Customer-server population dynamics in heavy traffic

Rami Atar* Prasenjit Karmakar* David Lipshutz*†

December 30, 2020

Abstract

We study a many-server queueing model with server vacations, where the population size dynamics of servers and customers are coupled: a server may leave for vacation only when no customers await, and the capacity available to customers is directly affected by the number of servers on vacation. We focus on scaling regimes in which server dynamics and queue dynamics fluctuate at matching time scales, so that their limiting dynamics are coupled. Specifically, we argue that interesting coupled dynamics occur in (a) the *Halfin-Whitt* regime, (b) the *nondegenerate slowdown* regime, and (c) the intermediate, *near Halfin-Whitt* regime; whereas the dynamics asymptotically decouple in the other heavy traffic regimes. We characterize the limiting dynamics, which are different for each scaling regime. We consider relevant respective performance measures for regimes (a) and (b) — namely, the probability of wait and the slowdown. While closed form formulas for these performance measures have been derived for models that do not accommodate server vacations, it is difficult to obtain closed form formulas for these performance measures in the setting with server vacations. Instead, we propose formulas that approximate these performance measures, and depend on the steady-state mean number of available servers and previously derived formulas for models without server vacations. We test the accuracy of these formulas numerically.

1 Introduction

Scaling limits for stochastic processing networks with a growing number of servers is an active research area. Since the pioneering work of Halfin and Whitt [13], the novel scaling that they introduced has attracted substantial interest, and also inspired the study of various other scaling regimes in which the number of servers grows to infinity in the limit. These studies include scaling limit results at the law of large numbers (or fluid) scale, as well as several distinct frameworks at the central limit theorem (or diffusion) scale. For a sample of such work, see [1], [2], [3], [8], [9], [11], [12], [14], [17], [22], [23], [26], [29], [30].

Queueing models with server vacations arise in computer communication systems and production engineering, and their mathematical analysis has a long history in the operations research literature; see the earlier survey [7], the recent book chapter [15, Ch. 10], and the recent work [22], as well as the references therein. Server vacations occur in models that accommodate primary and secondary classes of customers, where, from the viewpoint of primary

*Viterbi Faculty of Electrical Engineering, Technion — Israel Institute of Technology, Haifa 32000, Israel.

†Currently at the Center for Computational Neuroscience, Flatiron Institute, New York City, NY 10010, USA.

customers, a server working non-preemptively on secondary customers may equivalently be regarded as if it takes a vacation, as it is not available during that time. They also arise for a variety of other reasons, including machine breakdowns and maintenance. Kella and Whitt [19] make the distinction between models in which servers leave for vacations according to the state of the queue and ones in which vacations are triggered exogenously (see [15] for various other important distinctions and classifications of vacation models). In this paper we consider a model of the former type, where specifically, as in the case considered in a single-server setting in [19], a server may leave only when the queue is empty. In this case there is an interplay between the population size dynamics of servers and that of customers: the vacations are triggered by the state of the queue, and the queue dynamics is affected by the number of available servers. We study these dynamics at the diffusion scale in a class of heavy traffic many-server regimes, focusing on regimes in which the population size dynamics of customers and of servers fluctuate on the same time scale, so that the equations describing limiting dynamics remain coupled. Our first main contribution is to show that diffusion limits can indeed capture such coupled dynamics, and to classify regimes where it occurs.

To put these results in context some background on classification of heavy traffic regimes is necessary. A formulation of a continuum of heavy traffic regimes was introduced in [1], which contains as special cases the well-known *conventional* and *Halfin-Whitt* (HW) regimes. To introduce it, consider the N-server queue, let $\alpha \in [0,1]$ be a given parameter and let n denote a scaling parameter. Assume that the arrival rate is proportional to n and that the number of servers N_n is proportional to n^α. Impose a critical load condition by letting the total processing rate be nearly equal to the arrival rate. Then the individual service rate must be proportional to $n^{1-\alpha}$. For any α, a diffusion scaled process is obtained by scaling down the queue size by $n^{1/2}$. In this spectrum of heavy traffic regimes, the two endpoints, $\alpha = 0$ and $\alpha = 1$ give the conventional regime (with a fixed number of servers) and, respectively, the HW regime, with $O(n)$ servers and no acceleration of service times. A well-known property of the latter regime, that is unique among all heavy traffic regimes, is that the steady state probability that an arriving customer waits in the queue is asymptotic to a number strictly between 0 and 1. At the midpoint, $\alpha = 1/2$, one obtains the *nondegenerate slowdown* (NDS) regime. A unique property of it is that the time in queue and the time in service for a typical customer are of the same order of magnitude. The slowdown, defined as the ratio of sojourn time and service time for a typical customer is therefore nondegenerate in this regime (that is, it is asymptotic to a number strictly between 1 and ∞). The regimes where $\alpha \in (0, 1/2)$ and $\alpha \in (1/2, 1)$ were not given any names so far in the literature. In this paper we shall refer to them as the *near-conventional* and the *near-HW* regimes, respectively.

The aforementioned coupling between the dynamics of server population size and queue size is argued in this paper to occur for $\alpha \in [1/2, 1]$ but not for $\alpha \in [0, 1/2)$. That is, it occurs in the HW, the near-HW and the NDS regimes. The first setting we consider is of exponential vacation lengths. In this case we show that the scaling limits of the joint dynamics are governed by a pair of coupled one-dimensional equations. In each of the relevant regimes, the set of equations takes a different form:

(i) In the HW regime, the scaling limit is governed by a coupled stochastic differential equation (SDE) and ordinary differential equation (ODE) system.

(ii) In the near HW regime, the scaling limit is governed by a coupled SDE with reflection

(SDER) at zero and ODE system, where the boundary term that constrains the SDER to remain non-negative also appears as a term in the ODE.

(iii) In the NDS regime, the scaling limit is governed by a coupled SDER and birth-death process (BDP), where the positive jumps are driven by the boundary term for the SDER.

We also show that for $\alpha < 1/2$, scaling limits do not exhibit coupled dynamics. Furthermore, a more involved model is treated, in which vacation lengths follow a phase type distribution.

The second goal of this paper is to study the effect of server vacations on natural performance measures, in the special cases of HW and NDS. In these two cases there are performance measures that are particularly interesting to study. In the HW regime, it is the *probability of wait*, that is, the steady state probability that an arriving customer has to wait for service. This probability was shown in [13] to converge to a number strictly between 0 and 1, and an explicit formula was given for the limit. It is not a meaningful performance measure in any other regime $\alpha \in [0, 1)$, as in these regimes the limit is always 1. We are interested in the asymptotics of the probability of wait in presence of server vacations. The diffusion limit developed here can in principle make it possible to achieve this goal, however explicit expressions are hard to obtain for the two-dimensional dynamics (i). Instead, we propose a further approximation based on heuristics. This gives rise to a formula that is a variant of the original formula of [13].

Similarly, in the case of NDS, a property that distinguishes this regime from all regimes with $\alpha \in [0, 1/2) \cup (1/2, 1]$ is that the *slowdown*, defined as the ratio between expected sojourn time and expected service time in steady state, is asymptotic to a random variable strictly between 1 and ∞. A formula for the slowdown asymptotics was provided in [1] for the model without vacations, and it is of interest to explore how it varies in presence of vacations. Once again, an explicit expression is hard to obtain as it involves two-dimensional dynamics, and we turn instead to a heuristic argument. The heuristic gives rise to a variation of the formula from [1] obtained by introducing a correction term.

In both cases, we test the proposed formulas numerically. We provide arguments suggesting that the heuristic formulas are nearly accurate in specific parameter settings, and these arguments are validated by our numerical tests. The overall level of accuracy of the heuristic formulas is also discussed.

While there have been a number of works on exact analysis of queueing systems with server vacations (see, e.g., [7], [10], [22], [28]) or fluid limits of such queueing systems (see, e.g., [24]), there have been relatively few results on diffusion limits for queueing systems with server vacations, especially in the many-server regime. In terms of dependence of vacations on the state of the queue, the closest work to ours is the aforementioned [19], which addresses a single server setting. In their model, the server vacations each time the queue becomes empty. They also study the case that the server vacations according to an exogenous Poisson process (in which case there may be service interruptions). In both cases, the heavy traffic scaling limit is a Lévy processes with a secondary jump input. In follow up work [20] they prove decompositions for the stationary distributions of these processes. In the many-server setting, Pang and Whitt [23] prove diffusion limits in the HW regime in the case of exogenous server interruptions (see also [22] for a related work) that simultaneously affect a proportion of the servers. This is relevant for models in which exogenous events result in a large number of servers being out of service (e.g., system-wide computer crashes). The primary difference between our work and

the works [22, 23] is that our focus is on server vacations triggered by the state of the queue, which leads to a server-customer population dynamics.

The organization of this paper is as follows. Below, some mathematical notation used in this paper is introduced. In §2, the main model and scaling regimes are introduced, and the first main result is stated. Its proof is provided next in §3. An extension of the result to phase type vacation time distribution is presented and proved in §4. In §5, heuristic formulas are developed for performance measures in the HW and NDS regimes. Finally, numerical tests of the level of accuracy of these heuristics are then provided and discussed in the same section.

Notation

Let $\mathbb{N} = \{1, 2, \dots\}$ denote the positive integers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For $d \in \mathbb{N}$ let $\mathbb{R}^d$ denote d-dimensional Euclidean space. We use boldface letters to denote vectors. When $d = 1$ we suppress the superscript d and write $\mathbb{R}$ for the real numbers. Let $\mathbb{R}_+ = [0, \infty)$ denote the non-negative axis. For $a, b \in \mathbb{R}$, the maximum [resp., minimum] is denoted by $a \vee b$ [resp., $a \wedge b$]. For $a \in \mathbb{R}$, the positive [resp., negative] part is denoted by $a^+ = a \vee 0$ [resp., $a^- = (-a) \vee 0$]. For $a \in \mathbb{R}_+$, let $\lceil a \rceil = \min\{n \in \mathbb{N}_0 : n \geq a\}$. For $d \in \mathbb{N}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, let $\mathbf{x} \cdot \mathbf{y}$ and $\|\mathbf{x}\|$ denote the usual scalar product and ℓ_2 norm, respectively. Given a sequence $\{x_n\}$ in $\mathbb{R}_+$ and $\alpha \geq 0$, we say $x_n \sim n^\alpha$ if $n^{-\alpha} x_n \to C$ as $n \to \infty$ for some $C \in \mathbb{R}_+$.

For $f : \mathbb{R}_+ \to \mathbb{R}^d$, let $\|f\|_T = \sup_{t \in [0,T]} \|f(t)\|$, and, for $\theta \in (0, T)$, let

$$w_T(f, \theta) = \sup_{0 \leq s < u \leq s + \theta \leq T} \|f(u) - f(s)\|.$$

For a Polish space $\mathcal{S}$, let $\mathbb{C}_{\mathcal{S}}([0, T])$ and $\mathbb{D}_{\mathcal{S}}([0, T])$ denote the set of continuous and, respectively, cadlag functions $[0, T] \to \mathcal{S}$, which is endowed with the Skorokhod J_1-topology. Write $\mathbb{C}_{\mathcal{S}}$ and $\mathbb{D}_{\mathcal{S}}$ for the case where $[0, T]$ is replaced by $\mathbb{R}_+$. Write $X_n \Rightarrow X$ for convergence in distribution. A sequence of processes X_n with sample paths in $\mathbb{D}_{\mathcal{S}}$ is said to be *C-tight* if it is tight and every subsequential limit has, with probability 1, sample paths in $\mathbb{C}_{\mathcal{S}}$. Denote by ι the identity map on $\mathbb{R}_+$ defined by $\iota(t) = t$ for $t \in \mathbb{R}_+$.

A standard Brownian motion, or SBM for short, is a one-dimensional Brownian motion starting from zero, with zero drift and unit variance. We abbreviate "random variable" and "independent and identically distributed" with "RV" and, respectively, "IID".

2 The dynamic server population model

The queueing model consists of a single queue with multiple servers. Customers arrive according to a renewal process and are served in the order in which they arrive (i.e., first-come-first-serve). The service time at each server is exponentially distributed. Furthermore, when a server becomes idle it waits an exponentially distributed amount of time and then vacations (provided there are still no customers in queue). Servers spend an exponentially distributed amount of time vacationing before returning to service. The choice of the exponentially distribution for service times, idle times before vacationing, and vacationing times is chosen for mathematical tractability of the limiting process. In particular, we are able to obtain heuristic formulas which allow us to estimate relevant steady-state quantities. Allowing for general distributions would require a measure-valued state-descriptor (see, e.g., [18]). In §4 we consider phase type vacation time distributions.

2.1 Scaling regimes

We consider a sequence of queueing networks, indexed by $n \in \mathbb{N}$, that are built on a common probability space $(\Omega, \mathcal{F}, P)$. For $n \in \mathbb{N}$, let N^n denote the number of servers in the nth system, $\lambda^n > 0$ denote the inverse of the mean interarrival times of customers to the system and $\mu^n_{ind} > 0$ denote the inverse mean service time. We use 'ind' as a mnemonic for individual service rate. The parameter $\alpha \in [\frac{1}{2}, 1]$ will differentiate the different scaling regimes we consider. In particular, we assume

$$\lambda^n \sim n, \qquad N^n \sim n^\alpha, \qquad \mu^n_{ind} \sim n^{1-\alpha}.$$

Then the overall capacity of the servers, which is given by the product $\mu^n = \mu^n_{ind} N^n$, is of order n and is thus of the same order as the arrival rate λ^n. The three regimes we consider are as follows:

(i) $\alpha = 1$: HW regime.

(ii) $\alpha \in (\frac{1}{2}, 1)$: near HW regime.

(iii) $\alpha = \frac{1}{2}$: NDS regime.

2.2 Customer dynamics

For $n \in \mathbb{N}$ let $Q^n(t)$ denote the number of jobs in the buffer at time t. At any given time, a server can be in three possible states: busy, idle, or vacationing. The number of servers that are idle and vacationing at time t are denoted by $I^n(t)$ and $V^n(t)$, respectively. The number of busy servers is then given by

$$B^n(t) := N^n - I^n(t) - V^n(t), \tag{1}$$

and the number of customers in the system at time t, denoted by $X^n(t)$, is given by

$$X^n(t) = Q^n(t) + B^n(t) = Q^n(t) + N^n - I^n(t) - V^n(t). \tag{2}$$

The initial conditions $Q^n(0)$, $I^n(0)$ and $V^n(0)$ are $\mathbb{N}_0$-valued RVs represent the number of customers initially in the buffer, the number of servers initially idle and the number of servers initially vacationing, respectively.

