffusion processes. These can be used to obtain approximations for various steady-state performance measures of the system (e.g., involving the waiting time distribution or fraction of abandonments).

For our approach, we use the queue-based model proposed in Cui and Tezcan [10]. In that model, there are  $n$  servers, each being able to serve up to  $l$  customers simultaneously. Service times of customers are assumed to be exponential, where the service rate  $d_i$  of a server is a function of the number of customers  $i$  being served by that server. The multitasking effect is captured by the shape of the function  $i \mapsto d_i$ . Also, customer abandonment, because of customer impatience, is included. We primarily consider the quality-and-efficiency-driven (QED) scaling. In this regime, the arrival rate grows large, and the offered workload  $R$  is related to the number of servers  $n$  by the square-root principle  $n \approx R + \beta\sqrt{R}$ , with  $\beta$  a fixed parameter representing a service grade (see, e.g., Borst et al. [6], Gans et al. [17], Halfin and Whitt [20], Whitt [55]). For our analysis, we assume that  $i \mapsto d_i$  is increasing, thus covering a wide range of multitasking effects. This assumption is not too restrictive as the level of multitasking can be controlled by service system managers up to the point that multitasking is no longer efficient.

Using this setup, we derive diffusion limits covering a broad class of routing policies. For our first result, we consider a class  $\Pi$  of routing policies that, loosely speaking, send at least a small portion of the newly arrived customers to servers that have more than one free space. For each policy in  $\Pi$ , we obtain the same single-dimensional diffusion limit to determine staffing levels and characterize system performance. In particular, this shows that the diffusion results under QED scaling (and corresponding staffing levels and performance approximations) in Cui and Tezcan [10] are robust against errors or changes in the routing policy. Specifically, all policies in  $\Pi$  have the same asymptotic performance and are all asymptotically optimal by the result in Tezcan and Zhang [51]. As a counterpart, we show that a different asymptotic performance is obtained by the policy where a new customer is structurally sent to the most busy (available) server; this routing policy is called “most busy first” (MBF). Such a routing scenario can occur when a manager wants to maximize the number of free agents and turns out to be optimal when the multitasking effect is convex; we discuss this in more detail.

The class  $\Pi$  contains policies such as “least busy first” (LBF; also called *lightest load first* or *join the shortest queue*), “random available spot,” and “random available server” (also called *join idle queue*). These policies send a new customer to, respectively, the server with the smallest number of customers, a free spot selected uniformly at random over all free spots, and a random free server. The relevance of this class of policies lies in the trade-off between waiting time and communication overhead. In short, many routing policies require full information of the system’s state, and the necessary communication overhead becomes large when the number of servers increases, whereas random policies require little to no communication overhead. This trade-off between waiting time and overhead is a recent area of study in the so-called supermarket model, where arriving customers have to be routed to one of  $n$  parallel single-server queues (see, e.g., van der Boor et al. [52]).

For the proofs of our results, we use a novel approach compared with Cui and Tezcan [10]. The proof of the result for the policy class  $\Pi$  uses direct arguments, which can be easily explained on an intuitive level. In particular, we prove a state-space collapse result using a stochastic coupling technique. The proof of the MBF diffusion limit is considerably more involved and cannot be covered by the aforementioned technique or the approach in Cui and Tezcan [10]. Specifically, the result does not include a state-space collapse, and the limiting process is a reflected diffusion process instead. For the proof, we develop a technique based on a new type of regulator mapping associated with an extension of so-called generalized Skorokhod problems (see, e.g., Reed and Ward [42], Reed and Ward [43], Reed et al. [44]).

The policies that we consider are motivated by what we consider to be our secondary contribution: prelimit optimality results for two cases of the multitasking effect. To be precise, we show that routing customers to the least (most) busy available server is optimal, in terms of customers present and the number of abandonments, when the multitasking effect  $i \mapsto d_i$  is concave (convex). We emphasize that this strengthens the asymptotic optimality results for the LBF policy from Tezcan and Zhang [51], and the optimality criteria for the MBF policy have not been considered before. The stochastic ordering also applies in the reversed setting, showing that sending customers to the least (most) busy server yields the worst performance for a convex (concave) multitasking effect.

Nonheavy-traffic diffusion limits can also be established for systems operating under the LBF and MBF policies. The LBF case is covered in Cui and Tezcan [10], and we derive fluid and diffusion limits for the MBF case with subcritical workloads, relying on the same methods we use in the heavy-traffic scaling regime. The results for the MBF policy are particularly interesting because they show that by routing customers to the busiest server, a fraction of all servers will be empty. These servers can be used for other tasks (e.g., addressing other communication channels in contact centers or taking care of different patient groups in hospital wards).

Finally, we also cover the instances of service systems where there is the flexibility to transfer customers among servers during service. For this case, a routing policy should be interpreted as the allocation of customers among servers, which may change at any moment. In Legros and Jouini [34], it is shown that allocating customers to the least (most) busy server is optimal for concave (convex) multitasking effects. For these cases, we provide heavy-traffic diffusion limits from which favorable staffing levels can be derived. As in Legros and Jouini [34], we use the terms *shared work case* (SWC) and *nonshared work case* (NWC) to refer to, respectively, systems that allow customer transfers between servers and those that do not.

### 1.1. Organization and Notation

The paper is organized as follows. In Section 2, we start by introducing the model and discussing the asymptotic setup under which we derive the diffusion limits. Our primary contributions (i.e., the heavy-traffic diffusion limits for both classes  $\Pi$  and MBF) are presented in Section 3. Subsequently, diffusion limits for the SWC are given in Section 4. The LBF and MBF policies are further motivated in Section 5 by presenting optimality properties (i.e., our secondary contribution). Application of diffusion limits to approximate the probability of delay is discussed in Section 6, providing further intuitive insight into system performance compared with, for example, the Erlang A model. In Section 7, we study the diffusion-level behavior of our model in nonheavy-traffic regimes. In Sections 8 and 9, we present the proofs of the policy classes  $\Pi$  and MBF, respectively; the proofs of the other results can be found in the appendix. Numerical illustrations can be found in Section 10. Some straightforward model extensions are discussed in Section 11, whereas Section 12 concludes.

In this paper, all random elements are defined on a single probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . All vectors are understood as column vectors. We adopt the convention of denoting a stochastic process  $\{X(t), t \geq 0\}$  by  $X(\cdot)$ , and by  $X(t)$ , we denote the corresponding random variable at time  $t \geq 0$ . For  $d \in \mathbb{N}$ , we denote by  $D^d[0, \infty)$  the space of càdlàg functions  $f: \mathbb{R}_+ \rightarrow \mathbb{R}^d$ . We endow this space with the Skorohod topology; however, because all convergence results have continuous limits, we can also consider the uniform topology on compact time intervals. For a process  $X(\cdot)$  and  $t \geq 0$ , we denote by  $\|X(\cdot)\|_t$  the uniform norm of  $X(\cdot)$  on  $[0, t]$ . Weak convergence of a sequence of processes  $\{X_n(\cdot)\}_n \subset D^d[0, \infty)$  to a limit  $X(\cdot) \in D^d[0, \infty)$  is denoted  $X_n(\cdot) \Rightarrow X(\cdot)$ , and we write  $X_n(\cdot) \xrightarrow{\mathbb{P}} X(\cdot)$  when  $X_n(\cdot)$  converges to  $X(\cdot)$  in probability, with the uniform norm on compact time intervals. The order notation should be interpreted in a stochastic sense; that is, for a sequence of stochastic variables  $\{X_n\}_{n \geq 1}$  and function  $g(n)$ , we write  $X_n = o(g(n))$  if  $X_n/g(n)$  converges to zero in probability, and we write  $X_n = \mathcal{O}(g(n))$  if  $X_n/g(n)$  has a nontrivial distribution, as  $n \rightarrow \infty$ . Finally, for  $x \in \mathbb{R}$ , we write  $x^+ = \max\{x, 0\}$  and  $x^- = \max\{-x, 0\}$ .

### 1.2. Literature

We provide an overview of the streams of literature related to our work and describe how our results fit into these streams of literature.

*Processor sharing queues* are closely related to our work, as the way in which customers are served in a multitasking system can be interpreted as a type of (limited) processor sharing (or round robin). A key difference between limited processor sharing (LPS) and multitasking systems is that in the former, there is a single service entity for all customers, whereas in the latter, there are multiple servers, each managing a subset of the customers present. Also, in LPS, the total service rate is constant, whereas for multitasking systems, this depends on the number of customers in service. Some seminal work on processor sharing systems involves Kleinrock [33] and Yashkov [60]; more recent studies on LPS can be found in, for example, Zhang and Zwart [61] and Zhang et al. [62]. We refer to Altman et al. [2] for the study of routing customers to parallel processor sharing queues.



The main results in this paper are *many-server diffusion approximations* in the QED regime, for which there is a rich literature. The starting point for such diffusion limits is the celebrated paper by Halfin and Whitt [20]. Subsequently, many papers have been devoted to extensions: for example, by including customer abandonments (Dai et al. [12], Garnett et al. [18]), by focusing on dimensioning and/or constraint satisfaction (Borst et al. [6], Mandelbaum and Zeltyn [37]), by considering networks (Mandelbaum et al. [38]), by refinements of the staffing rules (Janssen et al. [27]), and by having time-dependent demand (Jennings et al. [28]). As the amount of literature is considerable, we refer to Gans et al. [17] for an early exposition of the QED and other regimes in call centers; see van Leeuwen et al. [53] for a recent partial review of queues in the QED regime containing many references.

Our methodology to prove convergence in the NWC with the MBF policy is based on the theory of *Skorokhod problems*. In simple terms, this theory describes a wide class of reflected processes as the image of a process with unconstrained paths under a continuous mapping. Therefore, by the continuous mapping theorem, deriving a diffusion limit for a reflected stochastic process reduces to deriving a diffusion limit for the unconstrained process. We point to Whitt [56, chapter 14] as a standard reference for this theory. This technique has frequently been used to obtain reflected Brownian models, building on Harrison and Williams [23], as approximations to queueing models; see, for example, Chen and Mandelbaum [9], Harrison and Williams [22], and Reiman [45] for a few examples. For reflected processes that include a position-dependent drift component, so-called generalized Skorokhod problems have been developed and employed in Reed and Ward [42], Reed and Ward [43], and Reed et al. [44]. For our case, this variant of the Skorokhod problem cannot be directly applied. Instead, we develop a modified version of the generalized Skorokhod problem to prove our results.

The literature on *multitasking, multiserver systems* is relatively small. Studies are primarily motivated by chat contact centers (Cui and Tezcan [10], Legros and Jouini [34], Luo and Zhang [36], Tezcan and Zhang [51]) and consider similar types of models. The recent paper (Legros and Jouini [34]) focuses on the routing decisions in both the SWC and NWC scenarios. For the SWC, they show that LBF (MBF) maximizes the service rate for concave (convex) multitasking effects. For the NWC, the authors consider the optimality of idling and further adopt a one-step policy improvement approach. The papers by Cui and Tezcan [10], Luo and Zhang [36], and Tezcan and Zhang [51] focus on asymptotic analysis in the many-server heavy-traffic regime. In Luo and Zhang [36], fluid approximations are established, and in Tezcan and Zhang [51], asymptotically optimal routing policies and staffing levels are given based on fluid models. In this paper, we strengthen the routing result by providing (nonasymptotically) optimal routing policies for convex and concave increasing multitasking effects. We also obtain a lower bound on the system performance. Our proofs regarding routing disciplines are based on weak majorizations and follow the lines of proof and stochastic couplings as given in Akgun et al. [1] and Sparaggis et al. [49]. In Cui and Tezcan [10], heavy-traffic diffusion limits are established for multitasking service systems to address the staffing problem. Their results are proven for what we refer to as the NWC with the LBF policy, whereas we consider a broader class of routing policies  $\Pi$  as well as the MBF policy for the NWC. Moreover, we also consider diffusion approximations for the SWC. Model extensions to general service and patience times can be found in Long et al. [35].

## 2. Model Definition and Asymptotic Framework

In this section, we introduce the model, the state description, and the routing policies, and we finish by presenting the asymptotic framework. For clarity of exposition, we introduce a basic model. The possible extensions and ramifications are discussed in Section 11.

### 2.1. Model Definition

The main focus of our study is an  $n$ -server queueing system, with  $n \in \mathbb{N}$ , to which customers arrive according to a Poisson process with rate  $\lambda > 0$ . Servers are able to serve up to  $I \in \mathbb{N}$  customers simultaneously to reflect the phenomenon of multitasking. A server is said to be at level  $i \in \{0, \dots, I\}$  when it is serving  $i$  customers. When there are  $i \in \{0, \dots, I\}$  customers at a server, each customer has a service time that is exponentially distributed with rate  $\mu_i > 0$ . In addition, we assume that every customer in service abandons the system after an exponentially distributed time with parameter  $\theta_s \geq 0$ . Consequently, for  $i \in \{1, \dots, I\}$ , the time until the next customer is removed from a level  $i$  server is exponentially distributed with parameter  $i(\mu_i + \theta_s) := d_i$ . When every server is at level  $I$ , an arriving customer will enter a queue, which is emptied according to a first come, first served discipline. We assume that all servers are at level  $I$  when there are at least  $nI$  customers in the system (i.e., the system is work conserving). Finally, we include customer impatience during waiting by imposing that each customer in the queue abandons the system after an exponential time with parameter  $\theta \geq 0$ .

Throughout the paper, we assume that  $d_i < d_j$  for  $i < j$  (i.e.,  $d_i$  is increasing in  $i$ ). This is a natural assumption as a system manager can control the level of multitasking up to the point where efficiency is lost (i.e., we expect that  $i\mu_i$

will be increasing in  $i$ ). By definition of  $d_i$ , we, therefore, expect that  $d_i$  is increasing in most practical applications. The assumption also plays a technical role in the proofs of our results.

## 2.2. System State and Policies

The state of the multitasking queueing system can be described by the stochastic process  $Z = \{Z(t), t \geq 0\}$ , with  $Z(t)$  defined as

$$Z(t) := (Q(t), n_0(t), n_1(t), \dots, n_{I-3}(t), n_{I-2}(t))^T,$$

where  $Q(t)$  is the total number of customers in the system, and for  $i \in \{0, 1, \dots, I\}$ ,  $n_i(t)$  is the number of level  $i$  servers, both at time  $t \geq 0$ . We note that the two components  $n_{I-1}(t)$  and  $n_I(t)$  are not required to describe the state of the system because  $n_I(t) = n$  and  $n_{I-1}(t) = 0$  when  $Q(t) \geq In$ , and for  $Q(t) < In$ , we can find  $n_i(t)$  and  $n_{I-1}(t)$  from

$$\begin{cases} n = \sum_{i=0}^I n_i(t) \\ Q(t) = \sum_{i=1}^I i n_i(t). \end{cases}$$

Hence, for all  $t \geq 0$ , we have

$$n_{I-1}(t) = \max\{In - Q(t), 0\} - \sum_{i=2}^I i n_{I-i}(t), \quad n_I(t) = n - \sum_{i=0}^{I-1} n_i(t). \quad (1)$$

As such, we can refer to  $n_i(t)$  and  $n_{I-1}(t)$  when speaking about  $Z(t)$ , even though they are not directly included in the state description.

The system dynamics depend on the routing policy. Three classical customer routing policies are the ones in which arriving customers are routed to (1) available servers with the least number of customers, (2) available servers with the most number of customers, and (3) random available servers that are not at level  $I$ . We refer to these policies as LBF, MBF, and *random-server* (RS) policies, respectively. We note that LBF and RS are, in fact, special cases of a wider class of policies  $\Pi$  that we introduce in Section 3.1.

Observe that  $Z(\cdot)$  is an irreducible Markov process with state space  $\{(Q, n_0, \dots, n_{I-2}) \in \mathbb{Z}_+ \times \{0, \dots, n\}^{I-2} : n_{I-1}, n_I \geq 0\}$ . It can be shown using a Foster–Lyapunov theorem (e.g., Robert [47, theorem 8.13]) that  $Z(\cdot)$  is positive recurrent when  $\theta > 0$  or when  $\lambda/(d_I n) < 1$  if  $\theta = 0$ . For  $\theta > 0$ , the intuition is that when a sufficiently large queue builds up, then the number of customers in the system is dominated by the occupancy process of an infinite-server queue, which is always stable. For  $\theta = 0$ , whenever there is a queue,  $Q(\cdot)$  behaves like an  $M/M/1$  queue with arrival rate  $\lambda$  and departure rate  $d_I n$ , which is stable if and only if  $\lambda/(d_I n) < 1$ .

Throughout the remainder of the paper, we assume that either  $\theta > 0$  or  $\lambda/(nd_I) < 1$ , in which case  $Z(\cdot)$  has a unique stationary distribution, which can be determined by solving the balance equations. For  $Q(\cdot) \geq In$ , the special structure of the Markov process may be used, whereas a system of equations needs to be solved for all states in case of an empty queue. This involves  $\binom{n+I}{n}$  states, which grow as an  $I$ th order polynomial in  $n$ . One cannot hope for an explicit expression for the stationary distribution. In addition, numerical solutions become unfeasible when  $n$  is large and would not give any deeper insight into the system behavior. Therefore, we focus on an asymptotic analysis of the multitasking system, involving diffusion limits.

**Remark 1.** Note that the system may also be considered as a generalization of an Erlang A model because in the case that  $d_i = id_1$ , for each  $i \in \{1, \dots, I\}$ , the system is equivalent to an  $M/M/In + M$  queue.

## 2.3. Heavy-Traffic Asymptotic Framework

After fixing a routing policy, we determine staffing levels by studying the heavy-traffic behavior of  $Z(\cdot)$  on a diffusion scale. We consider a scaling of the arrival rate  $\lambda \mapsto \lambda_n$ , depending on the number of servers  $n$ , such that the load per server  $\rho_n := \lambda_n/(d_I n) \uparrow 1$  as  $n \rightarrow \infty$ ; nonheavy-traffic scaling regimes for the multitasking system are considered in Section 7. As in Cui and Tezcan [10], we do not consider the overloaded regime, as the asymptotic behavior then coincides with the overloaded Erlang A model; see, for example, Whitt [57] for the Erlang A model in the efficiency-driven regime. There are many possible ways to achieve a heavy-traffic scaling, but we focus on the well-known QED regime (see, e.g., Cui and Tezcan [10], Gans et al. [17], Halfin and Whitt [20]). This regime balances the

system load and can be used to analyze many-server systems where the probability that an arriving customer has to wait is nontrivial in the stationary regime (cf. Puhalskii and Reiman [41]).

In their celebrated paper, Halfin and Whitt show (cf. Halfin and Whitt [20, proposition 1]) that, for the classical Erlang delay model with a fixed service rate, there is a unique way to scale the arrival rate with the number of servers such that the probability that an arriving customer has to wait is nontrivial (i.e., is strictly between zero and one). In this paper, we adopt a similar scaling, which has also been applied for the Erlang A model; see Garnett et al. [18]. More specifically, we consider a sequence of stochastic processes  $\{Z_n(\cdot)\}_{n \geq 1}$ , with  $Z_n(\cdot) = \{Z_n(t), t \geq 0\}$  as a multitasking system having  $n$  servers and

$$\lambda_n = d_I(n - \sqrt{n}\beta) \tag{2}$$

being the corresponding arrival rate, with  $\beta \in \mathbb{R}$ . Observe that when  $\theta = 0$ , we require that  $\beta > 0$  for the system to be stable. To emphasize their dependence on  $n$ , we write  $Q_n(\cdot)$  and  $n_{n,i}(\cdot)$ , for  $i \in \{0, \dots, I - 2\}$ , for the components of  $Z_n(\cdot)$ . The diffusion limits we derive for the multitasking system under this scaling regime can be related to the diffusion behavior of the Erlang A model as discussed in Section 6.

**Remark 2.** In the heavy-traffic scaling that we consider, it may appear that almost every server is close to the limit of the workload that it can handle as  $n \rightarrow \infty$ , which may be unrealistic depending on the application. We emphasize that this is not necessarily the case because the interpretation of  $I$  can be modified so that servers experience a preset maximum level of workload. For example, define  $I_1$  as the maximum level at which multitasking is efficient (i.e., the point up to which  $i\mu_i$  is increasing in  $i$ ) and  $I_2$  as the maximum level at which the workload per server remains below a prespecified threshold (or alternatively, the quality of service is high enough). The following cases of applications can be distinguished.

- i. The case  $I = I_1 = I_2$  applies to systems where full server utilization is desirable (e.g., multiprogramming computer systems).
- ii. The case  $I = I_2 < I_1$  refers to, for example, nonemergency healthcare and chat service systems, where a strict level  $I$  cap can be maintained by a system administrator and some modest waiting times are preferable to overloading the servers.
- iii. The case  $I = I_1 > I_2$  is related to systems where waiting should be avoided at any cost: for instance, emergency situations, such as emergency departments in hospitals.

Our diffusion analysis of the multitasking model in heavy-traffic regimes is primarily useful in cases (i) and (ii). In case (iii), nonheavy-traffic scaling regimes should be considered, which we do in Section 7.

On a fluid scale, all servers will be occupied in heavy traffic so that  $In$  will be an invariant state for the number of customers in the system (cf. Cui and Tezcan [10, theorem 2]). By centering around  $In$  in the first component of  $Z_n(\cdot)$ , its value represents either the queue length (when  $Q_n - In \geq 0$ ) or minus the number of available places (when  $Q_n - In < 0$ ). As  $n_I(\cdot)$  and  $n_{I-1}(\cdot)$  are not (directly) in the state description, there is no centering for the other components. To derive the diffusion limits, we consider a sequence of the scaled and centered processes  $\{\hat{Z}_n(\cdot)\}_{n \geq 1}$  given by

$$\hat{Z}_n(t) = \frac{(Z_n(t) - (In, 0, \dots, 0)^T)}{\sqrt{n}}, \quad t \geq 0, \tag{3}$$

with  $(In, 0, \dots, 0) \in \mathbb{R}^I$ . We denote the components of  $\hat{Z}_n(\cdot)$  by  $\hat{Q}_n(\cdot) = n^{-1/2}(Q_n(\cdot) - In)$  and  $\hat{n}_{n,i}(\cdot) = n^{-1/2}n_{n,i}(\cdot)$  for  $i \in \{0, \dots, I - 2\}$ .

For our diffusion limits, we require that the initial conditions converge in a suitable manner along with the dynamics of our process. This can be achieved in various ways and is mostly a technical condition. For our purposes, it is sufficient that the initial conditions converge weakly. To be specific, we assume that  $\hat{Z}_n(0)$  converges weakly to a random vector  $\hat{Z}(0)$  in  $\mathbb{R}^I$ , as  $n \rightarrow \infty$ , denoted by  $\hat{Z}_n(0) \Rightarrow \hat{Z}(0)$ , independently of the processes driving the evolution of  $Z_n(\cdot)$ . In addition, for each of our theorems, we assume that  $\hat{Z}(0)$  is a random variable taking on values in a space that matches the state space of the limiting diffusion process  $\hat{Z}(\cdot)$  corresponding to the specific theorem. When stating our theorems, we abbreviate these assumptions to the statement “suppose  $\hat{Z}_n(0) \Rightarrow \hat{Z}(0)$ .”

### 3. Heavy-Traffic Diffusion Limits

In this section, we present our main results (i.e., the diffusion limits for  $Z(\cdot)$  in the QED heavy-traffic scaling regime). We consider two fundamentally different cases. On the one hand, in Section 3.1, we provide the diffusion approximation under the class of policies as defined in Definition 1 (including the LBF and RS policies), giving rise to a one-dimensional limiting process. On the other hand, the diffusion limit in Section 3.2 for the MBF policy

is a multidimensional reflected diffusion process with nonhomogeneous drift characterized as the solution of a custom version of the Skorokhod problem.

The impact of the routing policies in terms of the probability of delay is deferred to Section 6. Proofs of the diffusion limits are given in Sections 8 and 9.

### 3.1. Diffusion Limits for the LBF and RS Policies

In the present section, we derive a diffusion approximation for  $Z(\cdot)$  under the LBF and RS policies. For both policies,  $\hat{Z}_n(\cdot)$  will have the same diffusion limit, showing that the policies have the same asymptotic performance. In fact, the proof for the LBF and RS policies can be easily extended to a wider class of policies, which we, therefore, consider instead.

**Definition 1** (Policy Class  $\Pi$ ). For some  $c > 0$  and with the convention that  $0/0 = 0$ , define  $\Pi$  as the set of policies so that given  $Z(t)$  at  $t \geq 0$ , we have the following assignment probability for an arriving customer at time  $t$ :

$$\mathbb{P}(\text{Route to server at level } I - 2 \text{ or below}) \geq \frac{ck(t)}{(Q(t) - In)^-},$$

where the process  $k(\cdot)$  is defined as

$$k(\cdot) := \sum_{i=0}^{I-2} (I - i - 1)n_i(\cdot). \tag{4}$$

Observe that at time  $t \geq 0$ ,  $k(t)$  is the number of customers missing at servers at level  $I - 2$  or below that would put all of these servers at level  $I - 1$ . It turns out that  $\hat{Z}_n(\cdot)$ , under all policies in  $\Pi$ , converges weakly to a diffusion process  $\hat{Z}(\cdot)$ , showing that all policies in  $\Pi$  have the same asymptotic performance.

It can be easily seen that LBF and RS policies are in  $\Pi$  because the probabilities of routing a customer to a server at level  $I - 2$  or below are  $\mathbb{1}_{\{k(t) > 0\}}$  and  $(k(t) + \sum_{i=0}^{I-2} n_i(t))/(Q(t) - In)^-$ , respectively. There are also other relevant policies in  $\Pi$ : for instance, the policy that sends customers to a random free spot; the relevant routing probability is equal to  $\sum_{i=0}^{I-2} (I - i)n_i(t)/(Q(t) - In)^-$ . As all these policies turn out to have the same asymptotic performance, it shows that in practice, one has to purposely avoid sending customers to level  $I - 2$  servers or below in order to obtain a different asymptotic performance in heavy traffic. In Section 3.2, we discuss the MBF policy, which is not in  $\Pi$  and has different asymptotic performance.

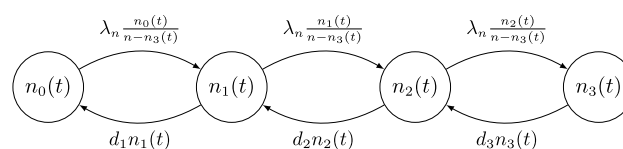
We start our exposition by providing some intuition for the main result. An explicit expression for the stationary distribution of the diffusion limit can be found in Section 6, which can be interpreted in terms of an Erlang A system.

**3.1.1. Intuition.** For the current routing policy, it turns out the service system, when in heavy traffic, exclusively has level  $I$  and level  $I - 1$  servers. To explain the reason behind this, let  $I = 3$ , and consider the system under the RS policy. If  $Q(t) \geq In$ , there can clearly only be level  $I$  servers. When  $Q(t) < In$ , the process is more complex; the rate at which servers at level  $i$  become servers at level  $j$ , for  $i \in \{0, \dots, 3\}$ ,  $j \in \{\max\{i - 1, 0\}, \min\{i + 1, 3\}\}$ , is represented schematically in Figure 1. Because  $Z_n(\cdot)$  is ergodic, it has a unique limiting distribution such that the rates in and out in Figure 1 should be balanced. For  $n_3(t)$ , this gives

$$d_3 n_3(t) = \lambda_n \frac{n_2(t)}{n - n_3(t)}. \tag{5}$$

Because  $\rho_n \uparrow 1$ , we have that for  $n$  large, the number of servers with a free space  $n - n_3(t)$  should be  $o(n)$ , and because we are considering a generalization of the M/M/n + M queue under the QED scaling, it is conceivable that  $n - n_3(t) = O(\sqrt{n})$ . As we have chosen  $\lambda_n = O(n)$ , (5) implies that  $n_2(t) = O(\sqrt{n})$  in the stationary regime. Repeating this argument, we get  $n_1(t) = O(1)$  and  $n_0(t) = O(\frac{1}{\sqrt{n}})$ . We deduce that when we scale the system by  $n^{-1/2}$ , only  $n_3(\cdot)$

**Figure 1.** Schematic representation process rates when  $Q(t) < 3n$ .



Downloaded from informas.org by [192.16.191.136] on 12 December 2024, at 05:52. For personal use only, all rights reserved.



and  $n_2(\cdot)$  will have nontrivial limits as  $n \rightarrow \infty$ . This argument can be generalized to arbitrary  $I \in \mathbb{N}$  and to every policy in  $\Pi$ , where for  $n$  large, the number of level  $i$  servers in the stationary regime is  $\mathcal{O}(n^{(i-I+2)/2})$ ,  $i \in \{0, \dots, I\}$ .

