Fluctuations around the mean-field for a large scale Erlang Loss system under the SQ(d) load balancing

Thirupathaiah Vasantam
Electrical and Computer Engineering
University of Waterloo, Canada
Email: tvasanta@uwaterloo.ca

Ravi R. Mazumdar
Electrical and Computer Engineering
University of Waterloo, Canada,
Email: mazum@uwaterloo.ca

***Abstract*—In this paper, we study the fluctuations of the transient and stationary empirical distributions around the mean-field for a large scale multi-server Erlang Loss system that has N servers. Jobs arrive according to a Poisson process with rate $N\lambda$ and each incoming job is dispatched by a central job dispatcher to the server with the minimum occupancy among d randomly chosen servers with ties broken uniformly at random. Previous works have studied the mean-field limit of this model and characterized the asymptotic behavior of the system when $N \to \infty$. In this paper, we focus on quantifying the resulting error when we approximate the transient and stationary behavior of the system when N is large by the mean-field of the system. We obtain functional central limit theorems (FCLTs) by studying the limit of a suitably scaled fluctuation process of the stochastic empirical process of the model with index N around the mean-field limit when $N \to \infty$. We show that for both the transient and stationary regimes, the limiting process is characterized by an Ornstein-Uhlenbeck (OU) process. We also show that the interchange of limits $\lim_{N\to\infty}\lim_{t\to\infty} = \lim_{t\to\infty}\lim_{N\to\infty}$ is valid under the CLT scaling. Finally, we exploit the FCLT to show that the gap between the exact average blocking probability of a job in the system with the number of servers N and the limiting average blocking probability which is a function of the fixed-point of the mean-field, is of the order $o(N^{-\frac{1}{2}})$ and thus establish the accuracy of the mean-field approximation for finite N.**

I. Introduction

In this paper, we study a large scale multi-server Erlang loss system that has a large number N (in the order of 10000) of parallel servers and a central job dispatcher which routes an incoming job to one of the servers for service according to a predefined routing policy. We assume that each server has ability to process a maximum of C jobs simultaneously, and there is no waiting room. If a job is routed to a server, the job will be accepted for service at the destination server if the occupancy (the number of progressing jobs) of the destination server is less than C, otherwise the job will be blocked or discarded. Each accepted job is processed at constant unit rate until its service is completed. The arrival process of jobs is a Poisson process with rate $N\lambda$ and the service times are assumed to be exponentially distributed with unit mean. The performance analysis of the above system model provides useful insights about the performance of cloud computing systems such as Microsoft's Azure [1] and Amazon EC2 [2] for which Erlang loss models are the appropriate mathematical abstraction as follows.

In cloud systems a large number of servers (typically of the order of tens of thousands) are maintained so that the incoming job requests are served efficiently. In these systems, customers request a variety of resources such as processor power, I/O bandwidth, disk etc. that are present at a large number of available servers. An incoming request is accepted at its destination server for service if the requested amount of resources are available, otherwise it is blocked or discarded. The resources allocated to a job will be released once the service of the job ends. This is akin to a loss model. Therefore the job dispatcher should make routing decisions cleverly so that the resulting average blocking probability is minimized.

In a homogeneous system where both servers and jobs are homogeneous, the obvious choice is the join-the-shortest-queue (JSQ) routing policy in which an arrival is routed to the server with the least occupancy among all the servers. This policy is very cumbersome in large cloud computing systems as it requires the information about occupancies of all the servers. However, for large-scale systems it has been shown that the randomized routing schemes in which an arrival is routed to the server with the most number of available resources from a set of few randomly chosen servers provide descent system performance but with a significant drop in the implementation complexity over the JSQ policy [3]–[6]. This policy is referred to as the SQ(d) policy, short for the power-of-d routing policy, according to which an arrival is routed to the server with the least occupancy out of the d randomly chosen servers with ties broken uniformly at random.

The SQ(d) scheme was first introduced in [4] for a large-scale multi-server server FCFS model with the assumption of $d = 2$. The exact analysis of the SQ(d) routing policy for the system with a finite valued parameter N is a challenging problem due to dependence amongst the servers introduced by the SQ(d) policy. In [4], they showed that the stationary distributions can be characterized explicitly when $N \to \infty$. This limit corresponds to a mean-field limit. In [3] the results were extended to the SQ(d) policy with $d \geq 2$ and it was shown that the case $d = 2$ provides most of the gains over the case of $d = 1$ whence the parlance "the power of 2" came about.

The impact of the SQ(d) policy was investigated for loss models similar to the one considered in this paper in [5]–[8] using mean-field techniques. We now provide a brief overview

of the main results of mean-field analysis of the homogeneous loss systems under the SQ(d) policy.

The identities of servers do not play any role and the system evolution can be described by the Markov process $\mathbf{x}^{(N)}(t) = (x_i^{(N)}(t), 0 \leq i \leq C)$ where $x_i^{(N)}(t)$ denotes the fraction of servers with at least i progressing jobs at time t. We now recall the following result from [6]. We use boldface letters to denote vectors.

Theorem 1. *For a deterministic $\mathbf{x}(0)$, if $\mathbf{x}^{(N)}(0)$ converges in distribution to $\mathbf{x}(0)$ as $N \to \infty$, then $(\mathbf{x}^{(N)}(t), t \geq 0)$ converges in distribution to $(\mathbf{x}(t), t \geq 0)$ where $(\mathbf{x}(t), t \geq 0)$ is a deterministic process and it is the unique solution of the following equations: for $\mathbf{h}(\mathbf{x}(t)) = (h_n(\mathbf{x}(t)), 1 \leq n \leq C)$,*

$$\frac{dx_n(t)}{dt} = h_n(\mathbf{x}(t)), \tag{1}$$

where

$$h_n(\mathbf{x}(t)) = \lambda(x_{n-1}^d(t) - x_n^d(t)) - n(x_n(t) - x_{n+1}(t)) \tag{2}$$

and $x_{C+1}(t) = 0$. The process $(\mathbf{x}(t), t \geq 0)$ is referred to as the mean-field limit and the equations (1)-(2) *are referred to as the mean-field equations with the initial point $\mathbf{x}(0)$.*

It was then shown that there exists a unique fixed-point $\boldsymbol{\pi} = (\pi_n, 0 \leq n \leq C)$ of the mean-field and it is globally stable. Furthermore, they showed the exchange of limits

$$\lim_{N\to\infty} \lim_{t\to\infty} \mathbf{x}^{(N)}(t) = \lim_{t\to\infty} \lim_{N\to\infty} \mathbf{x}^{(N)}(t). \tag{3}$$

Using these results, by exploiting exchangability of the underlying processes, independence of any finite set of servers as $N \to \infty$ was shown. Furthermore, it was shown that for a server, $\mathbf{x}(t)$ and $\boldsymbol{\pi}$ denote the distribution at time t and the stationary distribution, respectively, as $N \to \infty$. As a result, the average blocking probability of a job as $N \to \infty$ is equal to π_C^d.