Let $\{IA(l) : l \in \mathbb{N}\}$ be strictly positive IID RVs with mean 1 and variance $C^2_{IA} > 0$, and define

$$A^n(t) := \sup\left\{ l \geq 0 : \sum_{k=1}^{l} \frac{IA(k)}{\lambda^n} \leq t \right\}, \qquad t \geq 0\,. \tag{3}$$

Then $A^n(t)$ represents the number of customers that arrive in the interval $[0, t]$. We assume that, as $n \to \infty$,

$$\hat{\lambda}^n := n^{-\frac{1}{2}}(\lambda^n - n\lambda) \to \hat{\lambda}, \tag{4}$$

where $\lambda > 0$ and $\hat{\lambda} \in \mathbb{R}$ are constants. In the nth system the server pool consists of $N^n = \lceil n^\alpha \rceil$ servers. Each of the servers has IID exponential service times with parameter $\mu^n_{ind} = \mu^n_{ind}(\alpha)$.

Recall the overall capacity, given by the product $\mu^n = \mu^n_{ind} N^n$, is of order n. We assume that, as $n \to \infty$,

$$\hat{\mu}^n := n^{-\frac{1}{2}}(\mu^n - n\mu) = n^{-\frac{1}{2}}(\mu^n_{ind} N^n - n\mu) \to \hat{\mu}, \tag{5}$$

where $\mu > 0$ and $\hat{\mu} \in \mathbb{R}$ are constants. The critical load condition, $\mu = \lambda$, is assumed throughout this work. Fix a standard unit Poisson process S. Then the potential service process, denoted by S^n, is given by $S^n(t) := S(\mu^n t)$ for $t \geq 0$. The number of departing customers by time t, denoted $D^n(t)$, is given by

$$D^n(t) = S\left(\mu^n_{ind} \int_0^t B^n(s)ds\right). \tag{6}$$

Denoting by $J^n(t)$ the number of jobs routed to the service pool by time t (not counting the initial number), we also have the following balance equations, namely

$$Q^n(t) = Q^n(0) + A^n(t) - J^n(t), \tag{7}$$

$$I^n(t) + V^n(t) = I^n(0) + V^n(0) + D^n(t) - J^n(t). \tag{8}$$

Equation (7) expresses the fact that the queue length increases by one with each arrival and decreases by one on every routing of a job from the queue to the service pool. Equation (8) expresses the fact that the number of non-active servers (due to idleness or vacation) increases by one on every customer departure and decreases by one on every routing of a new job to the service pool (in particular, it does not vary when both events occur at the same time). A work conservation condition is in force, according to which servers may not be idle when there is work in the queue. This can be expressed as

$$\text{for every } t \geq 0,\ Q^n(t) > 0 \text{ implies } I^n(t) = 0. \tag{9}$$

It is assumed that the condition above holds even for $t = 0$, that is, if $Q^n(0) > 0$ then $I^n(0) = 0$, hence a constraint is implicitly assumed regarding the initial condition $(Q^n(0), I^n(0))$. However, no constraint is put on $V^n(0)$, which may take a positive value even if the initial queue length $Q^n(0)$ is positive.

2.3 Server dynamics

Our model for server vacations has positive parameters β^n and γ^n, which are assumed to satisfy

$$(\beta^n, \gamma^n) \to (\beta, \gamma) \quad \text{as} \quad n \to \infty, \tag{10}$$

for some $\beta \geq 0$ and $\gamma \geq 0$. Note that we allow for the degenerate cases where $\beta = 0$ and or $\gamma = 0$; see Remark 2.1(b). (In §4 we treat a more complicated model in which there are multiple vacationing states.) The model allows a server to start a vacation only if it is idle. It can be described as follows: an exponential clock operating at rate β^n is started when the server becomes idle, and if it ticks before the server is busy again, it goes on a vacation for a duration that is exponentially distribution with rate γ^n (that is, a vacationing server continues vacation until the exponential clock rings even if there are customers in the queue). Thanks to the assumed homogeneity of the servers, this mechanism can be modeled by working with the joint idleness process, and the total number of servers vacationing, rather than accounting for each server individually. To this end, let S_B and S_E be two standard Poisson processes,

where B and E are used as mnemonics for *beginning* and *end* of vacation. Then the counting processes associated with vacation beginnings and endings are

$$V_B^n(t) = S_B\left(\beta^n \int_0^t I^n(s)ds\right), \qquad V_E^n(t) = S_E\left(\gamma^n \int_0^t V^n(s)ds\right), \tag{11}$$

respectively. Thus the number of servers vacationing is given by

$$V^n(t) = V^n(0) + V_B^n(t) - V_E^n(t). \tag{12}$$

It is assumed that the five objects $(Q^n(0), I^n(0), V^n(0))$, $IA(\cdot)$, $S(\cdot)$, $S_B(\cdot)$ and $S_E(\cdot)$ are mutually independent.

2.4 Statement of main result

We can now state the main result on the model in the exponential vacation case, characterizing the scaling limit of (X^n, V^n). Define the diffusion scaled process

$$\hat{X}^n = \frac{X^n - N^n}{\sqrt{n}}, \tag{13}$$

where we recall that $N^n = \lceil n^\alpha \rceil$. Normalize the vacation population size process with scaling specific to α, namely

$$\tilde{V}^n = \frac{V^n}{n^{\alpha - \frac{1}{2}}}. \tag{14}$$

Set

$$b := \hat{\lambda} - \hat{\mu}, \qquad \sigma^2 := \mu(C_{IA}^2 + 1).$$

The statement of the result uses the following terminology. Given an $\mathbb{R}_+$-valued process $X = \{X(t), t \geq 0\}$, we say that a $\mathbb{R}_+$-valued process $L = \{L(t), t \geq 0\}$ is *a boundary term for X at zero* if a.s.,

(i) $L(0) = 0$,

(ii) the sample paths of L are non-decreasing, and

(iii) L can only increase when X is zero, i.e., $\int_{[0,\infty)} X(t)dL(t) = 0$.

Theorem 2.1 *Fix $\alpha \in [\frac{1}{2}, 1]$. Assume that the rescaled initial conditions of X^n and V^n converge, namely that $(\hat{X}^n(0), \tilde{V}^n(0)) \Rightarrow (X_0, V_0)$ as $n \to \infty$. In the case $\alpha \in [\frac{1}{2}, 1)$, assume also that $X_0 \geq 0$ a.s. Then $(\hat{X}^n, \tilde{V}^n) \Rightarrow (X, V)$ as $n \to \infty$, where the pair (X, V) satisfies coupled equations that depend on α as follows.*

(i) (HW regime) In the case $\alpha = 1$, the pair (X, V) takes values in $\mathbb{R} \times \mathbb{R}_+$ and forms a solution to the SDE-ODE system

$$\begin{cases} X(t) = X_0 + \displaystyle\int_0^t [b + \mu \max(-X(s), V(s))]ds + \sigma W(t), \\ \\ V(t) = V_0 + \displaystyle\int_0^t [\beta(X(s) + V(s))^- - \gamma V(s)]ds, \end{cases} \tag{15}$$

where W is an SBM, independent of (X_0, V_0).

(ii) (near-HW regime) In the case $\alpha \in (\frac{1}{2}, 1)$, the pair (X, V) takes values in $\mathbb{R}_+ \times \mathbb{R}_+$ and forms a solution to the SDER-ODE system

$$\begin{cases} X(t) = X_0 + \int_0^t [b + \mu V(s)]ds + \sigma W(t) + L(t), \\ \\ V(t) = V_0 - \gamma \int_0^t V(s)ds + \beta\mu^{-1} L(t), \end{cases} \tag{16}$$

where L is a boundary term for X at zero, and W is an SBM, independent of (X_0, V_0).

(iii) (NDS regime) In the case $\alpha = \frac{1}{2}$, the pair (X, V) takes values in $\mathbb{R}_+ \times \mathbb{Z}_+$ and forms a solution to the system

$$\begin{cases} X(t) = X_0 + \int_0^t [b + \mu V(s)]ds + \sigma W(t) + L(t), \\ \\ V(t) = V_0 - S_E\left(\gamma \int_0^t V(s)ds\right) + S_B(\beta\mu^{-1} L(t)), \end{cases} \tag{17}$$

where L is a boundary term for X at zero, W is an SBM, S_B and S_E are standard Poisson processes, and W, S_B, S_E and (X_0, V_0) are mutually independent.

In case (i) (resp., (ii), (iii)), the system of equations (15) (resp., (16), (17)) uniquely characterizes the law of the pair (X, V).

Remark 2.1 *(On the roles played by various parameters).*
(a) *It is argued in Appendix A.2 that for $\alpha \in [0, \frac{1}{2})$ (the conventional and near-conventional regimes) the unnormalized process V^n simply vanishes in the limit $n \to \infty$. Hence there can be no rescaling under which the pair of processes remains coupled. Thus the meaningful regimes for the model studied in this paper are only $\alpha \in [\frac{1}{2}, 1]$.*
(b) *Our limit theorem allows for the degenerate cases that $\beta = 0$ and/or $\gamma = 0$. When $\beta = \gamma = 0$ servers do not vacation or return to service and from (15)–(17) we see that $V(t) = V(0)$ for all $t \geq 0$. In (15) we recover the well known SDE for X (with drift $b + \mu V(0)$) studied in [13]. Similarly in (16) or (17) we recover the one-dimensional RBM for X (with drift $b + \mu V(0)$) obtained in [1] when the abandonment rate is set to zero. On the other hand, if $\beta = 0$ and $\gamma > 0$, then servers only return to service in the limit. In this case, $V(t) \to 0$ as $t \to \infty$ in all three scaling regimes. Finally, if $\beta > 0$ and $\gamma = 0$ then servers do not return to service in the limit. In this case, $V(\cdot)$ is a non-decreasing process in all three scaling regimes. Note that although the service capacity can only decrease with time, our results indicate that the system is still in heavy traffic, in the sense that the diffusion-scaled queue length does not blow up in finite time.*
(c) *The parameter α does not appear in any of the equations (15), (16), (17), but it plays a central role in that it determines the regime by controlling how N^n and μ^n_{ind} scale (as well as how V^n scales, see (14)), and it dictates which of these equations is valid in each case. Because these equations do not depend on α, the dynamics defined by (16) do not approach those of (15) (resp., (17)) as α tends to 1 (resp., 1/2). It should come as no surprise that such an*

interchange of limits may sometimes fail to hold. Even for the model with no vacations, the large n limit and the $\alpha \to 1$ limit do not commute: The dynamics defined by the first line of (16) *with $V \equiv 0$ do not approach those of the first line of* (15) *with $V \equiv 0$ as $\alpha \to 1$. (This cannot be said about the limit $\alpha \to 1/2$, because one obtains the same dynamics when setting $V = 0$ in the first line of* (16) *and in the first line of* (17)*; thus the interchange of limits $n \to \infty$ and $\alpha \to 1/2$ is valid in the model without vacations and invalid in the model with vacations).*

Remark 2.2 *(On the boundary term). Given a solution (X, V) to either* (16) *or* (17)*, let L be a boundary term for X at zero. Define the process $\xi = \{\xi(t), t \geq 0\}$ by*

$$\xi(t) = X_0 + \int_0^t [b + \mu \max(-X(s), V(s))]ds + \sigma W(t).$$

Then the pair (X, L) is a solution to the well known one-dimensional Skorokhod problem for ξ, and is explicitly given by $(X, L) = \Gamma(\xi)$, where $\Gamma = (\Gamma_1, \Gamma_2)$ is the one-dimensional Skorokhod map (see Appendix A.1). Moreover, as will become apparent from the proof, L is the weak limit of a suitably rescaled version of the cumulative idle process (that is, of L^n defined in (31)*).*

Remark 2.3 *(On the queue length and waiting time processes). It follows from Theorem 2.1 and its proof that the rescaled queue length process $\hat{Q}^n(t) := n^{-1/2}Q^n(t)$ also converges. Specifically, under the assumptions of Theorem 2.1, $(\hat{X}^n, \tilde{V}^n, \hat{Q}^n) \Rightarrow (X, V, Q)$, where*

$$Q(t) = (X(t) + V(t))^+ \text{ when } \alpha = 1 \text{ and } Q(t) = X(t)^+ \text{ when } \alpha \in [\tfrac{1}{2}, 1). \tag{18}$$

This statement is proved in Appendix A.3.

Next, the convergence of the rescaled waiting time process also follows. Specifically, let AT_t^n denote the time of the first arrival to occur after time t,

$$AT_t^n = \inf\{s > t : A^n(s) > A^n(s-)\}.$$

Let WT_t^n denote the time the customer arriving at AT_t^n waits in the queue before being assigned a server. A precise definition of this process in terms of the previously defined model quantities is $WT_t^n = 0$ if $Q^n(AT_t^n) = 0$, and otherwise

$$WT_t^n = \inf\{s > 0 : J^n(AT_t^n + s) = J^n(AT_t^n) + Q^n(AT_t^n)\}.$$

Namely, it is the time it takes all the $Q^n(AT_t^n)$ customers that are in the queue at time AT_t^n to leave the queue. The diffusive rescaling of this process is given by $\widehat{WT}_t^n = n^{1/2}\, WT_t^n$. Reiman's snapshot principle [27] relates the diffusion scale asymptotics of the waiting time process to that of the queue length process, under a heavy traffic condition. According to this principle, the weak limit of $\widehat{WT}_t^n$ is $\mu^{-1}Q_t$, where Q_t is the weak limit of the diffusion scaled queueing process. Since we have identified the latter as (18)*, we obtain the convergence of the processes $\widehat{WT}^n \Rightarrow WT$, where*

$$WT_t = \mu^{-1}(X(t) + V(t))^+ \text{ when } \alpha = 1 \text{ and } WT_t = \mu^{-1}X(t)^+ \text{ when } \alpha \in [\tfrac{1}{2}, 1).$$

A proof of Reiman's snapshot principle in various related contexts has been provided before, for example in [1] (see the online appendix) and [4] (see Section 5). Whereas the settings in these references are somewhat different, the proof in the current setting is very similar and is thus omitted.