**3.1.2. Diffusion Limit.** Based on the argument, the process  $\hat{Z}_n(\cdot)$ , for  $n$  large, behaves as if there were only level  $I$  and level  $I - 1$  servers, with jumps of size  $1/\sqrt{n}$ . For such a process, it is considerably easier to derive a diffusion limit because it behaves similar to a birth-death process. We formalize our result in the following theorem, the proof of which is in Section 8, where we also make the intuition rigorous.

**Theorem 1.** Suppose  $\hat{Z}_n(0) \Rightarrow \hat{Z}(0)$ . For any policy in  $\Pi$ , the process  $\hat{Z}_n(\cdot)$  converges weakly in  $D^I[0, \infty)$ , as  $n \rightarrow \infty$ , to the process

$$\hat{Z}(\cdot) = (\hat{Q}(\cdot), 0, \dots, 0)^\top,$$

where  $\hat{Q}(\cdot)$  is a one-dimensional diffusion process that has infinitesimal mean

$$m(x) = \begin{cases} -d_I\beta - \theta x, & x \geq 0, \\ -d_I\beta - (d_I - d_{I-1})x, & x < 0, \end{cases}$$

and constant diffusion

$$\sigma^2(x) = 2d_I.$$

**Remark 3.** For the result of Theorem 1 to remain valid, the constant  $c$  for the policies in  $\Pi$  may actually depend on  $n$ ; one can take  $\{c_n\}_n$  to be a sequence so that  $c_n$  is of the order  $1/\sqrt{n}$ . This shows that in many practical situations, sending only a small amount of customers to servers at level  $I - 2$  or below will provide the same asymptotic performance as sending all customers to these servers (if possible); one should really push customers toward level  $I - 2$  servers or below to obtain different asymptotic performance.

### 3.2. Diffusion Limit for the MBF Policy

We now focus on deriving a weak limit approximation for  $Z(\cdot)$  under the MBF policy. In contrast to the results of the LBF case, the limiting process  $\hat{Z}(\cdot)$  does not have nontrivial components  $\hat{n}_i(\cdot)$ ,  $i \in \{0, \dots, I - 2\}$  (i.e., there is no state-space collapse). We now first intuitively explain why difficulties arise under the MBF policy and review some theory about Skorokhod problems. Afterward, we present the main result of the section.

Let us first provide an intuitive argument on why there is no state-space collapse in this case and what the resulting diffusion limit should look like. Because of the MBF policy, arrivals are only assigned to a level  $j$  server, for  $j \in \{0, \dots, I - 2\}$ , when there are no available servers at a level higher than  $j$ . This means that when there are available servers at a level higher than  $j$ , the number of level  $j$  servers is increasing by one with rate  $d_{j+1}n_{j+1}(t)$  and decreasing by one with rate  $d_j n_j(t)$ . In particular, for  $j = I - 2$ ,  $n_j(t)$  increases only because of service completions at level  $I - 1$  servers. Because of the QED scaling, we expect  $n_{I-1}(t)$  to be  $\mathcal{O}(\sqrt{n})$  in steady state. It is conceivable that  $n_{I-2}(t)$  is  $\mathcal{O}(\sqrt{n})$  as well in steady state because  $n_{I-1}(\cdot)$  is positive for lengths of time that are of a constant order, whereas the number of departures at level  $I - 2$  servers is proportional to  $n_{I-2}(t)$ . After scaling by  $n^{-1/2}$ , these dynamics will appear as a drift that depends on the position of  $Z(\cdot)$  through  $n_{I-1}(\cdot)$  and  $n_{I-2}(\cdot)$ . Now, when  $n_{I-1}(\cdot)$  becomes zero, there will be a large drop in the number of level  $I - 2$  servers because of arrivals, being  $\mathcal{O}(n)$ , that are assigned to level  $I - 2$  servers. This phenomenon is like a regulator mapping, preventing  $n_{I-1}(\cdot)$  from becoming negative (i.e., reflecting  $n_{I-1}(\cdot)$  in zero), decreasing the number of level  $I - 2$  servers in the process. This line of reasoning carries over to all level  $j$  servers,  $j \in \{0, \dots, I - 2\}$ , through which we see that these are all  $\mathcal{O}(\sqrt{n})$  and are subject to a position dependent drift, and the term  $n_i(\cdot)$  is reflected in zero to prevent it from becoming negative, for  $i \in \{1, \dots, I - 1\}$ .

As indicated, we can consider  $\hat{Z}_n(\cdot)$  as a reflected process with a position-dependent drift. In order to derive the diffusion approximation, we rely on the theory of (generalized) Skorokhod problems. Let us review some theory on Skorokhod problems, starting with the following definition; for additional details on Skorokhod problems, we refer to Whitt [56].

**Definition 2 (Generalized Skorokhod Problem).** Given an element  $x(\cdot) \in D^d[0, \infty)$  with  $x(0) \geq 0$  for some  $d \in \mathbb{N}$ , a Lipschitz function  $f: \mathbb{R}_+^d \rightarrow \mathbb{R}^d$ , and a  $d \times d$  matrix  $R$ , we call a pair  $(z(\cdot), l(\cdot)) \in D^d[0, \infty) \times D^d[0, \infty)$  a solution to the generalized Skorokhod problem for  $x(\cdot)$ , with respect to  $R$ , if the following conditions are satisfied:

- i.  $z(t) = x(t) + \int_0^t f(z(s)) ds + Rl(t)$ ,  $t \geq 0$ ;
- ii.  $z_i(t) \geq 0$ ,  $i \in \{1, \dots, d\}$ ,  $t \geq 0$ ; and

- iii.  $l(\cdot)$  is such that for  $i \in \{1, \dots, d\}$ ,
  - a.  $l_i(0) = 0$ ,
  - b.  $l_i(\cdot)$  is nondecreasing, and
  - c.  $\int_0^t z_i(t) dl_i(t) = 0$ .

When  $f \equiv 0$ , it is well known that when  $R$  is a generalized  $\mathcal{M}$  matrix (see Definition 3), then there exists a unique pair  $(z(\cdot), l(\cdot))$  that solves the Skorokhod problem for  $x(\cdot)$ , with respect to  $R$ . In that case, we can associate a mapping  $(\phi, \psi) : D^d[0, \infty) \rightarrow D^d[0, \infty) \times D^d[0, \infty)$  to the Skorokhod problem by setting  $(\phi(x(\cdot)), \psi(x(\cdot))) = (z(\cdot), l(\cdot))$ . Moreover, it is known that this mapping is Lipschitz continuous. These results were first established in Harrison and Reiman [23], which is why a generalized  $\mathcal{M}$  matrix is also called a *Harrison–Reiman matrix*.

Now, when  $f$  is nontrivial, the corresponding term represents a drift that is a function of the position of the reflected process. In that case, when  $R$  is a generalized  $\mathcal{M}$  matrix, there still exists a unique solution to the generalized Skorokhod problem so that a mapping  $x(\cdot) \mapsto (z(\cdot), l(\cdot))$  can be associated, which is again Lipschitz continuous. The solution can be written as  $(\phi(M(x(\cdot))), \psi(M(x(\cdot))))$ , where  $M : D^d[0, \infty) \rightarrow D^d[0, \infty)$  is the function that maps  $x(\cdot) \mapsto v(\cdot)$ , with  $v(\cdot)$  solving

$$v(t) = x(t) + \int_0^t f(\phi(v(s))) ds, \quad \text{for all } t \geq 0.$$

This mapping is well defined and is Lipschitz continuous (Reed and Ward [42, lemma 1]). In the sequel, we continue calling this mapping  $M$  so as to emphasize in which part of the proof we will use this technique of including a drift.

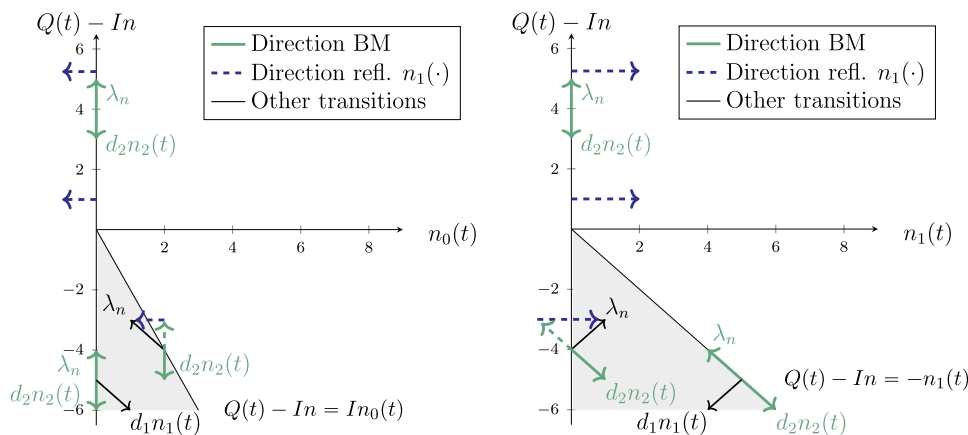
Skorokhod problems are often employed for deriving diffusion limits for nonnegative processes because a weak limit approximation for  $x(\cdot)$  also applies to  $z(\cdot)$  because of the continuous mapping theorem. We employ the same technique but require some modifications as our process  $Z(\cdot)$  does not fit the framework in Definition 2.

There are yet two issues specific to the process we consider. The first is that  $Z(\cdot)$  only requires  $I - 1$  directions of reflection, namely in the components  $n_i(\cdot)$ ,  $i \in \{1, \dots, I - 1\}$ , as was already alluded to in our heuristic argument at the beginning of this section. In the proof, we formalize the evolution of  $Z(\cdot)$  (see (28)), from which it can be readily checked that  $n_I(\cdot)$  and  $n_0(\cdot)$  stay nonnegative when  $n_i(\cdot) \geq 0$ . Because of this fact, we have more components than directions of reflection, which require an (straightforward) adaptation of the theory of (generalized) Skorokhod problems.

The second issue is that in the current state-space representation (including  $n_0(\cdot)$  but not  $n_{I-1}(\cdot)$ ), the matrix  $\tilde{R}$  associated with the reflection is not a generalized  $\mathcal{M}$  matrix. It is, therefore, a priori unclear whether a solution to the generalized Skorokhod problem exists (let alone that it is unique), so we cannot use the continuous mapping theorem. If we modify the state representation by adding  $n_{I-1}(\cdot)$  and dropping  $n_0(\cdot)$ , the associated reflection matrix  $R$  is a generalized  $\mathcal{M}$  matrix. With this representation, however, the direction of the Brownian motion that will drive the diffusion process as  $n \rightarrow \infty$  depends on the position of the diffusion process, making deriving the diffusion approximation a difficult task.

We have illustrated the second issue in Figure 2 for the case  $I = 2$  with the two different state representations  $(n_0(t), Q(t) - In)$  (left panel) and  $(n_1(t), Q(t) - In)$  (right panel); the shaded areas are the interior parts of the state

**Figure 2.** (Color online) Schematic representation of directions of Brownian motion (BM) and reflection (refl.) part for  $(n_0(t), Q(t) - In)$  plane (left panel) and  $(n_1(t), Q(t) - In)$  plane (right panel).



space where the queue is empty. In both panels, the solid arrows correspond to actual transitions. Note that the transitions associated with arrivals (with rate  $\lambda_n$ ) in the interior can be interpreted as the composition of a state-independent transition (green dashed arrows) and a reflection (blue dashed arrows). The green arrows represent jumps corresponding to the Brownian component as  $n \rightarrow \infty$ , whereas the blue arrows represent the direction of reflection of  $n_1(\cdot)$ . Now, for the state representation in the right panel of Figure 2, there is a natural boundary of reflection in the vertical axis, but the direction of the Brownian component depends on the state; in the left panel of Figure 2, the direction of the Brownian component is independent of the state, but there is a reflection in the diagonal. We tackle this second issue by using a custom version of the Skorokhod problem that combines the nice properties from both state-space representations, the details of which we leave until the proof.

Before we state the main theorem of this section, we introduce some additional notation. First, associated with the vector  $(Q(t), n_0(t), n_1(t), \dots, n_{I-2}(t))$ , let  $e_Q$  be the standard basis vector in  $\mathbb{R}^I$ , with its first element equal to one. Analogously, for  $i = 0, \dots, I - 2$ , we let  $e_{n_i}$  be the standard basis vector in  $\mathbb{R}^I$ , with its  $(i + 2)$ nd element equal to one. As such,  $e_Q$  and  $e_{n_i}$  correspond to a unit jump of the components  $Q(\cdot)$  and  $n_i(\cdot)$ ,  $i \in \{0, \dots, I - 2\}$ , respectively.

Second, we define the function  $f : \mathbb{R}^I \rightarrow \mathbb{R}^I$  as

$$f(\hat{Z}_n(t)) = -e_Q \left( \beta d_I - d_I \sum_{i=0}^{I-1} \hat{n}_i(t) + \theta(\hat{Q}_n(t))^+ \right) + \sum_{i=1}^{I-2} (-e_Q - e_{n_i} + e_{n_{i-1}}) d_i \hat{n}_i(t) + (-e_Q + e_{I-2}) d_{I-1} \hat{n}_{I-1}(t). \tag{6}$$

This function captures the nonhomogeneous drift of  $\hat{Z}_n(\cdot)$ . Note that  $f$  is Lipschitz continuous (because it is piecewise linear). Finally, let  $\tilde{R}$  be the  $(I - 1) \times (I - 1)$  matrix with  $-1$  on its diagonal,  $2$  on its first lower diagonal,  $-1$  on its second lower diagonal, and  $0$  elsewhere.

We are now ready to introduce the main result of this section.

**Theorem 2.** Suppose  $\hat{Z}_n(0) \Rightarrow \hat{Z}(0)$ . For the MBF policy,  $\hat{Z}_n(\cdot) \Rightarrow \hat{Z}(\cdot)$  as  $n \rightarrow \infty$ , where  $\hat{Z}(\cdot) = (\hat{Q}(\cdot), \hat{n}_0(\cdot), \dots, \hat{n}_{I-2}(\cdot))^\top$  is the unique process with values in  $\mathbb{R} \times \mathbb{R}_+^{I-1}$  satisfying the stochastic integral equation

$$\hat{Z}(t) = \hat{Z}(0) + e_Q B(t) + \int_0^t f(\hat{Z}(s)) ds + \begin{pmatrix} 0 & 0 \\ 0 & \tilde{R} \end{pmatrix} \hat{L}(\cdot), \tag{7}$$

where  $B(\cdot)$  is a scaled Brownian motion with  $\langle B \rangle_t = 2d_I t$  and the process  $\hat{L}(\cdot) = (\hat{L}_0(\cdot), \dots, \hat{L}_{I-1}(\cdot))^\top$  in  $D_+^I[0, \infty)$  is uniquely determined by  $\hat{L}_0(\cdot) \equiv 0$ ,  $\hat{n}_0(\cdot), \dots, \hat{n}_{I-1}(\cdot) \geq 0$ , and for  $i = 1, \dots, I - 1$ ,

- a.  $\hat{L}_i(0) = 0$ ,
  - b.  $\hat{L}_i(\cdot)$  is nondecreasing, and
  - c.  $\int_0^\infty \hat{n}_i(t) d\hat{L}_i(t) = 0$ ,
- with  $\hat{n}_{I-1}(t) := (-\hat{Q}(t))^+ - \sum_{i=2}^I i \hat{n}_{I-i}(t)$  for all  $t \geq 0$ .

## 4. Heavy-Traffic Diffusion Limits for the Shared Work Case

In Section 3, we considered policies in which customers are not allowed to be reallocated to another server during service; in Legros and Jouini [34], this is referred to as the NWC. We now consider the impact of adding the flexibility to reallocate customers during service, which we refer to as the SWC. For the SWC, we assume that customers may be (optimally) redistributed among the servers at any moment; a routing policy in the SWC should, therefore, be interpreted as the allocation of customers among servers. Observe that such a reallocation requires complete knowledge of the system state, and thus, random policies do not apply. Hence, for the SWC, we exclusively consider the LBF and MBF policies. Also, note that the state of the system is thus completely captured by  $Q(t)$ , which simplifies the analysis considerably.

We consider heavy-traffic diffusion limits for the LBF and MBF policies in the SWC; a better intuitive understanding of the impact of additional SWC flexibility can be found in Section 6.

### 4.1. Diffusion Limit for the LBF Policy

Under the LBF allocation in the SWC, the number of customers at each server is balanced as much as possible. Consequently, servers will be at most at two different levels. Similar to the NWC with the LBF policy, in the regime as  $\rho_n$  gets close to one, the number of available spaces should be  $\mathcal{O}(\sqrt{n})$  as a consequence of the QED scaling. This

implies that in the limit as  $n \rightarrow \infty$ , there are only level  $I - 1$  and level  $I$  servers. In the next theorem, we state the corresponding diffusion limit, which we prove in Appendix A.

**Theorem 3.** Suppose  $\hat{Z}_n(0) \Rightarrow \hat{Z}(0)$ . Then, in the SWC with the LBF policy, the process  $\hat{Z}_n(\cdot)$  converges weakly in  $D^I[0, \infty)$ , as  $n \rightarrow \infty$ , to the process

$$\hat{Z}(\cdot) = (\hat{Q}(\cdot), 0, \dots, 0)^\top,$$

where  $\hat{Q}(\cdot)$  is a diffusion process with distribution as the process  $\hat{Q}(\cdot)$  in Theorem 1.

As may be anticipated because of the intuition that servers are only at two different levels, the limiting process  $\hat{Z}(\cdot)$  is identical to the one found in Theorem 1 for the NWC. Hence, for larger systems in heavy traffic, there is no need for cooperation between servers. The LBF routing provides the same asymptotic performance in the NWC as in the SWC.

### 4.2. Diffusion Limit for the MBF Policy

Under the MBF allocation in the SWC, as many level  $I$  servers as possible are created, and there will be at most one nonempty server that is not at level  $I$ . Consequently, the process  $Q_n(\cdot)$  behaves as a birth-death process with birth rate  $\lambda_n$  and death rate  $d_I \lfloor Q(t)/I \rfloor + d_{g(Q(t))}$ , where  $g(x) = x - \lfloor x/I \rfloor I$ , for  $x \in \mathbb{R}$ . Moreover, this implies that in the considered heavy-traffic regime, there will only be level  $I$  and level 0 servers. This is formalized in the following theorem, which we prove in Appendix A.

**Theorem 4.** Suppose  $Z_n(0) \Rightarrow \hat{Z}(0)$ . Then, in the SWC with the MBF policy, the process  $\hat{Z}_n(\cdot)$  converges weakly in  $D^I[0, \infty)$ , as  $n \rightarrow \infty$ , to the process

$$\hat{Z}(\cdot) = \left( \hat{Q}(\cdot), \frac{(\hat{Q}(\cdot))^-}{I}, \dots, 0 \right)^\top,$$

where  $\hat{Q}(\cdot)$  is a one-dimensional diffusion process that has infinitesimal mean

$$m(x) = \begin{cases} -d_I \beta - \theta x, & x \geq 0, \\ -d_I \beta - \frac{d_I}{I} x, & x < 0, \end{cases}$$

and constant diffusion

$$\sigma^2(x) = 2d_I.$$

## 5. Optimality of Routing Policies LBF and MBF

The LBF and MBF are two intuitively appealing routing policies for arriving customers. This section provides additional theoretical support and is a primary source of motivation for studying LBF and MBF. Specifically, we identify conditions on the functional shape of  $i\mu_i$ , under which the LBF and MBF policies perform optimal (or worst) among all work-conserving policies. Thereby, they also provide bounds for the performance of other policies. The optimality is to be interpreted in the stochastic ordering sense. Hence, the optimality concerns comparisons of routing policies in the prelimit setting (i.e., for a fixed parameter setting with a finite number of servers). This is especially relevant because we have already shown in Section 3.1 that the asymptotic performance of the LBF policy is equal to that of all policies in the class  $\Pi$ . Optimal policies for other functional shapes of  $i \mapsto \mu_i$  are not studied in this paper. In Section 10, we do present a numerical example of a system in which  $i \mapsto \mu_i$  is neither convex nor concave in  $i$ .

Let us first introduce some notation. With  $m_i := i\mu_i$ , we say that  $m_i$  is concave in  $i$  when

$$m_i - m_{i-1} \geq m_{i+1} - m_i, \quad i = 1, \dots, I - 1.$$

Similarly,  $m_i$  is convex in  $i$  if

$$m_i - m_{i-1} \leq m_{i+1} - m_i, \quad i = 1, \dots, I - 1.$$

Note that for  $i = 0, \dots, I$ ,  $m_i$  can be seen as the value  $m(i)$  of a real-valued function  $m : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ . From this point of view, the notions of concavity and convexity align with, respectively, the notions of real-valued concave and convex functions (e.g., the setup considered in Legros and Jouini [34]). Observe that with our definition of  $d_i = i(\mu_i + \theta_s)$ , concavity (convexity) of  $m_i$  in  $i$  is equivalent to concavity (convexity) of  $d_i$  in  $i$ .



We only consider work-conserving policies (i.e., policies that route customers to servers whenever there is a server with fewer than  $I$  customers). Let  $\pi$  be some arbitrary work-conserving policy. We add a superscript  $\pi$  when a process is considered under strategy  $\pi$ . Denote by  $L^\pi(t)$  the total number of abandonments up to time  $t$  under policy  $\pi$ . The following theorem provides that the queue length and the number of abandonments of any work-conserving policy are bounded, in the sense of stochastic ordering, by LBF from below and by MBF from above if  $m_i$  is concave in  $i$ . Here, for two random variables  $X$  and  $Y$ , we write that  $X$  is stochastically larger than  $Y$ , denoted as  $X \geq_{st} Y$ , if  $\mathbb{P}(X > z) \geq \mathbb{P}(Y > z)$  for all  $z$ . Similarly, the process  $X(\cdot)$  is stochastically larger than process  $Y(\cdot)$ , denoted as  $X(\cdot) \geq_{st} Y(\cdot)$ , if  $(X(t_1), \dots, X(t_k)) \geq_{st} (Y(t_1), \dots, Y(t_k))$  for all  $0 \leq t_1 \leq \dots \leq t_k$  for every  $k \geq 1$ .

**Theorem 5.** *If  $m_i = i\mu_i$  is a concave function in  $i$ , then*

$$Q^{LBF}(\cdot) \leq_{st} Q^\pi(\cdot) \leq_{st} Q^{MBF}(\cdot), \tag{8}$$

$$L^{LBF}(\cdot) \leq_{st} L^\pi(\cdot) \leq_{st} L^{MBF}(\cdot) \tag{9}$$

for all work-conserving policies  $\pi$ , provided that the initial states under LBF,  $\pi$ , and MBF are identical.

The proof relies on weak submajorization arguments (see Akgun et al. [1], Sparagis et al. [49]). To do so, we use a different state representation than in the rest of the paper. Specifically, let  $N_i^\pi(t)$  be the number of customers at server  $i = 1, \dots, n$  at time  $t$  under policy  $\pi$ , and define  $N^\pi(t) = (N_1^\pi(t), \dots, N_n^\pi(t))$ . In the proof, we then focus on the process  $(Q^\pi(\cdot), N^\pi(\cdot))$ . This allows us to follow the lines of reasoning as presented in Sparagis et al. [49], which essentially focuses on  $N^\pi(\cdot)$ . Because of the similarities with Sparagis et al. [49], we present the proof in Appendix B.

The next theorem shows that a similar stochastic ordering result in case  $m_i$  is convex in  $i$ , but the LBF and MBF bounds are reversed. The proof is based on weak supermajorization and can also be found in Appendix B.

**Theorem 6.** *If  $m_i = i\mu_i$  is a convex function in  $i$ , then*

$$Q^{LBF}(\cdot) \geq_{st} Q^\pi(\cdot) \geq_{st} Q^{MBF}(\cdot), \tag{10}$$

$$L^{LBF}(\cdot) \geq_{st} L^\pi(\cdot) \geq_{st} L^{MBF}(\cdot) \tag{11}$$

for all work-conserving policies  $\pi$ , provided that the initial states under LBF,  $\pi$ , and MBF are identical.

**Remark 4.** For the SWC of Section 4, it already has been shown in Legros and Jouini [34, proposition 1] that LBF and MBF are optimal for  $m_i$  concave and convex in  $i$ , respectively. In particular, in case  $Q(t) < In$ , the allocation of customers that optimizes the overall service rate is given by the optimization problem  $\max \sum_{i=0}^I m_i n_i(t)$  subject to  $\sum_{i=0}^I i n_i(t) = Q(t)$ .

## 6. Application of Heavy-Traffic Diffusion Limits

The diffusion limits may be used as approximations for multitasking systems with a finite number of servers. Such approximations have become classical for the Erlang C and Erlang A models, where many performance measures involving waiting times and abandonments from the queue can be expressed in terms of the probability of delay (see Garnett et al. [18]). Therefore, we will focus on the probability of delay, as it is a primary building block for the approximation of waiting times and abandonment probabilities. Some other approximate performance measures, such as the sojourn time or the number of customers in the system, may be obtained in a similar fashion as shown in Cui and Tezcan [10, section 8].

Using the stationary distribution of the diffusion limit, the probability of delay in the Erlang C model is approximated by (cf. Halfin and Whitt [20])

$$HW(\beta) = \left( 1 + \frac{\beta\Phi(\beta)}{\phi(\beta)} \right)^{-1}, \tag{12}$$

where  $\Phi$  and  $\phi$  are the cumulative distribution function and probability density function of the standard normal distribution. The model with abandonments gives rise to another diffusion limit (Garnett et al. [18]) with the approximate probability of delay, also known as the Garnett delay function:

$$Garnett(\beta, \theta_{rat}) = \left( 1 + \sqrt{\theta_{rat}} \frac{h(\beta/\sqrt{\theta_{rat}})}{h(-\beta)} \right)^{-1}, \tag{13}$$

where  $\theta_{\text{rat}} = \theta/\mu$  and  $h(x) = \phi(x)/(1 - \Phi(x))$  is the hazard rate of the standard normal distribution; see Green et al. [19] for an appealing illustration of its application in service systems.

For all policies in  $\Pi$  (and LBF in the SWC), the diffusion limit is a generalized Ornstein–Uhlenbeck process. The stationary distribution of such a process allows for an explicit expression (cf. Browne and Whitt [7, equations 23–25]), which has been evaluated in Cui and Tezcan [10, theorem 7]. As this stationary distribution forms the basis for many approximations, such as the probability of delay as given, it is presented in the following proposition in case  $\theta > 0$ ; the case  $\theta = 0$  follows along the same lines.

