Noname manuscript No.
(will be inserted by the editor)

Choosing among heterogeneous server clouds

A. Karthik · Arpan Mukhopadhyay · Ravi R. Mazumdar

Received: date / Accepted: date

Abstract This paper considers a model of interest in cloud computing applications. We consider a multi-server system consisting of N heterogeneous servers. The servers are categorized into $M(\ll N\ )$ different types according to their service capabilities. Jobs having specific resource requirements arrive at the system according to a Poisson process with rate $N\lambda$. Upon each arrival, a small number of servers is sampled uniformly at random from each server type. The job is then routed to the sampled server with maximum vacancy per server-capacity. If a job cannot obtain the required amount of resources from the server to which it is assigned, then the job is discarded. We analyze the system in the limit as $N \to \infty$. This gives rise to a mean field, which we show has a unique fixed point and is globally attractive. Furthermore, as $N \to \infty$, the servers behave independently. The stationary tail probabilities of server occupancies are obtained from the stationary solution of the mean field. Numerical results suggest that the proposed scheme significantly reduces the average blocking probability compared to static schemes that probabilistically route jobs to servers in proportion to the number of servers of each type. Moreover the reduction in blocking holds even for systems at high load. For the limiting system in statistical equilibrium, our simulation results indicate that the occupancy distribution is insensitive to the holding time distribution and only depends on its mean.

A. Karthik
E-mail: k4ananth@uwaterloo.ca

A. Mukhopadhyay
E-mail: arpan.mukhopadhyay@uwaterloo.ca

Ravi R. Mazumdar
E-mail: mazum@uwaterloo.ca

Dept. of Electrical and Computer Engineering
University of Waterloo
Waterloo ON N2L 3G1, Canada

Keywords Heterogeneous servers · Multirate loss model · Mean field · Propagation of chaos

Mathematics Subject Classification (2000) MSC 60K35 · MSC 90B15 · 60J27 · 60J80

1 Introduction

The cloud computing paradigm has found many applications ranging from data centers and web server farms to next generation wireless communication technologies such as cloud radio access network (C-RAN) [12]. Offloading jobs onto clouds allows users computational flexibility without the need to maintain resources themselves. Many infrastructure-as-service cloud computing systems are now commercially available such as Amazon EC2 [1], Google Cloud [3], Microsoft Azure [6], and IBM Cloud [5]. Such cloud service providers own and operate servers, and sell computational resources to their users in terms of virtual machines (VMs), which are blocks of resource instances such as CPU and memory.

A cloud facility typically consists of thousands of server machines, and hence VMs. A job may require multiple VMs on a particular server. Since a server has only a finite amount of resources, a job arriving at a server may be unable to obtain the required number of VMs for its processing. If the job is delay sensitive, the time to wait for a free resource might be prohibitive. In such a case, the job is blocked, or dropped, and cannot be processed. A prime objective for a cloud service provider, in order to ensure a certain grade of service, is to reduce the the probability that a job request is blocked.

In addition to this customer-centric goal, it is also in the self-interest of the cloud service provider to maximize the efficient use of all its resources. An important way to achieve this is to regulate and equally distribute incoming service requests among all its resources, referred to as load balancing. In fact, major commercial service providers of today such as Amazon EC2, Google Cloud, and Microsoft Azure do implement load balancing [2,4,7]. These commercial services implement load balancing at two different levels. First, by means of a high level user-controlled inferface, and second, at a lower level, user-independent internal implementation. Our study in this paper relates to the latter.

We introduce and study a model for job assignment from the perspective of a primary dispatcher that provides service in assigning an incoming job to one of many commercially available server clouds for processing. Such clouds differ, both qualitatively and quantitatively, in the services they offer [1,3,6, 5]. Each such cloud typically consists of a large number of parallel processing homogeneous servers. Each server has a finite number of VMs that can be simultaneously used for job processing. Different clouds, however, can differ in the total number of VMs they contain as well as the number of VMs they employ per unit job. For example, cloud A might contain more number of VMs in all than another cloud B, whereas the number of VMs that cloud B uses

for a unit job might be smaller than that in cloud A. The primary dispatcher must therefore use this information for efficient job assignment to servers of different types.

It is well known that the blocking probability can be reduced by assigning arriving jobs to less congested servers [39,18,41]. Learning the states of all the servers before joining the least congested server, however, is infeasible due to the overhead incurred in such large systems. In such cases, randomized job assignment schemes offer practical alternatives [38,26,16]. In these schemes, a small, but random subset of server states is sampled and a job is assigned to the sampled server with the least occupancy. In models comprising of identical servers (homogeneous), randomly sampling just two servers and assigning a job to the server having the lower occupancy has been shown to drastically improve delay performance [38,26].

In this paper, we model the cloud system as a heterogeneous loss-system, which is motivated by the VM context, and consider a randomized scheme, referred to as maximum fractional vacancy (MFV) scheme, for job assignment. In this scheme, a small random subset of server states is sampled from each cloud. Jobs are then assigned to clouds whose sampled servers show the smallest occupancy per server-capacity.[1] We then analyse the MFV scheme and study its system performance. The key contribution of this paper is to show that when the number of servers of each type is large we can exploit mean field theory to characterize the performance precisely, and show that propagation of chaos or independence of servers holds in the heterogeneous context too. Moreover we show that the blocking experienced by jobs using such a randomized algorithm is very close to the lower bound on the blocking performance of such systems that can be achieved by any policy.

Randomized job assignment schemes have been primarily studied in the literature for a system consisting of N identical first come first serve (FCFS) servers, which is also referred to as the supermarket model. Most studies consider the so called shortest-queue-d (SQ(d)) scheme in which each job is assigned to the shortest of d randomly chosen queues. For $d \geq 2$, [38] showed, using the theory of operator semigroups, that the equilibrium queue sizes decay doubly exponentially in the limit as the system size increases (as $N \to \infty$). Mitzenmacher in [26,27] derived the same result using an extension of Kurtz's theorem [15]. Chaoticity on path space (or asymptotic independence among queue length processes) was established in [16] using empirical measures on the path space. The results of [38] were generalized to the case of open Jackson networks in [24].

The tradeoff between sampling cost of servers and the expected sojourn time seen by a customer in the supermarket model was studied under a game theoretic framework in [42]. Recently, in [30,29], the SQ(d) scheme was considered for a system of parallel processor sharing servers with heterogeneous service rates. It was shown that, in the heterogeneous setting, random sampling of d servers from the entire system reduces the stability region. However,

[1] In the context of loss systems, server capacity refers to the number of VMs a server has.

it can be recovered using the SQ(d) scheme over a randomly chosen server type.

Early works on mean field limits in routing problems in Erlang loss models include [23,8]. A large-system Erlang loss model for homogeneous servers was studied in [36]. It was shown here that the limiting system behaviour can be characterized using a system of differential equations. Simulation studies were used to show the super-exponential decay of the tail probabilities. Similar results were derived using an asymptotic independence ansatz in [31]. In [17], it was shown that loss models too exhibit propagation of chaos on the path space. Multi-server model for cloud systems with infinite waiting rooms, where jobs are queued till they obtain the required resources, was studied in the heavy traffic regime [22]. The assignment problem has also been in studied in the context of bin-packing problems under various constraints [34,9,25].

1.1 Main results

In this paper, we propose a new randomized scheme for job assignment in the heterogeneous scenario. In this scheme, upon arrival of a job, a small number of servers from each cloud is randomly sampled. The sampled servers are then compared based on their states and the arrival is assigned to the sampled server that has the highest vacancy per unit server-capacity.

This represents a scenario where a primary dispatcher first requests information from each cloud and then routes the job to the server that is likely to have the smallest blocking probability among the sampled servers. We analyze the performance of the proposed scheme in the limit as the system size $N \to \infty$ using the mean field approach. Our analysis shows the following.

- The stationary tail distribution of server occupancies, in a system with a large number of servers, can be characterized by means of a fixed point of a system of differential equations (mean field limit).
- We establish the existence and uniqueness of the equilibrium point of the mean field equations in the space of empirical tail measures. Our proof differs from the earlier works since closed form solutions cannot be obtained.
- We show that propagation of chaos holds at each finite time and also at the equilibrium. In that, we generalize the earlier results on propagation of chaos to systems where exchangeability holds only among servers of the same type.

We outline a method to numerically compute the fixed point of the mean field and hence obtain the blocking probability characterization of the limiting system under stationarity. We show that in the limit as $N \to \infty$, the stationary tail distribution satisfies a balance condition due to Whittle [40] and hence holds for general service times, a result that is confirmed via simulations. We compare the MFV scheme with a state independent scheme for the heterogeneous case, and observe that the MFV scheme clearly outperforms the state independent scheme. Finally we show that even picking a few servers

at each cluster of similar servers gives rise to blocking probabilities that are very close to the lower bound on blocking probabilities for such systems due to any policy, randomized or not, thus, showing the effectiveness and almost optimal behavior of such schemes.

1.2 Organization

The rest of the paper is organized as follows. In Section 2, we describe the system model, the MFV scheme, and our main result. We then analyze the MFV scheme using the mean field in Section 3. In Section 4, we generalize the model of Section 2 by introducing heterogeneous job classes, and show how stationary tail probabilities can be computed in this case, as well. In Section 5, numerical results are presented to benchmark the MFV scheme and to verify the accuracy of the theoretical results derived in the paper. Finally, we conclude the paper in Section 6 with a summary and a discussion on future work.

2 System model and main result

We consider a system comprising of N parallel processing servers which are partitioned into $M(\ll N)$ distinct types of server clouds. Let $\mathcal{J}=\{1,2,\ldots,M\}$ denote the index set of the cloud types, and let γ_j denote the fraction of servers of type j. A cloud of type $j\in\mathcal{J}$ contains $\gamma_j N$ servers, each having the same finite capacity, S_j, of the total number of virtual machines (VMs). A job in a server of type $j\in\mathcal{J}$ engages A_j of the S_j available VMs at the server. The tuple (S_j,A_j) captures the resource and processing capabilities of a type j server.

Jobs arrive at the system according to a Poisson process of rate $N\lambda$. Service times of jobs are exponentially distributed with a mean duration of $1/\mu$ units.[2] Further, service times are independent of one another, and also of the arrival process. Upon its arrival in the system, a job is routed to one of the N servers based on a routing scheme. If a server to which the job is routed has the required amount of resources to serve the job, then the processing of the job starts immediately. Otherwise, the job is discarded, or blocked. Resources used during the processing of a job are released upon its completion. We consider the following routing scheme in the rest of the paper, which we refer to as *maximum fractional vacancy scheme* (MFV).

In this scheme, a job is routed to a server based on its occupancy, that is, the number of current jobs at the server. A local dispatcher at cloud j samples d_j servers, uniformly at random, and conveys the smallest value, v_j,

[2] In this paper, we study the case of homogeneous job requests for which the assumption on mean service duration is valid. For our treatment of the general case of heterogenous job requests, see [28].

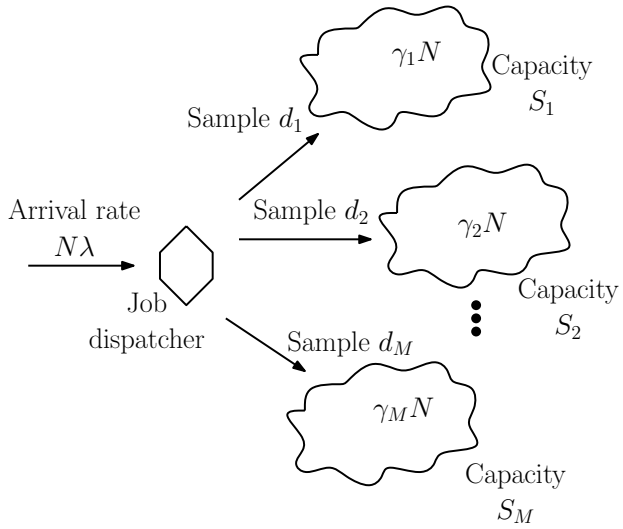

Fig. 1 System consisting of N parallel processor sharing (PS) servers, categorized into M types. There are $\gamma_j N$ servers of type j, each of which has a capacity S_j. Arrivals occur according to a Poisson process with rate $N\lambda$. For each arrival, the job dispatcher samples d_j servers of type j and routes the arrival to one of the sampled servers.

of the sampled server occupancies thus found to the primary dispatcher.[3] The primary dispatcher then routes the job to the cloud that offers the smallest value of v_j/C_j for all $j \in \mathcal{J}$, where C_j is defined as $C_j = \frac{S_j}{A_j}$. Ties are broken by preferring clouds with higher values of C_j. Without loss of generality, we suppose that

$$C_1 \leq C_2 \leq \ldots \leq C_M. \tag{1}$$

We observe that the number C_j is a measure of the total number of jobs that a type j server can simultaneously process. The scheme routes jobs to a sampled server having the least fractional occupancy, or equivalently, the highest fractional vacancy. This is illustrated in Figure 1.[4] We note that data models similar to the above have been employed in the research literature to address issues related to cloud services in various contexts [32,14,21,37].

