Mean Field and Propagation of Chaos in Multi-Class Heterogeneous Loss Models

Arpan Mukhopadhyay, A. Karthik, Ravi R. Mazumdar, Fabrice Guillemin

Abstract

We consider a system consisting of N parallel servers, where jobs with different resource requirements arrive and are assigned to the servers for processing. Each server has a finite resource capacity and therefore can serve only a finite number of jobs at a time. We assume that different servers have different resource capacities. A job is accepted for processing only if the resource requested by the job is 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 or unused resource among d randomly sampled servers. We consider the system in the limit where both the number of servers and the arrival rates of jobs are scaled by a large factor. This gives rise to a mean field analysis. We show that in the limiting system the servers behave independently – a property termed as *propagation of chaos*. Stationary tail probabilities of server occupancies are obtained from the stationary solution of the mean field which is shown to be unique and asymptotically stable. 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 as compared to static schemes that probabilistically route jobs to servers independently of their states.

I. Introduction

Consider jobs with different resource requirements, arriving at a multi-server system consisting of a large number of parallel servers. Each server has a finite resource capacity and therefore

A. Mukhopadhyay, and R. R. Mazumdar are with the department of Electrical and Computer Engineering, University of Waterloo, Canada. Email: {arpan.mukhopadhyay, mazum}@uwaterloo.ca

A. Karthik is with Qualcomm, Bangalore, India. Email: k4ananth@uwaterloo.ca

F. Guillemin is with Orange Labs, France. Email: fabrice.guillemin@orange.com

A preliminary version of this work was presented at the 27th International Teletraffic Congress (ITC 27), Belgium, 2015.

can process only a finite number of jobs at a time. Different servers are assumed to have different capacities. Upon arrival, each job is routed/assigned to a server where it is either accepted or blocked depending on the availability of resource requested by the job at the server. 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], [2]. 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. To meet different user demands the cloud operator allows users to choose from various types of VM's which differ by the amounts of resources they hold. We model situations [3], [4], where there is one bottleneck resource (e.g. memory). Hence, in our case the different VM's correspond to different amounts of the same resource. Each user requests a VM of specific type (e.g. large or small) 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 resource at the server. 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 average blocking probability of users 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 or VM requests to servers. In particular, we consider a model where each server has the capacity to serve a finite number VM requests at a time and different servers can have different capacities. 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 [5], [6] . 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 [2] where each server can process only a finite number of jobs at a time.

We consider a randomized scheme to assign VM requests to servers based on random sampling of d servers from the entire system. We show that assigning each request to the server having the maximum vacancy amongst the $d \geq 2$ sampled servers yields dramatic reduction in the blocking probability as compared to that in static assignment schemes ($d = 1$) where assignments are made independent of server states. Although, comparing the states of all servers would ideally

result in the best performance, such comparison involve high communication overhead due to large size of server farms. Moreover, we show that sampling too many servers is unnecessary since sampling only a few servers results in nearly optimal in terms of minimizing the average blocking probability of requests.

Related Literature: The routing scheme that we consider is a loss model analog of the *power-of-d* scheme considered in [7]–[9] for first-come-first-serve (FCFS) queues and in [10] for heterogeneous processor sharing (PS) servers. In the power-of-d scheme, each incoming job is assigned to the server having the least number of unfinished jobs among d servers, sampled randomly at the arrival instant of the job. Turner [11], [12] studied this scheme for a system of Erlang servers having infinite capacities in the large system limit. It was shown that in the large system limit, the system behavior can be characterized by a *mean field limit*, which satisfies a system of differential equations. The resulting tail distribution of server occupancies was shown to have a fast rate of decay even for small values of d. However, the existence and uniqueness of the equilibrium point of the mean field were not shown explicitly.

A recent work by Xie *et al.* [4] analyzed an Erlang loss system with identical (homogeneous) servers under the power-of-d scheme using mean field techniques. In [4], the existence, uniqueness, and (global) asymptotic stability of the equilibrium point of the mean field were established for the homogeneous (servers) case with single class of customers. For the homogeneous (servers) case with multiple class of customers, the paper derived a recursive relationship among the tail probabilities of the number of occupied resource units at each server in the limiting system. In the current paper, we generalize these results to the scenario where the servers have heterogeneous capacities. For this scenario, we establish independence of the servers in the limiting system through a milder requirement of *intra-type exchangeability* since exchangeability among servers of different capacities does not hold in this case. Such asymptotic independence of servers in the large system limit, also known as the *propagation of chaos* property, was studied earlier in the context of alternative routing by Graham and Méléard [13], [14] where the independence among servers was established on the path space of the processes of interest.

Contributions: In this paper, we analyze the performance of the power-of-d scheme for a multi-server system consisting of heterogeneous loss servers in the presence of multiple job classes using mean field techniques. The mean field limit, given by the solution of a system of ordinary differential equations, describes the behavior of the system when the arrival rates of different

classes of jobs and the number of servers in the system are scaled by the same large factor. We establish the existence of a unique equilibrium point of the mean field. The equilibrium point is shown to be globally asymptotically stable when all the arriving jobs belong to a single class. For the multi-class case, asymptotic stability is established for initial conditions sufficiently close to the equilibrium point. Furthermore, it is shown that in the limiting system any finite set of servers behave independently of each other (asymptotic independence) and the stationary distribution of states of a given server can be obtained from the unique equilibrium point of the mean field. To show asymptotic independence in the heterogeneous case, we introduce the notion of *intra-type exchangeability* where exchangeability holds only among servers having the same capacity. A bound on the rate of decay of stationary tail distribution of server occupancies similar to that obtained in [4] is found in the heterogeneous case. Numerical results show that the power-of-d scheme significantly reduces the average blocking probability of jobs in the system as compared to the static routing schemes and is nearly optimal in terms minimizing the average blocking probability.

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 consisting of N parallel servers, where jobs or VM requests arrive and request necessary resource for processing. We assume that there is only one bottleneck resource (e.g. memory) since such situations occur often in practice [3]. Each server holds a finite amount of the resource. The servers are categorized into M different types based on their resource capacities. Let $\mathcal{J} = \{1, 2, \ldots, M\}$ be the index set of server types. A server of type $j \in \mathcal{J}$ is assumed to hold C_j units of resource. Without loss of generality, we assume that the capacities are ordered as follows: $C_1 \leq C_2 \leq \ldots \leq C_M$. 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 or VM requests are categorized into L classes depending on their resource requirements.

Class $l \in \mathcal{L} = \{1, 2, \ldots, L\}$ VM requests require $A_l \geq 0$ units of resource and are assumed to arrive at the system according to a Poisson process with rate $N\lambda_l$ independent of the other classes. We denote by $\underline{A} = (A_1, A_2, \ldots, A_L)$ the L-dimensional vector of resource requirements. Upon arrival, a job is routed to one of the N servers according to the following routing scheme:

Power-of-d scheme: Upon arrival of each VM request, $d \geq 2$ *potential destination* servers are sampled uniformly at random from the set of N servers. The actual *destination* server for the arriving request is then chosen to be the server having the maximum vacancy or the maximum units of unused resource among the sampled servers. Ties among (sampled) 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 index (highest capacity). For example, if there are two type-j servers and one type $i < j$ server having the maximum vacancy among the sampled set of d servers, then any one of the two type j servers is chosen to be the destination server with probability $1/2$.

The destination server accepts the VM request if the necessary resource is available to process the request. If accepted, processing of the request begins immediately. Otherwise, the request is discarded or blocked and lost. We say that a server is in state $\underline{n} = (n_1, n_2, \ldots, n_L)$ when, for each $l \in \mathcal{L}$, there are n_l jobs of class l in progress at the server. Clearly, the set of admissible states for a type $j \in \mathcal{J}$ server is given by $\mathcal{S}_j = \left\{\underline{n} \in \mathbb{Z}_+^L : \underline{n} \cdot \underline{A} \leq C_j\right\}$, where $\mathbb{Z}_+$ denotes the set of all non-negative integers and $\underline{n} \cdot \underline{A} \triangleq \sum_{l=1}^{L} n_l A_l$. We define the set of blocking states $\mathcal{B}_j^{(l)}$ for class $l \in \mathcal{L}$ jobs at a server of type $j \in \mathcal{J}$ as the set of states in $\mathcal{S}_j$ for which the vacancy or the number of unused resource units is less than A_l, i.e., $\mathcal{B}_j^{(l)} = \left\{\underline{n} \in \mathbb{Z}_+^L : C_j - A_l < \underline{n} \cdot \underline{A} \leq C_j\right\}$. Clearly, a class-$l$ VM request is blocked only when all the d potential destination servers are in the blocking states for the arriving request. The service times of accepted job requests are assumed to be independent and exponentially distributed random variables with mean 1. The service times of jobs are assumed to be also independent of the inter-arrival times of the jobs. The resource held by a request is released immediately upon the completion of its service.

III. Main results

In this section, we state (without proof) the main results of this paper. 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. Without loss of much generality we assume that C_j and A_l are non-negative integers for each $j \in \mathcal{J}$ and $l \in \mathcal{L}$.

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

$$k\left(P_{k,j} - P_{k+1,j}\right) = \sum_{l=1}^{L} \frac{A_l \lambda_l}{\gamma_j} \left[\left(\sum_{i=1}^{j} \gamma_i P_{k-A_l+C_i-C_j,i} \; + \sum_{i=j+1}^{M} \gamma_i P_{k-A_l+C_i-C_j+1,i} \right)^d \right.$$
$$\left. - \left(\sum_{i=1}^{j-1} \gamma_i P_{k-A_l+C_i-C_j,i} + \sum_{i=j}^{M} \gamma_i P_{k-A_l+C_i-C_j+1,i} \right)^d \right], \quad (1)$$

for $1 \leq k \leq C_j$, with $P_{k,j} = 1$ 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 and their stationary occupancy distributions are insensitive to the service time distribution.

Remark *1 (Propagation of chaos):* We note that for finite system size N, the states of the servers are not independent of each other since at every arrival instant states of some randomly sampled servers are compared. However, in the limiting system ($N \to \infty$) the servers become mutually independent. This is known as the *propagation of chaos* or *asymptotic independence* property.

Remark *2 (Blocking probability):* Using the independence of servers stated above and the probabilities $P_{k,j}$ found by solving (1) we can compute the blocking probability $P_{\text{blocking}}^{(l)}$ of class-l requests in the limiting system as follows: The stationary probability that a server of type j is in the blocking state for a class l job is $P_{C_j-A_l+1,j}$ and the probability that it is sampled at an arrival instant is γ_j. Thus the total probability that a randomly sampled server is in the blocking state for class-l requests is $\sum_{j\in\mathcal{J}} \gamma_j P_{C_j-A_l+1,j}$. Since the servers in the limiting system are mutually independent, the probability that the class-l arrival is blocked is given by $P_{\text{blocking}}^{(l)} = \left(\sum_{j\in\mathcal{J}} \gamma_j P_{C_j-A_l+1,j}\right)^d$.

A lower bound on blocking probability: The blocking probability of requests averaged over all classes obtained from any given routing scheme can be lower bounded as follows. For an arbitrary job routing scheme, the blocking probability averaged over all classes is given by $P_{\text{blocking}}^{\text{avg}} = \frac{\sum_{l\in\mathcal{L}} \lambda_l P_{\text{blocking}}^{(l)}}{\sum_{l\in\mathcal{L}} \lambda_l}$, where $P_{\text{blocking}}^{(l)}$ denotes the blocking probability of class-l jobs under that scheme.

By Little's law the average number customer in the system is given by $(1-P^{\text{avg}}_{\text{blocking}})N\sum_{l\in\mathcal{L}}\lambda_l$. Now if type j servers can accommodate a maximum of $B_j = \max_{\underline{n}\in\mathcal{S}_j}\left(\sum_{l\in\mathcal{L}} n_l\right)$ jobs of all classes combined, then the average number of jobs in the entire system is upper bounded by $N\sum_{j\in\mathcal{J}}\gamma_j B_j$. We therefore have the following lower bound on the average blocking probability:

$$P^{\text{avg}}_{\text{blocking}} \geq \left(1-\frac{\sum_{j\in\mathcal{J}}\gamma_j B_j}{\sum_{l\in\mathcal{L}}\lambda_l}\right)_+ = \left(1-\frac{\lambda_{\text{crit}}}{\lambda}\right)_+, \tag{2}$$

where $\lambda_{\text{crit}} = \sum_{j\in\mathcal{J}}\gamma_j B_j$ is the critical load on the system, $\lambda = \sum_{l\in\mathcal{L}}\lambda_l$ is the total arrival rate per server, and $(w)_+ = \max(0,w)$. In Section VI, we compare the blocking probability for the power-of-d scheme with the lower bound derived above. We conclude that the power-of-d scheme is nearly optimal in terms of minimizing blocking probability even for small d.