**Proposition 1.** For  $\theta > 0$ , the diffusion process  $\hat{Q}(\cdot)$ , given in Theorem 1, has stationary probability measure  $\pi$  given by

$$\pi(x) = \begin{cases} \frac{1}{Cd_I} \exp\left\{\frac{\beta^2 d_I}{2\theta}\right\} \exp\left\{-\frac{1}{2d_I} \left(x + \frac{\beta d_I}{\theta}\right)^2\right\}, & x \geq 0, \\ \frac{1}{Cd_I} \exp\left\{\frac{\beta^2}{a}\right\} \exp\left\{-\frac{1}{2}a \left(x + \frac{\beta}{a}\right)^2\right\} & x < 0, \end{cases}$$

where  $0 < a = \left(1 - \frac{d_{I-1}}{d_I}\right)$  and  $C < \infty$  is the normalization constant

$$C = \frac{1}{d_I} \left[ \exp\left\{\frac{\beta^2}{2a}\right\} \sqrt{\frac{2\pi}{a}} \Phi\left(\frac{\beta}{\sqrt{a}}\right) + \exp\left\{\frac{\beta^2 d_I}{2\theta}\right\} \sqrt{\frac{2\pi d_I}{\theta}} \left(1 - \Phi\left(\beta \sqrt{\frac{d_I}{\theta}}\right)\right) \right].$$

Using Proposition 1, we may directly derive an approximation for the probability of delay by calculating  $\int_0^\infty \pi(x) dx$ . Specifically, after some basic calculus, we obtain

$$\mathbb{P}(\text{Delay}) \approx \left(1 + \frac{h\left(\beta/\sqrt{\frac{\theta}{d_I}}\right) \sqrt{\frac{\theta}{d_I}}}{h(-\beta/\sqrt{a}) \sqrt{a}}\right)^{-1} = \text{Garnett}\left(\frac{\beta}{\sqrt{a}}, \frac{\theta}{ad_I}\right).$$

For  $\theta = 0$ , the stationary distribution gives rise to the approximation

$$\mathbb{P}(\text{Delay}) \approx \text{HW}(\beta/\sqrt{a}). \quad (14)$$

These approximations allow for an intuitive interpretation and reveal how the system behaves in the QED regime. Let us first focus on the square-root safety staffing principle in the case of absence of abandonments. As argued before, there are only level  $I$  and level  $I - 1$  servers in the limiting regime. As all servers have (roughly) at least  $I - 1$  customers, these may be considered permanent customers. The variability is in terms of the number of level  $I$  servers as opposed to level  $I - 1$  servers, with  $d_I - d_{I-1} =: ad_I$  being the variable part of the service rate. Now, in line with the square-root safety staffing principle, we need at least  $R = \lambda/d_I$  servers to handle the total offered load. Excess capacity is needed to protect against stochastic variability (see, e.g., Borst et al. [6]). In this case, the variable part of the service rate can control the stochastic variability. Thus, the staffing level according to the square-root safety staffing principle is

$$n \approx R + \beta \sqrt{\frac{\lambda}{d_I - d_{I-1}}} = R + \beta \sqrt{\frac{\lambda}{ad_I}} = R + \frac{\beta}{\sqrt{a}} \sqrt{R},$$

explaining the factor  $\beta/\sqrt{a}$  in the Halfin–Whitt delay function in (14). With abandonments, there is a second term in the Garnett delay function, being the ratio  $\theta_{\text{rat}}$  between the abandonment rate and the service rate. As the service rate now corresponds to the difference between service at level  $I$  and service at level  $I - 1$ , we have  $\theta_{\text{rat}} = \theta/(d_I - d_{I-1}) = \theta/(ad_I)$ .

The MBF only allows for explicit expressions in the SWC. In that situation, the limiting process  $\hat{Q}(\cdot)$  is again a piecewise Ornstein–Uhlenbeck process, and therefore, an explicit expression for the stationary measure can be found, like in Proposition 1.

**Proposition 2.** For  $\theta > 0$ , the diffusion process  $\hat{Q}(\cdot)$ , given in Theorem 4, has stationary probability measure  $\pi$  given by

$$\pi(x) = \begin{cases} \frac{1}{Cd_I} \exp\left\{\frac{\beta^2 d_I}{2\theta}\right\} \exp\left\{-\frac{1}{2d_I} \left(x + \frac{\beta d_I}{\theta}\right)^2\right\}, & x \geq 0, \\ \frac{1}{Cd_I} \exp\{\beta^2 I\} \exp\left\{-\frac{11}{2I}(x + \beta I)^2\right\} & x < 0, \end{cases}$$

where  $C < \infty$  is the normalization constant

$$C = \frac{1}{d_I} \left[ \exp \left\{ \frac{1}{2} \beta^2 I \right\} \sqrt{2\pi I} \Phi(\beta\sqrt{I}) + \exp \left\{ \frac{\beta^2 d_I}{2\theta} \right\} \sqrt{\frac{2\pi d_I}{\theta}} \left( 1 - \Phi \left( \beta \sqrt{\frac{d_I}{\theta}} \right) \right) \right].$$

Again, we may relate the approximate probability of delay to the quality of service parameter  $\beta$  and the relative patience  $\theta_{\text{rat}}$ :

$$\mathbb{P}(\text{Delay}) = \left[ 1 + \frac{h \left( \beta / \sqrt{\frac{\theta I}{d_I}} \right)}{h(-\beta\sqrt{I})} \sqrt{\frac{\theta I}{d_I}} \right]^{-1} = \text{Garnett} \left( \beta\sqrt{I}, \frac{\theta}{d_I/I} \right).$$

Similar results may be derived for the case  $\theta = 0$ , yielding  $\mathbb{P}(\text{Delay}) \approx HW(\beta\sqrt{I})$ . In the absence of abandonments, the dynamics of the limiting diffusion process  $\hat{Q}(\cdot)$  are the same as for the Erlang delay model (cf. Halfin and Whitt [20, theorem 2]) but with the service rate scaled by a factor  $1/I$  (i.e., the limiting process  $\hat{Q}(\cdot)$  can be considered as the limiting process of an Erlang delay model with service rate  $d_I/I$  under the QED scaling). This may be explained by the fact that in the prelimit process, there is at most one server that is not at level 0 or  $I$ , which will not contribute in the limit. Therefore, the limiting process can be considered as if each customer is at a level  $I$  server, adding a marginal rate of  $d_I/I$  to the “total death rate.” Again,  $R = \lambda/d_I$  servers are required to handle the total offered load. As  $d_I/I$  is the marginal service rate per customer, the staffing level according to the square-root safety staffing principle is

$$n \approx R + \beta \sqrt{\frac{\lambda}{d_I/I}} = R + \beta\sqrt{I}\sqrt{R},$$

providing the factor  $\beta\sqrt{I}$  in the Halfin–Whitt delay function. For the case  $\theta > 0$ , the marginal service rate per customer of  $d_I/I$  directly provides  $\theta_{\text{rat}} = \frac{\theta}{d_I/I}$ , which appears as the second argument in the Garnett delay function.

## 7. Diffusion Limits in the Subcritical Regime

In this section, we discuss diffusion limits for the multitasking service system where the load is asymptotically strictly below one (i.e., nonheavy-traffic scaling regimes). These regimes are motivated by applications in which it is desired that servers operate at a fraction of their capacity (cf. Remark 2). We focus on the LBF and MBF policies for these subcritical regimes, of which the diffusion behavior for the former was studied in Cui and Tezcan [10], and we provide a theorem for the diffusion behavior of the latter. In addition, we show why the LBF diffusion behavior is no longer attained by a broad class of policies, such as in Theorem 1.

### 7.1. Asymptotic Framework

Our exposition starts with the introduction of the subcritical asymptotic scaling framework, for which we use the same notation as for the heavy-traffic scaling in Section 2.3. The new interpretation of this notation is limited to the current section and Appendix C. For  $\gamma \in (0, 1)$  and  $j \in \{0, \dots, I-1\}$ , define  $\lambda > 0$  by

$$\lambda := \gamma d_j + (1 - \gamma) d_{j+1},$$

and for  $n \geq 1$  and  $\beta \in \mathbb{R}$ , define  $\lambda_n$  as

$$\lambda_n = \lambda(n - \beta\sqrt{n}). \tag{15}$$

Throughout this section, we consider a sequence of processes  $\{Z_n(\cdot)\}_n$ , where the  $n$ th process corresponds to a multitasking system with  $n$  servers and arrival rate  $\lambda_n$  as in (15). Observe that for every choice of  $\gamma$ ,  $\lambda_n/(d_I n)$  converges to a number strictly below one as  $n \rightarrow \infty$ , which reflects the subcriticality of the system load under this scaling.

Under this subcritical scaling, the centering of the process has to be adjusted from the heavy-traffic case because we do not expect every server to be filled to capacity on a fluid scale in the limit as  $n \rightarrow \infty$ . In particular, the centering vector depends on the routing policy of assigning customers to servers. We introduce the notation  $\bar{Z}$  for the centering vector and consider the scaled and centered process  $\hat{Z}_n(\cdot)$  defined by

$$\hat{Z}_n(\cdot) := \frac{1}{\sqrt{n}}(Z_n(\cdot) - n\bar{Z}),$$

where  $\bar{Z} = (\bar{Q}, \bar{n}_0, \dots, \bar{n}_{I-2})$  and its elements  $\bar{Q}$  and  $\bar{n}_0, \dots, \bar{n}_{I-2}$  are defined differently for the LBF and MBF policies.

In the LBF case, we set

$$\bar{Q} = \gamma j + (1 - \gamma)(j + 1) \text{ and } \bar{n}_i = \gamma \delta_{ij} + (1 - \gamma) \delta_{i,j+1}, \quad i = 0, 1, \dots, I - 2,$$

where  $\delta_{ij} := \mathbb{1}_{\{i=j\}}$  is the Kronecker  $\delta$ . In the case of the MBF policy, we set

$$\bar{Q} = I \frac{\lambda}{d_I}, \quad \bar{n}_0 = 1 - \frac{\lambda}{d_I}, \text{ and } \bar{n}_i = 0, \quad i = 1, \dots, I - 2. \quad (16)$$

Note that  $\bar{Q}$  reflects that the load is balanced across servers in the LBF case, whereas for MBF, a fraction of servers is always busy, and a fraction of servers is always free. The components of  $\hat{Z}_n(\cdot)$  are still denoted by  $\hat{Q}_n(\cdot)$  and  $\hat{n}_{n,i}(\cdot)$  for  $i \in \{0, \dots, I - 2\}$ . Moreover, defining  $\bar{n}_{I-1} := \gamma \delta_{I-1,j} + (1 - \gamma) \delta_{I-1,j+1}$  and  $\bar{n}_I := (1 - \gamma) \delta_{I,j+1}$  in the LBF case and defining  $\bar{n}_{I-1} := 0$  and  $\bar{n}_I = \lambda/d_I$  in the MBF case, then it follows from (1) that the following identities hold in both the LBF and MBF cases:

$$\begin{aligned} \hat{n}_{I-1}(\cdot) &:= n^{-1/2}(n_{I-1}(\cdot) - n\bar{n}_{I-1}) = -\hat{Q}_n(\cdot) - \sum_{i=0}^{I-2} (I-i)\hat{n}_i(\cdot), \\ \hat{n}_I(\cdot) &:= n^{-1/2}(n_I(\cdot) - n\bar{n}_I) = -\sum_{i=0}^{I-1} \hat{n}_i(\cdot). \end{aligned} \quad (17)$$

To establish the diffusion limits for the LBF and MBF policies, we assume that

$$n^{-1}Z_n(0) \rightarrow \bar{Z} \text{ a.s. as } n \rightarrow \infty.$$

Moreover, we assume in the LBF case that  $\hat{Z}_n(0)$  converges weakly in  $\mathbb{R}^I$  to a random vector

$$\hat{Z}(0) = (j\hat{n}_j(0) + (j+1)\hat{n}_{j+1}(0), \hat{n}_0(0), \dots, \hat{n}_{I-2}(0))^\top,$$

as  $n \rightarrow \infty$ , with  $\hat{n}_i(0) = 0$  for  $i \notin \{j, j+1\}$ ,  $\hat{n}_{j+1}(0) = -\hat{n}_j(0)$ , and both  $\hat{n}_{I-1}(0)$  and  $\hat{n}_I(0)$  specified through (17). In the MBF case, we assume that  $\hat{Z}_n(0)$  converges weakly in  $\mathbb{R}^I$  to a random vector

$$\hat{Z}(0) = \left( \sum_{j=1}^I j\hat{n}_j(0), \hat{n}_0(0), \dots, \hat{n}_{I-2}(0) \right)^\top$$

as  $n \rightarrow \infty$ , with  $\hat{n}_i(0) \geq 0$  for  $i = 1, \dots, I - 2$ ,  $\hat{n}_{I-1}(0) \geq 0$  a random variable, and  $\hat{n}_I(0)$  specified through (17).

## 7.2. Diffusion Limit for LBF-Related Policies

Under the LBF policy and the assumptions on convergence of the initial conditions, the sequence  $\{\hat{Z}_n(\cdot)\}_n$  converges in distribution to a diffusion process. This result was proven in Cui and Tezcan [10, theorem 3]. We state it here for completeness and for our exposition on why there is no (strong) diffusion-level universality from a policy perspective as in Theorem 1.

**Theorem 7.** Suppose  $n^{-1}Z_n(0) \xrightarrow{\text{a.s.}} \bar{Z}$ , and suppose that  $\hat{Z}_n(0)$  converges weakly to  $\hat{Z}(0)$  in  $\mathbb{R}^I$ . Then, for the LBF policy,  $\hat{Z}_n(\cdot) \Rightarrow \hat{Z}(\cdot)$ , where  $\hat{Z}(\cdot) = (\hat{Q}(\cdot), \hat{n}_0(\cdot), \dots, \hat{n}_{I-2}(\cdot))^\top$  has initial condition  $\hat{Z}(0)$  and satisfies

$$\begin{aligned} \hat{n}_i(\cdot) &= 0 \text{ for } i \notin \{j, j+1\} \\ \hat{n}_{j+1}(\cdot) &= -\hat{n}_j(\cdot) \\ \hat{Q}(\cdot) &= j\hat{n}_j(\cdot) + (j+1)\hat{n}_{j+1}(\cdot), \end{aligned}$$

and  $\hat{n}_j(\cdot)$  is a diffusion process with infinitesimal mean

$$m(x) = \lambda\beta - (d_{j+1} - d_j)x, \quad x \in \mathbb{R}$$

and constant infinitesimal variance  $\sigma^2(x) = 2\lambda$ .

Theorem 7 tells us that under the LBF policy, a state-space collapse will occur in which only level  $j$  and level  $(j+1)$  servers are present in the system on both a fluid scale and a diffusion scale. The manner in which the initial conditions



converge makes this possible (e.g., it is shown in Cui and Tezcan [10, theorem 2] that  $\bar{Z}$  is the unique fixed point of the system on a fluid scale).

By deviating from heavy-traffic scaling regimes, we no longer hope to find a rich class  $\Pi$  of policies, such as in Theorem 1, with the same diffusion-level performance as the LBF policy. In the remainder of this section, we provide the arguments for this claim, where for conciseness, we choose a slightly informal presentation.

For a policy to have the same diffusion-level performance as the LBF policy, the number of servers above level  $(j + 1)$  and below level  $j$  should be  $o(\sqrt{n})$ , with probability increasing to one as  $n \rightarrow \infty$ . Focusing on the analysis of servers above level  $(j + 1)$  for now, consider the process

$$k_n^\circ(\cdot) := \sum_{i=j+2}^I n_{n,i}(\cdot)(i - j - 1),$$

which increases when a customer is routed to a server at level  $(j + 1)$  or above and decreases when a server at level  $(j + 2)$  or above finishes service. Here, we use that there will be no queue for subcritical scaling regimes with probability 1 in the limit as  $n \rightarrow \infty$ . If we suppose that a given policy for the  $n$ th system assigns customers to a server at level  $(j + 1)$  and above with probability  $p_n \in [0, 1]$ , then  $k_n^\circ(\cdot)$  increases by one with rate  $\lambda_n p_n$  and decreases with one with rate  $\sum_{i=j+2}^I d_i n_{n,i}(t)$ , which has a lower bound:

$$\sum_{i=j+2}^I d_i n_{n,i}(t) \geq d_{j+2} \sum_{i=j+2}^I n_{n,i}(t) > \frac{d_{j+2}}{I - (j + 1)} k_n^\circ(t).$$

The lower bound establishes that using a similar coupling argument as in the proof of Lemma 1, we can upper bound  $k_n^\circ(\cdot)$  by the occupation process of an infinite-server queue with arrival rate  $\lambda_n p_n$  and service rate  $d_{j+2}/(I - (j + 1))$ . Moreover, the same reasoning as in the aforementioned proof indicates that  $n^{-1/2} k_n^\circ(\cdot) \xrightarrow{\mathbb{P}} 0$  if  $\sqrt{n} p_n = o(1)$ . We expect this bound to be tight because the load (and thus, also the expected occupation) of the associated infinite-server system is  $O(\lambda_n p_n)$ , which is below  $O(\sqrt{n})$  only if  $\sqrt{n} p_n = o(1)$ .

The condition  $\sqrt{n} p_n = o(1)$  is sufficient in the subcritical scaling regime for a policy to achieve that no servers at level  $(j + 2)$  and above are present on a diffusion scale. We also expect that the condition is necessary because the lower bound for the departure rate at  $k_n^\circ(\cdot)$  is tight with respect to the order of  $n_{n,i}$ ,  $i = j + 2, \dots, I$ . In the case  $j + 1 = I - 1$ , it can be shown rigorously that the condition is necessary; the fact that there is only one level above  $(I - 1)$  simplifies the analysis compared with  $j + 1 \leq I - 2$ .

The arguments show that, in general, any class of policies resulting in the desired diffusion-level behavior of Theorem 7 would be significantly smaller compared with the class  $\Pi$  in Theorem 1. Specifically, randomized policies, such as random server or power of  $d$  (with  $d$  “small”), will no longer lead to a state-space collapse.

Given that a routing policy satisfies  $\sqrt{n} p_n = o(1)$ , one may wonder if routing the remaining arrivals randomly to servers below level  $(j + 1)$  still results in the diffusion-level state-space collapse of Theorem 7. To analyze this case, consider a policy satisfying  $\sqrt{n} p_n = o(1)$  and with  $q_n$  the probability of routing an arriving customer to a server at level  $(j - 1)$  or below. In the desired state-space collapse, we have  $n_j(t) \approx \gamma n + \epsilon$ , where  $\epsilon = O(\sqrt{n})$  with probability increasing to one as  $n$  grows large. To compensate for the rate  $d_j n_j$  at which servers at level  $(j - 1)$  increase,  $q_n$  has to be at least  $d_j \gamma n$ . Indeed, to obtain the desired state-space collapse, one expects that  $q_n \lambda / n$  should converge to a number strictly above  $\gamma$  because otherwise, there will be nontrivial diffusion- or fluid-level behavior at level  $(j - 1)$ . Hence, any class of policies leading to the desired state-space collapse will be highly similar to the LBF policy and will not be random at all.

To summarize, the heavy-traffic nature of the scaling in Theorem 1 implies that a wide range of routing policies have the same fluid and diffusion-level performance as the LBF policy. When turning away from heavy-traffic regimes, it turns out that there is a large difference between the LBF policy and other (conventional) policies in terms of their diffusion-level behavior.

### 7.3. Diffusion Limit for the MBF Policy

In this section, we provide the diffusion limit for the MBF policy under the subcritical scaling regime defined in (15). To prove the result, we also derive a fluid limit approximation for  $Z_n(\cdot)$  under the MBF policy, which we show to have a unique fixed point. We first state the diffusion limit result and thereafter, formulate the related fluid limit results. The proofs of all results in this section are given in Appendix C.

Recall that we denote  $e_Q$  for the first basis vector in  $\mathbb{R}^I$  and  $e_{n_i}$  for the subsequent basis vectors,  $i = 0, \dots, I - 2$ . Furthermore, define the function  $f^\circ : \mathbb{R}^I \rightarrow \mathbb{R}$  by

$$f^\circ(\hat{Z}_n(t)) = e_Q[-\beta\lambda - d_I \hat{n}_I(s)] + \sum_{i=1}^{I-2} (-e_Q - e_{n_i} + e_{n_{i+1}}) d_i \hat{n}_i(t) + (-e_Q + e_{I-2}) d_{I-1} \hat{n}_{I-1}(t).$$

Observe that  $f^\circ$  is Lipschitz continuous. The main result of this section is the following theorem.

**Theorem 8.** Suppose that  $n^{-1}Z_n(0) \xrightarrow{\text{a.s.}} \bar{Z}$ , and suppose that  $\hat{Z}_n(0)$  converges weakly to  $\hat{Z}(0)$  in  $\mathbb{R}^I$ . Then, for the MBF policy,  $\hat{Z}_n(\cdot) \Rightarrow \hat{Z}(\cdot)$  as  $n \rightarrow \infty$ , where  $\hat{Z}(\cdot) = (\hat{Q}(\cdot), \hat{n}_0(\cdot), \dots, \hat{n}_{I-2}(\cdot))^\top$  is the unique process with values in  $\mathbb{R} \times \mathbb{R}_+^{I-1}$  satisfying the stochastic integral equation

$$\hat{Z}(t) = \hat{Z}(0) + e_Q B(t) + \int_0^t f^\circ(\hat{Z}(s)) ds + \begin{pmatrix} 0 & 0 \\ 0 & \bar{R} \end{pmatrix} \hat{L}(t),$$

where  $B(\cdot)$  is a scaled Brownian motion with  $\langle B \rangle_t = 2\lambda t$  and  $\hat{L}(\cdot) = (\hat{L}_0(\cdot), \dots, \hat{L}_{I-1}(\cdot))^\top \in D_+^I[0, \infty)$  is uniquely determined by  $\hat{L}_0(\cdot) \equiv 0$  and the properties  $\hat{n}_0(\cdot), \dots, \hat{n}_{I-1}(\cdot) \geq 0$  and  $\hat{L}(\cdot)$  is such that for  $i = 1, \dots, I - 1$ ,

- $\hat{L}_i(0) = 0$ ,
  - $\hat{L}_i(\cdot)$  is nondecreasing, and
  - $\int_0^\infty \hat{n}_i(t) d\hat{L}_i(t) = 0$ ,
- with  $\hat{n}_{I-1}(\cdot) := (-\hat{Q}(\cdot))^+ - \sum_{i=2}^I in_{I-i}(\cdot)$  for all  $t \geq 0$ .

Theorem 8 shows that diffusion-level behavior different from that in Theorem 7 is obtained when customers are always routed to the most busy server. As a consequence of the MBF policy, every component of the diffusion process is nontrivial, in contrast to the LBF case. Moreover, from (16), we see that there are free servers on a fluid scale in this regime; this is further formalized in Proposition 4. This property is preferred in systems that operate in heavy traffic; more precisely, the property suggests that the system can achieve the same level of efficiency with fewer servers. In contrast, in systems in which servers preferable should operate at a fraction of their capacity, this property is undesirable, making MBF less attractive as a policy in that case. The subcritical nature of the scaling finally implies that the queue will always be empty on a diffusion scale, in contrast to Theorem 2.

To prove Theorem 8, we study the fluid-level behavior of the process  $Z_n(\cdot)$  under the MBF policy by considering the sequence  $\{n^{-1}Z_n(\cdot)\}_n$ . We show in Proposition 3 that this sequence is a.s. tight given that the initial conditions converge almost surely and identify a system of equations that each limit (of a subsubsequence) should satisfy. Afterward, we show in Proposition 4 that every fluid limit has the same unique fixed point and that under the assumptions on the sequence of initial states, there are exclusively level  $I$  and empty servers on a fluid scale as  $n \rightarrow \infty$ .

**Proposition 3.** Suppose that the sequence  $\{n^{-1}Z_n(0)\}_n$  converges almost surely to some random vector  $\bar{Z}(0) = (\bar{Q}(0), \bar{n}_0(0), \dots, \bar{n}_{I-2}(0))$ . Then, there exists a set  $\Omega' \subset \Omega$  with  $\mathbb{P}(\Omega') = 1$  such that every subsequence of  $\{n^{-1}Z_n(\cdot)\}_n$  has a further subsequence  $\{n_k^{-1}Z_{n_k}(\cdot)\}_k$  that converges everywhere on  $\Omega'$ . Moreover, every limit  $\bar{Z}(\cdot)$  of such a convergent subsequence with  $\bar{Z}(\cdot) = (\bar{Q}(\cdot), \bar{n}(\cdot), \dots, \bar{n}_{I-2}(\cdot))$  is Lipschitz continuous and satisfies for all  $t \geq 0$

$$\bar{Q}(t) = \bar{Q}(0) + \lambda t - \sum_{i=1}^I \int_0^t d_i \bar{n}_i(s) ds - \theta \int_0^t (\bar{Q}(s) - I)^+ ds, \quad (18)$$

$$\bar{n}_0(t) = \bar{n}_0(0) + \int_0^t d_1 \bar{n}_1(s) ds - \bar{L}_1(t), \quad (19)$$

$$\begin{aligned} \bar{n}_i(t) &= \bar{n}_i(0) + \int_0^t d_{i+1} \bar{n}_{i+1}(s) ds - \int_0^t d_i \bar{n}_i(s) ds \\ &\quad - \bar{L}_{i+1}(t) + 2\bar{L}_i(t) - \bar{L}_{i-1}(t) \mathbf{1}_{\{i \geq 2\}} \quad i = 1, \dots, I - 2, \end{aligned} \quad (20)$$

$$\bar{n}_{I-1}(t) = (\bar{Q}(t) - I)^+ - \sum_{i=0}^{I-2} (I - i) \bar{n}_i(t), \quad (21)$$

$$\sum_{i=0}^I \bar{n}_i(t) = 1, \quad (22)$$

$$\bar{Q}(t), \bar{n}_0(t), \dots, \bar{n}_I(t) \geq 0, \quad (23)$$

where  $\bar{L}(\cdot)$  is the Lipschitz-continuous,  $I$ -dimensional process satisfying

$$\bar{L}_i(t) = \int_0^t \lambda \mathbb{1}_{\{\sum_{k=i}^{I-1} \bar{n}_k(s)=0, \bar{Q}(s) < I\}} ds \quad i = 1, \dots, I-1, \quad (24)$$

$$\bar{L}_0(\cdot) \equiv 0. \quad (25)$$

**Proposition 4.** *The system of Equations (18)–(25) has a unique fixed point  $\bar{Z} = (\bar{Q}, \bar{n}_0, \dots, \bar{n}_{I-2})^\top$  given by (16) and with  $\bar{n}_{I-1} = 0$  and  $\bar{n}_I = \lambda/d_I$  as defined in the subsequent discussion. Consequently, if  $\{n^{-1}Z_n(0)\}_n$  converges almost surely to  $\bar{Z}$  as defined in (16), then*

$$n^{-1}Z_n(\cdot) \rightarrow \bar{Z} \quad \text{a.s.}$$

in  $D^I[0, \infty)$  uniformly on compact time intervals.

## 8. Proof of Diffusion Limit for Policies in $\Pi$ in the NWC

In this section, we prove Theorem 1. The proof consists of three steps.

1. We prove in Lemma 1 that for  $i \in \{0, \dots, I-2\}$ ,  $\hat{n}_{n,i}(\cdot) \xrightarrow{\mathbb{P}} 0$ , formalizing that, for large  $n$ ,  $\hat{Z}_n(\cdot)$  behaves like a multitasking system that exclusively has level  $I-1$  and level  $I$  servers, making jumps of size  $1/\sqrt{n}$ .
2. We show using Lemma 1 that for large  $n$ ,  $Q_n(\cdot)$  behaves like a birth-death process  $R_n(\cdot)$  with birth rate  $\lambda_n$  and death rate  $d_I n - (d_I - d_{I-1})(Q(t) - In)^- + \theta(Q(t) - In)^+$ . We prove this by coupling  $Q_n(\cdot)$  to  $R_n(\cdot)$  and show that for all  $T \geq 0$ ,  $\|R_n(\cdot) - Q_n(\cdot)\|_T \xrightarrow{\mathbb{P}} 0$ , as  $n \rightarrow \infty$ .
3. Finally, we prove that the process  $n^{-1/2}(R_n(\cdot) - In)$  converges weakly to the diffusion process  $\hat{Q}(\cdot)$  mentioned in Theorem 1. Because  $n_i(\cdot)$  and  $n_{i-1}(\cdot)$  are continuous transformations of  $Q(\cdot)$ , we obtain weak convergence of  $\hat{Z}_n(\cdot)$ .

