Mean-field analysis of loss models with mixed-Erlang distributions under Power-of-d routing

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

Arpan Mukhopadhyay
EPFL IC/LCA2
Lausanne, Switzerland
Email: arpan.mukhopadhyay@epfl.ch

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

***Abstract*—In this paper, we study the behavior of a large number N of parallel loss servers operating under a randomized power-of-d routing scheme for the arrivals. Such models are of importance in several cloud architectures that offer IaaS (Infrastructure as a Service). The probabilistic behavior of such models has previously been analyzed for jobs with exponential holding times. However, in most realistic applications, the assumption of exponential holding times does not hold and therefore it is of importance to understand the performance of the power-of-d routing scheme under more general holding time distributions. In this paper, we analyze the dynamics of the system under mixed-Erlang service time distributions since any distribution on $[0,\infty)$ can be approximated by the mixed-Erlang distribution with arbitrary accuracy. We focus on the limiting regime when $N \to \infty$. This leads to a mean-field dynamics that are significantly more difficult to analyze than the exponential case since the state of each server is multi-dimensional with no monotonicity properties. In particular we show that that the mean-field equation (MFE) has a unique fixed-point that corresponds to the fixed-point obtained with exponential assumptions on the holding times showing that the fixed-point is insensitive to the parameters of the mixed-Erlang distribution and only depends on the mean. This has important implications for practical systems.**

I. Introduction

Cloud platforms such as Amazon's EC2 [1] and Microsoft's Azure [2] etc., that provide Infrastructure-as-a-Service (IaaS) are becoming increasingly popular due to their ability to provide scalable and cost effective computing solutions. In recent years, the problem of designing efficient, low-complexity load balancing schemes for such systems has drawn considerable attention. Randomized load balancing schemes have been shown to be promising solutions [3]–[5] .The performance of these randomized schemes has been analyzed by modeling such systems as multi-server systems consisting of a large number of finite-capacity loss servers working in parallel and the key performance criterion considered is the average blocking probability of jobs. In this context, the power-of-d scheme, where an arriving job is routed to the server with maximum vacancy among d randomly sampled servers, is known to perform nearly optimally [3], [4], [6].

In analyzing the performance of randomized load balancing schemes for loss systems, one key assumption in the previous works was that the job holding times/service times are exponentially distributed. This simplifies the analysis considerably by allowing construction of a Markov process that keeps track of only the number of jobs at each server. In practice, however, job holding time distributions may have different tail characteristics [7], [8]. Statistical studies suggest that depending on the application the service times of jobs may have Log-Normal [9], or Gamma [10], or shifted exponential distributions [11]. Hence, it is important to investigate the implications of service time distributions on the performance of the power-of-d routing policy and the robustness in the statistical behavior. Motivated by this issue, in this paper we consider the service time distribution of the arriving jobs to be mixed-Erlang since any general distribution on $[0,\infty]$ can be approximated arbitrarily closely by mixed-Erlang distributions [12].

Unlike in the exponential service time case, to model the dynamics of the system as a Markov process, one has to also keep track of the progress of each ongoing job at each server. As a result, the analysis becomes considerably more challenging since the Markov process is now constructed on a multi-dimensional state space for each server.

Analyzing the multidimensional Markov chain is completely intractable. In this paper we focus on the regime where the number N of servers in the system approaches infinity along with external arrival rate ($N\lambda$) of the jobs. In such a regime, the Markov process $\mathbf{x}^N(t)$ describing the dynamics of servers will converge weakly to a deterministic limit $\mathbf{x}(t)$ as $N \to \infty$. The limiting process is known as the *mean-field limit* or simply the *mean-field* of the system and the governing equations are referred to as the mean-field equations (MFEs). The focus of this paper is on the equilibrium behavior of the mean-field.

In particular, we study the following:

- Existence of fixed-points for the mean-field.
- Uniqueness of the fixed point.
- The relation of the fixed point to the fixed point in the case of exponentially distributed holding times. Equivalence implies insensitivity.

The main contributions and observations of the paper are: 1) We establish the limit theorem for the convergence of the Markov process $\mathbf{x}^N(t)$[1]. This is an extension of the ODE method for the Markov processes with 1-dimensional underlying space established in [3], [4] to the Markov process with

[1]A preliminary version of this result was presented at ECQT 2016, Toulouse, France.

978-0-9883-0453-6/17 $31.00 © 2017 ITC
DOI 10.1109/ITC.2017.16

multi-dimensional countable underlying space. This follows routinely from the exponential case. However, the real challenge is the characterization of the equilibrium behavior due to the multi-dimensional nature of the underlying state-space for each queue. 2) We then show that the MFE has a unique equilibrium point, and finally 3) we show that the equilibrium point coincides with the equilibrium point in the exponential case establishing the insensitivity of the equilibrium of the MFE.

Establishing the fact that the equilibrium point corresponds to the stationary distribution at any queue is much harder since it is very difficult to establish global asymptotic stability (GAS) of the equilibrium. Extensive numerical studies show that the GAS property indeed holds but we can only conjecture that at this point. The difficulty is the lack of any monotonicity for the multi-dimensional Markov processes that represent the state at each server. It is important to note that the loss system under consideration is not insensitive to service time distributions for finite values of N as result of the dependencies among the servers in the power-of-d routing scheme. Hence, the insensitivity property holds only in the limiting regime. Our numerical results, however, show that even for finite N the system is nearly insensitive to service time distributions as was observed for processor sharing systems [13] (that are known to be insensitive too when inputs are Poisson) under join-the-shortest-queue policy.

The rest of the paper is organized as follows: We first discuss the related literature in Section II. The Section III describes the system model and the power-of-d policy. We then give the main results in Section IV. Proofs for the results are given in Section V. The numerical results are given in Section VI. Finally, we conclude in Section VII.

II. Related Literature

Although randomized routing schemes for queuing systems with exponential service time distributions were first studied two decades ago by using mean-field techniques, the complete analysis for queuing systems with general service time distributions is still an open question. We briefly discuss the related literature below.

The randomized routing schemes were first investigated in [14] in the context of bin packing models. The power-of-d scheme was considered for FCFS queues with exponential service time distributions in [15], [16]. The results were then extended to processor sharing (PS) queues in [17], [18] when service times are exponentially distributed. In [7], randomized routing schemes for queuing systems with general service time distributions when service disciplines are FCFS, PS, and LIFO were studied. The steady-state results were characterized assuming the asymptotic independence of servers. The asymptotic independence of servers for FCFS systems is established in [19] for the case when service time distributions have decreasing hazard rate functions. Recently the mean-field limit for FCFS systems with general service time distributions was established in [8] by considering age process of jobs that gives rise to a Markov model of the system. However, the steady-state behavior was not investigated.

Multi-server loss models under randomized routing schemes were first studied by Turner [20], [21] when job lengths are exponentially distributed by using mean-field techniques. The MFEs were used to characterize the limiting system and the resulting tail distribution of server occupancies was observed to decay rapidly even when there is a small number of routing choices for each arriving job. However, the existence and uniqueness of the fixed-point of the mean-field were not shown. In [3], the existence and uniqueness of the fixed-point of the mean-field for homogeneous loss model of Turner [20] was addressed. These results were extended to the heterogeneous case in [3] by assuming the asymptotic independence of servers. The asymptotic independence property of loss models had been studied earlier by [22], [23] in the context of alternative routing where the stronger property of propagation of chaos (independence) on the path space was shown.

The complete analysis for heterogeneous loss models under the power-of-d routing scheme was shown in [4]. They showed the existence and uniqueness of the stationary point of the mean-field, as well as the global asymptotic stability of the mean-field. Further, asymptotic independence of the distributions for every t was shown. The results were then extended to multi-class heterogeneous loss models in [6] where jobs belong to one of the several classes based on the amount of resources they use. All these works were based on the assumption that the job lengths are exponentially distributed. The main focus of this paper is to study large multi-server loss models with non-exponential service time distributions and see whether asymptotic insensitivity (insensitivity of the limiting system) holds for randomized power-of-d policies.

III. System Model

We consider a system of N parallel servers classified into K types with index set $\mathcal{K} = \{1, 2, \dots, K\}$ and a server of type k has the capacity to provide service to C_k jobs simultaneously. Without loss of generality we assume $C_1 \leq C_2 \dots \leq C_K$. The fraction of servers of type k out of N servers is fixed at β_k such that $\sum_{k=1}^{K} \beta_k = 1$. Jobs arrive according to Poisson process with rate $N\lambda$.