If we specialize (1) to the case where only a single class of jobs ($L=1$) requiring one unit of resource from all servers ($A=1$) arrive according to a Poisson process with rate $N\lambda$, then (1) simplifies to the following recursive relation:

$$P_{k,j} - P_{k+1,j} = \frac{\lambda}{\gamma_j k}\left[\left(\sum_{i=1}^{j}\gamma_i P_{k-1+C_i-C_j,i} + \sum_{i=j+1}^{M}\gamma_i P_{k+C_i-C_j,i}\right)^d - \left(\sum_{i=1}^{j-1}\gamma_i P_{k-1+C_i-C_j,i} + \sum_{i=j}^{M}\gamma_i P_{k+C_i-C_j,i}\right)^d\right]. \tag{3}$$

In this above, $P_{k,j}$ is the stationary probability that there are at least k jobs in progress at a type j server in the limiting system. Using (3), explicit upper bounds on the rate of decay of the tail probabilities $P_{k,j}$, $k\in\{0,1,2,\ldots,C_j\}$, $j\in\mathcal{J}$ can be obtained. This is done in the following proposition whose proof is similar to the proof of Theorem 2 of [4] for the homogeneous loss model.

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

$$\bar{P}_k = \frac{\lambda^{d^{k-k_0}-1}}{(\lfloor\lambda\rfloor+k-k_0)(\lfloor\lambda\rfloor+k-k_0-1)^d\ldots(\lfloor\lambda\rfloor+1)^{d^{k-k_0-1}}}, \tag{4}$$

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 for the single class case where each job requires unit resource we have

$\sum_{j=1}^{M} \gamma_j P_{k+C_j-C_M,j} \leq \bar{P}_k$ for $0 \leq k \leq C_M$. In particular, the average user blocking probability $P_{\text{blocking}}^{\text{avg}} = \left(\sum_{j\in\mathcal{J}} \gamma_j P_{C_j,j}\right)^d \leq \bar{P}_{C_M}^d$.

Proof: The proof is given in Appendix A. ■

The above proposition 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

In this section we provide detailed proofs of the main results discussed in Section III. An exact characterization of the stationary regime is difficult for finite N due to the fact that the arrival rate at a given server depends on the states of other servers. However, it is possible to analyze the system in the limit as the system size $N \to \infty$. Such a limit is known as the mean field limit [7], [15], [16]. We first introduce the notation and mathematical framework for the analysis.

Notations: The unit vector in $\mathbb{Z}_+^L$ with one in the r^{th} position is denoted by $\underline{e}_r$ and by $\underline{e}$ we denote the L-dimensional vector of all ones, i.e., $\underline{e} = (1, 1, \ldots, 1)$. Further, for each $j \in \mathcal{J}$ we denote the space of probability distributions on $\mathcal{S}_j$ by:

$$\mathcal{U}_j = \left\{ (g_{\underline{n}})_{\underline{n}\in\mathcal{S}_j} : g_{\underline{n}} \geq 0 \text{ for all } \underline{n} \in \mathcal{S}_j, \sum_{\underline{n}\in\mathcal{S}_j} g_{\underline{n}} = 1 \right\}.$$

The set of empirical probability distributions on $\mathcal{S}_j$ when the system size is N is denoted by $\mathcal{U}_j^{(N)}$, i.e., $\mathcal{U}_j^{(N)} = \left\{ (g_{\underline{n}})_{\underline{n}\in\mathcal{S}_j} \in \mathcal{U}_j : N\gamma_j g_{\underline{n}} \in \mathbb{Z}_+ \right\}$. 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 $j \in \mathcal{J}$. A point in the space $\mathcal{U}$ (or in $\mathcal{U}^{(N)}$) is denoted by $\mathbf{u} = (u_{n,j}, n \in \mathcal{S}_j, j \in \mathcal{J})$ with the understanding that for each $j \in \mathcal{J}$ the collection $(u_{\underline{n},j}, n \in \mathcal{S}_j)$ belongs to $\mathcal{U}_j$ ($\mathcal{U}_j^{(N)}$). Since for each $j \in \mathcal{J}$, the set $\mathcal{S}_j$ is finite and $\mathcal{U}_j$ denotes the space of probability measures on $\mathcal{S}_j$, the space $\mathcal{U} = \prod_{j\in\mathcal{J}} \mathcal{U}_j$ is convex and compact under any norm and all norms defined on this space are equivalent. We shall be using both the Euclidean norm and the norm induced by the metric ρ (defined below) according to our convenience: for any two points $\mathbf{u}, \mathbf{w}$ in $\mathcal{U}$ define $\rho(\mathbf{u}, \mathbf{w}) = \sup_{j\in\mathcal{J}} \sup_{\underline{n}\in\mathcal{S}_j} \frac{\left|u_{\underline{n},j} - w_{\underline{n},j}\right|}{(\underline{n}\cdot\underline{e})+1}$.

For a measure space $(H, \mathcal{H}, \mu)$ and a μ-integrable function $f : H \to \mathbb{R}$, we define duality brackets as $\langle f, \mu \rangle = \int f d\mu$. 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: For each $t \geq 0$ and $\underline{n} \in \mathcal{S}_j$, let $x^{(N)}_{\underline{n},j}(t)$ denote the fraction of type-j servers in state $\underline{n}$ at time t. We define $\mathbf{x}^{(N)}(t) = \left(x^{(N)}_{\underline{n},j}(t), n \in \mathcal{S}_j, j \in \mathcal{J}\right)$. Clearly, $\mathbf{x}^{(N)}(\cdot)$ is a Markov process with state space $\mathcal{U}^{(N)}$, i.e., for each $j \in \mathcal{J}$ the collection $\left(x^{(N)}_{\underline{n},j}(t), n \in \mathcal{S}_j\right)$ denotes the empirical probability distribution of states of type-j servers at time t. The generator $\mathbf{A}^{(N)}$ of the Markov process $\mathbf{x}^{(N)}(\cdot)$ acting on functions $\varphi : \mathcal{U}^{(N)} \to \mathbb{R}$ is given by $\mathbf{A}^{(N)}\varphi(\mathbf{u}) = \sum_{\mathbf{h} \neq \mathbf{u}} r(\mathbf{u} \to \mathbf{v})(\varphi(\mathbf{v}) - \varphi(\mathbf{u}))$, where $r(\mathbf{u} \to \mathbf{v})$ denotes the transition rate from state $\mathbf{u} \in \mathcal{U}^{(N)}$ to state $\mathbf{v} \in \mathcal{U}^{(N)}$. In the following lemma, we provide the expression for the generator.

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

$$\mathbf{A}^{(N)}\varphi(\mathbf{u}) = N \sum_{j \in \mathcal{J}} \sum_{\underline{n} \in \mathcal{S}_j} \sum_{l \in \mathcal{L}} \left[\lambda_l \frac{F(\underline{n} - \underline{e}_l, j, \mathbf{u})}{E(\underline{n} - \underline{e}_l, j, j, \mathbf{u})} \gamma_j u_{\underline{n} - \underline{e}_l, j} \left(\varphi(\mathbf{u} - \frac{\mathbf{e}(\underline{n} - \underline{e}_l, j)}{N\gamma_j} + \frac{\mathbf{e}(\underline{n}, j)}{N\gamma_j}) \right.\right.$$
$$\left.\left. - \varphi(\mathbf{u}) \right) + \gamma_j u_{\underline{n},j} n_l \left(\varphi(\mathbf{u} + \frac{\mathbf{e}(\underline{n} - \underline{e}_l, j)}{N\gamma_j} - \frac{\mathbf{e}(\underline{n}, j)}{N\gamma_j}) - \varphi(\mathbf{u}) \right) \right] I_{\underline{n} - \underline{e}_l \in \mathcal{S}_j}, \quad (5)$$

where I_A denotes the indicator function on the set A and for $\underline{n} \in \mathcal{S}_j$ and $i \in \mathcal{J}$ we have

$$E(\underline{n}, i, j, \mathbf{u}) = \gamma_i \sum_{\substack{\underline{n}' \in \mathcal{S}_i : \\ \underline{n}' \cdot \underline{A} = \underline{n} \cdot \underline{A} + C_i - C_j}} u_{\underline{n}', i} \quad (6)$$

$$G(\underline{n}, i, j, \mathbf{u}) = \gamma_i \sum_{\substack{\underline{n}' \in \mathcal{S}_i : \\ \underline{n}' \cdot \underline{A} > \underline{n} \cdot \underline{A} + C_i - C_j}} u_{\underline{n}', i} \quad (7)$$

$$GE(\underline{n}, i, j, \mathbf{u}) = G(\underline{n}, i, j, \mathbf{u}) + E(\underline{n}, i, j, \mathbf{u}). \quad (8)$$

and

$$F(\underline{n},j,\mathbf{u}) = \left(\sum_{i=1}^{j} GE(\underline{n},i,j,\mathbf{u}) + \sum_{i=j+1}^{M} G(\underline{n},i,j,\mathbf{u})\right)^d - \left(\sum_{i=1}^{j-1} GE(\underline{n},i,j,\mathbf{u}) + \sum_{i=j}^{M} G(\underline{n},i,j,\mathbf{u})\right)^d. \quad (9)$$

Proof: The proof is given in Appendix B. ■

Using the generator $\mathbf{A}^{(N)}$, we now show that that as $N \to \infty$, the sequence of processes $\left(\mathbf{x}^{(N)}(\cdot)\right)_N$ converges to a deterministic process.

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 $N \to \infty$. Further, $\mathbf{x}(t,\mathbf{u}_0)$ is given by the solution of the following system of differential equations*

$$\mathbf{x}(0,\mathbf{u}_0) = \mathbf{u}_0, \quad (10)$$

$$\dot{\mathbf{x}}(t,\mathbf{u}_0) = \mathbf{h}(\mathbf{x}(t,\mathbf{u}_0)), \quad (11)$$

where the mapping $\mathbf{h}$ is given by

$$h_{\underline{n},j}(\mathbf{x}) = \sum_{l\in\mathcal{L}} \left[\lambda_l \frac{F(\underline{n}-\underline{e}_l,j,\mathbf{x})}{E(\underline{n}-\underline{e}_l,j,j,\mathbf{x})} x_{\underline{n}-\underline{e}_l,j} - n_l x_{\underline{n},j}\right] I_{\underline{n}-\underline{e}_l\in\mathcal{S}_j} - \left[\lambda_l \frac{F(\underline{n},j,\mathbf{x})}{E(\underline{n},j,j,\mathbf{x})} x_{\underline{n},j} - (n_l+1) x_{\underline{n}+\underline{e}_l,j}\right] I_{\underline{n}+\underline{e}_l\in\mathcal{S}_j}, \quad (12)$$

for each $\underline{n} \in \mathcal{S}_j$ and $j \in \mathcal{J}$, where I_A denotes indicator function of the set A and $E(\underline{n},i,j,\mathbf{u})$, $F((\underline{n},j,\mathbf{u}))$ are as defined in Lemma 1.

Proof: The proof is given in Appendix C ■

The process $\mathbf{x}(\cdot,\mathbf{u}_0)$, describing the evolution of empirical distribution of server sates in the limiting system, is referred to as the *mean field*. It is important to characterize the properties of points $\boldsymbol{\pi} = (\pi_{\underline{n},j}, \underline{n} \in \mathcal{S}_j, j \in \mathcal{J})$ satisfying $\mathbf{h}(\boldsymbol{\pi}) = \mathbf{0}$. Such points are called the *equilibrium points* of the mean field since $\mathbf{x}(t,\boldsymbol{\pi}) = \boldsymbol{\pi}$ for all $t \geq 0$. Hence, by definition a stationary point $\boldsymbol{\pi}$ of the system (10)-(11) must satisfy $h_{\underline{n},j}(\boldsymbol{\pi}) = 0$, i.e., (from (12))

$$\sum_{l\in\mathcal{L}}\left[\lambda_l\frac{F(\underline{n}-\underline{e}_l,j,\boldsymbol{\pi})}{E(\underline{n}-\underline{e}_l,j,j,\boldsymbol{\pi})}\pi_{\underline{n}-\underline{e}_l,j}-n_l\pi_{\underline{n},j}\right]I_{\underline{n}-\underline{e}_l\in\mathcal{S}_j}$$
$$=\sum_{l\in\mathcal{L}}\left[\lambda_l\frac{F(\underline{n},j,\boldsymbol{\pi})}{E(\underline{n},j,j,\boldsymbol{\pi})}\pi_{\underline{n},j}-(n_l+1)\pi_{\underline{n}+\underline{e}_l,j}\right]I_{\underline{n}+\underline{e}_l\in\mathcal{S}_j}, \quad (13)$$

for $\underline{n}\in\mathcal{S}_j$ and $j\in\mathcal{J}$. In the next theorem, we show that there exists an equilibrium point $\boldsymbol{\pi}$ of the mean field $\mathbf{x}(\cdot)$ in the space $\mathcal{U}$.

Theorem *2:* There exists an equilibrium point $\boldsymbol{\pi}$ of the system (10)-(11) in the space $\mathcal{U}$.