Before we present the proof, we want to make the following remark.

**Remark 5.** In the proofs of the diffusion limits for  $\hat{Z}_n(\cdot)$  (not restricted to this section), we often utilize coupling techniques between one-dimensional stochastic processes that exclusively make jumps upward and downward, which are all of size 1. Each time we apply such a technique, the coupling concerns two of such processes, say  $X(\cdot)$  and  $Y(\cdot)$ , with  $X(\cdot)$  having rate at which jumps are made upward (downward) larger than or equal to that of  $Y(\cdot)$ . Now, the key insight is that the rate of  $Y(\cdot)$  can be obtained by randomly splitting the rate of  $X(\cdot)$ , and thus, we can couple the jumps of both processes such that  $X(\cdot)$  is greater (less) than or equal to  $Y(\cdot)$ . Or more precisely, we have  $X(t, \omega) \geq Y(t, \omega)$  ( $X(t, \omega) \leq Y(t, \omega)$ ) for all  $t \geq 0$ , for each  $\omega \in \Omega$ .

Let us now introduce and prove the following proposition.

**Lemma 1.** *Suppose  $\hat{Z}_n(0) \Rightarrow \hat{Z}(0)$ . For the process  $\hat{Z}_n(\cdot)$ , in the NWC with any policy in  $\Pi$ , we have as  $n \rightarrow \infty$ ,*

$$\hat{n}_{n,i}(\cdot) \xrightarrow{\mathbb{P}} 0, \quad \text{for all } i \in \{0, \dots, I-2\}.$$

**Proof.** Let  $\hat{k}_n(\cdot) = n^{-1/2}k_n(\cdot)$ , with  $k_n(\cdot)$  as in (4) associated with the  $n$ th process  $Z_n(\cdot)$ . To prove the statement, we prove the equivalent statement  $\hat{k}_n(\cdot) \xrightarrow{\mathbb{P}} 0$ , when  $n \rightarrow \infty$ .

For  $n$  fixed,  $k_n(\cdot)$  is equal to the number of free spaces minus the number of free servers. It is not hard to see that  $k_n(\cdot)$  increases by one with rate  $\sum_{i=0}^{I-1} d_i n_{n,i}(t) \leq d_{I-1}(Q_n(t) - In)^-$ . The rate down depends on the policy in  $\Pi$  and is upper bounded by  $\lambda_n \mathbb{1}_{\{k_n(t) > 0\}}$ . We proceed by defining simpler processes,  $Q'_n(\cdot)$  and  $k'_n(\cdot)$ , to bound  $Q_n(\cdot)$  and  $k_n(\cdot)$ .

First, define the process  $Q'_n(\cdot)$  with state space as  $Q_n(\cdot)$ , and  $Q'_n(0) = Q_n(0)$ , where  $Q'_n(\cdot)$  goes up by one with rate  $\lambda_n$  and goes down by one with rate  $d_I n + \theta(Q'_n(t) - In)^+$ . We couple the processes  $Q'_n(\cdot)$  and  $Q_n(\cdot)$  so that the jumps that increase the processes exactly match (i.e., the processes jump up at the same times). Moreover, if  $Q'_n(t) = Q_n(t)$ , then a jump downward of  $Q_n(\cdot)$  implies a jump downward of  $Q'_n(\cdot)$  at the same time. This coupling is possible because the rate at which  $Q'_n(\cdot)$  decreases is larger than the rate at which  $Q_n(\cdot)$  decreases. Moreover, it guarantees that  $Q'_n(t) \leq Q_n(t)$ , for all  $t \geq 0$ .

Now, to define  $k'_n(\cdot)$ , note that the rate at which  $k_n(\cdot)$  increases is bounded from above by  $d_{I-1}(Q_n(t) - In)^- \leq d_{I-1}(Q'_n(t) - In)^- \leq d_{I-1} \sup_{0 \leq s \leq t} (Q'_n(s) - In)^-$ . We know that for every policy in  $\Pi$ , the rate at which  $k_n(\cdot)$

decreases is bounded from below by

$$c\lambda_n \frac{k_n(t)}{\sup_{0 \leq s \leq t} (Q'_n(s) - In)^-},$$

for some constant  $c > 0$ . Now, define the process  $k'_n(\cdot)$  with state space as  $k_n(\cdot)$ , where  $k'_n(\cdot)$  increases by one with rate  $d_{I-1}M_n$  and decreases by one with rate  $c\lambda_n k'_n(t)/M_n$  for a sequence of positive constants  $\{M_n\}_n$  defined by  $M_n = n^{5/8}$ . In addition, we couple the process  $k'_n(\cdot)$  to  $k_n(\cdot)$  so that if  $k'_n(t) = k_n(t)$ , then  $k_n(\cdot)$  increases by one at time  $t$  only if  $k'_n(\cdot)$  increases by one at time  $t$ , and  $k'_n(\cdot)$  decreases by one at time  $t$  only if  $k_n(\cdot)$  decreases by one at time  $t$ , for all  $t \geq 0$ . Denote  $\hat{k}'_n(\cdot)$  for  $n^{-1/2}k'_n(\cdot)$ .

Now, let  $T > 0$  be arbitrary. On the event  $\{\sup_{0 \leq s \leq T} (Q'_n(s) - In)^- \leq M_n\}$ , the rates for jumps upward and downward of  $k_n(\cdot)$  are, respectively, upper and lower bounded by  $d_{I-1}M_n$  and  $c\lambda_n k_n(t)/M_n$ . Hence, on this event and because of the coupling between  $k_n(\cdot)$  and  $k'_n(\cdot)$ , we have that  $k_n(t) \leq k'_n(t)$ , for all  $t \in [0, T]$ . We proceed by proving that for each  $\epsilon > 0$ ,

$$\mathbb{P}\left(\sup_{0 \leq s \leq T} \hat{k}'_n(s) > \epsilon\right) \rightarrow 0, \quad n \rightarrow \infty.$$

Afterward, we prove that

$$\mathbb{P}\left(\sup_{0 \leq s \leq T} (Q'_n(s) - In)^- > M_n\right) \rightarrow 0, \quad n \rightarrow \infty. \quad (26)$$

Combining these statements with the arbitrary choice for  $T$  gives us that  $\hat{k}'_n(\cdot) \xrightarrow{\mathbb{P}} 0$ . By the coupling and conditioning on the event  $\{\sup_{0 \leq s \leq T} (Q'_n(s) - In)^- \leq M_n\}$ , we obtain  $\hat{k}_n(\cdot) \xrightarrow{\mathbb{P}} 0$ , completing the proof.

We start with the first statement, where we throughout work with processes on the interval  $[0, T]$ . Note that  $k'_n(\cdot)$  behaves as an M/M/ $\infty$  queue with offered load  $\rho_n^{k'} = d_{I-1}M_n^2/(c\lambda_n)$ , which is of the order  $n^{1/4}$ . Consider the process  $(k'_n(\cdot) - \rho_n^{k'})^2$ . By using Dynkin's formula (cf. Kallenberg [29, lemma 19.21]), we obtain the  $\mathbb{F}_n$ -martingale  $\mathbb{M}_n(\cdot)$ , defined by

$$\mathbb{M}_n(t) = (k'_n(t) - \rho_n^{k'})^2 - \int_0^t \frac{c\lambda_n k'_n(s)(1 - 2(k'_n(s) - \rho_n^{k'})) + d_{I-1}M_n(1 + 2(k'_n(s) - \rho_n^{k'}))}{M_n} ds,$$

where  $\mathbb{F}_n := \{\mathcal{F}_t^n\}_{t \geq 0}$  is the natural filtration. Also, for a constant  $K_n$  depending on  $n$ , define the stopping time  $\tau^{K_n} := \inf\{t \in [0, T] : k'_n(t) \geq K_n + \rho_n^{k'}\}$ . Because we are working on  $[0, T]$ , the stopping time is trivially bounded, and as such, the stopped process  $\mathbb{M}_n^{\tau^{K_n}}(\cdot)$ , defined through  $\mathbb{M}_n^{\tau^{K_n}}(t) := \mathbb{M}_n(t \wedge \tau^{K_n})$ , is an  $\mathbb{F}_n$ -martingale.

Now, by the definition of the stopping time, we have the following inequality:

$$\begin{aligned} K_n^2 \mathbb{P}\left(\sup_{0 \leq s \leq t} k'_n(s) \geq K_n + \rho_n^{k'}\right) &= \mathbb{E}[K_n^2 \mathbf{1}_{\{\sup_{0 \leq s \leq t} k'_n(s) \geq K_n + \rho_n^{k'}\}}] \\ &\leq \mathbb{E}\left[(k'_n(T \wedge \tau^{K_n}) - \rho_n^{k'})^2 \mathbf{1}_{\{\sup_{0 \leq s \leq t} k'_n(s) \geq K_n + \rho_n^{k'}\}}\right] \\ &\leq \mathbb{E}[(k'_n(T \wedge \tau^{K_n}) - \rho_n^{k'})^2]. \end{aligned}$$

By using that  $\mathbb{M}_n^{\tau^{K_n}}(\cdot)$  is a martingale, we find

$$\begin{aligned} \mathbb{E}[(k'_n(T \wedge \tau^{K_n}) - \rho_n^{k'})^2] &= \mathbb{E}[(k'_n(0 \wedge \tau^{K_n}) - \rho_n^{k'})^2] \\ &\quad + \mathbb{E} \int_0^{T \wedge \tau^{K_n}} \frac{c\lambda_n k'_n(s)(1 - 2(k'_n(s) - \rho_n^{k'})) + d_{I-1}M_n(1 + 2(k'_n(s) - \rho_n^{k'}))}{M_n} ds, \end{aligned}$$

which is upper bounded by

$$\mathbb{E}\left[\left(k'_n(0 \wedge \tau^{K_n}) - \rho_n^{k'}\right)^2\right] + \left(2d_{I-1}M_n + \frac{1}{8} \frac{c\lambda_n}{M_n}\right)T,$$



because the integrand is maximized in  $k'_n(s) = 1/4 + \rho_n^k$ , where it takes the value  $2d_{l-1}M_n + \frac{1c\lambda_n}{8M_n}$ . Summarizing, we have

$$\mathbb{P}\left(\sup_{0 \leq s \leq t} k'_n(s) \geq K_n + \rho_n^k\right) \leq \frac{1}{K_n^2} \left( \mathbb{E}\left[(k_n(0) - \rho_n^k)^2\right] + \left(2d_{l-1}M_n + \frac{1c\lambda_n}{8M_n}\right)T \right),$$

where we use that  $k'_n(0) = k_n(0)$ . Now, observe that when we choose  $K_n = \sqrt{n}$ , the right-hand side goes to zero when  $n \rightarrow \infty$ , proving our first claim. Here, we use the assumption that  $k_n(0) \Rightarrow 0$ , which can be strengthened to convergence in probability.

Now that we have proven  $\hat{k}'_n(\cdot) \xrightarrow{\mathbb{P}} 0$ , we show that  $(\sup_{0 \leq s \leq t} (Q'_n(s) - In)^-)^2/n$  is stochastically bounded (i.e., is of  $\mathcal{O}(1)$ ), which is sufficient to show our second claim because in that case, we have (26); the proof is complete.

To prove the stochastic boundedness, observe that for each  $T \geq 0$ , the process  $\mathbb{S}_n(\cdot) := \{\mathbb{S}_n(t) : t \in [0, T]\}$ , for

$$\mathbb{S}_n(t) := ((Q'_n(t) - In) - (\lambda_n - d_1n)t)^2 + (\lambda_n + d_1n)(T - t),$$

is a nonnegative  $\tilde{\mathbb{F}}_n$ -supermartingale on  $[0, T]$ , where  $\tilde{\mathbb{F}}_n := \{\tilde{\mathcal{F}}_t^n\}_{t \geq 0}$  is its natural filtration. This can be seen as follows. If  $\theta = 0$ , then the process  $\tilde{\mathbb{M}}_n(\cdot) := \{\tilde{\mathbb{M}}_n(t) : t \geq 0\}$ , for  $\tilde{\mathbb{M}}_n(t) := Q'_n(t) - In + (d_1n - \lambda_n)t$ , is an  $\tilde{\mathbb{F}}_n$ -martingale. Therefore  $(\tilde{\mathbb{M}}(\cdot))^2$  is a  $\tilde{\mathbb{F}}_n$ -submartingale, and by the Doob–Meyer decomposition (Karatzas and Shreve [30, theorem 4.10]), we obtain that

$$\{((Q'_n(t) - In) - (\lambda_n - d_1n)t)^2 - (\lambda_n + d_1n)t : t \geq 0\}$$

is an  $\tilde{\mathbb{F}}_n$ -martingale. Now, when  $\theta > 0$ , the process  $Q'_n(\cdot)$  has additional jumps downward, only when  $Q(t) - In > 0$ , in which case  $(Q'_n(t) - In) - (\lambda_n - d_1n)t = (Q'_n(t) - In) + d_1\sqrt{n}\beta t > 0$ , so that every jump downward because of abandonments also decreases the square term in  $\mathbb{S}(\cdot)$ . Therefore,

$$\{((Q'_n(t) - In) - (\lambda_n - d_1n)t)^2 - (\lambda_n + d_1n)t : t \geq 0\},$$

and this is indeed an  $\tilde{\mathbb{F}}_n$ -supermartingale on  $[0, \infty)$  when  $\theta > 0$ . We add a term  $(\lambda_n + d_1n)T$  that is constant on  $[0, T]$  so that  $\mathbb{S}(\cdot)$  is a nonnegative  $\tilde{\mathbb{F}}_n$ -supermartingale on  $[0, T]$ .

Now, by Doob’s supermartingale inequality (Revuz and Yor [46, chapter 1, exercise 1.15]), we obtain for each  $T \geq 0$  and  $t \in [0, T]$  and for each constant  $M > 0$ ,

$$\mathbb{P}\left(\sup_{0 \leq s \leq t} \mathbb{S}_n(s) \geq M\right) \leq \frac{1}{M} \mathbb{E}[\mathbb{S}_n(0)] = \frac{1}{M} ((Q'_n(0) - In)^2 + (\lambda_n + d_1n)T). \quad (27)$$

By assumption,  $n^{-1/2}(Q_n(0) - In)$  converges weakly so that the sequence is tight, and therefore, (27), with  $M = n$ , implies that  $(\sup_{0 \leq s \leq t} (Q'_n(s) - In)^-)^2/n$  is stochastically bounded, which completes the proof.  $\square$

We proceed by defining the sequence of processes  $\{R_n(\cdot)\}_n$ , where  $R_n(\cdot)$  is a birth-death process on the state space of  $Q_n(\cdot)$ , with birth rate  $\lambda_n$  and death rate  $v^{R_n}(t) = d_1n - (d_l - d_{l-1})(R_n(t) - In)^- + \theta(R_n(t) - In)^+$ , having initial condition  $Q_n(0)$ . In addition, the processes  $Q(\cdot)$  and  $R(\cdot)$  are coupled such that the jumps upward happen at identical times, and the jumps downward are such that when for some time  $t \geq 0$ , the death rate  $v^{Q_n}(t)$  of  $Q_n(\cdot)$  is larger than  $v^{R_n}(t)$ , we can only have a jump downward of  $R_n(\cdot)$  when  $Q_n(\cdot)$  jumps downward and vice versa when  $v^{R_n} \geq v^{Q_n}$ .

In the usual notation, we write  $\hat{R}_n(\cdot) = n^{-1/2}(R_n(\cdot) - In)$ . By using Lemma 1, we now have the following proposition.

**Proposition 5.** Assume the conditions of Lemma 1. Then, for all  $t \geq 0$  and all  $\epsilon > 0$ ,

$$\mathbb{P}(\|\hat{Z}_n(\cdot) - \hat{R}_n(\cdot)\|_t > \epsilon) \rightarrow 0, \quad \text{as } n \rightarrow \infty,$$

where the process  $\hat{R}_n(\cdot)$  is embedded into  $D^l[0, \infty)$  by adding zero components.

**Proof.** Because of the way that  $Q_n(\cdot)$  and  $R_n(\cdot)$  are coupled, it is immediate that for each  $t \geq 0$ , the difference  $|Q_n(t) - R_n(t)|$  can only increase when  $v^{Q_n}(t) > v^{R_n}(t)$  and  $Q(t) \leq R(t)$  or  $v^{Q_n}(t) < v^{R_n}(t)$  and  $Q(t) \geq R(t)$ .

Define the function  $f : \mathbb{R} \rightarrow \mathbb{R}$  by  $x \mapsto d_1n - (x - In)^-(d_l - d_{l-1}) + \theta(x - In)^+$ , then  $v^{R_n}(t) = f(R_n(t))$ , and  $v^{Q_n}(t) = f(Q_n(t)) + \epsilon_n(t)$ , where for each  $n$ ,  $\epsilon_n(\cdot) := \sum_{0 \leq i \leq l-2} (d_i + (l-i-1)d_l - (l-i)d_{l-1})n_i(\cdot)$ . Note that  $f$  is increasing and  $|f(Q_n(t)) - f(R_n(t))| \geq |Q_n(t) - R_n(t)| \min(\theta, d_l - d_{l-1}) := \ell |Q_n(t) - R_n(t)|$ , for all  $t \geq 0$ . Furthermore, we have  $|\epsilon_n(t)| \leq |k_n(t)|U$  for constant  $U := \max_{0 \leq i \leq l-2} \frac{|d_i + (l-i-1)d_l - (l-i)d_{l-1}|}{l-i-1}$  and all  $t \geq 0$ .

Because we know that  $k_n(\cdot)/\sqrt{n} \xrightarrow{\mathbb{P}} 0$ , as  $n \rightarrow \infty$ , we also immediately have  $\epsilon_n(\cdot)/\sqrt{n} \xrightarrow{\mathbb{P}} 0$ . This implies that the difference of  $Q_n(\cdot)$  and  $R_n(\cdot)$  can only increase when it is small. More precisely, for all  $t \geq 0$ , the difference

$|Q_n(t) - R_n(t)|$  cannot increase when  $|Q_n(t) - R_n(t)| \geq \sup_{0 \leq s \leq t} k_n(s)U/\ell$  because that would mean for  $Q_n(t) > R_n(t)$ , that

$$v^{Q_n}(t) - v^{R_n}(t) \geq f(Q_n(t)) - f(R_n(t)) - |\epsilon_n(t)| \geq \ell |Q_n(t) - R_n(t)| - \sup_{0 \leq s \leq t} k_n(s)U \geq 0,$$

and for  $R_n(t) > Q_n(t)$ , that

$$v^{R_n}(t) - v^{Q_n}(t) \geq f(R_n(t)) - f(Q_n(t)) - |\epsilon_n(t)| \geq \ell |Q_n(t) - R_n(t)| - \sup_{0 \leq s \leq t} k_n(s)U \geq 0,$$

which is not possible because of the coupling as we concluded before.

We conclude that  $\sup_{0 \leq t \leq t} |Q_n(t) - R_n(t)| < 1 + \sup_{0 \leq s \leq t} k_n(s)U/\ell$ , and so,  $|Q_n(\cdot) - R_n(\cdot)| \xrightarrow{\mathbb{P}} 0$ .  $\square$

**Proposition 6.** *The process  $\hat{R}_n(\cdot)$  converges weakly, as  $n \rightarrow \infty$ , to the one-dimensional diffusion process  $\hat{Q}(\cdot)$  of Theorem 1.*

**Proof.** The proof is a direct adaptation of the proof of Pang et al. [40, theorem 1.2] without a finite buffer for the system. We sketch the main steps.

First, we can write down an integral representation for  $\hat{R}_n(t)$ ,

$$\begin{aligned} \hat{R}_n(t) &= \hat{R}_n(0) + n^{-1/2}A_n(\lambda_n t) \\ &\quad - n^{-1/2}D_n \left( \int_0^t d_I n - (d_I - d_{I-1})(R_n(s) - In)^- + \theta(R_n(s) - In)^+ \, ds \right) \\ &= \hat{R}_n(0) + n^{-1/2}M_{A_n}(t) - n^{-1/2}M_{D_n}(t) - \beta d_I t \\ &\quad + \int_0^t (d_I - d_{I-1})(\hat{R}_n(s))^- - \theta(\hat{R}_n(s))^+ \, ds, \end{aligned}$$

where  $A_n(\cdot)$  and  $D_n(\cdot)$  are independent unit-rate Poisson processes and  $M_{A_n}(t)$  and  $M_{D_n}(\cdot)$  are the martingales obtained by subtracting the compensator of  $A_n(\cdot)$  and  $D_n(\cdot)$ , respectively (cf. Pang et al. [40, theorem 3.4] for the precise construction and relevant filtrations).

Because the functions  $x \mapsto (d_I - d_{I-1})(x)^-$  and  $x \mapsto \theta(x)^+$  are Lipschitz continuous, the integral representation used for  $\hat{R}_n(\cdot)$  is a continuous mapping from  $D^I[0, \infty)$  to  $D^I[0, \infty)$ . In particular,  $\hat{R}_n(\cdot)$  is the image of  $n^{-1/2}[M_{A_n}(\cdot) - M_{D_n}(\cdot)]$  under this continuous map (cf. Pang et al. [40, theorem 4.1]). Therefore, by proving that  $n^{-1/2}[M_{A_n}(\cdot) - M_{D_n}(\cdot)]$  converges weakly to a scaled Brownian motion  $B(\cdot)$  with  $\langle B \rangle_t = 2d_I t$ , the proof is complete because of an application of the continuous mapping theorem. This will be the focus of the remainder of the proof.

To obtain weak convergence of  $\mathbb{M}_n(\cdot) := n^{-1/2}[M_{A_n}(\cdot) - M_{D_n}(\cdot)]$  to  $B(\cdot)$ , we use the martingale functional central limit theorem (FCLT) (Jacod and Shiryaev [26, theorem VIII.3.22]). We check the conditions. For each  $n$ , the martingale  $\mathbb{M}_n(\cdot)$  is locally square integrable, and clearly,  $\Delta \mathbb{M}_n(t) = [\mathbb{M}_n(t) - \lim_{s \uparrow t} \mathbb{M}_n(s)] \rightarrow 0$  for all  $t > 0$  and all sample paths of  $\mathbb{M}_n(\cdot)$ . The quadratic variation of  $\mathbb{M}_n(\cdot)$  is upper bounded by

$$\frac{1}{n} \int_0^t \lambda_n + d_I n + (d_I - d_{I-1})(R_n(s) - In)^- + \theta(R_n(s) - In)^+ \, ds.$$

If we prove that  $(R_n(\cdot) - In)/n \xrightarrow{\mathbb{P}} 0$  as  $n \rightarrow \infty$ , we are done because by application of continuous mapping, we then obtain that  $(R_n(\cdot) - In)^+/n \xrightarrow{\mathbb{P}} 0$  and  $(R_n(\cdot) - In)^-/n \xrightarrow{\mathbb{P}} 0$ , as explained in Pang et al. [40, lemmas 4.2–4.4]. In turn, this allows us to conclude that  $\langle \mathbb{M} \rangle_t \rightarrow 2d_I t$  in probability for all  $t \geq 0$  as  $n \rightarrow \infty$ , completing the requirements for the martingale FCLT.

To show  $(R_n(\cdot) - In)/n \xrightarrow{\mathbb{P}} 0$ , it is sufficient if  $n^{-1}[M_{A_n}(\cdot) - M_{D_n}(\cdot)] \xrightarrow{\mathbb{P}} 0$ , as  $n \rightarrow \infty$ , as explained in Pang et al. [40, section 4.3]. At this point, we can transform the proof of Pang et al. [40, lemma 4.5] to our setting, which essentially states that the required limit is achieved by the strong law of large numbers (SLLN) for Poisson processes. To be precise, for  $M_{A_n}(\cdot)$ , a direct consequence of the SLLN is that

$$\|n^{-1}M_{A_n}(\cdot)\|_t \xrightarrow{\text{a.s.}} 0, \text{ for all } t \geq 0.$$

For  $M_{D_n}(\cdot)$ , the necessary limit is a bit harder to prove. For each  $n$ , consider the process  $\mathcal{U}_n(\cdot) := \{\mathcal{U}_n(\int_0^t \theta(R_n(s) - In)^+ \, ds) : t \geq 0\}$ , where  $\{\mathcal{U}_n(t) : t \geq 0\}$  is a unit-rate Poisson process. Denote  $M_{\mathcal{U}_n}(\cdot) := \{M_{\mathcal{U}_n}(t) : t \geq 0\}$ , with  $M_{\mathcal{U}_n}(t) := \mathcal{U}_n(\int_0^t \theta(R_n(s) - In)^+ \, ds) - \int_0^t \theta(R_n(s) - In)^+ \, ds$ , for the martingale associated with  $\mathcal{U}_n(\cdot)$ . For  $R_n(t)$ , we have the

rough upper bound  $R_n(t) \leq R_n(0) + A_n(\lambda_n t)$ , and therefore, we have

$$\frac{1}{n} \int_0^t R_n(s) - In \, ds \leq t \left( \frac{R_n(0) + A_n(\lambda_n t)}{n} \right).$$

By the strong law of large numbers for  $A(\cdot)$ , we have that for each  $T_1 > 0$ , there exists  $T_2 > 0$  such that

$$\mathbb{P} \left( \frac{1}{n} \int_0^{T_1} \theta(R_n(s) - In)^+ \, ds > T_2 \right) \rightarrow 0.$$

We now have for all  $\epsilon > 0$ ,

$$\begin{aligned} \mathbb{P} \left( \|M_{\mathcal{U}_n}(\cdot)\|_{T_1} > \epsilon \right) &= \mathbb{P} \left( \|M_{\mathcal{U}_n}(\cdot)\|_{T_1} > \epsilon, \frac{1}{n} \int_0^{T_1} \theta(R_n - In)^+ \, ds \leq T_2 \right) \\ &\quad + \mathbb{P} \left( \|M_{\mathcal{U}_n}(\cdot)\|_{T_1} > \epsilon, \frac{1}{n} \int_0^{T_1} \theta(R_n - In)^+ \, ds > T_2 \right) \\ &\leq \mathbb{P} \left( \sup_{0 \leq s \leq T_2} \frac{|\mathcal{U}_n(ns) - ns|}{n} > \epsilon \right) + \mathbb{P} \left( \frac{1}{n} \int_0^{T_1} \theta(R_n - In)^+ \, ds > T_2 \right) \\ &\rightarrow 0. \end{aligned}$$

In a similar way, because  $d_I n - (d_I - d_{I-1})(R_n(t) - In)^- \leq d_I n$ , one has  $R_n(t) \geq R_n(0) - N_1(d_I n t) - N_2(\int_0^t \theta(R_n(s) - In)^+ \, ds)$  for unit-rate Poisson processes  $N_1(\cdot)$  and  $N_2(\cdot)$  so that

$$\int_0^t (R_n(s) - In)^- \, ds \geq t \left( \frac{R_n(0) - In - N_1(d_I n t) - N_2(\int_0^t \theta(R_n(s) - In)^+ \, ds)}{n} \right).$$

Hence, using the approach and results, one finds that for each  $T_1 > 0$ , there is a  $T_3 > 0$  such that

$$\mathbb{P} \left( \frac{1}{n} \int_0^t d_I n + \theta(R_n(s) - In)^+ \, ds > T_3 \right) \rightarrow 0.$$

By splitting the probability  $\mathbb{P}(M_{D_n}(\cdot)_{T_1} > \epsilon)$  on the events  $\left\{ \frac{1}{n} \int_0^t d_I - (d_I - d_{I-1})(R_n(s) - In)^- + \theta(R_n(s) - In)^+ \, ds > T_2 + T_3 \right\}$  and its complement, we find that

$$M_{D_n}(\cdot) \xrightarrow{\mathbb{P}} 0, \text{ as } n \rightarrow \infty,$$

which concludes the proof.  $\square$

## 9. Proof of Diffusion Limit for the MBF Policy in the NWC

In this section, we prove Theorem 2. As discussed in Section 3.2,  $Z(\cdot)$  can be written as a process in which some components are reflected. In the proof, we make this precise by establishing that  $\hat{Z}_n(\cdot) = \Phi(M(\hat{X}_n(\cdot)))$ ; here,  $\Phi$  is a regulator mapping that is associated with a Skorokhod problem (which has to be tailored to our specific case),  $M$  is the mapping that incorporates the position-dependent drift of  $\hat{Z}_n(\cdot)$ , and  $\hat{X}_n(\cdot)$  is a scaled and centered version of a linear combination of counting processes driving the arrivals and departures in the system; see (33) for its definition. We show that the mapping  $\Phi \circ M$  is Lipschitz continuous and derive a weak limit approximation for  $\hat{X}_n(\cdot)$ ; application of the continuous mapping theorem gives a weak limit for  $\hat{Z}(\cdot)$ . The bulk of the proof is dedicated to the construction of the mapping  $\Phi \circ M$ .

