CrossMark

J. Appl. Prob. **58,** 428–448 (2021)
*doi:*10.1017/jpr.2020.99

SENSITIVITY OF MEAN-FIELD FLUCTUATIONS IN ERLANG LOSS MODELS WITH RANDOMIZED ROUTING

THIRUPATHAIAH VASANTAM,* *University of Massachusetts, Amherst*
RAVI R. MAZUMDAR,** *University of Waterloo*

Abstract

In this paper we study a large system of N servers, each with capacity to process at most C simultaneous jobs; an incoming job is routed to a server if it has the lowest occupancy amongst d (out of N) randomly selected servers. A job that is routed to a server with no vacancy is assumed to be blocked and lost. Such randomized policies are referred to JSQ(d) (Join the Shortest Queue out of d) policies. Under the assumption that jobs arrive according to a Poisson process with rate $N\lambda^{(N)}$ where $\lambda^{(N)} = \sigma - \frac{\beta}{\sqrt{N}}$, $\sigma \in \mathbb{R}_+$ and $\beta \in \mathbb{R}$, we establish functional central limit theorems for the fluctuation process in both the transient and stationary regimes when service time distributions are exponential. In particular, we show that the limit is an Ornstein–Uhlenbeck process whose mean and variance depend on the mean field of the considered model. Using this, we obtain approximations to the blocking probabilities for large N, where we can precisely estimate the accuracy of first-order approximations.

Keywords: Loss models; JSQ(d); Halfin–Whitt regime; functional central limit theorems; fluctuations; mean field

2010 Mathematics Subject Classification: Primary 60K35
Secondary 60F17; 60M20; 68M20

1. Introduction

This paper is motivated by the design of load-balancing algorithms for cloud computing systems such as Microsoft's Azure [12] and Amazon EC2 [1], where Erlang-type loss models are the appropriate mathematical abstraction. The models of interest are large numbers N of multi-server loss systems where each server has the capacity to process at most C jobs simultaneously, where C is independent of N. A job routed to a server will be accepted for service only if the occupancy (the number of progressing jobs) of the server is less than C, otherwise the routed job will be blocked from service and it is considered to be lost. If a routed job gets accepted at its destination, then its processing begins immediately at a constant unit rate until its service is completed. We assume that the service times are exponentially distributed with unit mean.

Received 11 July 2019; revision received 29 September 2020.
* Postal address: College of Information and Computer Sciences, Amherst, MA 01003, USA. E-mail address: tvasantam@cs.umass.edu
** Postal address: Department of Electrical and Computer Engineering, 200 University Ave W, Waterloo, ON N2L 3G1, Canada. E-mail address: mazum@uwaterloo.ca

© The Author(s), 2021. Published by Cambridge University Press on behalf of Applied Probability Trust.

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

We focus on randomized routing policies where arriving jobs or tasks are routed to the server with the least number of jobs amongst d servers chosen at random. These policies are referred to as JSQ(d) (Join the Shortest Queue out of d) policies. We consider the situation where the arrival rate to the system is given by a Poisson process with rate $N\lambda^{(N)}$ where $\lambda^{(N)} = \sigma - \frac{\beta}{\sqrt{N}}$, $\sigma \in \mathbb{R}_+$ and $\beta \in \mathbb{R}$. Thus, $\lambda^{(N)}$ is a fluctuation in the arrival rate that is nominally σ. The objective in this paper is to study the sensitivity of the blocking probabilities under JSQ(d). We also show that this result can then be exploited to obtain approximation errors.

A number of papers have studied the limiting behavior of blocking probabilities in Erlang loss models when $N \to \infty$ for both homogeneous systems [15, 22] and heterogeneous models (servers with differing capacities) [10, 16]. When N is infinite, the limiting distributions can be obtained via the mean-field limit of the empirical occupancy distribution, an approach that goes back to the work in [13, 20] for M/M/1 models incorporating JSQ(d) and was popularized as the *power-of-two* principle, where it was argued that most of the gains in the average delay are obtained when taking $d = 2$. The mean-field analysis also establishes the asymptotic statistical independence on path space [8] for the limiting occupancy processes. The results show that the blocking probability in the limiting system is very close to the theoretical lower bound on blocking that can be achieved by any non-anticipative policy, and thus approximates the optimal desirable behavior [10].

A key question is how well the mean-field limit describes the occupancy distribution and the blocking probabilities when N is finite but large. Recently, there have been a number of papers [6, 23] that have addressed this issue for M/M/1 queueing models for the $\beta = 0$ case where the limiting stationary distribution can be characterized explicitly as a double-exponential distribution. They used an approach based on Stein's method and showed that the rate of convergence of the empirical occupancy distribution to the mean-field distribution is $O\big(\frac{1}{\sqrt{N}}\big)$. In [6] a refined $O\big(\frac{1}{N}\big)$ term is also given. These approaches use Stein's method and exponential stability of the underlying mean-field equation to study the mean-squared error between the empirical distribution and the mean-field limit to characterize the rate of convergence. In [5], similar proof techniques based on Stein's method were used to show that for any twice-differentiable bounded function $f(\cdot)$ the gap between $\mathbb{E}[f(X^{(N)}(t))]$ and $f(x(t))$ is $O\big(\frac{1}{N}\big)$ in both transient and stationary regimes, where $X^{(N)}(t)$ and $x(t)$ are the considered empirical distributions and the mean-field limit, respectively. The results of [5] can be used for Erlang loss models for the special case $\beta = 0$ to conclude that the error between the average blocking probability of the system with N servers and the asymptotic blocking probability is $O\big(\frac{1}{N}\big)$. The results of [5] are not, however, applicable when $\beta \neq 0$, and are essentially weak convergence results.

Our approach is via the development of functional central limit theorems (FCLTs) for the variation of the empirical distribution around the mean field, which we term the *fluctuation process*. These limit theorems allow us to study both the transient and stationary fluctuations by showing convergence to an appropriate Ornstein–Uhlenbeck process whose drift and noise variance depend on the mean-field limit of the model. A by-product of our study of FCLTs is that the error between the average blocking probability of the system with N servers and the asymptotic blocking probability is $O\big(\frac{1}{\sqrt{N}}\big)$ for $\beta \neq 0$.

Recently, Eschenfeldt and Gamarnik [3] also studied the FCLT scaling of the queue occupancy process for a system of M/M/1 queues in the diffusion limit with JSQ; they showed that asymptotically the distribution concentrates on queues having up to two customers. They did not consider the distributional aspects or the mean-field issues.

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

Clearly, $\lambda^{(N)}$ is a perturbation of σ by $\frac{\beta}{\sqrt{N}}$. The Halfin–Whitt regime corresponds to the case when the offered load to the system is very close to the system capacity NC and corresponds to the special case when we choose $\sigma = C$, i.e. jobs arrive according to a Poisson process with rate $N\lambda^{(N)}$ where $\lambda^{(N)} = C - \frac{\beta}{\sqrt{N}}$. This implies that, for large N, $\frac{\lambda^{(N)}}{C} = 1 - \frac{\beta}{C\sqrt{N}}$ is very close to 1, implying that the system is critically loaded. We are interested in studying approximations to the blocking probability for the system when N is large but a finite value when JSQ(d) is used.

The interest in the Halfin–Whitt regime is because there is a phase change in the behavior of the blocking probabilities going from exponential decrease (in C) to a $\frac{1}{\sqrt{C}}$ scaling as C (see [7]) becomes large for the case of uniform routing. It can be shown that an equivalent result is valid for the case of complete resource pooling when N becomes large with fixed C, i.e. a loss system with a single server having an arrival rate of $N\lambda^{(N)}$ with server capacity NC. Although loss systems are stable for any finite average load, this phase change in blocking behavior has implications for dimensioning the system.

We first provide an overview of the system performance when the system is in the Halfin–Whitt regime. Clearly, the average blocking probability depends on how efficiently we use system resources. For example, let us consider the random routing case where an arrival is routed to a randomly selected server under the Halfin–Whitt regime. Then, the average blocking probability experienced by an arrival is the same as in the single-server loss system with capacity C where the jobs arrive as a Poisson process with intensity $\lambda^{(N)}$ (due to thinning). The average blocking probability is equal to $Er(\lambda^{(N)}, C)$, where $Er(\alpha, n)$ denotes the Erlang-B formula for Poisson arrivals with rate α and n servers. Since C is fixed, the average blocking probability converges to $Er(C,C)$ when $N \to \infty$. On the other hand, if we consider the complete resource pooling case in which an arrival is accepted for service as long as there is an empty spot at a server in the system, then the average blocking probability is given by $Er(N\lambda^{(N)}, NC)$. Then, from [21],

$$\lim_{N\to\infty} \sqrt{N}\big(Er\big(N\lambda^{(N)}, NC\big)\big) = \frac{\phi(\beta)}{\sqrt{C}\Phi(\beta)}, \tag{1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and distribution functions of a standard normal distribution.

Now, let $P^{(N)}_{\text{block}}$ be the average blocking probability of an arrival in the system with parameter N when the standard JSQ policy is used. It was shown in [14] that we obtain the same result, (1), as in the case of complete resource pooling, i.e.

$$\lim_{N\to\infty} \sqrt{N} P^{(N)}_{\text{block}} = \frac{\phi(\beta)}{\sqrt{C}\Phi(\beta)}. \tag{2}$$

This is to be expected since an arrival will not be blocked from service when there is an empty spot in the system, similar to the case of complete resource pooling. As a result, the average blocking probability under the JSQ policy is equal to $Er(N\lambda^{(N)}, NC)$. Under the influence of the JSQ(d) policy, an arrival could be blocked from service even if there is an empty spot in the system. Therefore we expect a decrease in the system utilization when we use the JSQ(d) policy. However, such a policy has less informational cost over the JSQ policy and is close to the optimal blocking that can be obtained with full resource sharing. It was shown in [14] that if d is also scaled with N, denoted by $d^{(N)}$, and if $\lim_{N\to\infty} \frac{d^{(N)}}{\sqrt{N}\log(N)} = \infty$, then we still obtain (2) for the JSQ($d^{(N)}$) scheme. It is thus of interest to know what happens when $d \geq 2$ is fixed and does not scale with N.

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

Our approach is similar to the FCLT approach in [9] that was carried out for M/M/1 first-come, first-serve queues with the JSQ(d) policy with $\lambda^{(N)} = b$, $b \in \mathbb{R}_+$. It was shown that suitably scaled fluctuations of the stochastic empirical occupancy process around the mean-field limit converge to an Ornstein–Uhlenbeck (OU) process in both the transient and stationary regimes as $N \to \infty$. However, the paper did not exploit this result further. In this paper we show how the limit theorems can be used to characterize the transient and steady-state system blocking probabilities and thus obtain approximation errors.

In preliminary work [19], we showed a similar FCLT result to [9] for the case of the loss model when $\lambda^{(N)} = b$, $b \in \mathbb{R}_+$. We then used the FCLT limit to show that the error between $P_{\text{block}}^{(N)}$ and the asymptotic blocking probability π_C^d is $o(N^{-\frac{1}{2}})$, where $\boldsymbol{\pi} = (\pi_i, 0 \le i \le C)$ is the fixed point of the corresponding mean field and π_C is the probability that a server is fully occupied when $N \to \infty$.

