Rate-Based Randomized Routing in Large Heterogeneous Processor Sharing Systems

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

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

***Abstract*—Randomized load balancing techniques are effective solutions to reduce mean waiting time of jobs in large web server farms, where obtaining state information of all the servers becomes costly. The classical power-of-two routing scheme, which has already been analyzed for systems of identical servers, requires the instantaneous state information of two randomly selected servers at each job-arrival instant. In this paper, we consider variants of the classical power-of-two scheme for multi-server systems where the servers may have different service rates. We modify the classical power-of-two scheme for the heterogeneous system so that it now incorporates server speeds into the criterion for server selection. We analytically characterize the stability region, stationary load distribution, and the mean sojourn time of jobs of this modified scheme. It is shown that, in the heterogeneous case, the stability region of the modified scheme may be a subset of the maximum achievable stability region. To improve the stability region, we propose and analyze another scheme which combines the power-of-two routing scheme with randomized state independent routing scheme. We show that this new scheme achieves the maximum stability region and results in the least mean sojourn time of jobs among all the schemes considered in the paper.**

I. Introduction

Information systems with large number of servers to process incoming job requests have become commonplace with the recent growth of data centres and web server farms [1]. The services provided by these systems such as online search, social networking etc. are extremely delay sensitive since a small increase in the average response time of jobs may cause significant loss of revenue and users [2]. Therefore, an important problem in this domain is to decide which server an incoming job should be routed to in order to reduce the mean response time of jobs.

The join-the-shortest-queue (JSQ) scheme, commonly used in small web server farms [1], [3], routes a new arrival to the server having the least number of unfinished jobs in the system. Recently, Gupta *et al* [4] showed that for a system of identical processor sharing (PS) servers JSQ is nearly optimal in terms of minimizing the mean sojourn time of jobs and results in near *insensitivity* of the system to the type of job length distribution.

There are two main disadvantages of using the JSQ scheme in a system consisting of a large number of servers with different service rates. Firstly, the JSQ scheme requires the queue length (number of unfinished jobs) information of all the servers in the system at each arrival epoch. Therefore, for a system having a large number of servers, the JSQ scheme incurs significant communication overhead between the router (job dispatcher) and the servers at each arrival instant of a new request. This results in an additional delay in the processing of each job. Secondly, in the JSQ scheme, server speeds are not taken into consideration in making routing decisions. In a system, where servers can have different service rates, the service rate of a server must also be taken into consideration in the criterion for server selection. This is because of the fact that a faster server with more number of unfinished jobs may be a better choice than a slower server with fewer unfinished jobs.

To avoid the above mentioned problems of the JSQ scheme we consider randomized routing schemes that combine scalability along with the server speed. Specifically, we consider variants of the classical power-of-two or the SQ(2) scheme [5]–[7] that applies the the JSQ rule to a subset of two randomly selected servers at each arrival instant.

We propose a modification of the SQ(2) scheme in which an incoming request is assigned to the server which provides the highest instantaneous processing rate per job among two randomly selected servers. We refer to this scheme as the HR(2) scheme.

In [5]–[9], the SQ(2) scheme was analyzed for homogeneous systems and exponential job length distribution using the mean-field approach. It was shown that, in the limit as the system size grows to infinity, the tail distribution of the server occupancies decay doubly exponentially. Similar results were obtained in [7], [9] for general job length distributions by assuming asymptotic independence of any finite subset of servers in the system. However, all these analyses were restricted to homogeneous system of servers.

Recently, in [10], we analyzed the performance of the classical SQ(2) scheme for the heterogeneous scenario. Along similar lines, in this paper, we present a detailed analysis of the proposed HR(2) scheme for the heterogeneous scenario. We analytically characterize the stability region, stationary load distribution and the mean sojourn time of jobs for the HR(2) scheme in the limit as the system size grows to infinity. It is shown that, as in the homogeneous case, the stationary tail

A part of this work was reported without proofs in a short paper at the ACM Sigmetrics 2014 conference.

978-0-9836283-9-2

distribution of server occupancies decays doubly exponentially and is *asymptotically insensitive* to the job size distribution. However, unlike the homogeneous case, in the heterogeneous case, the stability region the HR(2) scheme is shown to be a subset of the maximal stability region obtained by restricting the normalized arrival rate below the average capacity of the system.

To extend the stability region, we then propose and analyze a second scheme which combines the classical SQ(2) routing scheme with a state independent randomized routing scheme that chooses probabilities according to server classes to minimize the average sojourn time. In this scheme, upon arrival of a new job, a service rate is first chosen with a fixed probability from the set of possible service rates (speeds). The job is then routed to a server having the least number of unfinished jobs among two randomly selected servers having the chosen service rate. The probabilities of selecting different service rates are chosen so that the mean sojourn time of jobs in the system is minimized. We refer to this scheme as the hybrid SQ(2) scheme. It is shown that, by using the hybrid SQ(2) scheme, the maximal stability region can be recovered. We also analytically characterize the optimal probabilities for service rate selection for this scheme. Finally, numerical results are presented to compare the different schemes in terms of mean sojourn time of jobs and to demonstrate the accuracy of the derived asymptotic results in predicting the performance of large data centres. We conclude that the hybrid SQ(2) scheme outperforms all other the schemes considered in this paper and is *asymptotically insensitive* to job length distributions.

The rest of the paper is organized as follows. In Section II, we introduce the system model and describe the routing schemes studied in this paper. The routing schemes are then analyzed in Section III. Section IV provides numerical results to validate the theoretical results and to compare different routing schemes. The paper is finally concluded in Section V.

II. System Model

Consider a system consisting of N parallel processor sharing servers with heterogeneous service rates or capacities. The service rate, C, of a server is defined as the time rate at which it processes a single job assigned to it. If $x(t)$ jobs are present at a server of capacity C at time t, then the rate at which each job is processed at time t is given by $C/x(t)$. We assume that a server can have one of the M possible values of service rate from the set $\mathcal{C} = \{C_1, C_2, \ldots, C_M\}$. Define the index set $\mathcal{J} = \{1, 2, \ldots, M\}$. For each $j \in \mathcal{J}$, let the proportion of servers with service rate C_j be denoted by γ_j $(0 \leq \gamma_j \leq 1)$. Clearly, we have $\sum_{j=1}^{M} \gamma_j = 1$.

Jobs arrive at the system according to a Poisson process with rate $N\lambda$. Job lengths are assumed to be independent and identically distributed with a common finite mean $1/\mu$. The inter arrival times and the job lengths are assumed to be independent of each other. Upon arrival, a job is assigned to one of the N servers in the system by a router or job dispatcher. The job leaves the system after completion of its service at the server. We now discuss the routing schemes analyzed in this paper.

A. Scheme 1: Optimal state independent scheme

In this scheme, an incoming job is assigned to a server with a fixed probability, independent of the current states of the servers in the system. We denote by p_j, $j \in \mathcal{J}$, the probability that an incoming job is assigned to any one of the servers having capacity C_j. The probabilities p_j, $j \in \mathcal{J}$, are so chosen that the mean sojourn time of the jobs in the system is minimized under this scheme. Clearly, in this scheme, no communication is required between the router and the servers as the routing decisions are made independently of the state of the system. This scheme serves as a basis of comparison with the power-of-two type schemes that do require the knowledge of current states of some servers in the system.

B. Scheme 2: The HR(2) scheme

In this scheme, upon arrival of a new job, the router selects two servers uniformly at random from the set of N servers. The job is then assigned to the server that provides the highest instantaneous processing rate to each job present in it among the two selected servers. Note that the instantaneous processing rate of a job at a PS server is the server capacity divided by the instantaneous number of unfinished jobs at the server. If the two selected servers have the same instantaneous processing rate per job, then the job is assigned to any one of the servers with equal probability 0.5.