Proof: Consider a point $\mathbf{x}\in\mathcal{U}$. For each $j\in\mathcal{J}$, $l\in\mathcal{L}$ and $\underline{n}\in\mathcal{S}_j$, define

$$\lambda^{(l)}_{\underline{n},j}(\mathbf{x})=\lambda_l\frac{F(\underline{n},j,\mathbf{x})}{E(\underline{n},j,j,\mathbf{x})}>0. \quad (14)$$

Next, we define the quantities $y_{\underline{n},j}(\mathbf{x})$, $j\in\mathcal{J}$, $\underline{n}\in\mathcal{S}_j$ as the solution to the following system of linear equations

$$\sum_{l\in\mathcal{L}}\left[\lambda^{(l)}_{\underline{n}-\underline{e}_l,j}(\mathbf{x})y_{\underline{n}-\underline{e}_l,j}(\mathbf{x})-n_l y_{\underline{n},j}(\mathbf{x})\right]I_{\underline{n}-\underline{e}_l\in\mathcal{S}_j}$$
$$=\sum_{l\in\mathcal{L}}\left[\lambda^{(l)}_{\underline{n},j}(\mathbf{x})y_{\underline{n},j}(\mathbf{x})-(n_l+1)y_{\underline{n}+\underline{e}_l,j}(\mathbf{x})\right]I_{\underline{n}+\underline{e}_l\in\mathcal{S}_j}, \text{ for } j\in\mathcal{J} \text{ and } \underline{n}\in\mathcal{S}_j \quad (15)$$

and $\sum_{\underline{n}\in\mathcal{S}_j}y_{\underline{n},j}(\mathbf{x})=1$ for each $j\in\mathcal{J}$. Clearly, the solution $\mathbf{y}(\mathbf{x})=(y_{\underline{n},j}(\mathbf{x}),\underline{n}\in\mathcal{S}_j,j\in\mathcal{J})$ to the above set of linear equations satisfies

$$\lambda^{(l)}_{\underline{n}-\underline{e}_l,j}(\mathbf{x})y_{\underline{n}-\underline{e}_l,j}(\mathbf{x})I_{\underline{n}-\underline{e}_l\in\mathcal{S}_j}=n_l y_{\underline{n},j}(\mathbf{x}) \text{ for all } \underline{n}\in\mathcal{S}_j, j\in\mathcal{J}. \quad (16)$$

The above equations (which imply that $y_{\underline{n},j}(\mathbf{x})$ has the same sign for each $\underline{n}\in\mathcal{S}_j$ and $j\in\mathcal{J}$) together with $\sum_{\underline{n}\in\mathcal{S}_j}y_{\underline{n},j}(\mathbf{x})=1$ imply that $\mathbf{y}(\mathbf{x})\in\mathcal{U}$ for all $\mathbf{x}\in\mathcal{U}$. Furthermore, the map $\mathbf{x}\mapsto\mathbf{y}(\mathbf{x})$, as defined above, is continuous on the space $\mathcal{U}$. Since $\mathcal{U}$ is convex and compact, Brouwer's fixed point theorem guarantees the existence of a fixed point of the map $\mathbf{x}\mapsto\mathbf{y}(\mathbf{x})$. From (15), it is clear that any fixed point $\boldsymbol{\pi}$ of the map $\mathbf{x}\mapsto\mathbf{y}(\mathbf{x})$ satisfies (**??**) and hence is an equilibrium point of the mean field $\mathbf{x}(\cdot)$. This proves the existence of an equilibrium point $\boldsymbol{\pi}$ in $\mathcal{U}$ of the mean field $\mathbf{x}(\cdot)$. ■

Remark *3:* We note that for each $\mathbf{x} \in \mathcal{U}$, the mapping $\mathbf{x} \mapsto \mathbf{y}(\mathbf{x})$ satisfies (16). Hence, the fixed point $\boldsymbol{\pi}$ of the map satisfies

$$\lambda^{(l)}_{\underline{n}-\underline{e}_l,j}(\boldsymbol{\pi})\pi_{\underline{n}-\underline{e}_l,j} I_{\underline{n}-\underline{e}_l \in \mathcal{S}_j} = n_l \pi_{\underline{n},j} \text{ for all } \underline{n} \in \mathcal{S}_j, j \in \mathcal{J}. \tag{17}$$

We shall use this fact later in Section V.

We now focus on the single class case ($L = 1$) and show that the equilibrium point in this case is *unique* and *globally asymptotically stable*, i.e., for any $\mathbf{x}(0) \in \mathcal{U}$ we have $\lim_{t\to\infty} \mathbf{x}(t) = \boldsymbol{\pi}$, where $\boldsymbol{\pi}$ denotes the equilibrium point of the mean field.

Theorem *3:* For the single class case ($L = 1$), the mean field $\mathbf{x}(\cdot)$ has a unique, globally asymptotically stable equilibrium point $\boldsymbol{\pi} \in \mathcal{U}$.

Proof: We note that the uniqueness of the equilibrium point follows (by the uniqueness of limit) if one can show that any equilibrium point $\boldsymbol{\pi}$ is globally asymptotically stable. We now proceed to show the global asymptotic stability of any equilibrium of the mean field $\mathbf{x}(\cdot)$ for the single class case.

For the single class case, we assume without loss of generality that all incoming jobs require one unit of resource and they arrive according to a Poisson process with rate $N\lambda$, i.e., $A_1 = 1$ and $\lambda_1 = \lambda$. Hence, $\mathcal{S}_j = \{0, 1, \ldots, C_j\}$, $\mathcal{U}_j = \left\{(g_n)_{n\in\mathcal{S}_j} : g_n \geq 0, \sum_{n\in\mathcal{S}_j} g_n = 1\right\}$, $\mathcal{U} = \prod_{j\in\mathcal{J}} \mathcal{U}_j$. In this case, the mean field $\mathbf{x}(\cdot) = (x_{n,j}(\cdot), n \in \mathcal{S}_j, j \in \mathcal{J}) \in \mathcal{U}$ satisfies the following system of differential equations (from (10)-(11))

$$\mathbf{x}(0) = \mathbf{u}_0, \tag{18}$$

$$\dot{\mathbf{x}}(t) = \mathbf{h}(\mathbf{x}(t)), \tag{19}$$

where $\mathbf{h}$ is given by (specializing (12) to the case under consideration)

$$h_{n,j}(\mathbf{x}) = \left[\frac{\lambda}{\gamma_j} F(n-1, j, \mathbf{x}) - n x_{n,j}\right] I_{1\leq n\leq C_j} - \left[\frac{\lambda}{\gamma_j} F(n, j, \mathbf{x}) - (n+1) x_{n+1,j}\right] I_{0\leq n\leq C_j - 1}. \tag{20}$$

The mean field can equivalently be expressed in terms of the tail sums $\tilde{x}_{k,j}(t) = \sum_{n=k}^{C_j} x_{n,j}(t)$, $k \in \mathcal{S}_j$, $j \in \mathcal{J}$. We define $\tilde{\mathbf{x}}(t) = (\tilde{x}_{k,j}(t), k \in \mathcal{S}_j, j \in \mathcal{J})$. Hence, from (19) and (20) we have

$$\tilde{\mathbf{x}}(0) = \tilde{\mathbf{u}}_0, \tag{21}$$

$$\dot{\tilde{\mathbf{x}}}(t) = \tilde{\mathbf{h}}(\tilde{\mathbf{x}}(t)), \tag{22}$$

where the mapping $\tilde{\mathbf{h}} = \left(\tilde{h}_{k,j}, k \in \mathcal{S}_j, j \in \mathcal{J}\right)$ is given by $\tilde{h}_{0,j}(\tilde{\mathbf{x}}) = 0$ for all $j \in \mathcal{J}$ and for $1 \leq k \leq C_j$

$$\tilde{h}_{k,j}(\tilde{\mathbf{x}}) = \frac{\lambda}{\gamma_j}\left[\left(\sum_{i=1}^{j}\gamma_i\tilde{x}_{k-1+C_i-C_j,i} + \sum_{i=j+1}^{M}\gamma_i\tilde{x}_{k+C_i-C_j,i}\right)^d - \left(\sum_{i=1}^{j-1}\gamma_i\tilde{x}_{k-1+C_i-C_j,i} + \sum_{i=j}^{M}\gamma_i\tilde{x}_{k+C_i-C_j,i}\right)^d\right] - k\left(\tilde{x}_{k,j} - \tilde{x}_{k+1,j}\right) \tag{23}$$

We say that $\tilde{\mathbf{u}} \leq \tilde{\mathbf{u}}'$ if $\tilde{u}_{k,j} \leq \tilde{u}'_{k,j}$ for all $k \in \mathcal{S}_j$ and $j \in \mathcal{J}$. We first prove the following monotonicity property of the mean field with respect to the initial condition.

Lemma *2:* If $\tilde{\mathbf{u}}_0 \leq \tilde{\mathbf{u}}'_0$ then $\tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}_0) \leq \tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}'_0)$ for all $t \geq 0$.

Proof: Clearly, the right hand side of (23) is non-decreasing in $\tilde{x}_{n,i}(t)$ for all $(n,i) \neq (k,j)$. Hence, (23) defines a quasi-monotone system of differential equations. The proof of the lemma now follows directly from pages 70-74 of [17]. ■

We now define $z(t, \tilde{\mathbf{u}}_0) = \sum_{j\in\mathcal{J}}\gamma_j\sum_{k=1}^{C_j}\tilde{x}_{k,j}(t, \tilde{\mathbf{u}}_0)$. Clearly, $z(t, \tilde{\mathbf{u}}_0)$ denotes the mean number of customers in the limiting system at time t when the initial state is $\tilde{\mathbf{u}}_0$. Using (23) we obtain

$$\frac{dz(t, \tilde{\mathbf{u}}_0)}{dt} = \lambda\left(1 - \left(\sum_{j\in\mathcal{J}}\gamma_j\tilde{x}_{C_j,j}(t, \tilde{\mathbf{u}}_0)\right)^d\right) - z(t, \tilde{\mathbf{u}}_0). \tag{24}$$

Let $\tilde{\boldsymbol{\pi}}$ be an equilibrium point of the process $\tilde{\mathbf{x}}(\cdot)$. Hence, from (24) we have

$$\lambda\left(1 - \left(\sum_{j\in\mathcal{J}}\gamma_j\tilde{\pi}_{C_j,j}\right)^d\right) = z(t, \boldsymbol{\pi}) = \sum_{j\mathcal{J}}\gamma_j\sum_{k=1}^{C_j}\tilde{\pi}_{k,j} \tag{25}$$

Now, from Lemma 2 we have

$$\tilde{\mathbf{x}}(t, \min(\tilde{\mathbf{u}}_0, \tilde{\boldsymbol{\pi}})) \leq \tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}_0) \leq \tilde{\mathbf{x}}(t, \max(\tilde{\mathbf{u}}_0, \tilde{\boldsymbol{\pi}})), \tag{26}$$

where the maximum and the minimum are taken component-wise. Hence, to establish $\lim_{t\to\infty} \tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}_0) = \tilde{\boldsymbol{\pi}}$ for all $\tilde{\mathbf{u}}_0$, it is sufficient to show that the convergence holds for $\tilde{\mathbf{u}}_0 \geq \tilde{\boldsymbol{\pi}}$ and for $\tilde{\mathbf{u}}_0 \leq \tilde{\boldsymbol{\pi}}$.

To show $\tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}_0) \to \tilde{\boldsymbol{\pi}}$ for $\tilde{\mathbf{u}}_0 \geq \tilde{\boldsymbol{\pi}}$ it is sufficient to show that

$$\int_0^\infty (\tilde{x}_{n,j}(t, \tilde{\mathbf{u}}_0) - \tilde{\pi}_{n,j})\, dt < \infty, \text{ for all } j \in \mathcal{J}, 1 \leq n \leq C_j. \tag{27}$$

Similarly for $\tilde{\mathbf{u}}_0 \leq \tilde{\boldsymbol{\pi}}$ the convergence $\tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}_0) \to \tilde{\boldsymbol{\pi}}$ will follow if we can show that

$$\int_0^\infty (\tilde{\pi}_{n,j} - \tilde{x}_{n,j}(t, \tilde{\mathbf{u}}_0))\, dt < \infty, \text{ for all } j \in \mathcal{J}, 1 \leq n \leq C_j \tag{28}$$

We discuss the proof for $\tilde{\mathbf{u}}_0 \geq \tilde{\boldsymbol{\pi}}$. The proof for $\tilde{\mathbf{u}}_0 \leq \tilde{\boldsymbol{\pi}}$ follows similarly.