Unlike the case of FCFS models [3], [4] a closed form for $\boldsymbol{\pi}$ is not known. The solution can be obtained numerically as the fixed-point of a mapping. The fixed-point $\boldsymbol{\pi}$ is the unique solution to the following equation

$$\lambda(\pi_{n-1}^d - \pi_n^d) = n(\pi_n - \pi_{n+1}) \tag{4}$$

for $n \geq 1$ and $\pi_{C+1} = 0$. Then from (4), $\boldsymbol{\pi}$ is the stationary distribution of the single server Loss model with the state dependent arrival rates $(\hat{\lambda}_n, 0 \leq n \leq C)$ where $\hat{\lambda}_n = \lambda\frac{(\pi_n^d - \pi_{n+1}^d)}{(\pi_n - \pi_{n+1})}$. Let $\mathcal{P}(\{0, 1, \cdots, C\})$ be the set of probability measures on $\{0, 1, \cdots, C\}$. Then from [6], the fixed-point π can be computed as follows:

Lemma 1. *Let $\boldsymbol{\theta} = (\theta_n, 0 \leq n \leq C)$ be the unique probability distribution which is the unique fixed-point of the mapping $\Psi(\Phi)$ where the mappings $\Phi : \mathcal{P}(\{0, 1, \cdots, C\}) \mapsto \mathbf{R}_+^{C+1}$ and $\Psi : \mathbf{R}_+^{C+1} \mapsto \mathcal{P}(\{0, 1, \cdots, C\})$ are defined as follows: $\Phi((p_0, \cdots, p_C)) = (r_0, \cdots, r_C)$ and $\Psi((a_0, \cdots, a_C)) = (b_0, \cdots, b_C)$ such that*

$$r_n = \lambda\frac{(p_n^d - p_{n+1}^d)}{(p_n - p_{n+1})} \tag{5}$$

and

$$b_n = \left(\prod_{i=1}^{n} \left(\frac{a_{i-1}}{i}\right)\right) b_0, \tag{6}$$

for $n \geq 1$ and $\sum_{i=0}^{C} b_i = 1$. Then the fixed-point $\boldsymbol{\pi}$ of the mean-field is given by

$$\pi_n = \sum_{j=n}^{C} \theta_j, \tag{7}$$

$n \geq 1$.

We now provide a brief overview of previous related works in terms of the issues that they address. In [8], the randomized routing schemes for homogeneous loss models considered in this paper were first studied by using mean-field techniques. However, existence and uniqueness of a fixed-point of the mean-field were not shown. The existence and uniqueness of a fixed-point of the mean-field for homogeneous loss model of [8] was addressed in [5]–[7] by considering heterogeneous systems with an appropriate modification to the SQ(d) policy to account for server and job heterogeneity. In [5] the existence and uniqueness of a fixed-point was established under an asymptotic independence of servers ansatz while in [6] the asymptotic independence of servers and that the interchange of limits (3) was proved. In [7], they also obtained a theoretical lower bound the minimum average blocking achievable by any work conserving policy and then it was shown that the average blocking probability due to the SQ(d) policy is very close to the theoretical lower bound.

The problem of studying the accuracy of mean field approximations for finite systems was first studied in [9] for finite state CTMC models using Stein's method. It was shown that the mean square error between the stationary distribution of the system with parameter N and the fixed-point of the mean-field is $O(\frac{1}{N})$. Refined convergence rates for the mean field approximations have recently been addresses in [10] also using the Stein's method.

In this paper we develop functional central limit theorems to study the approximation of server distributions by the mean field. These results are similar in spirit to those in [11], functional central limit theorems (CLT) were obtained for the FCFS model but were not exploited further to characterize the system performance in any manner. A similar FCLT approach was also used by Hunt [12] to analyze large symmetric star loss networks under diverse routing where asymptotic independence between nodes was established. Recently there also has been interest in studying the behavior of the SQ(d) schemes in the context of heavy traffic limits of queueing systems as in [13], [14]. In [13] the interest is to show the convergence of the diffusion model to the JSQ as the sampling parameter goes to infinity. In [14] the approach is to obtain convergence rates to the mean field in the heavy traffic case or Halfin-Whitt regime. These results are in a different spirit to those in this paper where we exploit the fluctuation process to obtain the rate of convergence as a by-product.

Contributions of this work: The existing works [5]–[7] for loss models have focused on understanding the asymptotic behavior of the system for both the transient and stationary regimes when $N \to \infty$. It remains an open problem to understand the gap between the transient behavior of the system with the parameter N represented by $(\mathbf{x}^N(t), t \geq 0)$ and the asymptotic transient behavior $(\mathbf{x}(t), t \geq 0)$ obtained when $N \to \infty$. More importantly, it is of interest to quantify the resulting error when we approximate the average blocking probability of a job in the system with the index N by the asymptotic average blocking probability π_C^d. We address all these problems by studying the limiting behavior of the process $(\mathbf{z}^N(t), t \geq 0)$ as $N \to \infty$ where $\mathbf{z}^N(t) = \sqrt{N}(\mathbf{x}^{(N)}(t) - \mathbf{x}(t))$ for both the transient and stationary regimes. The process $(\mathbf{z}^N(t), t \geq 0)$ characterizes the fluctuation of the process $(\mathbf{x}^{(N)}(t), t \geq 0)$ around the mean-field $(\mathbf{x}(t), t \geq 0)$. We show that the process $(\mathbf{z}^N(t), t \geq 0)$ converges to a unique Ornstein-Uhlenbeck (OU) process for both the transient and stationary regimes. We then exploit the obtained functional central limit theorems to show that the gap between the average blocking probability in the system with the parameter N and π_C^d is $o(N^{-\frac{1}{2}})$ as $N \to \infty$.

Outline: The rest of the paper is organized as follows: We introduce the system model and notation in Section II In Section III, we give the main results of the paper. In this section, we first begin with some preliminary results on the transient regime and then we state the functional central limit theorem for the transient regime in Theorem 3. After that we present some preliminary results for the stationary regime and then we give the functional central limit theorem in Theorem 6. Finally, we conclude Section III with the result on the characterization of the error between the average blocking probability of the system with N servers and the resulting average blocking probability when $N \to \infty$. We then provide concluding remarks in Section IV. We have provided all the proofs in Appendix.

II. System model and Notation

A. System model

Consider a large-scale system with N Erlang loss servers where each server has capacity to process at most C jobs each with unit rate. We assume that servers have no buffers, and hence a job that arrives at a server with C progressing jobs is blocked or discarded. System has a central job dispatcher that has no buffer and we assume that the dispatcher routes an incoming arrival without any delay to a server according to the SQ(d) routing policy defined later in this section. In the case when the destination of an arrival has occupancy or the number of progressing jobs less than C, then the job is said to be accepted at the destination and it is then processed at unit rate. We assume that jobs arrive according to a Poisson process with rate $N\lambda$ and the job length distributions are exponential with mean equal to one.

Definition 1. *SQ(d) load balancing: Up on an arrival, the dispatcher selects d servers uniformly at random from the set of N servers*[1]*. These randomly chosen servers are referred to as the potential destination servers. The dispatcher then chooses the potential destination server with the minimum occupancy as the destination server with ties broken uniformly at random.*

B. Notation

Let $\mathcal{U}$ be the space defined as

$$\mathcal{U} \triangleq \{(u_0, u_1, \cdots, u_C) : u_0 = 1 \geq u_1 \geq \cdots \geq u_C \geq 0\}. \tag{8}$$

Clearly, we have $\mathbf{x}^{(N)}(t) \in \mathcal{U}$. The space $\mathcal{U}$ is equipped with the topology induced by the euclidean norm. We write an element of the form $(u_0, \cdots, u_C)$ as $\mathbf{u}$. The space $\mathcal{U}$ is equipped with the metric $\tau(\cdot,\cdot)$ defined as

$$\tau(\mathbf{u}, \mathbf{w}) = \|\mathbf{u} - \mathbf{w}\|_2 = \sqrt{\sum_{i=0}^{C} |u_i - w_i|^2}, \tag{9}$$

where $\mathbf{u} = (u_0, \cdots, u_C)$ and $\mathbf{w} = (w_0, \cdots, w_C)$. It can be checked that the space $\mathcal{U}$ is a Polish space. All the vectors in this paper are written by bold letters.

