The Power of Randomized Routing in Heterogeneous Loss Systems

Arpan Mukhopadhyay
Electrical and Computer Engineering
University of Waterloo, Canada
Email: arpan.mukhopadhyay@uwaterloo.ca

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

Fabrice Guillemin
Orange Labs, France,
Email: guillemin@orange.com

***Abstract*—Motivated by cloud computing applications, we consider a multi-server system, consisting of a large number of parallel servers, where jobs arrive according to a Poisson process and are assigned to the servers for processing. Each server has the capacity to process only a finite number of jobs simultaneously and different servers have different capacities. A job is accepted for processing only if there is a vacancy available at the server to which it is assigned. Otherwise, the job is discarded or blocked. We consider randomized schemes to assign jobs to servers with the aim of reducing the average blocking probability of jobs in the system. In particular, we consider a scheme that assigns an incoming job to the server having maximum available vacancy among d randomly sampled servers. We consider the system in the limit where both the number of servers and the arrival rate of jobs are scaled by a large factor. This gives rise to a mean field analysis. We show that in the limiting system servers behave independently. Stationary tail probabilities of server occupancies are obtained from the stationary solution of the mean field which is shown to be unique and globally attractive. We further characterize the rate of decay of the stationary tail probabilities. Numerical results suggest that the proposed scheme significantly reduces the average blocking probability of jobs compared to static schemes that probabilistically route jobs to servers independently of their states.**

I. Introduction

Consider a stream of jobs arriving at a multi-server system consisting of a large number of parallel servers each having finite amount of resource and no waiting room. The servers are categorized into different types according to the amounts of resources they hold. Each job, upon arrival, is routed/assigned to a server where it is either accepted or blocked depending on the availability of resources requested by the job. If accepted, the processing of the job begins immediately at the server. The objective is to design job routing/assigning schemes that reduce the average blocking probability of jobs in the system.

Such a model arises frequently in the context of cloud computing systems that provide infrastructure as a service [1]. A cloud service provider sells computing resources to its users in terms of virtual machines (VM's), that are computing instances consisting of various resources such as CPU, memory, storage etc. Each user requests a VM for a required amount of time. The VM request is then assigned to a physical machine (PM) or server where the request is either accepted or blocked depending on the availability of the requested VM. If accepted, the user holds the VM for the duration of its service after which it is released. Therefore, to maintain a certain quality of service, a cloud service provider should aim at reducing the blocking probability which measures the fraction of time a user is denied its requested resources.

The problem can be cast as a stochastic knapsack or a bin packing problem, where the blocking behavior is determined by the policy that assigns jobs to servers. In particular, we consider a model in which each server has the capacity to accommodate a finite number of job requests and different servers can have different capacity. We refer to this as the *heterogeneous loss system* model. There are other possible abstractions to model clouds that involve buffering of jobs in an infinite queue as was analyzed in [2], [3]. However the loss model is particularly relevant for the Infrastructure-as-a-service (IaaS) paradigm offered by Amazon's EC2 [1] and Microsoft's Azure [4] where each server can process only a finite number of jobs at a time.

We consider a randomized scheme to assign jobs to servers based on random sampling of $d \geq 2$ servers from the entire system. We show that assigning jobs to the server having the maximum vacancy amongst the d sampled servers yields dramatic reduction in the blocking probability as compared to that in static assignment schemes where job assignments are made independent of server states. Although assigning jobs based on the states of all the servers in the system would ideally minimize the blocking probability of jobs, such schemes involve high communication overhead due to large size of cloud computing systems. One can analyze the tradeoff between the cost of sampling and the improvement in performance as in [5]. However, in this paper we assume that the number of servers that are sampled (d) is fixed and is much smaller than the total number of servers.

Related Literature: The routing scheme that we consider is a loss model analog of the *power-of-d* scheme considered in [6]–[8] for first-come-first-serve (FCFS) queues and in [9], [10] for processor sharing (PS) servers. Turner [11], [12] studied this scheme for a system of identical Erlang servers with infinite capacity in the limit as the number servers $N \to \infty$. It was shown that in the large system limit the system behavior can be characterized by the so called *mean field equations* (differential equations) and the resulting tail distribution of server occupancies decay rapidly even when there is small number routing choices for each arriving job. The existence

and uniqueness of the stationary point of the mean field, however, were not shown and the rapid decay of stationary tail probabilities of the limiting system was demonstrated via simulation. A recent work by Xie *et al.* [13] analyzes the homogeneous Erlang loss model of Turner [11] and establishes the existence and uniqueness of the stationary point of the mean field. It assumes independence of servers in the limiting system. Such *asymptotic independence*, also known as the *propagation of chaos*, however, has been studied earlier in the context of alternative routing by Graham and Méléard [14], [15] where it is shown that loss models exhibit propagation of chaos on the path space.

The use of mean field techniques in analyzing large collections of interacting servers is not new. Early work on loss models in [16], [17] studied the hydrodynamic or mean field limits of the alternate routing problem in Erlang loss models. They were primarily interested in characterizing the mean field equations for homogeneous systems and did not address issues of uniqueness of the stationary solution, and convergence (stability) of the empirical occupancy distributions.

Contributions: In this paper, we not only generalize the results of Turner [12] to the case where servers are heterogeneous in nature but also rigorously establish the properties of mean field limit. In particular, we show that in the large system limit the system behavior can be characterized by a set of differential equations known as the mean field. We establish the uniqueness of the stationary point of the mean field and show that it is globally attractive. Moreover, we show that in the limiting system the servers become independent of each other. This result generalizes the notion of asymptotic independence to the heterogeneous case where statistical properties of the system are invariant with respect to permutation of states of servers of the same type. We also analytically characterize the rate of decay of stationary tail distribution of server occupancies in the limiting system. Numerical results suggest that the power-of-d scheme significantly reduces the average blocking probability of jobs in the system as compared to the static routing schemes.

The rest of the paper is organized as follows. In Section II, we introduce the system model and describe the routing scheme studied in this paper. We then present the main results in Section III. Section IV presents a detailed analysis of the randomized scheme. Section VI provides numerical results to compare different routing alternatives. Section VII concludes the paper with some remarks.

II. System model

We consider a system of N parallel servers where each server can host a finite number of virtual machines (VM's) and has no waiting room. The number of VM's a server can host is called the *capacity* of the server. The servers are categorized into M different types according to their capacities. Let $\mathcal{J} = \{1, 2, \ldots, M\}$ be the index set of server types. The capacity of type $j \in \mathcal{J}$ servers is denoted as C_j. Without loss of generality, we assume $C_1 \leq C_2 \leq \ldots \leq C_M$ holds. Furthermore, the fraction of type j servers in the system is assumed to be fixed and is denoted by $\gamma_j \in [0, 1]$ for all $j \in \mathcal{J}$. Clearly, we have $\sum_{j=1}^{M} \gamma_j = 1$.

Jobs arrive at the system according to a Poisson process of rate $N\lambda$. Each job requires one VM for its processing. Upon arrival, a job is routed to one of the N servers according to the following scheme.

Power-of-d scheme: Upon arrival of each job, $d \geq 2$ servers are sampled uniformly at random from the set of N servers. These sampled servers are called the *potential destination* servers for the arriving job. The job is then routed to the server having maximum vacancy or maximum number of unused VM's among the d sampled servers. Ties among servers of the same type are broken uniformly at random and ties across server types are broken by selecting the server type with the highest capacity (highest index). The server with the maximum vacancy, chosen after tie-breaking, is called the *destination* server for the arriving job. If the number of unused VM's at the destination server is more than or equal to one, then the processing of the job begins immediately at the destination server. Otherwise, the job is discarded or blocked and lost. Hence, it is clear from the description of the scheme that a job is discarded only when all the VM's of all d potential destination servers are occupied.