For $\tilde{\mathbf{u}}_0 \geq \tilde{\boldsymbol{\pi}}$ we have using Lemma 2 that $\tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}_0) \geq \tilde{\boldsymbol{\pi}}$ for all $t \geq 0$. Hence, to prove (27) it is sufficient to show that $\int_0^\infty \left(z(t, \tilde{\mathbf{u}}_0) - \sum_{j\in\mathcal{J}} \gamma_j \sum_{n=1}^{C_j} \tilde{\pi}_{n,j}) \right) dt < \infty$. We have

$$\begin{aligned}
&\int_0^\tau \left(z(t, \tilde{\mathbf{u}}_0) - \sum_{j\in\mathcal{J}} \gamma_j \sum_{n=1}^{C_j} \tilde{\pi}_{n,j} \right) dt = -\int_0^\tau \frac{dz(t, \tilde{\mathbf{u}}_0)}{dt} dt \\
&- \int_0^\tau \lambda \left(\left(\sum_{j\in\mathcal{J}} \gamma_j \tilde{\pi}_{C_j,j} \right)^d - \left(\sum_{j\in\mathcal{J}} \gamma_j \tilde{x}_{C_j,j}(t, \tilde{\mathbf{u}}_0) \right)^d \right) dt \\
&\leq (z(0, \tilde{\mathbf{u}}_0) - z(\tau, \tilde{\mathbf{u}}_0)) \leq z(0, \tilde{\mathbf{u}}_0)
\end{aligned}$$

where the first equality follows from (24) and (25); the second inequality follows since $\tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}_0) \geq \tilde{\boldsymbol{\pi}}$; the third inequality follows since $z(\tau, \tilde{\mathbf{u}}_0) \geq 0$ for all $\tau \geq 0$. Hence, the integral on the left hand side is uniformly bounded by a constant (independent of τ) for all $\tau \geq 0$. This implies that the integral must converge as $\tau \to \infty$. This completes the proof. ■

We now consider uniqueness and stability of the equilibrium point for the multi-class case, where the monotonicity property, similar to the one established in Lemma 2, does not hold [18]. Thus in the case ($L > 1$) we have to use a different argument.

Theorem *4:* For the multi-class case ($L > 1$) there exists a unique equilibrium point $\boldsymbol{\pi} \in \mathcal{U}$ of the mean field $\mathbf{x}(\cdot)$. Furthermore, the equilibrium point $\boldsymbol{\pi}$ is globally asymptotically stable.

Proof: We sketch a proof of the theorem below. We first express the mean field in terms of the tail sums $\tilde{x}_{\underline{k},j}(\cdot) = \sum_{\underline{n}\in\mathcal{S}_j : \underline{n}.\underline{A} \geq \underline{k}.\underline{A}} x_{\underline{n},j}(\cdot)$, $k \in \mathcal{S}_j$ assuming without loss of generality that

the vector $\underline{A}$ is such that for any two states $\underline{n}, \underline{n}' \in \mathcal{S}_j$ with $\underline{n} \neq \underline{n}'$ we have $\underline{n}.\underline{A} \neq \underline{n}'.\underline{A}$ [1]. We note that $\tilde{x}_{\underline{0},j}(t) = 1$ for all $j \in \mathcal{J}$, $t \geq 0$. Therefore, expressed in terms of the tail sums, the mean field is given by

$$\tilde{\mathbf{x}}(0) = \tilde{\mathbf{u}}_0, \tag{29}$$

$$\dot{\tilde{x}}_{\underline{k},j}(t) = \tilde{h}_{\underline{k},j}(\tilde{\mathbf{x}}(t)), \tag{30}$$

where $\tilde{\mathbf{x}}(\cdot) = (\tilde{x}_{\underline{k},j}(\cdot), \underline{k} \in \mathcal{S}_j \backslash \underline{0}, j \in \mathcal{J})$ and $\tilde{h}_{\underline{k},j}(\cdot) = \sum_{\underline{n} \in \mathcal{S}_j : \underline{n}.\underline{A} \geq \underline{k}.\underline{A}} h_{\underline{n},j}(\cdot)$ for $\underline{k} \in \mathcal{S}_j \backslash \underline{0}$, $j \in \mathcal{J}$. We define the mapping $\tilde{\mathbf{h}}(\cdot) = (\tilde{h}_{\underline{k},j}(\cdot), \underline{k} \in \mathcal{S}_j \backslash \underline{0}, j \in \mathcal{J})$. The next (second step) is to verify that the mapping $\tilde{\mathbf{h}}(\cdot)$, when seen as a mapping from $\mathbb{R}^{\sum_{j \in \mathcal{J}} |\mathcal{S}_j| - M}$ to itself, is proper, i.e., $\|\tilde{\mathbf{h}}(\tilde{\mathbf{x}})\|_2 \to \infty$ if $\|\tilde{\mathbf{x}}\|_2 \to \infty$. Finally, the third step is to verify that the Jacobian matrix $\tilde{\mathbf{J}}(\tilde{\mathbf{x}})$ of $\tilde{\mathbf{h}}$ evaluated at $\tilde{\mathbf{x}}$ has all its eigenvalues with negative real parts (Hurwitz) [19] (and hence non-singular) for all $\tilde{\mathbf{x}} \in \mathbb{R}^{\sum_{j \in \mathcal{J}} |\mathcal{S}_j| - M}$. The third step shows that the mapping $\tilde{\mathbf{h}}(\cdot)$ locally homeomorphic at every point in $\mathbb{R}^{\sum_{j \in \mathcal{J}} |\mathcal{S}_j| - M}$ [[20], Theorem 3.1.5, Page 113]. According to the Hadamard's global inverse function theorem [[20], Theorem 5.1.4 (i), Page 221], the second and the third step together imply that the mapping $\tilde{\mathbf{h}}(\cdot)$ is globally homeomorphic on $\mathbb{R}^{\sum_{j \in \mathcal{J}} |\mathcal{S}_j| - M}$, i.e, the inverse exists and is continuous at every point on $\mathbb{R}^{\sum_{j \in \mathcal{J}} |\mathcal{S}_j| - M}$. This implies in particular that $\tilde{\boldsymbol{\pi}} = \tilde{\mathbf{h}}^{-1}(\mathbf{0})$ is unique proving the uniqueness of the equilibrium point of the mean field. The Hurwitz property of the Jacobian matrix at the equilibrium, as established in the third step, also implies that the mean field $\tilde{\mathbf{x}}(t, \tilde{\mathbf{u}}_0)$ converges asymptotically to its unique equilibrium point $\tilde{\boldsymbol{\pi}}$ for all $\mathbf{u}_0$ sufficiently close to the equilibrium point $\tilde{\boldsymbol{\pi}}$. This corresponds to local asymptotic stability.

To show that the equilibrium is globally asymptotically stable we use the facts that the space of tail sums denoted by $\mathcal{V}$ is compact and the Jacobian $\tilde{\mathbf{J}}(\mathbf{x})$ is Hurwitz for all $\mathbf{x} \in \mathcal{V}$.

Consider the mean field $\tilde{\mathbf{x}}(\cdot)$ which satisfies the following equation

$$\frac{d\tilde{\mathbf{x}}(t)}{dt} = \tilde{\mathbf{h}}(\tilde{\mathbf{x}}(t)), \ \tilde{\mathbf{x}}(0) \in \mathcal{V},$$

[1] In case this does not hold, we can order the states of the servers of each type in the increasing order of their resource requirements. States having the same resource requirement can be ordered arbitrarily. The tail sums can then be defined according to the ordering of the states.

where $\tilde{\mathbf{h}}(\cdot)$ is a C^1 mapping on $\mathcal{V}$ and $\tilde{\mathbf{J}}(\cdot)$ denotes its Jacobian. From the Lyapunov theory of linear systems, we know that since $\tilde{\mathbf{J}}(\mathbf{x})$ is Hurwitz for each $\mathbf{x} \in \mathcal{V}$, there exists a unique positive definite matrix $\mathbf{P}(\mathbf{x})$ that solves the Lyapunov equation [19]:

$$\mathbf{P}(\mathbf{x})\tilde{\mathbf{J}}(\mathbf{x}) + \tilde{\mathbf{J}}(x)^T\mathbf{P}(\mathbf{x}) = -\mathbf{I} \tag{31}$$

where $\mathbf{I}$ is the identity matrix in $\mathbb{R}^{\sum_{j\in\mathcal{J}}|\mathcal{S}_j|-M}$. For convenience we denote $\sum_{j\in\mathcal{J}}|\mathcal{S}_j| - M$ as ν.

Let $\mathcal{S} = \{\mathbf{S} : \mathbf{S} \in S^{\nu}_{++}, ||\mathbf{S}|| = 1\}$ where S^{ν}_{++} denotes the convex cone of $\nu \times \nu$ positive definite matrices. By definition $\mathcal{S}$ is compact. Define:

$$g(\mathbf{x}) = \min_{\mathbf{y}\in\mathcal{V},\mathbf{S}\in\mathcal{S}} tr(\mathbf{S}(\mathbf{P}(\mathbf{x}) - \mathbf{P}(\mathbf{y})) \tag{32}$$

then $g(\cdot)$ is continuous for $\mathbf{x} \in \mathcal{V}$ and consider the following semi-infinite program (see [21])

$$\begin{aligned} \max_{\mathbf{x}\in\mathcal{V}} \ & ||\mathbf{P}(\mathbf{x})|| \\ g(\mathbf{x}) \geq \ & 0 \end{aligned}$$

where for $\mathbf{x} \in \mathcal{V}$, $\mathbf{P}(\mathbf{x})$ is the unique solution to the Lyapunov equation (31) and $||.||$ denotes the spectral norm given by $\lambda_{max}(\mathbf{P}(\mathbf{x}))$.

Since $g(\mathbf{x})$ is continuous and $\mathcal{V}$ is compact, there exists $\bar{\mathbf{x}} \in \mathcal{V}$ such that $\mathbf{P}(\bar{\mathbf{x}})$ solves the problem. Let $\mathbf{P} = \mathbf{P}(\bar{\mathbf{x}})$.

Now let $\tilde{\mathbf{P}}(\mathbf{x}, \delta)$ be a solution to (31) with the right hand side taken as $-(1+\delta)\mathbf{I}$ for $\delta > 0$ then $\tilde{\mathbf{P}}(\mathbf{x}, \delta) \geq \mathbf{P}(\mathbf{x})$ by the monotonicity of the Lyapunov equation. Moreover since for any $\mathbf{x}$

$$\tilde{\mathbf{P}}(\mathbf{x}, \delta) = (1+\delta)\int_0^{\infty} e^{\tilde{\mathbf{J}}(\mathbf{x})^T t} e^{\tilde{\mathbf{J}}(\mathbf{x})t} dt.$$

We observe that $\lambda_{max}(\tilde{\mathbf{P}}(\mathbf{x}, \delta)) = ||\tilde{\mathbf{P}}(\mathbf{x}, \delta)||$ is continuous and increasing in δ and goes to infinity as $\delta \to \infty$.

Therefore $\exists \delta_{\mathbf{x}} > 0$ such that $||\mathbf{P} - \tilde{\mathbf{P}}(\mathbf{x}, \delta_{\mathbf{x}})|| \leq \frac{\varepsilon}{2K}$, $\varepsilon \in (0, 1)$ since by construction $\mathbf{P} > \mathbf{P}(\mathbf{x})$ and where $K = \sup_{\mathbf{x}\in\mathcal{V}} ||\tilde{\mathbf{J}}(\mathbf{x})|| < \infty$ since $\mathcal{V}$ is convex and compact.

Hence for any $\mathbf{z} \in \mathbb{R}^{\nu}$ we have:

$$\begin{aligned} \mathbf{z}^T[\mathbf{P}\tilde{\mathbf{J}}(\mathbf{x}) + \tilde{\mathbf{J}}^T(\mathbf{x})\mathbf{P}]\mathbf{z} &= \mathbf{z}^T[(\mathbf{P} - \tilde{\mathbf{P}}(\mathbf{x}, \delta_{\mathbf{x}}))\tilde{\mathbf{J}}(\mathbf{x}) + \tilde{\mathbf{J}}^T(\mathbf{x})(\mathbf{P} - \tilde{\mathbf{P}}(\mathbf{x}, \delta_{\mathbf{x}}))]\mathbf{z} \\ &\quad + \mathbf{z}^T[\tilde{\mathbf{P}}(\mathbf{x}, \delta_{\mathbf{x}})\tilde{\mathbf{J}}(\mathbf{x}) + \tilde{\mathbf{J}}^T(\mathbf{x})\tilde{\mathbf{P}}(\mathbf{x}, \delta_{\mathbf{x}})]\mathbf{z} \\ &\leq \varepsilon||\mathbf{z}||^2 - (1+\delta_{\mathbf{x}})||\mathbf{z}||^2 \\ &< -c||\mathbf{z}||^2, \end{aligned}$$

where $c > 0$. Thus we have shown the existence of $\mathbf{P}$ such that $\mathbf{P}\tilde{\mathbf{J}}(\mathbf{x}) + \tilde{\mathbf{J}}^T(\mathbf{x})\mathbf{P} \leq -c\mathbf{I},\ c > 0$ for all $\mathbf{x} \in \mathcal{V}$.

Therefore the function: $V(\mathbf{x}(t)) = \tilde{\mathbf{h}}^T(\mathbf{x}(t))P\tilde{\mathbf{h}}(\mathbf{x}(t))$ serves as a *bona fide* Lyapunov function that is radially unbounded (proper) and

$$\frac{d}{dt}V(\mathbf{x}(t)) \leq -c||\tilde{\mathbf{h}}(\mathbf{x}(\mathbf{t}))||^2 < 0$$

for any $\mathbf{x} \neq \tilde{\boldsymbol{\pi}}$ in $\mathcal{V}$ by uniqueness of $\tilde{\boldsymbol{\pi}}$, and hence $\tilde{\boldsymbol{\pi}}$ is globally asymptotically stable.