We note that this scheme is the same as the classical power-of-two or SQ(2) scheme except that the metric for choosing the destination server is the instantaneous processing rate per job at a server instead of simply the instantaneous number of unfinished jobs. Clearly, the HR(2) scheme reduces to the SQ(2) scheme in the homogeneous case.

C. Scheme 3: Hybrid SQ(2) scheme

In this scheme, upon arrival of a new job, the router first chooses a capacity value C_j, $j \in \mathcal{J}$, with probability p_j. Two servers having the selected value of capacity are then chosen uniformly at random from set of available servers having that capacity. The job is then routed to the server having the least number of unfinished jobs among the two chosen servers. Ties are broken by tossing a fair coin. The probabilities p_j, for $j \in \mathcal{J}$, are chosen in such a way that the mean sojourn time of jobs in the system is minimized. Hence, this scheme combines the classical SQ(2) scheme with the probabilistic routing scheme.

III. Analysis

In this section, we present the analysis of the load balancing schemes described in the previous section. For Scheme 1 we only state, without proof, the key result. The detailed analysis can be found in [11]. The analyses of Scheme 2 and Scheme 3 are provided in detail in Section III-B and Scheme III-C, respectively.

978-0-9836283-9-2

A. Scheme 1: Optimal state independent scheme

In Scheme 1, a job is assigned to a server with a fixed probability, independent of the instantaneous states of the servers in the system. Hence, under this scheme, the system reduces to a set of independent parallel M/G/1 processor sharing servers. It follows directly from Proposition 1 of [11], that there exists routing probabilities p_j, $j \in \mathcal{J}$, for which the system is stable under Scheme 1 if and only if the following condition holds:

$$\lambda \in \Lambda = \left\{ 0 < \lambda < \mu \sum_{j=1}^{M} \gamma_j C_j \right\}. \tag{1}$$

In the above condition, the set Λ, obtained by restricting the normalized arrival rate (λ) below the average capacity of the system, is referred to as the maximal stability region. It was shown in [11] that, under condition (1), the routing probabilities p_j, $j \in \mathcal{J}$, can be chosen so that the mean sojourn time of jobs in the system is minimized. The analytical expressions for the optimal routing probabilities are given in Theorem 1 of [11] and, therefore, are not repeated here.

B. Scheme 2: The HR(2) scheme

In the HR(2) scheme, the job assignments are done based on the instantaneous states of two randomly selected servers in the system. Hence, unlike the state independent scheme, in this scheme, the arrival processes to the individual servers are *not* independent of each other. This makes the exact analytical computation of stationary load distribution very difficult for finite values of the system size N. However, the mean field approach outlined in [6], [12] and the propagation of chaos argument used in [7], [9], [13] allow us to analytically characterize the behaviour of the system under this scheme in the limit as $N \to \infty$. It will be later shown through simulations that such asymptotic analysis closely approximates the behaviour of a large but finite system of servers. Therefore, our analytical results are applicable to data centres that typically run thousands of servers.

We first derive the stability region of the system under the HR(2) scheme. We keep the proportions γ_j, $j \in \mathcal{J}$, fixed and increase the system size N to observe the evolution of stability region. Assume that, for each $j \in \mathcal{J}$, γ_j is a rational number in $[0, 1]$. Let N^* denote the minimum positive integer (> 2) such that $\gamma_j N^*$ is a positive integer for each $j \in \mathcal{J}$. Now, let Λ_k, $k \in \mathbb{N}$, denote the stability region of the system under Scheme 2 when there are $N = kN^*$ servers in the system. The following proposition characterizes the sets Λ_k for $k \in \mathbb{N}$.

Proposition *1:* For the sets Λ_k, $k \in \mathbb{N}$, and Λ defined in (1), we have

$$\Lambda \supseteq \Lambda_1 \supseteq \Lambda_2 \supseteq \dots \tag{2}$$

Furthermore, if $\Lambda_\infty = \cap_{k=1}^{\infty} \Lambda_k$ denotes the intersection of the sets Λ_k, $k \in \mathbb{N}$, then Λ_∞ is given by

$$\Lambda_\infty = \left\{ 0 < \lambda < \mu \min_{\mathcal{I} \subseteq \mathcal{J}} \left\{ \frac{\left(\sum_{j \in \mathcal{I}} \gamma_j C_j\right)}{\left(\sum_{j \in \mathcal{I}} \gamma_j\right)^2} \right\} \right\} \tag{3}$$

Proof: The proof is given in Appendix A. ■

Remark *1:* From the above proposition it is clear that for any finite value of N ($= kN^*$), the stability region under the HR(2) scheme (Λ_k) is a subset of that under Scheme 1 (Λ). Further, the stability region under the HR(2) scheme shrinks as N increases keeping the proportions γ_j, $j \in \mathcal{J}$, fixed. Finally, we note that the set Λ_∞ denotes the stability region of the system as $N \to \infty$.

Remark *2:* Under the notation $\rho_j = \lambda / \mu C_j$, it is easy to see that the condition $\lambda < \mu \min_{\mathcal{I} \subseteq \mathcal{J}} \left\{ \frac{\left(\sum_{j \in \mathcal{I}} \gamma_j C_j\right)}{\left(\sum_{j \in \mathcal{I}} \gamma_j\right)^2} \right\}$ in (3) can be equivalently expressed as

$$\sum_{j \in \mathcal{I}} \frac{\gamma_j}{\rho_j} > \left(\sum_{j \in \mathcal{I}} \gamma_j \right)^2 \quad \text{for all } \mathcal{I} \subseteq \mathcal{J}. \tag{4}$$

In view of Proposition 1, it is evident that if (4) holds (i.e., if $\lambda \in \Lambda_\infty$), then $\lambda \in \Lambda_k$ for all $k \in \mathbb{N}$. In other words, under (4), the system is stable under Scheme 2 for any N.

1) Mean Field Analysis: We now proceed to find the stationary load distribution of the limiting system under the assumption of exponential job length distribution with mean $1/\mu$. The following notation is used: For any $w \in \mathbb{R}$, $\lfloor w \rfloor$ denotes the greatest integer not exceeding w and $\lceil w \rceil$ denotes the smallest integer greater than or equal to w. Let $\mathbf{x}_N(t) = \left\{ x_n^{(j)}(t), 1 \le j \le M, n \in \mathbb{Z}_+ \right\}$ denote the state of the system at time t, where $x_n^{(j)}(t) = \frac{1}{N\gamma_j} \sum_{n' \ge n} y_{n'}^{(j)}(t)$ and $y_n^{(j)}(t)$ is the number of servers having capacity C_j with exactly n unfinished jobs. Hence, $x_n^{(j)}(t)$ denotes the fraction of servers having capacity C_j with at least n unfinished jobs. Clearly, for any N, the process $\mathbf{x}_N(t)$ is a Markov process. The state space of the process $\mathbf{x}_N(t)$ is given by $\prod_{j=1}^{M} \bar{\mathcal{U}}_N^{(j)}$, where $\bar{\mathcal{U}}_N^{(j)}$ is defined as follows:

$$\bar{\mathcal{U}}_N^{(j)} = \{ g = (g_n, n \in \mathbb{Z}_+) : g_0 = 1, \\ g_n \ge g_{n+1} \ge 0, N\gamma_j g_n \in \mathbb{N} \ \forall n \in \mathbb{Z}_+ \}. \tag{5}$$

We generalize the space $\bar{\mathcal{U}}_N^{(j)}$ to the space $\bar{\mathcal{U}}$ of decreasing sequences of positive real numbers in $[0, 1]$ by removing the last constraint in its definition (5). Hence, the space $\bar{\mathcal{U}}$ is defined as follows: $\bar{\mathcal{U}} = \{ g = (g_n, n \in \mathbb{Z}_+) : g_0 = 1, g_n \ge g_{n+1} \ge 0 \ \forall n \in \mathbb{Z}_+ \}$. This space will be required to study the limiting properties of the process $\mathbf{x}_N(t)$ as $N \to \infty$.