A job, if accepted, holds the requested VM for a random amount of time, exponentially distributed with mean 1. The service times of jobs are assumed to be independent of each other and independent of the inter-arrival times of the jobs. The VM held by a job is released immediately upon the completion of its service.

III. Main results

In this section, we state (without proof) the main results which are proved in the subsequent sections. Our results are asymptotic in the sense that they are derived in the limit as the system size $N \to \infty$ keeping the proportions γ_j, $j \in \mathcal{J}$, fixed. Such results are especially useful in the context of cloud computing systems since they typically run tens of thousands of servers.

Main results: For the model described in Section II, let $P_{k,j}^{(N)}$ denote the stationary probability that a server of type $j \in \mathcal{J}$ has at least k unfinished jobs. Then $P_{k,j}^{(N)}$ converges to $P_{k,j}$ as $N \to \infty$, where $P_{k,j}$ is the solution of the following recursive relation:

$$P_{k+1,j} - P_{k+2,j} = \frac{\lambda}{\gamma_j(k+1)} \left[\left(\sum_{i=1}^{j} \gamma_i P_{k+C_i-C_j,i} + \sum_{i=j+1}^{M} \gamma_i P_{k+C_i-C_j+1,i} \right)^d - \left(\sum_{i=1}^{j-1} \gamma_i P_{k+C_i-C_j,i} + \sum_{i=j}^{M} \gamma_i P_{k+C_i-C_j+1,i} \right)^d \right], \quad (1)$$

for $0 \leq k \leq C_j - 1$, $P_{k,j} = 1$ for $k \leq 0$, and $P_{C_j+1,j} = 0$ for all $j \in \mathcal{J}$. Furthermore, in the limit as $N \to \infty$ the servers become mutually independent.

Remark *1:* The independence of servers in the limiting system ($N \to \infty$), as stated above, can be used to compute the average blocking probability P_{blocking} of jobs in the limiting system in terms of the tail probabilities $P_{k,j}$ found by solving (1): The stationary probability that a server of type j is fully occupied is given by $P_{C_j,j}$ and the probability that it is sampled at an arrival instant is γ_j. Thus the total probability that a randomly sampled server is fully occupied is given by $\sum_{j\in\mathcal{J}} \gamma_j P_{C_j,j}$. Since the servers in the limiting system are mutually independent, the average blocking probability is given by $P_{\text{blocking}} = \left(\sum_{j\in\mathcal{J}} \gamma_j P_{C_j,j}\right)^d$.

Remark *2:* It can be readily verified that the results in [11], [13] for the homogeneous case can be recovered from (1) by setting $M = 1$, $C_1 = C$, $\gamma_1 = 1$, and using the notation $P_{k,1} = P_k$ for all k:

$$P_{k+1} - P_{k+2} = \frac{\lambda}{k+1}\left(P_k^d - P_{k+1}^d\right), \quad k = 0, 2, \ldots, C-1 \tag{2}$$

Furthermore, using (1) we obtain explicit bounds on the rate of decay of the tail probabilities $P_{k,j}$ as given in the following proposition.

Proposition *1:* Let $\left\{\bar{P}_k, 0 \leq k \leq C_M\right\}$ be defined as follows: $\bar{P}_k = 1$ for $0 \leq k \leq k_0$ and

$$\bar{P}_k = \frac{\lambda^{d^{k-k_0}-1}}{\prod_{l=0}^{k-k_0-1}(\lfloor\lambda\rfloor + k - k_0 - l)^{d^l}}, \tag{3}$$

for $k_0 + 1 \leq k \leq C_M$, where $k_0 = \lfloor\lambda\rfloor + C_M - C_1$, and $\lfloor y\rfloor$ denotes the greatest integer not exceeding y. Then $\sum_{j=1}^M \gamma_j P_{k+C_j-C_M,j} \leq \bar{P}_k$ for $0 \leq k \leq C_M$. In particular, the average user blocking probability $P_{\text{blocking}} = \left(\sum_{j\in\mathcal{J}} \gamma_j P_{C_j,j}\right)^d \leq \bar{P}_{C_M}^d$.

Proof: The proof is given in Appendix D. ∎

Proposition 1 shows that for $d \geq 2$ the quantity $\sum_{j=1}^M \gamma_j P_{k+C_j-C_M,j}$ decreases with k at a rate much faster than that for $d = 1$. This shows the efficacy of sampling a small number servers from the system over randomly sampling the destination server independent of the server states.

IV. Mean field analysis

The first step towards proving the main results discussed in Section III is to establish the mean field limit (a set of differential equations) that describes the behavior of the system as $N \to \infty$. In this section, we establish the mean field limit and show that the mean field has a unique and globally attractive stationary point. These results are used in Section V to prove the convergence of the stationary occupancy distribution to the stationary point of the mean field as $N \to \infty$ and to establish the independence of servers in the limiting system. We first introduce the notation and mathematical framework for the analysis.

Notations: Throughout the analysis we use $\mathbb{Z}$ and $\mathbb{Z}_+$ to denote the set of all integers and the set of non-negative integers, respectively. For each $j \in \mathcal{J}$, we define

$$\mathcal{U}_j = \big\{(g_n)_{n\in\mathbb{Z}} : g_n = 1 \text{ for } n \leq 0, g_n = 0 \text{ for } n > C_j, \text{ and } 1 = g_0 \geq g_1 \geq \ldots \geq g_{C_j} \geq 0 = g_{C_j+1}\big\}, \tag{4}$$

$$\mathcal{U}_j^{(N)} = \big\{(g_n)_{n\in\mathbb{Z}} \in \mathcal{U}_j : N\gamma_j g_n \in \mathbb{Z}_+ \forall n\big\}. \tag{5}$$

We will mainly be interested in the spaces $\mathcal{U} = \prod_{j\in\mathcal{J}} \mathcal{U}_j$ and $\mathcal{U}^{(N)} = \prod_{j\in\mathcal{J}} \mathcal{U}_j^{(N)}$, which are the Cartesian products of the spaces $\mathcal{U}_j$ and $\mathcal{U}_j^{(N)}$, respectively, over $j \in \mathcal{J}$. An element $\mathbf{u} = (u_{n,j}, n \in \mathbb{Z}, j \in \mathcal{J})$ is said to belong to $\mathcal{U}$ or $\mathcal{U}^{(N)}$ if the sequence $\{u_{n,j}, n \in \mathbb{Z}\}$ respectively belongs to $\mathcal{U}_j$ or $\mathcal{U}_j^{(N)}$, for each $j \in \mathcal{J}$. For $\mathbf{u}, \mathbf{w} \in \mathcal{U}$ we define the following distance metric

$$\rho(\mathbf{u}, \mathbf{w}) = \sup_{j\in\mathcal{J}} \sup_{n\geq 1} \left|\frac{u_{n,j} - w_{n,j}}{n+1}\right|. \tag{6}$$

It can be easily verified that under the metric defined in (6), the space $\mathcal{U}$ is compact (and hence complete and separable).

For a measure space $(H, \mathcal{H}, \mu_H)$ and a μ_H-integrable function $f : H \to \mathbb{R}$, we define duality brackets as $\langle f, \mu_H\rangle = \int f d\mu_H$. Law of a random variable X is denoted by $\mathcal{L}(X)$. Weak convergence (convergence in distribution) of a sequence of probability measures ν_n (random variables X_n) to a probability measure ν (random variable X) is denoted by $\nu_n \Rightarrow \nu$ ($X_n \Rightarrow X$).