Showing that $\tilde{\mathbf{h}}(\cdot)$ is proper and its Jacobian $\tilde{\mathbf{J}}(\cdot)$ is Hurwitz everywhere on $\mathcal{V}$ can be done for any $L > 1$ (multiple classes) and $M > 1$ (heterogeneous servers). However, for notational convenience, we illustrate these steps steps in an example. The steps can be shown to hold in the more general case at the cost of notational complexity.

Consider a system with parameters: $L = 2$, $M = 2$, $d = 2$ $C_1 = 3$, $C_2 = 4$, $A_1 = 2$, $A_2 = 3$. For the above parameter setting we have $\mathcal{S}_1 = \{(0,0),(1,0),(0,1)\}$ and $\mathcal{S}_2 = \{(0,0),(1,0),(0,1),(2,0)\}$. In this case the mean field can be expressed in terms of the vector of tail sums $\tilde{\mathbf{x}} = (\tilde{x}_{(1,0),1}, \tilde{x}_{(0,1),1}, \tilde{x}_{(1,0),2}, \tilde{x}_{(0,1),2}, \tilde{x}_{(2,0),2})$ as follows (we omit t for convenience)

$$\tilde{x}_{(0,0),1} = 1, \tag{33}$$

$$\frac{d\tilde{x}_{(1,0),1}}{dt} = h_{(1,0),1}(\tilde{\mathbf{x}}) = \frac{\lambda_1 + \lambda_2}{\gamma_1}\vartheta_1(\tilde{\mathbf{x}}) - \tilde{x}_{(1,0),1} \tag{34}$$

$$\frac{d\tilde{x}_{(0,1),1}}{dt} = h_{(0,1),1}(\tilde{\mathbf{x}}) = \frac{\lambda_2}{\gamma_1}\vartheta_1(\tilde{\mathbf{x}}) - \tilde{x}_{(0,1),1} \tag{35}$$

$$\tilde{x}_{(0,0),2} = 1, \tag{36}$$

$$\frac{d\tilde{x}_{(1,0),2}}{dt} = h_{(1,0),2}(\tilde{\mathbf{x}}) = \frac{\lambda_1 + \lambda_2}{\gamma_2}\vartheta_2(\tilde{\mathbf{x}}) - \tilde{x}_{(1,0),2} + \tilde{x}_{(2,0),2} \tag{37}$$

$$\frac{d\tilde{x}_{(0,1),2}}{dt} = h_{(0,1),2}(\tilde{\mathbf{x}}) = \frac{\lambda_2}{\gamma_2}\vartheta_2(\tilde{\mathbf{x}}) + \frac{\lambda_1}{\gamma_2}\vartheta_3(\tilde{\mathbf{x}}) - \tilde{x}_{(0,1),2} - \tilde{x}_{(2,0),2} \tag{38}$$

$$\frac{d\tilde{x}_{(2,0),2}}{dt} = h_{(2,0),2}(\tilde{\mathbf{x}}) = \frac{\lambda_1}{\gamma_2}\vartheta_3(\tilde{\mathbf{x}}) - 2\tilde{x}_{(2,0),2}, \tag{39}$$

where

$$\vartheta_1(\tilde{\mathbf{x}}) = (\gamma_1 + \gamma_2\tilde{x}_{(1,0),2})^2 - (\gamma_1\tilde{x}_{(1,0),1} + \gamma_2\tilde{x}_{(1,0),2})^2, \tag{40}$$

$$\vartheta_2(\tilde{\mathbf{x}}) = 1 - (\gamma_1 + \gamma_2\tilde{x}_{(1,0),2})^2, \tag{41}$$

$$\vartheta_3(\tilde{\mathbf{x}}) = (\gamma_1\tilde{x}_{(1,0),1} + \gamma_2\tilde{x}_{(1,0),2})^2 - (\gamma_1\tilde{x}_{(1,0),1} + \gamma_2\tilde{x}_{(0,1),2})^2. \tag{42}$$

Hence, each component of the mapping $\tilde{\mathbf{h}} = (h_{(1,0),1}, h_{(0,1),1}, h_{(1,0),2}, h_{(0,1),2}, h_{(2,0),2})$ is a polynomial on $\mathbb{R}^5$. It is easy to see from the expressions of the polynomials that if any subset of components of the vector $\tilde{\mathbf{x}}$ approaches to ∞, then at least one of the components of $\tilde{\mathbf{h}}$ approaches to ∞ or $-\infty$. Therefore, $\tilde{\mathbf{h}}$ is proper on $\mathbb{R}^5$ to itself. Finally, the Jacobian matrix $\tilde{\mathbf{J}}(\tilde{\mathbf{x}})$ of the map $\tilde{\mathbf{h}}$ computed at any point $\tilde{\mathbf{x}} \in \mathbb{R}^5$ is given by

$$\tilde{\mathbf{J}}(\tilde{\mathbf{x}}) = \begin{bmatrix} \begin{matrix}-2(\lambda_1+\lambda_2)(\gamma_1\tilde{x}_{(1,0),1}\\ +\gamma_2\tilde{x}_{(1,0),2})-1\end{matrix} & 0 & 2(\lambda_1+\lambda_2)\gamma_2(1-\tilde{x}_{(1,0),1}) & 0 & 0 \\ -2\lambda_2(\gamma_1\tilde{x}_{(1,0),1}+\gamma_2\tilde{x}_{(1,0),2}) & -1 & 2\lambda_2\gamma_2(1-\tilde{x}_{(1,0),1}) & 0 & 0 \\ 0 & 0 & -2(\lambda_1+\lambda_2)(\gamma_1+\gamma_2\tilde{x}_{(1,0),2})-1 & 0 & 1 \\ 2\lambda_1\gamma_1(\tilde{x}_{(1,0),2}-\tilde{x}_{(0,1),2}) & 0 & \begin{matrix}-2\lambda_2(\gamma_1+\gamma_2\tilde{x}_{(1,0),2})+\\ 2\lambda_1(\gamma_1\tilde{x}_{(1,0),1}+\gamma_2\tilde{x}_{(1,0),2})\end{matrix} & \begin{matrix}-2\lambda_1(\gamma_1\tilde{x}_{(1,0),1}\\ +\gamma_2\tilde{x}_{(1,0),2})-1\end{matrix} & -1 \\ 2\lambda_1\gamma_1(\tilde{x}_{(1,0),2}-\tilde{x}_{(0,1),2}) & 0 & 2\lambda_1(\gamma_1\tilde{x}_{(1,0),1}+\gamma_2\tilde{x}_{(1,0),2}) & \begin{matrix}-2\lambda_1(\gamma_1\tilde{x}_{(1,0),1}\\ +\gamma_2\tilde{x}_{(0,1),2})\end{matrix} & -2 \end{bmatrix}$$

A Routh-Hurwitz test [22] of the characteristic polynomial of the matrix $\tilde{\mathbf{J}}(\tilde{\mathbf{x}})$ then shows that all the eigenvalues of the matrix have strictly negative real parts. Therefore, $\tilde{\mathbf{J}}(\tilde{\mathbf{x}})$ is non-singular and Hurwitz everywhere in $\mathbb{R}^5$. Thus, we conclude that the system has a unique equilibrium point $\tilde{\boldsymbol{\pi}}$ and it is globally asymptotically stable. ∎

Remark *4:* The Hurwitz property of A shows that the convergence to the equilibrium point is exponential.

We note that for each N the process $\mathbf{x}^{(N)}(\cdot)$ is positive recurrent and hence has a unique stationary distribution $\boldsymbol{\pi}_N$. We denote by $\mathbf{x}^{(N)}(\infty)$ the random variable distributed according to $\boldsymbol{\pi}_N$. In the next theorem, we show that if the equilibrium point of the mean field is globally asymptotically stable then $\mathbf{x}^{(N)}(\infty)$ concentrates near the unique equilibrium point $\boldsymbol{\pi}$ of the mean field as $N \to \infty$.

Theorem *5: Let $\boldsymbol{\pi}_N$ denote the stationary distribution of the process $\mathbf{x}^{(N)}(\cdot)$. Then the sequence $(\boldsymbol{\pi}_N)_N$ converges weakly to $\delta_{\boldsymbol{\pi}}$ as $N \to \infty$ or equivalently $\mathbf{x}^{(N)}(\infty) \Rightarrow \boldsymbol{\pi}$*

Proof: We note that since the space $\mathcal{U}$ is compact, the sequence of probability measures on $(\boldsymbol{\pi}_N)_N$ is tight. Hence, Prohorov's theorem [23] implies that the sequence $(\boldsymbol{\pi}_N)_N$ has limit

points. We now show that all the limit points coincide with $\boldsymbol{\pi}$. Theorem 1 implies that any limit point of the sequence $(\boldsymbol{\pi}_N)_N$ is an invariant measure of the map $\mathbf{u}_0 \mapsto \mathbf{x}(t, \mathbf{u}_0)$. Theorems 3 and 4 guarantee the global asymptotic stability of the equilibrium point which implies that the unique invariant measure of the map $\mathbf{u}_0 \mapsto \mathbf{x}(t, \mathbf{u}_0)$ is $\delta_{\boldsymbol{\pi}}$. Therefore, any convergent subsequence of $(\boldsymbol{\pi}_N)_N$ converges to $\delta_{\boldsymbol{\pi}}$. This completes the proof. ■

V. Propagation of chaos

In this section, we focus on the states 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 which is formally known as the *propagation of chaos* or *asymptotic independence property*. We further show that the stationary distribution of server states in the limiting system is determined by the unique stationary point $\boldsymbol{\pi}$ of the system (10)-(11). To formally state the results we introduce the following notations.

- The state 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 $k \in \{1, 2, \ldots, N\gamma_j\}$, $j \in \mathcal{J}$.
- For each $j \in \mathcal{J}$ and $t \geq 0$, we denote by $x_j(t, \mathbf{u}_0)$, the distribution on $\mathcal{S}_j$ given by $x_j(t, \mathbf{u}_0) = (x_{\underline{n},j}(t, \mathbf{u}_0), n \in \mathcal{S}_j)$, where $\mathbf{x}(t, \mathbf{u}_0)$ is the mean field limit starting from $\mathbf{u}_0$. Further, we define $x_j(\infty, \mathbf{u}_0) = (\pi_{\underline{n},j}, n \in \mathcal{S}_j)$. We note that due to Theorem 2, $x_j(\infty, \mathbf{u}_0)$ does not depend the initial point $\mathbf{u}_0$.

Further, 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* $j \in \{1, 2, \ldots, M\}$.

Theorem *6: For the model considered in this paper, if the two conditions of Theorem 5 holds,* $\left\{q_{k,j}^{(N)}(0), 1 \leq k \leq N\gamma_j, 1 \leq j \leq M\right\}$ *is intra-type exchangeable, and* $\mathbf{x}^{(N)}(0) \Rightarrow \mathbf{u}_0 \in \mathcal{U}$ *as* $N \to \infty$*, then the following holds*

1) *For each fix* k *and* $t \in [0, \infty]$, $\mathcal{L}(q_{k,j}^{(N)}(t)) \Rightarrow x_j(t, \mathbf{u}_0)$ *as* $N \to \infty$.
2) *Fix positive integers* $r_1, r_2, \ldots, r_M$*. For each* $t \in [0, \infty]$,

$$\left\{q_{k,j}^{(N)}(t), 1 \leq k \leq r_j, 1 \leq j \leq M\right\} \Rightarrow \left\{U_{k,j}(t), 1 \leq k \leq r_j, 1 \leq j \leq M\right\}, \tag{43}$$

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 $x_j(t, \mathbf{u}_0)$ for all $1 \leq k \leq r_j$.

Proof: Note that the first part of Theorem 6 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 readily extends to any $M \geq 2$.