The FCLT approach has advantages over the Stein approach because it provides a process-level characterization of the scaled mean-squared error rather than just the rate at which the limit of the mean-squared error between the approximation of the stationary distribution for fixed N and the fixed point of the mean field goes to 0 as $N \to \infty$. More precisely, we show that the diffusion-scaled fluctuation process for our model converges to a limit which is an OU process with non-zero mean that depends on β and the fixed point $\boldsymbol{\pi}$ of the mean field in our model. We then exploit this result to show that $\lim_{N\to\infty} \sqrt{N}\big(P_{\text{block}}^{(N)} - \pi_C^d\big)$ goes to a limit that can be explicitly characterized in terms of β, $\boldsymbol{\pi}$, and C. We obtain results for both the transient and stationary occupancy distributions.

It is worth pointing out that the result we obtain is interesting when the overall system is in the Halfin–Whitt regime, i.e. when $\sigma = C$. The effect of the randomized SQ(d) (shortest queue) routing results in individual loss servers that are also critically loaded but whose blocking cannot be obtained from the classical Halfin–Whitt blocking limit; instead, the blocking is obtained from the fixed point of the mean field.

The rest of the paper is organized as follows. We first introduce the system model and provide some preliminary results in Section 2. We then give the main results of the paper in Section 3, and provide their proofs in Section 4. Finally, we conclude the paper in Section 5.

2. System model and preliminaries

2.1. System model

In this section we give details of the system model. We study a large-scale multi-server system with N Erlang loss servers and one central job dispatcher, which routes an incoming request to one of the servers according to a predefined load-balancing policy. A server accepts an incoming request if the occupancy or the number of progressing jobs of the server is less than C, a predefined integer value referred to as the capacity of a server. Otherwise, the request is blocked from service and it is considered to be discarded from the system. Furthermore, an accepted job is processed at a constant unit rate upon its acceptance for service until the job is completed. We assume that the service time distributions are exponential with unit mean. The job dispatcher uses the JSQ(d) load-balancing policy to dispatch the incoming jobs.

Definition 1. JSQ(d) load balancing: The job dispatcher routes an arriving job to the server with the least occupancy among d servers selected uniformly at random. Furthermore, the ties are assumed to be broken uniformly at random.

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

Remark 1. For our model, it does not matter whether the dispatcher samples d servers with or without replacement to dispatch an arrival as both methods lead to the same asymptotic results. The proof follows by the same arguments as [9, pages 11-12]. Hence, we assume that the dispatcher samples d servers with replacement upon an arrival to simplify the analysis.

The arrival process of jobs is a Poisson process with rate $N\lambda^{(N)}$, where the parameter $\lambda^{(N)} \in \mathbb{R}_+$ is defined such that, for $\sigma \in \mathbb{R}_+$ and $\beta \in \mathbb{R}$,

$$\lambda^{(N)} = \sigma - \frac{\beta}{\sqrt{N}}. \tag{3}$$

Clearly, $\lambda^{(N)}$ is a perturbation of σ by $\frac{\beta}{\sqrt{N}}$. For our model, we first show that there exists a functional law of large numbers limit referred to as the mean-field limit. Next, we establish an FCLT result, which is exploited to quantify the error between the actual blocking probability of the system with N servers and the asymptotic blocking probability expressed as a function of the unique fixed point of the mean field. The particular form of $\lambda^{(N)}$ in (3) subsumes an important special case of $\sigma = C$ that corresponds to the Halfin–Whitt heavy traffic regime as the resulting traffic intensity $\rho^{(N)} = \frac{\lambda^{(N)}}{C}$ approaches 1 as $N \to \infty$ and $\lim_{N\to\infty} \sqrt{N}(1 - \rho^{(N)}) = \frac{\beta}{C}$. In this case, the arrival rate of jobs, $N\lambda^{(N)}$, and the system capacity, NC, are related as $NC = N\lambda^{(N)} + \beta\sqrt{N}$, and they converge to ∞ as $N \to \infty$.

2.2. Notation

Since the job dispatcher uses only the knowledge of the server occupancies and does not use their identities to dispatch an arrival, we consider the Markov process $(\mathbf{X}^{(N)}(t), t \geq 0)$ to model the time evolution of the system, where $\mathbf{X}^{(N)}(t) = (\mathbf{X}_i^{(N)}(t), 0 \leq i \leq C)$, with $\mathbf{X}_i^{(N)}(t)$ denoting the fraction of the servers with at least i progressing jobs at time t. Let $\mathbb{U}$ be the space defined as

$$\mathbb{U} \triangleq \{(u_0, u_1, \ldots, u_C) : u_0 = 1 \geq u_1 \geq \cdots \geq u_C \geq 0\}.$$

It is evident that $\mathbf{X}^{(N)}(t)$ lies in the space $\mathbb{U}$. Without loss of generality, we write an element of the form $(u_0, \ldots, u_C)$ as $\mathbf{u}$. The space $\mathbb{U}$ is equipped with the metric generated by the Euclidean norm $\|\cdot\|_2$, defined as

$$\|\mathbf{u}\|_2 = \sqrt{\sum_{i=0}^{C} |u_i|^2}, \tag{4}$$

where $\mathbf{u} = (u_0, \ldots, u_C)$. It can be verified that the space $\mathbb{U}$ is a Polish space.

We study stochastic processes that are defined on $(\Omega, \mathbb{F}, \mathbb{P})$ with sample paths belonging to the space of right continuous functions with left limits (such functions are also called càdlàg functions). The space of càdlàg functions is equipped with the Skorohod J_1 topology. We write $\mathbf{Y}_n \Rightarrow \mathbf{Y}$ as $n \to \infty$ to indicate that a sequence of random elements $\{\mathbf{Y}_n\}_{n\geq 1}$ converges in distribution to a random element $\mathbf{Y}$. For two real-valued local martingales $(\mathbf{M}^1(t), t \geq 0)$ and $(\mathbf{M}^2(t), t \geq 0)$, let the covariation process be denoted by $(\langle \mathbf{M}^1, \mathbf{M}^2 \rangle_t, t \geq 0)$ and the quadratic variation process be denoted by $(\langle \mathbf{M}^1 \rangle_t, t \geq 0) = (\langle \mathbf{M}^1, \mathbf{M}^1 \rangle_t, t \geq 0)$.

2.3 Preliminaries

In this section we present a mathematical model of the main problem and give some preliminary results. We first begin with a discussion on the time evolution of the process

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

$(\mathbf{X}^{(N)}(t), t \geq 0)$. At an arrival instant t, if the system state is $\mathbf{b} = (b_0, \ldots, b_C)$, implying that the fraction of the servers with at least i progressing jobs at time t is equal to b_i for $0 \leq i \leq C$, then according to the JSQ(d) policy the destination server of the job will have occupancy n with probability $b_n^d - b_{n+1}^d$. Since the rate of the arrival process is $N\lambda^{(N)}$, the total instantaneous rate of arrivals to servers that have n jobs is equal to $N\lambda^{(N)}((\mathbf{X}_{n-1}^{(N)}(t))^d - (\mathbf{X}_n^{(N)}(t))^d)$. Furthermore, since the service times have exponential distributions with unit rate, the total instantaneous departure rate of jobs from servers with n progressing jobs is equal to $nN(\mathbf{X}_n^{(N)}(t) - \mathbf{X}_{n+1}^{(N)}(t))$. As a result, we can model the time evolution of the process $(\mathbf{X}^{(N)}(t), t \geq 0)$ by using random time change of a set of mutually independent unit-rate Poisson processes as in [17, Section 2.1], which we explain below.

Let $\{(\mathcal{N}_i(t), t \geq 0)\}_{i \geq 1}$ be a set of mutually independent unit-rate Poisson processes where $(\mathcal{N}_i(t), t \geq 0)$ is used to model the arrival process to servers that have $i-1$ progressing jobs. Similarly, let $\{(\mathcal{D}_i(t), t \geq 0)\}_{i \geq 1}$ be the collection of a set of mutually independent unit-rate Poisson processes where $(\mathcal{D}_i(t), t \geq 0)$ is used to model the departure process from servers that have i progressing jobs. Furthermore, the set of processes $\{(\mathcal{D}_i(t), t \geq 0)\}_{i \geq 1}$ is independent of the set of processes $\{(\mathcal{N}_i(t), t \geq 0)\}_{i \geq 1}$. Also, $\{(\mathcal{N}_i(t), t \geq 0)\}_{i \geq 1}$ and $\{(\mathcal{D}_i(t), t \geq 0)\}_{i \geq 1}$ are independent of $\mathbf{X}^{(N)}(0)$. Since the arrival process of jobs to the system is a Poisson process with rate $N\lambda^{(N)}$, and the service time distributions are exponential with unit mean, we can write $\mathbf{X}_0^{(N)}(t) = 1$ and, for $n \geq 1$,

$$\begin{aligned}\mathbf{X}_n^{(N)}(t) =& \mathbf{X}_n^{(N)}(0) + \frac{1}{N}\mathcal{N}_n\left(N\lambda^{(N)} \int_{s=0}^{t} \left(\left(\mathbf{X}_{n-1}^{(N)}(s)\right)^d - \left(\mathbf{X}_n^{(N)}(s)\right)^d\right) \mathrm{d}s\right) \\ & - \frac{1}{N}\mathcal{D}_n\left(Nn \int_{s=0}^{t} \left(\left(\mathbf{X}_n^{(N)}(s)\right) - \left(\mathbf{X}_{n+1}^{(N)}(s)\right)\right) \mathrm{d}s\right).\end{aligned} \tag{5}$$

We choose the filtration $(\mathcal{F}^{(N)}(t), t \geq 0)$ where

$$\begin{aligned}\mathcal{F}^{(N)}(t) =& \sigma\left(\mathbf{X}^{(N)}(0), \mathcal{N}_n\left(N\lambda^{(N)} \int_{s=0}^{r} \left(\left(\mathbf{X}_{n-1}^{(N)}(s)\right)^d - \left(\mathbf{X}_n^{(N)}(s)\right)^d\right) \mathrm{d}s\right),\right. \\ & \left. \mathcal{D}_n\left(Nn \int_{s=0}^{r} \left(\left(\mathbf{X}_n^{(N)}(s)\right) - \left(\mathbf{X}_{n+1}^{(N)}(s)\right)\right) \mathrm{d}s\right), n \geq 1, 0 \leq r \leq t\right),\end{aligned}$$

augmented by all the null sets.

We now present the results on the mean-field analysis of the model without proofs as they directly follow from the case of $\lambda^{(N)} = b$ for $b \in \mathbb{R}_+$, studied in [15]. The resulting mean-field equations (MFEs) in our case are the same as for $\lambda^{(N)} = b$ except that σ replaces b.