3 Proof of main result

3.1 Useful identities and preparatory lemmas

We first define some related scaled processes. Namely, let

$$\hat{A}^n(t) = \frac{A^n(t) - \lambda^n t}{\sqrt{n}}, \qquad \hat{S}^n(t) = \frac{S(nt) - nt}{\sqrt{n}}, \tag{19}$$

$$\hat{Q}^n(t) = \frac{Q^n(t)}{\sqrt{n}}, \qquad \tilde{I}^n(t) = \frac{I^n(t)}{n^{\alpha-\frac{1}{2}}} \tag{20}$$

(where $\hat{Q}^n$ has already been defined in Remark 2.3). The various convergence results in Theorem 2.1 are largely based on the fact that centered renewal processes (with finite second moments) satisfy a functional central limit theorem. In particular, Theorem 14.1 of [5] and the mutual independence of the processes $\hat{A}^n$ and $\hat{S}^n$ imply that these processes jointly converge to processes $\hat{A}$ and $\hat{S}$, which are mutually independent driftless BMs with diffusion coefficients $\sqrt{\lambda}C_{IA}$ and, respectively, 1.

We next develop several identities satisfied by the scaled processes. Using first (2), (7) and (8), and then (6), we obtain

$$\begin{aligned}\hat{X}^n(t) &= \hat{X}^n(0) + \frac{A^n(t) - D^n(t)}{\sqrt{n}} \\ &= \hat{X}^n(0) + \sigma W^n(t) + n^{-\frac{1}{2}}\lambda^n t - n^{-\frac{1}{2}}\mu^n_{ind}\int_0^t B^n(s)ds,\end{aligned}$$

where

$$W^n(t) = \frac{\hat{A}^n(t) - \hat{S}^n\left(n^{-1}\mu^n_{ind}\int_0^t B^n(s)ds\right)}{\sigma}. \tag{21}$$

Thus by (1), (4), (5) and the critical load condition $\lambda = \mu$, we have, denoting $b^n = \hat{\lambda}^n - \hat{\mu}^n$,

$$\hat{X}^n(t) = \hat{X}^n(0) + \sigma W^n(t) + b^n t + n^{-1}\mu^n \int_0^t (\tilde{I}^n(s) + \tilde{V}^n(s))ds. \tag{22}$$

By (2),

$$\hat{Q}^n = \hat{X}^n + n^{-\frac{1}{2}}(I^n + V^n) = \hat{X}^n + n^{\alpha-1}\tilde{I}^n + n^{\alpha-1}\tilde{V}^n.$$

In view of the non-idling condition (9), this gives the identities

$$(\hat{X}^n + n^{\alpha-1}\tilde{V}^n)^+ = \hat{Q}^n, \qquad (n^{1-\alpha}\hat{X}^n + \tilde{V}^n)^- = \tilde{I}^n. \tag{23}$$

(Notice that it is possible to have simultaneously $\hat{X}^n(t) < 0$ and $\hat{Q}^n(t) > 0$, unlike in the model without vacations). Define the process $Y^n = \{Y^n(t), t \geq 0\}$ by

$$Y^n := (\hat{X}^n, \tilde{V}^n), \tag{24}$$

and for a constant $c_0 > 0$, define the stopping time $\tau^n(c_0)$ by

$$\tau^n(c_0) := \inf\{t : \|Y^n(t)\| \geq c_0\}. \tag{25}$$

Lemma 3.1 *Let $\alpha \in [\frac{1}{2}, 1]$. Then the processes W^n are C-tight.*

Proof. Recalling that $B^n(t) \leq N^n$ by definition (1), and using the finiteness of the constant $c = \sup_n n^{-1}\mu^n_{ind}N^n < \infty$, we have $n^{-1}\mu^n_{ind}\int_0^t B^n(s)ds \leq ct$ for all $t \geq 0$ and $n \in \mathbb{N}$. Thus

$$w_T(W^n, \theta) \leq \sigma^{-1} w_T(\hat{A}^n, \theta) + \sigma^{-1} w_{cT}(\hat{S}^n, c\theta).$$

Since the centered renewal processes $\hat{A}^n$ and $\hat{S}^n$ are C-tight, it follows that the processes W^n are also C-tight. □

The following relations will be useful for cases (i) and (ii), i.e., for $\alpha \in (\frac{1}{2}, 1]$. For such α dividing by $n^{\alpha-\frac{1}{2}}$ in (11) and (12) yields the relation

$$\tilde{V}^n(t) = \tilde{V}^n(0) + \beta^n \int_0^t \tilde{I}^n(s)ds - \gamma^n \int_0^t \tilde{V}^n(s)ds + e^n(t), \tag{26}$$

where, with

$$e^n_B(u) = n^{-\alpha+\frac{1}{2}}\left[S_B(n^{\alpha-\frac{1}{2}}u) - n^{\alpha-\frac{1}{2}}u\right], \qquad e^n_E(u) = n^{-\alpha+\frac{1}{2}}\left[S_E(n^{\alpha-\frac{1}{2}}u) - n^{\alpha-\frac{1}{2}}u\right], \tag{27}$$

we have denoted

$$e^n(t) = e^n_B\left(\beta^n \int_0^t \tilde{I}^n(s)ds\right) - e^n_E\left(\gamma^n \int_0^t \tilde{V}^n(s)ds\right). \tag{28}$$

Lemma 3.2 *Let $\alpha \in (\frac{1}{2}, 1]$. Then for all $T < \infty$, $\|e^n_B\|_T + \|e^n_E\|_T \Rightarrow 0$ as $n \to \infty$.*

Proof. Let $T < \infty$. The convergence $\|e^n_B\|_T + \|e^n_E\|_T \Rightarrow 0$ follows from definition (27) and the functional law of large numbers for the Poisson processes S_B and S_E. □

Next, we record relations that will useful for cases (ii) and (iii), i.e., for $\alpha \in [\frac{1}{2}, 1)$. Let

$$\tilde{X}^n(t) = (\hat{X}^n(t))^+ \quad \text{and} \quad e^n_X(t) = (\hat{X}^n(t))^- = \tilde{X}^n(t) - \hat{X}^n(t). \tag{29}$$

Then $\tilde{X}^n$ is non-negative and by (22), $\tilde{X}^n = \xi^n + L^n$, where

$$\xi^n(t) = \hat{X}^n(0) + \int_0^t [b^n + n^{-1}\mu^n \tilde{V}^n(s)]ds + \sigma W^n(t) + e^n_X(t). \tag{30}$$

and

$$L^n(t) = n^{-1}\mu^n \int_0^t \tilde{I}^n(s)ds. \tag{31}$$

Furthermore, by (23), $\tilde{I}^n(t) > 0$ implies $\hat{X}^n(t) < 0$. As a result, by (31) and (29), we see that L^n is non-decreasing and L^n can only increase when $\tilde{X}^n$ is zero, i.e., $\int_0^\infty 1_{\{\tilde{X}^n(t)>0\}} dL^n(t) = 0$. Consequently, $(\tilde{X}^n, L^n)$ is the solution to the one-dimensional Skorokhod problem for ξ^n (see Appendix A.1), and so the pair can be expressed in terms of the one-dimensional Skorokhod map, as follows,

$$(\tilde{X}^n, L^n) = \Gamma(\xi^n). \tag{32}$$

Lemma 3.3 *Suppose $\alpha \in [\frac{1}{2}, 1)$. Then $e_X^n \Rightarrow 0$.*

Proof. Fix $\alpha \in [\frac{1}{2}, 1)$ and $T < \infty$. By assumption, the weak limit X_0 of $\hat{X}^n(0)$ is non-negative, so it suffices to show that for any $\varepsilon > 0$, $P(\Omega^{n,\varepsilon}) \to 0$, where

$$\Omega^{n,\varepsilon} = \left\{ \exists\, 0 \le s^n < t^n \le T : \hat{X}^n(s^n) > -2\varepsilon, \hat{X}^n(t^n) < -3\varepsilon, \sup_{t \in [s^n, t^n]} \hat{X}^n(t) < -\varepsilon \right\}.$$

By (22), on the event $\Omega^{n,\varepsilon}$, there exists $0 \le s^n < t^n \le T$ such that

$$-\varepsilon > \hat{X}^n(t^n) - \hat{X}(s^n) \ge -\sigma w_T(W^n, \delta^n) - c\delta^n + n^{-1}\mu^n \int_{s^n}^{t^n} (\tilde{I}^n(u) + \tilde{V}^n(u))du,$$

where $c = \sup_n \|b^n\| < \infty$ and $\delta^n = t^n - s^n$. It follows from (23) and the fact that $\hat{X}^n(t) < -\varepsilon$ on the interval $[s^n, t^n]$, that $\tilde{I}^n + \tilde{V}^n > n^{1-\alpha}\varepsilon$ on the interval. As a result, for all sufficiently large n, on $\Omega^{n,\varepsilon}$,

$$n^{-1}\mu^n \int_{s^n}^{t^n} (\tilde{I}^n(u) + \tilde{V}^n(u))du \ge c_1 \varepsilon n^{1-\alpha}\delta^n,$$

where $c_1 = \inf_n n^{\alpha-1}\mu^n_{ind} = \inf_n n^{-1}\mu^n > 0$. Thus

$$P(\Omega^{n,\varepsilon}) \le P(\sigma w_T(W^n, \delta^n) + c\delta^n \ge \varepsilon + c_1\varepsilon n^{1-\alpha}\delta^n).$$

Fix $\alpha' \in (0, 1-\alpha)$. Separating the cases $\delta^n < n^{-\alpha'}$ and $\delta^n \ge n^{-\alpha'}$, we obtain that, for all sufficiently large n,

$$P(\Omega^{n,\varepsilon}) \le P(\sigma w_T(W^n, n^{-\alpha'}) + cn^{-\alpha'} \ge \varepsilon) + P(2\sigma\|W^n\|_T + cT \ge c_1\varepsilon n^{1-\alpha-\alpha'}).$$

Both terms on the RHS converge to zero by the C-tightness of W^n shown in Lemma 3.1. Since ε and T are arbitrary, this shows that $e_X^n \Rightarrow 0$. □

3.2 Proof of Theorem 2.1

We can now prove our main scaling limit result. Throughout the proof, C^n denotes a generic sequence of constants that satisfy $C^n \to 1$, where by the term 'generic' we mean that the values the sequence takes may vary from one line to another. In addition, the symbol c denotes a generic positive constant (that, in particular, does not depend on n).

Proof of Theorem 2.1. The three regimes are treated separately. For each regime, our approach is as follows: (a) prove uniqueness in law of solutions to the limiting equations [i.e., (15), (16) or (17)], (b) express the equations for $(\hat{X}^n, \tilde{V}^n)$ in a form that resembles the limiting equations, (c) prove, for each T, tightness of $\|Y^n\|_T$, which implies C-tightness of the relevant processes, and (d) show that along any convergent subsequent, the limiting processes satisfy the limiting equations (for which uniqueness in law holds).

(i) The case $\alpha = 1$. We first establish uniqueness in law of solutions to the system of equations (15). Observe that (15) can be viewed as a degenerate SDE with Lipschitz drift and diffusion coefficients (the latter is constant). For such an SDE, pathwise uniqueness of solutions

holds, and consequently so does uniqueness in law. For the former see Ch. V, Theorem 7 of [25].

We now write relations for the $(\hat{X}^n, \tilde{V}^n)$ that closely resemble the limiting system (15). Setting $\alpha = 1$ in (23) shows that $\tilde{I}^n + \tilde{V}^n = \max(-\hat{X}^n, \tilde{V}^n)$, so

$$\tilde{I}^n(t) + \tilde{V}^n(t) \leq 2\|Y^n(t)\|. \tag{33}$$

In addition, substituting the relation into (22), and recalling relation (26) for $\tilde{V}^n$, yields the system of equations

$$\begin{cases} \hat{X}^n(t) = \hat{X}^n(0) + C^n\mu \displaystyle\int_0^t [b^n + \max(-\hat{X}^n(s), \tilde{V}^n(s))]ds + \sigma W^n(t), \\ \\ \tilde{V}^n(t) = \tilde{V}^n(0) + \displaystyle\int_0^t [\beta^n(\hat{X}^n(s) + \tilde{V}^n(s))^- - \gamma^n\tilde{V}^n(s)]ds + e^n(t). \end{cases} \tag{34}$$

Next, for fixed $T < \infty$, we prove tightness of $\|Y^n\|_T$. By (34) and the boundedness of b^n, β^n and γ^n, we have

$$\|Y^n\|_t \leq \|Y^n(0)\| + \sigma\|W^n\|_t + c_1 t + c_1 \int_0^t \|Y^n\|_s ds + \|e^n\|_t, \qquad t \geq 0. \tag{35}$$

for some fixed constant c_1 that does not depend on n. Appealing to Gronwall's lemma shows that, for $t \geq 0$,

$$\|Y^n\|_t \leq (\|Y^n(0)\| + \sigma\|W^n\|_t + c_1 t + \|e^n\|_t)\exp(c_1 t). \tag{36}$$

Recall that the RVs $\|Y^n(0)\| + \sigma\|W^n\|_T$ form a tight sequence by the assumed convergence of the initial conditions and the C-tightness of W^n shown in Lemma 3.1. Given $\varepsilon > 0$ let K be sufficiently large so that $\limsup_n P(\|Y^n(0)\| + \sigma\|W^n\|_T > K) < \varepsilon$. Choose $c_0 > (K + c_1 T + 1)\exp(c_1 T)$ and let $\tau^n = \tau^n(c_0)$ be defined as in (25). In this case, by (33) and the definition of τ^n, $\tilde{I}^n(t \wedge \tau^n) + \tilde{V}^n(t \wedge \tau^n) \leq 2c_0$ for all $t \geq 0$, so, in view of definition (28) for e^n and Lemma 3.2, we have $\|e^n\|_{t\wedge\tau^n} \leq \|e^n_B\|_{2c_0\beta^n} + \|e^n_E\|_{2c_0\gamma^n} \Rightarrow 0$ as $n \to \infty$. Therefore, by our choice of K,

$$\limsup_{n\to\infty} P(\|Y^n(0)\| + \sigma\|W^n\|_T + \|e^n\|_{T\wedge\tau^n} > K + 1) < \varepsilon.$$

From (36) and our choice of c_0 we see that on the event $\|Y^n(0)\| + \sigma\|W^n\|_T + \|e^n\|_{T\wedge\tau^n} \leq K+1$, we have $\|Y^n\|_{T\wedge\tau^n} < c_0$. Hence by definition (25), $\tau^n > T$ on that event. As a result, $\limsup_n P(\|Y^n\|_T > c_0) < \varepsilon$. Since ε is arbitrary, this shows that $\|Y^n\|_T$ forms a tight sequence of RVs.