We assume that the stochastic processes are random elements defined on $(\Omega, \mathbb{F}, \mathbb{P})$ with sample paths in the space of càdlàg functions that are right continuous with left limits. The space of càdlàg functions is equipped with the Skorohod J_1-topology. The stochastic processes are equipped with the Borel σ-algebra generated by the open sets under the Skorohod J_1-topology. A sequence of stochastic processes $\{X_n\}_{n\geq 1}$ where X_n is defined on $(\Omega_n, \mathbb{F}_n, \mathbb{P}_n)$ is said to converge in distribution to a stochastic process X defined on $(\Omega, \mathbb{F}, \mathbb{P})$, if for every bounded, continuous, and real valued functional F, we have $\lim_{n\to\infty} \mathbb{E}_n(F(X_n)) = \mathbb{E}(F(X))$ where the expectation operators $\mathbb{E}_n, \mathbb{E}$ are defined with respect to $\mathbb{P}_n, \mathbb{P}$, respectively. We denote the convergence of $\{X_n\}_{n\geq 1}$ in distribution to X by $X_n \Rightarrow X$. For two real valued local martingales $(M_t^1, t \geq 0)$ and $(M_t^2, t \geq 0)$, let $(< M^1, M^2 >_t, t \geq 0)$ and $(< M^1 >_t, t \geq 0) = (< M^1, M^1 >_t, t \geq 0)$ be the covariation and quadratic variation processes, respectively.

III. Main results

In this section, we provide the main results of the paper.

A. Results on the transient regime:

In this section, we present results on the transient regime. For $\mathbf{a} \in \mathcal{U}$, we first define a linear operator $H(\mathbf{a}) : \mathcal{V} \mapsto \mathcal{V}$, where $\mathcal{V}$ is defined as

$$\mathcal{V} \triangleq \{(b_0, \cdots, b_C) : b_0 = 0 \text{ and } b_i \in \mathbb{R}, 1 \leq i \leq C\}. \tag{10}$$

The space $\mathcal{V}$ is equipped with the topology induced by the euclidean norm.

In the rest of the paper, without loss of generality, we say that a process $(\mathbf{y}(t), t \geq 0)$ is a solution to the equations (1)-(2), we mean that its the unique solution with the initial point

[1]It can be shown that as $N \to \infty$, it does not matter whether servers are selected with replacement or without replacement. Hence the analysis assumes that servers are selected with replacement.

$\mathbf{y}(0)$. The linearization of (1)-(2) around a solution $(\mathbf{y}(t), t \geq 0)$ with an initial point $\mathbf{y}(0)$ is given by

$$\frac{d\mathbf{s}(t)}{dt} = H(\mathbf{y}(t))\mathbf{s}(t), \tag{11}$$

where for $\mathbf{a} \in \mathcal{U}$ and $\mathbf{b} \in \mathcal{V}$, $H(\mathbf{a}) : \mathcal{V} \mapsto \mathcal{V}$ is a linear operator defined as, for $n \geq 1$,

$$(H(\mathbf{a})\mathbf{b})_n = \lambda d a_{n-1}^{d-1} b_{n-1} - (\lambda d a_n^{d-1} + n) b_n + n b_{n+1}. \tag{12}$$

Any solution $(\mathbf{r}(t), t \geq 0)$ to (11) can be written as $\mathbf{r}(t) = \mathbf{w}(t) - \mathbf{y}(t)$, where $(\mathbf{w}(t), t \geq 0)$ is a solution to the equations (1)-(2). We can also write the linear operator $H(\mathbf{a})$ as a matrix in the canonical basis $(0,1,0,\cdots,0)$, $(0,0,1,0,\cdots,0)$, ..., $(0,0,\cdots,0,1)$ where the size of each vector is $C+1$ and we write $H(\mathbf{a})$ as the following matrix of size $C \times C$:

$$H(\mathbf{a}) = \begin{bmatrix} -\beta_1 & 1 & 0 & \cdots & 0 & 0 \\ \gamma_1 & -\beta_2 & 2 & \cdots & 0 & 0 \\ 0 & \gamma_2 & -\beta_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & -\beta_{C-1} & C-1 \\ 0 & 0 & 0 & \cdots & \gamma_{C-1} & -\beta_C \end{bmatrix},$$

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

Let us now define two operators $W_1, W_2 : \mathcal{U} \mapsto \mathcal{V}$ as follows: for $\mathbf{a} \in \mathcal{U}$, let $(W_1(\mathbf{a}))_0 = 0$, $(W_2(\mathbf{a}))_0 = 0$ and for $n \geq 1$,

$$(W_1(\mathbf{a}))_n = \lambda(a_{n-1}^d - a_n^d),\ (W_2(\mathbf{a}))_n = n(a_n - a_{n+1}). \tag{13}$$

Note that from (2) and (13), we have $h_n(\mathbf{a}) = (W_1(\mathbf{a}))_n - (W_2(\mathbf{a}))_n$. We now define the process $(\mathbf{M}(t), t \geq 0)$, where $\mathbf{M}(t) = (M_i(t), 0 \leq i \leq C)$ such that $\{(M_i(t), t \geq 0)\}_{i \in \{0,1,\cdots,C\}}$ are independent real valued continuous and centered Gaussian martingales, determined in law by their deterministic quadratic variation process given by, for $n \geq 0$,

$$< M_n >_t = \int_{s=0}^{t} ((W_1(\mathbf{x}(s))_n + (W_2(\mathbf{x}(s))_n)\, ds. \tag{14}$$

We have that both $\mathbf{M}(t)$ and $(< M_i >_t, 0 \leq i \leq C)$ have values in $\mathcal{V}$. From (14), for any $t > 0$, since $((W_1(\mathbf{a}))_n + (W_2(\mathbf{a}))_n)$ is uniformly bounded in n and $\mathbf{a}$ as $0 \leq a_i \leq 1$, the martingale $(\mathbf{M}(t), t \geq 0)$ is square integrable. As we will show later in this section, the process $(\mathbf{z}^{(N)}(t), t \geq 0)$ converges in distribution to a solution to the following inhomogeneous stochastic differential equation (SDE) given by, for $t \geq 0$,

$$\mathbf{z}(t) = \mathbf{z}(0) + \int_{s=0}^{t} H(\mathbf{x}_s)\mathbf{z}_s\, ds + \mathbf{M}(t). \tag{15}$$

A solution to (15) is an Ornstein-Uhlenbeck process (OU) process.

We first focus on the study of the SDE (15). In particular, we present the following results on existence and uniqueness of a solution to (15).

Theorem 2. *The following results are true:*

- *For $\mathbf{a} \in \mathcal{U}$, the operator norm of $H(\mathbf{a})$ is uniformly bounded in $\mathbf{a} \in \mathcal{U}$ and $B_H = \sqrt{32(\lambda^2 d^2 + C^2)}$ is a uniform bound on the operator norm of $H(\mathbf{a})$.*
- *If $\mathbb{E}\left[\|\mathbf{z}(0)\|_2^2\right] < \infty$, then there is a unique strong solution to (15) given by $\mathbf{z}_t = e^{\int_{s=0}^{t} H(\mathbf{x}(s))\, ds}\mathbf{z}(0) + \int_{s=0}^{t} e^{\int_{r=s}^{t} H(\mathbf{x}(r))\, dr}\, d\mathbf{M}(s)$ and also, $\mathbb{E}\left[\sup_{t \leq T} \|z(t)\|_2^2\right] < \infty$.*

The proof of the Theorem 2 is given in Appendix A.