For readability, we split the proof into the following four steps.

1. We employ a martingale representation for  $\hat{Z}_n(\cdot)$  to derive the diffusion limit. For this, we first explicitly express the evolution of the process  $Z(\cdot)$  as a linear combination of counting processes. The martingale representation is then obtained by compensating the counting processes (i.e., applying the Doob–Meyer decomposition).

2. We construct the Lipschitz mapping  $\Phi$ , which takes the role of a custom-made regulator mapping, without incorporating the drift of  $Z(\cdot)$ .
3. We show that we can write  $\hat{Z}_n(\cdot) = \Phi(M(\hat{X}_n(\cdot)))$  for the Lipschitz mapping  $M$  that includes the drift of  $\hat{Z}_n(\cdot)$  in the (generalized) Skorokhod problem, as explained in Section 3.2.
4. Finally, we prove that  $\hat{X}_n(\cdot) \Rightarrow B(\cdot)$ , with  $B(\cdot)$  a Brownian motion having  $\langle B(\cdot) \rangle_t = 2d_1t$ . By continuous mapping, we have weak convergence of  $\hat{Z}_n(\cdot)$ .

### 9.1. Step 1: Martingale Representation for $Z(\cdot)$

We start by constructing  $Z(\cdot)$  in terms of independent unit-rate Poisson processes, to which we apply a multiparameter random time change. This way of constructing queueing processes is convenient for obtaining scaling limits. We refer to Pang et al. [40, sections 2 and 3] for a careful treatment of such a construction for the M/M/ $\infty$  system and to Pang et al. [40, section 7] for the M/M/ $n + M$  queue, for which the details regarding the multiparameter random time change are based on Ethier and Kurtz [16, chapter 6, section 2].

For our multidimensional process, we multiply each counting process with a linear combination of  $e_Q$  and  $e_{n_i}$  (with  $e_Q$  and  $e_{n_i}$  introduced in Section 3.2) to obtain jumps occurring in the right coordinates. Recall that we adopted the notation of Pang et al. [40] by writing  $N(\int_0^t \nu(s) ds)$  for a counting process  $N(\cdot)$  with intensity  $\nu(\cdot)$  at time  $t \geq 0$ , where the integral in the argument is the time scaling that was applied to the unit-rate Poisson process  $\{N(t) : t \geq 0\}$ .

Let the number of servers  $n$  be fixed for now, and recall that  $F(\cdot) := I_n - Q(\cdot)$  is the number of free spaces in the system (a negative value indicating that there are no free spaces). Using the time-changed Poisson process representation, we have that  $Z(\cdot)$  satisfies

$$\begin{aligned}
 Z(t) = Z(0) &+ e_Q \left[ A(\lambda t) - D \left( \int_0^t d_I n_I(s) ds \right) - D^\circ \left( \int_0^t \theta(Q(s) - I_n)^+ ds \right) \right] \\
 &+ \sum_{i=1}^{I-2} (-e_Q - e_{n_i} + e_{n_{i-1}}) D_i \left( \int_0^t d_i n_i(s) ds \right) \\
 &+ (-e_Q + e_{I-2}) D_{I-1} \left( \int_0^t d_{I-1} n_{I-1}(s) ds \right) \\
 &+ \sum_{i=1}^{I-3} (-e_{n_{i+1}} + 2e_{n_i} - e_{n_{i-1}}) L_i \left( \int_0^t \lambda \mathbb{1}_{\{\sum_{k=i}^{I-1} n_k(s)=0, F(s)>0\}} \right) \\
 &+ (2e_{n_{I-2}} - e_{n_{I-3}}) L_{I-2} \left( \int_0^t \lambda \mathbb{1}_{\{n_{I-1}(s)+n_{I-2}(s)=0, F(s)>0\}} \right) \\
 &+ (-e_{n_{I-2}}) L_{I-1} \left( \int_0^t \lambda \mathbb{1}_{\{n_{I-1}(s)=0, F(s)>0\}} \right), \quad t \geq 0,
 \end{aligned} \tag{28}$$

where  $n_{I-1}(t)$  and  $n_i(t)$  are given as in Section 2.2, and  $A(\cdot)$ ,  $D(\cdot)$ ,  $D^\circ(\cdot)$ ,  $D_i(\cdot)$ , and  $L_i(\cdot)$  for  $i \in \{1, \dots, I-1\}$  are independent unit-rate Poisson processes. The process  $A(\cdot)$  counts the cumulative number of arriving customers. Changes in the number of servers at the different levels depend on their state at time  $t$ . For  $F(t) \leq 0$ , an arriving customer joins the queue, whereas an arrival is assigned to a level  $I-1$  server when  $F(t) > 0$  and  $n_{I-1}(t) > 0$ ; in that case, the components  $n_i(t)$ , for  $i = 0, \dots, I-2$ , do not change. Observe that an arrival can only be assigned to a level  $j \in \{1, \dots, I-1\}$  server when  $n_i(t) = 0$ , for all  $i > j$ . For each  $j$ , the process  $L_j(\cdot)$  corresponds to such a transition. To be precise,  $L_j(\cdot)$  multiplied by the corresponding unit vectors acts as a regulator to prevent  $n_j(\cdot)$  from becoming negative by sending an arrival to a level  $j-1$  server when  $n_i(t) = 0$  for all  $i \geq j$ . The processes  $D_i(\cdot)$  and  $D(\cdot)$  count the cumulative number of service completions of level  $i \in \{1, \dots, I-1\}$  and level  $I$  servers, respectively. Note that a service completion at a level  $i$  server provides an extra level  $i-1$  server and leads to one customer less in the system. The process  $D^\circ(\cdot)$  counts the total number of abandonments from the queue.

From now on, we consider the sequence  $\{\hat{Z}_n(\cdot)\}_n$ , as defined in Section 2.3. Note that (28) holds, with  $\lambda$  replaced by  $\lambda_n$ , and both sides of the equation are scaled by  $n^{-1/2}$ . Moreover, for the  $n$ th system, we add a subscript  $n$  in the

notation of the counting processes, and we denote  $\hat{L}_{n,i}(\cdot) = L_{n,i}(\cdot)/\sqrt{n}$ , for  $i \in \{1, \dots, I-1\}$ . We have

$$\begin{aligned} \hat{Z}_n(t) = & \hat{Z}_n(0) + n^{-1/2} \left( e_Q \left[ A_n(\lambda_n t) - D_n \left( \int_0^t d_I n_{n,I}(s) \, ds \right) - D_n^\circ \left( \int_0^t \theta(Q_n(s) - In)^+ \, ds \right) \right] \right. \\ & + \sum_{i=1}^{I-2} (-e_Q - e_{n_i} + e_{n_{i-1}}) D_{n,i} \left( \int_0^t d_i n_{n,i}(s) \, ds \right) + (-e_Q + e_{I-2}) D_{n,I-1} \left( \int_0^t d_{I-1} n_{n,I-1}(s) \, ds \right) \\ & + \sum_{i=1}^{I-3} (-e_{n_{i+1}} + 2e_{n_i} - e_{n_{i-1}}) \hat{L}_{n,i} \left( \int_0^t \lambda_n \mathbb{1}_{\{\sum_{k=i}^{I-1} n_{n,i}(s)=0, F(s)>0\}} \right) \\ & \left. + (2e_{n_{I-2}} - e_{n_{I-3}}) \hat{L}_{n,I-2} \left( \int_0^t \lambda_n \mathbb{1}_{\{n_{n,I-1}(s)+n_{n,I-2}(s)=0, F(s)>0\}} \right) + (-e_{n_{I-2}}) \hat{L}_{n,I-1} \left( \int_0^t \lambda_n \mathbb{1}_{\{n_{n,I-1}(s)=0, F(s)>0\}} \right) \right). \end{aligned} \quad (29)$$

We proceed by rewriting (29) into a martingale representation for  $\hat{Z}_n(\cdot)$ , similar to Pang et al. [40, sections 3 and 4]. To introduce martingales, we compensate all counting processes by the integral of their intensity; we refer to Pang et al. [40, section 7.1] for a detailed exposition of this technique relying on the multiparameter optional sampling theorem. For a counting process  $N(\cdot)$ , with intensity  $\nu(\cdot)$ , denote by  $M_N(\cdot)$  the associated martingale obtained by subtracting its compensator (i.e.,  $M_N(t) := N(t) - \int_0^t \nu(s) \, ds$ ,  $t \geq 0$ ).

For later use, we also write  $D_n(\cdot) = \tilde{D}_n(\cdot) - \tilde{D}_n^c(\cdot)$ , the difference of a counting process  $\tilde{D}_n(\cdot)$  with intensity  $d_I n$ , and a process  $\tilde{D}_n^c(\cdot)$  with intensity  $d_I(n - n_{n,I}(\cdot))$ . Applying the martingale representation yields

$$\begin{aligned} \hat{Z}_n(t) = & \hat{Z}_n(0) + n^{-1/2} e_Q [M_{A_n}(t) - M_{\tilde{D}_n}(t)] + n^{-1/2} e_Q [M_{\tilde{D}_n^c}(t) - M_{D_n^\circ}(t)] \\ & + n^{-1/2} \sum_{i=1}^{I-2} (-e_Q - e_{n_i} + e_{n_{i-1}}) M_{D_{n,i}}(t) + n^{-1/2} (-e_Q + e_{I-2}) M_{D_{n,I-1}}(t) \\ & + \int_0^t f(\hat{Z}_n(s)) \, ds + \begin{pmatrix} 0 & 0 \\ 0 & \tilde{R} \end{pmatrix} \begin{pmatrix} 0 \\ \hat{L}_n(t) \end{pmatrix}, \end{aligned} \quad (30)$$

where  $f$  is the function representing the drift as defined in (6),  $\hat{L}_n(\cdot)$  is the (column) vector-valued process  $(\hat{L}_{n,1}(\cdot), \dots, \hat{L}_{n,I-1}(\cdot))^\top$ , and  $\tilde{R}$  is the  $(I-1) \times (I-1)$  matrix as defined just above Theorem 2. As in Pang et al. [40], the martingales in this process are associated to the filtration  $\mathbb{F}_n := \{\mathcal{F}_{n,t}\}_{t \geq 0}$ , with

$$\begin{aligned} \mathcal{F}_{n,t} := & \sigma(Z_n(0), A_n(\lambda_n s), \tilde{D}_n(d_I n s), \tilde{D}_n^c \left( \int_0^s d_I(n - n_{n,i}(u)) \, du \right), \\ & D_n^\circ \left( \int_0^s \theta(Q_n(u) - In)^+ \, du \right), D_{n,i} \left( \int_0^s d_i n_{n,i}(u) \, du \right) : i \in \{1, \dots, I-1\}, s \leq t), \quad t \geq 0, \end{aligned}$$

augmented by including all  $\mathbb{P}$ -null sets of  $\mathcal{F}$ .

## 9.2. Step 2: Construction of Custom Regulator Mapping $\Phi$

To define our custom regulator map  $\Phi$ , we require some auxiliary definitions and results. We start by defining  $T : \mathbb{R}^I \rightarrow \mathbb{R}^I$  as the function given by

$$(x_1, \dots, x_I) \mapsto \left( x_1, x_3, \dots, x_I, \max(-x_1, 0) - \sum_{i=2}^I i x_{I-i+2} \right).$$

The goal of applying  $T$  is to switch from a state representation including  $n_0(\cdot)$  to one including  $n_{I-1}(\cdot)$  because  $T(Z(t) - (In, 0, \dots, 0)) = (Q(t) - In, n_1(t), \dots, n_{I-1}(t))$ . We use this map to prove that the reflection mapping we construct is well defined and continuous. The following straightforward lemma is for later use.

**Lemma 2.** *T is a bijection that is Lipschitz continuous.*

**Proof.** It is elementary to show that  $T$  is both injective and surjective. The Lipschitz property is a consequence of  $T$  being a piecewise linear function on  $\mathbb{R}^I$ .



The domain for  $T$  as defined is too large, and we have to verify that  $T$  is still a Lipschitz bijection when the domain is restricted to the state space of  $Z(\cdot)$ . Define  $S^1 := \{(F, n_0, \dots, n_{I-2}) \in \mathbb{R}^I : n_0, \dots, n_{I-2} \geq 0, (F)^+ - \sum_{i=2}^I in_{I-i} \geq 0\}$  and  $S^2 := \{(F, n_1, \dots, n_{I-1}) \in \mathbb{R}^I : n_1, \dots, n_{I-1} \geq 0, (F)^+ - \sum_{i=1}^{I-1} in_{I-i} \geq 0\}$ . It is now readily verified that  $T|_{S^1}$ , the restriction of  $T$  to  $S^1$ , is again a Lipschitz bijection onto  $S^2$ .

Next, we introduce the reflection matrix  $R$  associated with the process  $(Q(\cdot) - In, n_1(\cdot), \dots, n_{I-1}(\cdot))$ . Let  $R$  be a matrix in  $\mathbb{R}^{(I-1) \times (I-1)}$ , with every diagonal element equal to 2 and every upper and lower diagonal element equal to  $-1$  (i.e.,  $R_{ij} = 2\mathbb{1}_{\{i=j\}} - \mathbb{1}_{\{i=j+1\}} - \mathbb{1}_{\{i=j-1\}}$ , with  $i, j \in \{1, \dots, I-1\}$ ). For a matrix  $A$ , let  $|A|$  denote the matrix whose entries are the absolute values of  $A$ , and let  $\text{diag}(A)$  denote the diagonal matrix whose diagonal entries are taken from  $A$ .  $\square$

**Definition 3.** A  $d \times d$  matrix  $R$  is said to be a generalized  $\mathcal{M}$  matrix if

- i. each diagonal entry of  $R$  is positive and
  - ii. the spectral radius of  $|H|$  is less than one,
- where  $H$  is the  $d \times d$  matrix that satisfies

$$R = (Id - H)\text{diag}(R),$$

with  $Id$  the  $d \times d$  identity matrix.

It is now easily verified that the matrix  $R$  defined is a generalized  $\mathcal{M}$ -matrix; the first property of Definition 3 evidently holds, and the associated matrix  $|H|$  is substochastic; thus, it has a largest eigenvalue smaller than one.

As mentioned in Section 3.2, the diffusion limit is related to a Skorokhod problem. For our process, we need a slight modification of the usual Skorokhod setting. Denote  $D_+^{d+1}[0, \infty)$  for the elements  $x(\cdot) \in D_+^{d+1}[0, \infty)$  satisfying  $x(0) = 0$ . We adopt the notation that an element  $x(\cdot) \in D_+^{d+1}[0, \infty)$  is denoted  $(x_1(\cdot), x_{2,d+1}(\cdot))$ , where  $x_{2,d+1}(\cdot) \in D_+^d[0, \infty)$  denote the last  $d$  components of  $x(\cdot)$ . We refer to our version of the Skorokhod problem as the *coupled Skorokhod problem* because the number of reflection directions is fewer than the dimension of the process, whereas the processes, in general, cannot be decoupled.

**Definition 4** (Coupled Skorokhod Problem). Let  $x(\cdot) \in D_+^{d+1}[0, \infty)$ , and let  $R$  be a  $d \times d$  real matrix. A pair  $(z(\cdot), l(\cdot)) \in D_+^{d+1}[0, \infty) \times D_+^{d+1}[0, \infty)$  is a solution of the Skorokhod problem for  $x(\cdot)$  (with respect to  $R$ ) if the following conditions hold:

- i.  $\begin{pmatrix} z_1(t) \\ z_{2,d+1}(t) \end{pmatrix} = \begin{pmatrix} x_1(t) \\ x_{2,d+1}(t) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} 0 \\ l_{2,d+1}(t) \end{pmatrix}$ ,  $t \geq 0$ ;
- ii.  $z_{2,d+1}(t) \geq 0$ ,  $t \geq 0$ ; and
- iii.  $l(\cdot)$  is such that for  $i = 2, \dots, d+1$ ,
  - a.  $l_i(0) = 0$ ,
  - b.  $l_i(\cdot)$  is nondecreasing, and
  - c.  $\int_0^\infty z_i(t) dl_i(t) = 0$ .

We call  $(z(\cdot), l(\cdot))$  a *coupled  $R$  regulation* for  $x(\cdot)$ . The following result is a straightforward adaptation of Harrison and Reiman [23, theorem 1].

**Proposition 7.** Assume that the  $d \times d$  real matrix  $R$  is a generalized  $\mathcal{M}$  matrix. Then, for each  $x(\cdot) \in D_+^{d+1}[0, \infty)$ , a coupled  $R$  regulation for  $x(\cdot)$  exists and is unique.

In these settings, we could have taken  $D_+^{d+1}[0, \infty)$  as well when  $x(0)$  is in the state space associated with  $z(\cdot)$  (cf. Harrison and Reiman [23, theorem 1]); we, therefore, replace  $D_+^{d+1}[0, \infty)$  by  $D^{d+1}[0, \infty)$  in the remainder of this section. Because of Proposition 7, we can define a mapping associated with  $R$ ,  $(\phi, \psi) : D^{d+1}[0, \infty) \rightarrow D^{d+1}[0, \infty) \times D^{d+1}[0, \infty)$ , by  $x(\cdot) \mapsto (z(\cdot), l(\cdot))$ . The following proposition is an adaptation of Whitt [56, theorem 14.2.5], the proof of which goes along the same lines and is, therefore, omitted.

**Proposition 8.** Assume  $R$  is a generalized  $\mathcal{M}$  matrix. There exists a constant  $K > 0$ , which depends on  $R$ , such that for  $x, x' \in D_+^{d+1}$  and for each  $t \geq 0$ ,

$$\|\phi(x) - \phi(x')\|_t \vee \|\psi(x) - \psi(x')\|_t \leq K\|x - x'\|_t,$$

where for  $a, b \in \mathbb{R}$ ,  $a \vee b = \max\{a, b\}$ .

Now, we have the required theory to introduce the mapping  $\Phi$ . For the generalized  $\mathcal{M}$ -matrix  $R$ , define the function  $(\Phi, \Psi) : D^I[0, \infty) \rightarrow D^I[0, \infty) \times D^I[0, \infty)$  by

$$x(\cdot) \mapsto (T^{-1} \circ \phi \circ T(x(\cdot)), T^{-1} \circ \psi \circ T(x(\cdot))), \quad (31)$$

which is a well-defined function because  $T$  is a bijection by Lemma 2 and  $(\phi, \psi)$  are functions by Proposition 7. Moreover,  $(\Phi, \Psi)$  is Lipschitz because owing to Proposition 8 and again, Lemma 2, it is a composition of Lipschitz mappings. Finally, because  $T|_{\mathcal{S}^1}$  is a Lipschitz bijection onto  $\mathcal{S}^2$ , we have that  $\Phi(x) \in \mathcal{S}^1$  when  $\phi \circ T(x) \in \mathcal{S}^2$ .

### 9.3. Step 3: Establish That $\hat{Z}_n(\cdot) = \Phi(M(\hat{X}_n(\cdot)))$

This step is devoted to write  $\hat{Z}_n(\cdot) - \hat{Z}_n(0)$  as the image of a continuous mapping, acting on a martingale  $\hat{X}_n(\cdot)$ , the latter being the martingale part of  $\hat{Z}_n(\cdot)$  as defined in (33). In the next step, we show that  $\hat{X}_n(\cdot)$  converges weakly to a scaled Brownian motion, as  $n \rightarrow \infty$ , deriving a weak limit for  $\hat{Z}_n(\cdot)$ . As mentioned, we require the theory of regulator mappings to write  $\hat{Z}_n(\cdot) - \hat{Z}_n(0)$  as the image of a continuous map acting on  $\hat{X}_n$ . The functional form associated with regulator mappings is also apparent in (30), where the pair  $(\hat{Z}_n(\cdot), \hat{L}_n(\cdot))$  can be seen as a coupled  $\tilde{R}$  regulation, with  $\tilde{R}$  as defined in Section 3.2, of the process

$$n^{-1/2} \left[ e_Q[M_{A_n}(t) - M_{\tilde{D}_n}(t)] + e_Q[M_{\tilde{D}_n^c}(t) - M_{D_n^c}(t)] + \sum_{i=1}^{I-2} (-e_Q - e_{n_i} + e_{n_{i-1}})M_{D_{n,i}}(t) + (-e_Q + e_{I-2})M_{D_{n,I-1}}(t) + \int_0^t f(\sqrt{n}\hat{Z}_n(s)) ds \right]. \quad (32)$$

As indicated in Section 3.2, two issues arise. The first is that  $\hat{Z}_n(\cdot)$  appears in (32), which is not the case in the process  $(x_1(\cdot), x_{2,d+1}(\cdot))$  in Definition 4. This can be solved by considering a generalized version of the coupled Skorokhod problem (cf. Reed and Ward [42], Reed and Ward [43], Reed et al. [44]). The second issue is that  $\tilde{R}$  is not a generalized  $\mathcal{M}$  matrix. For this, we apply the function  $T$  and employ the fact that the matrix  $R$ , associated with the process  $(Q(\cdot) - In, n_1(\cdot), \dots, n_{I-1}(\cdot))$ , is a generalized  $\mathcal{M}$  matrix. We combine both elements to construct a “custom reflection mapping.”

For ease of notation, define the processes  $\hat{Y}_n(\cdot)$  and  $\hat{X}_n(\cdot)$  by

$$\begin{aligned} \hat{Y}_n(t) &:= \hat{Z}_n(t) - \begin{pmatrix} 0 & 0 \\ 0 & \tilde{R} \end{pmatrix} \begin{pmatrix} 0 \\ \hat{L}_n(t) \end{pmatrix}, \quad t \geq 0 \\ \hat{X}_n(t) &:= n^{-1/2} [e_Q[M_{A_n}(t) - M_{\tilde{D}_n}(t)] + e_Q[M_{\tilde{D}_n^c}(t) - M_{D_n^c}(t)] \\ &\quad + \sum_{i=1}^{I-2} (-e_Q - e_{n_i} + e_{n_{i-1}})M_{D_{n,i}}(t) + (-e_Q + e_{I-2})M_{D_{n,I-1}}(t)], \quad t \geq 0. \end{aligned} \quad (33)$$

It is now straightforward to verify that  $\hat{Z}_n(\cdot) = \Phi(\hat{Y}_n(\cdot))$ ; for the mapping  $\Phi$ , we defined in (31) because for  $t \geq 0$ ,

$$\begin{aligned} \hat{Z}_n(t) &= \hat{Y}_n(t) + \begin{pmatrix} 0 & 0 \\ 0 & \tilde{R} \end{pmatrix} \begin{pmatrix} 0 \\ \hat{L}_n(t) \end{pmatrix} = \hat{Y}_n(t) + T^{-1} \begin{pmatrix} 0 & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} 0 \\ \hat{L}_n(t) \end{pmatrix} \\ &= T^{-1} \left( T(\hat{Y}_n(t)) + \begin{pmatrix} 0 & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} 0 \\ \hat{L}_n(t) \end{pmatrix} \right) = T^{-1} \phi T(\hat{Y}_n(t)) = \Phi(\hat{Y}_n(t)), \end{aligned}$$

where the third equality holds as, for  $x, y \in \mathbb{R}^I$ ,  $T^{-1}(x + y) = T^{-1}x + T^{-1}y$ , when  $y_1 = 0$ . The properties (ii) and (iii) in Definition 4 then imply that

- ii'.  $\hat{n}_{n,0}(t), \dots, \hat{n}_{n,I-1}(t) \geq 0$ , for  $t \geq 0$ ; and
- iii'.  $\hat{L}_n(\cdot)$  is such that for  $i = 1, \dots, I - 1$ ,
  - a.  $\hat{L}_{n,i}(0) = 0$ ,
  - b.  $\hat{L}_{n,i}(\cdot)$  is nondecreasing, and
  - c.  $\int_0^\infty \hat{n}_{n,i}(t) d\hat{L}_{n,i}(t) = 0$ .

Property (ii') is true because property (ii) in Definition 4 implies that  $\hat{n}_{n,1}(t), \dots, \hat{n}_{n,I-1}(t) \geq 0$  for  $t \geq 0$ . The resulting process is in  $\mathcal{S}^2$ ; hence, applying  $T^{-1}$  gives that  $\hat{n}_0(t) \geq 0$ . Property (iii') is a direct consequence of property (iii) in Definition 4, which is preserved under  $T : \mathcal{S}^1 \rightarrow \mathcal{S}^2$ .

We now write  $\hat{Y}_n(\cdot)$  as the image under a continuous mapping acting on  $\hat{X}_n(\cdot)$ . For this, set  $\hat{Y}_n(\cdot) = M(\hat{Z}_n(0) + \hat{X}_n(\cdot))$ , with  $M : D^I[0, \infty) \rightarrow D^I[0, \infty)$ , the mapping that sets  $M(x(\cdot)) = v(\cdot)$ , for  $v(\cdot)$  solving the integral equation

$$v(t) = x(t) + \int_0^t f(\Phi(v(s))) \, ds, \quad \text{for all } t \geq 0. \quad (34)$$

The mapping  $M$  can be shown to be well defined and Lipschitz continuous, which is proven in Reed and Ward [42, lemma 1]. For completeness, we state the result, but we omit the proof.

**Proposition 9.** *Let  $h : \mathbb{R}^I \rightarrow \mathbb{R}^I$  be a Lipschitz continuous function. For any  $x(\cdot) \in D^I[0, \infty)$ , there is a unique  $v(\cdot) \in D^I[0, \infty)$  satisfying*

$$v(t) = x(t) + \int_0^t h(v(s)) \, ds, \quad \text{for all } t \geq 0. \quad (35)$$

The function associated with  $h$ ,  $M^h$ , that maps  $x(\cdot)$  to  $v(\cdot)$  is Lipschitz continuous w.r.t. the uniform norm on compact time intervals.

Because  $f \circ \Phi$  is Lipschitz continuous w.r.t. the uniform norm on compact intervals, we set  $M \equiv M^{f \circ \Phi}$ . Combining the fact that  $\Phi$  is a Lipschitz function with Proposition 9 yields our “custom regulator mapping,” which we summarize in the following proposition.