The aim of the analysis is to prove the weak convergence of the process $\mathbf{x}_N(t)$ as $N \to \infty$ to the deterministic process $\mathbf{u}(t) = \left\{ u_n^{(j)}(t), n \in \mathbb{Z}_+, 1 \le j \le M \right\}$, governed by the following system of differential equations:

$$\mathbf{u}(0) = \mathbf{g}, \tag{6}$$

$$\dot{\mathbf{u}}(t) = \mathbf{h}(\mathbf{u}(t)), \tag{7}$$

where $\mathbf{g} \in \bar{\mathcal{U}}^M$ and for $j \in \mathcal{J}$,

$$h_0^{(j)}(\mathbf{u}) = 0, \tag{8}$$

$$h_n^{(j)}(\mathbf{u}) = \lambda\left(u_{n-1}^{(j)} - u_n^{(j)}\right)\sum_{i=1}^{M}\gamma_i\left(u^{(i)}_{\lceil (n-1)C_i/C_j\rceil} + u^{(i)}_{\lfloor (n-1)C_i/C_j\rfloor+1}\right) - \mu C_j\left(u_n^{(j)} - u_{n+1}^{(j)}\right), \tag{9}$$

for all $n \geq 1$. In other words, we try to prove that if the distribution of $\mathbf{x}_N(0)$ converges to the Dirac measure concentrated at the point $\mathbf{g} \in \bar{\mathcal{U}}^M$ as $N \to \infty$, then for each $t \geq 0$ the distribution of $\mathbf{x}_N(t)$ converges to the Dirac measure concentrated at the point $\mathbf{u}(t)$ on the trajectory of (6)-(9). Proposition 4 along with Theorem 2.5 of Chapter 4 of [14] establishes this weak convergence. Moreover, we show that, under the stability condition (4), the stationary distribution π_N (which exists by Remark 2) of the process $\mathbf{x}_N(t)$ converges weakly to the Dirac measure concentrated at the point $\mathbf{P} = \left\{P_k^{(j)}, k \in \mathbb{Z}_+, 1 \leq j \leq M\right\}$ which is the unique equilibrium point of the differential system (6)-(9) obtained by solving the equation $\dot{\mathbf{u}}(t) = \mathbf{h}(\mathbf{u}(t)) = 0$. In the following lemma we first summarize some important properties of the equilibrium point $\mathbf{P}$ of the system (6)-(9).

Proposition *2:* If there exists a solution $\mathbf{P}$ of the equation $\mathbf{h}(\mathbf{P}) = 0$ such that for each $j \in \mathcal{J}$, $P_0^{(j)} = 1$ and $P_k^{(j)} \downarrow 0$ as $k \to \infty$, then

i) for each $k \in \mathbb{Z}_+$ and $j \in \mathcal{J}$

$$P_{k+1}^{(j)} = \rho_j \sum_{l=k}^{\infty}\sum_{i=1}^{M}\gamma_i\left(P_l^{(j)} - P_{l+1}^{(j)}\right) \times \left(P^{(i)}_{\lceil lC_i/C_j\rceil} + P^{(i)}_{\lfloor lC_i/C_j\rfloor+1}\right) \tag{10}$$

ii) for each $k \in \mathbb{Z}_+$ and $j \in \mathcal{J}$

$$\sum_{j=1}^{M}\frac{\gamma_j}{\rho_j}P_{k+1}^{(j)} = \left(\sum_{j=1}^{M}\gamma_j P_k^{(j)}\right)^2 \tag{11}$$

iii) the sequence $\left\{P_k^{(j)}, k \in \mathbb{Z}_+\right\}$ decreases doubly exponentially.

Proof: The proof is given in Appendix B. ∎

Remark *3:* A real sequence $\{z_n\}_{n\geq 1}$ is said to decrease doubly exponentially if and only if there exist positive constants L, $\omega < 1$, $\theta > 1$, and κ such that $z_n \leq \kappa\omega^{\theta^n}$ for all $n \geq L$. Hence, by definition, if a sequence $\{z_n\}_{n\geq 1}$ decays doubly exponentially, then it is summable, i.e., $\sum_{n=1}^{\infty} z_n < \infty$. Hence, in view of Proposition 2.iii), if there exists a solution $\mathbf{P}$ of the equation $\mathbf{h}(\mathbf{P}) = 0$ satisfying the hypothesis of Proposition 2, then it must be summable.

Before proving the weak convergence of the Markov process $\mathbf{x}_N(t)$ to the deterministic process $\mathbf{u}(t)$ defined by the systems (6)-(9), we show that the system indeed has a unique solution in $\bar{\mathcal{U}}^M$ and there exists a unique equilibrium point $\mathbf{P}$ of it satisfying $\sum_{k=1}^{\infty} P_k^{(j)} < \infty$ for each $j \in \mathcal{J}$. To do so, it is convenient to define the space $\mathcal{U}$ of tail probability distributions on $\mathbb{Z}_+$ having finite first moment as follows

$$\mathcal{U} = \left\{g = (g_n, n \in \mathbb{Z}_+) : g_0 = 1, \; g_n \geq g_{n+1} \geq 0 \; \forall n \in \mathbb{Z}_+, \sum_{n=0}^{\infty} g_n < \infty\right\}. \tag{12}$$

and the following norm on the spaces $\prod_{j=1}^{M}\bar{\mathcal{U}}_N^{(j)}$, $\bar{\mathcal{U}}^M$, and $\mathcal{U}^M$:

$$\|u\| = \sup_{1\leq j\leq M}\sup_{n\in\mathbb{Z}_+}\frac{\left|u_n^{(j)}\right|}{n+1}. \tag{13}$$

Note that the space $\bar{\mathcal{U}}^M$ is complete and compact under the above norm. Henceforth, this norm is understood when we refer to convergence or continuity in these spaces. The following proposition guarantees the existence and uniqueness of solution of the system (6)-(9) and its equilibrium point $\mathbf{P}$. To emphasize the dependence of the solution of the system (6)-(9) on the initial point $\mathbf{g}$, we shall, at times, denote the solution $\mathbf{u}(t)$ by $\mathbf{u}(t, \mathbf{g})$.

Proposition *3:* i) The system (6)-(9) has a unique solution, $\mathbf{u}(t, \mathbf{g})$, for all $t \geq 0$, in $\bar{\mathcal{U}}^M$ if $\mathbf{g} \in \bar{\mathcal{U}}^M$.

ii) Under condition (4), there exists a unique equilibrium point or fixed point $\mathbf{P}$ of the system (6)-(9) in the space $\mathcal{U}^M$. Therefore, $\mathbf{P}$ satisfies the properties stated in Lemma 2.

iii) Under condition (4),

$$\lim_{t\to\infty}\mathbf{u}(t, \mathbf{g}) = \mathbf{P} \text{ for all } \mathbf{g} \in \mathcal{U}^M. \tag{14}$$

Thus, there exists a unique probability measure π on $\mathcal{U}^M$ such that

$$\int f(\mathbf{g})d\pi(\mathbf{g}) = \int f(\mathbf{u}(t, \mathbf{g}))d\pi(\mathbf{g}) \tag{15}$$

for all $t \geq 0$, $f : \bar{\mathcal{U}}^M \to \mathbb{R}$; and $\pi = \delta_{\mathbf{P}}$. Here, $\delta_{\mathbf{P}}$ denotes the Dirac measure concentrated at the point $\mathbf{P}$.

Proof: The proof is given in Appendix C. ∎

We now proceed to establish the weak convergence as $N \to \infty$ of the process $\mathbf{x}_N(t)$ to the process $\mathbf{u}(t, \mathbf{g})$. This is done by showing that the generator of the process $\mathbf{x}_N(t)$ converges to the generator of the deterministic map $\mathbf{g} \mapsto \mathbf{u}(t, \mathbf{g})$ as $N \to \infty$. The weak convergence then follows from Theorem 2.5 of Chapter 4 of [14].