[3] For notational convenience, we assume that servers are sampled with replacement; the results in the paper remain unchanged even under the assumption that servers are sampled without replacement.

[4] From an implementation viewpoint, each local dispatcher j must periodically communicate v_j to the primary dispatcher. Thus, M positive integers must be communicated regularly to the central dispatcher. Furthermore, a local dispatcher j must update the value C_j at the central dispatcher, but this is required only if there is a change in C_j, which happens less frequently if at all, since these values are stored at the central dispatcher.

2.1 Main result

Let $\mathbb{N}$ denote the set of non-negative integers. For any $k \in \mathbb{N}$ and $i, j \in \mathcal{J}$, we define

$$\lfloor k \rfloor_{ij} = \left\lfloor \frac{C_j}{C_i} k \right\rfloor + 1, \tag{2}$$

$$\lceil k \rceil_{ij} = \left\lceil \frac{C_j}{C_i} k \right\rceil, \tag{3}$$

where $\lfloor x \rfloor$ denotes the greatest integer not exceeding x and $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Define $\theta_j = \lfloor C_j \rfloor$ for $j \in \mathcal{J}$. We are now ready to state the main result of the paper, which characterizes the system when the total number of servers is asymptotically large ($N \to \infty$).

Theorem 1 *Let $P_k^{(j)}(N)$ denote the stationary probability that a server of type j has at least k unfinished jobs, in a system of N servers. Under the MFV scheme, $P_k^{(j)}(N) \to P_k^{(j)}$, as $N \to \infty$, where for each $j \in \mathcal{J}$, $P_k^{(j)}$ satisfies:*

$$P_{k+1}^{(j)} - P_{k+2}^{(j)} = \frac{\lambda}{\mu \gamma_j (k+1)} \left(\left(P_k^{(j)} \right)^{d_j} - \left(P_{k+1}^{(j)} \right)^{d_j} \right) \times \prod_{i=1}^{j-1} \left(P_{\lceil k \rceil_{ji}}^{(i)} \right)^{d_i} \prod_{i=j+1}^{M} \left(P_{\lfloor k \rfloor_{ji}}^{(i)} \right)^{d_i}, \text{ for } 0 \le k \le \theta_j - 1, \tag{4}$$

where $P_0^{(j)} = 1$ and $P_k^{(j)} = 0$ for $k > \theta_j$.

Further, as $N \to \infty$, the stationary server occupancy distributions are independent. The blocking probability of the system is then given by $\prod_{j \in \mathcal{J}} (P_{\theta_j}^{(j)})^{d_j}$.

The following sections of the paper are dedicated to deriving the above characterization. As we shall see in the analysis, (4) governs the state of the system under equilibrium, as $N \to \infty$. Further, we show how $P_k^{(j)}$ can be easily computed. Thus, we can theoretically characterize and study the blocking probability of the system. Since commercial cloud systems typically contain a large number of servers, asymptotic analysis ($N \to \infty$) is natural and relevant for such systems. Moreover, we note that since arrivals at a given server depend on the states (occupancies) of other servers, obtaining the exact time evolution of the system is difficult. Large system analysis also aids analytical tractablility. The limiting system behaviour is known as the mean field limit [26,38,24].

3 The mean field

In this section we analyze the performance of the system with respect to the MFV scheme by studying its mean field. Further, we show that the mean field converges to a unique stationary point that is globally asymptotically stable. We also establish the asymptotic independence, or propogation of chaos, property of the server occupancies. We start with some notation.

3.1 Notation

We define the following real sequence spaces:

$$\mathcal{U}_{N,\theta_j}=\{\{g_n\}_{n\in\mathbb{N}} : 1=g_0\geq g_1\geq\ldots\geq g_{\theta_j}, g_n=0 \text{ for } n>\theta_j, \gamma_j g_n N\in\mathbb{N}\}, \tag{5}$$

$$\mathcal{U}_{\theta_j}=\{\{g_n\}_{n\in\mathbb{N}} : 1=g_0\geq g_1\geq\ldots\geq g_{\theta_j}, g_n=0 \text{ for } n>\theta_j\}. \tag{6}$$

Let $\mathcal{U}_N=\prod_{j\in\mathcal{J}}\mathcal{U}_{N,\theta_j}$ and $\mathcal{U}=\prod_{j\in\mathcal{J}}\mathcal{U}_{\theta_j}$ denote the Cartesian products of $\mathcal{U}_{N,\theta_j}$ and $\mathcal{U}_{\theta_j}$, respectively, over $j\in\mathcal{J}$. For $\mathbf{u},\mathbf{v}\in\mathcal{U}$, define the distance between them as

$$\|\mathbf{u}-\mathbf{v}\|=\sup_{j\in\mathcal{J}}\sup_{n\in\mathbb{N}}\left|\frac{u_n^{(j)}-v_n^{(j)}}{n+1}\right|. \tag{7}$$

Note that $\mathcal{U}$ is closed under the above metric, bounded, and finite-dimensional. Hence, under the metric defined in (7), the space $\mathcal{U}$ is compact (and hence complete and separable).

Let $(H,\mathcal{H},\mu_H)$ be a measure space and $f:H\to\mathbb{R}$ be a μ_H-integrable function. We define duality brackets as $\langle f,\mu_H\rangle=\int f d\mu_H$. We denote the weak convergence (convergence in distribution) of a sequence of probability measures P_n (random variables X_n) to a probability measure P (random variable X) by $P_n\Rightarrow P$ ($X_n\Rightarrow X$).

Let $\mathbf{x},\mathbf{x}',\mathbf{y}\in\mathcal{U}$. We denote $\mathbf{x}\leq\mathbf{x}'$ to mean $x_k^{(j)}\leq x'^{(j)}_k$ for all $j\in\mathcal{J}$ and $k\in\mathbb{N}$. Further, $\mathbf{y}=\min(\mathbf{x},\mathbf{x}')$ and $\mathbf{y}=\max(\mathbf{x},\mathbf{x}')$ means that $y_k^{(j)}=\min(x_k^{(j)},x'^{(j)}_k)$ and $y_k^{(j)}=\max(x_k^{(j)},x'^{(j)}_k)$, respectively, for all $j\in\mathcal{J}$ and $k\in\mathbb{N}$.

3.2 Analysis

We define the process

$$\mathbf{x}_N(t)=\left\{x_{N,n}^{(j)}(t), j\in\mathcal{J}, n\in\mathbb{N}\right\} \text{ for } t\geq 0, \tag{8}$$

where $x_{N,n}^{(j)}(t)$ denotes the fraction of type j servers having at least n unfinished jobs at time t. Thus $\left\{x_{N,n}^{(j)}(t), n\in\mathbb{N}\right\}$ denotes the empirical tail distribution of occupancy of type j servers at time t. Observe that $\mathbf{x}_N(t)\in\mathcal{U}_N$. In the following lemma, we evaluate the generator $\mathbf{A}_N$ associated with the process $\mathbf{x}_N(t)$.

Lemma 1 *Let $\mathbf{g}\in\mathcal{U}_N$ be any state of the process $\mathbf{x}_N(t)$ and $\mathbf{e}(n,j)\in\mathcal{U}_N$ be the unit vector with $e_n^{(j)}=1$ and $e_k^{(i)}=0$ if $i\neq j$ and $k\neq n$. The generator*

$\mathbf{A}_N$ *of the Markov process* $\mathbf{x}_N(t)$ *acting on functions* $f : \mathcal{U}_N \to \mathbb{R}$ *is given by*

$$\begin{aligned}\mathbf{A}_N f(\mathbf{g}) = N\lambda \sum_{j=1}^{M}\sum_{n=1}^{\theta_j}\left[\left(g_{n-1}^{(j)}\right)^{d_j} - \left(g_n^{(j)}\right)^{d_j}\right]\prod_{i=1}^{j-1}\left(g_{\lceil n-1\rceil_{ji}}^{(i)}\right)^{d_i} \\ \times \prod_{i=j+1}^{M}\left(g_{\lfloor n-1\rfloor_{ji}}^{(i)}\right)^{d_i}\left[f(\mathbf{g}+\frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{g})\right] \\ + \mu N\sum_{j=1}^{M}\sum_{n=1}^{\theta_j}\gamma_j n\left(g_n^{(j)} - g_{n+1}^{(j)}\right)\left[f(\mathbf{g}-\frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{g})\right]. \quad (9)\end{aligned}$$

Proof The proof is given in Appendix A.

We now state the main result of this section, which essentially captures the asymptotic behaviour of $\mathbf{x}_N(t)$, as $N \to \infty$. In particular, we employ the generator $\mathbf{A}_N$ to show that the process $\mathbf{x}_N(t)$ converges to a deterministic process as $N \to \infty$.

Theorem 2 *If* $\mathbf{x}_N(0)$ *converges in distribution to some constant* $\mathbf{g} \in \mathcal{U}$ *as* $N \to \infty$*, then the process* $\{\mathbf{x}_N(t)\}_{t\geq 0}$ *converges in distribution to a process* $\{\mathbf{u}(t)\}_{t\geq 0}$*, lying in the space* $\mathcal{U}$ *as* $N \to \infty$*. The process* $\mathbf{u}(t)$ *is given by the solution of the following system of differential equations*

$$\mathbf{u}(0) = \mathbf{g}, \quad (10)$$
$$\dot{\mathbf{u}}(t) = \mathbf{l}(\mathbf{u}(t)), \quad (11)$$

where the mapping $\mathbf{l} : \mathcal{U} \to \left(\mathbb{R}^{\mathbb{N}}\right)^M$ *is given by*

$$l_k^{(j)}(\mathbf{u}) = 0, \text{ for } k = 0 \text{ and } k > \theta_j, j \in \mathcal{J}, \quad (12)$$
$$\begin{aligned}l_k^{(j)}(\mathbf{u}) = \frac{\lambda}{\gamma_j}\left(\left(u_{k-1}^{(j)}\right)^{d_j} - \left(u_k^{(j)}\right)^{d_j}\right)\prod_{i=1}^{j-1}\left(u_{\lceil k-1\rceil_{ji}}^{(i)}\right)^{d_i}\prod_{i=j+1}^{M}\left(u_{\lfloor k-1\rfloor_{ji}}^{(i)}\right)^{d_i} \\ - k\mu\left(u_k^{(j)} - u_{k+1}^{(j)}\right), \text{ for } 1 \leq k \leq \theta_j, j \in \mathcal{J}. \quad (13)\end{aligned}$$

The process $\{\mathbf{u}(t)\}_{t\geq 0}$, defined in the theorem above, is referred to as the *mean field*. We first note that Theorem 2 implicitly assumes that the ordinary differential system (10)-(11) has a unique solution in the space $\mathcal{U}$. In the following proposition, we show that this is indeed the case. To emphasize the dependence of the solution $\mathbf{u}(t)$ on the initial point $\mathbf{g}$, we will often denote $\mathbf{u}(t)$ by $\mathbf{u}(t, \mathbf{g})$.

Proposition 1 *If* $\mathbf{g} \in \mathcal{U}$*, then the system* (10)-(11) *has a unique solution* $\mathbf{u}(t, \mathbf{g}) \in \mathcal{U}$*, for all* $t \geq 0$*.*

Proof The proof is given in Appendix B.

We will prove Theorem 2 using the theory of semigroup operators of Markov processes as in [38,24]. First, we recall the following from [15].

- For the process $\{\mathbf{x}_N(t)\}_{t\geq 0}$, the operator semigroup $\{\mathbf{T}_N(t)\}_{t\geq 0}$ acting on continuous functions $f : \mathcal{U}_N \to \mathbb{R}$ is defined as

$$\mathbf{T}_N(t)f(\mathbf{x}) = \mathbb{E}\left[f(\mathbf{x}_N(t))|\mathbf{x}_N(0)=\mathbf{x}\right] \quad \forall t \geq 0, \mathbf{x} \in \mathcal{U}_N.$$

- For the deterministic process $\{\mathbf{u}(t)\}_{t\geq 0}$, the transition semigroup $\{\mathbf{T}(t)\}_{t\geq 0}$ acting on continuous functions $f : \mathcal{U} \to \mathbb{R}$ is defined as

$$\mathbf{T}(t)f(\mathbf{x}) = f(\mathbf{u}(t,\mathbf{x})) \quad \forall t \geq 0, \mathbf{x} \in \mathcal{U}.$$

In the next proposition, we show that $\mathbf{T}_N(t)$ converges to $\mathbf{T}(t)$ uniformly on bounded intervals. This in conjunction with Theorem 2.11 of Chapter 4 of [15] proves Theorem 2.