We assume that the service times have mixed-Erlang distribution with finite mean $\frac{1}{\mu}$ where each job length is drawn from the Erlang distribution with i exponential phases with probability p_i, for $1 \leq i \leq M$ such that $\sum_{i=1}^{M} p_i = 1$. Each exponential phase is assumed to have rate μ_p. Therefore, we have,

$$\frac{1}{\mu} = \frac{\sum_{i=1}^{M} i p_i}{\mu_p}. \tag{1}$$

Let $\mathbb{Z}_+$ denote the set of non-negative integers. We assume that a job that is in progress at a server always has nonzero number of phases that it needs to finish to exit the server. At time t, state of a type k server with n jobs is denoted by $\underline{l} = (l_1, \dots, l_{C_k})$ where l_i is the number of remaining phases for i^{th} job, and further we have $l_m > 0$ for $m \leq n$ and $l_m = 0$

for $m > n$, and we say that the server has occupancy n and vacancy $C_k - n$. We denote the occupancy of a server of type k with state $\underline{l} = (l_1, \ldots, l_{C_k})$ which is the number of nonzero elements in $\underline{l}$ by $Z(\underline{l})$. For $k \in \mathcal{K}$, we define a set S_k which contains all possible server states for a type k server. We define

$$\begin{aligned} S_k = \{(b_1, b_2, \ldots, b_{C_k}) : \exists\, j \text{ such that} \\ b_i \in \{1, 2, \cdots, M\} \text{ for } i \leq j, b_i = 0 \text{ for } i > j\}\,. \end{aligned} \quad (2)$$

For $k \in \mathcal{K}$, we denote an element $(b_1, b_2, \ldots, b_{C_k}) \in S_k$ by $\underline{b} = (b_1, b_2, \ldots, b_{C_k})$. Size of the vector $\underline{b}$ is indicated by the type of the server. Except the elements in S_k for $k \in \mathcal{K}$, all other vectors are denoted by bold letters

An arriving job is routed to a server chosen according to the following randomized scheme.

Power-of-d scheme: When a job arrives, $d \geq 2$ servers are sampled uniformly at random from N servers with replacement[2]. These sampled servers are referred to as the potential destination servers. The job is then routed to the server with maximum vacancy among the potential destination servers. Ties among servers of the same type are broken by choosing a server uniformly at random and ties across different types are broken by choosing a server of the type with the highest index (highest capacity). The server to which the job is routed is referred to as the destination server. If the destination server of a job has no vacancy, then the job is dropped otherwise the job is accepted for service at the destination server.

IV. Main Results

In this section, we state the main results without proofs.

Notation: In order to model the system as Markov process, we consider a process $\mathbf{x}^N(t) = \left(x_{\underline{l},k}{}^N(t), \underline{l} \in S_k, k \in \mathcal{K}\right)$ where $x^N_{\underline{l},k}(t)$ denotes the fraction of type k servers (out of $N\beta_k$ servers) having state $\underline{l}$ at time t. Precisely, if $q^N_{i,k}(t)$ denotes the state of i^{th} server of type k at time t, then we have

$$x^N_{\underline{l},k}(t) = \frac{\sum_{i=1}^{N\beta_k} I_{\left\{q^N_{i,k}(t) = \underline{l}\right\}}}{N\beta_k}. \quad (3)$$

We next define the space of probability measures $\mathcal{V}^{(N)}, \mathcal{V}$ which contain the states of the Markov process $\mathbf{x}^N(\cdot)$ and the mean-field $\mathbf{x}(\cdot)$, respectively. For $k \in \mathcal{K}$,

$$\mathcal{V}_k = \left\{(b_{\underline{l}})_{\underline{l} \in S_k} : b_{\underline{l}} \geq 0, \forall\, \underline{l} \in S_k, \text{ and } \sum_{\underline{l} \in S_k} b_{\underline{l}} = 1\right\} \quad (4)$$

$$\mathcal{V}^{(N)}_k = \left\{(b_{\underline{l}})_{\underline{l} \in S_k} \in \mathcal{V}_k : N b_{\underline{l}} \in \mathbb{Z}_+, \forall\, \underline{l} \in S_k\right\}. \quad (5)$$

We are interested in the spaces that are Cartesian products of the spaces $\mathcal{V}_k$, $\mathcal{V}^{(N)}_k$ over $k \in \mathcal{K}$ denoted as $\mathcal{V} = \prod_{k \in \mathcal{K}} \mathcal{V}_k$ and $\mathcal{V}^{(N)} = \prod_{k \in \mathcal{K}} \mathcal{V}^{(N)}_k$, respectively. We say an element $\mathbf{u} = (u_{\underline{l},k}, \underline{l} \in S_k, k \in \mathcal{K})$ is in $\mathcal{V}$ or $\mathcal{V}^{(N)}$ if the sequence $\{u_{\underline{l},k}, \underline{l} \in S_k\}$ is in $\mathcal{V}_k$ or $\mathcal{V}^{(N)}_k$, respectively for each $k \in \mathcal{K}$.

To distinguish between the state of the process $\mathbf{x}^N(t)$ at time t and the state of a server at time t, we call the former as the process state and the latter as the server state from now onwards.

Let $R(\mathbf{u} \to \mathbf{w})$ denote the transition rate of the Markov process $\mathbf{x}^N(\cdot)$ from the process state $\mathbf{u} \in \mathcal{V}^{(N)}$ to the process state $\mathbf{w} \in \mathcal{V}^{(N)}$. Then the generator B^N of the Markov process $\mathbf{x}^N(\cdot)$ acting on function $f : \mathcal{V}^{(N)} \to \mathbb{R}$ is given by

$$B^N f(\mathbf{u}) = \sum_{\mathbf{w} \neq \mathbf{u}} R(\mathbf{u} \to \mathbf{w}) \left(f(\mathbf{w}) - f(\mathbf{u})\right). \quad (6)$$

Let the set of states of a type k server with occupancy at least n be denoted by $D_{n,k}$. Further, we define for $\mathbf{u} \in \mathcal{V}$

$$F_{n,k}(\boldsymbol{u}) = \sum_{\underline{l} \in D_{n,k}} u_{\underline{l},k}. \quad (7)$$

Further, for $n < C_k$, $k \in \mathcal{K}$ and $\mathbf{u} \in \mathcal{V}$, we define a function $\lambda^{(ME)}_{n,k}(\mathbf{u})$ to represent the arrival rate of jobs to a type k server that has n jobs when system state is $\mathbf{u}$ as follows,

$$\begin{aligned} \lambda^{(ME)}_{n,k}(\mathbf{u}) &= \frac{\lambda}{\beta_k \left(F_{n,k}(\mathbf{u}) - F_{n+1,k}(\mathbf{u})\right)} \\ &\times \left[\left(\sum_{i=1}^{k} \beta_i F_{C_i - C_k + n, i}(\mathbf{u}) + \sum_{i=k+1}^{K} \beta_i F_{C_i - C_k + n+1, i}(\mathbf{u})\right)^d \right. \\ &\left. - \left(\sum_{i=1}^{k-1} \beta_i F_{C_i - C_k + n, i}(\mathbf{u}) + \sum_{i=k}^{K} \beta_i F_{C_i - C_k + n+1, i}(\mathbf{u})\right)^d\right]. \end{aligned} \quad (8)$$

We define a metric $\tau(\mathbf{u}, \mathbf{w})$ for any $\mathbf{u}, \mathbf{w} \in \mathcal{V}$,

$$\tau(\mathbf{u}, \mathbf{w}) = \sum_{k \in \mathcal{K}} \sum_{\underline{l} \in S_k} \left| u_{\underline{l},k} - w_{\underline{l},k} \right|. \quad (9)$$

The space $\mathcal{V}$ is compact with respect to metric τ as it is closed and bounded subset of finite dimensional space.

For a measure space $(U, \mathcal{U}, \nu)$, and a ν-integrable function f, we define $\langle f, \nu \rangle = \int f d\nu$. Law of a random variable X is denoted by $\mathcal{L}(X)$. Then the weak convergence(convergence in distribution) of a sequence of probability measures ν_n (random variable X_n) to a probability measure ν (random variable X) is denoted by $\nu_n \Rightarrow \nu$ ($X_n \Rightarrow X$).

We state our first main result on establishing the mean-field limit. Using the expression for the generator B^N established in Lemma 1 of the Appendix, we show that the process $\mathbf{x}^N(\cdot)$ converges weakly to a deterministic process as $N \to \infty$ by showing the convergence of generators of Markov processes and the approach is same as that of [16]. We establish that if $\mathbf{x}^N(0)$ converges in distribution to $\mathbf{u}$, then $\mathbf{x}^N(\cdot)$ converges to mean-field limit $\mathbf{x}(\cdot, \mathbf{u})$. Here $\mathbf{u}$ indicates the initial point of the mean-field limit.