So far, we have studied the SDE (15). We now focus on the proof of the result that the process $(\mathbf{z}^{(N)}(t), t \geq 0)$ converges in distribution to the process $(\mathbf{z}(t), t \geq 0)$. We first derive the evolution equations of the process $(\mathbf{z}^{(N)}(t), t \geq 0)$. The probability that the destination server of a job that arrives to the system when the system state is $\mathbf{b} = (b_0, \cdots, b_C)$ has occupancy n, is equal to $\frac{(b_n^d - b_{n+1}^d)}{(b_n - b_{n+1})}$. Let $\{(\mathcal{N}_i(t), t \geq 0)\}_{i \geq 1}$ and $\{(\mathcal{D}_i(t), t \geq 0)\}_{i \geq 1}$ be the collection of mutually independent unit rate Poisson processes. We use $(\mathcal{N}_i(t), t \geq 0)$ to model the arrival process to servers that have $i-1$ progressing jobs. Similarly, we use $\{(\mathcal{D}_i(t), t \geq 0)\}_{i \geq 1}$ to model the departure process from servers that have i progressing jobs. Since the service time of each job is an exponentially distributed random variable with unit mean, from [15], we can write $\mathbf{x}_0^{(N)}(t) = 1$ and

$$\begin{aligned} \mathbf{x}_i^{(N)}(t) &= \mathbf{x}_i^{(N)}(0) \\ &\quad + \frac{1}{N}\mathcal{N}_i\Big(N\lambda \int_{s=0}^{t} ((\mathbf{x}_{i-1}^{(N)}(s))^d - (\mathbf{x}_i^{(N)}(s))^d)\, ds\Big) \\ &\quad - \frac{1}{N}\mathcal{D}_i\Big(N \int_{s=0}^{t} ((\mathbf{x}_i^{(N)}(s)) - (\mathbf{x}_{i+1}^{(N)}(s)))\, ds\Big), \end{aligned} \tag{16}$$

for all $i \geq 1$.

Let W be the operator defined as

$$W = W_1 - W_2. \tag{17}$$

We first define an independent square-integrable martingales $(\mathbf{M}^{(N)}(t), t \geq 0) = \{(M_i^{(N)}(t), t \geq 0)\}_{(i \in 0,1,\cdots,C)}$ such that $(\mathbf{M}^{(N)}(t), t \geq 0)$ is independent of $\mathbf{z}^{(N)}(0)$ and for $i \geq 1$,

$$< M_i^{(N)} >_t = \int_{s=0}^{t} ((W_1(\mathbf{x}^{(N)}(s)))_i + (W_2(\mathbf{x}^{(N)}(s)))_i)\, ds. \tag{18}$$

From (1), (2), (16), and the definition of $(\mathbf{z}^{(N)}(t), t \geq 0)$, as in [15], we can write

$$\begin{aligned} \mathbf{z}^{(N)}(t) = \mathbf{z}^{(N)}(0) + \int_{s=0}^{t} \sqrt{N}(W(\mathbf{x}^{(N)}(s)) - W(\mathbf{x}(s)))\, ds \\ + \mathbf{M}^{(N)}(t). \end{aligned} \tag{19}$$

We now state the following result which is used in the proof of the subsequent theorem.

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

$$\limsup_{N \to \infty} \mathbb{E}\left[\sup_{0 \leq t \leq T} \|\mathbf{z}^{(N)}(t)\|_2^2\right] < \infty \tag{20}$$

The proof is given in Appendix B.

We next present the main result on the convergence of the process $(\mathbf{z}^{(N)}(t), t \geq 0)$ as $N \to \infty$.

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 to* (15) *with the initial point* $\mathbf{z}(0)$.

The proof is given in Appendix C.

B. Results on the stationary regime:

In this section, we provide results on the stationary regime. In this case, the initial point of the mean-field $\mathbf{x}(0)$ coincides with fixed-point of the mean-field $\boldsymbol{\pi}$, and hence $\mathbf{x}(0) = \boldsymbol{\pi}$. Let $H(\boldsymbol{\pi}) = A$. Since $\boldsymbol{\pi}$ is the fixed-point of the mean-field, $W(\boldsymbol{\pi}) = W_1(\boldsymbol{\pi}) - W_2(\boldsymbol{\pi}) = 0$.

If we linearize (2) around $\boldsymbol{\pi}$, we get

$$\frac{d\mathbf{z}(t)}{dt} = A\mathbf{z}(t). \tag{21}$$

Let us now define $\mathbf{B}_t = (B_i(t), 0 \leq i \leq C)$ with $B_0(t) = 0$ such that $\{(B_i(t), t \geq 0)\}_{0 \leq i \leq C}$ are independent centered Brownian motions and $V_n = var(B_n(1)) = \mathbb{E}\left[B_n^2(1)\right] = 2n(\pi_n - \pi_{n+1})$. For $\mathbf{x}(0) = \boldsymbol{\pi}$, from (14), the martingales $(\mathbf{M}(t), t \geq 0)$ with $\mathbf{x}(0) = \boldsymbol{\pi}$ has the same law as $(\mathbf{B}(t), t \geq 0)$. The process $(\mathbf{B}(t), t \geq 0)$ has diagonal infinitesimal covariance matrix diag($\mathbf{V}$), where $\mathbf{V} = (V_n, 0 \leq n \leq C)$.

We now introduce the following SDE,

$$\mathbf{z}(t) = \mathbf{z}(0) + \int_{s=0}^{t} A\mathbf{z}(s)\, ds + \mathbf{B}(t). \tag{22}$$

Any solution to (22) is an OU process. We will show later in this section, the limit of $(\mathbf{z}^{(N)}(t), t \geq)$ in the stationary regime as $N \to \infty$ is the unique stationary OU process solving the SDE (22).

Then similar to the Theorem 2, by assuming that $\mathbf{z}(0)$ in (22) is an arbitrary initial value whose law is not an invariant law, we have the following result and the proof follows by the same arguments. Hence, it is omitted.

Theorem 4. *The following results are true:*

- *The operator norm of A is bounded.*
- *If* $\mathbb{E}\left[\|\mathbf{z}(0)\|_2^2\right] < \infty$, *then there is a unique strong solution to* (22) *given by* $\mathbf{z}(t) = e^{At}\mathbf{z}(0) + \int_{s=0}^{t} e^{A(t-s)}\, d\mathbf{B}(s)$ *and also,* $\mathbb{E}\left[\sup_{t \leq T} \|z(t)\|_2^2\right] < \infty$.

Note that the transpose of A denoted by A^* is the generator of a finite state birth-death process with killing rates. In state i for $1 \leq i \leq C$, the birth, death, and killing rates are equal to γ_i, $i-1$, and 1, respectively. Based on the spectral decomposition of A, there is a spectral gap for the operator A. The proof is similar to the proof of Theorem 2.9 of [11], and hence we skip the proof.

Lemma 3. *The unique solution* $(\mathbf{z}(t), t \geq 0)$ *where* $\mathbf{z}(t) = e^{At}\mathbf{z}(0)$ *to* (21) *satisfies that for some* $\delta > 0$ *and* $D < \infty$,

$$\|\mathbf{z}(t)\|_2 \leq e^{-\delta t} D \|\mathbf{z}(0)\|_2. \tag{23}$$

As a result, from Lemma 3 and the unique solution given in Theorem 4, the following result follows immediately. Hence, we skip the proof.