Analysis

Consider the process $\mathbf{x}^{(N)}(t) = \left(x_{n,j}^{(N)}(t), n \in \mathbb{Z}, j \in \mathcal{J}\right)$, where $x_{n,j}^{(N)}(t)$ denotes the fraction of type-j servers having at least n unfinished jobs (occupied VM's) at time $t \geq 0$. By convention, we set $x_{n,j}^{(N)}(t) = 1$ for $n \leq 0$, $j \in \mathcal{J}$. Clearly, $\mathbf{x}^{(N)}(\cdot)$ is a Markov process in the state space $\mathcal{U}^{(N)}$. Moreover, for each $j \in \mathcal{J}$, the collection $\left(x_{n,j}^{(N)}(t), n \in \mathbb{Z}_+\right)$ denotes the empirical tail distribution of occupancy of type-j servers at time t.

The generator $\mathbf{A}^{(N)}$ of the Markov process $\mathbf{x}^{(N)}(\cdot)$ acting on functions $f : \mathcal{U}^{(N)} \to \mathbb{R}$ is given by $\mathbf{A}^{(N)} f(\mathbf{u}) = \sum_{\mathbf{w}\neq\mathbf{u}} r(\mathbf{u} \to \mathbf{w})(f(\mathbf{w}) - f(\mathbf{u}))$, where $r(\mathbf{u} \to \mathbf{w})$ denotes the transition rate from state $\mathbf{u} \in \mathcal{U}^{(N)}$ to state $\mathbf{w} \in \mathcal{U}^{(N)}$. In the following lemma, we provide the expression for the generator $\mathbf{A}^{(N)}$.

Lemma *1:* Let $\mathbf{u} \in \mathcal{U}^{(N)}$ and $\mathbf{e}(n,j) = (e_{k,i}, k \in \mathbb{Z}, i \in \mathcal{J})$ be the unit vector with $e_{n,j} = 1$ and $e_{k,i} = 0$ if $k \neq n$ or $i \neq j$. The generator $\mathbf{A}^{(N)}$ of the Markov process $\mathbf{x}^{(N)}(\cdot)$ acting on functions $f : \mathcal{U}^{(N)} \to \mathbb{R}$ is given by

$$\mathbf{A}^{(N)}f(\mathbf{u}) = N\lambda \sum_{n=1}^{C_j}\sum_{j=1}^{M}\Bigg[\Bigg(\sum_{i=1}^{j}\gamma_i u_{n-1+C_i-C_j,i} + \sum_{i=j+1}^{M}\gamma_i u_{n+C_i-C_j,i}\Bigg)^d - \Bigg(\sum_{i=1}^{j-1}\gamma_i u_{n-1+C_i-C_j,i} + \sum_{i=j}^{M}\gamma_i u_{n+C_i-C_j,i}\Bigg)^d\Bigg] \times \Bigg(f(\mathbf{u}+\frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{u})\Bigg) + N \times \sum_{n=1}^{C_j}\sum_{j=1}^{M} n\gamma_j\left(u_{n,j}-u_{n+1,j}\right)\Bigg(f(\mathbf{u}-\frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{u})\Bigg). \tag{7}$$

Proof: We first consider the transition from a state $\mathbf{u} \in \mathcal{U}^{(N)}$ to the state $\mathbf{u} + \frac{\mathbf{e}(n,j)}{N\gamma_j}$, with $1 \le n \le C_j$. This transition occurs when an arrival joins a server of type j which had exactly $n-1$ occupied VM's just before the arrival. For this to occur all type $i \le j$ servers selected as potential destinations must have at least $n-1+C_i-C_j$ occupied VM's. Since there are $N\gamma_i u_{n-1+C_i-C_j,i}$ servers of type i with at least $n-1+C_i-C_j$ occupied VM's, the probability that one such server is selected as a potential destination is $\gamma_i u_{n-1+C_i-C_j,i}$. Similarly, for the transition to occur all type $i > j$ servers selected as potential destinations must have at least $n+C_i-C_j$ occupied VM's and the probability that such a server is selected as a potential destination for the arriving job is $\gamma_i u_{n+C_i-C_j,i}$. Hence, the total probability with which d servers with the above properties are selected as potential destinations is given by $\left(\sum_{i=1}^{j}\gamma_i u_{n-1+C_i-C_j,i} + \sum_{i=j+1}^{M}\gamma_i u_{n+C_i-C_j,i}\right)^d$. However, this includes the possibility that all the type j potential destination servers have occupancy strictly more than $n-1$. Hence, the probability with which d servers are sampled such that type j servers with exactly $n-1$ jobs are the ones with maximum vacancy is given by $\left(\sum_{i=1}^{j}\gamma_i u_{n-1+C_i-C_j,i} + \sum_{i=j+1}^{M}\gamma_i u_{n+C_i-C_j,i}\right)^d - \left(\sum_{i=1}^{j-1}\gamma_i u_{n-1+C_i-C_j,i} + \sum_{i=j}^{M}\gamma_i u_{n+C_i-C_j,i}\right)^d$. Therefore, the rate of transition from state $\mathbf{u}$ to $\mathbf{u} + \frac{\mathbf{e}(n,j)}{N\gamma_j}$ is $N\lambda\left[\left(\sum_{i=1}^{j}\gamma_i u_{n-1+C_i-C_j,i} + \sum_{i=j+1}^{M}\gamma_i u_{n+C_i-C_j,i}\right)^d - \left(\sum_{i=1}^{j-1}\gamma_i u_{n-1+C_i-C_j,i} + \sum_{i=j}^{M}\gamma_i u_{n+C_i-C_j,i}\right)^d\right]$.

Furthermore, the rate at which jobs depart from type j servers having exactly n jobs is $nN\gamma_j\left(u_{n,j}-u_{n+1,j}\right)$. The expression in (7) now follows directly from the definition of $\mathbf{A}^{(N)}$. ∎

Using the generator $\mathbf{A}^{(N)}$ derived in Lemma 1 we now show that the process $\mathbf{x}^{(N)}(\cdot)$ converges weakly to a deterministic process as $N \to \infty$,

Theorem *1:* If $\mathbf{x}^{(N)}(0)$ converges in distribution to some constant $\mathbf{u}_0 \in \mathcal{U}$ as $N \to \infty$, then the process $\mathbf{x}^{(N)}(\cdot)$ converges in distribution to a deterministic process $\mathbf{x}(\cdot, \mathbf{u}_0)$, lying in the space $\mathcal{U}$ as $N \to \infty$, where the process $\mathbf{x}(\cdot, \mathbf{u}_0)$ is the unique solution of the following system of differential equations

$$\mathbf{x}(0, \mathbf{u}_0) = \mathbf{u}_0, \tag{8}$$

$$\dot{\mathbf{x}}(t, \mathbf{u}_0) = \mathbf{h}(\mathbf{x}(t, \mathbf{u}_0)), \tag{9}$$

and the mapping $\mathbf{h} : \mathcal{U} \to \left(\mathbb{R}^{\mathbb{Z}}\right)^M$ is given by

$$h_{n,j}(\mathbf{x}) = 0, \text{ for } n \le 0 \text{ and } n > C_j \text{ and } j \in \mathcal{J}, \tag{10}$$

$$h_{n,j}(\mathbf{x}) = \frac{\lambda}{\gamma_j}\Bigg[\Bigg(\sum_{i=1}^{j}\gamma_i x_{n-1+C_i-C_j,i} + \sum_{i=j+1}^{M}\gamma_i x_{n+C_i-C_j,i}\Bigg)^d - \Bigg(\sum_{i=1}^{j-1}\gamma_i x_{n-1+C_i-C_j,i} + \sum_{i=j}^{M}\gamma_i x_{n+C_i-C_j,i}\Bigg)^d\Bigg] - n\left(x_{n,j}-x_{n+1,j}\right), \text{ for } 1 \le n \le C_j, j \in \mathcal{J} \tag{11}$$