Theorem *1:* If $\mathbf{x}^N(0)$ converges in distribution to a state $\mathbf{u} \in \mathcal{V}$, then the process $\mathbf{x}^N(\cdot)$ converges in distribution to a

[2]The asymptotic results, i.e., the MFE remains unchanged if sampled without replacement. However notationally the case we consider is more convenient to work with.

deterministic process $\mathbf{x}(\cdot,\mathbf{u})$ as $N\to\infty$ called the mean-field and it lies in the space $\mathcal{V}$. The process $\mathbf{x}(\cdot,\mathbf{u})$ is the unique solution of the following system of differential equations.

$$\mathbf{x}(0,\mathbf{u}) = \mathbf{u}, \tag{10}$$

$$\dot{x}_{\underline{l},k}(t,\mathbf{u}) = h_{\underline{l},k}(\mathbf{x}(t,\mathbf{u})), \tag{11}$$

and $\mathbf{h} = (h_{\underline{l},k}, \underline{l}\in S_k, k\in\mathcal{K})$ with the mapping $h_{\underline{l},k}:\mathcal{U}\to\mathbb{R}$ given by

$$\begin{aligned}h_{\underline{l},k}(\mathbf{x}) &= \sum_{b=1}^{Z(\underline{l})}\left(\frac{p_{l_b}}{Z(\underline{l})}\right)x_{(l_1,l_2,\ldots,l_{b-1},l_{b+1},\ldots,l_{C_k},0,k)}\\ &\times \lambda^{(ME)}_{Z(\underline{l})-1,k}(\mathbf{x}) - x_{\underline{l},k}\lambda^{(ME)}_{Z(\underline{l}),k}(\mathbf{x})I_{\{Z(\underline{l})<C_k\}}\\ &+\sum_{b=1}^{Z(\underline{l})+1}\mu_p I_{\{Z(\underline{l})<C_k\}}x_{(l_1,\ldots,l_{b-1},1,l_b,\ldots,l_{C_k-1},k)}\\ &+\sum_{b=1}^{Z(\underline{l})}\mu_p x_{(l_1,\ldots,l_{b-1},l_b+1,l_{b+1},\ldots,l_{C_k},k)} - Z(\underline{l})\mu_p x_{\underline{l},k}\end{aligned} \tag{12}$$

Remark *1:* To establish the mean-field limit, one can also use the theory developed for density-dependent Markov processes by Ethier and Kurtz [24, Chapter 11], or by martingale techniques as in [25]. Mitzenmacher used the Kurtz approach to establish the mean-field limit for supermarket models (parallel $./M/1$ servers) in [15] and while martingale techniques are used by Turner in [21]. Our approach is similar to the one used in [26].

In order to understand the uniqueness proof we compare the MFE (12) with the dynamics of the server occupancy in a loss system with a single server having the prespecified set of arrival rates where the arrival rate depends on the number of jobs that are in progress at the server.

Consider an Erlang loss system with single server of type k where jobs arrive according to a Poisson process with pre-specified state-dependent arrival rate $\eta_{j,k}$ when there are j jobs in progress at the server. We assume that service times of jobs are distributed according to mixed-Erlang distribution as in the system model. Let $w_{\underline{l},k}(t,\mathbf{v})$ denotes the probability that the server has state $\underline{l}$ at time t when initial distribution for server occupancies is given by $\mathbf{v}$. Then it can be checked that the dynamics of the probabilities for server occupancies $\mathbf{w}_k = (w_{\underline{l},k}, \underline{l}\in S_k)$ is given by

$$\mathbf{w}_k(0,\mathbf{v}) = \mathbf{v}, \tag{13}$$

For $\underline{l}\in S_k$, we have

$$\begin{aligned}\dot{w}_{\underline{l},k}(t,\mathbf{v}) &= \sum_{b=1}^{Z(\underline{l})}\left(\frac{p_{l_b}}{Z(\underline{l})}\right)w_{(l_1,l_2,\ldots,l_{b-1},l_{b+1},\ldots,l_{C_k},0,k)}(t,\mathbf{v})\\ &\times \eta_{Z(\underline{l})-1,k} - w_{\underline{l},k}(t,\mathbf{v})\eta_{Z(\underline{l}),k}I_{\{n<C_k\}}\\ &+\sum_{b=1}^{Z(\underline{l})+1}\mu_p I_{\left\{l_{C_k}=0\right\}}w_{(l_1,\ldots,l_{b-1},1,l_b,\ldots,l_{C_k-1},k)}(t,\mathbf{v})\\ &+\sum_{b=1}^{Z(\underline{l})}\mu_p w_{(l_1,\ldots,l_{b-1},l_b+1,l_{b+1},\ldots,l_{C_k},k)}(t,\mathbf{v}) - Z(\underline{l})\mu_p w_{\underline{l},k}(t,\mathbf{v}).\end{aligned} \tag{14}$$

Note that equation (14) represents the dynamics of a state-dependent linear Markov process whereas equation (12) represents the dynamics of a non-linear Markov process whose generator depends on the distribution of the state. It is clear that equations (12) and (14) differ only in the arrival rates (one being dependent only on the number of jobs while the second depends on the complete state of the system). The Markov process that represents the system in equation-(14) was shown to be ergodic in [27] and has unique steady-state distribution $\boldsymbol{\pi}_k^{(single)} = (\pi_{\underline{l},k}^{(single)}, \underline{l}\in S_k)$ given by, for $Z(\underline{l}) = n$, $\pi_{\underline{l},k}^{(single)}$ satisfies

$$\pi_{\underline{l},k}^{(single)} = m_{n,k}\prod_{i=1}^{n} a_{l_i}, \tag{15}$$

where

$$m_{n,k} = \left(\prod_{i=1}^{n}\frac{\eta_{i-1,k}}{i\mu}\right)m_{0,k}, \tag{16}$$

$$m_{0,k} = \left(1+\sum_{n=1}^{C_k}\left(\prod_{i=1}^{n}\frac{\eta_{i-1,k}}{i\mu}\right)\right)^{-1}. \tag{17}$$

By exploiting the analogy between equation (12) and equation (14), we characterize the equilibrium point $\boldsymbol{\pi}$ of the mean-field $\mathbf{x}(\cdot,\mathbf{u})$ which satisfies $\mathbf{h}(\boldsymbol{\pi}) = 0$. The proofs are given in the next section.

Theorem *2:* The mean-field $\mathbf{x}(\cdot,\mathbf{u})$ has an equilibrium point in the space $\mathcal{V}$.

We now state the main result on uniqueness of the fixed-point for the mean-field $\mathbf{x}(\cdot,\mathbf{u})$. When job lengths are exponentially distributed with mean $\frac{1}{\mu}$, it was shown in [4] that there exists unique fixed-point for the mean-field $\mathbf{y}(t)$ that is a limit of the sequence of processes $\mathbf{y}^N(t) = (y^N_{l,j}, 0\le l\le C_j)$ where $y^N_{l,j}(t)$ denotes the fraction of type j servers having l jobs at time t. Further, the unique fixed-point of the mean-field also represents the steady-state distribution for server occupancies of the limiting system (system with $N\to\infty$). Let $\boldsymbol{\pi}_k^{(exp)} = (\pi_{n,k}^{(exp)}, 0\le n\le C_k)$ denotes the unique steady-state distribution for the occupancy of a type k server and $\pi_{n,k}^{(exp)}$ is the probability that a type k server has n jobs in the limiting system in the exponential case with average job length $\frac{1}{\mu}$.

Theorem *3:* The mean-field $\mathbf{x}(\cdot,\mathbf{u})$ has unique fixed-point $\boldsymbol{\pi} = (\pi_{\underline{l},k}, \underline{l}\in S_k, k\in\mathcal{K})$ where for $\underline{l} = (l_1,l_2,\ldots,l_n,0,\ldots,0)\in S_k$ and $Z(\underline{l}) = n$, we have

$$\pi_{\underline{l},k} = \pi_{n,k}^{(exp)}\prod_{j=1}^{n} a_{l_j} \tag{18}$$

with $a_m = \frac{\sum_{j=m}^{M} p_j}{\sum_{j=1}^{M} jp_j}$. Further, the fixed-point is insensitive since we have, for $k\in\mathcal{K}$, $1\le n\le C_k$

$$\sum_{\underline{l}\in S_k:Z(\underline{l})=n}\pi_{\underline{l},k} = \pi_{n,k}^{(exp)}. \tag{19}$$

Remark *2:* Once the mean-field limit is established, we can establish the asymptotic independence of servers for any finite time $t \geq 0$ as in [6] and is hence omitted.