Proposition 2 *Let* $\mathbf{u}(t,\mathbf{g})$ *be the solution to the system* (10)-(11)*. For any continuous function* $f : \mathcal{U} \to \mathbb{R}$ *and* $t \geq 0$*,*

$$\lim_{N\to\infty} \sup_{\mathbf{g}\in\mathcal{U}_N} |\mathbf{T}_N(t)f(\mathbf{g}) - f(\mathbf{u}(t,\mathbf{g}))| = 0, \tag{14}$$

and the convergence is uniform in t *within any bounded interval.*

Proof The proof is given in Appendix C.

Remark 1 We note that Theorem 2 implies that if $\mathbf{x}_N(0) \Rightarrow \mathbf{g} \in \mathcal{U}_N$ as $N \to \infty$, then the following weaker convergence results also hold:

1. For each $t \geq 0$, $\mathbf{x}_N(t) \Rightarrow \mathbf{u}(t,\mathbf{g})$ as $N \to \infty$.
2. For each $t \geq 0$, $j \in \mathcal{J}$, and $k \in \mathbb{N}$, $x_{N,k}^{(j)}(t) \Rightarrow u_k^{(j)}(t,\mathbf{g})$ as $N \to \infty$.
3. For each $t \geq 0$, $j \in \mathcal{J}$, and $k \in \mathbb{N}$, $\mathbb{E}\left[x_{N,k}^{(j)}(t)\right] \to u_k^{(j)}(t,\mathbf{g})$ as $N \to \infty$.

The last assertion follows from the first since $x_{N,k}^{(j)}(t)$ is bounded for each N, j, k, t.

3.3 Properties of the mean field

In this section, we characterize some important properties of the mean field. In particular, we show that (10)-(11) has a unique globally asymptotically stable equilibrium point in $\mathcal{U}$.

Let $\mathbf{P}$ denote an equilibrium point of (10)-(11). Then, $\mathbf{P}$ satisfies $\mathbf{l}(\mathbf{P}) = \mathbf{0}$. The following proposition guarantees that there exists an equilibrium point of the system (10)-(11) $\mathcal{U}$.

Theorem 3 *There exists an equilibrium point* $\mathbf{P}$ *of the system* (10)-(11) *in the space* $\mathcal{U}$*.*

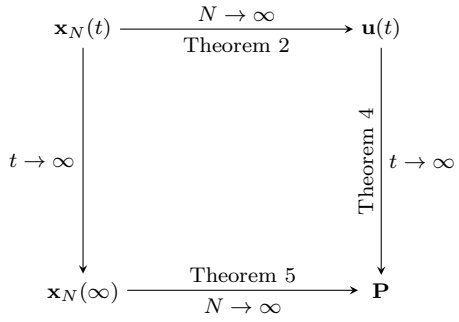

Fig. 2 Commutativity of limits

Proof The proof is given in Appendix D.

The next theorem shows that $\mathbf{P}$ is the unique globally asymptotically stable equilibrium point of the system (10)-(11) in the space $\mathcal{U}$.

Theorem 4

$$\lim_{t\to\infty} \mathbf{u}(t, \mathbf{g}) = \mathbf{P} \in \mathcal{U} \text{ for all } \mathbf{g} \in \mathcal{U}, \tag{15}$$

Hence, $\mathbf{P}$ *is a globally asymptotically stable fixed point of systems* (10)-(11). *Furthermore,* $\mathbf{P}$ *is the only equilibrium point of the above systems in the space* $\mathcal{U}$.

Proof The proof is given in Appendix E.

We now show that the stationary distribution of the process $\mathbf{x}_N(t)$ converges weakly to the Dirac measure concentrated at the unique equilibrium point of the mean field. Let π_N denote the stationary distribution of the process $\mathbf{x}_N(t)$. Since $\mathbf{x}_N(t)$ is positive recurrent, π_N exists and is unique.

Theorem 5 *We have*

$$\pi_N \Rightarrow \delta_{\mathbf{P}}, \text{ as } N \to \infty. \tag{16}$$

Proof The proof is given in Appendix F.

For each fixed N, let $\mathbf{x}_N(\infty)$ be a random variable distributed as π_N. By ergodicity, we have $\mathbf{x}_N(t) \Rightarrow \mathbf{x}_N(\infty)$ as $t \to \infty$. We have so far established that the interchange property indicated in Figure 2 holds. Note that the convergences indicated in the figure are in distribution.

We observe the following simple upper bounds on $P_{\theta_j}^{(j)}$.

Proposition 3 *Let* $\tau_j = \frac{\lambda}{\mu\gamma_j}$. *When* $d_j \geq 2$ *for each* $j \in \mathcal{J}$,

$$P_{\theta_j}^{(j)} \leq \frac{\tau_j^{\frac{d_j^{\theta_j - \lceil \tau_j \rceil + 1} - 1}{d_j - 1}}}{\prod_{k=0}^{\theta_j - \lceil \tau_j \rceil} (\theta_j - k)^{d_j^k}}. \tag{17}$$

When $d_j = 1$ for each $j \in \mathcal{J}$,

$$P_{\theta_j}^{(j)} \leq \frac{\tau_j^{\theta_j - \lceil \tau_j \rceil + 1}}{\prod_{k=0}^{\theta_j - \lceil \tau_j \rceil}(\theta_j - k)}. \tag{18}$$

Proof The proof is given in Appendix G.

Thus, we infer that when at least two servers are sampled from each server type, the tail blocking probability $\prod_{j \in \mathcal{J}} (P_{\theta_j}^{(j)})^{d_j}$ decays at a much faster rate than when a single server is sampled from each cloud. This behaviour is common in power-of-two randomized schemes [38,26,27].

3.4 Propagation of chaos

In this subsection, we focus on the occupancies of a given finite set of servers as $N \to \infty$. We show that as the system size grows the server occupancies become independent of each other. Such independence holds at any finite time and also at the equilibrium, provided that the initial server occupancies satisfy certain assumptions. This is formally known as the *propagation of chaos* [16, 35] or *asymptotic independence property* [11,10] in the literature.

To formally state the results we introduce the following notations. Let $q_N^{(j,k)}(t)$, for $j \in \mathcal{J}$ and $k \in \{1, 2, \ldots, N\gamma_j\}$, denote the occupancy of the k^{th} server of type j at time $t \geq 0$. By $q_N^{(j,k)}(\infty)$ we denote the occupancy of the k^{th} server of type j in equilibrium. Further, let $\chi_{N,n}^{(j)}(t)$, for $j \in \mathcal{J}$ and $n \in \mathbb{N}$, denote the fraction of type j servers having occupancy n at time $t \geq 0$. Define the process $\boldsymbol{\chi}_N(t) = \left\{\chi_{N,n}^{(j)}(t), j \in \mathcal{J}, n \in \mathbb{N}\right\}$. Clearly, $\chi_N^{(j)}(t) = \left\{\chi_{N,n}^{(j)}(t), n \in \mathbb{N}\right\}$ denotes the empirical distribution of occupancies of type j servers and for each n, j, we have $\chi_{N,n}^{(j)}(t) = x_{N,n}^{(j)}(t) - x_{N,(n+1)}^{(j)}(t)$. By $\chi_N^{(j)}(\infty)$ we will denote the empirical distribution occupancies for type j servers in equilibrium. Let the process $\mathbf{Q}(t) = \left\{Q_n^{(j)}(t), j \in \mathcal{J}, n \in \mathbb{N}\right\}$ be defined as $Q_n^{(j)}(t) = u_n^{(j)}(t) - u_{n+1}^{(j)}(t)$, for $t \in [0, \infty]$. Further, we denote by $Q^{(j)}(t)$ the distribution on $\mathbb{N}$ given by $Q^{(j)}(t) = \left\{Q_n^{(j)}, n \in \mathbb{N}\right\}$. We also define the following notion of exchangeable random variables.

Definition 1 Let $\left\{q_N^{(j,k)}, 1 \leq k \leq N\gamma_j, 1 \leq j \leq M\right\}$ denote a collection of N random variables among which $N\gamma_j$ belong to a particular class j and are indexed by k, where $1 \leq k \leq N\gamma_j$. The collection is called intra-class 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 the same class.

Proposition 4 *If the set $\left\{q_N^{(j,k)}(0), 1 \leq k \leq N\gamma_j, 1 \leq j \leq M\right\}$ is intra-class exchangeable and if $\mathbf{x}_N(0) \Rightarrow \mathbf{g} \in \mathcal{U}$ as $N \to \infty$, then the following holds*

1. *For each fix k and $t \in [0,\infty]$, $q_N^{(j,k)}(t) \Rightarrow U^{(j)}(t)$ as $N \to \infty$, where $U^{(j)}(t)$ is a random variable with distribution $Q^{(j)}(t)$.*
2. *Fix positive integers $r_1, r_2, \ldots, r_M$. For each $t \in [0,\infty]$,*

$$\left\{q_N^{(j,k)}, 1 \le k \le r_j, 1 \le j \le M\right\} \Rightarrow \left\{U^{(j,k)}(t), 1 \le k \le r_j, 1 \le j \le M\right\},$$

as $N \to \infty$, where $U^{(j,k)}(t)$, $1 \le k \le r_j, 1 \le j \le M$, are independent random variables with $U^{(j,k)}(t)$ having distribution $Q^{(j)}(t)$ for all $1 \le k \le r_j$.

Proof The proof is given in Appendix H.

Thus, the above proposition shows that in the limiting system server occupancies become independent of each other. It also shows that the stationary occupancy distribution of any type j server is given by $Q^{(j)}(\infty) = \left\{P_n^{(j)} - P_{n+1}^{(j)}, n \in \mathbb{N}\right\}$.

Remark 2 Since the server occupancies are asymptotically independent, the arrival process of jobs at any server in the limiting system is a state dependent Poisson process. The arrival rate of jobs at a server of type $j \in \mathcal{J}$, when its occupancy is k is given by (31) given in Appendix D. This equation can be explained as follows.

Consider a *tagged* type j server in the system and the arrivals that have the tagged server as one of its possible destinations. These arrivals constitute the *potential arrival process* at the tagged server. The probability that the tagged server is selected as a potential destination server for a new arrival is $\binom{N\gamma_j - 1}{d_j - 1} / \binom{N\gamma_j}{d_j} = \frac{d_j}{N\gamma_j}$. Thus, due to Poisson thinning, the potential arrival process to the tagged server is a Poisson process with rate $d_j / N\gamma_j \times N\lambda = \frac{d_j \lambda}{\gamma_j}$.

Consider the potential arrivals at the tagged server when its occupancy is k. This arrival actually joins the tagged server with probability $\frac{1}{x+1}$ when x other servers among the d_j servers of type j have occupancy k, all the d_i servers of type $i < j$ have at least occupancy $\lceil k \rceil_{ij}$, and all the d_i servers of type $i > j$ have at least occupancy $\lfloor k \rfloor_{ij}$. Thus, the total arrival rate $\lambda_k^{(j)}$ can be computed as

$$\lambda_k^{(j)} = \frac{d_j \lambda}{\gamma_j} \sum_{x=0}^{d_j - 1} \frac{1}{x+1} \binom{d_j - 1}{x} \left(P_k^{(j)} - P_{k+1}^{(j)}\right)^x \left(P_{k+1}^{(j)}\right)^{d_j - 1 - x} \times \prod_{i=1}^{j-1} \left(P_{\lceil k \rceil_{ij}}^{(i)}\right)^{d_i} \prod_{i=j+1}^{M} \left(P_{\lfloor k \rfloor_{ij}}^{(i)}\right)^{d_i}, \quad (19)$$

which simplifies to (31).

Hence, the equilibrium arrival rate at a given server depends on the stationary tail probabilities $P_k^{(j)}$, $k \in \mathbb{N}$ and $j \in \mathcal{J}$. The stationary tail probabilities

can in turn be expressed as functions of the arrival rate. Indeed, we note that (33) in Appendix D expresses the local balance equations in equilibrium, which hold under state dependent Poisson arrivals due to Theorems 3.10 and 3.14 of [20].