Proof: The proof is given in Appendix A. ∎

The process $\mathbf{x}(\cdot, \mathbf{u}_0)$, as defined by (8)-(9), is referred to as the *mean field*. We now focus on the *stationary points*, $\mathbf{P}$, of the mean field. By definition a stationary point $\mathbf{P}$ must satisfy $\mathbf{x}(t, \mathbf{P}) = \mathbf{P}$ for all $t \ge 0$. Hence, it must satisfy $h_{n,j}(\mathbf{P}) = 0$ for all $n \in \mathbb{Z}$ and $j \in \mathcal{J}$. In the next theorem we show that the mean field has a unique stationary point in the space $\mathcal{U}$ and it converges to the stationary point as $t \to \infty$ starting from any point in $\mathbf{u}_0 \in \mathcal{U}$.

Theorem *2:* There exists a unique stationary point $\mathbf{P}$ of the system (8)-(9) in the space $\mathcal{U}$. Moreover, for all $\mathbf{u}_0 \in \mathcal{U}$ we have

$$\lim_{t\to\infty}\mathbf{x}(t, \mathbf{u}_0) = \mathbf{P}, \tag{12}$$

Proof: The proof is given in Appendix B. ∎

We note that for each N the process $\mathbf{x}^{(N)}(\cdot)$ is positive recurrent and hence has a unique stationary distribution. Let $\pi^{(N)}$ denote its stationary distribution. We now show that as $N \to \infty$ the the stationary measure $\pi^{(N)}$ tends to concentrate on the unique stationary point $\mathbf{P}$ of the mean field.

Theorem *3:* The sequence of stationary measures $\left(\pi^{(N)}\right)_N$ converges weakly to the Dirac measure concentrated at the stationary point $\mathbf{P}$ of the system (8)-(9) as $N \to \infty$, i.e., $\pi^{(N)} \Rightarrow \delta_{\mathbf{P}}$.

Proof: Note that since the space $\mathcal{U}$ is compact, so is the space of probability measures on $\mathcal{U}$. Therefore, the sequence of probability measures $\left\{\pi^{(N)}\right\}_N$ must have limit points. Due to Theorem 1, any limit point of the sequence $\left(\pi^{(N)}\right)_N$ must be invariant under the map $\mathbf{g} \mapsto \mathbf{u}(t, \mathbf{g})$. Since by Theorem 2

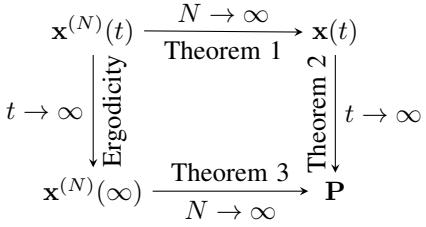

Fig. 1. Commutativity of limits

there exists only one stationary point $\mathbf{P}$ of the system (8)-(9) in $\mathcal{U}$, we conclude that every limit point of $\left(\pi^{(N)}\right)_N$ must coincide with $\delta_{\mathbf{P}}$. This completes the proof. ∎

Remark *3:* Let us define $\mathbf{x}^{(N)}(\infty)$ to be the random variable taking values in $\mathcal{U}^{(N)}$ with distribution $\pi^{(N)}$. By ergodicity of the process $\mathbf{x}^{(N)}(t)$ we have $\mathbf{x}^{(N)}(t) \Rightarrow \mathbf{x}^{(N)}(\infty)$ as $t \to \infty$. Moreover, Theorem 3 implies that $\mathbf{x}^{(N)}(\infty) \Rightarrow \mathbf{P}$ as $N \to \infty$. We have therefore established the convergences shown in Figure 1.

V. Propagation of chaos

In this section, we show that as the system size grows the server occupancies become independent of each other which is formally known as the *propagation of chaos* or *asymptotic independence property*. We further show that the stationary tail distribution of server occupancies in the limiting system is determined by the unique stationary point $\mathbf{P}$ of the mean field. To formally state the results we introduce the following notations.

- The occupancy of the k^{th} server of type j at a finite time $t \geq 0$ and at equilibrium are respectively denoted by the random variables $q_{k,j}^{(N)}(t)$ and $q_{k,j}^{(N)}(\infty)$, for $k \in \{1, 2, \ldots, N\gamma_j\}$, $j \in \mathcal{J}$.
- Define the process $\mathbf{Q}(t, \mathbf{u}) = (Q_{n,j}(t, \mathbf{u}), n \in \mathbb{Z}, j \in \mathcal{J})$ as $Q_{n,j}(t, \mathbf{u}) = x_{n,j}(t, \mathbf{u}) - x_{n+1,j}(t, \mathbf{u})$ for $t \in [0, \infty)$ and $Q_{n,j}(\infty, \mathbf{u}) = P_{n,j} - P_{n+1,j}$ for all $\mathbf{u} \in \mathcal{U}$. Furthermore, for each $j \in \mathcal{J}$ and $t \in [0, \infty]$, we denote by $Q_j(t, \mathbf{u})$ the distribution on $\mathbb{Z}$ given by $Q_j(t, \mathbf{u}) = (Q_{n,j}(t), n \in \mathbb{Z})$. We note that the $\mathbf{Q}(t, \mathbf{u})$ is uniquely determined by the mean field $\mathbf{x}(t, \mathbf{u})$.

Furthermore, we define the following notion of exchangeable random variables.

Definition Let $\left\{q_{k,j}^{(N)}, 1 \leq k \leq N\gamma_j, 1 \leq j \leq M\right\}$ denote a collection of N random variables classified into M different types. The collection is called intra-type exchangeable if the joint law of the collection is invariant under permutation of indices, $1 \leq k \leq N\gamma_j$, of random variables belonging to type j for each $j \in \{1, 2, \ldots, M\}$.

Proposition *2:* For the model considered in this paper, if $\left\{q_{k,j}^{(N)}(0), 1 \leq k \leq N\gamma_j, 1 \leq j \leq M\right\}$ is intra-type exchangeable and if $\mathbf{x}^{(N)}(0) \Rightarrow \mathbf{u} \in \mathcal{U}$ as $N \to \infty$, then the following holds

1) For each fix k and $t \in [0, \infty]$, $\mathcal{L}(q_{k,j}^{(N)}(t)) \Rightarrow Q_j(t, \mathbf{u})$ as $N \to \infty$.
2) Fix positive integers $r_1, r_2, \ldots, r_M$. For each $t \in [0, \infty]$,

$$\left\{q_{k,j}^{(N)}(t), 1 \leq k \leq r_j, 1 \leq j \leq M\right\} \Rightarrow \left\{U_{k,j}(t), 1 \leq k \leq r_j, 1 \leq j \leq M\right\},$$

as $N \to \infty$, where $U_{k,j}(t)$, $1 \leq k \leq r_j, 1 \leq j \leq M$, are independent random variables with $U_{k,j}(t)$ having distribution $Q_j(t)$ for all $1 \leq k \leq r_j$ and $j \in \mathcal{J}$.