Remark *3:* Establishing the GAS of the fixed-point $\boldsymbol{\pi}$ would imply that $\boldsymbol{\pi}$, in fact, represents the stationary distribution of the limiting system and the proof is same as that of the exponential case established in [4]. This would show insensitivity of the stationary distribution of the limiting system. However, since the mean-field $\mathbf{x}(\cdot, \mathbf{u})$ is not quasi-monotonic which was the key to proving GAS of the fixed-point in the exponential case, establishing the GAS of $\boldsymbol{\pi}$ is a challenging task. Numerical evidence shows this is true but a proof is as yet not available. If indeed GAS can be shown then one can establish the asymptotic independence of the servers even in the equilibrium as in [6].

Remark *4:* In a finite N system, since arriving jobs are routed to a server based on the server states of d randomly chosen servers, the actual arrival process into each server is not a Poisson process. Therefore the stationary distribution corresponding to server occupancies is not insensitive to the service time distribution in a finite N system. However, simulations in earlier works [3], [4] showed that the stationary distribution of the limiting system when $N \to \infty$ is insensitive to the service time distribution. This is referred to as the asymptotic insensitivity of the system. Although insensitivity has been shown to hold for state-dependent Erlang models, they correspond to linear Markov models, while in this case the process is a non-linear Markov process and thus a proof of insensitivity is needed that we will show next.

V. Proofs

In this section, we provide proofs for the results stated in Section IV.

Proof of Theorem 1

We first begin with the convergence to the mean field. We only provide the outlines of the proof. The crux of the proof is to first establish convergence of semi-groups of the Markov processes $\mathbf{x}^N(\cdot)$ as $N \to \infty$ and then by using [24, Theorem 2.11, p.172], conclude convergence of Markov processes $\mathbf{x}^N(\cdot)$ as $N \to \infty$. First it is important to note that if $\mathbf{u} \in \mathcal{V}$, then $\mathbf{x}(\cdot, \mathbf{u})$ always lies in $\mathcal{V}$. This follows from $\sum_{\underline{l} \in S_j} h_{\underline{l},j} = 0$ implying that $\sum_{\underline{l} \in S_j} x_{\underline{l},j}(t, \mathbf{u})$ always remains at one for $j \in \mathcal{K}$. If $x_{\underline{l},j}(t, \mathbf{u})$ is zero, then the equation (12) implies $\dot{x}_{\underline{l},j}(t, \mathbf{u}) \geq 0$. On the other hand, if $x_{\underline{l},j}(t, \mathbf{u})$ is equal to one, then we have $\dot{x}_{\underline{l},j}(t, \mathbf{u}) \leq 0$. Furthermore, $\mathbf{h}$ is a continuous mapping and therefore if $\mathbf{u} \in \mathcal{V}$ then $\mathbf{x}(\cdot, \mathbf{u})$ always lies in the space $\mathcal{V}$.

The next step is to show the uniqueness of solutions to MFEs. To show this we use Picard iteration as in [26]. For the iteration technique to converge it is sufficient to prove that the mapping $\mathbf{h}$ is a Lipschitz continuous. Let $|S_k|$ denotes the number of elements in the set S_k and $|S| = \sum_{k=1}^{K} |S_k|$. Then for $\mathbf{u} \in \mathcal{V}$, $\mathbf{v} \in \mathcal{V}$, we can find a constant $H = 2d^2\lambda |S| + 3\mu_p C_K$ such that $\tau(\mathbf{h}(\mathbf{u}), \mathbf{h}(\mathbf{v})) \leq H\tau(\mathbf{u}, \mathbf{v})$. Therefore for each initial point $\mathbf{u} \in \mathcal{V}$, there exists a unique process $\mathbf{x}(\cdot, \mathbf{u})$ satisfying the MFEs (10)-(12).

Furthermore, if $\mathbf{x}(0, \mathbf{u}) = \mathbf{u}$, we need to show that the solution $\mathbf{x}(t, \mathbf{u})$ is continuous w.r.t. initial point $\mathbf{u}$. Since the underlying space for the Markov process is finite dimensional, in the same way as in [26, Lemma 3.1], we can show that $\frac{\partial x_{\underline{l},k}(t,\mathbf{u})}{\partial u_{\underline{n},j}}$ exists which implies $\mathbf{x}(t, \mathbf{u})$ is continuous w.r.t. initial condition $\mathbf{u}$ on every compact interval $[0, t]$. Similarly, for $k \in \mathcal{K}$, $j_1, j_2 \in \mathcal{K}$ and $\underline{l} \in S_k$, $\underline{n} \in S_{j_1}$, $\underline{n'} \in S_{j_2}$ and [26, Lemma 3.1], it can be shown that $\frac{\partial^2 x_{\underline{l},k}(t,\mathbf{u})}{\partial^2 u_{\underline{n},j_1}}$, $\frac{\partial^2 x_{\underline{l},k}(t,\mathbf{u})}{\partial u_{\underline{n},j_1} \partial u_{\underline{n'},j_2}}$ also exist and are bounded.

Finally, we show $\mathbf{x}^N(\cdot)$ converges weakly to $\mathbf{x}(\cdot, \mathbf{u})$ by using convergence of semi group operators of Markov processes. The semi groups of operators $(\mathbf{T}(t), t \geq 0)$, $(\mathbf{T}_N(t), t \geq 0)$ of the processes $\mathbf{x}(\cdot)$, $\mathbf{x}^N(\cdot)$, respectively are given by

$$\mathbf{T}(t) f(\mathbf{u}) = f(\mathbf{x}(t, \mathbf{u})), \tag{20}$$

$$\mathbf{T}_N(t) f(\mathbf{u}) = \mathbb{E}\left[f(\mathbf{x}(t, \mathbf{u})) | (\mathbf{x}^N(0) = \mathbf{u})\right]. \tag{21}$$

Further, the generators for the semigroups $\mathbf{T}_N(t)$, $\mathbf{T}(t)$ are $B^N f(\mathbf{u})$, $Bf(\mathbf{u})$, respectively are $B^N f(\mathbf{u})$ given in (46) and

$$Bf(\mathbf{u}) = \frac{df(\mathbf{x}(t, \mathbf{u}))}{dt} | \{t = 0\}. \tag{22}$$

Let L be the set of bounded continuous functions $f : \mathcal{V} \to \mathcal{R}$ and D be the set of $f \in L$ for which $\frac{\partial \mathbf{f}(\mathbf{u})}{\partial u_{\underline{n},j}}$, $\frac{\partial^2 \mathbf{f}(\mathbf{u})}{\partial^2 u_{\underline{n},j}}$ and $\frac{\partial^2 \mathbf{f}(\mathbf{u})}{\partial u_{\underline{n},j} \partial u_{\underline{n'},j'}}$ exist and are uniformly bounded for any $j, j' \in \mathcal{K}$, $\underline{n} \in S_j$ and $\underline{n'} \in S_{j'}$. As in proof of Theorem 2 in [26], by observing that $B^N f \to Bf$ as $N \to \infty$ in sup norm for all $f \in D$, we have

$$\lim_{N \to \infty} \sup_{\mathbf{u} \in \mathcal{V}^{(N)}} |T_N(t) g(\mathbf{u}) - T(t) g(\mathbf{u})| = 0, \tag{23}$$

for every $g \in L$. Also $T(t)$ is a Feller semigroup because $T(t)\mathbb{I} = \mathbb{I}$ where $\mathbb{I}$ is a indicator function on $\mathcal{V}$ and $\mathbf{x}(t, \mathbf{u})$ is continuous w.r.t. $\mathbf{u}$. Hence, by using [24, Theorem 2.11, p.172], we conclude that if $\mathbf{x}^N(0) \Rightarrow \mathbf{u}$ as $N \to \infty$, then $\mathbf{x}^N(\cdot) \Rightarrow \mathbf{x}(\cdot, \mathbf{u})$ as $N \to \infty$. ■

Proof of Theorem 2

We first construct a continuous mapping in such a way that any fixed-point of the mapping also satisfies the steady-state MFEs. The existence of a fixed-point for the constructed mapping is guaranteed by the Brouwer's fixed-point theorem. This implies the existence of an equilibrium point for the mean-field in the space $\mathcal{V}$.