Theorem 1. *For $\mathbf{u} \in \mathbb{U}$, if $\mathbf{X}^{(N)}(0) \Rightarrow \mathbf{u}$ as $N \to \infty$, then $\left(\mathbf{X}^{(N)}(t), t \geq 0\right) \Rightarrow (\mathbf{x}(t, \mathbf{u}), t \geq 0)$ as $N \to \infty$, where $(\mathbf{x}(t, \mathbf{u}), t \geq 0) = (x_n(t, \mathbf{u}), t \geq 0,\ 0 \leq n \leq C)$ is the unique solution to the following: let $\mathbf{h}(\mathbf{x}(t, \mathbf{u})) = (h_n(\mathbf{x}(t, \mathbf{u})), 0 \leq n \leq C)$, where*

$$\mathbf{x}(0, \mathbf{u}) = \mathbf{u}, \qquad \frac{\mathrm{d}x_n(t, \mathbf{u})}{\mathrm{d}t} = h_n(\mathbf{x}(t, \mathbf{u})), \tag{6}$$

satisfying

$$h_0(\mathbf{x}(t, \mathbf{u})) = 0 \tag{7}$$

and, for $n \geq 1$,

$$h_n(\mathbf{x}(t, \mathbf{u})) = \sigma\left(x_{n-1}^d(t, \mathbf{u}) - x_n^d(t, \mathbf{u})\right) - n(x_n(t, \mathbf{u}) - x_{n+1}(t, \mathbf{u})), \tag{8}$$

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

with $x_0(t, \mathbf{u}) = 1$ *and* $x_{C+1}(t, \mathbf{u}) = 0$. *The deterministic process* $(\mathbf{x}(t, \mathbf{u}), t \geq 0)$ *is referred to as the mean-field limit and equations* (6)–(8) *are referred to as the mean-field equations with initial point* $\mathbf{u}$.

Without loss of generality, we say that if a process $(\mathbf{y}(t), t \geq 0)$ is a solution to the differential equations (6)–(8) then it is the unique generic solution with initial point $\mathbf{y}(0)$.

The mean field $(\mathbf{x}(t, \mathbf{u}), t \geq 0)$ has a unique global asymptotically stable fixed point $\boldsymbol{\pi} = (\pi_n, 0 \leq n \leq C)$ with $\pi_0 = 1$. Also, the following exchange of limits holds:

$$\lim_{N\to\infty} \lim_{t\to\infty} \mathbf{X}^{(N)}(t) = \lim_{t\to\infty} \lim_{N\to\infty} \mathbf{X}^{(N)}(t).$$

Using this, under the assumption of the exchangeability of initial states of servers, we can show the independence of any finite set of servers as $N \to \infty$. Also, it can be shown that as $N \to \infty$ a server's distribution equals $\mathbf{x}(t, \mathbf{u})$ and π at time t and in the stationary regime, respectively. As a result, π_C denotes the stationary probability that a server is fully occupied as $N \to \infty$. Since the dispatcher samples d servers upon arrival and selects one of them as the destination server, the stationary average blocking probability of a job as $N \to \infty$ is equal to π_C^d, where we use the fact that the chosen d servers are independent of each other. Our objective is to find the gap between the actual blocking probability of the system with N servers and π_C^d as a function of the parameters β, $\boldsymbol{\pi}$, σ, N, and C.

We can find $\boldsymbol{\pi}$ numerically as follows. The fixed point $\boldsymbol{\pi}$ is the unique solution to

$$\sigma\left(\pi_{n-1}^d - \pi_n^d\right) = n(\pi_n - \pi_{n+1})$$

for $n \geq 1$ and $\pi_{C+1} = 0$. Then, from there, we can also write

$$\sigma \frac{\left(\pi_{n-1}^d - \pi_n^d\right)}{(\pi_{n-1} - \pi_n)}(\pi_{n-1} - \pi_n) = n(\pi_n - \pi_{n+1}) \tag{9}$$

for $n \geq 1$ and $\pi_{C+1} = 0$. Let us define $\hat{\lambda}_n = \sigma \frac{\left(\pi_n^d - \pi_{n+1}^d\right)}{(\pi_n - \pi_{n+1})}$. Then, from (8), $\boldsymbol{\pi}$ is the stationary distribution of the single-server loss model with a Poisson arrival process of jobs having rate $\hat{\lambda}_n$ when there are n progressing jobs, and π_n is the probability that the server has at least n progressing jobs. Let $\mathbb{M}_1(\{0, 1, \ldots, C\})$ be the set of probability measures on $\{0, 1, \ldots, C\}$. Then, from [15], the fixed point $\boldsymbol{\pi}$ can be computed using the formula for the stationary distribution of a single-server loss system with state-dependent arrival rates. We first define two mappings, $\Theta : \mathbb{M}_1(\{0, 1, \ldots, C\}) \mapsto \mathbb{R}_+^{C+1}$ and $\widehat{\Xi} : \mathbb{R}_+^{C+1} \mapsto \mathbb{M}_1(\{0, 1, \ldots, C\})$, that are used in computing $\boldsymbol{\pi}$. For every $(p_0, \ldots, p_C) \in \mathbb{M}_1(\{0, 1, \ldots, C\})$, there exists $(r_0, \ldots, r_C) \in \mathbb{R}_+^{C+1}$ such that $\Theta((p_0, \ldots, p_C)) = (r_0, \ldots, r_C)$, where

$$r_n = \sigma \frac{\left(\left(\sum_{j=n}^C p_j\right)^d - \left(\sum_{i=n+1}^C p_i\right)^d\right)}{\left(\left(\sum_{j=n}^C p_j\right) - \left(\sum_{i=n+1}^C p_i\right)\right)}.$$

Similarly, for every $(b_0, \ldots, b_C) \in \mathbb{R}_+^{C+1}$, there exists $(a_0, \ldots, a_C) \in \mathbb{M}_1(\{0, 1, \ldots, C\})$ such that $\widehat{\Xi}((b_0, \ldots, b_C)) = (a_0, \ldots, a_C)$, where

$$a_n = \left(\prod_{i=1}^n \left(\frac{b_{i-1}}{i}\right)\right) a_0$$

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

for $n \geq 1$ and $\sum_{i=0}^{C} a_i = 1$. Then, $\boldsymbol{\pi}$ is the unique fixed point of the mapping $\widehat{\Xi}(\Theta)$, which can be computed numerically.

Our objective is to study the limit of the fluctuation process $(\mathbf{Z}^{(N)}(t), t \geq 0)$ as $N \to \infty$, where

$$\mathbf{Z}^{(N)}(t) = \sqrt{N}(\mathbf{X}^{(N)}(t) - \mathbf{x}(t, \mathbf{u})). \tag{10}$$

It can be checked that $\mathbf{Z}^{(N)}(t)$ lies in the space $\mathbb{V}$ defined as

$$\mathbb{V} \triangleq \{(r_0, \ldots, r_C) : r_0 = 0 \text{ and } r_i \in \mathbb{R}, 1 \leq i \leq C\}.$$

We equip the space $\mathbb{V}$ with the topology induced by the Euclidean norm (4). Our analysis uses the operator norm $\|\cdot\|_2$ defined as

$$\|K\|_2 = \sup_{\mathbf{v} \in \mathbb{V}} \frac{\|K\mathbf{v}\|_2}{\|\mathbf{v}\|_2},$$

where $K : \mathbb{V} \mapsto \mathbb{V}$ is a linear operator.

Next, we obtain the time evolution of the process $(\mathbf{Z}^{(N)}(t), t \geq 0)$ by using (5), (6), and (10). For this, we first define the following three useful operators $W_1, W_2, W_3 : \mathbb{U} \mapsto \mathbb{V}$. For $\mathbf{b} \in \mathbb{U}$ with $1 \leq i \leq 3$ we define $W_i(\mathbf{b}) = ((W_i(\mathbf{b}))_n, n \geq 0)$, where $(W_1(\mathbf{b}))_0 = 0$, $(W_2(\mathbf{b}))_0 = 0$, $(W_3(\mathbf{b}))_0 = 0$, and, for $1 \leq n \leq C$,

$$(W_1(\mathbf{b}))_n = \sigma\left(b_{n-1}^d - b_n^d\right), \quad (W_2(\mathbf{b}))_n = n(b_n - b_{n+1}), \quad (W_3(\mathbf{b}))_n = \beta\left(b_{n-1}^d - b_n^d\right). \tag{11}$$

From (8) and (11), we have $h_n(\mathbf{b}) = (W_1(\mathbf{b}))_n - (W_2(\mathbf{b}))_n$. Furthermore, let $W : \mathbb{U} \mapsto \mathbb{V}$ be the operator defined as $W = W_1 - W_2$. The operator W is Lipschitz continuous, satisfying, for all $\mathbf{a}, \mathbf{b} \in \mathbb{U}$, $\|W(\mathbf{a}) - W(\mathbf{b})\|_2 \leq B_W \|\mathbf{a} - \mathbf{b}\|_2$, where $B_W = 2d\sqrt{\sigma^2 + C^2}$.

Now we define a set of independent square-integrable martingales $(\mathbf{M}^{(N)}(t), t \geq 0) = \{(\mathbf{M}_i^{(N)}(t), t \geq 0)\}_{(i \in 0,1,\ldots,C)}$ adapted to the filtration $(\mathcal{F}^{(N)}(t), t \geq 0)$ such that $(\mathbf{M}^{(N)}(t), t \geq 0)$ is independent of $\mathbf{Z}^{(N)}(0)$ and, for $i \geq 1$,

$$\langle \mathbf{M}_i^{(N)} \rangle_t = \int_{s=0}^{t} \left(\left(W_1\left(\mathbf{X}^{(N)}(s)\right)\right)_i + \left(W_2\left(\mathbf{X}^{(N)}(s)\right)\right)_i - \frac{1}{\sqrt{N}} \left(W_3\left(\mathbf{X}^{(N)}(s)\right)\right)_i \right) \mathrm{d}s. \tag{12}$$

Then, from (5)–(8) and (10), we get

$$\begin{aligned} \mathbf{Z}^{(N)}(t) = &\mathbf{Z}^{(N)}(0) + \int_{s=0}^{t} \sqrt{N}\left(W\left(\mathbf{X}^{(N)}(s)\right) - W\left(\mathbf{x}(s, \mathbf{u})\right)\right) \mathrm{d}s \\ &- \int_{s=0}^{t} W_3\left(\mathbf{X}^{(N)}(s)\right) \mathrm{d}s + \mathbf{M}^{(N)}(t). \end{aligned} \tag{13}$$

3. Summary of main results

In this section we give the main results; their proofs are presented in Section 4. We present the results related to the transient regime and the stationary regime in Sections 3.1 and 3.2, respectively.

3.1. Transient regime

In this section we show that the process $(\mathbf{Z}^{(N)}(t), t \geq 0)$ converges to an OU process in the transient regime as $N \to \infty$. First, we begin with the following result that concludes stochastic

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

boundedness of the process $(\mathbf{Z}^{(N)}(t), t \geq 0)$ when $N \to \infty$. We use this property in proving the tightness of the sequence $\{(\mathbf{Z}^{(N)}(t), t \geq 0)\}_{N \geq 1}$.