Proof: The proof is given in Appendix C. ∎

Thus, Proposition 2 shows that in the limiting system server occupancies become independent of each other and the stationary occupancy distribution of any server of type-j server is given by $Q_j(\infty, \mathbf{u}) = \{P_{n,j} - P_{n+1,j}, n \in \mathbb{Z}\}$. In other words, the stationary probability that a server of type j in the limiting system has at least n occupied VM's is given by $P_{n,j}$. The stationary point $\mathbf{P}$ can be found by solving $\mathbf{h}(\mathbf{P}) = \mathbf{0}$ which corresponds to solving (1). However, we provide a simpler way to compute $\mathbf{P}$ which requires the following proposition.

Proposition *3:* In equilibrium, the arrival process of jobs at any given server in the limiting system is a state dependent Poisson process. Furthermore, in the equilibrium, the arrival rate of jobs to a server of type $j \in \mathcal{J}$ when it has occupancy n is given by

$$\lambda_{n,j} = \frac{\lambda}{\gamma_j \left(P_{n,j} - P_{n+1,j}\right)} \left[\left(\sum_{i=1}^{j} \gamma_i P_{n+C_i-C_j,i} + \sum_{i=j+1}^{M} \gamma_i P_{n+1+C_i-C_j,i} \right)^d - \left(\sum_{i=1}^{j-1} \gamma_i P_{n+C_i-C_j,i} + \sum_{i=j}^{M} \gamma_i P_{n+1+C_i-C_j,i} \right)^d \right] \quad (13)$$

for $0 \leq n \leq C_j - 1$ and $\lambda_{n,j} = 0$ for $n \geq C_j$.

Proof: Consider a *tagged* server of type j and the arrivals that have the tagged server as one of its potential destinations. These arrivals constitute the *potential arrival process* at the tagged server. The probability that the tagged server is sampled at the arrival instant of a job is $\frac{\binom{N-1}{d-1}}{\binom{N}{d}} = \frac{d}{N}$. Thus, due to Poisson thinning, the potential arrival process to the tagged server is a Poisson process with rate $\frac{d}{N} \times N\lambda = d\lambda$.

Next, we consider the arrivals that actually join the tagged server. These arrivals constitute the actual arrival process at the server. For finite N, this process is not Poisson since a potential arrival actually joins the tagged server depending on the number of jobs present at the other potential destination servers. However, as $N \to \infty$, due to the asymptotic independence property shown in Proposition 2 the occupancies of the sampled servers become independent of each other. As a result, in equilibrium the actual arrival process converges to a state

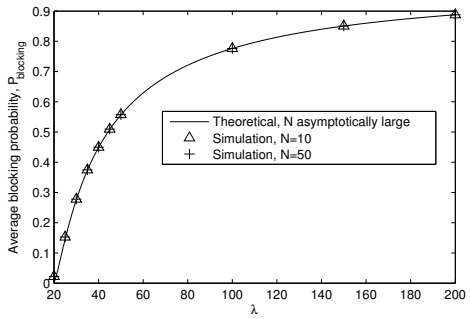

Fig. 2. Accuracy of mean field analysis of the power-of-d scheme: Average blocking probability as a function of λ for different values of N.

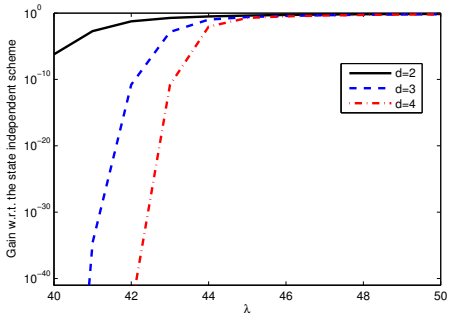

Fig. 3. Efficacy of the power-of-d scheme: Ratio of average blocking probability of the power-of-d scheme to that of the state independent scheme as a function of λ

dependent Poisson process as $N \to \infty$. Obtaining the arrival rates of the Poisson process, as given in (13), is a routine combinatorial exercise using the asymptotic independence property established in Theorem 2. ■

Hence, the above proposition shows that in equilibrium the arrival rate at a given server depends on the state of the server through stationary tail probabilities $P_{n,j}$, $n \in \mathbb{Z}_+$, $j \in \mathcal{J}$. The stationary tail probabilities can in turn be expressed in terms of the arrival rates. Indeed, in equilibrium the global balance equations (which hold even under state dependent Poisson arrivals due to Theorems 3.10 and 3.14 of [18]) yield

$$(P_{n,j} - P_{n+1,j})\,\lambda_{n,j} = (n+1)\,(P_{n+1,j} - P_{n+2,j})\,, \quad (14)$$

where $j \in \mathcal{J}, 0 \leq n \leq C_j - 1$. Thus, the stationary point $\mathbf{P}$ is a fixed point of the mapping $\Theta : \mathcal{U} \to \mathcal{U}$ defined as $\Theta(\mathbf{P}) = F(G(\mathbf{P}))$, where $G(\cdot)$ denotes the mapping defined by (13) from $\mathcal{U}$ to the space of possible arrival rates and $F(\cdot)$ denotes the mapping defined by (14) from the space of possible arrival rates to the space $\mathcal{U}$. We note that we have already shown the existence and uniqueness of fixed point of the map $\Theta(\cdot)$ in the proof of Theorem 1. Thus, the stationary point $\mathbf{P}$ can be computed using the standard fixed point iterations (i.e., by repeatedly applying the mapping $\Theta(\cdot)$ to some arbitrary point $\mathbf{Q} \in \mathcal{U}$).

Remark *4:* So far our results have been derived under the assumption that the service times of jobs are exponentially distributed. The conclusions of Proposition 3 continue to hold for any service time distribution due to the Whittle balance criterion [19], [20] that can be shown to hold for the stationary distribution (also see Theorems 3.10 and 3.14 of [18]). In view of the uniqueness of the stationary distribution and propagation of chaos this suggests that in the stationary regime the servers are asymptotically independent for general job size distributions. In Section VI, we provide numerical evidence to support insensitivity.

VI. Numerical Results

We first numerically investigate the accuracy of the asymptotic analysis presented in the paper in predicting the system performance for finite system size N. We set the following parameter values: $M = 2$, $d = 2$, $\gamma_1 = \gamma_2 = 0.5$, $C_1 = 20$, and $C_2 = 25$. All simulation results presented in this section are the average of 10,000 independent runs. In Figure 2, we plot the average blocking probability of users under the power-of-two scheme as a function of λ for $N = 10, 50$. We have also plotted the blocking probability obtained by solving the fixed point of (13) and (14). We observe that results obtained from the simulations match almost exactly with those obtained from the analysis. This leads us to believe that the mean-field results derived in this paper very accurately predict the behavior of the schemes even for moderate system sizes.

In Figure 3 we plot as a function of λ the ratio of the average blocking probability of the power-of-d scheme to that of the state independent routing scheme, in which an incoming job is routed to a server of type j with probability p_j independent of the states of the servers. We set $p_j = \gamma_j C_j / \sum_{i=1}^{M} \gamma_i C_i$ for all $j \in \mathcal{J}$ since in this case p_j is proportional to both γ_j and C_j.[1] The parameters were chosen as $M = 2$, $\gamma_1 = \gamma_2 = 0.5$, $N = 100$, $C_1 = 30$, and $C_2 = 60$. For this parameter setting the critical load is given by $\gamma_1 C_1 + \gamma_2 C_2 = 45$. Note that the y-axis is in log scale in 3. As can be seen from Figure 3, the average blocking probability under the power-of-d scheme is orders of magnitude lower than that under the state independent routing scheme even for small values of d. This shows the effectiveness of such a randomized strategies in reducing blocking for realistic systems which are typically operated below the critical load.