For the Markov process $\mathbf{x}_N(t)$, the generator $\mathbf{A}_N$ acting on functions $f : \prod_{j=1}^{M}\bar{\mathcal{U}}_N^{(j)} \to \mathbb{R}$ is defined as $\mathbf{A}_N f(\mathbf{g}) = \sum_{\mathbf{h}\neq\mathbf{g}} q_{\mathbf{gh}}\left(f(\mathbf{h}) - f(\mathbf{g})\right)$, where $q_{\mathbf{gh}}$, with $\mathbf{g}, \mathbf{h} \in \prod_{j=1}^{M}\bar{\mathcal{U}}_N^{(j)}$, denotes the transition rate from state $\mathbf{g}$ to state $\mathbf{h}$.

Lemma *1:* Let $\mathbf{g} \in \prod_{j=1}^{M} \bar{\mathcal{U}}_N^{(j)}$ and $\mathbf{e}(n,j) \in \prod_{j=1}^{M} \bar{\mathcal{U}}_N^{(j)}$ be a unit vector with $e_n^{(j)} = 1$ and $e_k^{(i)} = 0$ for all $i \neq j$, $k \neq n$. The generator $\mathbf{A}_N$ of the Markov process $\mathbf{x}_N(t)$ acting on functions $f : \prod_{j=1}^{M} \bar{\mathcal{U}}_N^{(j)} \to \mathbb{R}$ is given by

$$\begin{aligned}\mathbf{A}_N f(\mathbf{g}) &= \lambda N \sum_{n\geq 1}\sum_{j=1}^{M}\sum_{i=1}^{M} \gamma_i\gamma_j \left(g_{n-1}^{(j)} - g_n^{(j)}\right) \\ &\quad \times \left(g^{(i)}_{\lceil (n-1)C_i/C_j\rceil} + g^{(i)}_{\lfloor (n-1)C_i/C_j\rfloor+1}\right) \\ &\quad \times \left(f(\mathbf{g} + \frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{g})\right) \\ &+ N\sum_{n\geq 1}\sum_{j=1}^{M} \mu C_j\gamma_j \left(g_n^{(j)} - g_{n+1}^{(j)}\right)\left(f(\mathbf{g} - \frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{g})\right),\end{aligned}$$

Proof: The proof is given in Appendix D ∎

For $t \geq 0$, the transition semigroup operator $\mathbf{T}_N(t)$ generated by the operator $\mathbf{A}_N$ and acting on functions $f : \prod_{j=1}^{M} \bar{\mathcal{U}}_N^{(j)} \to \mathbb{R}$ is defined by $\mathbf{T}_N(t)f = \exp(t\mathbf{A}_N)\, f$. The following proposition establishes the convergence of the semigroup $\mathbf{T}_N(t)$ to the semigroup of the of the map $\mathbf{g} \mapsto \mathbf{u}(t,\mathbf{g})$.

Proposition *4:* For any continuous function $f : \bar{\mathcal{U}}^M \to \mathbb{R}$ and $t \geq 0$,

$$\lim_{N\to\infty} \sup_{\mathbf{g}\in\prod_{j=1}^{M}\bar{\mathcal{U}}_N^{(j)}} |\mathbf{T}_N(t)f(\mathbf{g}) - f(\mathbf{u}(t,\mathbf{g}))| = 0 \qquad (16)$$

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

Proof: The proof is given in Appendix E ∎

We now show that, under condition (4), the stationary distribution, π_N of the process weakly converges to $\delta_{\mathbf{P}}$, the Dirac measure concentrated at the unique equilibrium point $\mathbf{P} \in \mathcal{U}^M$ of the system (6)-(9).

Proposition *5:* Under the condition (4), the Markov process $\mathbf{x}_N(t)$ is positive recurrent for all N and hence has a unique invariant distribution π_N for each N. Moreover, $\pi_N \to \delta_{\mathbf{P}}$ weakly as $N \to \infty$, where $\delta_{\mathbf{P}}$ is as defined in Proposition 3, i.e., $\lim_{N\to\infty} \mathbb{E}_{\pi_N} f(\mathbf{g}) = f(\mathbf{P})$, for all continuous functions $f : \bar{\mathcal{U}}^M \to \mathbb{R}$.

Proof: The first part of the theorem is a direct consequence of Remark 2 following Proposition 1. The weak convergence of the stationary distributions π_N to $\delta_{\mathbf{P}}$ follows by using the same line of arguments as in the proof of Theorem 4.(ii) of [12]. ∎

Remark *4:* We have thus established the following weak convergence result:

$$\lim_{t\to\infty}\lim_{N\to\infty} \mathbf{x}_N(t) = \lim_{N\to\infty}\lim_{t\to\infty} \mathbf{x}_N(t) = \mathbf{P} \qquad (17)$$

Therefore, we conclude that, if job length distribution is exponential, then, in the large system limit ($N \to \infty$), the stationary tail distribution of the number of unfinished jobs at the servers is given by $\mathbf{P} \in \mathcal{U}^M$, which decays doubly exponentially for each $j \in \mathcal{J}$.

2) Insensitivity: We now show that the assumption of exponential that the job length distributions can be relaxed to allow for general job length distributions with finite mean and proceed to find the stationary load distribution of the limiting system for general job length distributions.

Suppose that as $N \to \infty$, the servers of the system become independent of each other. More formally, suppose that under condition (4), as $N \to \infty$ the joint stationary distribution of the numbers of unfinished jobs at the servers converges to a unique distribution Π on $\mathbb{Z}_+^\infty$. Furthermore, if $\Pi^{(B)}$ denotes the restriction Π to a finite subset B of servers and $\pi^{(n)}$ denotes the one dimensional marginal of Π for the n^{th} server, then

$$\Pi^{(B)} = \bigotimes_{n\in B} \pi^{(n)}. \qquad (18)$$

This is referred to as the *asymptotic independence* property which was originally conjectured in [7] for homogeneous systems. We show that the stationary load distribution under asymptotic independence coincides with the unique equilibrium point $\mathbf{P}$ of the system (6)-(9).

We now proceed to analyze the system under the above assumption. Consider a tagged server in the system. The server is chosen as one of the two possible destination servers for a new arrival with probability $\left(1 - \frac{\binom{N-1}{2}}{\binom{N}{2}}\right) = \frac{2}{N}$. Such an arrival is called a *potential arrival* to the tagged server. Hence, the potential arrival process to the tagged server is Poisson with rate $N\lambda \times \frac{2}{N} = 2\lambda$. The potential arrival is actually assigned to the tagged server depending on the state of the other possible destination server. Hence, unlike the potential arrival process, the actual arrival process to the tagged server is not Poisson. However, as $N \to \infty$, due to the asymptotic independence assumption, the two possible destination servers become independent of each other. As a result, the actual arrival process to the tagged server converges to a state dependent Poisson process as $N \to \infty$. The following proposition relates state-dependent arrival rates of the actual arrival process to the stationary load distribution.

Proposition *6:* For Scheme 2, under the condition (4), the arrival rate, $\lambda_k^{(j)}$, to a tagged server of capacity C_j when it has $k \in \mathbb{Z}_+$ unfinished jobs is given by

$$\lambda_k^{(j)} = \lambda \sum_{i=1}^{M} \gamma_i \left(P^{(i)}_{\lceil kC_i/C_j\rceil} + P^{(i)}_{\lfloor kC_i/C_j\rfloor+1}\right), \qquad (19)$$

where $P_k^{(j)}$, for $j \in \mathcal{J}$ and $k \in \mathbb{Z}_+$, denotes the stationary probability that a server with capacity C_j has at least k unfinished jobs. Further, we have $P_0^{(j)} = 1$, for all $j \in \mathcal{J}$ and $P_k^{(j)}$, for $k \in \mathbb{Z}_+$ and $j \in \mathcal{J}$, satisfy (10).