For $n < C_k$, $k \in \mathcal{K}$ and $\mathbf{u} \in \mathcal{V}$, we define $\zeta_{n,k}(\mathbf{u})$ which follows the recursive equations

$$\lambda_{n,k}^{(ME)}(\mathbf{u}) \zeta_{n,k}(\mathbf{u}) = \zeta_{n+1,k}(\mathbf{u})(n+1)\mu \tag{24}$$

such that $\sum_{n=0}^{C_k} \zeta_{n,k}(\mathbf{u}) = 1$. Then we have

$$\zeta_{0,k}(\mathbf{u}) = \left(1 + \sum_{n=1}^{C_k} \left(\prod_{i=1}^{n} \frac{\lambda_{i-1,k}^{(ME)}(\mathbf{u})}{i\mu}\right)\right)^{-1}, \tag{25}$$

$$\zeta_{n,k}(\mathbf{u}) = \left(\prod_{i=1}^{n} \frac{\lambda^{(ME)}_{i-1,k}(\mathbf{u})}{i\mu}\right)\zeta_{0,k}(\mathbf{u}). \tag{26}$$

Finally, for $i \in \{1, 2, \cdots, M\}$, let

$$a_i = \frac{\sum_{j=i}^{M} p_j}{\sum_{j=1}^{M} jp_j}. \tag{27}$$

Then we have $\sum_{i=1}^{M} a_i = 1$. For $\underline{l} = (l_1, \ldots, l_n, 0, \ldots, 0) \in S_k$ and $k \in \mathcal{K}$ with $Z(\underline{l}) = n$, let

$$\Gamma_{\underline{l},k}(\mathbf{u}) = \zeta_{n,k}(\mathbf{u}) \prod_{j=1}^{n} a_{l_j}. \tag{28}$$

We have $\mathbf{\Gamma}(\mathbf{u}) = (\Gamma_{\underline{l},k}(\mathbf{u}), \underline{l} \in S_k, k \in \mathcal{K}) \in \mathcal{V}$. Also the mapping $\mathbf{u} \to \mathbf{\Gamma}(\mathbf{u})$ is continuous over $\mathcal{V}$. Since $\mathcal{V}$ is compact, by Brouwer's fixed-point theorem there exists a fixed-point $\boldsymbol{\pi}^*$ in the space $\mathcal{V}$. Further, any fixed-point $\boldsymbol{\pi}^*$ of the mapping $\mathbf{u} \to \mathbf{\Gamma}(\mathbf{u})$ also satisfies $\mathbf{h}(\boldsymbol{\pi}^*) = 0$, where $\mathbf{h}$ is defined in (12). Hence, there exists an equilibrium point for the mean-field $\mathbf{x}(\cdot)$ in the space $\mathcal{V}$. ■

A. Proof of Theorem 3

From Theorem 2 there exists an equilibrium point for the mean-field. We now show that the fixed point is unique. We first recall the following two results that we use in the proof of uniqueness of the fixed-point. The Erlang loss system with single type k server having prespecified state-dependent arrival rates with dynamics as in equation (14) has a unique steady-state distribution.

Further, from [4], the mean-field limit for multi-server Erlang loss systems with the exponential service times under the power-of-d routing policy has unique fixed-point denoted by $\boldsymbol{\pi}^{(exp)} = (\pi^{(exp)}_{n,k}, 0 \le n \le C_k, k \in \mathcal{K})$. In exponential case, a type k server lies in state j, $0 \le j \le C_k$ if there are j jobs in progress at the server. It was also shown that the unique fixed-point $\boldsymbol{\pi}^{(exp)}$ of the mean-field is also the unique fixed-point of the mapping

$$\psi : \mathcal{U} \mapsto \mathcal{U} \tag{29}$$

where $\mathcal{U} = \{\mathbf{v} : (v_{j,k}, 0 \le j \le C_k, k \in \mathcal{K}) : v_{j,k} \ge 0, \sum_{j=0}^{C_k} v_{j,k} = 1, \forall k \in \mathcal{K}\}$ and the mapping ψ is defined as follows. For given $\mathbf{v} \in \mathcal{U}$, we first define the mapping Λ corresponding to the arrival rates as follows: $\lambda^{(exp)}_{C_k,k}(\mathbf{v}) = 0$ and for $n < C_k$,

$$\begin{aligned}\lambda^{(exp)}_{n,k}(\mathbf{v}) &= \frac{\lambda}{\beta_k\left(R_{n,k}(\mathbf{v}) - R_{n+1,k}(\mathbf{v})\right)} \\ &\times \left[\left(\sum_{i=1}^{k} \beta_i R_{C_i-C_k+n,i}(\mathbf{v}) + \sum_{i=k+1}^{K} \beta_i R_{C_i-C_k+n+1,i}(\mathbf{v})\right)^d \right. \\ &\left. - \left(\sum_{i=1}^{k-1} \beta_i R_{C_i-C_k+n,i}(\mathbf{v}) + \sum_{i=k}^{K} \beta_i R_{C_i-C_k+n+1,i}(\mathbf{v})\right)^d\right],\end{aligned} \tag{30}$$

where $R_{n,k}(\mathbf{v}) = \sum_{j=n}^{C_k} v_{j,k}$ and $v_{j,k} \ge 0$ such that $\sum_{j=0}^{C_k} v_{j,k} = 1$ for all $k \in \mathcal{K}$. For given set of arrival rates $\boldsymbol{\lambda}^{(exp)} = (\lambda^{(exp)}_{n,k}, 0 \le n \le C_k, k \in \mathcal{K})$, we define a mapping $\mathcal{H} : \boldsymbol{\lambda}^{(exp)} \mapsto \mathcal{U}$ that outputs an element $\mathbf{b} \in \mathcal{U}$ as follows

$$\lambda^{(exp)}_{n,k} b_{n,k} = (n+1)\mu b_{n+1,k} \tag{31}$$

such that $\sum_{j=0}^{C_k} b_{j,k} = 1$ for all $k \in \mathcal{K}$. Then the mapping ψ is defined as $\psi(\mathbf{v}) = \mathcal{H}(\boldsymbol{\lambda}^{(exp)}(\mathbf{v}))$.

We first show that any fixed-point $\boldsymbol{\phi} = (\phi_{\underline{l},k}, \underline{l} \in S_k, k \in \mathcal{K})$ of the mean-field must satisfy the following relation. For all $k \in \mathcal{K}$, for all $\underline{l} \in S_k$ with $Z(\underline{l}) = n$, $n \ge 0$ and let $\boldsymbol{\lambda}^{(ME)}_k(\boldsymbol{\phi}) = (\lambda^{(ME)}_{n,k}(\boldsymbol{\phi}), 0 \le n \le C_k)$, then

$$\phi_{\underline{l},k} = G_{n,k}(\boldsymbol{\lambda}^{(ME)}_k(\boldsymbol{\phi})) \prod_{i=1}^{n} a_{l_i} \tag{32}$$

where

$$G_{n,k}(\boldsymbol{\lambda}^{(ME)}_k(\boldsymbol{\phi})) = \left(\prod_{i=1}^{n} \frac{\lambda^{(ME)}_{i-1,k}(\boldsymbol{\phi})}{i\mu}\right) G_{0,k}(\boldsymbol{\lambda}^{(ME)}_k(\boldsymbol{\phi})), \tag{33}$$

$$G_{0,k}(\boldsymbol{\lambda}^{(ME)}_k(\boldsymbol{\phi})) = \left(1 + \sum_{n=1}^{C_k}\left(\prod_{i=1}^{n} \frac{\lambda^{(ME)}_{i-1,k}(\boldsymbol{\phi})}{i\mu}\right)\right)^{-1}. \tag{34}$$

We prove this by using the contradiction arguments as follows. From the stationary MFEs, any fixed-point $\boldsymbol{\phi}$ of the mean-field satisfies the equation

$$\begin{aligned}&\sum_{b=1}^{Z(\underline{l})}\left(\frac{p_{l_b}}{Z(\underline{l})}\right)\phi_{(l_1,l_2,\ldots l_{b-1},l_{b+1},\ldots,l_{C_k},0,k)}\lambda^{(ME)}_{Z(\underline{l})-1,k}(\boldsymbol{\phi}) \\ &-\phi_{\underline{l},k}\lambda^{(ME)}_{Z(\underline{l}),k}(\boldsymbol{\phi}) + \sum_{b=1}^{Z(\underline{l})+1}\mu_p I_{\{l_{C_k}=0\}}\phi_{(l_1,\ldots,l_{b-1},1,l_b,\ldots,l_{C_k-1},k)} \\ &+\sum_{b=1}^{Z(\underline{l})}\mu_p\phi_{(l_1,\ldots,l_{b-1},l_b+1,l_{b+1},\ldots,l_{C_k},k)} - Z(\underline{l})\mu_p\phi_{\underline{l},k} = 0.\end{aligned} \tag{35}$$