We now numerically verify the asymptotic insensitivity of the power-of-two scheme under different service time distributions. We set the following parameter values: $M = 2$, $d = 2$, $N = 100$, $\gamma_1 = \gamma_2 = 0.5$, $C_1 = 20$, and $C_2 = 25$. In Table I, average blocking probability is shown as a function of λ, for the following distributions.

1) *Constant*: We consider job length distribution having the cumulative distribution given by $F(x) = 0$ for $0 \leq x < 1$, and $F(x) = 1$, otherwise.
2) *Power law*: We consider job length distribution having

[1]The probabilities p_j, $j \in \mathcal{J}$, can be optimally chosen to minimize the average blocking probability. However, such optimal choice requires the knowledge of the arrival rate λ, which is difficult to estimate.

cumulative distribution function given by $F(x) = 1 - 1/4x^2$ for $x \geq \frac{1}{2}$ and $F(x) = 0$, otherwise.

Note that for each of the above distributions the average service time is 1. We see from Table I that the change in blocking probability is insignificant when the service time distribution is changed keeping the same mean.

TABLE I
INSENSITIVITY OF THE POWER-OF-d SCHEME

λ	Constant (Simulation)	Power Law (Simulation)
20	0.0087	0.0086
25	0.1467	0.1470
30	0.2758	0.2747
35	0.3733	0.3737
40	0.4490	0.4485
45	0.5085	0.5085

VII. CONCLUDING REMARKS

In this paper, we analyzed the power-of-two scheme for a heterogeneous multi-server Erlang loss system. We showed that in the large system limit the evolution of the empirical occupancy distribution can be characterized through its mean field limit. We established the existence and uniqueness of the stationary point of the mean field limit and showed that the stationary tail distribution of loads at each server decreases faster than that in the $M/G/C/C$ case. Furthermore, we showed that propagation of chaos holds for heterogeneous case through the requirement of type-based exchangeability.

A natural and more versatile model for clouds is one in which in addition to heterogeneous servers, the arriving jobs belong to heterogeneous classes with differing requirements. These are heterogeneous multi-rate loss models whose analysis is more challenging and will be the subject of a future work.

APPENDIX

A. Proof of Theorem 1

Due to space constraints we only provide the sketch of the proof. The proof consists of three main steps. The first step is to show that the sequence of Markov processes $\left(\mathbf{x}^{(N)}(\cdot)\right)_N$ is relatively compact. The second step is to show that there exists a unique process $\mathbf{x}(\cdot)$ satisfying (8)-(9). The third step is to show that the operator semigroup $\left(\mathbf{T}^{(N)}(t), t \geq 0\right)$ generated by $\mathbf{A}^{(N)}$ corresponding to the Markov process $\mathbf{x}^{(N)}(\cdot)$ converges to the operator semigroup of the process $\mathbf{x}(\cdot)$, i.e., $\lim_{N\to\infty} \sup_{\mathbf{u}\in\mathcal{U}^{(N)}} \left|\mathbf{T}^{(N)}(t)f(\mathbf{u}) - f(\mathbf{x}(t,\mathbf{u}))\right| = 0$. for all continuous functions $f : \mathcal{U} \to \mathbb{R}$, where the convergence is uniform in t within any bounded interval. The statement of Theorem 1 then follows from Corollary 8.7 of Chapter 4 of [21]. The proof of the first step follows mutatis mutandis from the arguments of Theorem 6.1 of [11]. The proof of the second step follows by showing that $\mathbf{h} : \mathcal{U} \to \left(\mathbb{R}^{\mathbb{Z}}\right)^M$ is Lipschitz continuous which is an easily consequence of the metric defined in (6). The third step follows from the observation that $\mathbf{A}^{(N)}f(\mathbf{u}) \to \frac{d}{dt}f(\mathbf{x}(t,\mathbf{u}))|_{t=0}$ as $N \to \infty$ uniformly in $\mathbf{u}$.

B. Proof of Theorem 2

For any point $\mathbf{x} \in \mathcal{U}$ we define for each $j \in \mathcal{J}$ and $n \in \mathbb{Z}$

$$\lambda_{n,j}(\mathbf{x}) = \frac{\lambda}{\gamma_j\left(x_{n,j} - x_{n+1,j}\right)}\left[\left(\sum_{i=1}^{j}\gamma_i x_{n+C_i-C_j,i} + \sum_{i=j+1}^{M}\gamma_i x_{n+1+C_i-C_j,i}\right)^d - \left(\sum_{i=1}^{j-1}\gamma_i x_{n+C_i-C_j,i} + \sum_{i=j}^{M}\gamma_i x_{n+1+C_i-C_j,i}\right)^d\right] \tag{15}$$

$$\lambda_{n,j}(\mathbf{x}) = 0, \text{ for } n \geq C_j \text{ and } n < 0. \tag{16}$$

Next, for each $j \in \mathcal{J}$ and $n \in \mathbb{Z}_+$, define the quantities $\pi_{n,j}(\mathbf{x})$ recursively as

$$\pi_{n+1,j}(\mathbf{x}) = \frac{\lambda_{n,j}(\mathbf{x})}{n+1}\pi_{n,j}(\mathbf{x}), \tag{17}$$

where $\pi_{0,j}(\mathbf{x}) = \left(1 + \sum_{n=0}^{C_j-1}\frac{\lambda_{n,j}(\mathbf{x})\lambda_{n-1,j}(\mathbf{x})\ldots\lambda_{0,j}(\mathbf{x})}{(n+1)!}\right)^{-1}$. Finally, for each $j \in \mathcal{J}$ and $n \in \mathbb{Z}_+$, define

$$y_{n,j}(\mathbf{x}) = \sum_{k\geq n}\pi_{k,j}(\mathbf{x}), \tag{18}$$

and $y_{n,j} = 1$ for $n \leq 0$. Clearly, $\mathbf{y} \in \mathcal{U}$ and the map $\mathbf{x} \mapsto \mathbf{y}(\mathbf{x})$, as defined above, is continuous in $\mathcal{U}$. Furthermore, from (11) it can be easily seen that any a fixed point, $\mathbf{P}$, of the map $\mathbf{x} \mapsto \mathbf{y}(\mathbf{x})$ must satisfy $\mathbf{h}(\mathbf{P}) = \mathbf{0}$. Now, since $\mathcal{U}$ is compact under the metric defined in (6), Brouwer's fixed point theorem guarantees the existence of a fixed point of the map $\mathbf{x} \mapsto \mathbf{y}(\mathbf{x})$. Hence, there exists a point $\mathbf{P} \in \mathcal{U}$ such that $\mathbf{h}(\mathbf{P}) = 0$. The uniqueness will follow from the uniqueness of limit in (12).

To prove (12), we first state the following lemma. We will write $\mathbf{u} \leq \mathbf{w}$ to mean that $u_{n,j} \leq w_{n,j}$ holds for all $n \in \mathbb{Z}$ and $j \in \mathcal{J}$.

Lemma *2:* If $\mathbf{u} \leq \mathbf{w}$ holds, for $\mathbf{u}, \mathbf{w} \in \mathcal{U}$, then $\mathbf{x}(t,\mathbf{u}) \leq \mathbf{x}(t,\mathbf{w})$ holds for all $t \geq 0$.