Theorem 5. *Any solution for the* (22) *converges to its unique invariant law as* $t \to \infty$. *The invariant law is the law of the Gaussian centered process* $\int_0^\infty e^{At}\, d\mathbf{B}(t)$ *that has the covariance matrix* $\int_0^\infty e^{At} diag(\mathbf{V}) e^{A^* t}\, dt$. *In stationary, there is a unique solution to* (22).

The following result on the exponential stability of the mean-field follows mutatis-mutandis the proof of Theorem 2.12 of [11]. Hence, the proof is omitted.

Lemma 4. *The following result is true:*

- *The mean-field* $(\mathbf{x}(t), t \geq 0)$ *satisfies that for some* $\delta > 0$ *and* $D < \infty$,

$$\|(x)(t) - \boldsymbol{\pi}\|_2 \leq e^{-\delta t} D \|\mathbf{x}(0) - \boldsymbol{\pi}\|_2. \tag{24}$$

We now state the following important result based on Lemma 4.

Lemma 5. *For* $\mathbf{z}^{(N)}(t) = \sqrt{N}(\mathbf{x}^{(N)}(t) - \boldsymbol{\pi})$, *If* $\limsup_{N \to \infty} \mathbb{E}\left[\|\mathbf{z}^{(N)}(0)\|_2^2\right] < \infty$, *then*

$$\limsup_{N \to \infty} \sup_{t \geq 0} \mathbb{E}\left[\|\mathbf{z}^{(N)}(t)\|_2^2\right] < \infty. \tag{25}$$

As a result, under the invariant laws, we have

$$\limsup_{N \to \infty} \mathbb{E}\left[\|\mathbf{z}^{(N)}(0)\|_2^2\right] < \infty. \tag{26}$$

The proof is given in Appendix D.

We now provide the main-result the functional central limit theorem in equilibrium.

Theorem 6. *Assume that the system with parameter N is in the stationary regime. Using Theorem 5, we show that the process* $(\mathbf{z}^{(N)}(t), t \geq 0)$ *converges in law to the unique stationary OU process which solves the equation* (22). *The sequence* $(\mathbf{z}^{(N)}(0)$ *in stationary converges in law to the invariant law of this process.*

The proof is given in Appendix E.

We now use the functional central limit theorems to characterize the error between the average blocking probability in the system with N servers and the average blocking probability of the limiting system given by π_C^d.

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

$$P_{block}^{(N)} - \pi_C^d = o(N^{-\frac{1}{2}}). \tag{27}$$

The proof is given in Appendix F.

The significance of the Theorem 7 is that although the exact blocking formula for $P_{block}^{(N)}$ is not known and it is difficult to characterize due to complex interactions between servers, but as $N \to \infty$ the blocking probability as a function of $\boldsymbol{\pi}$ which can be computed using the generalized Erlang formula as in Lemma 1, is very close to the actual value for a system with large number of servers N.

IV. Concluding remarks

In this paper, we have obtained functional central limit theorems by studying fluctuations of the stochastic empirical processes around the mean-field for both the transient and stationary regimes when $N \to \infty$. We showed that the limiting process is an OU process. A consequence of the functional central limit theorems of this paper is that the error between the actual average blocking probability in the system with N servers and the average blocking π_C^d of the limiting model when $N \to \infty$ is $o(N^{-\frac{1}{2}})$. In future work we will study the fluctuation process of the considered model in heavy traffic where the blocking probabilities are known to have a different behavior [16] in the context of regular Erlang loss models. This will enable us study what type of gains are achievable in heavily loaded systems. Furthermore it would be of interest to exploit these FCLT results to obtain bounds on the convergence rate of joint distributions to their mean field products as in [12]. Such results will provide a better understanding of mean-field approximations and their validity.

References

[1] "Microsoft Azure," http://www.microsoft.com/windowsazure/.

[2] "Amazon EC2," http://aws.amazon.com/ec2/.

[3] M. Mitzenmacher, "The power of two choices in randomized load balancing," *PhD Thesis, Berkeley*, 1996.

[4] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich, "Queueing system with selection of the shortest of two queues: an asymptotic approach," *Problems of Information Transmission*, vol. 32, no. 1, pp. 20–34, 1996.

[5] Q. Xie, X. Dong, Y. Lu, and R. Srikant, "Power of d choices for large-scale bin packing: A loss model," in *Proceedings of the 2015 ACM SIGMETRICS*, 2015, pp. 321–334.

[6] A. Mukhopadhyay, R. R. Mazumdar, and F. Guillemin, "The power of randomized routing in heterogeneous loss systems," in *Teletraffic Congress (ITC 27), 2015 27th International*, 2015, pp. 125–133.

[7] A. Mukhopadhyayay, A. Karthik, R. R. Mazumdar, and F. M. Guillemin, "Mean field and propagation of chaos in multi-class heterogeneous loss models," *Performance Evaluation*, vol. 91, pp. 117–131, September 2015.

[8] S. R. E. Turner, "Resource pooling in stochastic networks," *Ph.D. dissertation, University of Cambridge*, 1996.

[9] L. Ying, "On the approximation error of mean-field models," in *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, ser. SIGMETRICS '16. New York, NY, USA: ACM, 2016, pp. 285–297.

[10] N. Gast and B. Van Houdt, "A refined mean field approximation," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, no. 2, pp. 33:1–33:28, Dec. 2017.

[11] C. Graham, "Functional central limit theorems for a large network in which customers join the shortest of several queues," *Probability Theory and Related Fields*, vol. 131, no. 1, pp. 97–120, Jan 2005.

[12] P. J. Hunt, "Loss networks under diverse routing: The symmetric star network," *Advances in Applied Probability*, vol. 27, no. 1, pp. 255–272, 1995.

[13] D. Mukherjee, S. Borst, J. van Leeuwaarden, and P. Whiting, "Universality of power-of-d load balancing in many-server systems," *Stochastic Systems*, vol. 8, no. 4, pp. 265–292, 2018.

[14] L. Ying, "Stein's method for mean field approximations in light and heavy traffic regimes," *POMACS*, vol. 1, no. 1, pp. 12:1–12:27, 2017.

[15] G. Pang, R. Talreja, and W. Whitt, "Martingale proofs of many-server heavy-traffic limits for markovian queues," *Probab. Surveys*, vol. 4, pp. 193–267, 2007.

[16] P. Gazdzicki, I. Lambadaris, and R. Mazumdar, "Blocking probabilities for large multi-rate erlang loss systems," *Adv.Appl.Prob.*, vol. 25, pp. 997–1009, 1993.

[17] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*. John Wiley and Sons Ltd, 1985.

[18] A. Joffe and M. Metivier, "Weak convergence of sequences of semi-martingales with applications to multitype branching processes," *Advances in Applied Probability*, vol. 18, no. 1, pp. 20–65, 1986.

Appendix

A. Proof of Theorem 2

We first prove that $\|H(\mathbf{a})\|_2$ is uniformly bounded in $\mathbf{a} \in \mathcal{U}$. For $\mathbf{y} \in \mathcal{V}$, we have

$$\|H(\mathbf{a})\mathbf{y}\|_2^2 = \sum_{i=0}^{C} \left| \lambda d a_{n-1}^{d-1} y_{n-1} - (\lambda d a_n^{d-1} + n) y_n + n y_{n+1} \right|^2. \tag{28}$$

Since $a_i \leq 1$ for all $0 \leq i \leq C$, we get $\|H(\mathbf{a})\mathbf{y}\|_2^2 \leq 32(\lambda^2 d^2 + C^2)\|y\|_2^2$. Hence we get $\|H(\mathbf{a})\|_2^2 \leq 32(\lambda^2 d^2 + C^2)$. Hence B_H is a bound on the operator norm of $H(\mathbf{a})$ uniformly in $\mathbf{a}$.