Remark 3 So far our results have been obtained for exponential job length distributions. Under the hypothesis that asymptotic independence is observed in stationarity even under general job length distributions, the blocking probability of the system seems to depend only on the mean of the job length distribution (insensitivity). In [11] they conjecture that asymptotic independence holds for any local service discipline (rates only depend on current jobs in service) under general service time distributions. This would then imply that the arrivals to individual queues are state dependent Poisson processes that depend only on the state of that queue. This together with the result of Zachary [43] would then imply that insensitivity holds. However lacking a proof of asymptotic independence in the general service time case we can only claim insensitivity as a conjecture.In Section 5, we provide numerical evidence to support this hypothesis.

4 Heterogeneous jobs model

In this section, we relax the assumption that jobs have the same mean duration. This is a more realistic model for jobs; for example, some jobs might be computationally more intensive than other jobs and hence might have higher mean durations. We assume here that a job may belong to one of L types. Let $\mathcal{L} = \{1, \ldots, L\}$ denote the class of job-types. A job of type $l \in \mathcal{L}$ has a mean duration of $1/\mu_l$ units. A job of type $l \in \mathcal{L}$, in a server of type $j \in \mathcal{J}$ engages $A_l^{(j)}$ of the S_j available VMs at the server. The tuple $(S_j, A_l^{(j)})$ captures the resource and processing capabilities of a type j server with respect to a job of type $l \in \mathcal{L}$. Further, we assume that jobs of type $l \in \mathcal{L}$ arrive independently at the system according to a Poisson process of rate $N\lambda_l$. Note that the MFV scheme still compares the fractional total occupancies of servers for job assignment.

We remark that a mean field analysis similar to the one presented in Sec. 3 can be carried out for this model as well, at the cost of more notation. Moreover, it can be shown that the server occupancies are asymptotically independent. Let $p^{(j)}(k)$ denote the stationary probability that a server of type $j \in \mathcal{J}$, has at least k busy VMs. Following the approach of Kaufman-Roberts recursion [19,33] we can show that $p^{(j)}(k)$ statisfy the following recursions for

each $j \in \mathcal{J}$:

$$\sum_{l\in\mathcal{L}} \frac{\lambda_l}{\gamma_j\mu_l}\left[\left(p^{(j)}\left(k-A_l^{(j)}\right)\right)^{d_j} - \left(p^{(j)}\left(k+1-A_l^{(j)}\right)\right)^{d_j}\right] \times \prod_{i=1}^{j-1}\left(p^{(i)}\left(\left\lceil k-A_l^{(j)}\right\rceil_{ji}\right)\right)^{d_i} \prod_{i=j+1}^{M}\left(p^{(i)}\left(\left\lfloor k-A_l^{(j)}\right\rfloor_{ji}\right)\right)^{d_i} = k(p^{(j)}(k)-p^{(j)}(k+1)), 0\le k\le S_j, \quad (20)$$

where $p^{(j)}(k) = 1$ for $k \le 0$, and $p^{(j)}(k) = 0$ for $k > S_j$. From this, the blocking probability $P_{\mathrm{b},l}^{(j)}$ of a job of type $l \in \mathcal{L}$ at a server of type $j \in \mathcal{J}$ is given by $p^{(j)}\left(S_j - A_l^{(j)} + 1\right)$, and the system blocking probability of a type l job is thus calculated as $\prod_{j\in\mathcal{J}}\left(P_{\mathrm{b},l}^{(j)}\right)^{d_j}$.

5 Numerical results

We first present simulation results that verify the asymptotic analysis presented in the paper. In Figure 3, we plot the average blocking probability of the system as a function of the arrival rate λ for $N = 50$ and $N = 10$. Also shown is the blocking probability computed as $\prod_{j\in\mathcal{J}}(P_{\theta_j}^{(j)})^{d_j}$, where $P_{\theta_j}^{(j)}$ is obtained as the solution to the fixed point equation (4). Specifically, the unique fixed point $\mathbf{P}$ was computed by a repeated application of the map presented in Appendix D. For the simulation set up, we consider $\theta_1 = 20$, $\theta_2 = 25$, $A_1 = 2$, $A_2 = 3$, $1/\mu = 1$, $\gamma_1 = \gamma_2 = 0.5$, and $d_1 = d_2 = 2$. We note the match of the results from the fixed point analysis with the simulations for both the values of N. This supports the approach of asymptotic analysis, via the mean field, that we have employed in the paper, and provides a theoretical alternative to characterize the system blocking.

Next, we compare state-dependent job assignment schemes with state-independent job assignment schemes. In Figure 4, we plot the average blocking probability, as a function of the normalized arrival rate, seen by the system under the MFV scheme, which is state-dependent, and two state-independent schemes referred to as static routing-1 and static routing-2, in which jobs are routed to clouds with fixed probabilities. For the set up, we consider $N = 50$, $\theta_1 = 20$, $\theta_2 = 25$, $A_1 = 2$, $A_2 = 3$, $1/\mu = 1$, $\gamma_1 = \gamma_2 = 0.5$, and $d_1 = d_2 = 2$. In the static routing-1 scheme, a job is assigned to a cloud j with probability $\gamma_j\theta_j/(\sum_{i\in\mathcal{J}}\gamma_i\theta_i)$. Note that this scheme is independent of the arrival rate λ. In the static routing-2 scheme, the routing probabilities are set to the optimal values obtained by numerical evauation. We note the routing probabilites in this scheme are indeed a function of the arrival rate; the job dispatcher must to know the arrival rate apriori in order to implement this scheme. We observe the MFV outperforms both the state-independent schemes. The MFV obtains 36.8% better blocking performance than the static routing-2 scheme

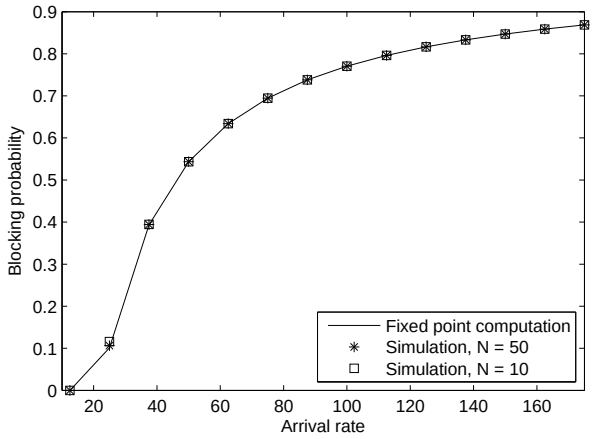

Fig. 3 Blocking probability as a function of the arrival rate.

at the normalized arrival rate of 0.5. This shows the advantages of simple, randomized state-dependent job assignment schemes over the state-independent ones.

Also shown in Figure 4 is a theoretical lower bound on the system blocking probability P_{b} of any job assignment scheme, which is obtained by the following argument. The effective arrival rate of jobs at the system equals $N\lambda(1-P_{\mathrm{b}})$, and by Little's law, the average number of jobs in the system is given by $N\lambda(1-P_{\mathrm{b}})/\mu = \sum_{j\in\mathcal{J}} N\gamma_j \sum_{k=1}^{\theta_j} P_k^{(j)} \leq \sum_{j\in\mathcal{J}} N\gamma_j\theta_j$. Thus, we have $P_{\mathrm{b}} \geq \max\{0, 1-\mu\frac{\sum_{j\in\mathcal{J}}\gamma_j\theta_j}{\lambda}\}$. We denote $\lambda_c = \mu\sum_{j\in\mathcal{J}}\gamma_j\theta_j$ as the critical arrival rate. It is the largest arrival rate at with the lower bound evaluates to 0. We observe from the figure that the MFV scheme has a system blocking probability that is very close to the theoretical lower bound, which hints that the MFV scheme is very close to optimal. We explore this further in Figure 5 and Figure 6 for two different system settings having $\gamma_1 = \gamma_2 = 0.5$ and $d_1 = d_2 = 2$. We observe that in both cases, the margin between the simulation values and the corresponding lower bounds is not greater than 10%. Thus, the MFV scheme closely follows the optimal assignment scheme.

In Figure 7, we consider a system of $N = 100$ servers such that $\theta_1 = 20$, $\theta_2 = 40$, and $\mu = 1/3$. We plot the blocking probability for two different sampling configurations as a function of the arrival rate near the region of critical arrival rate, which in this case is $\lambda_c = 10$. We observe from the plot that more number of samples yields smaller blocking probabilities. This is because of the greater chance of choosing a shorter queue, when more than one server is sampled.

We now numerically investigate the behaviour of the MFV scheme under different job length distributions. In Table 1, blocking probability under the scheme is shown as a function of normalized λ, for the following distributions.

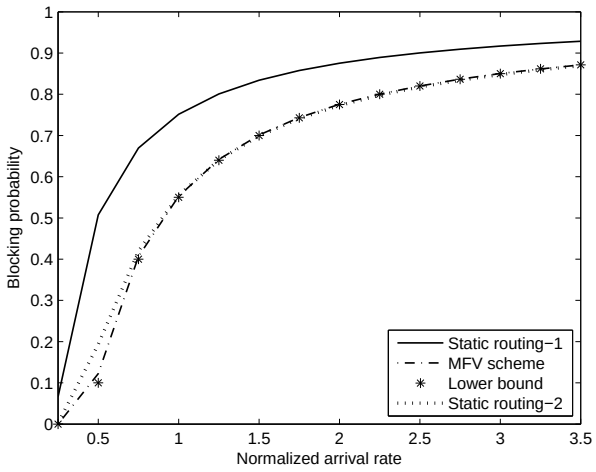

Fig. 4 Blocking probability as a function of the normalized arrival rate.

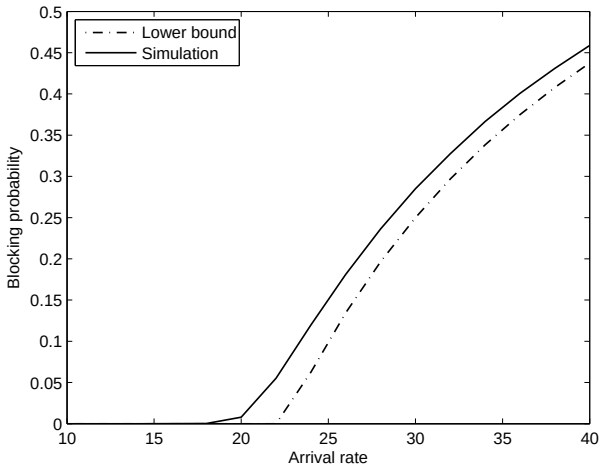

Fig. 5 Comparison with lower bound in the under-loaded regime. $N = 50$, $\theta_1 = 25$, $\theta_2 = 20$.

Table 1 Insensitivity of MFV scheme

λ	Theoretical	Power law	Constant
0.4	0.00002	0.000224	0.000184
0.5	0.12130	0.122268	0.123114
0.6	0.26166	0.261557	0.261806
0.7	0.36470	0.364175	0.364757
0.8	0.44283	0.443080	0.442825
0.9	0.50398	0.504142	0.504038

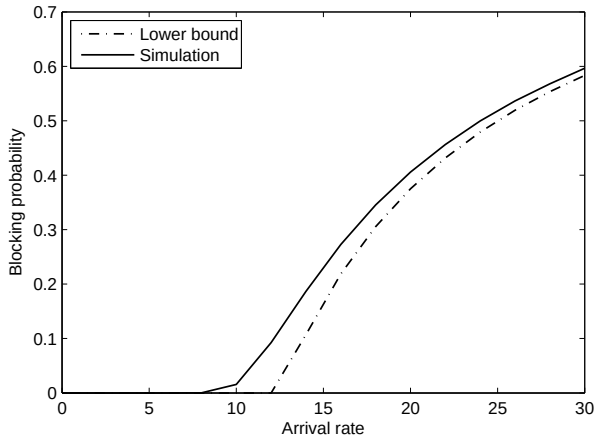

Fig. 6 Comparison with lower bound in the under-loaded regime. $N = 100$, $\theta_1 = 5$, $\theta_2 = 20$.

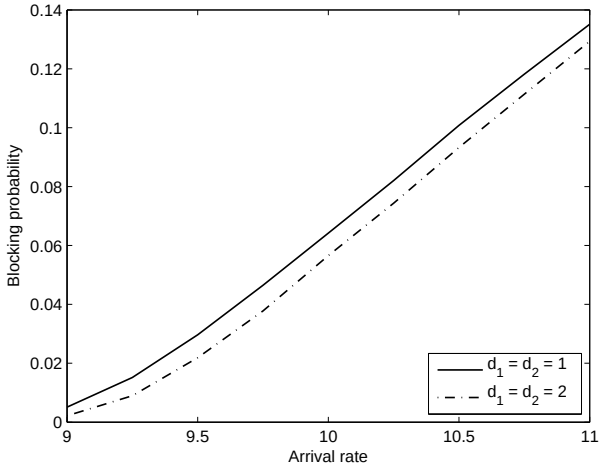

Fig. 7 Comparison with different sampling numbers. $N = 100$, $\theta_1 = 20$, $\theta_2 = 40$, $\mu = 1/3$.