Proof: Let a potential arrival occur at the tagged server having capacity C_j when the number of unfinished jobs at the server is k. Further, let the second possible destination server have capacity C_i. Under the above scenario, the potential arrival actually joins the tagged server with probability 1 if the number unfinished jobs, l,

at the other possible destination server satisfies $\frac{l}{C_i} > \frac{k}{C_j}$ or equivalently $l \geq \lfloor kC_i/C_j \rfloor + 1$. The potential arrival joins the server with capacity C_j with probability 0.5 when $l/C_i = k/C_j$ or equivalently $l = kC_i/C_j$. Since a server having capacity C_i is chosen as a possible destination server with probability γ_i, the total probability that the potential arrival joins the server with capacity C_j in state k is $\sum_{j=1}^{M} \gamma_j \left(0.5\left(P^{(j)}_{\lceil kC_i/C_j \rceil} - P^{(j)}_{\lfloor kC_i/C_j \rfloor + 1}\right) + P^{(j)}_{\lfloor kC_i/C_j \rfloor + 1}\right) = 0.5 \sum_{j=1}^{M} \gamma_j \left(P^{(j)}_{\lceil kC_i/C_j \rceil} + P^{(j)}_{\lfloor kC_i/C_k \rfloor + 1}\right)$. Therefore, the rate at which arrivals occur to the given server at state k is given by $\lambda^{(j)}_k = 2\lambda \times 0.5 \sum_{j=1}^{M} \gamma_j \left(P^{(j)}_{\lceil kC_i/C_j \rceil} + P^{(j)}_{\lfloor kC_i/C_k \rfloor + 1}\right)$. This simplifies to (19).

Since processor sharing is a symmetric service discipline, it follows from Theorems 3.10 and 3.14 of [15] and also from Theorem 4.2 of [4] that the detailed balance equations hold for state dependent Poisson arrival rates and any job length distribution. Hence, for $k \in \mathbb{Z}_+$ and $j \in \mathcal{J}$, we have $P^{(j)}_{k+1} - P^{(j)}_{k+2} = \frac{\lambda^{(j)}_k}{\mu C_j}(P^{(j)}_k - P^{(j)}_{k+1}),$ for $j \in \mathcal{J}$. Substituting the value of $\lambda^{(j)}_k$ from (19) and upon further simplification we get (10). ∎

Remark *5:* Hence, from Proposition 6, we conclude that the stationary tail distribution of loads at each server in the limiting system remains the same for all job length distribution so long as the mean of the distribution remains unchanged. We refer to this property as the *asymptotic insensitivity* property.

Remark *6:* The long run probability that a user joins a server with capacity C_j is given by $\frac{N\gamma_j \bar{\lambda}^{(j)}}{N\lambda} = \frac{\gamma_j \bar{\lambda}^{(j)}}{\lambda}$, where, $\bar{\lambda}^{(j)} = \sum_{k=0}^{\infty} \lambda^{(j)}_k \left(P^{(j)}_k - P^{(j)}_{k+1}\right)$ denotes the average arrival rate to a server having capacity C_j. From (19) and (10), we obtain that $\frac{\gamma_j \bar{\lambda}^{(j)}}{\lambda} = \gamma_j \frac{P^{(j)}_1}{\rho_j}$ for each $j \in \mathcal{J}$. Thus, the long run probability that a job is routed to a server with capacity C_j is $\gamma_j \frac{P^{(j)}_1}{\rho_j}$.

We now find the relationship between the mean sojourn time of jobs under Scheme 2 and the stationary tail distribution derived in (10).

Proposition *7:* The mean sojourn time of jobs, $\bar{T}$, under Scheme 2 is given by $\bar{T} = \frac{1}{\lambda} \sum_{j=1}^{M} \sum_{k=1}^{\infty} \gamma_j P^{(j)}_k$, where $P^{(j)}_k$, $k \in \mathbb{Z}_+$ and $j \in \mathcal{J}$, are as given in Proposition 6.

Proof: Let $\bar{T}_j$ denote the mean sojourn time of a user given that it has joined a server having capacity C_j. Now, the expected number of users at a server having capacity C_j is given by $\sum_{k=1}^{\infty} P^{(j)}_k$. Let the average arrival rate at the server be denoted by $\bar{\lambda}^{(j)}$. Thus, applying Little's formula we have $\bar{T}_j = \frac{\sum_{k=1}^{\infty} P^{(j)}_k}{\bar{\lambda}^{(j)}}$

As discussed in the second remark, the long run probability that a user joins a server having capacity C_j is $\frac{\gamma_j \bar{\lambda}^{(j)}}{\lambda}$. Therefore, the overall mean sojourn time is given by $\bar{T} = \sum_{j=1}^{M} \frac{\gamma_j \bar{\lambda}^{(j)}}{\lambda} \bar{T}_j = \frac{1}{\lambda} \sum_{j=1}^{M} \sum_{k=1}^{\infty} \gamma_j P^{(j)}_k$. ∎

C. Scheme 3: The hybrid SQ(2) scheme

In this scheme, for each $j \in \mathcal{J}$, the service rate $C_j \in \mathcal{C}$ is selected for a new arrival with probability p_j. Hence, the Poisson arrival rate to the set of $N\gamma_j$ servers having capacity C_j is $p_j N\lambda$. Hence, under this scheme the system can be viewed as being composed of M independent parallel homogeneous subsystems each working under the classical SQ(2) scheme. The j^{th} ($j \in \mathcal{J}$) subsystem has $N\gamma_j$ servers of capacity C_j and the total input rate at this subsystem is $p_j N\lambda$. Define $\rho_j = \frac{p_j \lambda}{\gamma_j \mu C_j}$. From the results of [6]–[8], [16], the system is stable if and only if $\rho_j < 1$ for all $j \in \mathcal{J}$. The necessary and sufficient condition which guarantees the existence of routing probabilities p_j, $j \in \mathcal{J}$ such that the system is stable is given by the following proposition.

Proposition *8:* There exists probabilities p_j, $j \in \mathcal{J}$, for which the system is stable under Scheme 3 if and only if $\lambda \in \Lambda$

Proof: Let us assume that $\lambda \in \Lambda$ holds. Now let $p_i = \frac{\gamma_i C_i}{\sum_{j=1}^{M} \gamma_j C_j}$, for all $i \in \mathcal{J}$. Using these values of p_i, $i \in \mathcal{J}$, we have $\rho_i = \frac{\lambda}{\mu \sum_{j=1}^{M} \gamma_j C_j} < 1$. Hence, condition $\lambda \in \Lambda$ is sufficient.

Now let $\frac{\lambda}{\mu \sum_{j=1}^{M} \gamma_j C_j} \geq 1$. For stability we must have $\rho_i < 1$ for all $i \in \mathcal{J}$. Hence, $\frac{\lambda}{\mu \sum_{j=1}^{M} \rho_j \gamma_j C_j} > 1$ which contradicts the fact that $\sum_{j=1}^{M} p_j = 1$ or $\frac{\lambda}{\mu \sum_{j=1}^{M} \rho_j \gamma_j C_j} = 1$. Hence, condition $\lambda \in \Lambda$ is necessary. ∎

Remark *7:* We observe that the stability regions for the hybrid SQ(2) scheme and Scheme 1 are the same. Therefore, the hybrid SQ(2) scheme achieves the maximal stability region.