We now give the proof of the uniqueness of a solution by using the Gronwall's Lemma. Let $(\mathbf{z}^{(1)}(t), t \geq 0)$ and $(\mathbf{z}^{(2)}(t), t \geq 0)$ be two solutions with initial points $\mathbf{z}^{(1)}(0)$ and $\mathbf{z}^{(2)}(0)$, respectively. Then

$$\mathbf{z}^{(1)}(t) - \mathbf{z}^{(2)}(t) = \mathbf{z}^{(1)}(0) - \mathbf{z}^{(2)}(0) + \int_{s=0}^{t} H(\mathbf{x}_s)(\mathbf{z}^{(1)}(s) - \mathbf{z}^{(2)}(s))\, ds. \tag{29}$$

Then

$$\|\mathbf{z}^{(1)}(t) - \mathbf{z}^{(2)}(t)\|_2 \leq \|\mathbf{z}^{(1)}(0) - \mathbf{z}^{(2)}(0)\|_2 + B_H \int_{s=0}^{t} \|\mathbf{z}^{(1)}(s) - \mathbf{z}^{(2)}(s)\|_2\, ds \tag{30}$$

By the Gronwall's Lemma,

$$\|\mathbf{z}^{(1)}(t) - \mathbf{z}^{(2)}(t)\|_2 \leq e^{B_H t}\|\mathbf{z}^{(1)}(0) - \mathbf{z}^{(2)}(0)\|_2 \tag{31}$$

Hence,

$$\mathbb{E}\left[\|\mathbf{z}^{(1)}(t) - \mathbf{z}^{(2)}(t)\|_2^2\right] \leq e^{2B_H t}\mathbb{E}\left[\|\mathbf{z}^{(1)}(0) - \mathbf{z}^{(2)}(0)\|_2^2\right]. \tag{32}$$

If $\mathbf{z}^{(1)}(0) = \mathbf{z}^{(2)}(0)$, then $\mathbf{z}^{(1)}(t) = \mathbf{z}^{(2)}(t)$ *a.s.* for all rational t. Since $(\mathbf{z}^{(1)}(t), t \geq 0)$ and $(\mathbf{z}^{(2)}(t), t \geq 0)$ have continuous sample paths, $\mathbf{z}^{(1)}(t) = \mathbf{z}^{(2)}(t)$ *a.s.* for all $t \geq 0$.

Finally, we give the proof of the claim $\mathbb{E}\left[\sup_{t \leq T} \|\mathbf{z}(t)\|_2^2\right] < \infty$. The proof uses the Gronwall's lemma, and Doob's L^2 inequality. It can be seen that

$$\|\mathbf{z}(t)\|_2 \leq \|\mathbf{z}(0)\|_2 + B_H \int_{s=0}^{t} \|\mathbf{z}(s)\|_2\, ds + \mathbf{M}(\mathbf{t})_2. \tag{33}$$

Therefore, for any $T > 0$,

$$\|\mathbf{z}(t)\|_2^2 \leq 3\left(\|\mathbf{z}(0)\|_2^2 + B_H^2 \Big(\int_{s=0}^{t} \|\mathbf{z}(s)\|_2\, ds \Big)^2 + \|\mathbf{M}(t)\|_2^2 \right). \tag{34}$$

By applying the Cauchy-Schwartz inequality,

$$\|\mathbf{z}(t)\|_2^2 \leq 3\left(\|\mathbf{z}(0)\|_2^2 + B_H^2 t \int_{s=0}^{t} \|\mathbf{z}(s)\|_2^2\, ds + \|\mathbf{M}(t)\|_2^2 \right). \tag{35}$$

By using the Gronwall's Lemma,

$$\|\mathbf{z}(t)\|_2^2 \leq 3\left(\|\mathbf{z}(0)\|_2^2 + \|\mathbf{M}(t)\|_2^2 \right) e^{3B_H^2 t}. \tag{36}$$

Hence, we have

$$\sup_{0\le t\le T}\|\mathbf{z}(t)\|_2^2 \le 3\left(\|\mathbf{z}(0)\|_2^2+\sup_{0\le t\le T}\|\mathbf{M}(t)\|_2^2\right)e^{3B_H^2T^2}. \tag{37}$$

We can write

$$\mathbb{E}\left[\sup_{0\le t\le T}\|\mathbf{z}(t)\|_2^2\right] \le 3e^{3B_H^2T^2}\left(\mathbb{E}\left[\|\mathbf{z}(0)\|_2^2\right] + \mathbb{E}\left[\sup_{0\le t\le T}\|\mathbf{M}(t)\|_2^2\right]\right). \tag{38}$$

By using the Doob's inequality

$$\mathbb{E}\left[\sup_{0\le t\le T}\|\mathbf{z}(t)\|_2^2\right] \le 3e^{3B_H^2T^2}\left(\mathbb{E}\left[\|\mathbf{z}(0)\|_2^2\right] + 4\mathbb{E}\left[\|\mathbf{M}(T)\|_2^2\right]\right). \tag{39}$$

Therefore

$$\mathbb{E}\left[\sup_{0\le t\le T}\|\mathbf{z}(t)\|_2^2\right] \le 3e^{3B_H^2T^2}\left(\mathbb{E}\left[\|\mathbf{z}(0)\|_2^2\right] + 4\mathbb{E}\left[\sum_{i=1}^{C}< M_i >_T\right]\right). \tag{40}$$

Since $< M_i >_T$ is finite for every T and i, $\mathbb{E}\left[\sup_{0\le t\le T}\|\mathbf{z}(t)\|_2^2\right] < \infty$.

From the SDE (15), by applying the Ito's formula to the function $g(t,\mathbf{a}) = e^{-\int_{r=0}^{t}H(\mathbf{x}(r))\,dr}\mathbf{a}$, we obtain the proposed solution and it is well-defined.

B. Proof of Lemma 2

We first recall that the process $(\mathbf{z}^{(N)}(t), t\ge 0)$ satisfies

$$\mathbf{z}^{(N)}(t) = \mathbf{z}^{(N)}(0) + \int_{s=0}^{t}\sqrt{N}(W(\mathbf{x}^{(N)}(s)) - W(\mathbf{x}(s)))\,ds + \mathbf{M}^{(N)}(t), \tag{41}$$

where for $i \ge 1$,

$$< M_i^{(N)} >_t = \int_{s=0}^{t}((W_1(\mathbf{x}^{(N)}(s)))_i + (W_2(\mathbf{x}^{(N)}(s)))_i)\,ds. \tag{42}$$

It can be verified that the operator W is Lipschitz continuous and there exists a constant B_W such that for all $\mathbf{u},\mathbf{v}\in\mathcal{U}$

$$\|W(\mathbf{u}) - W(\mathbf{v})\|_2 \le B_W\|\mathbf{u}-\mathbf{v}\|_2. \tag{43}$$

From (41), we can write

$$\|\mathbf{z}^{(N)}(t)\|_2 \le \|\mathbf{z}^{(N)}(0)\|_2 + B_W\int_{s=0}^{t}\|\mathbf{z}^{(N)}(s)\|_2\,ds + \|\mathbf{M}^{(N)}(t)\|_2, \tag{44}$$

By the Gronwall's Lemma,

$$\|\mathbf{z}^{(N)}(t)\|_2 \le (\|\mathbf{z}^{(N)}(0)\|_2 + \|\mathbf{M}^{(N)}(t)\|_2)e^{B_Wt}, \tag{45}$$