1. *Constant*: We consider job length distribution having the cumulative distribution given by $F(x) = 0$ for $0 \leq x < 1/\mu$, and $F(x) = 1$, otherwise.
2. *Power law*: We consider job length distribution having cumulative distribution function given by $F(x) = 1 - 1/4\mu^2x^2$ for $x \geq \frac{1}{2\mu}$ and $F(x) = 0$, otherwise.

Note that both the above distributions have the same mean $1/\mu$. Same simulation parameters as stated earlier were used. We observe that there is insignificant change in the blocking probability of the system when the job length distribution type is changed, while keeping the mean constant.

6 Conclusions

In this paper, we analyzed the MFV 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. Furthermore, we showed that propagation of chaos holds for heterogeneous case through the requirement of type-based exchangeability. The foregoing results were shown for the exponential job length distribution. An interesting avenue to explore is if these can be replicated for generic job length distributions, as well. Our numerical results hint that asymptotic independence is exhibited even under generic job length distributions.

Further, we observed that the MFV scheme is very nearly optimal in performance. This shows the effectiveness of such simple randomized assignment schemes. We remark that similar state dependent assignment schemes, for instance, assigning to servers with smallest vacancy, or least occupancy, must have similar system behaviour.

A Proof of Lemma 1

We recall the generator $\mathbf{A}_N$ of the semigroup $\{\mathbf{T}_N(t)\}_{t\geq 0}$ acting on functions $f : \mathcal{U}_N \to \mathbb{R}$ is given by $\mathbf{A}_N f(\mathbf{g}) = \sum_{\mathbf{h}\neq\mathbf{g}} q_{\mathbf{gh}} \left(f(\mathbf{h}) - f(\mathbf{g})\right)$, where $q_{\mathbf{gh}}$, with $\mathbf{g}, \mathbf{h} \in \mathcal{U}_N$, denotes the transition rate from state $\mathbf{g}$ to state $\mathbf{h}$ [15].

Consider an arrival at time t that joins a server of type j with exactly $n-1$ unfinished jobs, when the state of the system is $\mathbf{g}$. This corresponds to the transition from state $\mathbf{g}$ to the state $\mathbf{g} + \frac{\mathbf{e}(n,j)}{N\gamma_j}$ since the fraction of type j servers with at least n unfinished jobs increases by $1/N\gamma_j$, whereas the empirical tail occupancies of the other servers remain unchanged. This transition occurs in the MFV scheme only when the following conditions are satisfied:

- Among the d_j sampled servers of type j, at least one has exactly $n-1$ jobs and the rest of them have at least n jobs. Since there are $N\gamma_j g_{n-1}^{(j)}$ and $N\gamma_j g_n^{(j)}$ servers with at least $n-1$ and n jobs, respectively, uniform sampling of the type j servers result in a probability of $\frac{(N\gamma_j g_{n-1}^{(j)})^{d_j} - (N\gamma_j g_n^{(j)})^{d_j}}{(N\gamma_j)^{d_j}} = (g_{n-1}^{(j)})^{d_j} - (g_n^{(j)})^{d_j}$ for this case.
- For each $i < j$, all the d_i sampled servers of type i have fractional occupancies satisfying $i/C_i \geq (n-1)/C_j$; equivalently, sampled servers of type i must have at least $\lceil n-1 \rceil_{ji}$ jobs. Proceeding as above, the probability of this case is $\prod_{i=1}^{j-1} (g_{\lceil n-1 \rceil_{ji}}^{(i)})^{d_i}$.
- For each $i > j$, all the d_i servers of type i have fractional occupancies satisfying $i/C_i > (n-1)/C_j$; equivalently, sampled servers of type i must have at least $\lfloor n-1 \rfloor_{ji}$ jobs. The probability of this case is then calculated as $\prod_{i=j+1}^{M} (g_{\lfloor n-1 \rfloor_{ji}}^{(i)})^{d_i}$.

Thus, the probability with which an arrival joins a type j server with exactly $n-1$ jobs is given by $\left(\left(g_{n-1}^{(j)}\right)^{d_j} - \left(g_n^{(j)}\right)^{d_j}\right) \prod_{i=1}^{j-1} \left(g_{\lceil n-1 \rceil_{ji}}^{(i)}\right)^{d_i} \prod_{i=j+1}^{M} \left(g_{\lfloor n-1 \rfloor_{ji}}^{(i)}\right)^{d_i}$. Since the arrival rate of jobs is $N\lambda$, the rate of the above transition is given by

$$q_{\mathbf{g},\mathbf{g}+\frac{\mathbf{e}(n,j)}{N\gamma_j}} = N\lambda \left[\left(g_{n-1}^{(j)}\right)^{d_j} - \left(g_n^{(j)}\right)^{d_j}\right] \prod_{i=1}^{j-1} \left(g_{\lceil n-1 \rceil_{ji}}^{(i)}\right)^{d_i} \prod_{i=j+1}^{M} \left(g_{\lfloor n-1 \rfloor_{ji}}^{(i)}\right)^{d_i}.$$

Finally, the rate at which jobs depart from a server of type j having exactly n jobs is $\mu n N\gamma_j \left(g_n^{(j)} - g_{n+1}^{(j)}\right)$. This corresponds to the transition from state $\mathbf{g}$ to the state $\mathbf{g} -$

$\frac{\mathbf{e}(n,j)}{N\gamma_j}$ since the fraction of type j servers with at least n unfinished jobs decreases by $1/N\gamma_j$, whereas the empirical tail occupancies of the other servers remain unchanged. The expression (9) now follows from the definition of $\mathbf{A}_N$.

Remark 4 We observe that the foregoing expressions are obtained under the assumption that the $d_j, j \in \mathcal{J}$ servers are sampled with replacement. If, however, they were sampled without replacement, the probability of each of the aforementioned cases changes as follows.

- Servers of the tagged type j: The probability that at least one of them has exactly $n-1$ jobs and the rest of them have at least n jobs becomes $\left(\binom{N\gamma_j g_{n-1}^{(j)}}{d_j} - \binom{N\gamma_j g_n^{(j)}}{d_j}\right)/\binom{N\gamma_j}{d_j}$.
- Servers of type $i < j$: The probability that each of them has $\lceil n-1 \rceil_{ji}$ jobs at least is given by $\prod_{i=1}^{j-1} \binom{N\gamma_i g_{\lceil n-1 \rceil_{ji}}^{(i)}}{d_i}/\binom{N\gamma_i}{d_i}$.
- Servers of type $i > j$: The probability that each of them has $\lfloor n-1 \rfloor_{ji}$ jobs at least is given by $\prod_{i=1}^{j-1} \binom{N\gamma_i g_{\lfloor n-1 \rfloor_{ji}}^{(i)}}{d_i}/\binom{N\gamma_i}{d_i}$.

Consequently, we obtain a different form for the generator of $\mathbf{x}_N(t)$, say $\mathbf{A'}_N$.[5] We note that in the limit as $N \to \infty$, the probabilities in each of the above cases reduce, respectively, to exactly those obtained when servers are sampled with replacement. This fact when used in a parallel development leading up to (30) shows $\lim_{N\to\infty} \mathbf{A'}_N f(\mathbf{g}) = \frac{d}{dt} f(\mathbf{u}(t,\mathbf{g}))|_{t=0}$. Hence, even when servers are sampled without replacement, the same mean field, as in the case of sampling with replacement, is obtained. We consider the latter case for the analysis of MFV for ease of notation; it leads to no change in any of the asymptotic results presented in the paper.

B Proof of Proposition 1

Define $\phi : \mathbb{R} \to [0,1]$ as $\phi(x) = [\min\{x,1\}]_+$, where $[z]_+ = \max\{0,z\}$ and consider the following modification of the system (10)-(11):

$$\mathbf{u}(0) = \mathbf{g}, \tag{21}$$

$$\dot{\mathbf{u}}(t) = \hat{\mathbf{l}}(\mathbf{u}(t)), \tag{22}$$

where the mapping $\hat{\mathbf{l}} : \left(\mathbb{R}^{\mathbb{N}}\right)^M \to \left(\mathbb{R}^{\mathbb{N}}\right)^M$ is given by

$$\hat{l}_k^{(j)}(\mathbf{u}) = 0 \text{ for } k = 0 \text{ and } k > \theta_j, j \in \mathcal{J}, \tag{23}$$

$$\hat{l}_k^{(j)}(\mathbf{u}) = \frac{\lambda}{\gamma_j}\left[\left(\phi(u_{k-1}^{(j)})\right)^{d_j} - \left(\phi(u_k^{(j)})\right)^{d_j}\right]_+ \prod_{i=1}^{j-1}\left(\phi(u_{\lceil k-1 \rceil_{ji}}^{(i)})\right)^{d_i} \prod_{i=j+1}^{M}\left(\phi(u_{\lfloor k-1 \rfloor_{ji}}^{(i)})\right)^{d_i} - k\mu\left[\phi(u_k^{(j)}) - \phi(u_{k+1}^{(j)})\right]_+, \text{ for } 1 \le k \le \theta_j, j \in \mathcal{J}. \tag{24}$$

[5] In fact, for $f : \mathcal{U}_N \to \mathbb{R}$,

$$\mathbf{A'}_N f(\mathbf{g}) = N\lambda \sum_{j=1}^{M}\sum_{n=1}^{\theta_j} \frac{\binom{N\gamma_j g_{n-1}^{(j)}}{d_j} - \binom{N\gamma_j g_n^{(j)}}{d_j}}{\binom{N\gamma_j}{d_j}} \prod_{i=1}^{j-1} \frac{\binom{N\gamma_i g_{\lceil n-1 \rceil_{ji}}^{(i)}}{d_i}}{\binom{N\gamma_i}{d_i}} \prod_{i=1}^{j-1} \frac{\binom{N\gamma_i g_{\lfloor n-1 \rfloor_{ji}}^{(i)}}{d_i}}{\binom{N\gamma_i}{d_i}}$$
$$\times \left[f(\mathbf{g} + \frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{g})\right] + \mu N \sum_{j=1}^{M}\sum_{n=1}^{\theta_j} \gamma_j n \left(g_n^{(j)} - g_{n+1}^{(j)}\right)\left[f(\mathbf{g} - \frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{g})\right].$$

Clearly, the right hand side of (11) and (24) are equal if $\mathbf{u} \in \mathcal{U}$. Therefore, the two systems must have identical solutions in $\mathcal{U}$. Also if $\mathbf{g} \in \mathcal{U}$, then any solution of the modified system remains within $\mathcal{U}$. This is because of the facts that if $u_n^{(j)}(t) = u_{n+1}^{(j)}(t)$ for some j, n, t in (24) then $\hat{l}_n^{(j)}(\mathbf{u}(t)) \geq 0$ and $\hat{l}_{n+1}^{(j)}(\mathbf{u}(t)) \leq 0$, and if $u_n^{(j)}(t) = 0$ for some j, n, t, then $\hat{l}_n^{(j)}(\mathbf{u}(t)) \geq 0$. Hence, to prove the uniqueness of solution of (10)-(11), we need to show that the modified system (21)-(22) has a unique solution in $(\mathbb{R}^{\mathbb{Z}_+})^M$. We now extend the distance metric defined in (7) to the space $(\mathbb{R}^{\mathbb{N}})^M$.

Using the metric defined in (7) and the facts that $|x_+ - y_+| \leq |x - y|$ for any $x, y \in \mathbb{R}$, $\left|a_1 b_1^m - a_2 b_2^m\right| \leq |a_1 - a_2| + m\,|b_1 - b_2|$ for any $a_1, a_2, b_1, b_2 \in [0,1]$, and $|\phi(x) - \phi(y)| \leq |x - y|$ for any $x, y \in \mathbb{R}$ we obtain

$$\|\hat{\mathbf{l}}(\mathbf{u})\| \leq K_1, \tag{25}$$

$$\|\hat{\mathbf{l}}(\mathbf{u}) - \hat{\mathbf{l}}(\mathbf{w})\| \leq K_2 \|\mathbf{u} - \mathbf{w}\|, \tag{26}$$

where $\mathbf{u}, \mathbf{w} \in (\mathbb{R}^{\mathbb{N}})^M$, K_1 and K_2 are constants defined as $K_1 = \frac{\lambda}{\min_{j\in\mathcal{J}} \gamma_j} + \mu\theta_M$ and $K_2 = 4M\lambda \frac{\max_{j\in\mathcal{J}} d_j}{\min_{j\in\mathcal{J}} \gamma_j} + 3\mu\theta_M$. The uniqueness now follows from inequalities (25) and (26) by using Picard's iteration technique since $(\mathbb{R}^{\mathbb{N}})^M$ is complete under the metric defined in (7). □

C Proof of Proposition 2

We prove Proposition 2 by showing that the generators $\mathbf{A}_N$ of the corresponding semigroups converge as $N \to \infty$ to the generator $\mathbf{A}$ of the deterministic process $\mathbf{u}(t, \mathbf{g})$.