Having shown tightness of the RVs $\|Y^n\|_T$, we now establish that $(W^n, e^n) \Rightarrow (W, 0)$. Using (1), (33) and the fact that $N^n = n$, we have $1 - 2n^{-\frac{1}{2}}\|Y^n\|_T \leq n^{-1}B^n(t) \leq 1$ for all $0 \leq t \leq T$ and $n \in \mathbb{N}$. As a result, $\|n^{-1}B^n - 1\|_T \Rightarrow 0$ as $n \to \infty$. Using this, along with the convergence $\mu^n_{ind} \to \mu$ as $n \to \infty$, in the definition (21) for W^n shows that $W^n(\cdot) \Rightarrow \sigma^{-1}(\hat{A}(\cdot) - \hat{S}(\mu\cdot))$. Note that the latter process is a driftless BM with diffusion coeffient given by $\sigma^{-1}(\lambda C^2_{IA} + \mu)^{1/2} = \sigma^{-1}\lambda^{1/2}(C^2_{IA} + 1)^{1/2} = 1$. Hence the limit is equal in distribution to the SBM W. Next, observe that the RVs $\int_0^T \tilde{I}^n(s)ds$ and $\int_0^T \tilde{V}^n(s)ds$, which appear as arguments of e^n_B and e^n_E in expression (28) for $e^n(T)$, form tight sequences of RVs, as follows from (33). This, along with Lemma 3.2, yields $\|e^n\|_T \Rightarrow 0$.

Now, by (34), for every $T > 0$ and $0 < \theta \leq 1$,

$$w_T(\hat{X}^n, \theta) \leq C^n \mu(|b^n| + \|\hat{X}^n\|_{T+1} + \|\tilde{V}^n\|_{T+1})\theta + \sigma w_T(W^n, \theta),$$

$$w_T(\tilde{V}^n, \theta) \leq \beta^n(\|\hat{X}^n\|_{T+1} + \|\tilde{V}^n\|_{T+1})\theta + 2\|e^n\|_{T+1}.$$

Hence by the established tightness of $\|Y^n\|_{T+1}$ and the convergence $\|e^n\|_{T+1} \Rightarrow 0$, it is seen that the sequence $(\hat{X}^n, \tilde{V}^n)$ is C-tight. Taking limits in (34) along any convergent subsequence, and using that $(\hat{X}^n(0), \tilde{V}^n(0)) \Rightarrow (X_0, V_0)$ by assumption, shows that any limit (X, V, W) of $(\hat{X}^n, \tilde{V}^n, W^n)$ satisfies (15), where we have used that $(C^n, b^n, \beta^n, \gamma^n) \to (1, b, \beta, \gamma)$. Since uniqueness in law for solutions to (15) holds, this completes the proof in the case $\alpha = 1$.

(ii) The case $\alpha \in (\frac{1}{2}, 1)$. As in the previous case we first establish uniqueness in law of solutions to the system of equations (16), which can be viewed as a degenerate SDER on the domain $\mathbb{R}_+ \times \mathbb{R}$ (the non-negativity of the initial condition V_0 and the non-decreasing property of L imply that $V(t) \geq 0$ for all $t \geq 0$; hence there is no necessity to consider the SDER on the smaller domain $\mathbb{R}^2_+$). The reflection vector field takes the constant value $(1, \beta\mu^{-1})$ on the boundary $\{0\} \times \mathbb{R}$. Theorem 4.3 of [21] covers such an SDE with reflection on a bounded domain and provides pathwise uniqueness. A standard localization argument yields pathwise uniqueness for the unbounded domain at hand. This shows that uniqueness in law holds for solutions of (16).

Next, we write relations for the pair $(\hat{X}^n, \tilde{V}^n)$ that closely resemble the limiting system (16). By (22) and (26), for $t \geq 0$,

$$\begin{cases} \hat{X}^n(t) = \hat{X}^n(0) + \displaystyle\int_0^t [b^n + C^n \mu \tilde{V}^n(s)]ds + \sigma W^n(t) + L^n(t), \\ \\ \tilde{V}^n(t) = \tilde{V}^n(0) - \gamma^n \displaystyle\int_0^t \tilde{V}^n(s)ds + C^n \beta^n \mu^{-1} L^n(t) + e^n(t), \end{cases} \tag{37}$$

where L^n is defined as in (31) and we recall that relation (32) holds.

We now turn to the proof that for fixed $T < \infty$ the RVs $\|Y^n\|_T$ are tight. Fix $T < \infty$. By (30) for ξ^n, we have, for all $t \geq 0$,

$$\|\xi^n\|_t \leq \|\hat{X}^n(0)\| + ct + c\int_0^t \|Y^n\|_s ds + c\|W^n\|_t + \|e^n_X\|_t. \tag{38}$$

By (32) and the Lipschitz continuity of Γ_2 (see Proposition A.1), $L^n(t) \leq \|\xi^n\|_t$ for all $t \geq 0$. Thus, using (37), we get, for any $t \geq 0$,

$$\|Y^n\|_t \leq c_2 \left(Z^n + \|e^n\|_t + \|e^n_X\|_t\right) + c_2 \int_0^t \|Y^n\|_s ds,$$

where c_2 is a fixed constant that does not depend on n and

$$Z^n := \|Y^n(0)\| + \|W^n\|_T + T. \tag{39}$$

Hence by Gronwall's lemma, for $t \in [0, T]$,

$$\|Y^n\|_t \leq c_2 \left(Z^n + \|e^n_X\|_t + \|e^n\|_t\right) \exp\left(c_2 t\right). \tag{40}$$

By the assumed tightness of the initial conditions and C-tightness of W^n shown in Lemma 3.1, the RVs Z^n are tight. Given $\varepsilon > 0$ let K be sufficiently large so that $\limsup_n P(Z^n \geq K) < \varepsilon$. Choose $c_0 > c_2(K+1)\exp(c_2 T)$ and let $\tau^n = \tau^n(c_0)$. Since $\tilde{V}^n(T \wedge \tau^n) \leq \|Y^n(T \wedge \tau^n)\| \leq c_0$, it holds that the RVs $\int_0^{T\wedge\tau^n} \tilde{V}^n(s)ds$ are tight. In addition, by (31), the fact that $L^n(t) \leq \|\xi^n\|_t$ for all $t \geq 0$, and (38),

$$n^{-1}\mu^n \int_0^{T\wedge\tau^n} \tilde{I}^n(s)ds = L^n(T \wedge \tau^n) \leq \|\hat{X}^n(0)\| + c(1+c_0)T + c\|W^n\|_{T\wedge\tau^n} + \|e_X^n\|_{T\wedge\tau^n}.$$

Since $\hat{X}^n(0)$ are tight by assumption, W^n are C-tight by Lemma 3.1 and $\|e_X^n\|_{T\wedge\tau^n} \Rightarrow 0$ by Lemma 3.3, it follows that the RVs $\int_0^{T\wedge\tau^n} \tilde{I}^n(s)ds$ are tight. In view of definition (28) for e^n and Lemma 3.2, we see that $\|e^n\|_{T\wedge\tau^n} \Rightarrow 0$ as $n \to \infty$. Thus, by our choice of K,

$$\limsup_{n\to\infty} P(Z^n + \|e_X^n\|_{T\wedge\tau^n} + \|e^n\|_{T\wedge\tau^n} > K+1) < \varepsilon.$$

From (40) and our choice of c_0 we see that on the event $Z^n + \|e_X^n\|_{T\wedge\tau^n} + \|e^n\|_{T\wedge\tau^n} \leq K+1$, we have $\|Y^n\|_{T\wedge\tau^n} < c_0$. Hence by definition (25), $\tau^n > T$ on that event. As a result, $\limsup_n P(\|Y^n\|_T > c_0) < \varepsilon$. Since ε is arbitrary, this shows that $\|Y^n\|_T$ forms a tight sequence of RVs.

Having shown tightness of the RVs $\|Y^n\|_T$, it follows that the RVs $\int_0^T \tilde{V}^n(s)ds$ are tight. Hence, by (41), the RVs ξ^n are C-tight, and since $(\tilde{X}^n, L^n) = \Gamma(\xi^n)$, it follows from Proposition A.1 that the RVs $(\tilde{X}^n, L^n)$ are also C-tight. Consequently, since $e_X^n \Rightarrow 0$ (Lemma 3.3), the RVs $\hat{X}^n$ are also C-tight. Using that $L^n(t) = n^{-1}\mu^n \int_0^t \tilde{I}^n(s)ds$ and that $n^{-1}\mu^n \to \mu$ as $n \to \infty$, we see that the RVs $\int_0^t \tilde{I}^n(s)ds$ are tight. Thus, the definition (28) for e^n gives $e^n \Rightarrow 0$ as $n \to \infty$. Finally, using (1) and (31) yields

$$N^n - c(\|Y^n\|_t + L^n(t)) \leq \int_0^t B^n(s)ds \leq N^n.$$

As a result, $\|n^{-\alpha}B^n - 1\|_T \Rightarrow 0$ as $n \to \infty$. Using this, along with the convergence $n^{-1+\alpha}\mu_{ind}^n \to \mu$, in the definition (21) for W^n shows that $W^n \Rightarrow W$. Taking limits in (37) and (32) along any convergent subsequence, and using that $(\hat{X}(0), \tilde{V}(0)) \Rightarrow (X_0, V_0)$ by assumption, the Skorokhod representation theorem, the continuity of Γ and the convergence $e_X^n \Rightarrow 0$, shows that any limit (X, X, V, W, L, ξ) of $(\hat{X}^n, \tilde{X}^n, \tilde{V}^n, W^n, L^n, \xi^n)$ satisfies the equation (16), and $(X, L) = \Gamma(\xi)$, which, by definition, implies that L is a boundary term for X. Since uniqueness in law holds for solutions of (16), this completes the proof in the case $\alpha \in (\frac{1}{2}, 1)$.

(iii) The case $\alpha = \frac{1}{2}$. First note that thanks to the non-negativity of X_0, the continuity of the second and third terms in the first equation in (17) and the oscillation inequality for the SM (Proposition A.1), the processes X and L have continuous sample paths a.s.. Also, by the second equation in (17), V has piecewise constant, right-continuous sample paths. Uniqueness in law of solutions to the system of the equations (17) follows from pathwise uniqueness, which we now establish. That is, given W, S_B, S_E and (X_0, V_0), then any two solutions of (17) are equal a.s. To this end, let $\Sigma = (X, L, V)$ and $\Sigma' = (X', L', V')$ be two solutions, let $\tau = \inf\{t : \Sigma(t) \neq \Sigma'(t)\}$, and consider the event $\tau < \infty$. By definition, $(X(t), V(t)) = (X'(t), V'(t))$ for $t \in [0, \tau)$. We first show that $V(\tau) = V'(\tau)$. To see this must

hold, suppose V has a jump of size -1 at τ. Then necessarily $S_E(u) = S_E(u-) + 1$, where $u = \gamma \int_0^\tau V(s)ds$. However, $V = V'$ on $[0, \tau)$, this implies that V' also has a jump of size -1 at τ. A similar argument holds for a jump of size $+1$. This shows that V and V' agree on $[0, \tau]$ whenever $\tau < \infty$. Since V is piecewise constant with right-continuous sample paths and $V(\tau) = V'(\tau)$, there exists $\varepsilon > 0$ (depending on the sample path) such that $V(t) = V'(t)$ for all $t \in [0, \tau + \varepsilon)$. It follows from the expression for (X, L) and (X', L') in terms of the one-dimensional Skorokhod map that they are also equal on $[0, \tau + \varepsilon)$, thus contradicting the definition of τ. With this contradiction thus obtained, we must have $\tau = \infty$ a.s., so pathwise uniqueness holds.

Next, we write relations for the $(\hat{X}^n, \tilde{V}^n)$ that closely resemble the limiting system (17). By (22) and (11)–(12),

$$\begin{cases} \hat{X}^n(t) = \hat{X}^n(0) + \displaystyle\int_0^t [b^n + C^n \mu V^n(s)]ds + \sigma W^n(t) + L^n(t), \\ \\ V^n(t) = V^n(0) - S_E\left(\gamma^n \displaystyle\int_0^t V^n(s)ds\right) + S_B\left(C^n \beta^n \mu^{-1} L^n(t)\right), \end{cases} \tag{41}$$

where L^n is defined as in (31). In addition we recall that (32) holds.

We now turn to the proof of tightness of the RVs $\|Y^n\|_T$. The main difference between this case and the case $\alpha \in (\frac{1}{2}, 1)$ is the treatment of the equation that governs V^n, where we note that $\tilde{V}^n = V^n$ in this case. We argue as follows. By (30), for all $t \geq 0$,

$$\|\xi^n\|_t \leq c_4 M^n + c_4 \int_0^t V^n(s)ds, \tag{42}$$

where $M^n := \|\hat{X}^n(0)\| + T + \|W^n\|_T + \|e_X^n\|_T$ and $c_4 \geq 1$ is a suitable constant that does not depend on n. Hence by (41) and the fact that $L^n(t) \leq \|\xi^n\|_t$,

$$\|V^n\|_t \leq V^n(0) + S_B(c\|\xi^n\|_t) \leq V^n(0) + S_B\left(c_4 M^n + c_4 \int_0^t V^n(s)ds\right), \tag{43}$$

where we have chosen c_4 to be possibly larger. By the convergence of the initial conditions $(V^n(0), \hat{X}^n(0)) \Rightarrow (V_0, X_0)$, the C-tightness of W^n (Lemma 3.1) and the convergence $e_X^n \Rightarrow 0$ (Lemma 3.3), it follows that the RVs $V^n(0) + c_4 M^n$ are tight. Let $\varepsilon > 0$ and choose $K < \infty$ sufficiently large such that $P(V^n(0) + c_4 M^n \geq K) < \varepsilon$ for all n. In addition, since S_B is a Poisson process, by the functional law of large numbers, $\|k^{-1} S_B(k\cdot) - \iota(\cdot)\|_u \Rightarrow 0$ as $k \to \infty$, for $u = 1 + 2c_4 T e^{c_4 T}$. Thus, by choosing $K < \infty$ possibly larger, we can ensure that $P(\Omega^n) > 1 - 2\varepsilon$, where

$$\Omega^n := \left\{V^n(0) + c_4 M^n \leq K \text{ and } S_B(t) < t + K \text{ for all } t \leq K + 2c_4 K T e^{c_4 T}\right\}.$$

Let $c_0 := 2K + 4c_4 K T c^{e_4 T} + 2K e^{c_4 T}$ and $\tau^n = \tau^n(c_0)$. Then $c_4 M^n + c_4 \int_0^t V^n(s) \leq K + 2c_4 K T e^{c_4 T}$ for all $t \leq T \wedge \tau^n$. Thus, on the event Ω^n, we have, for all $t \leq T$,

$$\|V^n\|_{t \wedge \tau^n} \leq V^n(0) + c_4 M + c_4 \int_0^{t \wedge \tau^n} V^n(s)ds + K \leq 2K + c_4 \int_0^{t \wedge \tau^n} V^n(s)ds.$$

Hence, by Gronwall's lemma, on this event, $\|V^n\|_{T\wedge\tau^n} \leq 2Ke^{c_4T}$. Moreover, by (32), the Lipschitz continuity of the SM (Proposition A.1) and (42), on this event we have

$$\|\hat{X}^n\|_{T\wedge\tau^n} \leq 2\|\xi^n\|_{T\wedge\tau^n} \leq 2c_4M^n + 2c_4\int_0^t V^n(s)ds \leq 2K + 4c_4KTe^{c_4T}.$$

Combining the bounds on V^n and $\hat{X}^n$ gives $\|Y^n\|_{T\wedge\tau^n} \leq 2K + 4c_4KTe^{c_4T} + 2Ke^{c_4T}$. Thus, $\tau^n > T$ on Ω^n. Since ε is arbitrary, this shows the tightness of the RVs $\|Y^n\|_T$.