Lemma 1. *For any $T > 0$, if $\limsup_{N\to\infty} \mathbb{E}\left[\|\mathbf{Z}^{(N)}(0)\|_2^2\right] < \infty$, then*

$$\limsup_{N\to\infty} \mathbb{E}\left[\sup_{0\leq t\leq T} \|\mathbf{Z}^{(N)}(t)\|_2^2\right] < \infty. \tag{14}$$

Proof. See Section 4.1. □

As we show later in this section, any limit point of $(\mathbf{Z}^{(N)}(t), t \geq 0)$ is a solution to the stochastic differential equation (SDE) defined in (20). Next, we introduce some notation that is used in (20). The proposed SDE depends on $(\mathbf{s}(t), t \geq 0)$, which is a solution of (18). Let $(\mathbf{w}(t), t \geq 0)$ be a generic solution of (6)–(8) with an initial point $\mathbf{w}(0)$, where $\mathbf{w}(t) = (w_n(t), 0 \leq n \leq C+1)$ satisfying $w_0(t) = 1$ and $w_{C+1}(t) = 0$. Then we have

$$\frac{\mathrm{d}w_n(t)}{\mathrm{d}t} = h_n(\mathbf{w}(t)), \tag{15}$$

where

$$h_0(\mathbf{w}(t)) = 0 \tag{16}$$

and, for $1 \leq n \leq C$,

$$h_n(\mathbf{w}(t)) = \sigma\left(w_{n-1}^d(t) - w_n^d(t)\right) - n(w_n(t) - w_{n+1}(t)). \tag{17}$$

By linearizing (15)–(17) around a solution $(\mathbf{r}(t), t \geq 0)$ of (15)–(17) with an initial point $\mathbf{r}(0)$, we get

$$\frac{\mathrm{d}\mathbf{s}(t)}{\mathrm{d}t} = H(\mathbf{r}(t))\mathbf{s}(t), \tag{18}$$

where, for $\mathbf{a} \in \mathbb{U}$ and $\mathbf{b} \in \mathbb{V}$, the linear operator $H(\mathbf{a}) : \mathbb{V} \mapsto \mathbb{V}$ is defined as

$$(H(\mathbf{a})\mathbf{b})_n = \sigma d a_{n-1}^{d-1} b_{n-1} - \left(\sigma d a_n^{d-1} + n\right) b_n + n b_{n+1},$$

$n \geq 1$. Note that any solution $(\mathbf{s}(t), t \geq 0)$ of (18) satisfies $\mathbf{s}(t) = \mathbf{w}(t) - \mathbf{r}(t)$, where $(\mathbf{w}(t), t \geq 0)$ is a solution of (15)–(17) with an initial point $\mathbf{w}(0)$.

We will show that the limit of the sequence $\{(\mathbf{Z}^{(N)}(t), t \geq 0)\}_{N \geq 1}$ depends on the process $(\mathbf{s}(t), t \geq 0)$ when $(\mathbf{r}(t), t \geq 0)$ in (18) is replaced with the mean field $(\mathbf{x}(t, \mathbf{u}), t \geq 0)$. The operator $H(\mathbf{a})$ is a matrix in the canonical basis $(0, 1, 0, \ldots, 0)$, $(0, 0, 1, 0, \ldots, 0)$, …, $(0, 0, \ldots, 0, 1)$, where the dimension of each vector is $C+1$. We can write $H(\mathbf{a})$ as the following matrix of size $C \times C$:

$$H(\mathbf{a}) = \begin{bmatrix} -\nu_1 & 1 & 0 & 0 & \cdots & 0 \\ \gamma_1 & -\nu_2 & 2 & 0 & \cdots & 0 \\ 0 & \gamma_2 & -\nu_3 & 3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_{C-2} & -\nu_{C-1} & C-1 \\ 0 & 0 & \cdots & 0 & \gamma_{C-1} & -\nu_C \end{bmatrix},$$

where $\gamma_i = \sigma d a_i^{d-1}$ and $\nu_i = \gamma_i + i$, $1 \leq i \leq C$.

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

Let $(\mathbf{M}(t), t \geq 0) = \{(\mathbf{M}_i(t), t \geq 0)\}_{i \in \{0,1,\ldots,C\}}$ be a collection of mutually independent real-valued continuous and centered Gaussian martingales, determined in law by their deterministic quadratic variation process

$$\langle \mathbf{M}_n \rangle_t = \int_{s=0}^{t} ((W_1(\mathbf{x}(s, \mathbf{u}))_n + (W_2(\mathbf{x}(s, \mathbf{u}))_n) \, \mathrm{d}s \tag{19}$$

for $n \geq 0$. Note that both $\mathbf{M}(t)$ and $(\langle \mathbf{M}_i \rangle_t, 0 \leq i \leq C)$ lie in $\mathbb{V}$. From (19), the martingale $(\mathbf{M}(t), t \geq 0)$ is square integrable since $((W_1(\mathbf{b}))_n + (W_2(\mathbf{b}))_n)$ for $\mathbf{b} \in \mathbb{U}$ is uniformly bounded in n and $\mathbf{b}$ because $0 \leq b_i \leq 1$ for $0 \leq i \leq C$.

Now we introduce an SDE; we later show that the limit of the process $(\mathbf{Z}^{(N)}(t), t \geq 0)$ as $N \to \infty$ in the transient regime is a unique solution of this SDE.

Definition 2. SDE for the transient regime: Let $(\mathbf{Z}(t), t \geq 0)$ be a solution of the SDE

$$\mathbf{Z}(t) = \mathbf{Z}(0) + \int_{s=0}^{t} H(\mathbf{x}(s, \mathbf{u}))\mathbf{Z}(s) \, \mathrm{d}s - \int_{s=0}^{t} W_3(\mathbf{x}(s, \mathbf{u})) \, \mathrm{d}s + \mathbf{M}(t). \tag{20}$$

The solution of (20) is an OU process. Next, we study the SDE (20).

Theorem 2.

(i) *For* $\mathbf{a} \in \mathbb{U}$, *the linear operator* $H(\mathbf{a})$ *satisfies* $\|H(\mathbf{a})\|_2 < B_H$, *where* $B_H = \sqrt{32(\sigma^2 d^2 + C^2)}$.

(ii) *If* $\mathbb{E}\left[\|\mathbf{Z}(0)\|_2^2\right] < \infty$, *then there exists a unique strong solution to* (20) *denoted by* $(\mathbf{Z}(t), t \geq 0)$ *that satisfies* $\mathbb{E}\left[\sup_{t \leq T} \|\mathbf{Z}(t)\|_2^2\right] < \infty$.

The proof of Theorem 2 follows by arguments similar to the proof of [19, Theorem 2], and hence we omit the proof here.

Now we present the main result on the transient regime.

Theorem 3. *If* $\mathbf{Z}^{(N)}(0) \Rightarrow \mathbf{Z}(0)$, *then* $(\mathbf{Z}^{(N)}(t), t \geq 0) \Rightarrow (\mathbf{Z}(t), t \geq 0)$, *where* $(\mathbf{Z}(t), t \geq 0)$ *is the unique solution of* (20) *with the initial point* $\mathbf{Z}(0)$.

Proof. See Section 4.2. □

Remark 2. For a constant $a \in \mathbb{R}_+$, if $\lambda^{(N)} = a$ then, for $n \geq 1$,

$$(W_1(\mathbf{b}))_n = a(b_{n-1}^d - b_n^d), \qquad (W_2(\mathbf{b}))_n = n(b_n - b_{n+1}), \qquad (W_3(\mathbf{b}))_n = 0.$$

As a result, from the SDE (20), we recover the following SDE obtained in [19] for the $\lambda^{(N)} = a$ case:

$$\mathbf{Z}(t) = \mathbf{Z}(0) + \int_{s=0}^{t} H(\mathbf{x}(s, \mathbf{u}))\mathbf{Z}(s) \, \mathrm{d}s + \mathbf{M}(t).$$

3.2. Stationary regime

In this section we present results pertaining to the stationary regime. Here, the mean field is located at $\boldsymbol{\pi}$, and hence we assume that $\mathbf{u} = \boldsymbol{\pi}$. We recall that $\boldsymbol{\pi}$ satisfies $W(\boldsymbol{\pi}) = W_1(\boldsymbol{\pi}) - W_2(\boldsymbol{\pi}) = 0$.

Our objective is to show that the sequence of processes $\{(\mathbf{Z}^{(N)}(t), t \geq 0)\}_{N \geq 1}$ converges to a limit in the stationary regime as $N \to \infty$, where $(\mathbf{Z}^{(N)}(t), t \geq 0)$ is the process defined in (10).

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

First, we will show that the sequence $\{\mathbf{Z}^{(N)}(t)\}_{N\geq 1}$ is relatively compact in the stationary regime by using another process $(\mathbf{Q}^{(N)}(t), t \geq 0)$, where $\mathbf{Q}^{(N)}(t) = \sqrt{N}(\mathbf{X}^{(N)}(t) - \boldsymbol{\pi})$. We introduce an SDE, (22), and show that there exists a unique solution to this SDE with a unique invariant law. We then use this result to prove that any limit point of the process $(\mathbf{Z}^{(N)}(t), t \geq 0)$ as $N \to \infty$ in the stationary regime is a stationary OU process with the same invariant law as that of the solution of the proposed SDE, (22). As a result, it would imply that $\{(\mathbf{Z}^{(N)}(t), t \geq 0)\}_{N\geq 1}$ converges as $N \to \infty$ in the stationary regime to the solution of (22).

Next, we state the exponential stability of the mean field in Lemma 2, which we later use in the proof of the result stated in Lemma 3.

Lemma 2. *There exist* $\delta_1 > 0$ *and* $D_3 < \infty$ *such that, for all* $\mathbf{u} \in \mathbb{U}$, *the mean field* $(\mathbf{x}(t, \mathbf{u}), t \geq 0)$ *satisfies* $\|\mathbf{x}(t, \mathbf{u}) - \boldsymbol{\pi}\|_2 \leq \mathrm{e}^{-\delta_1 t} D_3 \|\mathbf{u} - \boldsymbol{\pi}\|_2$.

Proof. See Section 4.6. □

The following result shows the tightness of $\left\{\mathbf{Z}^{(N)}(t)\right\}_{N\geq 1}$ in the stationary regime.

Lemma 3. *If* $\limsup_{N\to\infty} \mathbb{E}\left[\|\mathbf{Q}^{(N)}(0)\|_2^2\right] < \infty$, *then*

$$\limsup_{N\to\infty} \sup_{t\geq 0} \mathbb{E}\left[\|\mathbf{Q}^{(N)}(t)\|_2^2\right] < \infty.$$

Consequently, in the stationary regime corresponding to $t = \infty$,

$$\limsup_{N\to\infty} \mathbb{E}\left[\|\mathbf{Z}^{(N)}(\infty)\|_2^2\right] < \infty.$$

Proof. See Section 4.3. □

Next, we state the SDE that is used to obtain the limit of the sequence $\{(\mathbf{Z}^{(N)}(t), t \geq 0)\}_{N\geq 1}$ in the stationary regime. First, we linearize (8) around $\boldsymbol{\pi}$ to obtain a process $(\mathbf{s}(t), t \geq 0)$ satisfying

$$\frac{\mathrm{d}\mathbf{s}(t)}{\mathrm{d}t} = H(\boldsymbol{\pi})\mathbf{s}(t). \tag{21}$$

Let $\mathbf{B}(t) = (\mathbf{B}_i(t), 0 \leq i \leq C)$ with $\mathbf{B}_0(t) = 0$, where $\{(\mathbf{B}_i(t), t \geq 0)\}_{0\leq i\leq C}$ are independent centered Brownian motions and $\mathbb{E}\left[\mathbf{B}_i^2(1)\right] = V_i = \mathrm{var}(\mathbf{B}_i(1)) = 2i(\pi_i - \pi_{i+1})$, $i \geq 1$. The infinitesimal covariance matrix of $(\mathbf{B}(t), t \geq 0)$ is the diagonal matrix $\mathrm{diag}(\mathbf{V})$, where $\mathbf{V} = (V_n, 0 \leq n \leq C)$. From (19), the martingales $(\mathbf{M}(t), t \geq 0)$ with $\mathbf{x}(0, \boldsymbol{\pi}) = \boldsymbol{\pi}$ have the same law as $(\mathbf{B}(t), t \geq 0)$. Now we define the following SDE, which is used to study the process $(\mathbf{Z}(t), t \geq 0)$ in the stationary regime.