Due to the dynamics of the system (power-of-d scheme) the joint law of the collection $\{q_{k,j}^{(N)}(t), 1 \leq k \leq N\gamma_j, 1 \leq j \leq 2\}$ depends only on the empirical distribution of states at time t given by $\mathbf{x}^{(N)}(t)$. Hence, permuting states among servers of the same type does not affect the joint law of the collection. Therefore, $\{q_{k,j}^{(N)}(t), 1 \leq k \leq N\gamma_j, 1 \leq j \leq 2\}$ is intra-type exchangeable for all $t \in [0, \infty]$. Now, given that $\mathbf{x}^{(N)}(0) \Rightarrow \mathbf{u}_0 \in \mathcal{U}$ as $N \to \infty$ we know from Theorem 1 and Theorem 5 that $\mathbf{x}^{(N)}(t) \Rightarrow \mathbf{x}(t, \mathbf{u}_0)$ as $N \to \infty$ for all $t \in [0, \infty]$. Henceforth, we will omit the variables t and $\mathbf{u}_0$ in our calculations since they hold for all $t \in [0, \infty]$ and all $\mathbf{u}_0 \in \mathcal{U}$. To prove the independence, it is sufficient to show that the following holds:

$$\mathbb{E}\left[\prod_{k=1}^{r_1} \phi_k\left(q_{k,1}^{(N)}\right) \prod_{k=1}^{r_2} \psi_k\left(q_{k,2}^{(N)}\right)\right] \to \prod_{k=1}^{r_1} \langle \phi_k, x_1\rangle \prod_{k=1}^{r_2} \langle \psi_k, x_2\rangle \text{ as } N \to \infty \tag{44}$$

for all bounded mappings $\phi_k : \mathcal{S}_1 \to \mathbb{R}_+$ and $\psi_k : \mathcal{S}_2 \to \mathbb{R}_+$. We have

$$\begin{aligned}
&\left|\mathbb{E}\left[\prod_{k=1}^{r_1} \phi_k\left(q_{k,1}^{(N)}\right) \prod_{k=1}^{r_2} \psi_k\left(q_{k,2}^{(N)}\right)\right] - \prod_{k=1}^{r_1} \langle \phi_k, x_1\rangle \prod_{k=1}^{r_2} \langle \psi_k, x_2\rangle\right| \\
&\qquad \leq \left|\mathbb{E}\left[\prod_{k=1}^{r_1} \phi_k\left(q_{k,1}^{(N)}\right) \prod_{k=1}^{r_2} \psi_k\left(q_{k,2}^{(N)}\right)\right] - \mathbb{E}\left[\prod_{k=1}^{r_1} \langle \phi_k, x_1^{(N)}\rangle \prod_{k=1}^{r_2} \langle \psi_k, x_2^{(N)}\rangle\right]\right| \\
&\qquad\qquad + \left|\mathbb{E}\left[\prod_{k=1}^{r_1} \langle \phi_k, x_1^{(N)}\rangle \prod_{k=1}^{r_2} \langle \psi_k, x_2^{(N)}\rangle\right] - \prod_{k=1}^{r_1} \langle \phi_k, x_1\rangle \prod_{k=1}^{r_2} \langle \psi_k, x_2\rangle\right|,
\end{aligned} \tag{45}$$

where $x_j^{(N)} = (x_{\underline{n},j}, \underline{n} \in \mathcal{S}_j)$ is the random probability measure on $\mathcal{S}_j$ induced by the process $\mathbf{x}^{(N)}$. We note that the second term on the right hand side of the above inequality vanishes as $N \to \infty$ because of the following facts: $x_j^{(N)} \Rightarrow x_j$ as $N \to \infty$ for $j = 1, 2$; x_1 and x_2 are deterministic; $x_j^{(N)}$ is a bounded random vector for $j = 1, 2$. 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

$$\mathbb{E}\left[\prod_{k=1}^{r_1}\phi_k\left(q_{k,1}^{(N)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_{k,2}^{(N)}\right)\right] = \frac{1}{(N\gamma_1)_{r_1}(N\gamma_2)_{r_2}}\times \mathbb{E}\left[\sum_{\substack{\sigma\in P(r_1,N\gamma_1)\\ \sigma'\in P(r_2,N\gamma_2)}}\prod_{k=1}^{r_1}\phi_k\left(q_{\sigma(k),1}^{(N)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_{\sigma'(k),2}^{(N)}\right)\right] \quad (46)$$

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. Also, by definition of $x_j^{(N)}$ we have

$$\mathbb{E}\left[\prod_{k=1}^{r_1}\langle\phi_k, x_1^{(N)}\rangle\prod_{k=1}^{r_2}\langle\psi_k, x_2^{(N)}\rangle\right] = \mathbb{E}\left[\left(\prod_{k=1}^{r_1}\frac{1}{N\gamma_1}\sum_{l=1}^{N\gamma_1}\phi_k\left(q_{l,1}^{(N)}\right)\right)\left(\prod_{k=1}^{r_2}\frac{1}{N\gamma_2}\sum_{l=1}^{N\gamma_2}\psi_k\left(q_{l,2}^{(N)}\right)\right)\right] \quad (47)$$

Hence, the first term on the right hand side of (45) can be bounded as follows

$$\begin{aligned}
&\left|\mathbb{E}\left[\prod_{k=1}^{r_1}\phi_k\left(q_{k,1}^{(N)}\right)\prod_{k=1}^{r_2}\psi_k\left(q_{2,k}^{(N)}\right)\right] - \mathbb{E}\left[\prod_{k=1}^{r_1}\langle\phi_k, x_1^{(N)}\rangle\prod_{k=1}^{r_2}\langle\psi_k, x_2^{(N)}\rangle\right]\right| \\
&\leq (N\gamma_1)_{r_1}(N\gamma_2)_{r_2}\left(\frac{1}{(N\gamma_1)_{r_1}(N\gamma_2)_{r_2}} - \frac{1}{(N\gamma_1)^{r_1}(N\gamma_2)^{r_2}}\right)B^{r_1+r_2} \\
&+ \left((N\gamma_1)^{r_1}(N\gamma_2)^{r_2} - (N\gamma_1)_{r_1}(N\gamma_2)_{r_2}\right)\frac{B^{r_1+r_2}}{(N\gamma_1)^{r_1}(N\gamma_2)^{r_2}} \\
&\leq 2B^{r_1+r_2}\left(1 - \frac{(N\gamma_1)_{r_1}(N\gamma_2)_{r_2}}{(N\gamma_1)^{r_1}(N\gamma_2)^{r_2}}\right) \to 0 \text{ as } N\to\infty
\end{aligned}$$

where B is a constant such that $\|\phi_k\|_\infty < B$ for $k = 1,2,\dots,r_1$ and $\|\psi_k\|_\infty < B$ for $k = 1,2,\dots,r_2$. This completes the proof. ■

Remark *5:* We note that any initial condition that specifies $\mathbf{x}_{\underline{n},j}^{(N)}(0) = g_{\underline{n},j}$ for all N, $\underline{n}\in\mathcal{S}_j$, and $j\in\mathcal{J}$, satisfies the conditions in Theorem 6. In particular, the conditions in Theorem 6 are satisfied when all the servers are empty at $t=0$.

Thus, the above theorem shows that in the limiting system any finite set of servers become independent of each other and the stationary distribution of states of any server of type-j is

given by $\pi_j = \{\pi_{\underline{n},j}, n \in \mathcal{S}_j\}$. The following proposition shows (using the independence of the servers in the limiting system) that in equilibrium the arrivals of class $l \in \mathcal{L}$ at any given server of type $j \in \mathcal{J}$ in the limiting system form a state dependent Poisson process whose rates are given by $\lambda_{\underline{n},j}(\boldsymbol{\pi})$, $\underline{n} \in \mathcal{S}_j$, where $\lambda_{\underline{n},j}(\mathbf{x})$ for $\mathbf{x} \in \mathcal{U}$ are as defined in (14).

Proposition *2:* In equilibrium, the arrival process of jobs at any given server in the limiting system is a state dependent Poisson process. Furthermore, the equilibrium arrival rate of class-l jobs at a server of type $j \in \mathcal{J}$, when it is in state $\underline{n} \in \mathcal{S}_j$, is given by

$$\lambda^{(l)}_{\underline{n},j}(\boldsymbol{\pi}) = \lambda_l \frac{F(\underline{n}, j, \boldsymbol{\pi})}{E(\underline{n}, j, j, \boldsymbol{\pi})}, \tag{48}$$

where $\underline{n} \in \mathcal{S}_j$ is such that $\underline{n} + \underline{e}_l \in \mathcal{S}_j$ and $F(\underline{n}, j, \boldsymbol{\pi})$, $E(\underline{n}, j, j, \boldsymbol{\pi})$ are as defined in Lemma 1.

Proof: The proof is given in the Appendix D. ■

From Remark 3, we already know that the equilibrium point $\boldsymbol{\pi}$ of the mean field satisfies

$$\lambda^{(l)}_{\underline{n}-\underline{e}_l,j}(\boldsymbol{\pi})\pi_{\underline{n}-\underline{e}_l,j} I_{\underline{n}-\underline{e}_l \in \mathcal{S}_j} = n_l \pi_{\underline{n},j} \text{ for } \underline{n} \in \mathcal{S}_j \text{ and } l \in \mathcal{L}. \tag{49}$$

Now, since (by Proposition 2) $\lambda^{(l)}_{\underline{n}-\underline{e}_l,j}(\boldsymbol{\pi})$ is the equilibrium arrival rate of class-l jobs at a server of type j in state $\underline{n} - \underline{e}_l \in \mathcal{S}_j$, the above equations can be interpreted as the detailed balance equations that equate the transition rates between the states $\underline{n} - \underline{e}_l$ and $\underline{n}$ for each $\underline{n}, \underline{n} - \underline{e}_l \in \mathcal{S}_j$, $j \in \mathcal{J}$, $l \in \mathcal{L}$. Using the detailed balance equations, we now find a recursive relationship among the stationary tail probabilities of the number of occupied resource units as in [24], [25] at each server in the limiting system. This allows efficient computation of the blocking probabilities for each class of jobs.

Proposition *3: Let $P_{k,j}$, for $1 \leq k \leq C_j$ and $j \in \mathcal{J}$, denote the stationary probability that a server in the limiting system has at least k units of occupied resources, i.e., $P_{k,j} = \sum_{\underline{n} \in \mathcal{S}_j:\ \underline{n}.\underline{A} \geq k} \pi_{\underline{n},j}$. Then $P_{k,j}$ satisfies* (1) *for $0 \leq k \leq C_j - 1$, where $P_{k,j} = 1$ for $k \leq 0$, and $P_{C_j+1,j} = 0$ for all $j \in \mathcal{J}$.*

Proof: For $j \in \mathcal{J}$, $0 \leq k \leq C_j$, we define the set $\mathcal{D}_{k,j}$ as $\mathcal{D}_{k,j} = \{\underline{n} \in \mathcal{S}_j : \underline{n} \cdot \underline{A} = k\}$. Thus, $\mathcal{D}_{k,j}$ denotes the set of states in $\mathcal{S}_j$ for which the total number occupied VM's at a type j server is exactly k. We note that for all $\underline{n} \in \mathcal{D}_{k,j}$ such that $\underline{n} - \underline{e}_l \in \mathcal{S}_j$, we have $G(\underline{n}-\underline{e}_l, i, j, \boldsymbol{\pi}) = \gamma_i P_{k-A_l+C_i-C_j+1,i}$ and $E(\underline{n}-\underline{e}_l, i, j, \boldsymbol{\pi}) = \gamma_i(P_{k-A_l+C_i-C_j,i} - P_{k-A_l+C_i-C_j+1,i})$. Thus, for all $\underline{n} \in \mathcal{D}_{k,j}$ such that $\underline{n} - \underline{e}_l \in \mathcal{S}_j$ we have

$$\lambda^{(l)}_{\underline{n}-\underline{e}_l,j} = \frac{\lambda_l}{\gamma_j \left(P_{k-A_l,j} - P_{k-A_l,j}\right)} \left[\left(\sum_{i=1}^{j} \gamma_i P_{k-A_l+C_i-C_j,i} \; + \sum_{i=j+1}^{M} \gamma_i P_{k-A_l+C_i-C_j+1,i} \right)^d \right.$$
$$\left. - \left(\sum_{i=1}^{j-1} \gamma_i P_{k-A_l+C_i-C_j,i} + \sum_{i=j}^{M} \gamma_i P_{k-A_l+C_i-C_j+1,i} \right)^d \right]. \quad (50)$$

Now from (49) we have

$$\sum_{l\in\mathcal{L}} \sum_{\underline{n}\in\mathcal{D}_{k,j}} A_l \lambda^{(l)}_{\underline{n}-\underline{e}_l,j} \pi_{\underline{n}-\underline{e}_l,j} I_{\underline{n}-\underline{e}_l\in\mathcal{S}_j} = \sum_{l\in\mathcal{L}} \sum_{\underline{n}\in\mathcal{D}_{k,j}} n_l A_l \pi_{\underline{n},j} \text{ for } \underline{n}\in\mathcal{S}^j \text{ and } l\in\mathcal{L} \quad (51)$$