Having shown tightness of the RVs $\|Y^n\|_T$, we can argue exactly as in case (ii) to conclude that $(\hat{X}^n, \tilde{X}^n, V^n, L^n, \xi^n)$ are C-tight. Taking limits in (41) and (32) along any convergent subsequence, and using that $(\hat{X}(0), V(0)) \Rightarrow (X_0, V_0)$ by assumption, the Skorokhod representation theorem, the continuity of Γ and the convergence $e_X^n \Rightarrow 0$, shows that any limit (X, X, V, W, L, ξ) of $(\hat{X}^n, \tilde{X}^n, V^n, W^n, L^n, \xi^n)$ satisfies the equations (17), and $(X, L) = \Gamma(\xi)$, which, by definition, implies that L is a boundary term for X. Since uniqueness in law holds for solutions of (17), this completes the proof in the case $\alpha = \frac{1}{2}$. □

4 Multi-stage vacation model

In this section we consider a generalization of the model that allows for multiple vacationing states. Let $m \geq 2$ denote the number of vacationing states. In the multi-stage model, servers may take the following states: busy, idle, and vacationing state i for $i = 1, \ldots, m$. Let $\mathbb{M} = \{1, \ldots, m\}$. At time $t \geq 0$ let $I^n(t)$ denote the number of idle servers and $U_i^n(t)$ denote the number of vacationing servers in state i, for $i \in \mathbb{M}$. Let $\mathbf{U}^n(t) = (U_1^n(t), \ldots, U_m^n(t))$. Then $V^n(t) = \mathbf{U}^n(t) \cdot \mathbf{1} = U_1^n(t) + \cdots + U_m^n(t)$ denotes the total number of vacationing servers and $B^n(t)$, defined as in (1), denotes the number of busy servers.

The customer dynamics described in §2.2 still hold; in particular, equations (1)–(9) hold. The server dynamics, however, are no longer described by (11)–(12). For the multi-stage setting, fix vectors $\boldsymbol{\beta}^n, \boldsymbol{\gamma}^n \in \mathbb{R}_+^m$. Here β_i^n denotes the rate at which idling servers transition to vacationing state i and γ_i^n denotes the rate at which vacationing servers in state i return to idling, for $i \in \mathbb{M}$. Fix an $m \times m$ transition rate matrix $R^n = (r_{ij}^n)$ so that r_{ij}^n denotes the rate at which vacationing servers transition from state i to j, for $i \neq j \in \mathbb{M}$, and $r_{ii}^n = -\sum_{j\neq i} r_{ij}^n$. We assume there exist vectors $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}_+^m$ and an $m \times m$ matrix R such that

$$(\boldsymbol{\beta}^n, \boldsymbol{\gamma}^n, R^n) \to (\boldsymbol{\beta}, \boldsymbol{\gamma}, R) \quad \text{as} \quad n \to \infty. \tag{44}$$

As in §2 we allow for the degenerate cases that $\boldsymbol{\beta} = \mathbf{0}$ and/or $\boldsymbol{\gamma} = \mathbf{0}$. Let $\mathbb{M}_0 = \mathbb{M} \cup \{0\}$ and S_{ij}, for $i \neq j \in \mathbb{M}_0$, be independent unit Poisson processes. For $i \in \mathbb{M}$, define

$$U_{i,B}^n(t) = S_{0i}\left(\beta_i^n \int_0^t I^n(s)ds\right) + \sum_{j\neq i} S_{ji}\left(r_{ji}^n \int_0^t U_j^n(s)ds\right), \tag{45}$$

$$U_{i,E}^n(t) = S_{i0}\left(\gamma_i^n \int_0^t U_i^n(s)ds\right) + \sum_{j\neq i} S_{ij}\left(r_{ij}^n \int_0^t U_i^n(s)ds\right). \tag{46}$$

Then

$$U_i^n(t) = U_i^n(0) + U_{i,B}^n(t) - U_{i,E}^n(t). \tag{47}$$

It is assumed that the objects $(Q^n(0), I^n(0), \mathbf{U}^n(0))$, $IA(\cdot)$, $S(\cdot)$ and $S_{ij}(\cdot)$, $i \neq j \in \mathbb{M}_0$, are mutually independent. Define $\hat{X}^n$ as in (13) and define $\tilde{\mathbf{U}}^n$ by

$$\tilde{\mathbf{U}}^n = \frac{\mathbf{U}^n}{n^{\alpha-\frac{1}{2}}}. \tag{48}$$

Our main result for the multi-stage vacation model characterizes the limit of the scaled pair $(\hat{X}^n, \tilde{\mathbf{U}}^n)$.

4.1 Statement of main result

Let $G^n = \mathrm{diag}(\gamma^n)$ denote the $m \times m$ diagonal matrix satisfying $G^n_{ii} = \gamma^n_i$ for $i \in \mathbb{M}$. Recall the definition of a boundary term given prior to Theorem 2.1. We can now state our main result on limits of the scaled queue-server processes in the case of multi-stage vacations.

Theorem 4.2 *Fix $\alpha \in [\frac{1}{2}, 1]$. Assume that the rescaled initial conditions of X^n and V^n converge, namely that $(\hat{X}^n(0), \tilde{\mathbf{U}}^n(0)) \Rightarrow (X_0, \mathbf{U}_0)$. In the case $\alpha \in [\frac{1}{2}, 1)$, assume also that $X_0 \geq 0$ a.s. Then $(\hat{X}^n, \tilde{\mathbf{U}}^n) \Rightarrow (X, \mathbf{U})$, where the law of the limit process $(X, \mathbf{U})$ depends on α and is specified in what follows.*

(i) (HW regime) In the case $\alpha = 1$, the pair $(X, \mathbf{U})$ takes values in $\mathbb{R} \times \mathbb{R}^m_+$ and forms a solution to the SDE-ODE system

$$\begin{cases} X(t) = X_0 + \displaystyle\int_0^t [b + \mu \max(-X(s), V(s))]ds + \sigma W(t), \\ \\ \mathbf{U}(t) = \mathbf{U}_0 + \displaystyle\int_0^t [\boldsymbol{\beta}(X(s) + V(s))^- + (R - G)\mathbf{U}(s)]ds, \end{cases} \tag{49}$$

where $V = U \cdot \mathbf{1}$ and W is a SBM, independent of $(X_0, \mathbf{U}_0)$.

(ii) (near-HW regime) In the case $\alpha \in (\frac{1}{2}, 1)$, the pair $(X, \mathbf{U})$ takes values in $\mathbb{R}_+ \times \mathbb{R}_+$ and forms a solution to the SDER-ODE system

$$\begin{cases} X(t) = X_0 + \displaystyle\int_0^t [b + \mu V(s)]ds + \sigma W(t) + L(t), \\ \\ \mathbf{U}(t) = \mathbf{U}_0 + \displaystyle\int_0^t (R - G)\mathbf{U}(s)ds + \boldsymbol{\beta}\mu^{-1}L(t), \end{cases} \tag{50}$$

where $V = U \cdot \mathbf{1}$, L is a boundary term for X at zero, and W is a SBM, independent of $(X_0, \mathbf{U}_0)$.

(iii) (NDS regime) In the case $\alpha = \frac{1}{2}$, the pair $(X, \mathbf{U})$ takes values in $\mathbb{R}_+ \times \mathbb{Z}_+$ and forms a

solution to the system

$$\begin{cases} X(t) = X_0 + \int_0^t [b + \mu V(s)]ds + \sigma W(t) + L(t), \\ \\ U_i(t) = U_{0i} - \sum_{j\in\mathbb{M}\setminus\{i\}} S_{ij}\left(r_{ij}\int_0^t U_i(s)ds\right) + \sum_{j\in\mathbb{M}\setminus\{i\}} S_{ji}\left(r_{ji}\int_0^t U_j(s)ds\right) \\ \qquad\qquad + S_{0i}(\beta_i\mu^{-1}L(t)) - S_{i0}\left(\gamma_i\int_0^t U_i(s)ds\right), \end{cases} \tag{51}$$

where $V = U\cdot\mathbf{1}$, L is a boundary term for X at zero, W is a SBM, S_B and S_E are standard Poisson processes, and W, S_{ij}, $i \neq j \in \mathbb{M}_0$, and $(X_0, \mathbf{U}_0)$ are mutually independent.

(iv) In case (i) (resp., (ii), (iii)), the system of equations (49) *(resp.,* (50), (51)*) uniquely characterizes the law of the pair $(X, \mathbf{U})$.*

Remark 4.4 *The equations for X in the multi-stage vacation model are exactly the same as the equations for X in the singe-stage vacation model owing to the fact that the customer dynamics only require information about the total number of servers on vacation.*

4.2 Proof of Theorem 4.2

Define $\hat{A}^n$, $\hat{S}^n$, $\hat{Q}^n$ and $\tilde{I}^n$ as in (19)–(20). Then, as stated there, the centered and scaled processes $\hat{A}^n$ and $\hat{S}^n$ converge to driftless BMs with diffusion coefficients $\sqrt{\lambda}C_{IA}$ and, respectively, 1. Define

$$\tilde{V}^n = \frac{V^n}{n^{1-\alpha}} = \tilde{\mathbf{U}}^n \cdot \mathbf{1}.$$

Then (22) holds with W^n defined as in (21), and relations (23) hold. Define Y^n and $\tau^n(c_0)$, for $c_0 > 0$, as in (24)–(25). Then Lemma 3.1 holds by the exact same argument.

We now derive equations for $\tilde{\mathbf{U}}^n$ for cases (i) and (ii), i.e., for $\alpha \in (\frac{1}{2}, 1]$. Dividing by $n^{\alpha-\frac{1}{2}}$ in (45)–(47), we obtain

$$\begin{aligned} \tilde{U}_i^n(t) = \tilde{U}_i^n(0) + \beta_i^n \int_0^t \tilde{I}^n(s)ds + \sum_{j\in\mathbb{M}\setminus\{i\}} r_{ji}^n \int_0^t \tilde{U}_j^n(s)ds \\ - \gamma_i^n \int_0^t \tilde{U}_i^n(s)ds - \sum_{j\in\mathbb{M}\setminus\{i\}} r_{ij}^n \int_0^t \tilde{U}_i^n(s)ds + e_i^n(t), \end{aligned} \tag{52}$$

where, with

$$e_{ij}^n(u) = n^{-\alpha+\frac{1}{2}}\left[S_{ij}\left(n^{\alpha-\frac{1}{2}}u\right) - n^{\alpha-\frac{1}{2}}u\right], \tag{53}$$

for $i \neq j \in \mathbb{M}_0$, we have denoted

$$\begin{aligned} e_i^n(t) = e_{0i}^n\left(\beta_i^n \int_0^t \tilde{I}^n(s)ds\right) + \sum_{j\in\mathbb{M}\setminus\{i\}} e_{ji}^n\left(r_{ji}^n \int_0^t \tilde{U}_j^n(s)ds\right) \\ - e_{i0}^n\left(\gamma_i^n \int_0^t \tilde{U}_i^n(s)ds\right) - \sum_{j\in\mathbb{M}\setminus\{i\}} e_{ij}^n\left(r_{ij}^n \int_0^t \tilde{U}_i^n(s)ds\right). \end{aligned} \tag{54}$$

Let

$$\mathbf{e}^n(t) = (e_1^n(t), \ldots, e_m^n(t)), \qquad t \geq 0.$$

Lemma 4.4 *Suppose $\alpha \in (\frac{1}{2}, 1]$. Then $e_{ij}^n \Rightarrow 0$ for each $i \neq j \in \mathbb{M}_0$.*

Proof. This follows from definition (53) and the functional law of large numbers. □

Next, in cases (ii) and (iii), i.e., for $\alpha \in [\frac{1}{2}, 1)$, define $\tilde{X}^n$ and e_X^n as in (29), and define ξ^n and L^n as in (30) and, respectively, (31). Then, by the same argument presented there, (32) holds, as does Lemma 3.3.

We can now prove Theorem 4.2. The convention from the previous section regarding the notation C^n and c is kept.

Proof of Theorem 4.2. The proof follows a similar structure to the proof of Theorem 2.1 and many of the arguments are identical or quite similar. Here we describe the new aspects of the proof and refer the reader to the proof of Theorem 2.1 when the arguments are identical. As in the proof of Theorem 2.1, the different regimes are treated separately.

The case $\alpha = 1$. Uniqueness in law of solutions to the system of equations (49) follows from the fact that the system can be viewed as a degenerate SDE with Lipschitz drift and diffusion coefficients, for which pathwise uniqueness of solutions holds.