Suppose there exists a fixed-point $\boldsymbol{\phi}^* = (\phi^*_{\underline{l},k}, \underline{l} \in S_k, k \in \mathcal{K})$ for the mean-field such that for some $q \in \mathcal{K}$, there exists a state $\mathbf{b} \in S_q$ such that equation (32) is not true. Let $\boldsymbol{\phi}^*_q = (\phi^*_{\underline{l},q}, \underline{l} \in S_q)$. For given $\boldsymbol{\phi}^*$, we first compute the set of arrival rates $\boldsymbol{\lambda}^{(ME)}(\boldsymbol{\phi}^*) = (\lambda^{(ME)}_{j,k}(\boldsymbol{\phi}^*), 0 \le j \le C_k, k \in \mathcal{K})$. Now consider an Erlang loss system with single server of type q and job lengths are distributed according to mixed-Erlang distribution as considered in the system model. Suppose jobs arrive according to a Poisson process with state-dependent arrival rates $\boldsymbol{\lambda}^{(ME)}_q(\boldsymbol{\phi}^*) = (\lambda^{(ME)}_{n,q}(\boldsymbol{\phi}^*), 0 \le n \le C_q)$ and $\lambda^{(ME)}_{n,q}(\boldsymbol{\phi}^*)$ be the arrival rate when server has n jobs. Then the dynamics of the probabilities for server occupancies is given by equation (14) where type k is replaced by type q and $\eta_{j,k}$ is replaced by $\lambda^{(ME)}_{j,q}(\boldsymbol{\phi}^*)$ for $0 \le j \le C_q$. Then for this system, let the unique steady-state distribution is denoted $\boldsymbol{\pi}^{(single)}_q(\boldsymbol{\lambda}^{(ME)}_q(\boldsymbol{\phi}^*)) = (\pi^{(single)}_{n,q}(\boldsymbol{\lambda}^{(ME)}_q(\boldsymbol{\phi}^*)), 0 \le n \le$

C_q). Furthermore. $\pi_{n,q}^{(single)}(\boldsymbol{\lambda}_q^{(ME)}(\boldsymbol{\phi^*}))$ is given by equation (15) where k is replaced by q and $\eta_{j,k}$ is replaced by $\lambda_{j,q}^{(ME)}(\boldsymbol{\phi^*})$ for $0 \leq j \leq C_q$.

However, since $\boldsymbol{\phi}^*$ is a fixed-point of the mean-field, from equation (35), it can be seen that $\boldsymbol{\phi}_q^* = (\phi^*_{\underline{l},q}, \underline{l} \in S_q)$ also satisfies the stationary system of equations (with dynamics as in equation (14)) of Erlang loss system with single server of type q having state-dependent arrival rates $\boldsymbol{\lambda_q^{(ME)}}(\boldsymbol{\phi}^*)$. Therefore $\boldsymbol{\phi}_q^*$ is also another stationary distribution for Erlang loss system with single server of type q having prespecified state-dependent arrival rates $\boldsymbol{\lambda_q^{(ME)}}(\boldsymbol{\phi}^*)$. This is in contradiction to the fact that an Erlang loss system with the prespecified state-dependent arrival rates always has unique stationary distribution which is shown in [27]. Therefore any fixed-point $\boldsymbol{\phi}$ of the mean-field must satisfy equation (32). For any fixed-point $\boldsymbol{\phi}$, let $Q_{n,k} = \sum_{\underline{l} \in S_k : Z(\underline{l}) = n} \phi_{\underline{l},k}$, $\boldsymbol{Q} = (Q_{n,k}, 0 \leq n \leq C_k)$, then we have $\lambda_{n,k}^{(ME)}(\boldsymbol{\phi}) = \lambda_{n,k}^{(exp)}(\boldsymbol{Q})$. Further, from equation (32), by summing over all $\underline{l} \in S_k$ such that $Z(\underline{l}) = n$ and since $\sum_{j=1}^M a_j = 1$, we have

$$Q_{n,k} = G_{n,k}(\boldsymbol{\lambda}_k^{(ME)}(\boldsymbol{\phi})). \tag{36}$$

From the definition of $G_{n,k}(\boldsymbol{\lambda}_k^{(ME)}(\boldsymbol{\phi}))$, we have

$$\lambda_{n,k}^{(ME)}(\boldsymbol{\phi}) G_{n,k}(\boldsymbol{\lambda}_k^{(ME)}(\boldsymbol{\phi})) = (n+1)\mu G_{n+1,k}(\boldsymbol{\lambda}_k^{(ME)}(\boldsymbol{\phi})) \tag{37}$$

and hence, we have

$$Q_{n,k} \lambda_{n,k}^{(exp)}(\boldsymbol{Q}) = (n+1)\mu Q_{n+1,k}. \tag{38}$$

Therefore from the definition of the mapping ψ in equation (29), $\boldsymbol{Q}$ must be the unique fixed-point of the mapping ψ and hence $\boldsymbol{Q} = \boldsymbol{\pi}^{(exp)}$. Therefore for every fixed-point $\boldsymbol{\phi}$ of the mean-field, we have,

$$\sum_{\underline{l} \in S_k, Z(\underline{l}) = n} \phi_{\underline{l},k} = \pi_{n,k}^{(exp)} \tag{39}$$

concluding insensitivity of the fixed-point of the mean-field. Further, from equation (32), for every fixed $\boldsymbol{\phi}$ of the mean-field, for $\underline{l} \in S_k$ with $Z(\underline{l}) = n$, we have

$$\phi_{\underline{l},k} = \pi_{n,k}^{(exp)} \prod_{j=1}^{n} a_{l_j}. \tag{40}$$

Since $\boldsymbol{\pi}^{(exp)}$ is unique and a_i is constant for all i, we conclude that there exists unique fixed-point $\boldsymbol{\pi}$ as defined in equation (18) for the mean-field. This completes the proof. ■

VI. Numerical Results

In this section we discuss the convergence of the stationary distributions as $N \to \infty$ and the issue of GAS of the fixed-point of the mean-field.

The results given correspond to system parameters as follows: There are four types of servers in the system ($K = 4$) with fraction of servers of types are fixed at $\beta_1 = 0.3, \beta_2 = .2, \beta_3 = 0.3, \beta_4 = .2$. The capacities of servers of the various types are assumed to be $C_1 = 2, C_2 = 3, C_3 = 4, C_4 = 5$. The job lengths are assumed to have mean one, i.e. $\mu = 1$ and service times are drawn from a mixed-Erlang distribution with parameters $M = 3$, $p_1 = .3, p_2 = 0.3, p_3 = 0.4$.

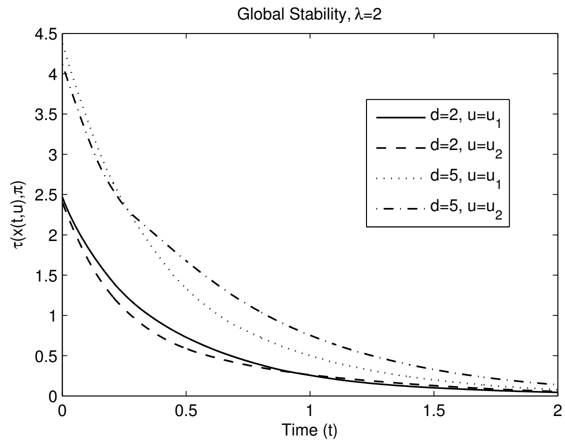

Fig. 1. Convergence of mean-field to the fixed-point

When service times follow mixed-Erlang distribution, for $\lambda = 2$, we plot the distance between the mean-field $\mathbf{x}(t, \mathbf{u})$ which is numerically evaluated from equations (10)-(12) by using Euler's method with step size 10^{-3} and the unique fixed-point $\boldsymbol{\pi}$ in Figure 1. It is observed that for $d = 2, 5$ and for two different initial points $\mathbf{u}_1, \mathbf{u}_2$, the mean-field $\mathbf{x}(t, \mathbf{u})$ converges to $\boldsymbol{\pi}$. Note that the computed $\boldsymbol{\pi}$ depends on the chosen value of d. This supports GAS of the mean-field.

In Table I, we provide simulation results that support convergence of the stationary distributions to the fixed-point of the mean-field as $N \to \infty$. Let $\boldsymbol{\pi}_{ME}^N = (\pi_{ME}^N(i,j), 0 \leq i \leq C_j, j \in \mathcal{K})$ where $\pi_{ME}^N(i,j)$ denotes the stationary probability that a type j server has i jobs in a system with N servers. Similarly, let $\boldsymbol{\pi}_C^N$, $\boldsymbol{\pi}_{PL}^N$ and $\boldsymbol{\pi}_{LN}^N$ denote the stationary distribution when job lengths are constant, drawn from power law distribution with CDF $G(y) = 1 - \frac{1}{4y^2}$ for $y \geq \frac{1}{2}$ and zero otherwise, a log normal distribution with mean one and variance $\sigma^2 = 4$, respectively.