Now, the LHS of the above equation can be simplified as follows:

$$\sum_{l\in\mathcal{L}} \sum_{\underline{n}\in\mathcal{D}_{k,j}} A_l \lambda^{(l)}_{\underline{n}-\underline{e}_l,j} \pi_{\underline{n}-\underline{e}_l,j} I_{\underline{n}-\underline{e}_l\in\mathcal{S}_j} = \sum_{l\in\mathcal{L}} A_l \lambda^{(l)}_{\underline{n}-\underline{e}_l,j} \sum_{\underline{n}\in\mathcal{D}_{k,j}} \pi_{\underline{n}-\underline{e}_l,j} I_{\underline{n}-\underline{e}_l\in\mathcal{S}_j}$$
$$= \sum_{l\in\mathcal{L}} A_l \frac{\lambda_l}{\gamma_j \left(P_{k-A_l,j} - P_{k-A_l+1,j}\right)} \left[\left(\sum_{i=1}^{j} \gamma_i P_{k-A_l+C_i-C_j,i} \; + \sum_{i=j+1}^{M} \gamma_i P_{k-A_l+C_i-C_j+1,i} \right)^d \right.$$
$$\left. - \left(\sum_{i=1}^{j-1} \gamma_i P_{k-A_l+C_i-C_j,i} + \sum_{i=j}^{M} \gamma_i P_{k-A_l+C_i-C_j+1,i} \right)^d \right] \left(P_{k-A_l,j} - P_{k-A_l+1,j}\right)$$
$$= \sum_{l\in\mathcal{L}} A_l \frac{\lambda_l}{\gamma_j} \left[\left(\sum_{i=1}^{j} \gamma_i P_{k-A_l+C_i-C_j,i} \; + \sum_{i=j+1}^{M} \gamma_i P_{k-A_l+C_i-C_j+1,i} \right)^d \right.$$
$$\left. - \left(\sum_{i=1}^{j-1} \gamma_i P_{k-A_l+C_i-C_j,i} + \sum_{i=j}^{M} \gamma_i P_{k-A_l+C_i-C_j+1,i} \right)^d \right]$$

The second equality follows since $\sum_{\underline{n}\in\mathcal{D}_{k,j}} \pi_{\underline{n}-\underline{e}_l,j} I_{\underline{n}-\underline{e}_l\in\mathcal{S}_j} = \left(P_{k-A_l,j} - P_{k-A_l+1,j}\right)$. Similarly, the RHS can be simplified as

$$\sum_{l\in\mathcal{L}} \sum_{\underline{n}\in\mathcal{D}_{k,j}} n_l A_l \pi_{\underline{n},j} = \sum_{\underline{n}\in\mathcal{D}_{k,j}} \pi_{\underline{n},j} \sum_{l\in\mathcal{L}} n_l A_l = \sum_{\underline{n}\in\mathcal{D}_{k,j}} \pi_{\underline{n},j} k = k\left(P_{k,j} - P_{k+1,j}\right).$$

This completes the proof. ■

Remark *6 (Insensitivity):* All the results, discussed so far in this section, have been obtained assuming that the service time distribution of the incoming jobs is exponential. The same results

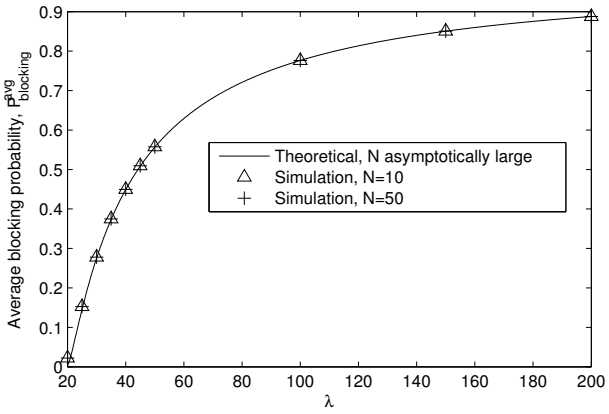


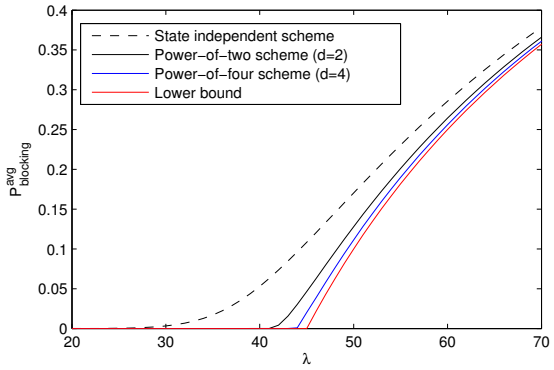
Fig. 1. Accuracy of mean field analysis of power-of-d scheme: Average blocking probability as a function of λ for different values of N.

can be shown to hold for any service time distribution if asymptotic independence of the servers is assumed to hold for general service time distributions. The asymptotic independence property was conjectured to hold for homogeneous systems with local service disciplines and general service time distributions in [26]. The proof of this remains as an open problem.

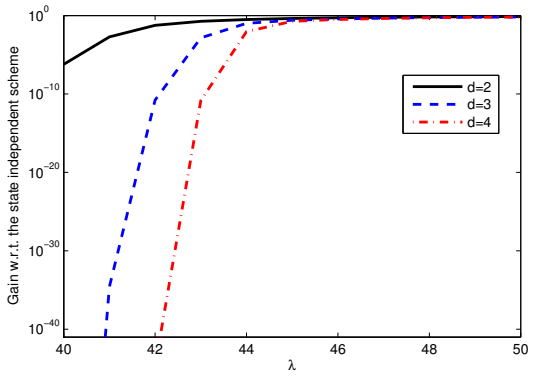
Under the assumption of asymptotic independence of servers for general service time distributions, the statement of Proposition 2 continues to hold, i.e., the equilibrium arrival process at each server in the limiting system is a state dependent Poisson process whose rates are given by (48). This implies that the detailed balance equations given by (49) also hold for general service time distributions. Since the servers in the system are loss servers, the detailed balance condition implies that that the stationary distribution of each server in the limiting system is *insensitive* to service time distributions (see Theorem 1 of [27]). We refer to this property as the *asymptotic insensitivity* of the system. Thus, asymptotic insensitivity of the system holds under the hypothesis of asymptotic independence of the servers, the proof of which remains as an open problem. In the next section, we provide numerical evidences to support insensitivity.

VI. Numerical Results

We first 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: $L = 1$, $A_1 = A = 1$, $M = 2$, $\gamma_1 = \gamma_2 = 0.5$, $C_1 = 20$, $C_2 = 25$, and $d = 2$. All simulation results



(a) Average blocking probability as a function of λ.



(b) Ratio of average blocking probability of the power-of-d scheme to that of the state independent scheme as a function of λ

Fig. 2. Efficacy of the power-of-d scheme

presented in this section are the average of 10,000 independent runs. We use λ to denote the arrival rate of jobs. In Figure 1, we plot the average blocking probability of requests 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 (3). 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 power-of-d scheme even for moderate system sizes.

In Figure 2 we compare for $L = 1$ the average blocking probability of the power-of-d scheme with 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 $p_j = \gamma_j C_j / \sum_{i=1}^{M} \gamma_i C_i$ since in this case p_j is proportional to both γ_j and C_j.[2] The parameters were chosen as $M = 2$, $N = 100$, $\gamma_1 = \gamma_2 = 0.5$, $C_1 = 30$, and $C_2 = 60$. For this parameter setting, the critical load is given by $\lambda_{\text{crit}} = \gamma_1 C_1 + \gamma_2 C_2 = 45$. In Figure 2(a), we plot the average blocking probability of the two schemes as a function of the arrival rate λ. We have also plotted the lower bound obtained from (2). In Figure 2(b), we plot the ratio of average blocking probability of the power-of-d scheme to that of the state independent scheme. Note that the

[2]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.

y-axis is in the log scale. From Figure 2(a) we observe that the average blocking probability obtained for $d = 4$ is almost equal to that of the lower bound. We also observe from Figure 2(b), that the average blocking probability under the power-of-d scheme is orders of magnitude lower than that under the state independent routing scheme around $\lambda = \lambda_{\text{cap}}$. This shows the efficacy of such randomized strategies in reducing blocking for realistic systems which are typically operated near the critical load.

We now numerically confirm the 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 and 2) *Power law*: We consider job length distribution having 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. This supports the fact the under the power-of-d scheme the system is insensitive to the service time distribution in the limit as $N \to \infty$ (asymptotic insensitivity).

TABLE I

ASYMPTOTIC 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-d scheme for multi-class heterogeneous Erlang loss systems with a large number of servers. We showed that in the large system limit the evolution of the empirical occupancy distribution can be characterized through its mean field limit.

Furthermore, we showed that propagation of chaos holds for heterogeneous case through the requirement of intra-type exchangeability.

VIII. Acknowledgements

We thank Prof. R. Srikant for pointing out a gap in the original version of the paper pertaining to the proof of uniqueness and asymptotic convergence to the fixed point. We wish to thank Prof. C. Nielsen for bringing the Krasovskii-Lyapunov theorem to our attention and Prof. H. Wolkowitz for very helpful discussions.

References

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

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

[3] V. Gupta and A. Radovanovic, "Online stochastic bin packing." *Corr*: abs/1211.2687, 2012.

[4] 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*.

[5] 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.

[6] S. T. Maguluri, R. Srikant, and L. Ying, "Heavy traffic optimal resource allocation algorithms for cloud computing clusters," in *Proceedings of 24th International Teletraffic Congress (ITC 24)*, 2012.

[7] 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.

[8] M. Mitzenmacher, "The power of two choices in randomized load balancing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 12, no. 10, pp. 1094–1104, 2001.

[9] A. Ganesh, S. Lilienthal, D. Manjunath, A. Proutiere, and F. Simatos, "Load balancing via random local search in closed and open systems," *Queueing Systems*, 2012.

[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*, 2015, 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] 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.

[14] 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.

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

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

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

[18] P. Nain, "Qualitative properties of the Erlang blocking model with heterogeneous user requirements," *Queueing Systems*, no. 2, pp. 189–206, 1990.

[19] H. K. Khalil, *Nonlinear systems*. Macmillan Publishing Company, New York, 1992.

[20] M. S. Berger, *Nonlinearity and Functional Analysis: Lectures on Nonlinear Problems in Mathematical Analysis*. Academic Press, 1977.

[21] R. Hettich and K. O. Kortanek, "Semi-infinite programming: theory, methods, and applications," *SIAM Rev.*, vol. 35, pp. 380–429, 1993.

[22] B.Kuo, *Automatic Control Systems*. Prentice-Hall, 1982.

[23] P. Billingsley, *Convergence of Probability Measures*. Wiley, 1968.

[24] J. Kaufman, "Blocking in a shared resource environment," *IEEE Transactions on Communications*, vol. 29, no. 10, pp. 1474–1481, 1981.

[25] J. W. Roberts, "A service system with heterogeneous user requirement," in *Performance of Data Communications Systems and Their Applications*, pp. 423–431, 1981.

[26] M. Bramson, Y. Lu, and B. Prabhakar, "Asymptotic independence of queues under randomized load balancing," *Queueing Systems*, vol. 71, no. 3, pp. 247–292, 2012.

[27] S. Zachary, "A note on insensitivity in stochastic networks," *Journal of Applied Probability*, vol. 44, no. 1, pp. 238–248, 2007.

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

APPENDIX

A. Proof of Proposition 1

Using (3) it can be shown that the following recursive relation holds for $0 \leq k \leq C_M - 1$

$$\sum_{j\in\mathcal{J}}(k+1+C_j-C_M)_+\gamma_j(P_{k+1+C_j-C_M,j}-P_{k+2+C_j-C_M,j}) = \lambda\left[\left(\sum_{j=1}^{M}\gamma_j P_{k+C_j-C_M,j}\right)^d-\left(\sum_{j=1}^{M}\gamma_j P_{k+1+C_j-C_M,j}\right)^d\right], \quad (52)$$

where $(y)_+ = \max(0, y)$. From (52) the following can be shown to hold for $0 \leq k \leq C_M - 1$ using backward induction starting at $k = C_M - 1$.

$$\sum_{j=1}^{M}(k+1+C_j-C_M)_+\gamma_j P_{k+1+C_j-C_M,j} \leq \lambda\left(\sum_{j=1}^{M}\gamma_j P_{k+C_j-C_M,j}\right)^d. \quad (53)$$

From (53) it is clear that

$$\left(\sum_{j=1}^{M}\gamma_j P_{k+1+C_j-C_M,j}\right) \leq \frac{\lambda}{k+(C_1-C_M)+1}\left(\sum_{j=1}^{M}\gamma_j P_{k+C_j-C_M,j}\right)^d, \quad (54)$$

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

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

This completes the proof. ∎

B. Proof of Lemma 1

We first consider the transition of the system from the state $\mathbf{u} \in \mathcal{U}^{(N)}$ at t^- to the state $\mathbf{u} - \frac{\mathbf{e}(\underline{n}-\underline{e}_l,j)}{N\gamma_j} + \frac{\mathbf{e}(\underline{n},j)}{N\gamma_j}$ at t, where $\underline{n} \in \mathcal{S}_j$. This transition occurs when an arrival of class $l \in \mathcal{L}$

at time t joins a type-j server which was in state $\underline{n}-\underline{e}_l$ at time t^- (just before the arrival). Let k of the d sampled servers be of type j with state $\underline{n}'$ satisfying $\underline{n}'\cdot\underline{A}=(\underline{n}-\underline{e}_l)\cdot\underline{A}$. For the transition to occur, we must have $k\geq 1$ and one among the k servers must be in state $\underline{n}-\underline{e}_l$. Since there are $N\gamma_j u_{\underline{n}-\underline{e}_l,j}$ and $N\times E(\underline{n}-\underline{e}_l,j,j,u)$ servers of type j in states $\underline{n}-\underline{e}_l$ ans $\underline{n}'$ satisfying $\underline{n}'\cdot\underline{A}=(\underline{n}-\underline{e}_l)\cdot\underline{A}$, respectively, the probability of sampling k such servers is $\binom{k}{1}\gamma_j u_{\underline{n}-\underline{e}_l,j}E^{k-1}(\underline{n}-\underline{e}_l,j,j,u)$. In this case, since there are k servers with equal vacancy, the arrival joins a server with state $\underline{n}-\underline{e}_l$ with probability $1/k$. The other $d-k$ sampled servers must satisfy either of the following two conditions:

- If the sampled server is of type $i<j$, then its state $\underline{n}'$ must satisfy $C_i-\underline{n}'\cdot\underline{A}\leq C_j-(\underline{n}-\underline{e}_l)\cdot\underline{A}$, or, $\underline{n}'\cdot\underline{A}\geq(\underline{n}-\underline{e}_l)\cdot\underline{A}+C_i-C_j$. The number of type i servers in a state satisfying the above relation is $N\times GE(\underline{n}-\underline{e}_l,i,j,\mathbf{u})$. Since servers are sampled uniformly at random, the probability with which one of these servers is sampled is $GE(\underline{n}-\underline{e}_l,i,j,\mathbf{u})$.
- If the sampled server is of type $i\geq j$, then its state $\underline{n}'$ must satisfy $C_i-\underline{n}'\cdot\underline{A}<C_j-(\underline{n}-\underline{e}_l)\cdot\underline{A}$, or, $\underline{n}'\cdot\underline{A}>(\underline{n}-\underline{e}_l)\cdot\underline{A}+C_i-C_j$. Using the similar argument as before, the probability with which such a server is sampled is $G(\underline{n}-\underline{e}_l,i,j,\mathbf{u})$.