Next, we write relations for $(\hat{X}^n, \tilde{\mathbf{U}}^n)$ that are similar to (49). As in the single stage setting, equation for $\hat{X}^n$ in (34) holds. After substituting the relation (23) (with $\alpha = 1$) into equation (52) for $\tilde{\mathbf{U}}^n$ with relation and recalling that $R = (r_{ij})$ and $G = \mathrm{diag}(\gamma^n)$, we arrive at the system of equations

$$\begin{cases} \hat{X}^n(t) = \hat{X}^n(0) + C^n \mu \displaystyle\int_0^t [b^n + \max(-\hat{X}^n(s), \tilde{V}^n(s))] ds + \sigma W^n(t), \\ \\ \tilde{\mathbf{U}}^n(t) = \tilde{\mathbf{U}}^n(0) + \displaystyle\int_0^t [\beta^n (\hat{X}^n(s) + \tilde{V}^n(s))^- + (R - G)\mathbf{U}^n(s)] ds + \mathbf{e}^n(t). \end{cases} \tag{55}$$

We now prove tightness of $\|Y^n\|_T$ for all $T < \infty$. Let $T < \infty$. From (55) we see that (35) holds with a possibly different choice of c_1. The remaining argument that the RVs $\|Y^n\|_T$ are tight follows exactly as in the single stage setting, except that

$$\|\mathbf{e}^n\|_{t \wedge \tau^n} \leq \sum_{i \neq j \in \mathbb{M}_0} \|e_{ij}^n\|_{2c_0 \tilde{c}} \Rightarrow 0,$$

with $\tilde{c} = \sup_n \max_{i,j} (\beta_i^n, \gamma_i^n, r_{ij}^n)$, follows from Lemma 4.4 instead of Lemma 3.2. Having established tightness of the RVs $\|Y^n\|_T$ for all $T < \infty$, we argue exactly as in the single stage setting that $(W^n, e^n) \Rightarrow (W, 0)$ (again using Lemma 4.4 instead of Lemma 3.2) and the RVs $\int_0^T \tilde{I}^n(s) ds$ and $\int_0^T \tilde{V}^n(s) ds$ form tight sequences of RVs. Thus, in view of Lemma 4.4, $\mathbf{e}^n \Rightarrow 0$. Taking limits in (55) along any convergent subsequence and using that $(\hat{X}^n(0), \tilde{\mathbf{U}}^n(0)) \Rightarrow (X_0, \mathbf{U}_0)$ by assumption, shows that any limit $(X, \mathbf{U}, W)$ of $(\hat{X}^n, \tilde{\mathbf{U}}^n, W^n)$ satisfies (49), where we have used that $(C^n, b^n, \boldsymbol{\beta}^n, \boldsymbol{\gamma}^n, R^n) \to (1, b, \boldsymbol{\beta}, \boldsymbol{\gamma}, R)$. Since uniqueness in law of solutions to (49) holds, this completes the proof.

The case $\alpha \in (\frac{1}{2}, 1)$. Uniqueness in law of solutions to the system (50), which can be viewed as a degenerate SDER on the domain $\mathbb{R}_+ \times \mathbb{R}^m$, with constant reflection vector $(1, \beta_1\mu^{-1}, \ldots, \beta_m\mu^{-1})$ on the boundary $\{0\} \times \mathbb{R}^m$. Since pathwise uniqueness of such an SDER holds (again via a localization argument), this implies that uniqueness in law holds for solution of (50).

By (22) and (52),

$$\begin{cases} \hat{X}^n(t) = \hat{X}^n(0) + \int_0^t [b^n + C^n\mu\tilde{V}^n(s)]ds + \sigma W^n(t) + L^n(t), \\ \\ \tilde{\mathbf{U}}^n(t) = \tilde{\mathbf{U}}^n(0) + \int_0^t (R-G)\tilde{\mathbf{U}}^n(s)ds + C^n\boldsymbol{\beta}^n\mu^{-1}L^n(t) + \mathbf{e}^n(t), \end{cases} \tag{56}$$

where L^n is defined as in (31) and we recall that relation (32) holds.

The proof that the RVs $\|Y^n\|_T$ are tight for fixed $T < \infty$ follows the same argument as in the single stage setting and, as in that case, implies C-tightness of ξ^n, $\tilde{X}^n$, L^n and $\hat{X}^n$, and the convergence $W^n \Rightarrow W$. Taking limits in (56) along any convergent subsequence, and using the convergence of the initial conditions, the Skorokhod representation theorem and the continuity of Γ shows that any limit (X, X, U, W, L, ξ) of $(\hat{X}^n, \tilde{X}^n, \tilde{\mathbf{U}}^n, W^n, L^n, \xi^n)$ satisfies equation (50) and $(X, L) = \Gamma(\xi)$, so L is a boundary term for X. Since uniqueness in law holds for solutions to (50), this completes the proof in the case $\alpha \in (\frac{1}{2}, 1)$.

The case $\alpha = \frac{1}{2}$. Uniqueness in law of solutions to (51) follows from pathwise uniqueness, which is shown using an argument analogous to the one in the proof of Theorem 2.1. By (22) and (45)–(47),

$$\begin{cases} \hat{X}^n(t) = \hat{X}^n(0) + \int_0^t [b^n + C^n\mu V^n(s)]ds + \sigma W^n(t) + L^n(t), \\ \\ U_i^n(t) = U_i^n(0) - \sum_{j\in\mathbb{M}\setminus\{i\}} S_{ij}\left(r_{ij}^n \int_0^t U_i^n(s)ds\right) + \sum_{j\in\mathbb{M}\setminus\{i\}} S_{ji}\left(r_{ji}^n \int_0^t U_j^n(s)ds\right) \\ \qquad\qquad + S_{0i}(C^n\beta_i\mu^{-1}L^n(t)) - S_{i0}\left(\gamma_i^n \int_0^t U_i^n(s)ds\right), \end{cases} \tag{57}$$

where L^n is defined as in (31). In addition, we recall that (32) holds.

Next, we prove tightness of the sequence $\|Y^n\|_T$ for fixed $T < \infty$. As in the single stage server setting, (42) holds. Hence, by (57), for $i \in \mathbb{M}$,

$$\|V^n\|_t = \|U_1^n + \cdots + U_m^n\|_t \le V^n(0) + \sum_{i\in\mathbb{M}} S_{0i}\left(c_4M^n + c_4\int_0^t U_i^n(s)ds\right), \tag{58}$$

where we have chosen c_4 possibly larger. As in the single stage setting, the RVs $V^n(0) + c_4M^n$ are tight. Let $\varepsilon > 0$ and choose $K < \infty$ sufficiently large such that $P(V^n(0) + c_4M^n \ge K) < \varepsilon$ for all n. In addition, we can choose $K < \infty$ possibly larger so that $P(\Omega^n) > 1 - 2\varepsilon$, where

$$\Omega^n = \left\{V^n(0) + c_4M^n \le K \text{ and } S_{0i}(t) < t + K \text{ for all } t \le K + 2c_4KTe^{c_4T} \text{ and } i \in \mathbb{M}\right\}.$$

The remainder of the proof proceeds exactly as in the single-stage vacation setting. □

5 Heuristic formulas for steady-state performance

The goal of this section is to quantify the effect of the server vacations on performance measures that are of particular interest in the HW and the NDS regimes. In the HW regime (i.e., when $\alpha = 1$), it was shown in [13] that the probability of wait (POW) converges to a nondegenerate limit strictly between 0 and 1, where as in all other regimes (i.e., when $0 \leq \alpha < 1$), POW converges to 1. Furthermore, an explicit formula for the limit of POW was derived in [13]. As argued in [30], POW is an especially useful performance measure because it requires no scaling by a function on n in this regime. It is desirable to understand how this limit probability differs in the case of server vacations of the form that we study. Potentially, this information could be obtained from the steady state distribution of the diffusion limit of Theorem 2.1(i). However, one does not expect an explicit formula for this two-dimensional dynamics. Instead, we develop in §5.1 a formula based on Theorem 2.1(i) and on a heuristic argument in which we substitute the steady-state mean number of available servers into the explicit formula derived in [13]. We provide numerical tests of the accuracy of this heuristic formula.

In the NDS regime (i.e., when $\alpha = 1/2$), the slowdown is a performance measure that is particularly important. The slowdown converges (without any rescaling) to a number strictly between 1 and $+\infty$ in the NDS regime, whereas it converges to either 1 or $+\infty$ in all other regimes (i.e., when $\alpha \neq 1/2$). A formula for the slowdown in absence of server vacations was given in [1]. Again, although an expression for the slowdown, based on the steady state distribution of the coupled pair of Theorem 2.1(iii), could be developed, it would not be explicit. Thus instead we develop, in §5.2, a variant of the formula from [1], based on Theorem 2.1(iii) and a heuristic argument similar to the one used for the HW regime, accompanied by numerical tests.

In both the HW and NDS regimes, the heuristic and the numerics are concerned with the single stage vacation model.

5.1 Steady-state probability of waiting for service in the HW regime

In the original setting of [13] for the model with no vacations, the diffusion limit is given by the SDE

$$X(t) = X(0) + \int_0^t [b + \mu(X(s))^-]ds + \sigma W(t), \tag{59}$$

which is precisely equation (15) with $V = 0$. In the case $b < 0$, this Markov process has a unique invariant distribution, and so the steady state probability $POW_0 := P(X(t) > 0)$ is well defined as a quantity that does not depend on t. We sometimes write this in an informal fashion as $P(X(\infty) > 0)$. It is also proved there that this probability is the $n \to \infty$ limit of the nth system probability of wait. Throughout, $b < 0$ is assumed. Denote by Φ the standard normal cumulative distribution function. Then for the model without vacations, the formula reads

$$POW_0 = \frac{1}{1 + \sqrt{2\pi} b_1 \Phi(b_1) \exp(b_1^2/2)} \qquad \text{where} \qquad b_1 = \sqrt{\frac{2}{\mu}} \frac{|b|}{\sigma}. \tag{60}$$

For the original expression from [13], see Th. 4 there; a corrected version of this formula appeared in [30, eq. (1.1)] (although the original formula is correct in the special case $\sigma = \sqrt{2}$,

corresponding to the M/M/n limit law). Formula (60) above follows the corrected version, albeit with somewhat different notation.

We now go back to the model that accommodates vacations, in which case the limit is given by (15). For each $n \in \mathbb{N}$, define $Y^n(t) = Q^n(t) - I^n(t)$ for all $t \geq 0$, so that, by the non-idling policy, $Y^n(t) > 0$ if and only if $Q^n(t) > 0$. On this event, an arriving customer necessarily waits in the queue. Define the scaled process $\hat{Y}^n$ by

$$\hat{Y}^n(t) = \frac{Y^n(t)}{\sqrt{n}}, \qquad t \geq 0.$$

Then (recalling $\alpha = 1$)

$$\hat{Y}^n(t) = \hat{X}^n(t) + \tilde{V}^n(t), \qquad t \geq 0,$$

and so $\hat{Y}^n \Rightarrow Y$ as $n \to \infty$, where $Y(t) = X(t) + V(t)$ for all $t \geq 0$. We are therefore led to study the probability $P(Y(t) > 0)$ at steady state (where it is independent of t). We have from (15)

$$\begin{cases} Y(t) = Y_0 + \displaystyle\int_0^t [b + (\mu+\beta)(Y(s))^- + (\mu-\gamma)V(s)]ds + \sigma W(t), \\ \\ V(t) = V_0 + \displaystyle\int_0^t [\beta(Y(s))^- - \gamma V(s)]ds. \end{cases} \tag{61}$$

In general an exact expression is beyond the scope of this work. Instead, we use heuristics to estimate the probability, as follows. Denote by $y, v \geq 0$ the a.s. averages

$$y = \lim_{t\to\infty} \frac{1}{t}\int_0^t (Y(s))^- ds, \qquad v = \lim_{t\to\infty} \frac{1}{t}\int_0^t V(s)ds.$$

From (61) we obtain that $b + (\mu+\beta)y + (\mu-\gamma)v = 0$ and $\beta y = \gamma v$. Solving for y and v we have

$$y = \frac{\gamma|b|}{\mu(\gamma+\beta)}, \qquad v = \frac{\beta|b|}{\mu(\gamma+\beta)}.$$

The quantity v is the steady state value of the process V. The main heuristic step is now to substitute v in for $V(s)$ in (61). This yields the (uncoupled) SDE

$$\tilde{Y}(t) = Y_0 + \int_0^t [\tilde{b} + \tilde{\mu}(\tilde{Y}(s))^-]ds + \sigma W(t), \tag{62}$$

where we denote

$$\tilde{b} = \frac{\gamma(\mu+\beta)}{\mu(\gamma+\beta)}b, \qquad \tilde{\mu} = \mu + \beta.$$

Then $P(\tilde{Y}(\infty) > 0) \approx P(Y(\infty) > 0)$ with equality holding when $\mu = \gamma$ since the SDE for Y in (43) becomes uncoupled from the ODE for V. In what follows we will use the notation $POW = P(Y(\infty) > 0)$ and $\widetilde{POW} = P(\tilde{Y}(\infty) > 0)$ for the exact and, respectively, approximate performance measure.

For an explicit expression for $\widetilde{POW}$ we only need to notice the similarity of (62) to (59). That is, $\tilde{Y}$ satisfies the same equation that X satisfies in the model with no vacations, but

with different parameters, hence $P(\tilde{Y}(\infty) > 0)$ can be obtained from (60) upon modifying the parameters. This gives

$$\widetilde{POW} = \frac{1}{1 + \sqrt{2\pi} b_2 \Phi(b_2) \exp(b_2^2/2)} \qquad \text{where} \qquad b_2 = \sqrt{\frac{2}{\tilde{\mu}}} \frac{|\tilde{b}|}{\sigma} = \frac{\gamma\sqrt{2(\mu+\beta)}}{\mu(\gamma+\beta)} \frac{|b|}{\sigma}. \tag{63}$$

The main heuristic step is replacing the stochastic process $V(t)$ by its average. Hence it is expected that the approximation is good when the time scale of vacation lengths is long compared to the time scale at which the queue length fluctuates, which corresponds to the parameter regime $\beta, \gamma \ll \mu$. In this case, $\tilde{\mu} \approx \mu$ and $\tilde{b} \approx rb$, where $r = \gamma/(\gamma + \beta)$, so the limiting dynamics are well approximated by the limiting dynamics of the many-server system without vacations and with scaled drift rb. We return to this point when discussing the simulation results.