Definition 3. An SDE for the stationary regime: Let $(\mathbf{Q}(t), t \geq 0)$ be a solution to the SDE

$$\mathbf{Q}(t) = \mathbf{Q}(0) + \int_{s=0}^{t} H(\boldsymbol{\pi})\mathbf{Q}(s)\,\mathrm{d}s - \int_{s=0}^{t} W_3(\boldsymbol{\pi})\,\mathrm{d}s + \mathbf{B}(t). \tag{22}$$

Then, (22) defines an OU process whose drift and variance depend on $\boldsymbol{\pi}$.

For an arbitrary $\mathbf{Q}(0)$ in (22), we have the following result. The proof follows by the same arguments as in the proof of Theorem 2, and hence is omitted.

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

Theorem 4.

(i) $\|H(\boldsymbol{\pi})\|_2$ *is bounded.*

(ii) *If* $\mathbb{E}\left[\|\mathbf{Q}(0)\|_2^2\right] < \infty$, *then* $(\mathbf{Q}(t), t \geq 0)$ *given by*

$$\mathbf{Q}(t) = \mathrm{e}^{H(\boldsymbol{\pi})t}\mathbf{Q}(0) - \int_{s=0}^{t} \mathrm{e}^{H(\boldsymbol{\pi})(t-s)} W_3(\boldsymbol{\pi})\,\mathrm{d}s + \int_{s=0}^{t} \mathrm{e}^{H(\boldsymbol{\pi})(t-s)}\,\mathrm{d}\mathbf{B}(s)$$

is the unique strong solution to (22). *Furthermore,* $\mathbb{E}\left[\sup_{t\leq T} \|\mathbf{Q}(t)\|_2^2\right] < \infty$.

We point out that the transpose $H(\boldsymbol{\pi})^*$ of $H(\boldsymbol{\pi})$ is the generator of a finite-state birth–death process, and the birth, death, and killing rates in state i $(1 \leq i \leq C)$ are γ_i, $i-1$, and 1, respectively. Let $\mathcal{I}$ be the identity matrix of dimension $C \times C$. Then, since $H(\boldsymbol{\pi})^* + \mathcal{I}$ is the generator of a birth–death process with zero killing rates, all the eigenvalues of $H(\boldsymbol{\pi})^* + \mathcal{I}$ are negative [11]. Hence, all the eigenvalues of $H(\boldsymbol{\pi})$ are less than -1. As a consequence, we have the following result due to the fact that all the eigenvalues are negative.

Lemma 4. *The unique solution to* (21) *is given by* $(\mathbf{s}(t), t \geq 0)$, *where* $\mathbf{s}(t) = \mathrm{e}^{H(\boldsymbol{\pi})t}\mathbf{s}(0)$. *Furthermore,* $(\mathbf{s}(t), t \geq 0)$ *satisfies, for some* $\delta_2 > 0$ *and* $D_4 < \infty$,

$$\|\mathbf{s}(t)\|_2 \leq \mathrm{e}^{-\delta_2 t} D_4 \|\mathbf{s}(0)\|_2.$$

From Lemma 4 and the unique solution given in Theorem 4, the following result follows immediately. Hence, we omit the proof.

Theorem 5. *The unique solution of* (22) *as* $t \to \infty$ *has the invariant law coinciding with the law of a stationary Gaussian process with mean* $\int_0^\infty \mathrm{e}^{H(\boldsymbol{\pi})s} W_3(\boldsymbol{\pi})\,\mathrm{d}s$ *and covariance matrix* $\int_0^\infty \mathrm{e}^{H(\boldsymbol{\pi})s}\mathrm{diag}(\mathbf{V})\mathrm{e}^{H(\boldsymbol{\pi})^* s}\,\mathrm{d}s$.

We are now ready to state the main result on the FCLT for the stationary regime.

Theorem 6. *Under the assumption that the system with index N is in the stationary regime, the sequence* $\{(\mathbf{Z}^{(N)}(t), t \geq 0)\}_{N\geq 1}$ *as* $N \to \infty$ *converges in law to the unique stationary OU process which solves* (22).

Furthermore, the limit of the sequence $\{(\mathbf{Z}^{(N)}(0))\}_{N\geq 1}$ *in the stationary regime has the same law as the invariant law of the solution of* (22).

Proof. See Section 4.4. □

Now we use Theorem 6 to provide an approximation to the average blocking probability in the system with N servers.

Theorem 7. *Let* $P_{\text{block}}^{(N)}$ *be the average blocking probability in the system with N servers. Then,*

$$P_{\text{block}}^{(N)} = \pi_C^d - \frac{1}{\sigma\sqrt{N}}\left(\sum_{i=0}^{C} i(\kappa_i - \kappa_{i+1})\right) - \frac{\beta}{\sigma\sqrt{N}}\left(1 - \pi_C^d\right) + o\left(N^{-\frac{1}{2}}\right), \tag{23}$$

where the vector $\kappa = (\kappa_i, 0 \leq i \leq C) = \int_0^\infty \mathrm{e}^{H(\boldsymbol{\pi})s} W_3(\boldsymbol{\pi})\,\mathrm{d}s$ *is the mean of the unique solution of* (22) *in the stationary regime.*

Proof. See Section 4.5. □

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

Remark 3. From Theorem 7, we have

(i) If $\sigma = C$, the result (23) corresponds to the Halfin–Whitt regime. In this case, we have

$$\lim_{N\to\infty} \sqrt{N}\big(P_{\text{block}}^{(N)} - \pi_C^d\big) = -\frac{1}{\sigma}\left(\sum_{i=0}^{C} i(\kappa_i - \kappa_{i+1})\right) - \frac{\beta}{\sigma}(1 - \pi_C^d).$$

(ii) If $\beta = 0$, then $W_3(\boldsymbol{\pi}) = 0$. As a result,

$$\lim_{N\to\infty} \sqrt{N}\big(P_{\text{block}}^{(N)} - \pi_C^d\big) = 0.$$

The results for the case $\beta = 0$ were presented in [19].

The significance of Theorem 7 is that although the exact blocking formula for $P_{\text{block}}^{(N)}$ is not known, and is also difficult to characterize due to the complex interactions between servers, when N becomes large we can compute approximations to the blocking probability as a function of $\boldsymbol{\pi}$, β, N, σ, and C.

4. Proofs of main results

4.1. Proof of Lemma 1

We have

$$\begin{aligned}\mathbf{Z}^{(N)}(t) =& \mathbf{Z}^{(N)}(0) + \int_{s=0}^{t} \sqrt{N}(W(\mathbf{X}^{(N)}(s)) - W(\mathbf{x}(s, \mathbf{u}))\,\mathrm{d}s \\ &- \int_{s=0}^{t} W_3(\mathbf{X}^{(N)}(s))\,\mathrm{d}s + \mathbf{M}^{(N)}(t)\end{aligned} \tag{24}$$

and, for $n \geq 1$,

$$\langle \mathbf{M}_n^{(N)} \rangle_t = \int_{s=0}^{t} ((W_1(\mathbf{X}^{(N)}(s)))_n + (W_2(\mathbf{X}^{(N)}(s)))_n - \frac{1}{\sqrt{N}}(W_3(\mathbf{X}^{(N)}(s)))_n)\,\mathrm{d}s.$$

From (24), we obtain

$$\|\mathbf{Z}^{(N)}(t)\|_2 \leq \|\mathbf{Z}^{(N)}(0)\|_2 + B_W \int_{s=0}^{t} \|\mathbf{Z}^{(N)}(s)\|_2\,\mathrm{d}s + 2C\beta t + \|\mathbf{M}^{(N)}(t)\|_2.$$

By using Gronwall's lemma,

$$\|\mathbf{Z}^{(N)}(t)\|_2 \leq \big(\|\mathbf{Z}^{(N)}(0)\|_2 + 2C\beta t + \|\mathbf{M}^{(N)}(t)\|_2\big)\mathrm{e}^{B_W t}.$$

As a result, we get

$$\|\mathbf{Z}^{(N)}(t)\|_2^2 \leq 3\big(\|\mathbf{Z}^{(N)}(0)\|_2^2 + 4C^2\beta^2 t^2 + \|\mathbf{M}^{(N)}(t)\|_2^2\big)\mathrm{e}^{2B_W t}.$$

For $T > 0$, we have

$$\sup_{0\leq t\leq T} \|\mathbf{Z}^{(N)}(t)\|_2^2 \leq 3\Big(\|\mathbf{Z}^{(N)}(0)\|_2^2 + 4C^2\beta^2 T^2 + \sup_{0\leq t\leq T} \|\mathbf{M}^{(N)}(t)\|_2^2\Big)\mathrm{e}^{2B_W T}.$$

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

Finally, Doob's inequality implies that

$$\mathbb{E}\left[\sup_{0\le t\le T}\|\mathbf{Z}^{(N)}(t)\|_2^2\right] \le 3\mathrm{e}^{2B_W T}\left(\mathbb{E}\left[\|\mathbf{Z}^{(N)}(0)\|_2^2\right] + 4C^2\beta^2T^2 + 4\mathbb{E}\left[\sum_{i=1}^{C}\langle \mathbf{M}_i^{(N)}\rangle_T\right]\right).$$

Since $\sup_{N\ge 1}\mathbb{E}\left[\sum_{i=1}^{C}\langle \mathbf{M}_i^{(N)}\rangle_T\right] < \infty$, we conclude that

$$\limsup_{N\to\infty}\mathbb{E}\left[\sup_{0\le t\le T}\|\mathbf{Z}^{(N)}(t)\|_2^2\right] < \infty.$$

4.2 Proof of Theorem 3

We recall that since the space $\mathbb{V}$ is a Polish space, the space of càdlàg functions under the Skhorohod topology is a Polish space [4, Theorem 5.6, p. 121]. Hence, from Prohorov's theorem [4], tightness is equivalent to relative compactness. Therefore, it is enough to show the tightness, and then we need to show that every limiting point has the same law as the unique OU process that solves (20) with the initial point $\mathbf{Z}(0)$.

We use Theorem 4.1 of [4, p. 354] to show the tightness of $(\mathbf{Z}^{(N)}(t), t \ge 0)$. First, we establish several useful preliminary results.