**Proposition 10.** *For every  $x(\cdot) \in D^{d+1}[0, \infty)$ , with  $x(0) \in \mathbb{R} \times \mathbb{R}_+^{I-1}$ , there exists a unique pair  $(z(\cdot), l(\cdot)) \in D^{d+1}[0, \infty) \times D^{d+1}[0, \infty)$  such that*

$$(z(\cdot), l(\cdot)) = (\Phi \circ M(x(\cdot)), \Psi \circ M(x(\cdot))),$$

with  $(\Phi \circ M, \Psi \circ M) : D^I[0, \infty) \rightarrow D^I[0, \infty) \times D^I[0, \infty)$  a Lipschitz continuous mapping. In addition, when  $\phi \circ T \circ M(x(\cdot)) \in \mathcal{S}^2$ , then  $z_i(\cdot) \geq 0$  for  $i \geq 2$ , and  $l(\cdot)$  satisfies

- a.  $l(0) = 0$ ,
- b.  $l_i(\cdot)$  is nondecreasing for  $i = 1, \dots, I$ , and
- c.  $\int_0^\infty z_i(t) \, dl_i(t) = 0$  for  $i = 2, \dots, I$ .

#### 9.4. Step 4: Weak Limit for $\hat{X}_n(\cdot)$

With the definition of  $\Phi$  and  $M$ , we may now complete the proof of Theorem 2 by deriving a weak limit for  $\hat{X}_n(\cdot)$ . We show that  $\hat{X}_n(\cdot) \Rightarrow B(\cdot)$  in  $D^I[0, \infty)$  with the uniform norm on compact time intervals. This completes the proof because the remainder is an application of the mapping theorem (Billingsley [5, theorem 2.7]) to  $\hat{Z}_n(0) + \hat{X}_n(\cdot)$ . First, write  $\hat{X}_n(\cdot) = \hat{\Xi}_n(\cdot) + \hat{\Gamma}_n(\cdot)$ , where

$$\begin{aligned} \hat{\Xi}_n(t) &= \frac{1}{\sqrt{n}} e_Q [M_{A_n}(t) - M_{\bar{D}_n}(t)] \\ \hat{\Gamma}_n(t) &= \frac{1}{\sqrt{n}} (e_Q [M_{\bar{D}_n^c}(t) - M_{D_n^c}(t)] + \sum_{i=1}^{I-2} (-e_Q - e_{n_i} + e_{n_{i-1}}) M_{D_{n,i}}(t) \\ &\quad + (-e_Q + e_{I-2}) M_{D_{n,I-1}}(t)), \end{aligned}$$

with the martingales,  $\hat{\Xi}_n(\cdot)$  and  $\hat{\Gamma}_n(\cdot)$ , depending on  $n$  through the definition of  $\hat{X}_n(\cdot)$ . We will show that  $\hat{\Xi}_n(\cdot) \Rightarrow B(\cdot)$  and  $\hat{\Gamma}_n(\cdot) \xrightarrow{\mathbb{P}} 0$ .

First, we show the latter statement. Note that  $\hat{\Gamma}_n(\cdot)$  is a locally square integrable martingale, having predictable quadratic variation matrix  $\langle \hat{\Gamma}_n \rangle_t$ , whose diagonal entries  $\langle \hat{\Gamma}_n^{(i,i)} \rangle_t$ ,  $i \in \{1, \dots, I\}$ , have the (rough) upper bound

$$\langle \hat{\Gamma}_n^{(i,i)} \rangle_t \leq \frac{1}{n} \int_0^t \left( d_I(n - n_{n,I}(s)) + \theta(Q_n(s) - In)^+ + \sum_{i=1}^{I-2} 3(d_i n_{n,i}(s)) + 2d_{I-1} n_{n,I-1}(s) \right) ds,$$

and where the (off-diagonal) predictable quadratic covariations can be bounded by

$$|\langle \hat{\Gamma}_n^{(i,j)} \rangle_t|^2 \leq \langle \hat{\Gamma}_n^{(i,i)} \rangle_t \langle \hat{\Gamma}_n^{(j,j)} \rangle_t, \quad i, j \in \{1, \dots, I\} \, t \geq 0,$$

by application of the Cauchy–Schwarz inequality. Therefore, if  $\langle \hat{\Gamma}_n^{(i,i)} \rangle_t \rightarrow 0$  in probability for all  $t \geq 0$  and each  $i \in \{1, \dots, I\}$ , then  $\langle \hat{\Gamma}_n^{(i,j)} \rangle_t \rightarrow 0$  in probability for each  $t \geq 0$  and all  $i, j \in \{1, \dots, I\}$  as well.

Because  $i \mapsto d_i$  is increasing and because  $(n - n_{n,i}(s)) = \sum_{i=0}^{I-1} n_{n,i}(s) \leq (F_n(s))^+$ , we get a further upper bound for the diagonal elements of  $\langle \hat{\Gamma}_n \rangle_t$ ,

$$\langle \hat{\Gamma}_n^{(i,i)} \rangle_t \leq \frac{6}{n} (d_I + \theta) \int_0^t (F_n(s))^+ + (F_n(s))^- ds,$$

where we use that  $\theta(Q_n(s) - In)^+ = \theta(F_n(s))^-$ .

Now, recognize that the process  $F_n(\cdot)$  increases by one with birth rate  $\sum_{i=1}^I d_i n_{n,i}(t) + \theta(F_n(t))^- \leq d_I n + \theta(F_n(t))^-$  and decreases by one with rate  $\lambda_n t$  at time  $t \geq 0$ . With the upper bound for the birth rate, we recognize an Erlang A system. Using the proof of Proposition 6, inspired by Pang et al. [40], we obtain that  $(F_n(\cdot))^+ / n \xrightarrow{\mathbb{P}} 0$  and  $(F_n(\cdot))^- / n \rightarrow 0$  such that  $\langle \hat{\Gamma}_n \rangle_t \rightarrow 0$  in probability for all  $t \geq 0$ . By application of the martingale FCLT (Jacod and Shiryaev [26, theorem VIII.3.22]), we get  $\hat{\Gamma}_n(\cdot) \Rightarrow 0$ , and by Slutsky, we get  $\hat{\Gamma}_n(\cdot) \xrightarrow{\mathbb{P}} 0$ , which shows our claim.

We continue by proving the claim for  $\hat{\Xi}_n(\cdot)$ , for which we again use the martingale FCLT. For all  $n$ , the process  $\hat{\Xi}_n(\cdot)$  is a locally square integrable martingale. We check the conditions of Jacod and Shiryaev [26, theorem VIII.3.22]. First,  $\Delta \hat{\Xi}_n(t) = \hat{\Xi}_n(t) - \hat{\Xi}_n(t^-) \leq 1/\sqrt{n} \rightarrow 0$  for all  $\omega \in \Omega$ . Second, we have for all  $t \geq 0$ ,

$$\langle \hat{\Xi}_n^{(1,1)} \rangle_t = \frac{1}{n} \int_0^t \lambda_n + d_I n ds \xrightarrow{\text{a.s.}} 2d_I t,$$

with all the other components of  $\langle \hat{\Xi}_n \rangle_t$  being zero. The conditions for the martingale FCLT are, therefore, satisfied, and we conclude that  $\hat{\Xi}_n(\cdot) \Rightarrow e_Q B(\cdot)$ , as  $n \rightarrow \infty$ . This completes step 4 and finalizes the proof of Theorem 2.

## 10. Numerical Illustration

In this section, we investigate the performance of the multitasking system as a function of the number of servers. As indicated in Section 6, we focus on the probability of delay, as this is a primary building block in many approximations. Our aim is (i) to verify the impact of the number of servers on the probability of delay and give a rough idea at what number of servers  $n$  our diffusion approximations become accurate and (ii) to get some insight into the impact of the routing policy.

For our numerical experiments, we take  $I = 4$  and the following three shapes of the multitasking effect:  $d_i^{(\text{cc})} = 1.25\sqrt{i}$  (concave),  $d_i^{(\text{cv})} = 0.25i^2$  (convex) for  $i = 0, 1, \dots, I$ , and  $d^{(\text{mx})} = (0, 0.5, 1.5, 2.25, 2.75)$  (mixed). The arrival rate is given by (2), where we take  $\beta = 0.5$ . As the number of servers, we take  $n \in \{2, 4, 6, 8, 10, 15, 20, 40, 60, 80, 100\}$ , and for the abandonment rate from the queue, we take  $\theta = 0.2$ . For each parameter combination, we use a warm-up period of 5,000 arrivals, and the simulation is run for 1,000 batches of 50,000 arrivals (i.e., 50 million arrivals in total per parameter combination) such that confidence intervals for the simulated probability of delay are reasonably small. As the width of the confidence intervals was somewhat larger, we used batches of 100,000 arrivals in case of MBF and  $n \geq 40$ .

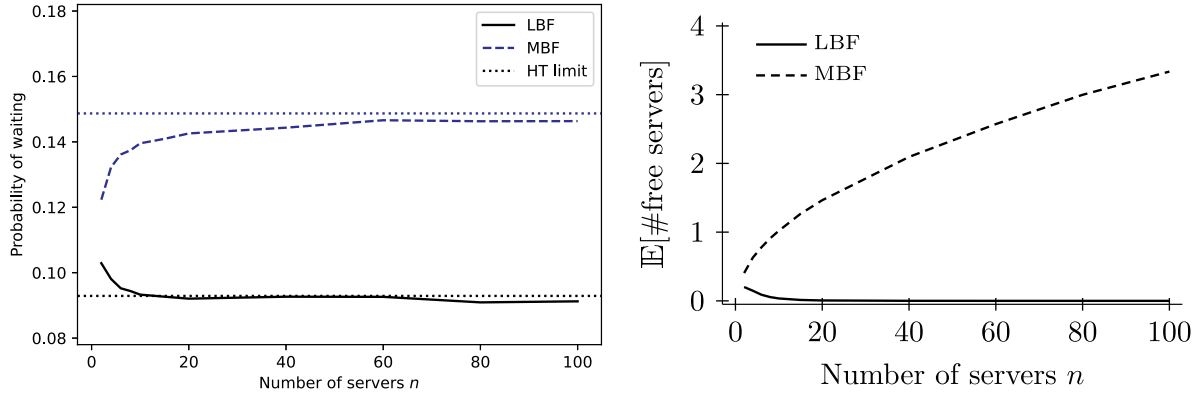
The approximation of the probability of waiting can be found in Section 6, and for the LBF policy, it can be expressed in terms of the Garnett delay function (see (13)). Let  $\hat{\mathbb{P}}(\text{Delay})$  denote the simulated probability of delay. For LBF, we define the absolute and relative errors between the simulation results and heavy-traffic limit as follows:

$$\Delta_{\text{diff}} = \text{Garnett} \left( \frac{\beta}{\sqrt{a}}, \frac{\theta}{ad_I} \right) - \hat{\mathbb{P}}(\text{Delay}), \quad \Delta_{\text{rel}} = 100\% \times \frac{\Delta_{\text{diff}}}{\hat{\mathbb{P}}(\text{Delay})}$$

with  $a = 1 - d_{I-1}/d_I$ . For MBF, the diffusion process only has a tractable stationary distribution in the SWC and not in the NWC. To determine the errors, we replace the heavy-traffic limit with the simulation results for  $n = 200$  based on 100 million arrivals.

The probability of waiting for a concave multitasking effect  $d^{(\text{cc})}$  as a function of the number of servers is presented in the left panel of Figure 3 for the LBF (solid line) and MBF (dashed line) policies; the corresponding heavy-traffic limits (simulation for  $n = 200$  in case of MBF) are indicated by dotted lines. In line with the optimality results of Theorem 5, LBF outperforms MBF for every  $n$ . Using the heavy-traffic limit, the approximate probability of waiting for LBF is roughly 0.0929. We see that probabilities of delay converge to their heavy-traffic limits. The small differences for larger  $n$  are because of randomness in the simulation, with the half-width of the confidence interval being about 0.0015 for 100 million arrivals. The speed of convergence appears to be quite high for LBF. The absolute ( $\Delta_{\text{diff}}$ ) and relative ( $\Delta_{\text{rel}}$ ) errors between the simulated values and their limiting behavior can be found in Table 1. For LBF, already for  $n = 6$ , the probability of waiting is reasonably close to its heavy-traffic limit. The speed of

**Figure 3.** (Color online) Probability of waiting (left panel) and expected number of free servers (right panel) as a function of  $n$  for concave  $d^{(cc)}$  multitasking effect (LBF is shown as solid lines, and MBF is shown as dashed lines; heavy-traffic (HT) limits are shown as dotted lines).



convergence for MBF is somewhat slower in this case, which is partly because of the choice of  $d^{(cc)}$ . However, although the relative errors remain significant until  $n=20$ , the actual differences are modest, with deviations in the probability of delay of less than 0.01 from  $n=10$  on.

Although MBF is the worst policy in terms of waiting times for concave multitasking effects, the MBF routing policy has the advantage over LBF of creating free servers. This is depicted in the right panel of Figure 3, where the long-run expected number of level 0 servers is plotted as a function of  $n$ . Clearly, for LBF, the number of level 0 servers tends to zero very fast, in line with the state-space collapse. For MBF, the expected number of free servers grows roughly as a square root of  $n$ , matching the diffusion behavior. Hence, when using an MBF policy, servers may, for example, be available to perform other tasks.

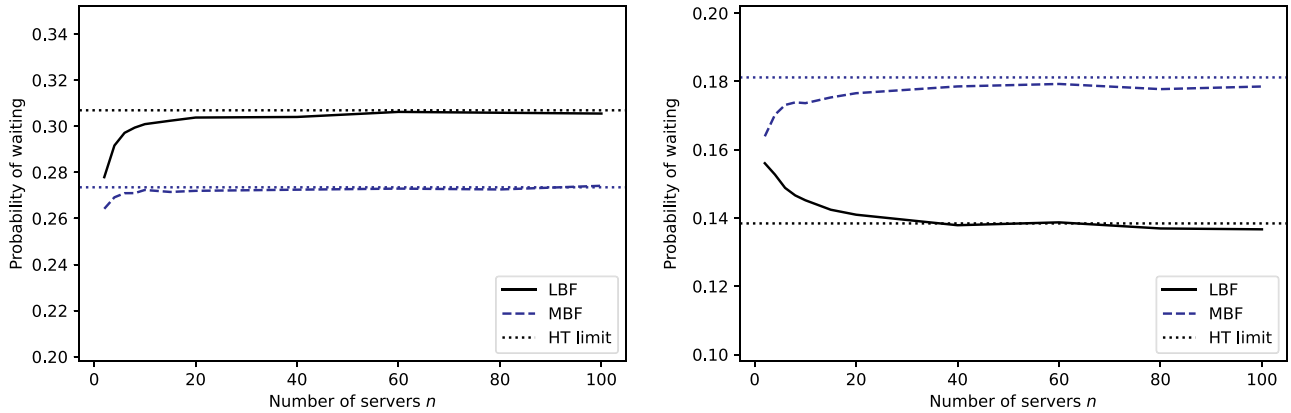
In Figure 4, the probabilities of waiting are visualized for convex  $d^{(cv)}$  (left panel) and mixed  $d^{(mx)}$  (right panel) multitasking effects. For  $d^{(cv)}$ , MBF is optimal for every  $n$  according to Theorem 6, as can also be seen in the left panel of Figure 4. For  $d^{(mx)}$ , there is no result for optimality, but LBF gives lower waiting times than MBF in this case. This can be explained by the smaller value of  $d_4^{(mx)} - d_3^{(mx)} = 0.5$  compared with  $d_4^{(mx)}/4 = 11/16$  (see Section 6). The probability of waiting converges again to its heavy-traffic limit, as the differences fall inside the confidence bounds of the simulation. For the speed of convergence, we see roughly similar behavior as for the case of concave  $d^{(cc)}$ . The speed of convergence for MBF is considerably faster for both  $d^{(cv)}$  and  $d^{(mx)}$  than for  $d^{(cc)}$ ; see also Table 1. However, the convergence seems slightly slower for LBF in the case of  $d^{(mx)}$ . We note that the heavy-traffic limits for LBF based on the Garnett delay function are now 0.3069 and 0.1384 for  $d^{(cv)}$  and  $d^{(mx)}$ , respectively. Moreover,

**Table 1.** Deviations in the probability of waiting between smaller systems and its limiting behavior.

$n$ and error	Concave $d^{(cc)}$		Convex $d^{(cv)}$		Mixed $d^{(mx)}$	
	LBF	MBF	LBF	MBF	LBF	MBF
2						
$\Delta_{diff}$	-0.01	0.0264	0.029	0.0093	-0.0176	0.0172
$\Delta_{rel}$ (%)	-9.71	21.56	10.43	3.53	-11.29	10.52
6						
$\Delta_{diff}$	-0.0023	0.0126	0.0098	0.0026	-0.0104	0.0081
$\Delta_{rel}$ (%)	-2.43	9.22	3.29	0.94	-6.98	4.66
10						
$\Delta_{diff}$	-0.0004	0.0091	0.006	0.0012	-0.0068	0.0075
$\Delta_{rel}$ (%)	-0.41	6.55	1.99	0.42	-4.66	4.34
20						
$\Delta_{diff}$	0.0008	0.0061	0.0031	0.0015	-0.0025	0.0046
$\Delta_{rel}$ (%)	0.92	4.29	1.03	0.56	-1.8	2.62
100						
$\Delta_{diff}$	0.0017	0.0023	0.0014	-0.0007	0.0018	0.0027
$\Delta_{rel}$ (%)	1.83	1.61	0.46	-0.24	1.28	1.49



**Figure 4.** (Color online) Probability of waiting for convex  $d^{(cv)}$  (left panel) and mixed  $d^{(mx)}$  (right panel) multitasking effects (LBF is shown as solid lines, and MBF is shown as dashed lines; heavy-traffic (HT) limits are shown as dotted lines).



the expected number of free servers follows a similar pattern as in the right panel of Figure 3 (and are, therefore, omitted).

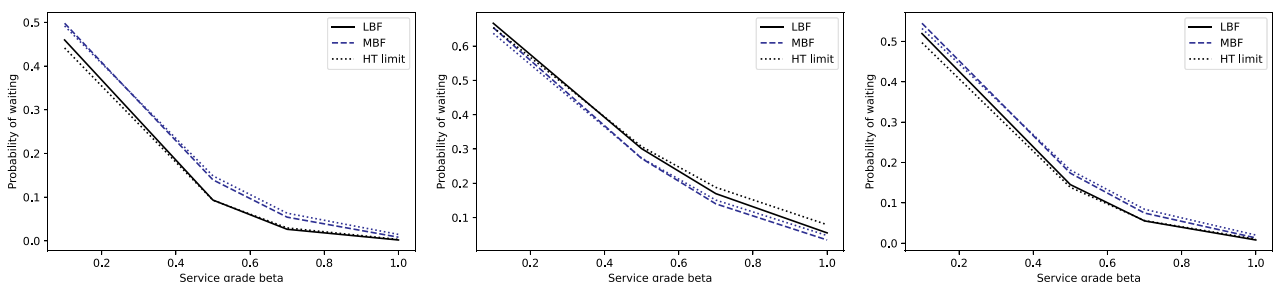
To investigate the sensitivity of the delay probability to the grade of service parameter  $\beta$ , we varied  $\beta \in \{0.1, 0.5, 0.7, 1\}$ . We now only focus on the case  $n = 10$ , as the results indicate that we then start to approach the heavy-traffic limit, but the approximation is clearly not yet excellent, yielding some modest errors. The probabilities of waiting are visualized in Figure 5 for LBF (black solid lines), MBF (blue dashed lines), and the corresponding limits (dotted lines) for  $d^{(cc)}$  (left panel),  $d^{(cv)}$  (center panel), and  $d^{(mx)}$  (right panel). Clearly, the choice of  $\beta$  has a considerable impact on the probability of waiting. Overall, we see that the simulations coincide reasonably well with the approximations, already for  $n = 10$ . For  $\beta = 1$ , the probabilities of waiting and the absolute errors are small (below 0.01 for  $d^{(cc)}$  and  $d^{(mx)}$ ); only  $d^{(cv)}$  gives rise to somewhat larger errors and probabilities. However, the relative errors blow up because of these small probabilities. Albeit  $\beta = 0.1$  may also sometimes provide somewhat larger errors (at most  $-4.34\%$  for  $d^{(mx)}$ ), the relative performance of the approximation typically degrades as  $\beta$  grows beyond 0.5, as expected.

Finally, we consider the impact of  $d_I - d_{I-1}$  and  $d_{I-1} - d_{I-2}$ ; as indicated in Cui and Tezcan [10], the approximation may be poor when the former is small and the latter is large. In particular, we consider two extreme scenarios  $d^{(SL)} = (0, 0.5, 1.5, 3.4, 3.5)$  and  $d^{(LS)} = (0, 0.5, 1.5, 1.6, 3.5)$ . We take  $\beta = 0.5$ ,  $\theta = 0.2$ , and  $n \in \{10, 20, 40\}$ . The simulated probabilities of waiting and the corresponding errors for LBF and MBF can be found Tables 2 and 3 for  $d^{(SL)}$  and  $d^{(LS)}$ , respectively. As expected, the error for  $d^{(SL)}$  is large for LBF, as for the large value of  $d_{I-1} - d_{I-2}$  compared with  $d_I - d_{I-1}$ , the variance becomes large and the convergence becomes slow as noted (Cui and Tezcan [10]). Observe that this problem does not occur for MBF, as the variable part in the service rate (cf. Section 6) is not primarily determined by  $d_I - d_{I-1}$ . Moreover, the heavy-traffic approximations work well in case  $d_I - d_{I-1}$  is large and  $d_{I-1} - d_{I-2}$  is small; see Table 3.

## 11. Extensions and Ramifications

In this section, we discuss two extensions that are rather straightforward to incorporate into the basic model introduced in Section 2.1.

**Figure 5.** (Color online) Probability of waiting for different values of  $\beta$  for concave (left panel), convex (center panel), and mixed (right panel) multitasking effects. HT, heavy traffic.



**Table 2.** Deviations in the probability of waiting for  $d_I - d_{I-1}$  small and  $d_{I-1} - d_{I-2}$  large (i.e.,  $d^{(SL)}$ ).

$n$	LBF			MBF		
	$\hat{\mathbb{P}}(\text{Delay})$	$\Delta_{\text{diff}}$	$\Delta_{\text{rel}} (\%)$	$\hat{\mathbb{P}}(\text{Delay})$	$\Delta_{\text{diff}}$	$\Delta_{\text{rel}} (\%)$
10	0.0886	-0.0872	-98.37	0.1671	-0.0002	-0.10
20	0.0576	-0.0562	-97.49	0.1679	-0.0010	-0.57
40	0.0328	-0.0314	-95.59	0.1686	-0.0016	-0.97

### 11.1. More General Arrival Processes

For the sake of readability, we assumed that arrivals occur according to a Poisson process, but our results remain valid with more general arrival processes. First, the optimality results remain valid, as we did not need to make any assumption regarding the arrival process. Second, the diffusion limits can also be extended using the line of reasoning as explained in Pang et al. [40, section 7.3]. Let  $A_n(\cdot)$  be the arrival process in model  $n$ , and let  $\hat{A}_n(t) = n^{-1/2}(A_n(t) - \lambda_n t)$ , for  $t \geq 0$ , be the associated scaled arrival process. Assume that  $\hat{A}_n(\cdot) \Rightarrow \hat{A}(\cdot)$  in  $D[0, \infty)$ , with  $\hat{A}(\cdot)$  a Brownian motion with drift 0 and variance  $d_I c_a^2$  in the heavy-traffic scaling regime. This holds, for instance, if arrivals occur according to a renewal process where the squared coefficient of variation of the interarrival times is  $c_a^2$ . Under these assumptions, all diffusion limits remain valid where the diffusion coefficient  $\sigma^2(x)$  should be replaced by  $\sigma^2(x) = d_I(1 + c_a^2)$  in Theorems 1–4 and by  $\sigma^2(x) = \lambda(1 + c_a^2)$  in Theorems 7 and 8.

### 11.2. Finite Buffer

Another model extension is to consider the case when there is a maximum to the number of customers that can be in the queue. In this case,  $Q(t) - In \leq \kappa\sqrt{n}$ , for some  $\kappa > 0$ . Clearly, the optimality results carry over. For Theorems 1–4, the results remain valid but with the limiting diffusion being reflected from above at level  $\kappa$ . From a technical point, this can be handled by composing the continuous functions that are already utilized with the continuous function associated with the regulator mapping to the reflection in  $\kappa$  (cf. Pang et al. [40, theorem 7.3]). As there is no queue at the diffusion level in the subcritical regime, Theorems 7 and 8 remain valid.

## 12. Conclusions

In this paper, we studied the interplay of routing decisions, multitasking effects, and preferred staffing levels for a queueing system in which each server can serve multiple customers at once. For a broad class of routing policies and multitasking effects, heavy-traffic diffusion limits are derived for the queue-length process, from which appropriate staffing levels can be determined. Our first main result showed that for a wide class of policies, the limiting diffusion process is identical to that of the least-busy-first routing policy, in which each server is effectively fully or almost fully occupied. The one-dimensional limiting process can also be interpreted in terms of the standard Erlang A model. Our second main result shows that routing to the most busy (nonfull) server first is a policy that achieves a different diffusion limit. In the case that customers cannot be reallocated, the diffusion limit corresponds to a non-standard multidimensional reflected diffusion process with nonhomogeneous drift. For the proof, we constructed an extended version of the regulator mappings associated with generalized Skorokhod problems. We also showed that the two extreme policies of routing to the least busy and most busy server achieve best and worst performance in the case that the total service rate of a server is a concave or convex function of the number of customers in service at that server. The diffusion limits for systems that are not in heavy traffic are somewhat different. The class of policies providing similar behavior as the least-busy-first policy is then considerably smaller; when routing to the most-busy server, there will be empty servers at the fluid scale.

**Table 3.** Deviations in the probability of waiting for  $d_I - d_{I-1}$  large and  $d_{I-1} - d_{I-2}$  small (i.e.,  $d^{(LS)}$ ).

$n$	LBF			MBF		
	$\hat{\mathbb{P}}(\text{Delay})$	$\Delta_{\text{diff}}$	$\Delta_{\text{rel}} (\%)$	$\hat{\mathbb{P}}(\text{Delay})$	$\Delta_{\text{diff}}$	$\Delta_{\text{rel}} (\%)$
10	0.3367	0.0096	2.85	0.2708	0.0024	0.87
20	0.3426	0.0037	1.07	0.2724	0.0008	0.29
40	0.3460	0.0003	0.08	0.2734	-0.0002	-0.08

There are several model refinements possible and directions for future research. First, our primary focus was on diffusion limits in heavy traffic. Randomized policies, such as random-server or power-of- $d$  policies, will have different fluid and diffusion limits than routing to the least busy server in nonheavy-traffic scaling regimes. Depending on the application area, it might be of interest to derive diffusion limits for such randomized policies in the subcritical scaling regime. Another direction of interest is to have a more refined model for the service process. In the current setting, service occurs in a processor sharing fashion, but in practice, service may occur according to an on-off process; see, for example, Véricourt and Jennings [54] and Yankovic and Green [59] for a healthcare example where patients alternate between needing help or not. In fact, Campello et al. [8] study a rich class of models for case managers that incorporate such elements. A topic for further research is to derive diffusion limits for such two-layered systems. Finally, we were able to derive optimal routing of customers for concave and convex service rates. As indicated in Kc [31], the service rate may resemble an inverse U-shaped function. Determining the optimal assignment policy is of interest, along with its diffusion limit. It seems plausible that either routing to the most-busy server is optimal (in case the inverse U ends at its top) or the heavy-traffic limit will be the same as least busy first.

### Appendix A. Proofs SWC

In this section, we present the proofs of Theorems 3 and 4, which are the heavy-traffic diffusion limits in the SWC, for both the MBF and LBF policies. Recall that in the SWC, the state of  $Z_n(\cdot)$  is completely characterized by the state of  $Q_n(\cdot)$ .