First, we show that the solution $\mathbf{u}(t, \mathbf{g})$ of (10)-(11) is smooth with respect to the initial point $\mathbf{g}$ and its partial derivatives are bounded.

Lemma 2 *Let $\mathbf{u}(t, \mathbf{g})$ denote the solution of (10)-(11). For each j, n, j', n', i, k, and $t \geq 0$, the partial derivatives $\frac{\partial \mathbf{u}(t,\mathbf{g})}{\partial g_n^{(j)}}$, $\frac{\partial^2 \mathbf{u}(t,\mathbf{g})}{\partial {g_n^{(j)}}^2}$, and $\frac{\partial^2 \mathbf{u}(t,\mathbf{g})}{\partial g_n^{(j)} \partial g_{n'}^{(j')}}$ exist for $\mathbf{g} \in \mathcal{U}$ and satisfy*

$$\left| \frac{\partial u_k^{(i)}(t, \mathbf{g})}{\partial g_n^{(j)}} \right| \leq \exp(B_1 t) \tag{27}$$

and

$$\left| \frac{\partial^2 u_k^{(i)}(t, \mathbf{g})}{\partial {g_n^{(j)}}^2} \right|, \left| \frac{\partial^2 u_k^{(i)}(t, \mathbf{g})}{\partial g_n^{(j)} \partial g_{n'}^{(j')}} \right| \leq \frac{B_2}{B_1} (\exp(2B_1 t) - \exp(B_1 t)), \tag{28}$$

where $B_1 = \frac{2\lambda \max_{j\in\mathcal{J}} d_j}{\min_{j\in\mathcal{J}} \gamma_j} + 2\mu\theta_M \left(\max_{j\in\mathcal{J}}\right)$, and $B_2 = \frac{2\lambda \left(\max_{j\in\mathcal{J}} d_j\right)^2}{\min_{j\in\mathcal{J}} \gamma_j}$.

Proof The proof follows the same line of arguments as the proof of Lemma 3.2 of [24]. Fix j, n, $\mathbf{g}$ and define $\mathbf{u}'(t) = \partial \mathbf{u}(t, \mathbf{g}) / \partial g_n^{(j)}$. If this partial derivative exists, then $\mathbf{u}'(t)$ must satisfy ${u'}_0^{(i)}(t) = 0$, ${u'}_k^{(i)}(0) = \delta_{ij}\delta_{kn}$. Further, by differentiating (13) we obtain (variable t is supressed for simplifying notation)

$$\frac{d{u'}_k^{(i)}}{dt} = \frac{\lambda}{\gamma_j} d_i \left(\left(u_{k-1}^{(i)}\right)^{d_i - 1} {u'}_{k-1}^{(i)} - \left(u_k^{(i)}\right)^{d_i - 1} {u'}_k^{(i)} \right) \prod_{l=1}^{i-1} \left(u_{\lceil k-1 \rceil_{il}}^{(l)}\right)^{d_l} \times \prod_{l=i+1}^{M} \left(u_{\lfloor k-1 \rfloor_{il}}^{(l)}\right)^{d_l} - \mu k \left({u'}_k^{(i)} - {u'}_{k+1}^{(i)}\right). \tag{29}$$

Conversely, if $\mathbf{u}'(t)$ is a solution of the system above, then it must be the required partial derivative. Using Lemma 3.1 of [24] with $a = B$, $b_0 = 0$, and $c = 1$ and the fact that $\left|u_r^{(k)}\right| \leq 1$ for all k,r it can be shown that $\frac{\partial u_r^{(k)}(t,\mathbf{g})}{\partial g_n^{(j)}}$ exists and is bounded as given by (27).

Similarly, by differentiating (29) again with respect to $g_n^{(j)}$ and $g_{n'}^{(j')}$, we get the system of equations in $\frac{\partial^2 u_r^{(k)}(t,\mathbf{g})}{\partial {g_n^{(j)}}^2}$ and $\frac{\partial^2 u_r^{(k)}(t,\mathbf{g})}{\partial g_n^{(j)} \partial g_{n'}^{(j')}}$, respectively. Lemma 3.1 of [24] can be applied again to this system to show that the second order partial derivatives also exist and are bounded as given by (28).

Next, we show convergence of the generators $\mathbf{A}_N$. Note that the following arguments are along the same lines of the proof of Theorem 2 in [24]. We repeat them here for completeness.

Let L denote the set of real continuous functions on $\mathcal{U}$, and let D denote the set of $f \in L$ such that the partial derivatives $\frac{\partial f(\mathbf{g})}{\partial g_n^{(j)}}$, $\frac{\partial^2 f(\mathbf{g})}{\partial {g_n^{(j)}}^2}$, and $\frac{\partial^2 f(\mathbf{g})}{\partial g_n^{(j)} \partial g_{n'}^{(j')}}$ exist for all $\mathbf{g}, j, j', n, n'$, and are uniformly bounded. Using the norm (7) on $\mathcal{U}$ and the sup-norm on L, we note that D is dense in L. Further, for $f \in D$, we have

$$N\left(f(\mathbf{g} + \frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{g})\right) \to \frac{1}{\gamma_j}\frac{\partial f(\mathbf{g})}{\partial g_j(n)},$$
$$N\gamma_j\left(f(\mathbf{g} - \frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{g})\right) \to -\frac{\partial f(\mathbf{g})}{\partial g_j(n)},$$

uniformly in $\mathbf{g} \in \mathcal{U}$, which upon substitution in (9) of Lemma 1 yields

$$\begin{aligned}
&\lim_{N\to\infty} \mathbf{A}_N f(\mathbf{g}) \\
&= \sum_{j=1}^{M}\sum_{n=1}^{\theta_j} \frac{\partial f(\mathbf{g})}{\partial g_j(n)} \frac{\lambda}{\gamma j}\left[\left(g_{n-1}^{(j)}\right)^{d_j} - \left(g_n^{(j)}\right)^{d_j}\right] \prod_{i=1}^{j-1}\left(g_{\lceil n-1\rceil_{ji}}^{(i)}\right)^{d_i} \prod_{i=j+1}^{M}\left(g_{\lfloor n-1\rfloor_{ji}}^{(i)}\right)^{d_i} \\
&\quad - \sum_{j=1}^{M}\sum_{n=1}^{\theta_j} \frac{\partial f(\mathbf{g})}{\partial g_j(n)} \mu n\left(g_n^{(j)} - g_{n+1}^{(j)}\right), \\
&= \sum_{j=1}^{M}\sum_{n=1}^{\theta_j} \frac{\partial f(\mathbf{g})}{\partial g_j(n)} \left(\frac{\lambda}{\gamma j}\left[\left(g_{n-1}^{(j)}\right)^{d_j} - \left(g_n^{(j)}\right)^{d_j}\right] \prod_{i=1}^{j-1}\left(g_{\lceil n-1\rceil_{ji}}^{(i)}\right)^{d_i} \prod_{i=j+1}^{M}\left(g_{\lfloor n-1\rfloor_{ji}}^{(i)}\right)^{d_i}\right. \\
&\qquad\qquad \left. - \mu n\left(g_n^{(j)} - g_{n+1}^{(j)}\right)\right), \\
&= \frac{d}{dt} f(\mathbf{u}(t,\mathbf{g}))\Big|_{t=0}, \qquad (30)
\end{aligned}$$

uniformly in $\mathbf{g} \in \mathcal{U}$.

We define a semigroup of operators $\mathbf{T}(t), t \geq 0$ in L by setting $\mathbf{T}(t)f(\mathbf{g}) = f(\mathbf{u}(t,\mathbf{g}))$. Observe that the generator $\mathbf{A}$ of this semigroup is given by $\mathbf{A}f(\mathbf{g}) = \lim_{t\downarrow 0} \frac{\mathbf{T}(t)f(\mathbf{g}) - f(\mathbf{g})}{t} = \frac{d}{dt} f(\mathbf{u}(t,\mathbf{g}))|_{t=0}$, which coincides with the RHS of (30).[6] Thus, we obtain $\mathbf{A}_N f \to \mathbf{A}f$, in the sup norm for all $f \in D$.

Next, define $D_0 \subset D$ as those funcitons in D that depend on finitely many variables $g_j(n)$. By the norm defined in (7) we note that D_0 is dense in D, and hence in L. Further, by Lemma 2, $f_0 \in D_0 \implies \mathbf{T}(f_0) \in D$. In addition, we note that the corresponding semigroups $\mathbf{T}_N(t)$ and $\mathbf{T}(t)$ are continuous and contracting in the space of continuous real functions on $\mathcal{U}$. These facts, along with Proposition 3.3 and Theorem 6.1 of [15] gives the desired result.

[6] Recall that the generator $\mathbf{A}$ of the semigroup $\{\mathbf{T}(t)\}_{t\geq 0}$ acting on functions $f : \mathcal{U} \to \mathbb{R}$ having bounded partial derivatives is given by $\mathbf{A}f(\mathbf{g}) = \lim_{t\downarrow 0} \frac{\mathbf{T}(t)f(\mathbf{g}) - f(\mathbf{g})}{t}$ [15].

D Proof of Theorem 3

Consider a point $\mathbf{x} \in \mathcal{U}$. For each $j \in \mathcal{J}$ and $k \in \mathbb{N}$, define

$$\lambda_k^{(j)} = \frac{\lambda}{\gamma_j} \frac{\left(x_k^{(j)}\right)^{d_j} - \left(x_{k+1}^{(j)}\right)^{d_j}}{x_k^{(j)} - x_{k+1}^{(j)}} \prod_{i=1}^{j-1} \left(x_{\lceil k \rceil_{ji}}^{(i)}\right)^{d_i} \prod_{i=j+1}^{M} \left(x_{\lfloor k \rfloor_{ji}}^{(i)}\right)^{d_i}, \text{ for } 0 \leq k \leq \theta_j - 1, \tag{31}$$

$$\lambda_k^{(j)} = 0, \text{ for } k \geq \theta_j. \tag{32}$$

Next, for each $j \in \mathcal{J}$ and $k \in \mathbb{N}$, define

$$\pi_{k+1}^{(j)} = \frac{\lambda_k^{(j)}}{(k+1)\mu} \pi_k^{(j)}, \tag{33}$$

where $\pi_0^{(j)} = \left(1 + \sum_{k=0}^{\theta_j - 1} \frac{\lambda_k^{(j)} \lambda_{k-1}^{(j)} \ldots \lambda_0^{(j)}}{(\mu)^{k+1}(k+1)!}\right)^{-1}$. Finally, for each $j \in \mathcal{J}$ and $k \in \mathbb{N}$, define

$$y_k^{(j)} = \sum_{n \geq k} \pi_n^{(j)}. \tag{34}$$

Clearly, $\mathbf{y} \in \mathcal{U}$. The map $\mathbf{x} \longmapsto \mathbf{y}$, as defined above, is continuous in $\mathcal{U}$. Further, since $\mathcal{U}$ is compact under the metric defined in (7), Brower's fixed point theorem shows that a fixed point $\mathbf{P}$ exists. Substituting $\mathbf{P}$ for $\mathbf{x}$ in (31) and using the fact that $\pi_k^{(j)} = P_k^{(j)} - P_{k+1}^{(j)}$ in the balance equations (33), and comparing (33) with (13), we see that $\mathbf{P}$ satisfies $\mathbf{l}(\mathbf{P}) = \mathbf{0}$. This shows that $\mathbf{P}$ is a fixed point.

E Proof of Theorem 4

We note that for $\mathbf{g}, \mathbf{g}' \in \mathcal{U}_N$ such that $\mathbf{g} \leq \mathbf{g}'$, we have $\mathbf{u}(t, \mathbf{g}) \leq \mathbf{u}(t, \mathbf{g}')$ for all $t \geq 0$. This is because (10)-(11) show that $du_k^{(j)}/dt$ is non-decreasing in $u_n^{(i)}$ for $n \neq k$ and $i \neq j$ [13]. Since this implies that

$$\mathbf{u}(t, \min(\mathbf{g}, \mathbf{P})) \leq \mathbf{u}(t, \mathbf{g}) \leq \mathbf{u}(t, \max(\mathbf{g}, \mathbf{P})),$$

it is sufficient to consider the two cases: $\mathbf{g} \geq \mathbf{P}$ and $\mathbf{g} \leq \mathbf{P}$.