Remark 5.5 *One of the referees of this paper pointed out that a simplification also occurs in the opposite parameter regime, where $\beta, \gamma \gg \mu$. In this regime the vacation time scale is very short compared to the queue length fluctuations time scale. In this case, V of* (61) *arrives at its quasistationary state $v(t)$ defined via the relation $\beta(Y(t))^- = \gamma v(t)$. Substituting in the equation for Y, one obtains the approximation*

$$Y(t) = Y_0 + \int_0^t [b + \mu(1 + \beta\gamma^{-1})(Y(s))^-]ds + \sigma W(t).$$

Accordingly, in this regime the formula (60) *should be updated by replacing the parameter μ by $\mu(1 + \beta\gamma^{-1})$.*

5.2 Steady-state slowdown in the NDS regime

It is well known that the steady state distribution of

$$X(t) = x + bt + \sigma W(t) + L(t), \tag{64}$$

where L is the boundary term for X at zero and $b < 0$, is exponential with mean $EX(\infty) = \sigma^2/(2|b|)$. Based on this and the fact that the leading term in the expected service time is given by $\mu\sqrt{n}$, the limiting slowdown (SD) in the NDS regime, for the model with no vacationing servers, was computed in [1] to give

$$SD_0 = 1 + EX(\infty) = 1 + \frac{\sigma^2}{2|b|}. \tag{65}$$

For the model with vacations, the relevant quantity for similar considerations is

$$SD = 1 + EX(\infty),$$

where (X, V) is a solution to (17), and, throughout, $b < 0$ is assumed. As before, we do not aim at an exact calculation because an explicit expression is not expected; however, again one can proceed via a heuristic argument. Define

$$\ell = \lim_{t\to\infty} \frac{L(t)}{t}, \qquad v = \lim_{t\to\infty} \frac{1}{t} \int_0^t V(s)ds.$$

Then by (17), $b + \mu v + \ell = 0$ and $-\gamma v + \beta \ell \mu^{-1} = 0$. Solving yields

$$v = \frac{|b|}{\mu(1 + \beta^{-1}\gamma)} > 0.$$

The main heuristic step is again to substitute v in for $V(s)$ in (17). This yields the approximation

$$\tilde{X}(t) = X_0 + \tilde{b}t + \sigma W(t) + \tilde{L}(t),$$

where

$$\tilde{b} = \frac{b}{1 + \beta\gamma^{-1}}.$$

Thus $\tilde{X}$, that approximates X, is merely a reflected BM, and therefore $E\tilde{X}(\infty) = \sigma^2/(2|\tilde{b}|)$. Hence we obtain an approximation $\widetilde{SD}$ for SD in the form

$$\widetilde{SD} = 1 + E\tilde{X}(\infty) = 1 + \frac{\sigma^2}{2|\tilde{b}|} = 1 + \frac{\sigma^2}{2|b|}(1 + \beta\gamma^{-1}). \tag{66}$$

As in the case of §5.1, the approximation is expected to improve as the mean length of the vacations grows, for the same reasons. By a similar approximation, we see that the limiting dynamics are well approximated by the limiting dynamics of the many-server system without vacations and with a scaled drift rb.

5.3 Numerics

In both cases, a standard Euler-Maruyama method is used to simulate the SDE [16]. The Skorokhod constraining mechanism, associated with the boundary term L, is treated by projecting X back to zero whenever its iteration assumes a negative value. The birth-death processes are treated by drawing Bernoulli RVs, with suitable state dependent bias, to dictate upward and downward jumps.

The time step parameter is denoted by δ, the number of steps by N, and the length of the simulated time interval is thus given by $T = N\delta$.

Sample paths First we present some sample paths of each of the coupled pairs (15) and (17), where the coupling between the processes X and V is apparent. The time step parameter is taken as $\delta = 10^{-3}$. Here the number of steps is $N = 2 \times 10^5$, corresponding to a time interval of length $T = 200$. Sample paths for equation (15), corresponding to the HW regime, are shown in parts (a) and (b) of Figure 1, where $\sigma = 1$ and $\sigma = 3$, respectively (the remaining parameters are taken to be $b = -0.3$, $\beta = 2$, $\gamma = 0.1$ $\mu = 2$). In both (a) and (b) we notice clearly the effect of X on V. Namely, an increase of V occurs when $X + V$ is negative. It is also noticeable that X reaches greater values in (b), where the diffusion coefficient is greater, than in (a).

For the NDS case, the sample paths of equation (17) are shown in parts (c) and (d) of Figure 1. Again, σ takes the two values 1 and 3, respectively (and the remaining parameters are now $b = -3$, $\beta = 2$, $\gamma = 0.1$ $\mu = 2$). Here, the effect of each process on the other is visible. Upward jumps of V occur only when the diffusion process visits zero, and its downward jumps occur only on excursions of X away from zero. The effect of V on X is particularly sharp in

(c): on each excursion of X, the path has strong tendency to increase when $V = 2$, but it decreases rapidly when $V = 1$. A similar dependence of the structure of the excursions on the value of V occurs in (d), where X has a strong negative drift when V becomes zero. Finally, as in the previous case, when σ is greater (that is, in (d)), X reaches higher values on excursions than when σ is smaller.

The HW regime We now use simulations to estimate the quality of the heuristic prediction (63) under various conditions.

As reference for the level of accuracy, we consider the model with $V = 0$ for which a theoretical value of POW_0 is known, and compare it to simulation runs. This we do for the set of parameters $\beta = 2$, $\gamma = 0.1$, $b = -2$, $\mu = 1$, $\sigma = 3$. The theoretical value of POW_0 (given by formula (60)) and the simulation results of 8 runs with $N = 2 \times 10^8$ steps (for the case with $V = 0$, i.e., based on sample paths of (59)) are summarized in Table 1.

POW_0	1	2	3	4	5	6	7	8	max. dev.
0.2470	0.2487	0.2473	0.2487	0.2441	0.2437	0.2459	0.2456	0.2459	0.0033

Table 1: *Simulation results for estimating POW_0 for $N = 2 \times 10^8$ steps. The theoretical value is shown on the left. The maximum absolute deviation from the theoretical result is shown on the right.*

Among these 8 runs, the maximum absolute deviation away from the theoretical value is 0.0033, that is, less than 0.4%. This figure is sufficient for the purposes of this study, and therefore in the simulations described below we keep this value of N.

We control the length of vacations by varying γ. The larger γ is, the shorter is the expected length of vacations. Since the heuristic is based on substituting the long run average of V for V in the X dynamics, it is expected that for long vacations (small γ) the heuristic provides accurate predictions. We also recall that when $\gamma = \mu$, the equation for X decouples from that of V and the heuristic prediction is exact.

The results of our simulation are shown in Figure 2. These four graphs show the simulation results of POW and the heuristic prediction $\widetilde{POW}$ of formula (63) as a function of γ, for two values of μ and two values of σ. Specifically, γ ranges between 0 and 1, and the remaining parameters are taken as $b = -2$, $\beta = 2$, and $\mu \in \{0.5, 1\}$ and $\sigma \in \{1, 3\}$.

The arguments given above suggest that the graphs of POW and $\widetilde{POW}$ should meet at two points, namely $\gamma = 0$ and when $\gamma = \mu$. This is seen very clearly in all parts of Figure 2. As for the level of accuracy, it is overall very good in cases (a), (c) and (d), and is somewhat less satisfactory in case (b). For the actual numerical values of the maximal error sizes, see the figure description. Overall, in all these cases the error is no greater than 5%.

Finally, the general behavior observed, where POW (simulated and predicted) is decreasing as γ increases, is explained by the fact that when the vacations are long (γ small), the system has effectively less service capacity, consequently it is more loaded, and the probability of wait must increase.

The NDS regime In the case of the NDS regime, the simulations are aimed at testing the accuracy of the prediction of formula (66) for the slowdown.

Again we start by considering the reference model with $V = 0$, for which there is a precise formula for the slowdown. The parameters are taken to be $\beta = 5$, $\gamma = 3$, $b = 6$, $\mu = 2$, $\sigma = 3$. It turns out that δ must be calibrated. With $\delta = 10^{-3}$ and $N = 10^8$, there is a significant bias between the simulation and theoretical value, explained by the inaccuracy introduced by the constraining mechanism at zero, as an approximation for the boundary term L. Whereas the size of this error converges to zero as $\delta \to 0$ by theoretical results, the actual error for the above value of δ is too large for our purposes. When we reduce to $\delta = 10^{-4}$ and keep $N = 10^8$, the bias is considerably smaller. The values of 8 runs (based on simulating sample paths of (64)) appear in Table 2, along with the theoretical value (given by formula (65)) and the maximal absolute error.

SD_0	1	2	3	4	5	6	7	8	max. dev.
1.7500	1.7288	1.7340	1.7352	1.7220	1.7410	1.7231	1.7382	1.7402	0.0280

Table 2: *Simulation results for estimating SD_0 for $N = 10^8$ steps. The theoretical value is shown on the left. The maximum absolute deviation from the theoretical result is shown on the right.*

The maximal relative error is 0.0160, that is less than 2%, and is sufficient for our purposes. In what follows we keep these values of δ and N.

Figure 3 shows the simulation results of SD and the heuristic prediction $\widetilde{SD}$ of formula (66) as a function of γ, for two values of μ and two values of σ. The parameters were chosen differently than in the HW case. Our concern here was to calibrate the parameters so as to reach mean delay and mean service time of similar order. This occurs when the slowdown is not too far from the value 2. Specifically, γ ranges between 1 and 10, and the remaining parameters are $b = -6$, $\beta = 5$, and $\mu \in \{2, 4\}$ and $\sigma \in \{2, 3\}$.

Overall, the accuracy of the heuristic prediction is worse than in the HW case, with relative errors reaching as high as 20% in some cases (see description of Figure 3), although in parts of the ranges considered the relative error is considerably smaller. In lack of better approximations, these estimates may be useful in applications as first order approximations.

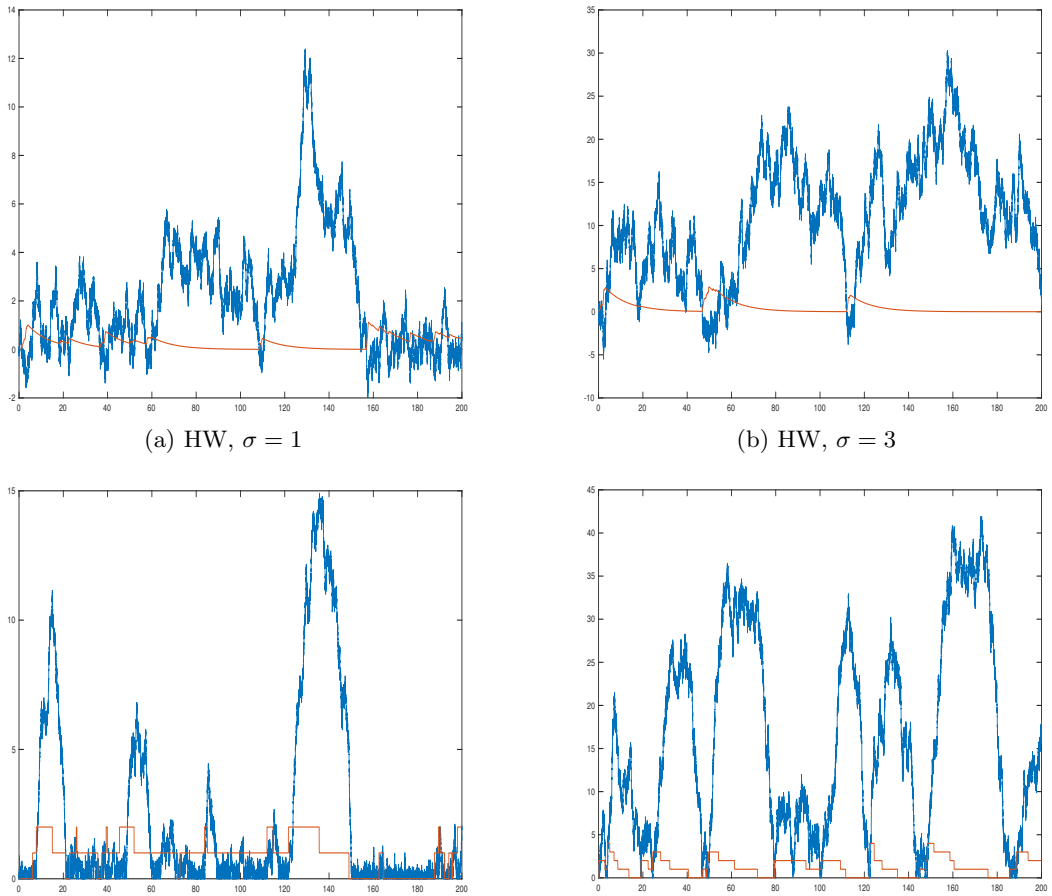
A Appendix

A.1 One-dimensional Skorokhod problem

In this appendix we briefly review some well known properties of the one-dimensional Skorokhod problem. For proofs of the results here, see [6, Ch. 8].

Definition A.1 *Given $x \in \mathbb{D}_+([0,\infty),\mathbb{R})$ we say that a pair $(z, y) \in \mathbb{D}([0,\infty),\mathbb{R}_+)\times\mathbb{D}_0([0,\infty),\mathbb{R}_+)$ satisfies the one-dimensional Skorokhod problem for x if the following conditions hold,*

1. *$z(t) = x(t) + y(t)$ for all $t \geq 0$;*
2. *y is non-decreasing and can only increase when z is zero, i.e., $\int_0^t z(s)dy(s) = 0$, $t \geq 0$.*