We use the following metric to measure the distance between $\mathbf{u} = (u(i,j), 0 \leq i \leq C_j, j \in \mathcal{K})$ and $\mathbf{w} = (w(i,j), 0 \leq i \leq C_j, j \in \mathcal{K})$

$$\rho^*(\mathbf{u}, \mathbf{w}) = \sum_{i,j} |u(i,j) - w(i,j)| \tag{41}$$

Recently, in the exponential case, the error in approximating the stationary distribution $\boldsymbol{\pi}_E^N$ of a system with N servers with the fixed-point $\boldsymbol{\pi^{(exp)}}$ of the mean-field was studied by Stein's method in [28] where it was shown that if $\mathbf{W}^N = (W_{j,k}^N, 0 \leq j \leq C_k, k \in \mathcal{K})$ where $W_{j,k}^N$ denotes the fraction of type k severs having j jobs when servers' occupancies are sampled according to the stationary distribution $\boldsymbol{\pi}_E^N$, then

$$d^*(\boldsymbol{\pi}_E^N, \boldsymbol{\pi^{(exp)}}) := \mathbb{E}_{\boldsymbol{\pi}_E^N}[(\rho^*(\mathbf{W}^N, \boldsymbol{\pi^{(exp)}}))^2] = \mathcal{O}\left(\frac{1}{N}\right). \tag{42}$$

Our simulation results in Table I show that equation (42) is true even when $\boldsymbol{\pi}_E^N$ is replaced by $\boldsymbol{\pi}_{ME}^N$ $\boldsymbol{\pi}_C^N$, $\boldsymbol{\pi}_{PL}^N$ and $\boldsymbol{\pi}_{LN}^N$.

A further point to note from Table I is that even for small N the mean-square distance does not change much for different distributions implying near insensitivity for even for finite N. Similar behavior was also observed earlier for processor sharing systems under join-the-shortest-queue policy in [13].

TABLE I
CONVERGENCE OF THE STATIONARY DISTRIBUTIONS FOR $d = 2$, $\lambda = 0.5$

N	$d^*(\boldsymbol{\pi}_{ME}^N, \boldsymbol{\pi}^{(exp)})$	$d^*(\boldsymbol{\pi}_{C}^N, \boldsymbol{\pi}^{(exp)})$	$d^*(\boldsymbol{\pi}_{PL}^N, \boldsymbol{\pi}^{(exp)})$	$d^*(\boldsymbol{\pi}_{LN}^N, \boldsymbol{\pi}^{(exp)})$
10	0.8686	0.8678	0.8674	0.8681
20	0.4336	0.4339	0.4330	0.4347
50	0.1738	0.1738	0.1735	0.1732
100	0.0867	0.0868	0.0866	0.0867
200	0.0434	0.0434	0.0433	0.0434
500	0.0173	0.0173	0.0173	0.0173
1000	0.0086	0.0086	0.0086	0.0086

Finally we study the behavior of the system for small numbers of servers which in theory are not insensitive due to dependence across all servers that implies non-Poissonian inputs. For simplicity we consider a homogeneous system with capacity of servers equal to C. Table II lists total normalized deviation of the distribution for different service time models from stationary-distribution $\boldsymbol{\pi}_E^N$ for exponential service times. The metric we used is:

$$\nu(\boldsymbol{\pi}_E^N, \boldsymbol{\pi}_A^N) = \sum_{i=0}^{C} \frac{\left|\pi_E^N(i) - \pi_A^N(i)\right|}{\pi_E^N(i)}. \tag{43}$$

The system parameters chosen were $C = 2$, $\lambda = 1.8$, $d = 2$, $\mu = 1$ and the service times taken from power law distribution with the CDF $G(y) = 1 - \left(\frac{7}{12y}\right)^{2.4}$ for $y \geq \frac{7}{12}$ and zero otherwise and the parameters for the remaining distributions are same as that we used for Table I.

TABLE II
NEAR INSENSITIVITY OF FINITE N SYSTEMS

N	$\nu(\boldsymbol{\pi}_E^N, \boldsymbol{\pi}_C^N)$	$\nu(\boldsymbol{\pi}_E^N, \boldsymbol{\pi}_{LN}^N)$	$\nu(\boldsymbol{\pi}_E^N, \boldsymbol{\pi}_{PL}^N)$
3	0.0149	0.0030	0.0060
5	0.0071	0.0029	0.0026
10	0.0024	0.0026	0.0008
100	0.0013	0.0007	0.0011

From the table it is clear that the normalized deviations (although small) are non-zero for small values of N and decrease with increase in N. This clearly illustrates a service-time distribution dependent behavior that implies that the system is not insensitive for finite N. From a practical point of view since the magnitudes of the deviations are small the system can be taken to be nearly insensitive.

VII. CONCLUSIONS

In this paper, we have shown the insensitivity of the fixed-point of the MFE of multi-server loss models with mixed-Erlang holding times under power-of-d routing. This result provides strong evidence of insensitivity of the fixed-point of the mean-field for general service times. To claim that the fixed-point of the mean-field represents the stationary distribution for server occupancies, we need to establish the GAS of the fixed-point of the mean-field which appears very difficult due to the lack of quasi-monotonicity which was the key in proving the GAS for the exponential case [4]. However, the numerical evaluation of MFEs by using Euler's method supports the GAS of the MFE that will be pursued in the future.

REFERENCES

[1] "Amazon EC2." http://aws.amazon.com/ec2/.
[2] "Microsoft Azure." http://www.microsoft.com/windowsazure/.
[3] 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*, pp. 321–334, 2015.
[4] 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*, pp. 125–133, 2015.
[5] A. L. Stolyar, "Large-scale heterogeneous service systems with general packing constraints," *Advances in Applied Probability*, vol. 49, no. 1, 2017.
[6] 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.
[7] M. Bramson, Y. Lu, and B. Prabhakar, "Randomized load balancing with general service time distributions," in *Proceedings of ACM SIGMETRICS*, pp. 275–286, 2010.
[8] R. Aghajani, X. Li, and K. Ramanan, "Mean-field Dynamics of Load-Balancing Networks with General Service Distributions," *ArXiv e-prints*, Dec. 2015.
[9] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, "Statistical analysis of a telephone call center," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 36–50, 2005.
[10] P. Kolesar, "Stalking the endangered cat: A queueing analysis of congestion at automatic teller machines," *Interfaces*, vol. 14, no. 6, pp. 16–26, 1984.
[11] S. Chen, Y. Sun, U. Kozat, L. Huang, P. Sinha, G. Liang, X. Liu, and N. Shroff, "When queueing meets coding: Optimal-latency data retrieving scheme in storage clouds," in *Proceedings of IEEE INFOCOM*, pp. 1042–1050, April 2014.
[12] S. Asmussen, *Applied Probability and Queues*, vol. 51 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2003.
[13] V. Gupta, M. Harchol Balter, K. Sigman, and W. Whitt, "Analysis of Join-the-shortest-queue routing for web server farms," *Perform. Eval.*, vol. 64, pp. 1062–1081, Oct. 2007.
[14] Y. Azar, A. Z. Broder, A. R. Karlin, and E. Upfal, "Balanced allocations," *SIAM J. Comput.*, vol. 29, pp. 180–200, Sept. 1999.
[15] M. Mitzenmacher, "The power of two choices in randomized load balancing," *PhD Thesis, Berkeley*, 1996.
[16] 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.
[17] A. Mukhopadhyay and R. R. Mazumdar, "Rate-based randomized routing in large heterogeneous processor sharing systems," in *Proceedings of 26th International Teletraffic Congress (ITC 26)*, 2014.
[18] A. Mukhopadhyay and R. R. Mazumdar, "Analysis of randomized join-the-shortest-queue (jsq) schemes in large heterogeneous processor sharing systems," *IEEE Transactions on Control of Network Systems*, vol. 3(2), pp. 116–126, 2016.
[19] M. Bramson, Y. Lu, and B. Prabhakar, "Asymptotic independence of queues under randomized load balancing," *Queueing Systems*, vol. 71, no. 3, pp. 247–292, 2012.
[20] S. R. E. Turner, "Resource pooling in stochastic networks," *Ph.D. dissertation, University of Cambridge*, 1996.
[21] S. R. E. Turner, "The effect of increasing routing choice on resource pooling," *Probability in the Engineering and Informational Sciences*, vol. 12, pp. 109–124, 1998.

[22] C. Graham and S. Méléard, "Propagation of chaos for a fully connected loss network with alternate routing," *Stochastic Processes and their Applications*, vol. 44, no. 1, pp. 159–180, 1993.