Since $\mathbf{Z}^{(N)}(0) \Rightarrow \mathbf{Z}(0)$, it implies that the sequence $\mathbf{Z}^{(N)}(0)$ is tight. Let $\tilde{B}(r)$ be the closed ball with radius r centered at $\mathbf{0}$. For every $\varepsilon > 0$, there exists $r_\varepsilon < \infty$ such that $\mathbb{P}(\mathbf{Z}^{(N)}(0) \in \tilde{B}(r_\varepsilon)) > 1-\varepsilon$ for all $N \ge 1$. We now define a random variable $\mathbf{X}^{(N,\varepsilon)}(0)$ such that it coincides with $\mathbf{X}^{(N)}(0)$ on $\{\mathbf{Z}^{(N)}(0) \in \tilde{B}(r_\varepsilon)\}$ and $\mathbf{Z}^{(N,\varepsilon)}(0)$ is uniformly bounded in N on $\{\mathbf{Z}^{(N)}(0) \notin \tilde{B}(r_\varepsilon)\}$. Then, by using coupling arguments, the processes $(\mathbf{Z}^{(N,\varepsilon)}(t), t \ge 0)$ and $(\mathbf{Z}^{(N)}(t), t \ge 0)$ coincide on $\{\mathbf{Z}^{(N)}(0) \in \tilde{B}(r_\varepsilon)\}$. Hence, without loss of generality, we assume that $\mathbf{Z}^{(N)}(0)$ is uniformly bounded in N. As a consequence, the result stated in Lemma 1 can be used in the rest of the proof.

Next, we state the following useful result from [9, Lemma 3.3]. For a and h in $\mathbb{R}$, let $\hat{B}(a,h) = (a+h)^d - a^d - da^{d-1}h$. Then, if both a and $a+h$ lie in $[0,1]$, we have

$$0 \le \hat{B}(a,h) \le h^d + (2^d - d - 2)ah^2. \tag{25}$$

We now define a mapping $\hat{G} : \mathbb{U}\times\mathbb{V} \mapsto \mathbb{V}$ as follows: for $\mathbf{r}\in\mathbb{U}$ and $\mathbf{y}\in\mathbb{V}$,

$$(\hat{G}(\mathbf{r},\mathbf{y}))_n = \sigma\hat{B}(r_{n-1}, y_{n-1}) - \sigma\hat{B}(r_n, y_n).$$

Then, if $\mathbf{r}+\mathbf{y}\in\mathbb{U}$, we have

$$W(\mathbf{r}+\mathbf{y}) - W(\mathbf{r}) = H(\mathbf{r})\mathbf{y} + \hat{G}(\mathbf{r},\mathbf{y}). \tag{26}$$

Note that since $\mathbf{Z}^{(N)}(t) = \sqrt{N}(\mathbf{X}^{(N)}(t) - \mathbf{x}(t,\mathbf{u}))$, we have

$$\mathbf{X}^{(N)}(t) = \mathbf{x}(t,\mathbf{u}) + \frac{\mathbf{Z}^{(N)}(t)}{\sqrt{N}}.$$

Here, $\mathbf{X}^{(N)}(t)$, $\mathbf{x}(t,\mathbf{u})\in\mathbb{U}$ and $\frac{\mathbf{Z}^{(N)}(t)}{\sqrt{N}}\in\mathbb{V}$. Hence, we can write the following equation from (26):

$$W(\mathbf{X}^{(N)}(t)) - W(\mathbf{x}(t,\mathbf{u})) = H(\mathbf{x}(t,\mathbf{u}))\frac{\mathbf{Z}^{(N)}(t)}{\sqrt{N}} + \hat{G}\left(\mathbf{x}(t,\mathbf{u}), \frac{\mathbf{Z}^{(N)}(t)}{\sqrt{N}}\right). \tag{27}$$

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

Now we show that the conditions of [4, Theorem 4.1, p. 354] are satisfied. Let us write $\mathbf{B}_n(t)$ and $\mathbf{A}_n^{i,j}(t)$ of [4, Theorem 4.1, p. 354] as $\mathbf{D}_{(N)}(t)$ and $\mathbf{A}_{(N)}^{i,j}(t)$, respectively. Then, from (13) we can write

$$\mathbf{D}_{(N)}(t) = \int_{s=0}^{t} \sqrt{N}(W(\mathbf{X}^{(N)}(s)) - W(\mathbf{x}(s,\mathbf{u}))\,\mathrm{d}s - \int_{s=0}^{t} W_3(\mathbf{X}^{(N)}(s))\,\mathrm{d}s.$$

From (27), we can also write

$$\begin{aligned}\mathbf{D}_{(N)}(t) = &\int_{s=0}^{t} H(\mathbf{x}(s,\mathbf{u}))\mathbf{Z}^{(N)}(s)\,\mathrm{d}s + \int_{s=0}^{t} \sqrt{N}\hat{G}\left(\mathbf{x}(s,\mathbf{u}), \frac{\mathbf{Z}^{(N)}(s)}{\sqrt{N}}\right)\,\mathrm{d}s \\ &- \int_{s=0}^{t} W_3(\mathbf{X}^{(N)}(s))\,\mathrm{d}s.\end{aligned}$$

Also, from (12) we have

$$\mathbf{A}_{(N)}^{i,i}(t) = \int_{s=0}^{t} \left((W_1(\mathbf{X}^{(N)}(s)))_i + (W_2(\mathbf{X}^{(N)}(s)))_i - \frac{1}{\sqrt{N}}(W_3(\mathbf{X}^{(N)}(s)))_i\right)\,\mathrm{d}s,$$

and $\mathbf{A}_{(N)}^{i,j}(t) = 0$ for $i \neq j$. Let us define

$$\begin{aligned}\mathbf{D}(t) &= \int_{s=0}^{t} H(\mathbf{x}(s,\mathbf{u}))\mathbf{Z}(s)\,\mathrm{d}s - \int_{s=0}^{t} W_3(\mathbf{x}(s,\mathbf{u}))\,\mathrm{d}s, \\ A^{i,i}(t) &= \int_{s=0}^{t} ((W_1(\mathbf{x}(s,\mathbf{u})))_i + (W_2(\mathbf{x}(s,\mathbf{u})))_i)\,\mathrm{d}s,\end{aligned}$$

with $A^{i,j}(t) = 0$ for $i \neq j$.

Since $(\mathbf{Z}_i^{(N)}(t), t \geq 0)$ has jumps of size $\frac{1}{\sqrt{N}}$, and from the continuity of $\mathbf{D}_{(N)}(t)$ and $\mathbf{A}_{(N)}^{i,j}(t)$ in t, conditions (4.3)–(4.5) of [4, Theorem 4.1, p. 354] are valid. From condition (4.6) of [4, Theorem 4.1, p. 354], we need to show that, for $T > 0$, $\mathbf{V}_{(N)}(T) \to 0$ in probability as $N \to \infty$, where $\mathbf{V}_{(N)}(t) = \mathbf{D}_{(N)}(t) - \mathbf{D}(t)$ is given by

$$\begin{aligned}\mathbf{V}_{(N)}(t) = &\left(\int_{s=0}^{t} H(\mathbf{x}(t,\mathbf{u}))\mathbf{Z}^{(N)}(t)\,\mathrm{d}s + \int_{s=0}^{t} \sqrt{N}\hat{G}\left(\mathbf{x}(t,\mathbf{u}), \frac{\mathbf{Z}^{(N)}(t)}{\sqrt{N}}\right)\mathrm{d}s\right. \\ &\left. - \int_{s=0}^{t} W_3\left(\mathbf{X}^{(N)}(s)\right)\mathrm{d}s\right) \\ &- \left(\int_{s=0}^{t} H(\mathbf{x}(s,\mathbf{u}))\mathbf{Z}^{(N)}(s)\,\mathrm{d}s - \int_{s=0}^{t} W_3(\mathbf{x}(s,\mathbf{u}))\,\mathrm{d}s\right).\end{aligned}$$

From (14) and (25), we have $\int_{s=0}^{T} \sqrt{N}\hat{G}\left(\mathbf{x}(s,\mathbf{u}), \frac{\mathbf{Z}^{(N)}(s)}{\sqrt{N}}\right)\mathrm{d}s \to 0$ in probability. From the existence of the mean-field limit, it follows that $\int_{s=0}^{T} W_3(\mathbf{X}^{(N)}(s))\,\mathrm{d}s - \int_{s=0}^{T} W_3(\mathbf{x}(s,\mathbf{u}))\,\mathrm{d}s \to 0$ in probability. Thus, condition (4.6) of [4, Theorem 4.1, p. 354] is also true. Condition (4.7) of [4, Theorem 4.1, p. 354] is also true since $\mathbf{A}_{(N)}^{i,i}(T) - A^{i,i}(T) \to 0$ in probability from the existence of the mean-field limit. Finally, from the fact that the SDE (20) has a unique solution, the sequence $\{(\mathbf{Z}^{(N)}(t), t \geq 0)\}_{N\geq 1}$ converges to the unique solution of the SDE (20) as $N \to \infty$.

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

4.3. Proof of Lemma 3

The proof is based on Lemma 2. Although our main objective is to establish the convergence of $\{\mathbf{Z}^{(N)}(t)\}_{N\geq 1}$ as $N\to\infty$ in the stationary regime, we prove the tightness of the stationary sequence $\{\mathbf{Z}^{(N)}(\infty)\}_{N\geq 1}$ by studying an alternative process $(\mathbf{Q}^{(N)}(t), t\geq 0)$ with the help of Lemma 2, where $\mathbf{Q}^{(N)}(t)=\sqrt{N}(\mathbf{X}^{(N)}(t)-\boldsymbol{\pi})$. Let us write the solution to the mean-field equation (8) at time h with the initial point $\mathbf{v}$ as $\mathbf{y}(h,\mathbf{v})$. We have

$$\mathbf{y}(h,\mathbf{v})=\mathbf{v}+\int_{s=0}^{h} W(\mathbf{y}(s,\mathbf{v}))\,\mathrm{d}s.$$

Also, the process $(\mathbf{X}^{(N)}(t), t\geq 0)$ satisfies

$$\mathbf{X}^{(N)}(t)=\mathbf{X}^{(N)}(0)+\int_{s=0}^{t} W(\mathbf{X}^{(N)}(s))\,\mathrm{d}s-\int_{s=0}^{t}\frac{1}{\sqrt{N}}W_3\big(\mathbf{X}^{(N)}(s)\big)\,\mathrm{d}s+\frac{\mathbf{M}^{(N)}(t)}{\sqrt{N}}. \tag{28}$$

For $t_0\geq 0$, we obtain

$$\mathbf{Q}^{(N)}(t_0+h)=\sqrt{N}\big(\mathbf{X}^{(N)}(t_0+h)-\mathbf{y}\big(h,\mathbf{X}^{(N)}(t_0)\big)\big)\ +\sqrt{N}\big(\mathbf{y}\big(h,\mathbf{X}^{(N)}(t_0)\big)-\boldsymbol{\pi}\big).$$

By defining $\mathbf{Z}^{(N)}(t_0,h)=\sqrt{N}\big(\mathbf{X}^{(N)}(t_0+h)-\mathbf{y}\big(h,\mathbf{X}^{(N)}(t_0)\big)\big)$, we obtain

$$\mathbf{Q}^{(N)}(t_0+h)=\mathbf{Z}^{(N)}(t_0,h)+\sqrt{N}\big(\mathbf{y}\big(h,\mathbf{X}^{(N)}(t_0)\big)-\boldsymbol{\pi}\big). \tag{29}$$