Proof: The proof directly follows by noting that $dx_{n,j}(t)/dt = h_{n,j}(\mathbf{x})$, given by (11), is non-decreasing in $x_{l,i}$ for $l \neq k$ and $i \neq j$ (see [22]). ■

Clearly, Lemma 2 implies $\mathbf{x}(t, \min(\mathbf{u},\mathbf{P})) \leq \mathbf{x}(t,\mathbf{u}) \leq \mathbf{x}(t, \max(\mathbf{u},\mathbf{P}))$ for all $t \geq 0$. Hence, to prove (12), it is sufficient to show that the convergence holds for the upper and lower bounds. Therefore, it is sufficient to prove the convergence given by (12) holds for $\mathbf{u} \geq \mathbf{P}$ and for $\mathbf{u} \leq \mathbf{P}$.

Since the derivative of $x_{n,j}(t,\mathbf{u})$ is bounded for all $j \in \mathcal{J}$, the convergence $\mathbf{x}(t,\mathbf{u}) \to \mathbf{P}$ for all $\mathbf{u} \geq \mathbf{P}$ will hold if for each $j \in \mathcal{J}, n \geq 1$ we have $\int_0^\infty (x_{n,j}(t,\mathbf{u}) - P_{n,j})\,dt < \infty$. Similarly, for all $\mathbf{u} \leq \mathbf{P}$ the convergence holds if for each $j \in \mathcal{J}, n \geq 1$ we have $\int_0^\infty (P_{n,j} - x_{n,j}(t,\mathbf{u}))\,dt < \infty$. We prove the case for $\mathbf{u} \geq \mathbf{P}$. The proof for $\mathbf{u} \leq \mathbf{P}$ follows similarly.

Let $v(t,\mathbf{u}) = \sum_{j\in\mathcal{J}}\gamma_j\sum_{n\geq 1}x_{n,j}(t,\mathbf{u})$ and $v(\mathbf{u}) = \sum_{j\in\mathcal{J}}\gamma_j\sum_{n\geq 1}u_{n,j}$ for each $u\in\mathcal{U}$. From (10) and (11) we obtain

$$\frac{dv(t,\mathbf{u})}{dt} = \lambda\left(1-\left(\sum_{j=1}^{M}\gamma_j x_{C_j,j}(t,\mathbf{u})\right)^2\right) - v(t,\mathbf{u}). \quad (19)$$

We first need to check that for each $\mathbf{u}\geq\mathbf{P}$, the quantity $v(t,\mathbf{u})$ is bounded uniformly in t. If $\mathbf{u}\geq\mathbf{P}$, then Lemma 2 implies $\mathbf{x}(t,\mathbf{u})\geq\mathbf{x}(t,\mathbf{P})=\mathbf{P}$. Hence, we have

$$\lambda\left(1-\left(\sum_{j=1}^{M}\gamma_j x_{C_j,j}(t,\mathbf{u})\right)^2\right) - v(t,\mathbf{u}) \leq \lambda\left(1-\left(\sum_{j=1}^{M}\gamma_j P_{C_j,j}\right)^2\right) - v(\mathbf{P}) = 0 \quad (20)$$

Thus, from (19) we have $dv(t,\mathbf{u})/dt\leq 0$. Hence, we have $0\leq v(t,\mathbf{u})\leq v(0,\mathbf{u}) = v(\mathbf{u})$ for all $t\geq 0$.

To prove $\int_0^\infty (x_{n,j}(t,\mathbf{u})-P_{n,j})\,dt < \infty$ holds for all $j\in\mathcal{J}, n\geq 1$ it is sufficient to show that $\int_0^\infty (v(t,\mathbf{u})-v(\mathbf{P}))\,dt<\infty$. We have

$$\int_0^\tau (v(t,\mathbf{u})-v(\mathbf{P}))\,dt = -\int_0^\tau \frac{dv(t,\mathbf{u})}{dt}dt - \lambda\int_0^\tau\left(\left(\sum_{j=1}^{M}\gamma_j x_{C_j,j}(t,\mathbf{u})\right)^2 - \left(\sum_{j=1}^{M}\gamma_j P_{C_j,j}\right)^2\right)dt \leq (v(\mathbf{u})-v(\tau,\mathbf{u})). \text{ (since } \mathbf{x}(t,\mathbf{u})\geq\mathbf{P})$$

Since the right hand side is bounded by a constant for all τ, the integral on the left hand side must converge as $\tau\to\infty$. This completes the proof.

C. Proof of Proposition 2

Note that the first part of Proposition 2 is a special case of the second part. Hence, it is sufficient to prove the second part. We will provide a proof for the $M=2$ case. The proof can be readily generalized to any $M\geq 2$.

Define the process $\boldsymbol{\chi}^{(N)}(t) = \left\{\chi_{n,j}^{(N)}(t), n\in\mathbb{Z}, j\in\mathcal{J}\right\}$ as $\chi_{n,j}^{(N)}(t) = x_{n,j}^{(N)}(t) - x_{n+1,j}^{(N)}(t)$ for $t\in[0,\infty]$. Thus, $\chi_{n,j}^{(N)}(t)$ and $\chi_{n,j}^{(N)}(\infty)$ denote the fraction of type j servers having occupancy n at a finite time $t\geq 0$ and in equilibrium, respectively. We use $\chi_j^{(N)}(t) = \left\{\chi_{n,j}^{(N)}(t), n\in\mathbb{Z}\right\}$ to denote the empirical distribution occupancies of type j servers at time $t\in[0,\infty]$.

The dynamics of the system (power-of-d scheme) and the intra-type exchangeability of $\{q_{k,j}^{(N)}(0), 1\leq k\leq N\gamma_j, 1\leq j\leq M\}$ implies that the collection $\{q_{k,j}^{(N)}(t), 1\leq k\leq N\gamma_j, 1\leq j\leq M\}$ is intra-type exchangeable for all $t\in[0,\infty]$. Also from Theorem 1 we have $\left(\boldsymbol{\chi}^{(N)}(t), t\geq 0\right)\Rightarrow(\mathbf{Q}(t,\mathbf{u}), t\geq 0)$ as $N\to\infty$. Henceforth, we will omit the variables t and $\mathbf{u}$ in our calculations, which hold for all $t\in[0,\infty]$ and $\mathbf{u}\in\mathcal{U}$.

To prove the proposition, it is sufficient to show that the following holds: $\mathbb{E}\left[\prod_{k=1}^{r_1}\phi_k\left(q_{k,1}^{(N)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_{k,2}^{(N)}\right)\right] \to \prod_{k=1}^{r_1}\langle\phi_k,Q_1\rangle\prod_{k=1}^{r_2}\langle\psi_k,Q_2\rangle$, for all bounded mappings $\phi_k,\psi_k:\mathbb{Z}_+\to\mathbb{R}_+$ as $N\to\infty$. We have

$$\begin{aligned}
&\left|\mathbb{E}\left[\prod_{k=1}^{r_1}\phi_k\left(q_{k,1}^{(N)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_{k,2}^{(N)}\right)\right] - \prod_{k=1}^{r_1}\langle\phi_k,Q_1\rangle\prod_{k=1}^{r_2}\langle\psi_k,Q_2\rangle\right| \\
&\leq \left|\mathbb{E}\left[\prod_{k=1}^{r_1}\phi_k\left(q_{k,1}^{(N)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_{k,2}^{(N)}\right)\right] - \mathbb{E}\left[\prod_{k=1}^{r_1}\langle\phi_k,\chi_1^{(N)}\rangle\prod_{k=1}^{r_2}\langle\psi_k,\chi_2^{(N)}\rangle\right]\right| \\
&\quad + \left|\mathbb{E}\left[\prod_{k=1}^{r_1}\langle\phi_k,\chi_1^{(N)}\rangle\prod_{k=1}^{r_2}\langle\psi_k,\chi_2^{(N)}\rangle\right] - \prod_{k=1}^{r_1}\langle\phi_k,Q_1\rangle\prod_{k=1}^{r_2}\langle\psi_k,Q_2\rangle\right|. \quad (21)
\end{aligned}$$