[23] C. Graham and S. Méléard, "Stochastic particle approximations for generalized boltzmann models and convergence estimates," *The Annals of Probability*, vol. 28, no. 1, pp. 115–132, 1997.

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

[25] D. Stroock and S. R. S. Varadhan, *Multidimensional diffusion processes*, vol. 233 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin-New York, 1979.

[26] J. B. Martin and Y. M. Suhov, "Fast Jackson networks," *Ann. Appl. Probab.*, vol. 9, pp. 854–870, 08 1999.

[27] S. L. Brumelle, "A generalization of Erlang's loss system to state dependent arrival and service rates," *Mathematics of Operations Research*, vol. 3, no. 1, pp. 10–16, 1978.

[28] 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*, SIGMETRICS '16, pp. 285–297, ACM, 2016.

APPENDIX

We define the following notation which is used in computing the expression for the generator B^N. For $\underline{n} \in S_k$, let $\tilde{\mathbf{e}}(\underline{n}; i) = (\tilde{e}_{\underline{l},k}, \underline{l} \in S_k, k \in \mathcal{K})$ where

$$\tilde{e}_{\underline{l},k} = \begin{cases} 1 & \text{if } \underline{l} = \underline{n} \textit{ and } i = k \\ 0 & \text{otherwise.} \end{cases} \tag{44}$$

Similarly, let $\mathbf{e}_{i,k}$ be the vector in $\mathbb{R}^{C_k}$ with

$$\mathbf{e}_{i,k}(j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \tag{45}$$

Lemma *1:* For $\mathbf{u} \in \mathcal{V}^{(N)}$, the generator B^N of the process $\mathbf{x}^N(\cdot)$ acting on function $f : \mathcal{V}^{(N)} \to \mathbb{R}$ is given by

$$\begin{aligned} B^N f(\mathbf{u}) = N\lambda \sum_{k=1}^{K} \sum_{\underline{l} \in S_k} \sum_{r=1}^{Z(\underline{l})+1} \sum_{m=1}^{M} \left(\frac{p_m}{Z(\underline{l})+1} \right) \\ \times \beta_k \lambda^{(ME)}_{Z(\underline{l}),k}(\mathbf{u}) u_{\underline{l},k} \\ \times \left(f\left(\mathbf{u} - \frac{\tilde{\mathbf{e}}(\underline{l}; k)}{N\beta_k} + \frac{\tilde{\mathbf{e}}(l_1, \ldots, l_{r-1}, m, l_r, l_{r+1}, \ldots, l_{C_k-1}; k)}{N\beta_k} \right) \right. \\ \left. - f(\mathbf{u}) \right) I_{\{Z(\underline{l}) < C_k\}} + \sum_{k=1}^{K} \sum_{\underline{l} \in S_k} \sum_{i=1}^{Z(\underline{l})} N\beta_k \mu_p u_{\underline{l},k} \\ \left(f\left(\mathbf{u} - \frac{\tilde{\mathbf{e}}(\underline{l}; k)}{N\beta_k} + \frac{\tilde{\mathbf{e}}(\underline{l} - \mathbf{e}_{i,k}; k)}{N\beta_k} \right) - f(\mathbf{u}) \right) I_{\{l_i > 1\}} \\ + \sum_{k=1}^{K} \sum_{\underline{l} \in S_k} \sum_{i=1}^{Z(\underline{l})} N\beta_k \mu_p u_{\underline{l},k} \\ \times \left(f\left(\mathbf{u} - \frac{\tilde{\mathbf{e}}(\underline{l}; k)}{N\beta_k} + \frac{\tilde{\mathbf{e}}(l_1, l_2, \ldots, l_{i-1}, l_{i+1}, \ldots, l_{C_k}, 0; k)}{N\beta_k} \right) \right. \\ \left. - f(\mathbf{u}) \right) I_{\{l_i = 1\}} \end{aligned} \tag{46}$$

Proof First consider transition from the process state $\mathbf{u}$ to the process state $\mathbf{u} - \frac{\tilde{\mathbf{e}}(\underline{l}; k)}{N\beta_k} + \frac{\tilde{\mathbf{e}}(l_1, \ldots, l_{r-1}, m, l_r, l_{r+1}, \ldots, l_{C_k-1}; k)}{N\beta_k}$. This corresponds to transition of the process state when an arriving job with job length drawn from the Erlang distribution with m phases enters a server of type k with server state $\underline{l}$ and the index of new job is r. Since the destination server of the arriving job has state $\underline{l}$, twhen the new job enters the destination server, the fraction of type k servers having state $\underline{l}$ decreases by $\frac{1}{N\beta_k}$ and the fraction of type k servers having state $(l_1, \ldots, l_{r-1}, m, l_r, l_{r+1}, \ldots, l_{C_k-1})$ increases by $\frac{1}{N\beta_k}$. From power-of-d policy, all the potential destination servers must have vacancy at most $C_k - Z(\underline{l})$ and a server of type k with vacancy $C_k - Z(\underline{l})$ should be selected as the destination server. The number of type i servers with state $\underline{l}$ is equal to $N\beta_i u_{\underline{l},i}$. Therefore a potential destination server is a type i server with state $\underline{l}$ with probability $\beta_i u_{\underline{l},i}$.

Suppose j servers of the potential d destination servers are type k having occupancy $Z(\underline{l})$ and remaining $d - j$ are either type i, $i < k$ servers having vacancy at most $C_i - C_k + Z(\underline{l})$ or type i, $i \geq k$ servers having vacancy less than $C_i - C_k + Z(\underline{l})$ and out of j type k potential destination servers with occupancy $Z(\underline{l})$, assume z servers have state $\underline{l}$. Then the probability that the destination server is a type k with state $\underline{l}$ is given by

$$\begin{aligned} \binom{d}{j} \binom{j}{z} \left(\frac{z}{j} \right) (\beta_k u_{\underline{l},k})^z \\ \times \left(\beta_k F_{Z(\underline{l}),k}(\mathbf{u}) - \beta_k F_{Z(\underline{l})+1,k}(\mathbf{u}) - \beta_k u_{\underline{l},k} \right)^{j-z} \\ \times \left(\sum_{i=1}^{k-1} \beta_i F_{C_i - C_k + Z(\underline{l}),i}(\mathbf{u}) + \sum_{i=k}^{K} \beta_i F_{C_i - C_k + Z(\underline{l})+1,i}(\mathbf{u}) \right)^{d-j}. \end{aligned}$$

Further, the new job has job length drawn from Erlang distribution with m phases with probability p_m and the new job takes the index r with probability $\frac{1}{Z(\underline{l})+1}$. Finally by summing over all the possible values of j and z, upon arrival of a new job, by simplification, since jobs arrive at a rate $N\lambda$, the transition rate from the process state $\mathbf{u}$ to the process state $\mathbf{u} - \frac{\tilde{\mathbf{e}}(\underline{l}; k)}{N\beta_k} + \frac{\tilde{\mathbf{e}}(l_1, \ldots, l_{r-1}, m, l_r, l_{r+1}, \ldots, l_{C_k-1}; k)}{N\beta_k}$ is equal to

$$N\lambda \left(\frac{p_m}{Z(\underline{l})+1} \right) u_{\underline{l},k} \left(\frac{\beta_k \lambda^{(ME)}_{Z(\underline{l}),k}(\mathbf{u})}{\lambda} \right)$$

Finally, using the definition of the generator and summing over all possible values of k, r, m and $\underline{l}$, we get the first term of right hand side of the equation (46).

Further, consider transitions from the process state $\mathbf{u}$ when the current phase of a job expires. Suppose the current phase of i^{th} job of a server of type k with state $\underline{l}$ expires. Then if $l_i = 1$, the i^{th} job exits the server otherwise l_i changes to $l_i - 1$. In the first case, the process state changes to $\mathbf{u} - \frac{\tilde{\mathbf{e}}(\underline{l}; k)}{N\beta_k} + \frac{\tilde{\mathbf{e}}(l_1, l_2, \ldots, l_{i-1}, l_{i+1}, \ldots, l_{C_k}, 0; k)}{N\beta_k}$ at a rate $N\beta_k \mu_p u_{\underline{l},k}$. Similarly, the second case results in changing the process state from $\mathbf{u}$ to $\mathbf{u} - \frac{\tilde{\mathbf{e}}(\underline{l}; k)}{N\beta_k} + \frac{\tilde{\mathbf{e}}(\underline{l} - \mathbf{e}_{i,k}; k)}{N\beta_k}$ at a rate $N\beta_k \mu_p u_{\underline{l},k}$. Finally, using the definition of the generator, and summing over all possible values of k, $\underline{l}$ and i, we get the second term and third term of the right hand side of the equation (46). ■