Hence,

$$\|\mathbf{z}^{(N)}(t)\|_2^2 \le 2(\|\mathbf{z}^{(N)}(0)\|_2^2 + \|\mathbf{M}^{(N)}(t)\|_2^2)e^{2B_Wt}, \tag{46}$$

Therefore

$$\sup_{0\le t\le T}\|\mathbf{z}^{(N)}(t)\|_2^2 \le 2(\|\mathbf{z}^{(N)}(0)\|_2^2 + \sup_{0\le t\le T}\|\mathbf{M}^{(N)}(t)\|_2^2)e^{2B_WT}, \tag{47}$$

By using the Doob's inequality,

$$\mathbb{E}\left[\sup_{0\le t\le T}\|\mathbf{z}^{(N)}(t)\|_2^2\right] \le 2e^{2B_WT}\left(\mathbb{E}\left[\|\mathbf{z}^{(N)}(0)\|_2^2\right] + 4\mathbb{E}\left[\sum_{i=1}^{C}< M_i^{(N)} >_T\right]\right), \tag{48}$$

Since $\mathbb{E}\left[\sum_{i=1}^{C}< M_i^{(N)} >_T\right]$ is uniformly bounded in N, it implies that $\limsup_{N\to\infty}\mathbb{E}\left[\sup_{0\le t\le T}\|\mathbf{z}^{(N)}(t)\|_2^2\right] < \infty$.

C. Proof of Theorem 3

We recall that since the space $\mathcal{V}$ is Polish, the space of càdlàg functions under the Skhorohod topology is a Polish space. Hence from the Prohorov's theorem [17], tightness is equivalent to relative compactness. Therefore we first need to show the tightness and then we need to show that every limiting point has the same Law as the unique OU process with the initial point $\mathbf{z}(0)$.

To show the tightness of $(\mathbf{z}^{(N)}(t), t\ge 0)$, we need to show that the Theorem 4.1 of [17, page 354] holds. For this, we first establish several useful results.

Since $\mathbf{z}^{(N)}(0) \Rightarrow \mathbf{z}(0)$, it implies that the sequence $\mathbf{z}^{(N)}(0)$ is tight. Let $Q(r)$ be the closed ball with radius r centered at $\mathbf{0}$. For every $\epsilon > 0$, there exists $r_\epsilon < \infty$ such that $\mathbb{P}(\mathbf{z}^{(N)}(0)\in Q(r_\epsilon)) > 1-\epsilon$ for all $N\ge 1$. We now define a random variable $\mathbf{x}^{(N,\epsilon)}(0)$ such that it coincides with $\mathbf{x}^{(N)}(0)$ on $\{\mathbf{z}^{(N)}(0)\in Q(r_\epsilon)\}$ and $\mathbf{z}^{(N,\epsilon)}(0)$ is uniformly bounded in N on $\{\mathbf{z}^{(N)}(0)\notin Q(r_\epsilon)\}$. Then by using coupling arguments, the processes $(\mathbf{z}^{(N,\epsilon)}(t), t\ge 0)$ and $(\mathbf{z}^{(N)}(t), t\ge 0)$ coincide on $\{\mathbf{z}^{(N)}(0)\in Q(r_\epsilon)\}$. Hence, without loss of generality, we assume that $\mathbf{z}^{(N)}(0)$ is uniformly bounded in N. As a result, the result stated in Lemma 2 can be used in the rest of the proof.

We next recall the following useful result from [11, Lemma 3.3]. For a and h in $\mathbb{R}$, let $B(a,h) = (a+h)^d - a^d - da^{d-1}h$, then if both a and $a+h$ lie in $[0,1]$, we have

$$0 \le B(a,h) \le h^d + (2^d - d - 2)ah^2. \tag{49}$$

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

$$(G(\mathbf{r},\mathbf{y}))_n = \lambda B(r_{n-1},y_{n-1}) - \lambda B(r_n,y_n). \tag{50}$$

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

$$W(\mathbf{r}+\mathbf{y}) - W(\mathbf{r}) = H(\mathbf{r})\mathbf{y} + G(\mathbf{r},\mathbf{y}). \quad (51)$$

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

$$\mathbf{x}^{(N)}(t) = \mathbf{x}(t) + \frac{\mathbf{z}^{(N)}(t)}{\sqrt{N}}. \quad (52)$$

Here, $\mathbf{x}^{(N)}(t)$, $\mathbf{x}(t) \in \mathcal{U}$ and $\frac{\mathbf{z}^{(N)}(t)}{\sqrt{N}} \in \mathcal{V}$. Hence, from (51), we have

$$W(\mathbf{x}^{(N)}(t) - W(\mathbf{x}(t)) = H(\mathbf{x}(t))\frac{\mathbf{z}^{(N)}(t)}{\sqrt{N}} + G\left(\mathbf{x}(t), \frac{\mathbf{z}^{(N)}(t)}{\sqrt{N}}\right). \quad (53)$$

Finally, since $(z_i^{(N)}(t), t \geq 0)$ has jumps of size $\frac{1}{\sqrt{N}}$ and the fact that the SDE (15) has a unique solution, by using (53),(49), and Theorem 2, we get that $(\mathbf{z}^{(N)}(t), t \geq 0)$ converges to the unique solution of the SDE (15). The proof follows by the same arguments as in the proof of the Theorem 4.1 on page 354 in [17] or Theorem 3.3.1 of [18]. This completes the proof.

D. Proof of Lemma 5

Let $\mathbf{y}_h(\mathbf{v})$ be the solution to the mean-field equation (2) at time h with an initial point $\mathbf{v}$. Since we are working in the stationary regime, we consider that the mean-field is at its fixed-point. Hence, we have

$$\mathbf{z}^N(t) = \sqrt{N}(\mathbf{x}^{(N)}(t) - \boldsymbol{\pi}). \quad (54)$$

Then we can write, for $t_0 \geq 0$,

$$\mathbf{z}^N(t_0+h) = \sqrt{N}(\mathbf{x}^{(N)}(t_0+h) - \mathbf{y}_h(\mathbf{x}^{(N)}(t_0))) + \sqrt{N}(\mathbf{y}_h(\mathbf{x}^{(N)}(t_0)) - \boldsymbol{\pi}). \quad (55)$$

Let $\mathbf{z}^N(t_0,h) = \sqrt{N}(\mathbf{x}^{(N)}(t_0+h) - \mathbf{y}_h(\mathbf{x}^{(N)}(t_0)))$. Then we have

$$\mathbf{z}^N(t_0+h) = \mathbf{z}^N(t_0,h) + \sqrt{N}(\mathbf{y}_h(\mathbf{x}^{(N)}(t_0)) - \boldsymbol{\pi}). \quad (56)$$

As a consequence, the following relationship holds from Lemma 4:

$$\|\mathbf{z}^N(t_0+h)\|_2 \leq \|\mathbf{z}^N(t_0,h)\|_2 + e^{-\delta h} D \|\mathbf{z}^{(N)}(t_0)\|_2. \quad (57)$$

Also, we have

$$\|\mathbf{x}^{(N)}(t_0+h)\|_2 \leq \|\boldsymbol{\pi}\|_2 + N^{-\frac{1}{2}}\|\mathbf{z}^{(N)}(t_0+h)\|_2 \quad (58)$$