**Proof of Theorem 3.** As LBF balances the number of customers among servers, there will only be level  $\lceil Q(t)/n \rceil$  and level  $\lceil Q(t)/n \rceil - 1$  servers, for  $Q(t) < In$ . It may be easily verified that, for  $Q(t) < In$ ,

$$n_{\lceil Q(t)/n \rceil - 1}(t) = n \left\lceil \frac{Q(t)}{n} \right\rceil - Q(t), \quad n_{\lceil Q(t)/n \rceil}(t) = n - n_{\lceil Q(t)/n \rceil - 1}(t), \quad (\text{A.1})$$

with the convention that  $n_{-1}(t) = 0$ . When  $Q_n(t) \geq (I - 1)n$ , there are exclusively level  $I$  and level  $I - 1$  servers, the latter type only when  $Q_n(t) < In$ , and therefore, the rate at which  $Q_n(\cdot)$  jumps down is given by  $d_I n + d_{I-1}(n - n_I) + \theta(Q_n(t) - In)^+$ .

Let us now consider a sequence of processes  $\{R_n^\circ(\cdot)\}_n$ , where  $R_n^\circ(\cdot)$  has the same rates and state space as the process  $R_n(\cdot)$  that we defined above Proposition 5 and  $R_n^\circ(0) = Q_n(0)$ . In addition, we couple  $Q_n(\cdot)$  and  $R_n^\circ(\cdot)$  in such a way that when  $Q_n(t) = R_n^\circ(t) > (I - 1)n$ , then both processes jump at the same time, which is possible because in this case, the infinitesimal rates of both processes are equal. We observe now that  $Q_n(t) = R_n^\circ(t)$ , when  $\inf_{0 \leq s \leq t} R_n^\circ(s) > (I - 1)n$ , for all  $t \geq 0$ .

Now, consider again  $\hat{R}_n^\circ(\cdot)$ , with  $\hat{R}_n^\circ(t) := n^{-1/2}(R_n^\circ(t) - In)$ . Because of Proposition 6,  $\hat{R}_n^\circ(\cdot) \Rightarrow \hat{Q}(\cdot)$ , with  $\hat{Q}(\cdot)$  the diffusion process given in Theorem 1. By continuity of the infimum, we have that  $\inf_{0 \leq s \leq t} \hat{R}_n^\circ(s)$  converges weakly on a process level as well, and therefore, for all  $t \geq 0$ ,

$$\mathbb{P} \left( \inf_{0 \leq s \leq t} \hat{R}_n^\circ(s) < -\sqrt{n} \right) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

This implies that as  $n \rightarrow \infty$ , the uniform distance between  $\hat{R}_n^\circ(\cdot)$  and  $\hat{Q}_n(\cdot)$  goes to zero in probability: that is, for each  $\epsilon > 0$  and all  $T \geq 0$ ,

$$\mathbb{P} \left( \|\hat{Q}_n(\cdot) - \hat{R}_n^\circ(\cdot)\|_T > \epsilon \right) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

We conclude, by Slutsky's lemma, that  $\hat{Q}_n(\cdot) \Rightarrow \hat{Q}(\cdot)$  as well. By continuous mapping and the expressions in (A.1), we obtain weak convergence of  $\hat{Z}_n(\cdot)$  to  $\hat{Z}(\cdot)$ .  $\square$

**Proof of Theorem 4.** Under MBF, there are as many level  $I$  servers as possible in addition to at most one nonempty server that is not at level  $I$ . Writing the state of the  $n$ th system as a function of the number of free spaces  $Q(t) - In$ , we obtain, for  $Q(t) < In$  and  $i \in \{1, \dots, I - 1\}$ ,

$$n_i(t) = n - \left\lceil \frac{Q(t) - In}{I} \right\rceil, \quad n_i(t) = \mathbf{1}_{\{\tilde{g}(Q(t) - In) = i\}}, \quad n_0(t) = \left\lfloor \frac{Q(t) - In}{I} \right\rfloor,$$

where  $i \in \{1, \dots, I - 1\}$  and  $\tilde{g}(x) := \lceil \frac{x}{I} \rceil I - x^-$ . This gives us that

$$\hat{Z}_n(t) = n^{-1/2} \left( Q(t) - In, \left\lceil \frac{Q(t) - In}{I} \right\rceil, \mathbf{1}_{\{\tilde{g}(Q(t) - In) = 1\}}, \dots, \mathbf{1}_{\{\tilde{g}(Q(t) - In) = I - 2\}} \right).$$

By adding and subtracting  $n^{-1/2} \frac{Q(t) - In}{I}$  in the second component, we obtain

$$\hat{Z}_n(t) = n^{-1/2} \left( Q(t) - In, \frac{Q(t) - In}{I}, 0, \dots, 0 \right) + \hat{H}_n(t), \quad (\text{A.2})$$

where

$$\hat{H}_n(t) = n^{-1/2} \left( 0, \left\lfloor \frac{(Q(t) - In)^-}{I} \right\rfloor - \frac{(Q(t) - In)^-}{I}, \mathbb{1}_{\{\bar{g}(Q(t)-In)=1\}}, \dots, \mathbb{1}_{\{\bar{g}(Q(t)-In)=l-2\}} \right).$$

Because the components of  $\hat{H}_n(\cdot)$  are bounded by  $n^{-1/2}$ , we have  $\hat{H}_n(\cdot) \xrightarrow{\mathbb{P}} 0$ , as  $n \rightarrow \infty$ . In addition,  $x \mapsto \frac{x^-}{I}$  is a continuous function. By the mapping theorem and using (A.2), it is sufficient if we show that  $n^{-1/2}(Q(\cdot) - In)$  converges weakly to  $\hat{Q}(\cdot)$ . This last statement is easily proven using the approach in the proof of Proposition 6.  $\square$

### Appendix B. Proof of Stochastic Ordering Results

The proofs in this appendix rely on weak submajorization and weak supermajorization arguments. First, we give the definitions of  $\prec_w$ ,  $\prec^w$ , and  $\prec$ . Let  $x, y$  be two integer-valued  $n$ -dimensional vectors; denote by  $x_{[i]}$  the  $i$ th largest element in  $x$ ; and let  $x_{(i)}$  be the  $i$ th smallest element in  $x$ . Weak submajorization is defined as

$$x \prec_w y \quad \text{if} \quad \sum_{i=1}^k x_{[i]} \leq \sum_{i=1}^k y_{[i]}, \quad k = 1, \dots, n.$$

Informally speaking,  $x \prec_w y$  means that  $x$  is more balanced and smaller than  $y$ . Similarly, weak supermajorization is defined as

$$x \prec^w y \quad \text{if} \quad \sum_{i=1}^k x_{(i)} \geq \sum_{i=1}^k y_{(i)}, \quad k = 1, \dots, n.$$

Furthermore,  $x$  is said to be majorized by  $y$ , denoted as  $x \prec y$  if, next to  $\sum_{i=1}^k x_{[i]} \leq \sum_{i=1}^k y_{[i]}$  for all  $k = 1, \dots, n - 1$ , it also holds that  $\sum_{i=1}^n x_{[i]} = \sum_{i=1}^n y_{[i]}$ . Preservations properties of weak submajorization, weak supermajorization, and majorizations can be found in Akgun et al. [1], Marshall et al. [39, theorem 5.A.1], and Sparaggis et al. [49].

For some routing strategy  $\pi$ , let  $N_{[i]}^\pi(t)$  denote the  $i$ th busiest server, and let  $N_{(i)}^\pi(t)$  be the  $i$ th least busy server,  $i \in \{1, \dots, n\}$ . The results in Theorems 5 and 6 concern the comparison of routing policies, whereas all other parameters (such as the number of servers) remain fixed.

**Proof of Theorem 5.** The proof is by conditioning on the event times and using forward induction. We first consider policies *LBF* and  $\pi$  and only focus on the first stochastic inequality relations in (8) and (9). To this end, we couple the processes, and we will show that

$$N^{LBF}(t) \prec_w N^\pi(t), \tag{B.1}$$

$$Q^{LBF}(t) \leq Q^\pi(t), \tag{B.2}$$

$$L^{LBF}(t) \leq L^\pi(t), \tag{B.3}$$

for all  $t \geq 0$  over all sample paths.

We condition on  $t_k$ , where  $t_k$  are the ordered epochs of arrival times, service completions, and abandonments. The arrival processes are coupled such that arrival epochs in the two systems coincide. The service completions are coupled using the construction in Sparaggis et al. [49], and the coupling can be described as follows. For some policy  $\gamma$ , let  $\mu^\gamma(t) = \sum_{i=1}^n m_{N_{[i]}^\gamma(t)}$  be the total service rate at time  $t$ . After an event has occurred, say at time  $t$ , we schedule a next potential service completion event according to an exponential distribution with rate  $\max\{\mu^{LBF}(t), \mu^\pi(t)\}$ . This potential service completion occurs at the  $l$ th busiest server under policy  $\gamma = LBF, \pi$  if

$$\sum_{i=1}^{l-1} m_{N_{[i]}^\gamma(t)} < \phi(t) \leq \sum_{i=1}^l m_{N_{[i]}^\gamma(t)},$$

where  $\phi(t)$  is drawn uniformly from the interval  $(0, \max\{\mu^{LBF}(t), \mu^\pi(t)\})$ . Finally, we couple each abandonment under policy *LBF* with an abandonment under policy  $\pi$ . Specifically, for abandonments from the queue, the patience of the  $i$ th customers waiting in the queue (if present) is coupled. For abandonments from the server, we follow a similar construction as in Akgun et al. [1, section 3], where customers are labeled at each server from the busiest server to the least busy server. That is, under policy  $\gamma$ , customers  $1, 2, \dots, N_{[1]}^\gamma(t)$  are the customers at the busiest server; then, customers  $N_{[1]}^\gamma(t) + 1, \dots, N_{[1]}^\gamma(t) + N_{[2]}^\gamma(t)$  are the customers at the second busiest server and so on. A potential abandonment of a label  $p$  customer under *LBF* corresponds to a potential abandonment of a label  $p$  customer under  $\pi$ .

Starting the induction is trivial, as (B.1), (B.2), and (B.3) evidently hold for  $t = 0$  because of identical initial conditions. Assume now that the inequality relations hold through  $t = t_k$ . Because the system state does not change for  $t_k \leq t < t_{k+1}$ , it remains to be shown that the inequalities hold for  $t = t_{k+1}$ . We treat the arrival, service completion, abandonment from queue, and abandonment from server events separately.

Case 1 (Arrival). Clearly, the induction hypothesis implies that  $Q^{LBF}(t_{k+1}) \leq Q^\pi(t_{k+1})$  and  $L^{LBF}(t_{k+1}) \leq L^\pi(t_{k+1})$ . If  $N_{[n]}^\pi(t) = I$ , then all servers are occupied at time  $t$  under policy  $\pi$ , clearly providing (B.1). Else, policy *LBF* sends the arrival to the  $n$ th largest queue, yielding (B.1) by invoking Akgun et al. [1, lemma 1] or Sparaggis et al. [49, lemma 2].

Case 2 (Service completion). First note that (B.3) obviously still holds. There are three cases. When a service completion occurs only under policy *LBF*, then the inequality relations clearly hold. Now, suppose that only a service completion occurs under policy  $\pi$ . Observe that this can only happen when  $N_{[n]}^{LBF} < I$ . When  $Q^\pi(t) \geq In$ , then (B.1) and (B.2) hold trivially. For  $Q^\pi(t) < In$ , suppose that the service completion occurs at the  $v$ th busiest server under policy  $\pi$ . Then, for all  $l \geq v$ ,

$$\sum_{i=1}^l m_{N_{[i]}^\pi(t_k)} \geq \sum_{i=1}^v m_{N_{[i]}^\pi(t_k)} \geq \phi(t_k) > \sum_{i=1}^n m_{N_{[i]}^{LBF}(t_k)} \geq \sum_{i=1}^l m_{N_{[i]}^{LBF}(t_k)}. \quad (\text{B.4})$$

For a concave function  $m_i$ , we next verify (as in Sparaggis et al. [49]) that

$$\sum_{i=1}^l m_{N_{[i]}^\pi(t)} > \sum_{i=1}^l m_{N_{[i]}^{LBF}(t)} \rightarrow \sum_{i=1}^l N_{[i]}^\pi(t) > \sum_{i=1}^l N_{[i]}^{LBF}(t). \quad (\text{B.5})$$

To show (B.5), suppose instead that  $\sum_{i=1}^l N_{[i]}^\pi(t) = \sum_{i=1}^l N_{[i]}^{LBF}(t)$ , providing  $(N_{[1]}^{LBF}(t), \dots, N_{[l]}^{LBF}(t)) < (N_{[1]}^\pi(t), \dots, N_{[l]}^\pi(t))$ . By the second part of Marshall et al. [39, theorem 5.A.1], it then holds that  $\sum_{i=1}^l m_{N_{[i]}^{LBF}(t)} \geq \sum_{i=1}^l m_{N_{[i]}^\pi(t)}$ . This contradicts (B.4) and thus, yields (B.5). As  $\sum_{i=1}^l N_{[i]}^\pi(t) > \sum_{i=1}^l N_{[i]}^{LBF}(t)$  for all  $l = v, \dots, n$ , application of Akgun et al. [1, lemma 1] or Sparaggis et al. [49, lemma 2] implies (B.1) at  $t_{k+1}$ . Because all customers are being served in this scenario, (B.2) also follows directly.

Now, suppose that a service completion occurs under both policies. The inequality relations are trivial in case  $Q^\pi(t) \geq In$ . Otherwise, suppose that the service completion is at the  $u$ th busiest server under *LBF* and at the  $v$ th busiest server under  $\pi$ . If  $v \geq u$ , then the equality relations are because of Akgun et al. [1, lemma 1] or Sparaggis et al. [49, lemma 2]. Else, let  $v < u$ . This means that, for  $l = v, \dots, u - 1$ ,

$$\sum_{i=1}^l m_{N_{[i]}^\pi(t_k)} \geq \phi(t_k) > \sum_{i=1}^l m_{N_{[i]}^{LBF}(t_k)}.$$

This implies that because of (B.5),  $\sum_{i=1}^l N_{[i]}^\pi(t) > \sum_{i=1}^l N_{[i]}^{LBF}(t)$ , for  $l = v, \dots, u - 1$ . Akgun et al. [1, corollary 1] yields (B.1) at  $t_{k+1}$ . Again, (B.2) follows directly as all customers are being served.

Case 3 (Abandonment from queue). Because  $Q^{LBF}(t) \leq Q^\pi(t)$  and the coupling of abandonments, it holds that no abandonment occurs (e.g., when  $Q^\pi \leq In$ ), abandonments occur under both policies, or an abandonment occurs only under policy  $\pi$ . For the first two cases, the equality relations are trivially satisfied. In the third case, if a customer abandons only under policy  $\pi$ , then  $Q^{LBF}(t) < Q^\pi(t)$ , and it follows directly that  $Q^{LBF}(t_{k+1}) = Q^{LBF}(t) \leq Q^\pi(t) - 1 = Q^\pi(t_{k+1})$ . Equations (B.1) and (B.3) are evident.

Case 4 (Abandonment from server). Because  $N_{[n]}^{LBF}(t) \leq N_{[n]}^\pi(t)$  and the coupling of abandonments from service, it holds that no abandonment occurs, abandonments occur under both policies, or an abandonment occurs only under policy  $\pi$  such that (B.3) is satisfied. The first case is trivial. For the second case, suppose that the abandonment is at the  $u$ th busiest server under *LBF* and at the  $v$ th busiest server under  $\pi$ . The case  $u = v$  follows again from Akgun et al. [1, lemma 1] or Sparaggis et al. [49, lemma 2]. If  $v < u$ , then because of the labeling, it holds that  $\sum_{i=1}^l N_{[i]}^{LBF}(t) < \sum_{i=1}^l N_{[i]}^\pi(t)$ , for all  $v \leq l < u$ . Hence, for  $v \leq l < u$ ,

$$\sum_{i=1}^l N_{[i]}^{LBF}(t_{k+1}) = \sum_{i=1}^l N_{[i]}^{LBF}(t) \leq \sum_{i=1}^l N_{[i]}^\pi(t) - 1 = \sum_{i=1}^l N_{[i]}^\pi(t_{k+1}).$$

Equations (B.1) and (B.2) are then evident. The third case follows from similar arguments, where the labeling yields that  $N_{[n]}^{LBF}(t) < N_{[n]}^\pi(t)$ .

Removal of conditioning on arrival times, service completions, and abandonments completes the proof of the first stochastic inequality relations in (8) and (9).

Now, consider policies  $\pi$  and *MBF* and the second stochastic inequality relations in (8) and (9). Using coupling and forward induction on event times again, we show that  $N^\pi(t) \prec_w N^{MBF}(t)$ ,  $Q^\pi(t) \leq Q^{MBF}(t)$ , and  $L^\pi(t) \leq L^{MBF}(t)$  for all  $t \geq 0$  over all sample paths. The events of service completions and abandonments are handled exactly as described for policies *LBF* and  $\pi$ . If the next event is an arrival and  $Q^{MBF}(t_k) \geq In$ , then the inequality relations are also clearly satisfied. Now, suppose that the next event is an arrival at the  $u$ th busiest server under policy  $\pi$  and at the  $v$ th busiest server under policy *MBF*. If  $u \geq v$ , then the inequalities are because of Akgun et al. [1, lemma 1] or Sparaggis et al. [49, lemma 2]. Else, let  $u < v$ . Under *MBF*, we label the queues such that  $N_{[v-1]}^{MBF}(t) = I$ . Let  $u' \leq u$  be such that  $N_{[u']}^\pi(t) = N_{[u'+1]}^\pi(t) = \dots = N_{[u]}^\pi(t)$  and either  $u' = 1$  or  $N_{[u'-1]}^\pi(t) > N_{[u']}^\pi(t)$ . Observe that because of the labeling and the *MBF* policy,  $\sum_{i=1}^l N_{[i]}^\pi(t) < \sum_{i=1}^l N_{[i]}^{MBF}(t)$  for  $l = u', \dots, v - 1$ . Hence, for  $l = u', \dots, v - 1$ , we have

$$\sum_{i=1}^l N_{[i]}^\pi(t_{k+1}) = \sum_{i=1}^l N_{[i]}^\pi(t) + 1 \leq \sum_{i=1}^l N_{[i]}^{MBF}(t) = \sum_{i=1}^l N_{[i]}^{MBF}(t_{k+1}).$$



For  $l = 1, \dots, u' - 1$  and  $l = v, \dots, n$ , we have

$$\sum_{i=1}^l N_{[i]}^{\pi}(t_{k+1}) = \sum_{i=1}^l N_{[i]}^{\pi}(t) + \mathbf{1}_{\{l \geq v\}} \leq \sum_{i=1}^l N_{[i]}^{MBF}(t) + \mathbf{1}_{\{l \geq v\}} = \sum_{i=1}^l N_{[i]}^{MBF}(t_{k+1}).$$

This yields  $N^{\pi}(t_{k+1}) \prec_w N^{MBF}(t_{k+1})$ . Removal of the conditioning on event times completes the proof.  $\square$

**Proof of Theorem 6.** We use the same sample path approach as in Theorem 5 and follow the lines of Sparaggis et al. [49] again (also showing large similarity with the proof of Theorem 5). We consider the relation between  $\pi$  and  $MBF$ ; similar arguments can be used for the relation between  $\pi$  and  $LBF$ . By coupling the processes, we will show that

$$N^{\pi}(t) \prec^w N^{MBF}(t), \quad (B.6)$$

$$Q^{\pi}(t) \geq Q^{MBF}(t), \quad (B.7)$$

$$L^{\pi}(t) \geq L^{MBF}(t) \quad (B.8)$$

for all  $t \geq 0$  over all sample paths. Again, the relations will be established at time  $t_{k+1}$ , assuming that they hold at time  $t_k$ , where  $t_k$  is the  $k$ th event time on the considered sample path. Arrivals and abandonments from the queue are coupled as in the proof of Theorem 5. The construction of service completion events is slightly modified and follows Sparaggis et al. [49]. In particular, after an event at time  $t$ , the next potential service completion is scheduled after an exponential time with rate  $\max\{\mu^{\pi}(t), \mu^{MBF}(t)\}$ . This potential service completion occurs at the  $l$ th least busy server under policy  $\gamma = \pi, MBF$ , if  $\sum_{i=1}^{l-1} m_{N_{(i)}^{\gamma}(t)} < \psi(t) \leq \sum_{i=1}^l m_{N_{(i)}^{\gamma}(t)}$ , where  $\psi(t)$  is drawn uniformly from the interval  $(0, \max\{\mu^{\pi}(t), \mu^{MBF}(t)\}]$ . For abandonments from the server, we label customers at each server from the least busy server to the busiest server. We treat the arrival, service completion, and abandonment (from the queue or the server) events separately.

Case 1 (Arrival). When  $N_{(1)}^{\pi}(t) = l$ , all servers are fully occupied at time  $t$  under policy  $\pi$ , and the inequality relations are clearly satisfied. Otherwise, suppose that the next event is an arrival at the  $u$ th least busy server under policy  $\pi$  and at the  $v$ th least busy server under policy  $MBF$ . If  $u \leq v$ , then the inequalities follow from Akgun et al. [1, lemma 2] or Sparaggis et al. [49, lemma 1] (using that  $x_{(i)} = x_{[n+1-i]}$ ). So, suppose that  $u > v$ , and label the servers such that for  $MBF$ , it holds that  $N_{(v+1)}^{MBF}(t) = l$ . Let  $u' \geq u$  be such that  $N_{(u)}^{\pi}(t) = N_{(u+1)}^{\pi}(t) = \dots = N_{(u')}^{\pi}(t)$  and  $N_{(u'+1)}^{\pi}(t) > N_{(u')}^{\pi}(t)$  or  $u' = n$ . Because of the  $MBF$  policy, it follows that  $\sum_{i=1}^n N_{(i)}^{\pi}(t) < \sum_{i=1}^n N_{(i)}^{MBF}(t)$ , for  $l = v + 1, \dots, u'$ . As  $\sum_{i=1}^n N_{(i)}^{\pi}(t) \geq \sum_{i=1}^n N_{(i)}^{MBF}(t)$ , we have  $\sum_{i=1}^l N_{(i)}^{\pi}(t) > \sum_{i=1}^l N_{(i)}^{MBF}(t)$  for  $l = v, \dots, u' - 1$ . Hence, as in the proof of Theorem 5, we obtain, for  $l = v, \dots, u' - 1$ ,

$$\sum_{i=1}^l N_{(i)}^{\pi}(t_{k+1}) = \sum_{i=1}^l N_{(i)}^{\pi}(t) \geq \sum_{i=1}^l N_{(i)}^{MBF}(t) + 1 = \sum_{i=1}^l N_{(i)}^{MBF}(t_{k+1}).$$

For  $l = 1, \dots, v - 1$  and  $l = u', \dots, n$ , we have

$$\sum_{i=1}^l N_{(i)}^{\pi}(t_{k+1}) = \sum_{i=1}^l N_{(i)}^{\pi}(t) + \mathbf{1}_{\{l \geq u'\}} \geq \sum_{i=1}^l N_{(i)}^{MBF}(t) + \mathbf{1}_{\{l \geq u'\}} = \sum_{i=1}^l N_{(i)}^{MBF}(t_{k+1}).$$

This yields  $N^{\pi}(t_{k+1}) \prec_w N^{MBF}(t_{k+1})$ . The other two inequalities are straightforward.

Case 2 (Service completion). There are three cases again. First, when there is only a service completion under policy  $MBF$ , then the relations clearly hold. Second, suppose there is only a service completion under policy  $\pi$  at the  $v$ th least busy server. Observe that having a service completion only under  $\pi$  means  $N_{(1)}^{MBF}(t) < l$  so that the case  $Q^{\pi} \geq ln$  is trivial again. From the coupling we have, for  $l \geq v$ ,

$$\sum_{i=1}^l m_{N_{(i)}^{\pi}(t_k)} \geq \sum_{i=1}^v m_{N_{(i)}^{\pi}(t_k)} \geq \psi(t_k) > \sum_{i=1}^n m_{N_{(i)}^{MBF}(t_k)} \geq \sum_{i=1}^l m_{N_{(i)}^{MBF}(t_k)}.$$

Similar to (B.5), this implies  $\sum_{i=1}^l N_{(i)}^{\pi}(t) > \sum_{i=1}^l N_{(i)}^{MBF}(t)$  for all  $l \geq v$ , where we now use the convexity of  $d$  and the first part of Marshall et al. [39, theorem 5.A.1]. Application of Akgun et al. [1, lemma 2] or Sparaggis et al. [49, lemma 1] provides the inequality relations for this case. Third, the case where a service completion occurs under both  $\pi$  and  $MBF$  follows similar arguments and the corresponding case in Theorem 5.

Cases 3 and 4 (Abandonment). This is similar to the arguments for cases 3 and 4 in Theorem 5.

Removal of conditioning on event times finishes the proof.  $\square$

## Appendix C. Proofs of the Subcritical Regime

In this section, we prove the results in Section 7.3. The results strongly rely on variants of the integral representation and continuous mapping result in Proposition 10; we use a variant for the fluid limit result and a different variant for the diffusion limit result. To start this section, we summarize the result for the fluid limit case.

Let  $\bar{f} : \mathbb{R}^I \rightarrow \mathbb{R}^I$  be the function defined by

$$\begin{aligned} \bar{f}(n^{-1}Z_n(t)) := & e_Q \left[ \lambda - \frac{d_I n_I(s)}{n} - \theta \left( \frac{Q_n(s)}{n} - I \right)^+ \right] \\ & + \sum_{i=1}^{I-2} (-e_Q - e_{n_i} + e_{n_{i+1}}) d_i \frac{n_i(t)}{n} + (-e_Q + e_{I-2}) d_{I-1} \frac{n_{I-1}(t)}{n}. \end{aligned}$$

Furthermore, for  $n \in \mathbb{N}$ , define  $\bar{X}_n(\cdot)$  by

$$\begin{aligned} \bar{X}_n(t) := & e_Q [M_{A_n}(t) - M_{D_n}(t) - M_{D_n^c}(t)] + \sum_{i=1}^{I-2} (-e_Q - e_{n_i} + e_{n_{i+1}}) M_{D_{n,i}}(t) \\ & + (-e_Q + e_{I-2}) M_{D_{n,I-1}}(t), \end{aligned} \tag{C.1}$$

with  $M_{A_n}(\cdot)$ ,  $M_{D_n}(\cdot)$ ,  $M_{D_n^c}(\cdot)$ , and  $M_{D_{n,i}}(\cdot)$  the martingales obtained by compensating the counting processes  $A_n(\cdot)$ ,  $D_n(\cdot)$ ,  $D_n^c(\cdot)$ , and  $D_{n,i}(\cdot)$ ,  $i = 1, \dots, I-1$  defined during step 1 in Section 9. Observe that  $\bar{X}(\cdot) = \sqrt{n} \bar{X}_n(\cdot)$ , with the latter process defined in (33), the difference being that we do not decompose  $D_n(\cdot)$  in  $\bar{X}_n(\cdot)$ . Now, it follows from the construction leading to Proposition 10 that for every  $n$ ,  $n^{-1}Z_n(\cdot)$  satisfies the integral equation

$$n^{-1}Z_n(t) = n^{-1}Z_n(0) + n^{-1}\bar{X}_n(t) + \int_0^t \bar{f}(n^{-1}Z_n(s)) \, ds + \begin{pmatrix} 0 & 0 \\ 0 & \bar{R} \end{pmatrix} n^{-1}L_n(t) - e_Q \frac{\lambda \beta \sqrt{nt}}{n},$$

where  $n^{-1}L_n(\cdot) = n^{-1}(L_{n,0}(\cdot), \dots, L_{n,I-1}(\cdot))^\top$  is the unique process defined through

$$L_{n,i}(t) = \int_0^t \mathbb{1}_{\{\sum_{k=1}^{I-1} n_{n,k}(s) = 0, Q_n(s) < In\}} \, dA_n(s) \quad i = 1, \dots, I-1,$$

and  $L_{n,0}(\cdot) \equiv 0$ . Furthermore, for every  $n$ , the pair  $(n^{-1}Z_n(\cdot) - n^{-1}Z_n(0), n^{-1}L_n(\cdot))$  is the image of  $n^{-1}\bar{X}_n(\cdot) - n^{-1}(\lambda \beta \sqrt{nt})$  under a Lipschitz continuous map from  $D^I[0, \infty)$  to  $D^I[0, \infty) \times D^I[0, \infty)$ . With these results, we are set to prove Proposition 3.