Define $v(t, \mathbf{g}) = \sum_{j \in \mathcal{J}} (\gamma_j / \lambda) \sum_{k=1}^{\theta_j} u_k^{(j)}(t, \mathbf{g})$. We will show that for each $\mathbf{g}$, the quantity $v(t, \mathbf{g})$ is bounded uniformly in t. If $\mathbf{g} \leq \mathbf{P}$, then we have $\mathbf{u}(t, \mathbf{g}) \leq \mathbf{u}(t, \mathbf{P}) = \mathbf{P}$ for all $t \geq 0$. Hence, $v(t, \mathbf{g}) \leq \sum_{j \in \mathcal{J}} (\gamma_j / \lambda) \sum_{k=1}^{\theta_j} P_k^{(j)}$ for all $t \geq 0$. On the other hand, if $\mathbf{g} \geq \mathbf{P}$, then $\mathbf{u}(t, \mathbf{g}) \geq \mathbf{u}(t, \mathbf{P}) = \mathbf{P}$. Adding the set of equations in (11) first over k and then over j, and simplifying further yields:

$$\begin{aligned} \frac{dv(t, \mathbf{g})}{dt} &= 1 - \prod_{j \in \mathcal{J}} (u_{\theta_j}^{(j)}(t, \mathbf{g}))^{d_j} - \sum_{j \in \mathcal{J}} \frac{\mu \gamma_j}{\lambda} \sum_{k=1}^{\theta_j} u_k^{(j)}(t, \mathbf{g}), \\ &\leq 1 - \prod_{j \in \mathcal{J}} (P_{\theta_j}^{(j)})^{d_j} - \sum_{j \in \mathcal{J}} \frac{\mu \gamma_j}{\lambda} \sum_{k=1}^{\theta_j} P_k^{(j)}, \\ &= 0, \end{aligned} \tag{35}$$

where the last equality follows because $\mathbf{P}$ is a fixed point. Thus, we get $v(t, \mathbf{g}) \leq v(0, \mathbf{g})$, for all $t \geq 0$.

Since the derivative of $u_n^{(j)}(t)$ is bounded for all $j \in \mathcal{J}$, the convergence $\mathbf{u}(t,\mathbf{g}) \to \mathbf{P}$ will follow from

$$\int_0^\infty \left(u_k^{(j)}(t,\mathbf{g}) - P_k^{(j)}\right) dt < \infty,\ j \in \mathcal{J}, k \geq 1 \tag{36}$$

in the case $\mathbf{g} \geq \mathbf{P}$, and from

$$\int_0^\infty \left(P_k^{(j)} - u_k^{(j)}(t,\mathbf{g})\right) dt < \infty,\ j \in \mathcal{J}, k \geq 1 \tag{37}$$

in the case $\mathbf{g} \leq \mathbf{P}$. Both the bounds can be shown similarly. We discuss the proof of (36). To prove (36) it is sufficient to show that

$$\int_0^\infty \sum_{j\in\mathcal{J}} \frac{\mu\gamma_j}{\lambda} \sum_{k=1}^{\theta_j} \left(u_k^{(j)}(t,\mathbf{g}) - P_k^{(j)}\right) dt < \infty. \tag{38}$$

Rearranging (35) and using the fact that $\mathbf{P}$ is a fixed point gives

$$\begin{aligned}\sum_{j\in\mathcal{J}} \frac{\mu\gamma_j}{\lambda} \sum_{k=1}^{\theta_j} \left(u_k^{(j)}(t,\mathbf{g}) - P_k^{(j)}\right) &= -\prod_{j\in\mathcal{J}} (u_{\theta_j}^{(j)}(t,\mathbf{g}))^{d_j} + \prod_{j\in\mathcal{J}} (P_{\theta_j}^{(j)})^{d_j} - \frac{dv(t,\mathbf{g})}{dt},\\ &\leq -\frac{dv(t,\mathbf{g})}{dt},\end{aligned}$$

where the last inequality is due to the fact that $\mathbf{u}(t,\mathbf{g}) \geq \mathbf{P}$. Thus,

$$\int_0^\tau \sum_{j\in\mathcal{J}} \frac{\mu\gamma_j}{\lambda} \sum_{k=1}^{\theta_j} \left(u_k^{(j)}(t,\mathbf{g}) - P_k^{(j)}\right) dt \leq v(0,\mathbf{g}) - v(\tau,\mathbf{g}).$$

Since $v(t,\mathbf{g})$ is uniformly bounded in t, the right hand side of the above is bounded for all $\tau \geq 0$. Thus, taking $\tau \to \infty$ in the above gives (38).

F Proof of Theorem 5

We recall that, for a given N, the stationary (invariant) distribution of $\mathbf{x}_N(t) \in \mathcal{U}_N$ is denoted by $\pi_N \in \mathcal{U}_N$. Consider starting the CTMC $\mathbf{x}_N(t)$ according the initial distribution π_N, that is, $\mathbf{x}_N(t) = \pi_N$. Since $\mathcal{U}_N$ is compact, so is the space of probability measures on $\mathcal{U}_N$. Therefore, $\lim_{N\to\infty} \pi_N = \pi$ exists. Further, Theorem 2 shows that $\lim_{N\to\infty} \mathbf{x}_N(t) = \lim_{N\to\infty} \pi_N = \pi$ satisfies (10) and (11). Since $\{\pi_N\}_N$ are all invariant distributions, π trivially satisfies $\mathbf{l}(\pi) = \mathbf{0}$. Hence, π is a stationary point of the system of equations (10) and (11). Using Theorem 4, which shows the unicity of the stationary point, we obtain the desired result.

G Proof of Proposition 3

Since $\mathbf{l}(\mathbf{P}) = 0$, the following must hold for all $j \in \mathcal{J}$:

$$\begin{aligned}P_{k+1}^{(j)} - P_{k+2}^{(j)} = \frac{\lambda}{\mu\gamma_j(k+1)} &\left(\left(P_k^{(j)}\right)^{d_j} - \left(P_{k+1}^{(j)}\right)^{d_j}\right)\\ &\times \prod_{i=1}^{j-1} \left(P_{\lceil k\rceil_{ji}}^{(i)}\right)^{d_i} \prod_{i=j+1}^{M} \left(P_{\lfloor k\rfloor_{ji}}^{(i)}\right)^{d_i},\ \text{for } 0 \leq k \leq \theta_j - 1,\end{aligned} \tag{39}$$

where $P_0^{(j)} = 1$ and $P_k^{(j)} = 0$ for $k > \theta_j$.

For some $j \in \mathcal{J}$, putting $k = \theta_j - 1$ in the above, we get

$$\begin{aligned} P_{\theta_j}^{(j)} &= \frac{\lambda}{\mu\gamma_j\theta_j}\left(\left(P_{\theta_j-1}^{(j)}\right)^{d_j} - \left(P_{\theta_j}^{(j)}\right)^{d_j}\right)\prod_{i=1}^{j-1}\left(P_{\lceil\theta_j-1\rceil_{ji}}^{(i)}\right)^{d_i}\prod_{i=j+1}^{M}\left(P_{\lfloor\theta_j-1\rfloor_{ji}}^{(i)}\right)^{d_i}, \\ &\leq \frac{\lambda}{\mu\gamma_j\theta_j}\left(P_{\theta_j-1}^{(j)}\right)^{d_j}. \end{aligned}$$

Next, for $k = \theta_j - 2$, we have

$$\begin{aligned} P_{\theta_j-1}^{(j)} &= P_{\theta_j}^{(j)} + \frac{\lambda}{\mu\gamma_j(\theta_j-1)}\left(\left(P_{\theta_j-2}^{(j)}\right)^{d_j} - \left(P_{\theta_j-1}^{(j)}\right)^{d_j}\right) \\ &\qquad \times\prod_{i=1}^{j-1}\left(P_{\lceil\theta_j-2\rceil_{ji}}^{(i)}\right)^{d_i}\prod_{i=j+1}^{M}\left(P_{\lfloor\theta_j-2\rfloor_{ji}}^{(i)}\right)^{d_i}, \\ &\leq \frac{\lambda}{\mu\gamma_j\theta_j}\left(P_{\theta_j-1}^{(j)}\right)^{d_j} + \frac{\lambda}{\mu\gamma_j(\theta_j-1)}\left(\left(P_{\theta_j-2}^{(j)}\right)^{d_j} - \left(P_{\theta_j-1}^{(j)}\right)^{d_j}\right), \\ &= \frac{\lambda}{\mu\gamma_j(\theta_j-1)}\left(P_{\theta_j-2}^{(j)}\right)^{d_j} - \frac{\lambda}{\mu\gamma_j}\left(\frac{1}{\theta_j-1} - \frac{1}{\theta_j}\right)\left(P_{\theta_j-1}^{(j)}\right)^{d_j}, \\ &\leq \frac{\lambda}{\mu\gamma_j(\theta_j-1)}\left(P_{\theta_j-2}^{(j)}\right)^{d_j}. \end{aligned}$$

Proceeding in the above manner, we obtain

$$P_k^{(j)} \leq \begin{cases} 1, & \text{for } 0 \leq k < \left\lceil \frac{\lambda}{\mu\gamma_j} \right\rceil, \\ \frac{\lambda}{\mu\gamma_j k}(P_{k-1}^{(j)})^{d_j}, & \text{for } \left\lceil \frac{\lambda}{\mu\gamma_j} \right\rceil \leq k \leq \theta_j \end{cases}.$$

Expanding the above for the case of $P_{\theta_j}^{(j)}$ and simplifying further, we obtain (17). Proceeding on similar lines with $d_j = 1, \forall j \in \mathcal{J}$ in the above, we obtain (18).

H Proof of Proposition 4

Note that the first part of the proposition 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 $M \geq 2$.

Due to the dynamics of the system (under MFV scheme) and the hypothesis of the proposition $\{q_N^{(j,k)}(t), 1 \leq k \leq N\gamma_j, 1 \leq j \leq M\}$ is intra-class exchangeable for all $t \in [0, \infty]$. The hypothesis of the proposition also implies that $\boldsymbol{\chi}_N(t) \Rightarrow \mathbf{Q}(t)$ as $N \to \infty$ for all $t \in [0, \infty]$. Henceforth, we will omit the variable t in our calculations, which hold for all $t \in [0, \infty]$.

For the case $M = 2$, it is sufficient to show that the following convergence holds as $N \to \infty$.

$$\mathbb{E}\left[\prod_{k=1}^{r_1}\phi_k\left(q_N^{(1,k)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_N^{(2,k)}\right)\right] \to \prod_{k=1}^{r_1}\langle\phi_k, Q^{(1)}\rangle\prod_{k=1}^{r_2}\langle\psi_k, Q^{(2)}\rangle \tag{40}$$

for all bounded mappings $\phi_k, \psi_k : \mathbb{N} \to \mathbb{R}_+$. Now we have

$$\left|\mathbb{E}\left[\prod_{k=1}^{r_1}\phi_k\left(q_N^{(1,k)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_N^{(2,k)}\right)\right]-\prod_{k=1}^{r_1}\langle\phi_k,Q^{(1)}\rangle\prod_{k=1}^{r_2}\langle\psi_k,Q^{(2)}\rangle\right|$$
$$\leq\left|\mathbb{E}\left[\prod_{k=1}^{r_1}\phi_k\left(q_N^{(1,k)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_N^{(2,k)}\right)\right]-\mathbb{E}\left[\prod_{k=1}^{r_1}\langle\phi_k,\chi_N^{(1)}\rangle\prod_{k=1}^{r_2}\langle\psi_k,\chi_N^{(2)}\rangle\right]\right|$$
$$+\left|\mathbb{E}\left[\prod_{k=1}^{r_1}\langle\phi_k,\chi_N^{(1)}\rangle\prod_{k=1}^{r_2}\langle\psi_k,\chi_N^{(2)}\rangle\right]-\prod_{k=1}^{r_1}\langle\phi_k,Q^{(1)}\rangle\prod_{k=1}^{r_2}\langle\psi_k,Q^{(2)}\rangle\right|. \quad (41)$$