Then from (41), (57), (58), for $T \geq 0$, we can find using Gronwall's Lemma some constant K_T such that

$$\sup_{0\leq h\leq T} \|\mathbf{z}^{(N)}(t_0,h)\|_2 \leq K_T\big(N^{-\frac{1}{2}}\|\boldsymbol{\pi}\|_2 + N^{-1}D\|\mathbf{z}^{(N)}(t_0)\|_2 + \sup_{0\leq h\leq T}\|\mathbf{M}^{(N)}(t_0+h) - \mathbf{M}^{(N)}(t_0)\|_2\big). \quad (59)$$

As a result, using (57) and (59), there exists a constant S_T such that we have

$$\mathbb{E}\left[\|\mathbf{z}^{(N)}(t_0+h)\|_2^2\right] \leq S_T + 2(K_T N^{-1} + e^{-\delta h})^2 D^2 \mathbb{E}\left[\|\mathbf{z}^{(N)}(t_0)\|_2^2\right]. \quad (60)$$

We now select a large value of T such that $8e^{-2\delta T}D^2 \leq \epsilon < 1$. Then fo all $N \geq K_T e^{\delta T}$ and an integer m, we get

$$\mathbb{E}\left[\|\mathbf{z}^{(N)}((m+1)T)\|_2^2\right] \leq S_T + \epsilon \mathbb{E}\left[\|\mathbf{z}^{(N)}(mT)\|_2^2\right]. \quad (61)$$

By using the induction method, we can write

$$\mathbb{E}\left[\|\mathbf{z}^{(N)}(mT)\|_2^2\right] \leq S_T(\sum_{j=1}^{m}\epsilon^{j-1}) + \epsilon^m \mathbb{E}\left[\|\mathbf{z}^{(N)}(0)\|_2^2\right] \leq \frac{S_T}{1-\epsilon} + \mathbb{E}\left[\|\mathbf{z}^{(N)}(0)\|_2^2\right]. \quad (62)$$

By using (60),

$$\sup_{0\leq h\leq T} \mathbb{E}\left[\|\mathbf{z}^{(N)}(mT+h)\|_2^2\right] \leq S_T + 8D^2\mathbb{E}\left[\|\mathbf{z}^{(N)}(mT)\|_2^2\right] \quad (63)$$

Hence, from (62), we have

$$\sup_{0\leq h\leq T} \mathbb{E}\left[\|\mathbf{z}^{(N)}(mT+h)\|_2^2\right] \leq S_T + 8D^2\left(\frac{S_T}{1-\epsilon} + \mathbb{E}\left[\|\mathbf{z}^{(N)}(0)\|_2^2\right]\right). \quad (64)$$

Since m is arbitrary, we have

$$\sup_{t\geq 0} \mathbb{E}\left[\|\mathbf{z}^{(N)}(t)\|_2^2\right] \leq S_T + 8D^2\left(\frac{S_T}{1-\epsilon} + \mathbb{E}\left[\|\mathbf{z}^{(N)}(0)\|_2^2\right]\right). \quad (65)$$

Let $\mathbf{z}^{(N)}(\infty)$ be a random variable with the invariant law, then from Ergodicity and Fatou Lemma, we have

$$\mathbb{E}\left[\|\mathbf{z}^{(N)}(\infty)\|_2^2\right] \leq \liminf_{t\geq 0}\mathbb{E}\left[\|\mathbf{z}^{(N)}(t)\|_2^2\right] \leq \sup_{t\geq 0}\mathbb{E}\left[\|\mathbf{z}^{(N)}(t)\|_2^2\right]. \quad (66)$$

Therefore, to obtain the invariant law bound, we need to show $\limsup_{N\to\infty}\mathbb{E}\left[\|\mathbf{z}^{(N)}(\infty)\|_2^2\right] < \infty$. This follows if we can find $\mathbf{x}^{(N)}(0)$ such that $\limsup_{N\to\infty}\mathbb{E}\left[\|\mathbf{z}^{(N)}(0)\|_2^2\right] < \infty$ holds. We can select $\mathbf{x}^{(N)}(0)$ such that for $n \geq 1$, $x_n^{(N)}(0) = \frac{i}{N}$ and $\frac{-1}{2N} \leq \pi_n - \frac{i}{N} \leq \frac{1}{2N}$. Then $\mathbb{E}\left[\|\mathbf{z}^{(N)}(0)\|_2^2\right] < \frac{C}{4N}$. Hence $\limsup_{N\to\infty}\mathbb{E}\left[\|\mathbf{z}^{(N)}(0)\|_2^2\right] = 0$. The invariant bound follows by selecting the above $\mathbf{z}^{(N)}(0)$. This completes the proof.

E. Proof of Theorem 6

From Lemma 5 and Markov inequality, the sequence $\{\mathbf{z}^{(N)}(0)\}$ is tight. As a result, from Prohorov theorem, the sequence $\{\mathbf{z}^{(N)}(0)\}$ is relatively compact. Consider a converging subsequence and let $\mathbf{z}^{(\infty)}(0)$ be its limiting point and it is square integrable. Then from Theorem 3, the considered converging subsequence converges in law to the unique OU process $(\mathbf{z}^{(\infty)}(t), t \geq 0)$ satisfying the SDE (22) with the initial point $\mathbf{z}^{(\infty)}(0)$. But, we know from [17, Lemma 7.7 and Theorem 7.8, page 131] that the limit of a sequence of stationary processes is stationary. Hence the law of $(\mathbf{z}^{(\infty)}(t), t \geq 0)$ should be the unique law of the stationary OU process solving the SDE (22). This argument applies for every converging subsequence. Hence, the sequence $\{(\mathbf{z}^{(N)}(t), t \geq 0)\}$ in stationary converges to the unique stationary OU process solving the SDE (22). This completes the proof.

F. Proof of Theorem 7

Let $Y^{(N)}(\infty)$ be a random variable with the invariant law denoting the number of progressing jobs in the system. Then from the Little's law, we have

$$(N\lambda)(1 - P_{block}^{(N)}) = \mathbb{E}\left[Y^{(N)}(\infty)\right]. \tag{67}$$

Let $\mathbf{x}^{(N)}(\infty) = (x_i^{(N)}(\infty), 0 \leq i \leq C)$ be a random variable with the invariant law where $x_i^{(N)}$ denotes the fraction of servers with at least i progressing jobs. It can be seen that

$$Y^{(N)}(\infty) = \sum_{i=0}^{C} Ni(x_i^{(N)}(\infty) - x_{i+1}^{(N)}(\infty)) \tag{68}$$

Hence, we have

$$(N\lambda)(1 - P_{block}^{(N)}) = \mathbb{E}\left[\sum_{i=0}^{C} Ni(x_i^{(N)}(\infty) - x_{i+1}^{(N)}(\infty))\right]. \tag{69}$$

It implies that

$$\lambda(1 - P_{block}^{(N)}) = \sum_{i=0}^{C} i\mathbb{E}\left[x_i^{(N)}(\infty) - x_{i+1}^{(N)}(\infty)\right]. \tag{70}$$

Also, from Theorem 6, since the diffusion limit in stationary has zero mean, we have

$$\mathbb{E}\left[\mathbf{x}^{(N)}(\infty)\right] - \boldsymbol{\pi} = o(N^{-\frac{1}{2}}). \tag{71}$$

As a result, we have

$$\lambda(1 - P_{block}^{(N)}) = \sum_{i=0}^{C} i(\pi_i - \pi_{i+1}) + o(N^{-\frac{1}{2}}). \tag{72}$$

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

$$\sum_{i=0}^{C} i(\pi_i - \pi_{i+1}) = \lambda(1 - \pi_C^d). \tag{73}$$

By using the above equation, we get

$$P_{block}^{(N)} - \pi_C^d = o(N^{-\frac{1}{2}}). \tag{74}$$

This completes the proof.