Note that the second term on the right hand side of the above inequality vanishes as $N\to\infty$ since $\chi_j^{(N)}\Rightarrow Q_j$ as $N\to\infty$ for $j=1,2$ and Q_1 and Q_2 are constants. Now, due to intra-type exchangeability the permutation of states between servers belonging to the same class does not affect the joint distribution. Hence, we have

$$\mathbb{E}\left[\prod_{k=1}^{r_1}\phi_k\left(q_{k,1}^{(N)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_{k,2}^{(N)}\right)\right] = \frac{1}{(N\gamma_1)_{r_1}(N\gamma_2)_{r_2}}\times \mathbb{E}\left[\sum_{\substack{\sigma\in P(r_1,N\gamma_1)\\ \sigma'\in P(r_2,N\gamma_2)}}\prod_{k=1}^{r_1}\phi_k\left(q_{\sigma(k),1}^{(N)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_{\sigma'(k),2}^{(N)}\right)\right] \quad (22)$$

where $(N)_k = N(N-1)\ldots(N-k+1)$, and $P(r,n)$ denotes the set of all permutations of the numbers $\{1,2,\ldots,n\}$ taken r at a time. Also, by definition of $\chi_j^{(N)}$ we have

$$\begin{aligned}
&\mathbb{E}\left[\prod_{k=1}^{r_1}\langle\phi_k,\chi_1^{(N)}\rangle\prod_{k=1}^{r_2}\langle\psi_k,\chi_2^{(N)}\rangle\right] \\
&= \mathbb{E}\left[\left(\prod_{k=1}^{r_1}\frac{1}{N\gamma_1}\sum_{l=1}^{N\gamma_1}\phi_k\left(q_{l,1}^{(N)}\right)\right)\times\left(\prod_{k=1}^{r_2}\frac{1}{N\gamma_2}\sum_{l=1}^{N\gamma_2}\psi_k\left(q_{l,2}^{(N)}\right)\right)\right] \quad (23)
\end{aligned}$$

Hence, the first term on the right hand side of (21) can be bounded as follows

$$\left| \mathbb{E}\left[\prod_{k=1}^{r_1}\phi_k\left(q_{k,1}^{(N)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_{2,k}^{(N)}\right)\right] - \mathbb{E}\left[\prod_{k=1}^{r_1}\langle\phi_k,\chi_1^{(N)}\rangle\prod_{k=1}^{r_2}\langle\psi_k,\chi_2^{(N)}\rangle\right]\right| \leq 2B^{r_1+r_2}\left(1-\frac{(N\gamma_1)_{r_1}(N\gamma_2)_{r_2}}{(N\gamma_1)^{r_1}(N\gamma_2)^{r_2}}\right) \quad \to 0 \text{ as } N\to\infty$$

where $\max\left(\|\phi_k\|_\infty, \|\psi_k\|_\infty\right) = B$. This completes the proof. ∎

D. Proof of Proposition 1

Using (1) it can easily verified that the following recursive relation holds for $0 \leq k \leq C_M - 1$

$$\sum_{j\in\mathcal{J}}(k+1+C_j-C_M)_+\gamma_j(P_{k+1+C_j-C_M,j}-P_{k+2+C_j-C_M,j}) = \lambda\left[\left(\sum_{j=1}^{M}\gamma_j P_{k+C_j-C_M,j}\right)^d - \left(\sum_{j=1}^{M}\gamma_j P_{k+1+C_j-C_M,j}\right)^d\right], \quad (24)$$

where $(y)_+ = \max(0,y)$. From (24) the following can be shown to hold for $0 \leq k \leq C_M - 1$ using backward induction starting at $k = C_M - 1$.

$$\sum_{j=1}^{M}(k+1+C_j-C_M)_+\gamma_j P_{k+1+C_j-C_M,j} \leq \lambda\left(\sum_{j=1}^{M}\gamma_j P_{k+C_j-C_M,j}\right)^d. \quad (25)$$

From (25) it is clear that

$$\left(\sum_{j=1}^{M}\gamma_j P_{k+1+C_j-C_M,j}\right) \leq \frac{\lambda}{k+(C_1-C_M)+1} \times \left(\sum_{j=1}^{M}\gamma_j P_{k+C_j-C_M,j}\right)^d, \quad (26)$$

for $C_M - C_1 \leq k \leq C_M - 1$. Now, for $0 \leq k \leq k_0$, we have $\bar{P}_k = 1 \geq \left(\sum_{j=1}^{M}\gamma_j P_{k+C_j-C_M,j}\right)$. Assume that $\left(\sum_{j=1}^{M}\gamma_j P_{k+C_j-C_M,j}\right) \leq \bar{P}_k$ holds for some $k \geq k_0$. Using induction, we will now show that the inequality must hold for $k+1$. We have

$$\left(\sum_{j=1}^{M}\gamma_j P_{k+1+C_j-C_M,j}\right) \leq \frac{\lambda}{k+(C_1-C_M)+1} \times \left(\sum_{j=1}^{M}\gamma_j P_{k+C_j-C_M,j}\right)^d \leq \frac{\lambda}{k+(C_1-C_M)+1}\bar{P}_k^d = \bar{P}_{k+1}.$$

This completes the proof. ∎

REFERENCES

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

[2] S. T. Maguluri, R. Srikant, and L. Ying, "Stochastic models of load balancing and scheduling in cloud computing clusters," in *Proceedings of IEEE INFOCOM*, 2012.

[3] A. L. Stolyar and Y. Zhong, "A large-scale service system with packing constraints: Minimizing the number of occupied servers," *SIGMETRICS Perform. Eval. Rev.*, vol. 41, no. 1, pp. 41–52, 2013.

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

[5] J. Xu and B. Hajek, "The supermarket game," *Stochastic Systems*, vol. 3, no. 2, pp. 405–441, 2013.

[6] 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.

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

[8] J. B. Martin and Y. M. Suhov, "Fast Jackson networks," *Annals of Applied Probability*, vol. 9, no. 3, pp. 854–870, 1999.

[9] 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.

[10] 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*, to appear.

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

[12] 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.

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

[14] 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.

[15] 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.

[16] V. Marbukh, "Loss circuit switched communication network—performance analysis and dynamic routing," *Queueing Systems Theory Appl.*, vol. 13, no. 1-3, pp. 111–141, 1993.

[17] V. Anantharam, "A mean field limit for a lattice caricature of dynamic routing in circuit switched networks," *Ann. Appl. Probab.*, vol. 1, no. 4, pp. 481–503, 1991.

[18] F. P. Kelly, *Reversibility and Stochastic Networks*. John Wiley and Sons Ltd, 1979.

[19] P. Whittle, "Partial balance and insensitivity," *Journal of Applied Probability*, vol. 22, no. 1, pp. 168–176, 1985.

[20] J. W. Cohen, "On up and down crossings," *Journal of Applied Probability*, vol. 14, pp. 405–410, 1977.

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

[22] K. Deimling, "Ordinary differential equations in Banach spaces," vol. 596 of *Lecture Notes in Mathematics*, Springer Berlin, 1977.