Note that the second term on the right hand side of the above inequality vanishes as $N\to\infty$ since $\chi_N^{(j)}\Rightarrow Q^{(j)}$ as $N\to\infty$ for $j=1,2$ and $Q^{(1)}$ and $Q^{(2)}$ are constants. Since intra-class exchangeability implies that the permutation of states of servers of the same class does not change their joint distribution, we can average over all the possible states and thus write

$$\mathbb{E}\left[\prod_{k=1}^{r_1}\phi_k\left(q_N^{(1,k)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_N^{(2,k)}\right)\right]=\frac{1}{(N\gamma_1)_{r_1}(N\gamma_2)_{r_2}}$$
$$\times\,\mathbb{E}\left[\sum_{\sigma\in P(r_1,N\gamma_1)}\sum_{\sigma'\in P(r_1,N\gamma_1)}\prod_{k=1}^{r_1}\phi_k\left(q_N^{(1,\sigma(k))}\right)\prod_{k=1}^{r_2}\psi_k\left(q_N^{(2,\sigma'(k))}\right)\right], \quad (42)$$

where $(N)_k=N(N-1)\dots(N-k+1)$, and $P(r,n)$ denotes the set of all permutations of the numbers $\{1,2,\dots,n\}$ taken r at a time. In the following we let $C(r,n)$ denote the set of all r-tuples formed from elements of $\{1,2,\dots,n\}$. Thus, $|P(r,n)|=(n)_r$ and $|C(r,n)|=n^r$. Further, we define $D(r,n)=C(r,n)\backslash P(r,n)$. Proceeding, from the definition of $\chi_N^{(j)}$ we have

$$\mathbb{E}\left[\prod_{k=1}^{r_1}\langle\phi_k,\chi_N^{(1)}\rangle\prod_{k=1}^{r_2}\langle\psi_k,\chi_N^{(2)}\rangle\right]$$
$$=\mathbb{E}\left[\left(\prod_{k=1}^{r_1}\frac{1}{N\gamma_1}\sum_{l=1}^{N\gamma_1}\phi_k\left(q_N^{(1,l)}\right)\right)\left(\prod_{k=1}^{r_2}\frac{1}{N\gamma_2}\sum_{l=1}^{N\gamma_2}\psi_k\left(q_N^{(2,l)}\right)\right)\right],$$
$$=\mathbb{E}\left[\frac{1}{(N\gamma_1)^{r_1}}\sum_{\sigma\in C(r_1,N\gamma_1)}\prod_{k=1}^{r_1}\phi_k\left(q_N^{(1,\sigma(k))}\right)\frac{1}{(N\gamma_2)^{r_2}}\sum_{\sigma'\in C(r_2,N\gamma_2)}\prod_{k=1}^{r_2}\psi_k\left(q_N^{(1,\sigma'(k))}\right)\right],$$
$$=\frac{1}{(N\gamma_1)^{r_1}(N\gamma_2)^{r_2}}\mathbb{E}\left[\sum_{\sigma\in P(r_1,N\gamma_1)}\sum_{\sigma'\in P(r_2,N\gamma_2)}\prod_{k=1}^{r_1}\phi_k\left(q_N^{(1,\sigma(k))}\right)\prod_{k=1}^{r_2}\psi_k\left(q_N^{(1,\sigma'(k))}\right)\right]$$
$$+\frac{1}{(N\gamma_1)^{r_1}(N\gamma_2)^{r_2}}\mathbb{E}\left[\sum_{\sigma\in D(r_1,N\gamma_1)}\sum_{\sigma'\in D(r_2,N\gamma_2)}\prod_{k=1}^{r_1}\phi_k\left(q_N^{(1,\sigma(k))}\right)\prod_{k=1}^{r_2}\psi_k\left(q_N^{(1,\sigma'(k))}\right)\right]. \quad (43)$$

From (42) and (43), we have

$$
\begin{aligned}
&\left|\mathbb{E}\left[\prod_{k=1}^{r_1}\phi_k\left(q_N^{(1,k)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_N^{(2,k)}\right)\right]-\mathbb{E}\left[\prod_{k=1}^{r_1}\langle\phi_k,\chi_N^{(1)}\rangle\prod_{k=1}^{r_2}\langle\psi_k,\chi_N^{(2)}\rangle\right]\right| \\
&=\left|\left(\frac{1}{(N\gamma_1)_{r_1}(N\gamma_2)_{r_2}}-\frac{1}{(N\gamma_1)^{r_1}(N\gamma_2)^{r_2}}\right)\right. \\
&\qquad\times\mathbb{E}\left[\sum_{\sigma\in P(r_1,N\gamma_1)}\sum_{\sigma'\in P(r_2,N\gamma_2)}\prod_{k=1}^{r_1}\phi_k\left(q_N^{(1,\sigma(k))}\right)\prod_{k=1}^{r_2}\psi_k\left(q_N^{(1,\sigma'(k))}\right)\right] \\
&\quad\left.+\frac{1}{(N\gamma_1)^{r_1}(N\gamma_2)^{r_2}}\mathbb{E}\left[\sum_{\sigma\in D(r_1,N\gamma_1)}\sum_{\sigma'\in D(r_2,N\gamma_2)}\prod_{k=1}^{r_1}\phi_k\left(q_N^{(1,\sigma(k))}\right)\prod_{k=1}^{r_2}\psi_k\left(q_N^{(1,\sigma'(k))}\right)\right]\right|, \\
&\le\left|\left(\frac{1}{(N\gamma_1)_{r_1}(N\gamma_2)_{r_2}}-\frac{1}{(N\gamma_1)^{r_1}(N\gamma_2)^{r_2}}\right)|P(r_1,N\gamma_1)||P(r_2,N\gamma_2)|B^{r_1+r_2}\right. \\
&\qquad\left.+\frac{1}{(N\gamma_1)^{r_1}(N\gamma_2)^{r_2}}\left(|C(r_1,N\gamma_1)||C(r_2,N\gamma_2)|-|P(r_1,N\gamma_1)||P(r_2,N\gamma_2)|\right)B^{r_1+r_2}\right|, \\
&\le 2B^{r_1+r_2}\left(1-\frac{(N\gamma_1)_{r_1}(N\gamma_2)_{r_2}}{(N\gamma_1)^{r_1}(N\gamma_2)^{r_2}}\right), \\
&\to 0 \text{ as } N\to\infty,
\end{aligned}
$$

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

References

1. Amazon EC2. http://aws.amazon.com/ec2/
2. Amazon EC2 load balancing. http://docs.aws.amazon.com/ElasticLoadBalancing/latest/DeveloperGuide/elastic-load-balancing.html
3. Google cloud. https://cloud.google.com/
4. Google Cloud load balancing. https://cloud.google.com/compute/docs/load-balancing-and-autoscaling
5. IBM Cloud. http://www.ibm.com/cloud-computing/us/en/
6. Microsoft azure. http://www.microsoft.com/windowsazure/
7. Microsoft Azure load balancing. https://azure.microsoft.com/en-in/documentation/articles/load-balancer-overview
8. Anantharam, V.: A mean field limit for a lattice caricature of dynamic routing in circuit switched networks. The Annals of Applied Probability pp. 481–503 (1991)
9. Bansal, N., Caprara, A., Sviridenko, M.: A new approximation method for set covering problems, with applications to multidimensional bin packing. SIAM J. Comput pp. 1256–1278 (2009)
10. Bramson, M., Lu, Y., Prabhakar, B.: Randomized load balancing with general service time distributions. In: Proceedings of ACM SIGMETRICS, pp. 275–286 (2010)
11. Bramson, M., Lu, Y., Prabhakar, B.: Asymptotic independence of queues under randomized load balancing. Queueing Systems **71**(3), 247–292 (2012)
12. Cai, Y., Yu, F., Bu, S.: Cloud radio access networks (C-RAN) in mobile cloud computing systems. In: Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on, pp. 369–374 (2014)
13. Deimling, K.: Ordinary differential equations in Banach spaces. Lecture notes in mathematics. Springer-Verlag (1977)
14. Deng, W., Liu, F., Jin, H., Li, B., Li, D.: Harnessing renewable energy in cloud datacenters: opportunities and challenges. Network, IEEE **28**(1), 48–55 (2014)
15. Ethier, S.N., Kurtz, T.G.: Markov Processes: Characterization and Convergence. John Wiley and Sons Ltd (1985)

16. Graham, C.: Chaoticity on path space for a queueing network with selection of shortest queue among several. Journal of Applied Probability **37**(1), 198–211 (2000)
17. Graham, C., Mlard, S.: Propagation of chaos for a fully connected loss network with alternate routing. Stochastic Processes and their Applications **44**(1), 159–180 (1993)
18. Gupta, V., Balter, M.H., Sigman, K., Whitt, W.: Analysis of join-the-shortest-queue routing for web server farms. Performance Evaluation **64**(9-12), 1062–1081 (2007)
19. Kaufman, J.: Blocking in a shared resource environment. Communications, IEEE Transactions on **29**(10), 1474–1481 (1981)
20. Kelly, F.P.: Reversibility and Stochastic Networks. John Wiley and Sons Ltd (1979)
21. Maguluri, S.T., Srikant, R.: Scheduling jobs with unknown duration in clouds. Networking, IEEE/ACM Transactions on **22**(6), 1938–1951 (2014)
22. Maguluri, S.T., Srikant, R., Ying, L.: Stochastic models of load balancing and scheduling in cloud computing clusters. In: Proceedings of IEEE INFOCOM (2012)
23. Marbukh, V.: Loss circuit switched communication network-performance analysis and dynamic routing. Queueing systems **13**(1-3), 111–141 (1993)
24. Martin, J.B., Suhov, Y.M.: Fast jackson networks. Annals of Applied Probability **9**(3), 854–870 (1999)
25. Meng, X., Pappas, V., Zhang, L.: Improving the scalability of data center networks with traffic-aware virtual machine placement. In: Proceedings of the 29th Conference on Information Communications, INFOCOM'10, pp. 1154–1162 (2010)
26. Mitzenmacher, M.: The power of two choices in randomized load balancing. PhD Thesis, Berkeley (1996)
27. Mitzenmacher, M.: The power of two choices in randomized load balancing. IEEE Transactions on Parallel and Distributed Systems **12**(10), 1094–1104 (2001)
28. Mukhopadhyay, A., Karthik, A., Mazumdar, R.R., Guillemin, F.: Mean field and propagation of chaos in multi-class heterogeneous loss models. Performance Evaluation **91**, 117–131 (2015)
29. Mukhopadhyay, A., Mazumdar, R.R.: Analysis of load balancing in large heterogeneous processor sharing systems. ArXiv:1311.5806 [cs.DC]
30. Mukhopadhyay, A., Mazumdar, R.R.: Rate-based randomized routing in large heterogeneous processor sharing systems. In: 26th International Teletraffic Congress (ITC 24), pp. 1–9 (2014)
31. Q. Xie , X. Dong, Y. Lu, R. Srikant: Power of d choices for large-scale bin packing: A loss model. In: ACM Sigmetrics 2015, to appear
32. Rastegarfar, H., Rusch, L.A., Leon-Garcia, A.: Optical load-balancing tradeoffs in wavelength-routing cloud data centers. Journal of Optical Communications and Networking **7**(4), 286–300 (2015)
33. Roberts, J.W.: A service system with heterogeneous user requirements. In: Performance of Data Communications systems and their applications, vol. 29, pp. 423–431 (1981)
34. Stolyar, A.L., Zhong, Y.: A large-scale service system with packing constraints: Minimizing the number of occupied servers. SIGMETRICS Perform. Eval. Rev. **41**(1), 41–52 (2013)
35. Sznitman, A.S.: Propagation of chaos. In: École d'été de probabilites de Saint-Flour XIX - 1989, *Lecture Notes in Mathematics*, vol. 1464, pp. 165–251. Springer Berlin Heidelberg (1991)
36. Turner, S.R.E.: The effect of increasing routing choice on resource pooling. Probability in the Engineering and Informational Sciences **12**, 109–124 (1998)
37. Ungureanu, V., Melamed, B., Katehakis, M.: Effective load balancing for cluster-based servers employing job preemption. Performance Evaluation **65**(8), 606–622 (2008)
38. Vvedenskaya, N.D., Dobrushin, R.L., Karpelevich, F.I.: Queueing system with selection of the shortest of two queues: an asymptotic approach. Problems of Information Transmission **32**(1), 20–34 (1996)
39. Weber, R.R.: On the optimal assignment of customers to parallel servers. Journal of Applied Probability **15**, 406–413 (1978)
40. Whittle, P.: Partial balance and insensitivity. Journal of Applied Probability **22**(1), 168–176 (1985)
41. Winston, W.: Optimality of the shortest line discipline. Journal of Applied Probability **14**(1), 181–189 (1977)

42. Xu, J., Hajek, B.: The supermarket game. Stochastic Systems **3**(2), 405–441 (2013)
43. Zachary, S.: A note on insensitivity in stochastic networks. J. Appl. Probab. **44**(1), 238–248 (2007)