Thus the total probability with which the incoming arrival joins a server of type j in state $\underline{n}-\underline{e}_l$ is $\sum_{k=1}^{d}\binom{k}{1}\frac{1}{k}\gamma_j u_{\underline{n}-\underline{e}_l,j}E^{k-1}(\underline{n}-\underline{e}_l,j,j,u)\left(\sum_{i=1}^{j-1}GE(\underline{n},i,j,\mathbf{u})+\sum_{i=j}^{M}G(\underline{n},i,j,\mathbf{u})\right)^{d-k}$ which simplifies to $\frac{F(\underline{n}-\underline{e}_l,j,\mathbf{u})}{E(\underline{n}-\underline{e}_l,j,j,\mathbf{u})}\gamma_j u_{\underline{n}-\underline{e}_l,j}$. Since the arrival rate of class-l jobs is $N\lambda_l$, the rate of transition from the state $\mathbf{u}$ to the state $\mathbf{u}-\frac{\mathbf{e}(\underline{n}-\underline{e}_l,j)}{N\gamma_j}+\frac{\mathbf{e}(\underline{n},j)}{N\gamma_j}$ is given by

$$r\left(\mathbf{u}\to\mathbf{u}-\frac{\mathbf{e}(\underline{n}-\underline{e}_l,j)}{N\gamma_j}+\frac{\mathbf{e}(\underline{n},j)}{N\gamma_j}\right)=N\lambda_l\frac{F(\underline{n}-\underline{e}_l,j,\mathbf{u})}{E(\underline{n}-\underline{e}_l,j,j,\mathbf{u})}\gamma_j u_{\underline{n}-\underline{e}_l,j}. \tag{55}$$

Next, we consider the transition from the state $\mathbf{u}\in\mathcal{U}^{(N)}$ to the state $\mathbf{u}+\frac{\mathbf{e}(\underline{n}-\underline{e}_l,j)}{N\gamma_j}-\frac{\mathbf{e}(\underline{n},j)}{N\gamma_j}$, where $\underline{n}\in\mathcal{S}_j$. This transition occurs when a job of class $l\in\mathcal{L}$ leaves a type $j\in\mathcal{J}$ server in state $\underline{n}$. The number of type-j servers in state $\underline{n}$ when the system is in state $\mathbf{u}$ is $N\gamma_j u_{\underline{n},j}$. From each of these servers, the rate at which class-l jobs depart is n_l. Hence, the rate of transition from the state $\mathbf{u}$ to the state $\mathbf{u}+\frac{\mathbf{e}(\underline{n}-\underline{e}_l,j)}{N\gamma_j}-\frac{\mathbf{e}(\underline{n},j)}{N\gamma_j}$ is given by

$$r\left(\mathbf{u}\to\mathbf{u}+\frac{\mathbf{e}(\underline{n}-\underline{e}_l,j)}{N\gamma_j}-\frac{\mathbf{e}(\underline{n},j)}{N\gamma_j}\right)=N\gamma_j u_{\underline{n},j}n_l \tag{56}$$

The expression (5) now follows directly from the definition of $\mathbf{A}^{(N)}$.

C. Proof of Theorem 1

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 (10)-(11). 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.,

$$\lim_{N\to\infty} \sup_{\mathbf{u}\in\mathcal{U}^{(N)}} \left|\mathbf{T}^{(N)}(t)f(\mathbf{u}) - f(\mathbf{x}(t,\mathbf{u}))\right| = 0, \tag{57}$$

for all continuous functions $f : \mathcal{U} \to \mathbb{R}$, where the convergence is uniform in t within any bounded interval. Combining these three steps, the statement of Theorem 1 follows from Corollary 8.7 of Chapter 4 of [28].

The proof of the first part is essentially the same as the proof of Theorem 6.1 of [11] but we give the details here for the completeness of the paper. We first recall that the metric ρ defined on the space $\mathcal{U}$ is given by:

$$\rho(\mathbf{u},\mathbf{w}) = \sup_{j\in\mathcal{J}} \sup_{\underline{n}\in\mathcal{S}_j} \frac{|u_{\underline{n},j} - w_{\underline{n},j}|}{(\underline{n}\cdot\underline{e})+1}. \tag{58}$$

Now, to prove relative compactness of the sequence of processes $\left\{\mathbf{x}^{(N)}(\cdot)\right\}$ we need to satisfy the following three conditions of Theorem 8.6 of Chapter 3 of [28]:

- For every $\eta > 0$ and rational $t \geq 0$, there exists a compact set $\Gamma_{\eta,t}$ such that

$$\liminf_{N\to\infty} \mathbb{P}\left[\inf\left\{\rho(\mathbf{x}^{(N)}(t), y) : y \in \Gamma_{\eta,t}\right\} < \eta\right] \geq 1 - \eta. \tag{59}$$

  This is condition (7.7) of Chapter 3 of [28].

- For all $T > 0$, there exists $\beta > 0$, $C > 0$, and $\theta > 1$, such that for all N and all $0 \leq h \leq t \leq T + 1$,

$$\mathbb{E}\left[\rho^{\beta/2}(\mathbf{x}^{(N)}(t+h), \mathbf{x}^{(N)}(t))\rho^{\beta/2}(\mathbf{x}^{(N)}(t), \mathbf{x}^{(N)}(t-h))\right] \leq Ch^{\theta}. \tag{60}$$

  This is condition (8.37) of Chapter 3 of [28] which by Theorem 8.8 of Chapter 3 of [28] implies condition (8.28) of Theorem 8.6 of Chapter 3 of [28].

- For that β,

$$\lim_{\delta\to 0}\limsup_{N\to\infty}\mathbb{E}\left[\rho^{\beta}(\mathbf{x}^{(N)}(\delta),\mathbf{x}^{(N)}(0))\right]=0. \tag{61}$$

This is condition (8.30) of Chapter 3 of [28].

Now, (59) follows immediately by choosing $\Gamma_{\eta,t}=\mathcal{U}$ for all $\eta>0$ and $t\geq 0$ since $\mathcal{U}$ is compact and the process $\mathbf{x}^{(N)}(t)$ lies in the space $\mathcal{U}^{(N)}\subseteq\mathcal{U}$. Clearly, the change under the metric ρ is bounded above by that under the sup metric. Due to arrival or departure one of $x^{(N)}_{\underline{n},j}$ is increased by $1/N\gamma_j$ and one is decreased by $1/N\gamma_j$. This changes $\mathbf{x}^N$ by at most $1/(N(\min_{j\in\mathcal{J}}\gamma_j))$ in the sup metric. We next need to bound the number of arrivals and departures that can occur in an interval of length h. The arrivals form a Poisson process whose rate is bounded above by $N\lambda$, where $\lambda=\sum_{l\in\mathcal{L}}\lambda_l$ and the number of departures is bounded above by the number of events in a Poisson process of rate NB where $B=\max_{j\in\mathcal{J}}\max_{\underline{n}\in\mathcal{S}_j}(\underline{n}\cdot\underline{e})$. Therefore, the total number of events in an interval of length h is bounded above by a Poisson random variable with mean $(\lambda+B)Nh$. Thus, $\rho(\mathbf{x}^{(N)}(t+h),\mathbf{x}^{(N)}(t))$ is bounded above by $1/(N(\min_{j\in\mathcal{J}}\gamma_j))$ times a Poisson random variable with mean $(\lambda+B)Nh$. Now we set $\beta=2$. Then using the Markov property of $\mathbf{x}^{(N)}$ we have

$$\begin{aligned}&\mathbb{E}\left[\rho^{\beta/2}(\mathbf{x}^{(N)}(t+h),\mathbf{x}^{(N)}(t))\rho^{\beta/2}(\mathbf{x}^{(N)}(t),\mathbf{x}^{(N)}(t-h))\right]\\&\quad=\mathbb{E}\left[\rho^{\beta/2}(\mathbf{x}^{(N)}(t+h),\mathbf{x}^{(N)}(t))\right]\mathbb{E}\left[\rho^{\beta/2}(\mathbf{x}^{(N)}(t),\mathbf{x}^{(N)}(t-h))\right]\leq\frac{(\lambda+B)^2}{(\min_{j\in\mathcal{J}}\gamma_j)^2}h^2.\end{aligned} \tag{62}$$

Hence, (60) holds for all necessary T,t, and h with $C=\frac{(\lambda+B)^2}{(\min_{j\in\mathcal{J}}\gamma_j)^2}$ and $\theta=2$. It remains to prove (61) for $\beta=2$. This is easy to see since $\mathbb{E}\left[\rho^{\beta}(\mathbf{x}^{(N)}(\delta),\mathbf{x}^{(N)}(0))\right]\leq\frac{(\lambda+B)N\delta+(\lambda+B)^2N^2\delta^2}{N^2(\min_{j\in\mathcal{J}}\gamma_j)^2}$. This completes the proof of relative compactness.

The proof of the second part follows if the mapping $\mathbf{h}$ is shown to be Lipschitz continuous, i.e., there exists $K>0$ such that $\rho\left(\mathbf{h}(\mathbf{u}),\mathbf{h}(\mathbf{v})\right)\leq K\rho\left(\mathbf{u},\mathbf{v}\right)$. For any $\mathbf{u},\mathbf{v}\in\mathcal{U}$ let the L_1-distance between $\mathbf{u}$ and $\mathbf{v}$ be defined as $\|u-v\|=\sum_{j\in\mathcal{J}}\sum_{\underline{n}\in\mathcal{S}_j}|u_{\underline{n},j}-v_{\underline{n},j}|$. Clearly, we have $\frac{\|\mathbf{u}-\mathbf{v}\|}{(B+1)S}\leq\rho(\mathbf{u},\mathbf{v})\leq\|\mathbf{u}-\mathbf{v}\|$, where $S=\sum_{j\in\mathcal{J}}|\mathcal{S}_j|$ and B is as defined before. Thus, these two metrics are equivalent. It is therefore, sufficient to show $\|\mathbf{h}(\mathbf{u})-\mathbf{h}(\mathbf{v})\|\leq K\|\mathbf{u}-\mathbf{v}\|$ for some $K>0$. But this is satisfied with $K=2B+2\lambda d+8\lambda S(d-1)$ as can be seen using (12) and the L_1-norm.

The proof of the third part essentially the same as the proof of Theorem 2 of [16]. We omit the details and mention only the key observation used in the proof which is the fact that as $N \to \infty$ we have $\mathbf{A}^{(N)} f(\mathbf{u}) \to \frac{d}{dt} f(\mathbf{x}(t, \mathbf{u}))|_{t=0}$ uniformly in $\mathbf{u}$ for all $f : \mathcal{U} \to \mathbb{R}$ such that $f(\mathbf{u})$ has bounded partial derivatives of first and second order with respect to each component of $\mathbf{u}$.

D. Proof of Proposition 2

Consider a *tagged* server of type j and the class-l 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 of class-l jobs to the tagged server is a Poisson process with rate $\frac{d}{N} \times N\lambda_l = d\lambda_l$.

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 to the tagged server actually joins the server depending on the number of jobs present at the other possible destination server. However, as $N \to \infty$, due to the asymptotic independence property shown in Proposition 6 the occupancies of the sampled servers become independent of each other. As a result, in equilibrium the actual arrival process converges to a state dependent Poisson process as $N \to \infty$. Now the arrival rates of the Poisson process, as given in (48), can be computed following the similar line of arguments as in the proof of Lemma 1.