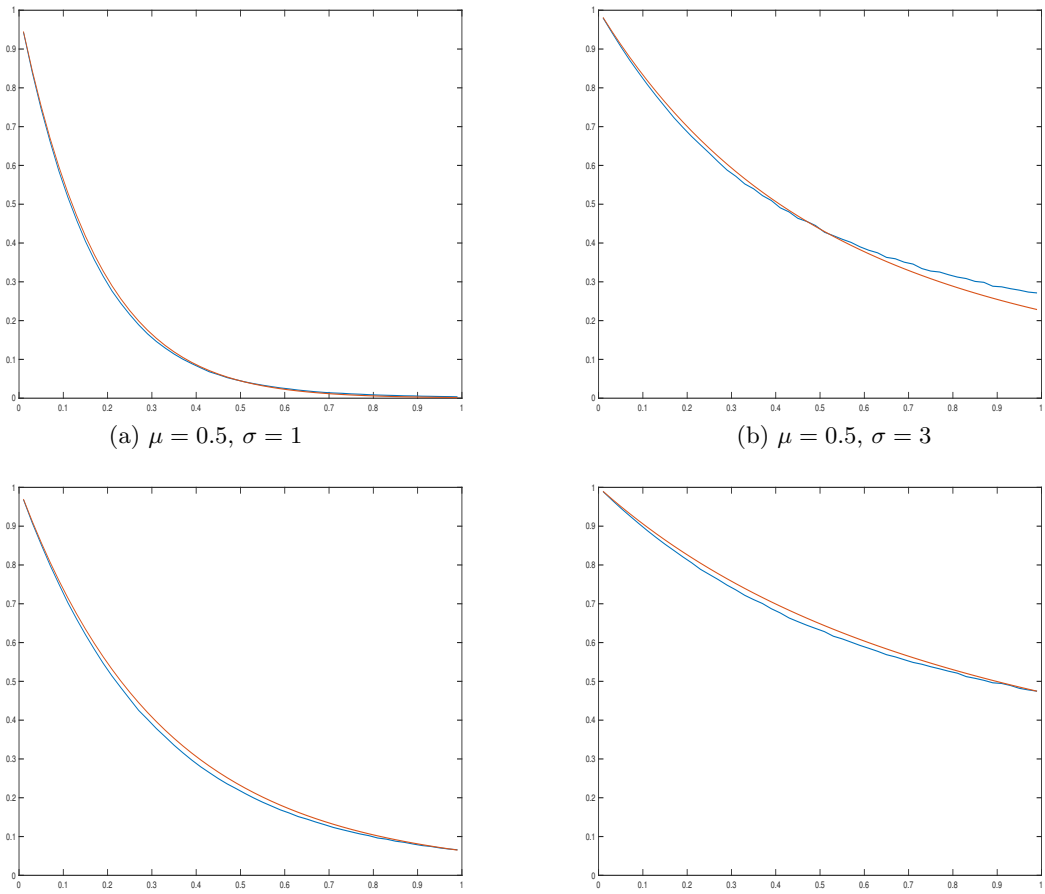
(a) HW, $\sigma = 1$

(b) HW, $\sigma = 3$

(c) NDS, $\sigma = 1$

(d) NDS, $\sigma = 3$

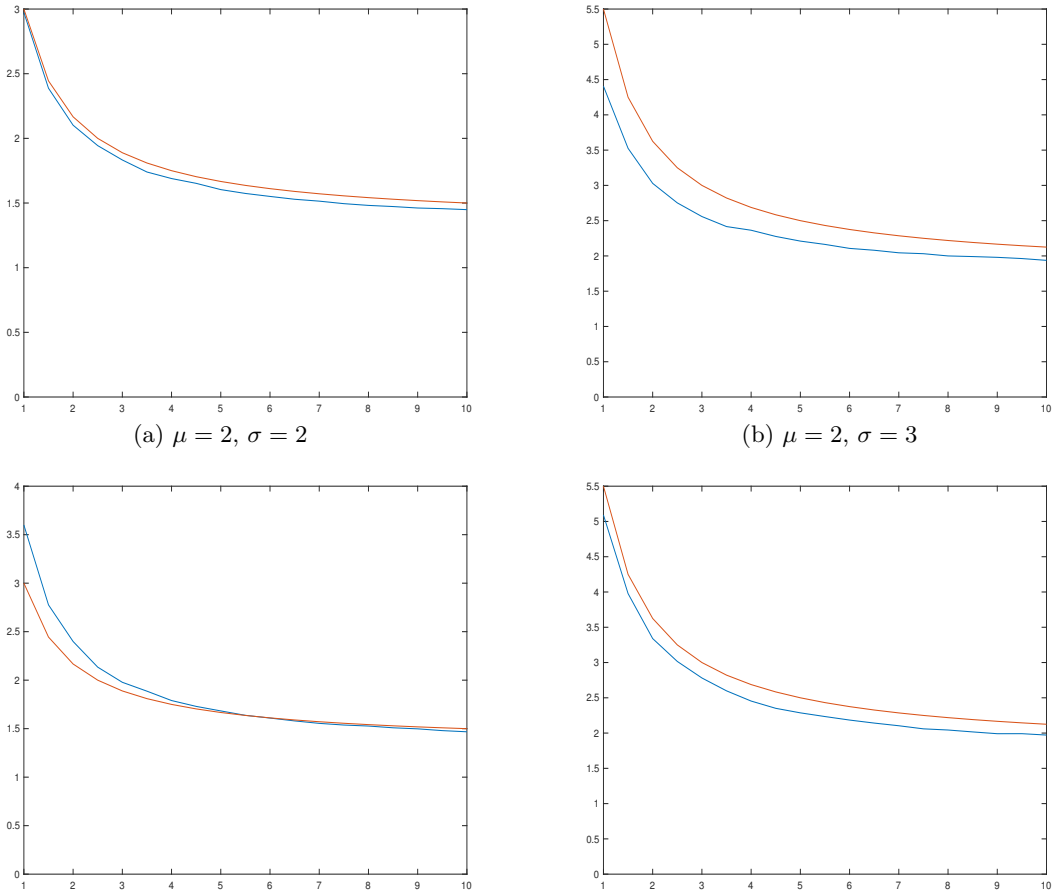
Figure 1: *Parts (a) and (b) [resp., (c) and (d)] show sample paths of the pair X (blue) and V (orange) corresponding to the HW [resp., NDS] limit law, for different values of σ.*



(a) $\mu = 0.5$, $\sigma = 1$ (b) $\mu = 0.5$, $\sigma = 3$

(c) $\mu = 1$, $\sigma = 1$ (d) $\mu = 1$, $\sigma = 3$

Figure 2: *POW obtained by simulation (blue), and* $\widetilde{POW}$ *of formula* (63) *(orange) as a function of* γ *in the range* 0.01 *to* 1*, for different values of* μ *and* σ*. Maximum absolute error: (a)* 0.013*, (b)* 0.042*, (c)* 0.017*, (d)* 0.020*.*



(a) $\mu = 2$, $\sigma = 2$ (b) $\mu = 2$, $\sigma = 3$

(c) $\mu = 4$, $\sigma = 2$ (d) $\mu = 4$, $\sigma = 3$

Figure 3: *SD obtained by simulation (blue), and* $\widetilde{SD}$ *of formula* (63) *(orange) as a function of* γ *in the range* 1 *to* 10*, for different values of* μ *and* σ*. Maximum absolute [resp., relative] error: (a)* 0.0696 *[*0.0390*], (b)* 1.0931 *[*0.1988*], (c)* 0.5996 *[*0.1999*], (d)* 0.4081 *[*0.0900*].*

Proposition A.1 *Given $x \in \mathbb{D}_+([0,\infty),\mathbb{R})$ there exists a unique solution (z,y) of the one-dimensional Skorokhod problem for h given by $(z,y) = (\Gamma_1,\Gamma_2)(h)$, where, for $t \geq 0$,*

$$\Gamma_1(x)(t) = x(t) + \Gamma_2(x)(t), \tag{67}$$

$$\Gamma_2(x)(t) = \sup_{0\leq s\leq t} (x(s))^-. \tag{68}$$

Consequently, the following properties hold:

1. *Oscillation inequality: given $x \in \mathbb{D}_+([0,\infty),\mathbb{R})$ and $0 \leq s < t < \infty$,*

$$Osc(\Gamma_1(x),[s,t]) \leq Osc(x,[s,t]) \qquad \text{and} \qquad Osc(\Gamma_2(x),[s,t]) \leq Osc(x,[s,t]). \tag{69}$$

2. *Lipschitz continuity: for $x_1, x_2 \in \mathbb{D}_+([0,\infty),\mathbb{R})$ and $t \geq 0$,*

$$\sup_{0\leq s\leq t} |\Gamma_1(x_1)(s) - \Gamma_1(x_2)(s)| \leq 2 \sup_{0\leq s\leq t} |x_1(s) - x_2(s)|, \tag{70}$$

$$\sup_{0\leq s\leq t} |\Gamma_2(x_1)(s) - \Gamma_2(x_2)(s)| \leq \sup_{0\leq s\leq t} |x_1(s) - x_2(s)|. \tag{71}$$

A.2 Nonexistence of relevant scaling for $\alpha \in [0, \frac{1}{2})$

Here we provide an argument showing that for α in the range $[0, \frac{1}{2})$ there can be no rescaling of the server population process under which the pair of processes (queue length, server population) remains asymptotically coupled. This is argued by proving the following claim: *Given any $T \in (0,\infty)$, the unnormalized process V^n, if started at zero, remains zero on the interval $[0,T]$ with probability tending to 1 as $n \to 0$.*

To prove the claim, let us first show that Lemma 3.3 remains valid for this range of α. By (29) and (13), $0 \leq e_X^n(t) = n^{-1/2}(N^n - X^n(t))^+ \leq n^{-1/2}N^n \leq n^{\alpha-1/2} \to 0$.

Next consider the equation (12) for $V^n(t)$ with initial condition $V^n(0) = 0$. Let s_1^n denote the first time when V^n assumes the value 1. Our goal is to show that $P(s_1^n \leq T) \to 0$.

In equation (30), the term $\int_0^t n^{-1}\mu^n \tilde{V}^n(s)ds$ vanishes for all $t \leq s_1^n$. The remaining terms in (30) are C-tight (recall $e_X^n \Rightarrow 0$), and thus $\|\xi^n\|_{s_1^n\wedge T}$ is a tight sequence of RVs. As a result of (32), this is true also for $\|L^n\|_{s_1^n\wedge T}$. By (31) and (20) and the convergence μ^n/n to a positive constant, we obtain that $k_n := n^{-\alpha+\frac{1}{2}} \int_0^{s_1^n\wedge T} I^n(s)ds$ is a tight sequence of RVs. By the equation (12) for V^n, and the definition of s_1^n, $S_B(\beta^n \int_0^{s_1^n} I^n(s)ds) = 1$. Consequently $\int_0^{s_1^n} I^n(s)ds \geq c\tau_1$, where τ_1 is the first jump time of S_B, that is specifically an exponential with parameter 1. Combining these facts,

$$P(s_1^n \leq T) \leq P\left(\int_0^{s_1^n\wedge T} I^n(s)ds \geq c\tau_1\right) = P(k_n \geq cn^{-\alpha+\frac{1}{2}}\tau_1) \to 0,$$

by the tightness of k_n. □

A.3 Proof of convergence stated in Remark 2.3

By the convergence of $\tilde{V}^n$ asserted in Theorem 2.1, the term $n^{\alpha-1}\tilde{V}^n$ that appears in identity (23) is equal to $\tilde{V}^n$ when $\alpha = 1$, and converges to zero when $\alpha < 1$. The convergence

$(\hat{X}^n, \tilde{V}^n, \hat{Q}^n) \Rightarrow (X, V, Q)$ stated in Remark 2.3 follows, with Q defined differently according as $\alpha = 1$ or $\alpha < 1$, as in (18).

Acknowledgment. The authors are grateful to Professor Ward Whitt for referring them to [30] for formula (60). They are also grateful to two referees for comments that have greatly improved the presentation of this work. Research of RA supported in part by the ISF (grant 1184/16). Research of DL supported in part by a Zuckerman fellowship.

References

[1] R. Atar. A diffusion regime with nondegenerate slowdown. *Operations Research*, 60(2):490–500, 2012.

[2] R. Atar and I. Gurvich. Scheduling parallel servers in the nondegenerate slowdown diffusion regime: Asymptotic optimality results. *The Annals of Applied Probability*, 24(2):760–810, 2014.

[3] R. Atar, A. Mandelbaum, and M. I. Reiman. Scheduling a multi class queue with many exponential servers: asymptotic optimality in heavy traffic. *Ann. Appl. Probab.*, 14(3):1084–1134, 2004.

[4] R. Atar and M. Shifrin. An asymptotic optimality result for the multiclass queue with finite buffers in heavy traffic. *Stochastic Systems*, 4(2):556–603, 2015.

[5] P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, second edition, 1999.

[6] K. L. Chung and R. J. Williams. *Introduction to Stochastic Integration.* Birkhaüser, Boston, 1990.

[7] B. T. Doshi. Queueing systems with vacations: a survey. *Queueing Systems*, 1(1):29–66, 1986.

[8] D. Gamarnik and D. A. Goldberg. Steady-state G/G/N queue in the Halfin–Whitt regime. *The Annals of Applied Probability*, 23(6):2382–2419, 2013.

[9] D. Gamarnik and A. L. Stolyar. Multiclass multiserver queueing system in the halfin–whitt heavy traffic regime: Asymptotics of the stationary distribution. *Queueing Systems*, 71(1-2):25–51, 2012.

[10] A. Gandhi, S. Doroudi, M. Harchol-Balter, and A. Scheller-Wolf. Exact analysis of the m/m/k/setup class of markov chains via recursive renewal reward. *Queueing Systems*, 77(2):177–209, 2014.

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

[12] V. Gupta and N. Walton. Load balancing in the nondegenerate slowdown regime. *Operations Research*, 67 (1):281–294, 2019.

[13] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations research*, 29(3):567–588, 1981.

[14] J. M. Harrison and A. Zeevi. Dynamic scheduling of a multiclass queue in the halfin-whitt heavy traffic regime. *Operations Research*, 52(2):243–257, 2004.

[15] R. Hassin. *Rational Queueing.* Chapman and Hall/CRC, 2016.

[16] D. J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, 43(3):525–546, 2001.

[17] H. Kaspi and K. Ramanan. Law of large numbers limits for many-server queues. *Ann. Appl. Probab.*, 21 (1):33–114, 2011.

[18] H. Kaspi, K. Ramanan, et al. Spde limits of many-server queues. *The Annals of Applied Probability*, 23 (1):145–229, 2013.

[19] O. Kella and W. Whitt. Diffusion approximations for queues with server vacations. *Advances in Applied Probability*, 22(3):706–729, 1990.

[20] O. Kella and W. Whitt. Queues with server vacations and Lévy processes with secondary jump input. *The Annals of Applied Probability*, 1(1):104–117, 1991.

[21] P.-L. Lions and A.-S. Sznitman. Stochastic differential equations with reflecting boundary conditions. *Communications on Pure and Applied Mathematics*, 37(4):511–537, 1984.

[22] H. Lu, G. Pang, and Y. Zhou. $G/GI/N(+GI)$ queues with service interruptions in the Halfin-Whitt regime. *Mathematical Methods of Operations Research*, 83(1):127–160, 2016.

[23] G. Pang and W. Whitt. Heavy-traffic limits for many-server queues with service interruptions. *Queueing Systems*, 61(2-3):167, 2009.

[24] J. Pender and T. Phung-Duc. A law of large numbers for m/m/c/delayoff-setup queues with nonstationary arrivals. In *International conference on analytical and stochastic modeling techniques and applications*, pages 253–268. Springer, 2016.

[25] P. Protter. *Stochastic Differential Equations*. Springer, Berlin, Heidelberg, 2005.

[26] A. A. Puhalskii and M. I. Reiman. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 32(2):564–595, 2000.

[27] M. I. Reiman. The heavy traffic diffusion approximation for sojourn times in Jackson networks. In *Applied probability–computer science: the interface*, pages 409–421. Springer, 1982.

[28] N. Tian, Q.-L. Li, and J. Gao. Conditional stochastic decompositions in the m/m/c queue with server vacations. *Stochastic Models*, 15(2):367–377, 1999.

[29] M. van der Boor, S. C. Borst, J. S. van Leeuwaarden, and D. Mukherjee. Scalable load balancing in networked systems: A survey of recent advances. *arXiv preprint arXiv:1806.05444*, 2018.

[30] W. Whitt. A diffusion approximation for the G/GI/n/m queue. *Operations Research*, 52(6):922–941, 2004.