Henceforth we assume that $\lambda \in \Lambda$ holds. We intend to find a vector $\mathbf{p}^* = \left\{p^*_j, j \in \mathcal{J}\right\}$ or equivalently a vector $\boldsymbol{\rho}^* = \left\{\rho^*_j, j \in \mathcal{J}\right\}$ such that the mean sojourn time of jobs in the limiting system is minimized. As in Proposition 7, it can be shown that the mean sojourn time of jobs in the limiting system under Scheme 3 is given by $\bar{T} = \frac{1}{\lambda} \sum_{j=1}^{M} \sum_{k=1}^{\infty} \gamma_j P^{(j)}_k$, where $P^{(j)}_k$, $j \in \mathcal{J}$ and $k \in \mathbb{Z}_+$, denotes the stationary probability that a server with capacity C_j in the limiting system has at least k unfinished jobs. From the results of [6]–[8] it is known that, for $j \in \mathcal{J}$ and $k \in \mathbb{Z}_+$, $P^{(j)}_k = \rho_j^{2^k - 1}$. Therefore, the mean sojourn time minimization problem can be formulated in terms of the loads ρ_j, $j \in \mathcal{J}$, as follows:

$$\begin{aligned} \underset{\boldsymbol{\rho}}{\text{Minimize}} \quad & \frac{1}{\lambda} \sum_{j \in \mathcal{J}} \gamma_j \sum_{k=1}^{\infty} \rho_j^{2^k - 1} \\ \text{subject to} \quad & 0 \leq \rho_j < 1, \text{for all } j \in \mathcal{J} \\ & \sum_{j \in \mathcal{J}} \gamma_j C_j \rho_j = \frac{\lambda}{\mu}. \end{aligned} \tag{20}$$

To characterize the solution of (20), we assume without loss of generality that the server capacities are ordered as follows: $C_1 \geq C_2 \geq \ldots \geq C_M$ Further, let $\mathcal{J}_{opt} \subseteq \mathcal{J}$ denote the index set of server capacities being used in the optimal scheme.

Proposition *9:* Let $\Phi : \mathbb{R}_+ \to [0, 1)$ be the inverse of the monotone mapping $\Phi^{-1} : [0, 1) \to \mathbb{R}_+$ defined as $\Phi^{-1}(\rho) = \sum_{k=1}^{\infty} \left(2^k - 1\right) \rho^{2^k - 2} < \sum_{x=1}^{\infty} x\rho^{x-1} < \infty$ for $0 < \rho < 1$. Further, for each $j \in \mathcal{J}$, let $\Psi_j : \mathbb{R}_+ \to \mathbb{R}_+$ denote the inverse of the monotone mapping $\Psi_j^{-1} : \mathbb{R}_+ \to \mathbb{R}_+$ defined

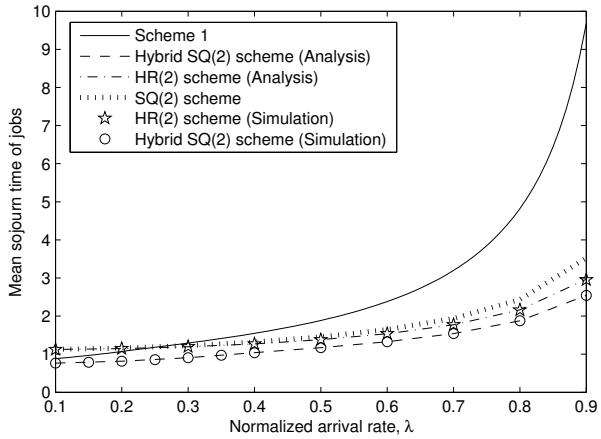

Fig. 1. Mean sojourn time jobs as a function of λ for $C_1 = 2/3$, $C_2 = 4/3$, $N = 200$ and $\gamma_1 = \gamma_2 = 1/2$

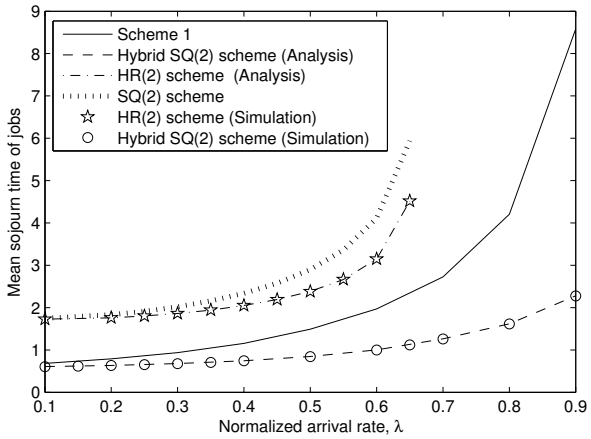

Fig. 2. Mean sojourn time jobs as a function of λ for $C_1 = 1/3$, $C_2 = 5/3$, $N = 200$, and $\gamma_1 = \gamma_2 = 1/2$

as $\Psi_j^{-1}(\theta) = \mu \sum_{i=1}^{j} \gamma_i C_i \Phi(\theta C_i)$. The index set of server capacities used in the optimal SQ(2) scheme is then given by $\mathcal{J}_{\text{opt}} = (1, 2, \ldots, j^*)$, where j^* is given by

$$j^* = \sup \left\{ j \in \mathcal{J} : \frac{1}{C_j} < \Psi_j(\lambda) \right\}. \tag{21}$$

Moreover, the optimal traffic intensities ρ_i^*, for $i \in \mathcal{J}$ satisfy

$$\rho_i^* = \begin{cases} \Phi(\Psi_{j*}(\lambda) C_i), & \text{if } i \in \mathcal{J}_{\text{opt}} \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

The proposition can be proved by solving the convex optimization problem (20) using Lagrange multipliers. We omit the proof due to space constraints.

The optimal routing probabilities p_j^*, $j \in \mathcal{J}$, and the minimum mean sojourn time $\bar{T}^*$ can be found from Proposition 9 by using the relations $\rho_j^* = \frac{p_j^* \lambda}{\gamma_j \mu C_j}$ and $\bar{T}^* = \frac{1}{\lambda} \sum_{i=1}^{j*} \gamma_i \sum_{k=1}^{\infty} (\rho_i^*)^{2^k - 1}$, respectively.

IV. Numerical Results

In this section, we present simulation results to compare the different load balancing schemes considered in this paper. The results also indicate the accuracy of the asymptotic analyses of the HR(2) scheme and the hybrid SQ(2) scheme in predicting their performance in a large but finite system of servers. We set $\mu = 1$ in all our simulations.

We choose $C_1 = 4/3$, $C_2 = 2/3$, $N = 200$ and $\gamma_1 = \gamma_2 = \frac{1}{2}$. Using conditions (1) and (4) it is found that for the above parameter setting the stability regions of all three schemes coincide and is given by $\lambda < 1$. In Figure 1, we plot the mean sojourn time jobs in the system as a function of the normalized arrival rate, λ, for the three schemes and also for the classical SQ(2) scheme. It is observed from the plot that the HR(2) scheme performs better than Scheme 1 for higher values of λ. It also outperforms the classical SQ(2) scheme. We see that among all the schemes, the hybrid SQ(2) scheme results in the least mean sojourn time of jobs in the system.

The performance of the HR(2) scheme need not always be better than that of Scheme 1. To demonstrate this fact we choose a second set of parameter values as follows: $C_1 = 5/3$, $C_2 = 1/3$, $N = 200$, and $\gamma_1 = \gamma_2 = 1/2$. Under this parameter setting, the stability region for Scheme 1 and the hybrid SQ(2) scheme are found to be equal to $\lambda < 1$. For the HR(2) scheme, however, the stability region is found to be $\lambda < 2/3$. In Figure 2, we plot the average response time of jobs as a function of λ for the three schemes and for the classical SQ(2) scheme. As expected, we observe that the HR(2) scheme still performs better than the classical SQ(2) scheme. However, in this case, the mean response time of jobs is lower in Scheme 1 than in the HR(2) scheme. Again, the hybrid SQ(2) scheme outperforms all the other schemes.