Also, from (4.3),

$$\mathbf{Z}^{(N)}(t_0,h)=\sqrt{N}\big(\mathbf{X}^{(N)}(t_0+h)-\mathbf{X}^{(N)}(t_0)\big)-\sqrt{N}\int_{s=0}^{t} W\big(\mathbf{y}\big(s,\mathbf{X}^{(N)}(t_0)\big)\big)\,\mathrm{d}s. \tag{30}$$

From (28), we have

$$\begin{aligned}\mathbf{X}^{(N)}(t_0+h)-\mathbf{X}^{(N)}(t_0)=&\int_{s=t_0}^{t_0+h} W(\mathbf{X}^{(N)}(s))\,\mathrm{d}s+\frac{\mathbf{M}^{(N)}(t_0+h)-\mathbf{M}^{(N)}(t_0)}{\sqrt{N}}\\&-\frac{1}{\sqrt{N}}\int_{s=t_0}^{t_0+h} W_3(\mathbf{X}^{(N)}(s))\,\mathrm{d}s.\end{aligned} \tag{31}$$

Then, by using (30) and (31), we get

$$\begin{aligned}\mathbf{Z}^{(N)}(t_0,h)=&\sqrt{N}\int_{s=0}^{h} W\big(\mathbf{X}^{(N)}(t_0+s)\big)\,\mathrm{d}s-\sqrt{N}\int_{s=0}^{h} W\big(\mathbf{y}\big(s,\mathbf{x}^{(N)}(t_0)\big)\big)\,\mathrm{d}s\\&+\big(\mathbf{M}^{(N)}(t_0+h)-\mathbf{M}^{(N)}(t_0)\big)-\int_{s=t_0}^{t_0+h} W_3\big(\mathbf{X}^{(N)}(s)\big)\,\mathrm{d}s.\end{aligned}$$

After simplifications, we can write

$$\begin{aligned}\|\mathbf{Z}^{(N)}(t_0,h)\|_2\leq&\sqrt{N}\int_{s=0}^{h}\big\|W\big(\mathbf{X}^{(N)}(t_0+s)\big)-W\big(\mathbf{y}\big(s,\mathbf{X}^{(N)}(t_0)\big)\big)\big\|_2\,\mathrm{d}s\\&+\|\mathbf{M}^{(N)}(t_0+h)-\mathbf{M}^{(N)}(t_0)\|_2+|\beta|d\sqrt{C}h.\end{aligned}$$

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

Hence, we obtain

$$\|\mathbf{Z}^{(N)}(t_0, h)\|_2 \le B_W \int_{s=0}^{h} \|\mathbf{Z}^{(N)}(t_0, s)\|_2 \, \mathrm{d}s + |\beta| d\sqrt{C}h + \|\mathbf{M}^{(N)}(t_0 + h) - \mathbf{M}^{(N)}(t_0)\|_2.$$

For any $T \ge 0$, Gronwall's inequality implies that there exists a constant S_T such that

$$\sup_{0\le h\le T} \|\mathbf{Z}^{(N)}(t_0, h)\|_2 \le S_T\Big(S_T + \sup_{0\le h\le T} \|\mathbf{M}^{(N)}(t_0 + h) - \mathbf{M}^{(N)}(t_0)\|_2\Big). \tag{32}$$

By using Lemma 2, we can write, from (29),

$$\|\mathbf{Q}^{(N)}(t_0 + h)\|_2 \le \|\mathbf{Z}^{(N)}(t_0, h)\|_2 + \mathrm{e}^{-\delta_1 h} D_3 \|\mathbf{Q}^{(N)}(t_0)\|_2. \tag{33}$$

From (32) and (33), we obtain

$$\|\mathbf{Q}^{(N)}(t_0 + h)\|_2 \le S_T\Big(S_T + \sup_{0\le h\le T} \|\mathbf{M}^{(N)}(t_0 + h) - \mathbf{M}^{(N)}(t_0)\|_2\Big) + \mathrm{e}^{-\delta_1 h} D_3 \|\mathbf{Q}^{(N)}(t_0)\|_2.$$

As a result, we can find some constant L_T as a function of T such that, for $0 \le h \le T$,

$$\mathbb{E}\left[\|\mathbf{Q}^{(N)}(t_0 + h)\|_2^2\right] \le L_T + 3\mathrm{e}^{-2\delta_1 h} D_3^2 \mathbb{E}\left[\|\mathbf{Q}^{(N)}(t_0)\|_2^2\right]. \tag{34}$$

We now select a large value of T such that $3\mathrm{e}^{-2\delta_1 T} D_3^2 \le \varepsilon < 1$. Then, for all $N \ge 1$ and an integer m, we get

$$\mathbb{E}\left[\|\mathbf{Q}^{(N)}((m+1)T)\|_2^2\right] \le L_T + \varepsilon \mathbb{E}\left[\|\mathbf{Q}^{(N)}(mT)\|_2^2\right].$$

By using the induction method, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{Q}^{(N)}(mT)\|_2^2\right] &\le L_T\left(\sum_{j=1}^{m} \varepsilon^{j-1}\right) + \varepsilon^m \mathbb{E}\left[\|\mathbf{Q}^{(N)}(0)\|_2^2\right] \\ &\le \frac{L_T}{1-\varepsilon} + \mathbb{E}\left[\|\mathbf{Q}^{(N)}(0)\|_2^2\right].\end{aligned} \tag{35}$$

However, from (34),

$$\sup_{0\le h\le T} \mathbb{E}\left[\|\mathbf{Q}^{(N)}(mT + h)\|_2^2\right] \le L_T + 3D_3^2 \mathbb{E}\left[\|\mathbf{Q}^{(N)}(mT)\|_2^2\right].$$

As a consequence, (35) implies

$$\sup_{0\le h\le T} \mathbb{E}\left[\|\mathbf{Q}^{(N)}(mT + h)\|_2^2\right] \le L_T + 3D_3^2\left(\frac{L_T}{1-\varepsilon} + \mathbb{E}\left[\|\mathbf{Q}^{(N)}(0)\|_2^2\right]\right).$$

Since m is arbitrary, we conclude that

$$\sup_{t\ge 0} \mathbb{E}\left[\|\mathbf{Q}^{(N)}(t)\|_2^2\right] \le L_T + 3D_3^2\left(\frac{L_T}{1-\varepsilon} + \mathbb{E}\left[\|\mathbf{Q}^{(N)}(0)\|_2^2\right]\right).$$

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

From ergodicity and Fatou's Lemma [4, p. 492], the stationary random variable $\mathbf{Z}^{(N)}(\infty)$ satisfies

$$\mathbb{E}\left[\|\mathbf{Z}^{(N)}(\infty)\|_2^2\right] \le \liminf_{t\ge 0} \mathbb{E}\left[\|\mathbf{Q}^{(N)}(t)\|_2^2\right] \le \sup_{t\ge 0} \mathbb{E}\left[\|\mathbf{Q}^{(N)}(t)\|_2^2\right].$$

Finally, to show that $\limsup_{N\to\infty} \mathbb{E}\left[\|\mathbf{Z}^{(N)}(\infty)\|_2^2\right] < \infty$, we need to find an $\mathbf{X}^{(N)}(0)$ such that $\limsup_{N\to\infty} \mathbb{E}\left[\|\mathbf{Q}^{(N)}(0)\|_2^2\right] < \infty$. For $n \ge 1$, if we select $\mathbf{X}_n^{(N)}(0) = \frac{j}{N}$ so that $\frac{-1}{2N} \le \pi_n - \frac{j}{N} \le \frac{1}{2N}$, then $\mathbb{E}\left[\|\mathbf{Q}^{(N)}(0)\|_2^2\right] < \frac{C}{4N}$. Hence, completing the proof, $\limsup_{N\to\infty}$ $\mathbb{E}\left[\|\mathbf{Q}^{(N)}(0)\|_2^2\right] = 0$.

4.4. Proof of Theorem 6

From Lemma 3 and the Markov inequality, the sequence $\{\mathbf{Z}^{(N)}(0)\}_{N\ge 1}$ is tight. As a result, from the Prohorov theorem [4, p. 104], the sequence $\{\mathbf{Z}^{(N)}(0)\}_{N\ge 1}$ is relatively compact. Consider a converging subsequence and let $\mathbf{Z}^{(\infty)}(0)$ be its limiting point, which is square integrable. Then, from Theorem 3, the considered converging subsequence converges in law to the unique OU process $(\mathbf{Z}^{(\infty)}(t), t \ge 0)$ that solves the SDE (22) with initial point $\mathbf{Z}^{(\infty)}(0)$. But, we know from [4, Lemma 7.7 and Theorem 7.8, p. 131] that the limit of a sequence of stationary processes is stationary. Hence, the law of $(\mathbf{Z}^{(\infty)}(t), t \ge 0)$ should be the unique law of the stationary OU process solving the SDE (22). This argument applies for every converging subsequence. Hence, the sequence $\{(\mathbf{Z}^{(N)}(t), t \ge 0)\}_{N\ge 1}$ in the stationary regime converges to the unique stationary OU process solving the SDE (22). This completes the proof.

4.5. Proof of Theorem 7

The proof is based on Little's law [2, Theorem 4.1], and Theorems 5 and 6. Let us consider a random variable $\tilde{\mathbf{S}}^{(N)}(\infty)$ which denotes the number of progressing jobs in the system in the stationary regime. From Little's law, we obtain $(N\lambda^{(N)})\left(1 - P_{\text{block}}^{(N)}\right) = \mathbb{E}\left[\tilde{\mathbf{S}}^{(N)}(\infty)\right]$. We now consider a random variable $\mathbf{X}^{(N)}(\infty) = (\mathbf{X}_n^{(N)}(\infty), 0 \le n \le C)$, where $\mathbf{X}_n^{(N)}(\infty)$ is the fraction of servers with at least n progressing jobs in the stationary regime. Then, we can write $\tilde{\mathbf{S}}^{(N)}(\infty) = \sum_{n=0}^{C} Nn(\mathbf{X}_n^{(N)}(\infty) - \mathbf{X}_{n+1}^{(N)}(\infty))$. Hence, we have

$$\left(N\lambda^{(N)}\right)\left(1 - P_{\text{block}}^{(N)}\right) = \mathbb{E}\left[\sum_{n=0}^{C} Nn\left(\mathbf{X}_n^{(N)}(\infty) - \mathbf{X}_{n+1}^{(N)}(\infty)\right)\right].$$

Therefore, we obtain

$$\lambda^{(N)}\left(1 - P_{\text{block}}^{(N)}\right) = \sum_{n=0}^{C} n\mathbb{E}\left[\mathbf{X}_n^{(N)}(\infty) - \mathbf{X}_{n+1}^{(N)}(\infty)\right]. \tag{36}$$