**Proof of Proposition 3.** For the proof, we first establish the convergence result along a subsequence  $\{n_k\}_k$  on a set  $\Omega'$  of probability 1. Afterward, we show that any limit corresponding to such a subsequence satisfies (18)–(25) and is Lipschitz.

Note that (by assumption)  $n^{-1}Z_n(0) \rightarrow \bar{Z}(0)$  almost surely and that the function  $t \mapsto \lambda \beta \sqrt{nt}/n$  converges to zero uniformly on compacts for every  $\omega$ . Therefore, to establish the convergence result, it is sufficient to show that  $n^{-1}X_n(\cdot)$  converges almost surely along a subsequence because convergence of  $n^{-1}Z_n(\cdot)$  along the same subsequence follows by continuous mapping.

Now, we show that  $n^{-1}X_n(\cdot)$  converges to zero in probability using the martingale FCLT. Observe that  $n^{-1}X_n(\cdot)$  is a locally square integrable martingale for each  $n$  and that its jumps vanish with probability 1 as  $n \rightarrow \infty$ . Moreover, the quadratic variation of  $n^{-1}X_n(\cdot)$  converges to zero uniform on compact sets almost surely because its components are increasing and converge to zero pointwise for every  $t \geq 0$ . To be precise, it suffices to show that the diagonal terms of the quadratic variation process go to zero (cf. step 4 of Section 9), and for each such component  $\langle n^{-1}X_n^{(i,i)} \rangle_t$  with  $i = 1, \dots, I$ , we have

$$\begin{aligned} \langle n^{-1}X_n^{(i,i)} \rangle_t & \leq \frac{1}{n^2} \int_0^t \lambda_n - d_I n_I(s) \, ds + \frac{1}{n^2} \int_0^t \theta(Q_n(s) - In)^+ \, ds + \frac{1}{n^2} \cdot d_I \int_0^t \sum_{i=1}^{I-1} n_{n,i}(s) \, ds \\ & \leq o(1) + \frac{1}{n^2} \int_0^t \theta(Q_n(0) + A_n(s)) \, ds + \frac{1}{n} \cdot d_I \int_0^t \frac{n}{n} \, ds \xrightarrow{\text{a.s.}} 0, \quad \text{as } n \rightarrow \infty, \end{aligned}$$

where we used that  $A_n(s)/n \rightarrow \lambda s$  almost surely for every  $s \geq 0$ ,  $\{Q_n(0)/n\}_n$  is a convergent sequence, and we apply dominated convergence. This shows that the sequence  $\{n^{-1}X_n(\cdot)\}_n$  satisfies the conditions for the martingale FCLT; hence,  $n^{-1}X_n(\cdot) \Rightarrow 0$  as  $n \rightarrow \infty$ , which implies that  $n^{-1}Z_n(\cdot) \xrightarrow{\mathbb{P}} \bar{Z}(0)$  as  $n \rightarrow \infty$ . Consequently, any subsequence of  $\{(n_k)^{-1}X_{n_k}(\cdot)\}_k$  also converges to zero in probability, and hence, it has a further subsequence  $\{(n_k)^{-1}X_{n_k}(\cdot)\}_k$  that converges to zero almost surely, which proves the convergence claim.

We proceed by showing that any limit  $(\bar{Z}(\cdot), \bar{L}(\cdot))$  satisfies (18)–(25). First, observe that any such limit satisfies

$$\bar{Z}(t) = \bar{Z}(0) + 0 + \int_0^t \bar{f}(\bar{Z}(s)) \, ds + \begin{pmatrix} 0 & 0 \\ 0 & \bar{R} \end{pmatrix} \bar{L}(t)$$

because we used the continuous mapping theorem, where  $\bar{L}(\cdot) = (\bar{L}_0(\cdot), \dots, \bar{L}_{I-1}(\cdot))^\top$ , with  $\bar{L}_0(\cdot) \equiv 0$  because of uniform convergence. Therefore, (18)–(20) immediately follow. Furthermore, (21) and (22) follow because  $n_{n,I-1}(\cdot)$  and  $n_{n,i}(\cdot)$  are continuous in  $(Z_n(\cdot), Z_n(0), L_n(\cdot))$ . Also,  $Q_n(\cdot), n_{n,i}(\cdot) \geq 0$  for all  $n$  and  $i$ , which establishes (23) because of uniform convergence. In a similar way, (25) follows from the conditions on  $L_n(\cdot)$  for every  $n$ . Finally, observe that  $\bar{L}(\cdot)$  in (24) satisfies conditions (a)–(c) from the mapping  $\Phi \circ M$  used in Proposition 10. Because of (18)–(23), such a process is unique, and therefore, any limit  $\bar{L}(\cdot)$  should satisfy (24).

To finish the proof, we show that any solution  $(\bar{Z}(\cdot), \bar{L}(\cdot))$  to (18)–(25) has Lipschitz continuous paths. First,  $\bar{L}(\cdot)$  is Lipschitz because its components are integrals over uniformly bounded functions. Then, by (22),  $\bar{n}_i(s) \leq 1$  for all  $i$ . Hence, the functions  $t \mapsto \int_0^t d_i \bar{n}_i(s) \, ds$  are also Lipschitz, and thus,  $\bar{n}_i(\cdot)$  is Lipschitz for every  $i \in \{0, \dots, I-2\}$  because they are a linear combination of

Lipschitz functions. Now, observe that the remaining functions  $\bar{n}_{l-1}(\cdot)$  and  $\bar{n}_l(\cdot)$  are Lipschitz if  $\bar{Q}(\cdot)$  is Lipschitz. The latter is true because for any initial condition  $\bar{Q}(0)$ ,  $\bar{Q}(\cdot)$  is locally Lipschitz. The term  $\int_0^t (\bar{Q}(s) - I)^+ ds$  has a theoretically unbounded integrand. However, because  $\bar{Q}(\cdot)$  is differentiable almost everywhere on  $[0, \epsilon)$  for some  $\epsilon > 0$ , we know that when  $\bar{Q}(0) > I$ , its derivative is strictly negative. Consequently, the term  $(\bar{Q}(s) - I)^+$  is uniformly bounded given  $\bar{Q}(0)$  and  $\bar{Q}(\cdot)$  is Lipschitz.

We proceed by proving Proposition 4, for which we use the following lemma (Dai and Weiss [11, lemma 2.2]). Denote  $\dot{f}(t)$  for the derivative of a function  $f$  on  $\mathbb{R}_+$  that is differentiable at  $t \in \mathbb{R}_+$ .

**Lemma C.1.** *If  $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is Lipschitz continuous, then at every point  $t \in \mathbb{R}_+$  where  $\dot{f}(t)$  exists,  $f(t) = 0$  implies  $\dot{f}(t) = 0$ .*

**Proof of Proposition 4.** The proof is divided in two parts. In the first part, we prove that every solution to (18)–(25) has a unique fixed point  $\bar{Z}$ , and in the second part, we prove that under the additional assumption  $n^{-1}Z_n(0) \xrightarrow{a.s.} \bar{Z}$ , the sequence  $\{n^{-1}Z_n(\cdot)\}_n$  converges to this fixed point almost surely.

To begin the first part, recall that because  $\bar{Z}(\cdot)$  and  $\bar{L}(\cdot)$  are Lipschitz, they are differentiable almost everywhere. For the remainder of this first part of the proof, we consider only points  $t \in [0, \infty)$ , where both are differentiable. By plugging  $\bar{Z}$  into (18)–(25), it can be checked that it is indeed a fixed point. Therefore, we focus on showing that it is unique by showing that every other value of  $\bar{Z}(\cdot)$  has nonzero derivative.

First, observe that if  $\bar{Q}(t) > I$ , in which case  $\bar{n}_l(t) = 1$  by (21)–(23) and by (18), we have

$$\dot{\bar{Q}}(t) = -d_l + \lambda - \theta(\bar{Q}(t) - I)^+ < 0,$$

every fixed point should satisfy  $\bar{Q}(t) \leq I$  for all  $t$ .

We now claim that  $\bar{n}_{k+1}(t) > 0$  implies  $\bar{n}_i(t) > 0$  for every  $i \in \{1, \dots, k\}$ . To establish this claim, suppose the converse, namely that  $\bar{n}_1(t) + \dots + \bar{n}_k(t) = 0$ ,  $\bar{n}_{k+1}(t) > 0$  for some  $k + 1 \leq l - 1$ . Then, using Lemma C.1 and (20),

$$0 = \bar{n}_k(t) = d_{k+1}\bar{n}_{k+1}(t) - d_k\bar{n}_k(t) - \dot{\bar{L}}_{k+1}(t) + \dot{\bar{L}}_k(t) - \dot{\bar{L}}_{k-1}(t) = d_{k+1}\bar{n}_{k+1}(t) - d_k\bar{n}_k(t),$$

where we use that (24) with  $\bar{n}_{k+1}(t) > 0$  implies that  $\dot{\bar{L}}_i(t) = 0$  for  $i \leq k + 1$ . However,  $d_{k+1}\bar{n}_{k+1}(t) - d_k\bar{n}_k(t) = d_{k+1}\bar{n}_{k+1}(t) > 0$ , which is a contradiction.

We use this intermediate result to show that any fixed point satisfies  $\bar{n}_0(t) = \bar{n}_l(t) = 1$ . To this end, consider the function  $f(t) = \sum_{i=1}^l \bar{n}_i(t)$ . Using (19) and (22), we have

$$\dot{f}(t) = -\dot{\bar{n}}_0(t) = -d_1\bar{n}_1(t) + \dot{\bar{L}}_1(t).$$

Now, by our claim,  $\sum_{i=1}^{l-1} \bar{n}_i(t) > 0$  implies that  $\bar{n}_1(t) > 0$ , which implies  $\dot{f}(t) < 0$ . Consequently, any invariant state should satisfy  $\sum_{i=1}^{l-1} \bar{n}_i(t) = 0$ , or equivalently,  $\bar{n}_0(t) + \bar{n}_l(t) = 1$ .

Finally, if  $\bar{Q}(t) \leq I$  and  $\bar{n}_0(t) + \bar{n}_l(t) = 1$ , using again (18), we find that

$$\dot{\bar{Q}}(t) = \lambda - d_l\bar{n}_l(t).$$

Hence,  $\dot{\bar{Q}}(t) < 0$  if  $\bar{n}_l(t) > \lambda/d_l$  and  $\dot{\bar{Q}}(t) > 0$  if  $\bar{n}_l(t) < \lambda/d_l$ . Concluding, any invariant state should satisfy  $\bar{n}_0(t) + \bar{n}_l(t) = 1$  and  $\bar{n}_l(t) = \lambda/d_l$ , which is the vector  $\bar{Z}$ .

Now, we focus on the second part of the proof. Note that because  $\bar{Z}$  is a fixed point, any solution  $(\bar{Z}(\cdot), \bar{L}(\cdot))^T$  to (18)–(25) with initial value  $(\bar{Z}, 0)^T$  should satisfy  $\bar{Z}(t) = \bar{Z}$  for all  $t \geq 0$ . By assumption,  $\{n^{-1}Z_n(0)\}_n$  converges almost surely to  $\bar{Z}$ , and by Proposition 3, every subsequence of  $\{n^{-1}Z_n(\cdot)\}_n$  has a further subsequence converging to a solution of (18)–(25), which should then be the constant process  $\bar{Z}$ . Consequently, by the almost sure convergence variant of Billingsley [5, theorem 2.6], the sequence  $\{n^{-1}Z_n(\cdot)\}_n$  converges almost surely to  $\bar{Z}$ , uniformly on compact time intervals.

We finish this section with the proof of Theorem 8, for which we rely again heavily on a variant of the continuous mapping construction in and leading up to Proposition 10. For sake of brevity, the details of this result for the subcritical diffusion scaling in the MBF case are summarized at the beginning of the proof.  $\square$

**Proof of Theorem 8.** The proof follows the same lines as that of Theorem 2. From the construction in Section 9, it can be shown that for every  $n$ ,

$$\hat{Z}_n(t) = \hat{Z}_n(0) + \hat{X}_n(t) - n^{-1/2}D_n^\circ(t) + \int_0^t f^\circ(\hat{Z}_n(s)) ds + \begin{pmatrix} 0 & 0 \\ 0 & \tilde{R} \end{pmatrix} \hat{L}_n(t),$$

with the pair  $(\hat{Z}_n(\cdot), \hat{L}_n(\cdot))$  being the image of  $\hat{X}_n(\cdot)$  under a continuous mapping and  $\hat{X}_n(\cdot)$  the martingale given by

$$\hat{X}_n(t) = \frac{1}{\sqrt{n}} \left[ e_Q[M_{A_n}(t) - M_{D_n}(t)] + \sum_{i=1}^{l-2} (-e_Q - e_{n_i} + e_{n_{i-1}})M_{D_{n,i}}(t) + (-e_Q + e_{l-2})M_{D_{n,l-1}}(t) \right] \quad (C.2)$$

for the definitions given at the beginning of this section and in Section 7. Therefore, to prove the result, it is sufficient to show that  $\hat{X}_n(\cdot) \Rightarrow B(\cdot)$  and that  $n^{-1/2}D_n^\circ(\cdot) \rightarrow 0$  as  $n \rightarrow \infty$ . We begin with the latter.

Observe that  $n_{n,i}(t) \geq \lambda n/d_i$  when  $Q_n(t) \geq n(l-1) + \lceil \lambda n/d_i \rceil$ . Therefore,  $Q_n(t)$  has a death rate of at least  $\lambda n$  in that case. Consider the birth-death process  $Q'_n(\cdot)$  with initial condition  $Q_n(0)$ , birth rate  $\lambda_n$ , and death rate equal to that of  $Q_n(\cdot)$  when  $Q'_n(t) \leq n(l-1) + \lambda/d_i n$  and equal to  $\lambda n$  otherwise. Moreover,  $Q_n(\cdot)$  and  $Q'_n(\cdot)$  are coupled such that all jumps upward are at identical

times for both processes and jumps downward are at identical times when  $Q'_n(t) \leq n(I-1) + \lambda/d_I n$ . With this construction,  $Q'_n(t) \geq Q_n(t)$  for all  $t \geq 0$ . Moreover, given that  $Q'_n(\cdot) \geq n(I-1) + \lambda/d_I$ , we have that  $Q'_n(\cdot)$  is a birth-death process with birth rate  $\lambda_n$ , which is strictly smaller than its death rate  $\lambda n$ . Hence, a similar argument as the one we used in Lemma 1 for the  $S_n(\cdot)$  shows that  $(Q_n(\cdot) - In)^+ \xrightarrow{\mathbb{P}} 0$  as  $n \rightarrow \infty$ . Consequently,  $n^{-1/2}D_n^o(\cdot) \xrightarrow{\mathbb{P}} 0$  as  $n \rightarrow \infty$ .

To show that  $\hat{X}_n(\cdot) \Rightarrow B(\cdot)$ , we apply the martingale FCLT to the sequence  $\{n^{-1/2}X_n(\cdot)\}_n$ . For each  $n$ ,  $\hat{X}_n(\cdot)$  is square integrable, and its jumps vanish with probability 1 as  $n \rightarrow \infty$ . Thus, if we can establish the correct limit for  $\langle \hat{X}_n \rangle_t$ , we obtain weak convergence of the sequence  $\{n^{-1/2}X_n(\cdot)\}_n$  to the Brownian term in Theorem 8, and the proof is complete.  $\square$

In the remainder, we show that  $\langle \hat{X}_n \rangle_t \rightarrow 2\lambda te_Q^T$  in probability for all  $t \geq 0$  as  $n \rightarrow \infty$  (i.e., except for the quadratic variation of  $\hat{Q}_n(\cdot)$ , all quadratic (co-)variation processes will vanish in probability as  $n \rightarrow \infty$ ). Denoting the matrix  $\langle \hat{X}_n \rangle_t$  by  $\langle \hat{X}_n^{(i,j)} \rangle_t$ , for  $i, j = 1, \dots, I$ , we use again that, by Cauchy–Schwarz, we have

$$|\langle \hat{X}_n^{(i,j)} \rangle_t|^2 \leq \langle \hat{X}_n^{(i,i)} \rangle_t \langle \hat{X}_n^{(j,j)} \rangle_t, \quad t \geq 0.$$

Consequently, if we can show  $\langle \hat{X}_n^{(j,j)} \rangle_t \rightarrow 0$  in probability for every  $t \geq 0$  and  $j \geq 2$ , together with  $\langle \hat{X}_n^{(1,1)} \rangle_t \rightarrow 2\lambda t$  in probability for every  $t \geq 0$ , then we are done.

Consider  $\langle \hat{X}_n^{(1,1)} \rangle_t$  first. Focusing on the terms with factor  $e_Q$  in (C.2), we have the lower bound

$$\langle \hat{X}_n^{(1,1)} \rangle_t \geq \frac{1}{n} \int_0^t \lambda_n + d_I n_I(s) \, ds = \int_0^t \lambda + d_I n^{-1} n_I(s) - \lambda \beta n^{-1/2} \, ds.$$

From Proposition 4, we have that  $n^{-1}n_I(s) \rightarrow \lambda/d_I$  almost surely for every  $s \geq 0$  as  $n \rightarrow \infty$ . Hence,  $\lambda_n + d_I n_I(s) \rightarrow 2\lambda$  for every  $s \geq 0$  as  $n \rightarrow \infty$ , and by dominated convergence, we have  $\frac{1}{n} \int_0^t \lambda_n + d_I n_I(s) \, ds \rightarrow 2\lambda t$  almost surely as  $n \rightarrow \infty$ .

An upper bound is given by

$$\langle \hat{X}_n^{(1,1)} \rangle_t \leq \frac{1}{n} \int_0^t \lambda_n + d_I n_I(s) \, ds + 3d_I \sum_{i=1}^{I-1} \int_0^t n^{-1} n_i(s) \, ds.$$

Using again Proposition 4, we have that  $n^{-1}n_i(s)$  converges to zero almost surely for each  $i \in \{1, \dots, I-1\}$ . Dominated convergence then gives the same limit for the upper bound as for the lower bound (with convergence in probability). Combining the upper and lower bounds gives us the desired conclusion that  $\langle \hat{X}_n^{(1,1)} \rangle_t$  converges in probability to  $2\lambda t$ .

The argument also shows the desired claim for  $\langle \hat{X}_n^{(j,j)} \rangle_t$  with  $j \geq 2$  because we have the upper bound

$$\langle \hat{X}_n^{(j,j)} \rangle_t \leq 3d_I \sum_{i=1}^{I-1} \int_0^t n^{-1} n_i(s) \, ds, \quad j = 2, \dots, I,$$

and we have shown the right-hand side to converge to zero in probability as  $n \rightarrow \infty$ .

## References

- [1] Akgun OT, Righter R, Wolff R (2011) Multiple-server system with flexible arrivals. *Adv. Appl. Probab.* 43(4):985–1004.
- [2] Altman E, Ayesta U, Prabhu BJ (2011) Load balancing in processor sharing systems. *Telecomm. Systems* 47(1–2):35–48.
- [3] Berry Jaeger JA, Tucker AL (2017) Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Sci.* 63(4):1042–1062.
- [4] Bertrand JM, Van Ooijen H (2002) Workload based order release and productivity: A missing link. *Production Planning Control* 13(7):665–678.
- [5] Billingsley P (2013) *Convergence of Probability Measures* (John Wiley & Sons, New York).
- [6] Borst S, Mandelbaum A, Reiman MI (2004) Dimensioning large call centers. *Oper. Res.* 52(1):17–34.
- [7] Browne S, Whitt W (1995) Piecewise-linear diffusion processes. Dshalalow J, ed. *Advances in Queueing: Theory, Methods, and Open Problems* (CRC Press, Boca Raton), 463–480.
- [8] Campello F, Ingolfsson A, Shumsky RA (2017) Queueing models of case managers. *Management Sci.* 63(3):882–900.
- [9] Chen H, Mandelbaum A (1991) Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *Ann. Probab.* 19(4):1463–1519.
- [10] Cui L, Tezcan T (2016) Approximations for chat service systems using many-server diffusion limits. *Math. Oper. Res.* 41(3):775–807.
- [11] Dai JG, Weiss G (1996) Stability and instability of fluid models for reentrant lines. *Math. Oper. Res.* 21(1):115–134.
- [12] Dai JG, He S, Tezcan T (2010) Many-server diffusion limits for G/Ph/n+ GI queues. *Ann. Appl. Probab.* 20(5):1854–1890.
- [13] Delasay M, Ingolfsson A, Kolfal B, Schultz K (2019) Load effect on service times. *Eur. J. Oper. Res.* 279(3):673–686.
- [14] Douglas HE, Raban MZ, Walter SR, Westbrook JI (2017) Improving our understanding of multi-tasking in healthcare: Drawing together the cognitive psychology and healthcare literature. *Appl. Ergonomics* 59:45–55.
- [15] Elkhuizen SG, Bor G, Smeenk M, Klazinga NS, Bakker PJ (2007) Capacity management of nursing staff as a vehicle for organizational improvement. *BMC Health Service Res.* 7(1):196–205.
- [16] Ethier SN, Kurtz TG (2009) *Markov Processes: Characterization and Convergence*, vol. 282 (John Wiley & Sons, Hoboken, NJ).
- [17] Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.
- [18] Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.
- [19] Green LV, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* 16(1):13–39.



- [20] Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.
- [21] Hall R (2012) Bed assignment and bed management. Hall R, ed. *Handbook of Healthcare System Scheduling*, International Series in Operations Research & Management Science, vol. 168 (Springer, Boston), 177–200.
- [22] Harrison J, Williams R (1996) A multiclass closed queueing network with unconventional heavy traffic behavior. *Ann. Appl. Probab.* 6(1):1–47.
- [23] Harrison JM, Reiman MI (1981) Reflected Brownian motion on an orthant. *Ann. Probab.* 9(2):302–308.
- [24] Hasija S, Pinker E, Shumsky RA (2010) OM practice—Work expands to fill the time available: Capacity estimation and staffing under Parkinson’s law. *Manufacturing Service Oper. Management* 12(1):1–18.
- [25] Horváth IA, Scully Z, Van Houdt B (2019) Mean field analysis of join-below-threshold load balancing for resource sharing servers. *Proc. ACM Measurement Anal. Comput. Systems*, vol. 3, 1–21.
- [26] Jacod J, Shiryaev A (2013) *Limit Theorems for Stochastic Processes*, vol. 288 (Springer, Berlin).
- [27] Janssen A, Van Leeuwen JS, Zwart B (2011) Refining square-root safety staffing by expanding Erlang C. *Oper. Res.* 59(6):1512–1522.
- [28] Jennings OB, Mandelbaum A, Massey WA, Whitt W (1996) Server staffing to meet time-varying demand. *Management Sci.* 42(10):1383–1394.
- [29] Kallenberg O (2006) *Foundations of Modern Probability* (Springer, Cham, Switzerland).
- [30] Karatzas I, Shreve SE (1998) Brownian motion. *Brownian Motion and Stochastic Calculus* (Springer, New York), 47–127.
- [31] Kc DS (2014) Does multitasking improve performance? Evidence from the emergency department. *Manufacturing Service Oper. Management* 16(2):168–183.
- [32] Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Sci.* 55(9):1486–1498.
- [33] Kleinrock L (1976) *Queueing Systems, Volume 2. Computer Applications* (Wiley, New York).
- [34] Legros B, Jouini O (2019) On the scheduling of operations in a chat contact center. *Eur. J. Oper. Res.* 274(1):303–316.
- [35] Long Z, Tezcan T, Zhang J (2018) Customer service chat systems with general service and patience times. Preprint, submitted June 24, <http://dx.doi.org/10.2139/ssrn.3201743>.
- [36] Luo J, Zhang J (2013) Staffing and control of instant messaging contact centers. *Oper. Res.* 61(2):328–343.
- [37] Mandelbaum A, Zeltyn S (2009) Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.* 57(5):1189–1205.
- [38] Mandelbaum A, Massey WA, Reiman MI (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30(1–2):149–201.
- [39] Marshall AW, Olkin I, Arnold BC (2011) *Inequalities: Theory of Majorization and Its Applications*, 2nd ed. (Springer, Cham, Switzerland).
- [40] Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys* 4:193–267.
- [41] Puhalskii AA, Reiman MI (2000) The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* 32(2):564–595.
- [42] Reed J, Ward AR (2004) A diffusion approximation for a generalized Jackson network with reneging. *Proc. 42nd Annual Allerton Conf. Comm. Control Comput.* (Curran Associates, Inc., Red Hook, NY).
- [43] Reed J, Ward AR (2008) Approximating the GI/GI/1+ GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Math. Oper. Res.* 33(3):606–644.
- [44] Reed J, Ward A, Zhan D (2013) On the generalized drift Skorokhod problem in one dimension. *J. Appl. Probab.* 50(1):16–28.
- [45] Reiman MI (1984) Open queueing networks in heavy traffic. *Math. Oper. Res.* 9(3):441–458.
- [46] Revuz D, Yor M (2013) *Continuous Martingales and Brownian Motion*, vol. 293 (Springer Science & Business Media, New York).
- [47] Robert P (2013) *Stochastic Networks and Queues*, vol. 52 (Springer Science & Business Media, New York).
- [48] Schroeder B, Harchol-Balter M, Iyengar A, Nahum E, Wierman A (2006) How to determine a good multi-programming level for external scheduling. *2nd Internat. Conf. Data Engrg. (ICDE’06)* (IEEE, Piscataway, NJ).
- [49] Sparaggis PD, Towsley D, Cassandras C (1993) Extremal properties of the shortest/longest non-full queue policies in finite-capacity systems with state-dependent service rates. *J. Appl. Probab.* 30(1):223–236.
- [50] Tan TF, Netessine S (2014) When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Sci.* 60(6):1574–1593.
- [51] Tezcan T, Zhang J (2014) Routing and staffing in customer service chat systems with impatient customers. *Oper. Res.* 62(4):943–956.
- [52] van der Boer MV, Borst SC, Van Leeuwen JS, Mukherjee D (2022) Scalable load balancing in networked systems: A survey of recent advances. *SIAM Rev.* 64(3):554–622.
- [53] van Leeuwen JS, Mathijsen BW, Zwart B (2019) Economies-of-scale in many-server queueing systems: Tutorial and partial review of the QED Halfin-Whitt heavy-traffic regime. *SIAM Rev.* 61(3):403–440.
- [54] Véricourt Fd, Jennings OB (2011) Nurse staffing in medical units: A queueing perspective. *Oper. Res.* 59(6):1320–1331.
- [55] Whitt W (1992) Understanding the efficiency of multi-server service systems. *Management Sci.* 38(5):708–723.
- [56] Whitt W (2002) *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues* (Springer, New York).
- [57] Whitt W (2004) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* 50(10):1449–1461.
- [58] Wickens CD, Hollands JG, Banbury S, Parasuraman R (2015) *Engineering Psychology and Human Performance* (Psychology Press, New York).
- [59] Yankovic N, Green LV (2011) Identifying good nursing levels: A queueing approach. *Oper. Res.* 59(4):942–955.
- [60] Yashkov S (1987) Processor-sharing queues: Some progress in analysis. *Queueing Systems* 2(1):1–17.
- [61] Zhang J, Zwart B (2008) Steady state approximations of limited processor sharing queues in heavy traffic. *Queueing Systems* 60(3–4):227–246.
- [62] Zhang J, Dai J, Zwart B (2011) Diffusion limits of limited processor sharing queues. *Ann. Appl. Probab.* 21(2):745–799.