In Figures 1 and 2, we also observe a good match between the analysis and simulation results for the HR(2) scheme and the hybrid SQ(2) scheme for $N = 200$. The deviation of the asymptotic results from the simulation results is only about 4% for $N = 20$. This implies that the asymptotic results derived in the paper are not only applicable to data centres that typically run thousands of servers but also to server farms running moderately large number of servers.

Now we investigate the asymptotic insensitivity of the HR(2) scheme. In Table I, we show that the mean sojourn time of jobs obtained for the parameter setting $C_1 = 4/3$, $C_2 = 2/3$, $N = 200$ and $\gamma_1 = \gamma_2 = \frac{1}{2}$. We choose the following two job length distributions: i) constant, with distribution satisfying $F(x) = 0$ for $0 \leq x < 1$, and $F(x) = 1$, otherwise. ii) power law, with distribution satisfying $F(x) = 1 - 1/4x^2$ for $x \geq \frac{1}{2}$ and $F(x) = 0$, otherwise. It is seen that there is almost no change in the mean sojourn time of jobs when the job length distribution type is changed. The results, therefore, justify the asymptotic independence assumption stated in III-B.

TABLE I
INSENSITIVITY OF THE HR(2) SCHEME

λ	Mean sojourn time $\bar{T}$	Constant	Power Law
	Theoretical	Simulation	Simulation
0.2	1.1479	1.1475	1.1481
0.3	1.1956	1.1957	1.1960
0.5	1.3801	1.3800	1.3798
0.7	1.7735	1.7743	1.7745
0.8	2.1641	2.1628	2.1650
0.9	2.9604	2.9704	2.9710

V. CONCLUSION

In this paper, we considered randomized load balancing schemes for large heterogeneous processor sharing systems. The variant of the classical SQ(2) scheme, in which routing decisions are based on instantaneous processing rate per job, was analyzed in the limit as the system size grows to infinity. It was shown that the stationary tail distribution of loads at each server decreases doubly exponentially and is insensitive to the type of job length distribution as in the homogeneous case. However, unlike the homogeneous case, this scheme has a smaller stability region than the optimal state independent scheme in the heterogeneous case. The stability region of the optimal state independent scheme can be fully recovered by using a scheme that combines the SQ(2) scheme with the state independent scheme. The three schemes were compared on the basis of the mean sojourn time of jobs in the system. We conclude that the combination of the SQ(2) scheme with the state independent scheme results in the least mean sojourn time of jobs among all the schemes considered.

APPENDIX

A. Proof of Proposition 1

From condition (1.2) of [16] we found that for a value of N, the system is stable under Scheme 2 if the following condition is satisfied:

$$\max_{\mathcal{B}\subseteq\mathcal{S}}\left\{\left(\sum_{i\in\mathcal{B}}C_{(i)}\right)^{-1}\frac{N\lambda}{\mu}\frac{\binom{|\mathcal{B}|}{2}}{\binom{N}{2}}\right\}<1, \tag{23}$$

where $\mathcal{S}=\{1,2,\ldots,N\}$ denotes the index set of servers, $\mathcal{B}\subseteq\mathcal{S}$ and $C_{(k)}\in\mathcal{C}$ denotes the capacity of the k^{th} server in the system. Thus, for $N=kN^*$, the set Λ_k is given by

$$\Lambda_k=\left\{\lambda>0:\left(\sum_{i\in\mathcal{B}}C_{(i)}\right)^{-1}\frac{N\lambda}{\mu}\frac{\binom{|\mathcal{B}|}{2}}{\binom{N}{2}}<1\ \forall\ \mathcal{B}\subseteq\mathcal{S}_k\right\} \tag{24}$$

where $\mathcal{S}_k=\{1,2,\ldots,kN^*\}$. Clearly, for integers l and k, with $l\geq k$, we have $\mathcal{S}_k\subseteq\mathcal{S}_l$. Hence, if $\mathcal{B}\subseteq\mathcal{S}_k$, then $\mathcal{B}\subseteq\mathcal{S}_l$. Therefore, from (24) it is clear that $\lambda\in\Lambda_l$ implies $\lambda\in\Lambda_k$. Consequently, for $l\geq k$ we have $\Lambda_k\supseteq\Lambda_l$. Further, if we set $\mathcal{B}=\mathcal{S}$ in (23) then we get (1). Hence, for all $k\in\mathbb{N}$, $\Lambda\supseteq\Lambda_k$. This proves (2).

To prove (3), let us consider a finite value of N and a set $\mathcal{I}\subseteq\mathcal{J}$. Let $\mathcal{B}_\mathcal{I}\subseteq\mathcal{S}$ be a subset of servers in which there are a_i ($0<a_i\leq N\gamma_i$) servers of capacity C_i for each $i\in\mathcal{I}$. It can be easily checked that $\frac{\left(\sum_{i\in\mathcal{I}}a_i\right)\left(\sum_{i\in\mathcal{I}}a_i-1\right)}{\sum_{i\in\mathcal{I}}a_iC_i}$ is an increasing function in each of the variables a_i. Therefore, we have

$$\begin{aligned}\left(\sum_{i\in\mathcal{B}_\mathcal{I}}C_{(i)}\right)^{-1}\frac{N\lambda}{\mu}\frac{\binom{|\mathcal{B}_\mathcal{I}|}{2}}{\binom{N}{2}}&=\frac{\lambda}{\mu}\frac{\left(\sum_{i\in\mathcal{I}}a_i\right)\left(\sum_{i\in\mathcal{I}}a_i-1\right)}{\left(\sum_{i\in\mathcal{I}}a_iC_i\right)(N-1)}\\&\leq\frac{\lambda}{\mu}\frac{\left(\sum_{i\in\mathcal{I}}N\gamma_i\right)\left(\sum_{i\in\mathcal{I}}N\gamma_i-1\right)}{\left(\sum_{i\in\mathcal{I}}N\gamma_iC_i\right)(N-1)}\\&\leq\frac{\lambda}{\mu}\frac{\left(\sum_{i\in\mathcal{I}}N\gamma_i\right)\left(\sum_{i\in\mathcal{I}}N\gamma_i\right)}{\left(\sum_{i\in\mathcal{I}}N\gamma_iC_i\right)(N)}\\&=\frac{\lambda}{\mu}\frac{\left(\sum_{i\in\mathcal{I}}\gamma_i\right)^2}{\left(\sum_{i\in\mathcal{I}}\gamma_iC_i\right)}\end{aligned}$$

Hence, $\lambda\in\Lambda_\infty$ implies $\left(\sum_{i\in\mathcal{B}_\mathcal{I}}C_{(i)}\right)^{-1}\frac{N\lambda}{\mu}\frac{\binom{|\mathcal{B}_\mathcal{I}|}{2}}{\binom{N}{2}}<1$. As this is true for any $\mathcal{I}\subseteq\mathcal{J}$ and any N, we have that $\Lambda_\infty\subseteq\Lambda_k$ for all $k\in\mathbb{N}$. Hence, $\Lambda_\infty\subseteq\cap_{k=1}^\infty\Lambda_k$. To prove the reverse inclusion, consider $\lambda\in\cap_{k=1}^\infty\Lambda_k$. For $\mathcal{I}\subseteq\mathcal{J}$, consider a set $\mathcal{B}_\mathcal{I}^{(N)}$ which contains all the $N\gamma_i$ servers of capacity C_i for each $i\in\mathcal{I}$. Since $\lambda\in\Lambda_k$ for all $k\in\mathbb{N}$, we have $\lim_{N\to\infty}\left(\sum_{i\in\mathcal{B}_\mathcal{I}}C_{(i)}\right)^{-1}\frac{N\lambda}{\mu}\frac{\binom{|\mathcal{B}_\mathcal{I}^{(N)}|}{2}}{\binom{N}{2}}<1$, which is equivalent to the condition $\frac{\lambda}{\mu}\frac{\left(\sum_{i\in\mathcal{I}}\gamma_i\right)^2}{\left(\sum_{i\in\mathcal{I}}\gamma_iC_i\right)}<1$. As this is true for all $\mathcal{I}\subseteq\mathcal{J}$, we have $\lambda\in\Lambda_\infty$. Hence, $\Lambda_\infty=\cap_{k=1}^\infty\Lambda_k$ as required.