Also, from Theorem 5, Lemma 3, and Theorem 6, since the diffusion limit in the stationary regime has the mean vector $\kappa = \int_0^\infty \mathrm{e}^{H(\boldsymbol{\pi})s} W_3(\boldsymbol{\pi})\, \mathrm{d}s$ and $\limsup_{N\to\infty} \mathbb{E}\left[\|\mathbf{Z}^{(N)}(\infty)\|^2\right] < \infty$, we have $\lim_{N\to\infty} \sqrt{N}\left(\mathbb{E}\left[\mathbf{X}^{(N)}(\infty)\right] - \boldsymbol{\pi}\right) = \kappa$. Therefore,

$$\mathbb{E}\left[\mathbf{X}^{(N)}(\infty)\right] - \boldsymbol{\pi} = \frac{\kappa}{\sqrt{N}} + o\left(N^{-\frac{1}{2}}\right). \tag{37}$$

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

Then, from (36) and (37), we get

$$\lambda^{(N)}\left(1-P_{\text{block}}^{(N)}\right)=\sum_{n=0}^{C} n(\pi_n-\pi_{n+1})+\frac{1}{\sqrt{N}}\sum_{n=0}^{C} n(\kappa_n-\kappa_{n+1})+o\left(N^{-\frac{1}{2}}\right). \tag{38}$$

But, from the stationary mean-field equations, the fixed point satisfies

$$\pi_0=1, \qquad \sum_{n=0}^{C} n(\pi_n-\pi_{n+1})=\sigma\left(1-\pi_C^d\right). \tag{39}$$

Then, from (38) and (39), we can write

$$1-P_{\text{block}}^{(N)}=\frac{1}{\lambda^{(N)}}\left[\sigma(1-\pi_C^d)+\frac{1}{\sqrt{N}}\sum_{n=0}^{C} n(\kappa_n-\kappa_{n+1})\right]+o\left(N^{-\frac{1}{2}}\right).$$

However, by using $\lambda^{(N)}=\sigma-\frac{\beta}{\sqrt{N}}$ and the fact that $\frac{\beta}{\sigma\sqrt{N}}<1$, we obtain

$$1-P_{\text{block}}^{(N)}=\frac{1}{\sigma}\left(1+\frac{\beta}{\sigma\sqrt{N}}+o\left(N^{-\frac{1}{2}}\right)\right)\left[\sigma\left(1-\pi_C^d\right)+\frac{1}{\sqrt{N}}\sum_{n=0}^{C} n(\kappa_n-\kappa_{n+1})\right]+o\left(N^{-\frac{1}{2}}\right).$$

After some simple calculations, we obtain

$$P_{\text{block}}^{(N)}=\pi_C^d-\frac{1}{\sigma\sqrt{N}}\left(\sum_{n=0}^{C} n(\kappa_n-\kappa_{n+1})\right)-\frac{\beta}{\sigma\sqrt{N}}\left(1-\pi_C^d\right)+o\left(N^{-\frac{1}{2}}\right).$$

This completes the proof.

4.6. Proof of Lemma 2

The proof uses the quasi-monotonicity of the mean field. Let us write the unique solution to the MFEs with the initial point $\mathbf{v}$ as $(\mathbf{y}(t,\mathbf{v}), t\geq 0)$. From the quasi-monotonicity of the mean field, we have

$$\mathbf{y}(t,\min(\mathbf{v},\boldsymbol{\pi}))\leq\mathbf{y}(t,\mathbf{v})\leq\mathbf{y}(t,\max(\mathbf{v},\boldsymbol{\pi})). \tag{40}$$

For $\mathbf{a},\mathbf{b}\in\mathbb{U}$, let $\|\mathbf{a}-\mathbf{b}\|_1=\sum_{i=0}^{C}|a_i-b_i|$. From Lemma 4 of [22], since $\min(\mathbf{v},\boldsymbol{\pi})\leq\boldsymbol{\pi}$ and $\max(\mathbf{v},\boldsymbol{\pi})\geq\boldsymbol{\pi}$, we have

$$\begin{aligned}\|\mathbf{y}(t,\min(\mathbf{v},\boldsymbol{\pi}))-\boldsymbol{\pi}\|_1&\leq e^{-t}\|\min(\mathbf{v},\boldsymbol{\pi})-\boldsymbol{\pi}\|_1,\\ \|\mathbf{y}(t,\max(\mathbf{v},\boldsymbol{\pi}))-\boldsymbol{\pi}\|_1&\leq e^{-t}\|\max(\mathbf{v},\boldsymbol{\pi})-\boldsymbol{\pi}\|_1.\end{aligned} \tag{41}$$

For $t\geq 0$, let us define two sets $V_+(t)$ and $V_-(t)$ as

$$\begin{aligned}V_+(t)&=\{i: y_i(t,\mathbf{v})\geq\pi_i\},\\ V_-(t)&=\{i: y_i(t,\mathbf{v})<\pi_i\}.\end{aligned}$$

Then, we can write

$$\|\mathbf{y}(t,\mathbf{v})-\boldsymbol{\pi}\|_1=\sum_{i\in V_+(t)}(y_i(t,\mathbf{v})-\pi_i)+\sum_{j\in V_-(t)}(\pi_j-y_j(t,\mathbf{v})).$$

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

From (40) and (41), we write

$$\begin{aligned}\|\mathbf{y}(t,\mathbf{v})-\boldsymbol{\pi}\|_1 &\le \sum_{i\in V_+(t)}(y_i(t,\max(\mathbf{v},\boldsymbol{\pi}))-\pi_i)+\sum_{j\in V_-(t)}(\pi_j-y_j(t,\min(\mathbf{v},\boldsymbol{\pi})))\\ &\le \|\mathbf{y}(t,\max(\mathbf{v},\boldsymbol{\pi}))-\boldsymbol{\pi}\|_1+\|\mathbf{y}(t,\min(\mathbf{v},\boldsymbol{\pi}))-\boldsymbol{\pi}\|_1,\\ &\le \mathrm{e}^{-t}\|\max(\mathbf{v},\boldsymbol{\pi})-\boldsymbol{\pi}\|_1+\mathrm{e}^{-t}\|\min(\mathbf{v},\boldsymbol{\pi})-\boldsymbol{\pi}\|_1\\ &= \mathrm{e}^{-t}\|\mathbf{v}-\boldsymbol{\pi}\|_1.\end{aligned}$$

Finally, the result follows from the fact that the norms $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent.

5. Conclusions

We have established an FCLT satisfied by the fluctuation process around the mean field of the occupancy distribution in both transient and stationary regimes for loss models with exponential service time distributions under the assumption that $\lambda^{(N)}=\sigma-\frac{\beta}{\sqrt{N}}$. The proof used the global stability of the mean field. We then showed that the limiting diffusion process is an OU process that is controlled by the mean-field limit. We then showed how Little's law and the FCLT can be combined to show that $P^{(N)}_{\text{block}}-\pi^d_C$ is $O(N^{-\frac{1}{2}})$ if $\beta\neq 0$, with the constants given in terms of the mean field and the parameters. These techniques may be used to study other routing strategies such as a randomized join below threshold strategy, or more generally other occupancy-based routing strategies for which the global asymptotic stability of the mean field can be established (see [18] for such models).

Acknowledgement

This research was supported in part by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). Cette recherche a été subventionneé par le Conseil de recherches en sciences naturelles et génie du Canada (CRSNG).

References

[1] Amazon EC2. (2021) http://aws.amazon.com/ec2/.
[2] Asmussen, S. (2003). *Applied Probability and Queues*. Springer, New York.
[3] Eschenfeldt, P. and Gamarnik, D. (2018). Join the shortest queue with many servers. The heavy-traffic asymptotics. *Math. Operat. Res.* **43**, 867–886.
[4] Ethier, S. N. and Kurtz, T. G. (1985). *Markov Processes: Characterization and Convergence*. John Wiley, New York.
[5] Gast, N. (2017). Expected values estimated via mean-field approximation are $1/N$-accurate. *Proc. ACM Meas. Anal. Comput. Sys.* **1**, 17.
[6] Gast, N. and Van Houdt, B. (2017). A refined mean field approximation. *Proc. ACM Meas. Anal. Comput. Sys.* 1, 33.
[7] Gazdzicki, P., Lambadaris, I. and Mazumdar, R. (1993). Blocking probabilities for large multi-rate Erlang loss systems. *Adv. Appl. Prob.* **25**, 997–1009.
[8] Graham, C. (2000). Chaoticity on path space for a queueing network with selection of the shortest queue among several. *J. Appl. Prob.* **37**, 198–211.
[9] Graham, C. (2005). Functional central limit theorems for a large network in which customers join the shortest of several queues. *Prob. Theory Relat. Fields* **131**, 97–120.
[10] Karthik, A., Mukhopadhyay, A. and Mazumdar, R. R. (2017). Choosing among heterogeneous server clouds. *Queueing Systems* **85**, 1–29.
[11] Ledermann, W., Reuter, G. E. H. and Mahler, K. (1954). Spectral theory for the differential equations of simple birth and death processes. *Phil. Trans. R. Soc. London A* **246**, 321–369.
[12] Microsoft Azure. (2021) http://www.microsoft.com/windowsazure/.

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press

[13] MITZENMACHER, M. (1996). The power of two choices in randomized load balancing. PhD thesis, University of California, Berkeley.
[14] MUKHERJEE, D., BORST, S. C., VAN LEEUWAARDEN, J. AND WHITING, P. A. (2016). *Asymptotic optimality of power-of-d load balancing in large-scale systems*. Preprint, arXiv:1612.00723.
[15] MUKHOPADHYAY, A., MAZUMDAR, R. R. AND GUILLEMIN, F. (2015). The power of randomized routing in heterogeneous loss systems. In *Proc. 27th Int. Teletraffic Congress*, pp. 125–133.
[16] MUKHOPADHYAY, A., KARTHIK, A., MAZUMDAR, R. R. AND GUILLEMIN, F. M. (2015). Mean field and propagation of chaos in multi-class heterogeneous loss models. *Performance Evaluation* **91**, 117–131.
[17] PANG, G., TALREJA, R. AND WHITT, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Prob. Surv.* **4**, 193–267.
[18] VASANTAM, T. AND MAZUMDAR, R. R. (2018) On occupancy based randomized routing schemes in large systems of shared servers. In *Proc. 30th Int. Teletraffic Congress*, pp. 28–36.
[19] VASANTAM, T. AND MAZUMDAR, R. R. (2019). Fluctuations around the mean-field for a large scale Erlang loss system under the SQ(d) load balancing. In *Proc. 31st Int. Teletraffic Congress*, pp. 1–9.
[20] VVEDENSKAYA, N. D., DOBRUSHIN, R. L. AND KARPELEVICH, F. I. (1996). Queueing system with selection of the shortest of two queues: an asymptotic approach. *Probl. Inform. Transm.* **32**, 20–34.
[21] WHITT, W. (1984). Heavy-traffic approximations for service systems with blocking. *AT&T Bell Lab. Tech. J.* **63**, 689–708.
[22] XIE, Q., DONG, X., LU, Y. AND SRIKANT, R. (2015). Power of d choices for large-scale bin packing: a loss model. In *Proc. 2015 ACM SIGMETRICS*, pp. 321–334.
[23] YING, L. (2016). On the approximation error of mean-field models. In *Proc. 2016 ACM SIGMETRICS*, pp. 285–297.

https://doi.org/10.1017/jpr.2020.99 Published online by Cambridge University Press