B. Proof of Proposition 2

i) Let $\mathbf{P}$ satisfy the hypothesis of the proposition. Hence, from (9), we have that, for each $l\in\mathbb{Z}_+$ and $j\in\mathcal{J}$,

$$\begin{aligned}P_{l+1}^{(j)}-P_{l+2}^{(j)}=\rho_j\left(P_l^{(j)}-P_{l+1}^{(j)}\right)\\\times\sum_{i=1}^M\gamma_i\left(P_{\{lC_i/C_j\}}^{(i)}+P_{[lC_i/C_j]+1}^{(i)}\right).\end{aligned} \tag{25}$$

Since by hypothesis $P_l^{(j)}\to0$ as $l\to\infty$, adding the above equations for $l\geq k$ yields (10) upon simplification.

ii) Equation (11) is a direct consequence of (10).

iii) From (11) we obtain $\frac{\gamma_j}{\rho_j}P_{k+1}^{(j)}\leq\left(\sum_{j=1}^M\gamma_jP_k^{(j)}\right)^2\leq\left(\tilde{P}_k\right)^2$, where $\tilde{P}_k=\max_{1\leq j\leq M}P_k^{(j)}$. Thus, we have $P_{k+1}^{(j)}\leq\delta\tilde{P}_k$, where $\delta=\tilde{P}_k\max_{1\leq j\leq M}(\rho_j/\gamma_j)$. Since by hypothesis, for each j, $P_k^{(j)}\to0$ as $k\to\infty$, one can choose k sufficiently large such that $\delta<1$. Hence, we have $\left(\max_{1\leq j\leq M}P_{k+1}^{(j)}\right)\leq\delta\tilde{P}_k$. Similarly we have, $\left(\max_{1\leq j\leq M}P_{k+n}^{(j)}\right)\leq\delta^{2^n-1}\tilde{P}_k$. This proves that the sequence $\left\{P_k^{(j)},k\in\mathbb{Z}_+\right\}$ decreases doubly exponentially for each j.

C. Proof of Proposition 3

i) Define $\theta(x) = [\min(x,1)]_+$, where $[z]_+ = \max(0,z)$. Now, we consider the following modification of (6)-(9): $\mathbf{u}(0) = \mathbf{g}$, $\dot{\mathbf{u}}(t) = \tilde{\mathbf{h}}(\mathbf{u}(t))$, where for $1 \le j \le M$, $\tilde{h}_0^{(j)}(\mathbf{u}) = 0$ and

$$\tilde{h}_n^{(j)}(\mathbf{u}) = \lambda \left[\theta(u_{n-1}^{(j)}) - \theta(u_n^{(j)})\right]_+ \times \sum_{i=1}^{M} \gamma_i \left(\theta(u^{(i)}_{\{(n-1)C_i/C_j\}}) + \theta(u^{(i)}_{[(n-1)C_i/C_j]+1})\right) - \mu C_j \left[\theta(u_n^{(j)}) - \theta(u_{n+1}^{(j)})\right]_+ \quad (26)$$

for all $n \ge 1$. Note that the right hand side of (9) and (26) are equal if $\mathbf{u} \in \bar{\mathcal{U}}^M$. Therefore, the two systems have the same solution in $\bar{\mathcal{U}}^M$. Also if $\mathbf{g} \in \bar{\mathcal{U}}^M$, then any solution of the modified system remains within $\bar{\mathcal{U}}^M$. This is because of the facts that if $u_n^{(j)}(t) = u_{n+1}^{(j)}(t)$ for some j, n, t, then $\tilde{h}_n^{(j)}(\mathbf{u}(t)) \ge 0$ and $\tilde{h}_{n+1}^{(j)}(\mathbf{u}(t)) \le 0$, and if $u_n^{(j)}(t) = 0$ for some j, n, t, then $\tilde{h}_n^{(j)}(\mathbf{u}) \ge 0$. Now, using Picard's successive approximation method, it can be shown that the modified system has a unique solution in $(\mathbb{R}^{\mathbb{Z}_+})^M$. Hence, the system (6)-(9) has a unique solution in $\bar{\mathcal{U}}^M$.

ii) For ease of exposition we outline the proof for the $M = 2$ case. The proof can be extended to any $M \ge 2$.

We note that if there exists $\mathbf{P} \in \bar{\mathcal{U}}^M$ such that the sequences $\left\{P_l^{(1)}, l \in \mathbb{Z}_+\right\}$ and $\left\{P_l^{(2)}, l \in \mathbb{Z}_+\right\}$ satisfy the recursive relation (25) for all $l \in \mathbb{Z}_+$, $j = 1, 2$, then it must be an equilibrium point of the system (6)-(9). Moreover, if $P_l^{(1)}, P_l^{(2)} \downarrow 0$ as $l \to \infty$, then by Proposition 2.iii), such $\mathbf{P}$ must also lie in the space $\mathcal{U}^M$. We now construct the sequences $\left\{P_l^{(1)}(\alpha), l \in \mathbb{Z}_+\right\}$ and $\left\{P_l^{(2)}(\alpha), l \in \mathbb{Z}_+\right\}$ as functions of the real variable α as follows: $P_0^{(1)}(\alpha) = P_0^{(2)}(\alpha) = 1$, $P_1^{(1)}(\alpha) = \alpha$, $P_1^{(2)}(\alpha) = \frac{\rho_2}{\gamma_2}\left(1 - \frac{\gamma_1}{\rho_1}\alpha\right)$, and for $l \ge 2$ the terms $P_l^{(1)}(\alpha)$ and $P_l^{(2)}(\alpha)$ are defined by the recursive relation (25). It can be shown using the method of induction that, under condition (4), there exists a value of $\alpha \in (0,1)$, such that both $\left\{P_l^{(1)}(\alpha), l \in \mathbb{Z}_+\right\}$ and $\left\{P_l^{(2)}(\alpha), l \in \mathbb{Z}_+\right\}$ are non-negative, decreasing sequences converging to 0. Hence, there exists $\alpha \in (0,1)$ such that $\mathbf{P}(\alpha) = \left\{P_l^{(j)}(\alpha), l \in \mathbb{Z}_+, j = 1, 2\right\} \in \mathcal{U}^M$ is a fixed point of (6)-(9). The uniqueness follows from the uniqueness of the limit in Proposition 3. iii).

iii) The proof is similar to the proof of Theorem 1.(iii) of [12] and hence is omitted to conserve space.

D. Proof of Lemma 1

It is easy to see that the transition rate from the state $\mathbf{g}$ to the state $\mathbf{g} - \mathbf{e}(n,j)/N\gamma_j$, where $n \ge 1$, is given by $\mu C_j N \gamma_j \left[g^{(j)}(n) - g^{(j)}(n+1)\right]$. Similarly, the transition rate from state $\mathbf{g}$ to the state $\mathbf{g} + \mathbf{e}(n,j)/N\gamma_j$, where $n \ge 1$, is given by $\lambda N \left(g_{n-1}^{(j)} - g_n^{(j)}\right) \sum_{i=1}^{M} \gamma_i \gamma_j \left(g^{(i)}_{\left\{\frac{(n-1)C_i}{C_j}\right\}} + g^{(i)}_{\left[\frac{(n-1)C_i}{C_j}\right]+1}\right)$. Therefore, the result follows from the definition of the generator $\mathbf{A}_N$.

E. Proof of Proposition 4

The proof is along the lines similar to the the proof of Theorem 2 of [12]. We omit the details and mention the key point that for a function $f : \bar{\mathcal{U}}^M \to \mathbb{R}$ whose derivatives $\frac{\partial f(\mathbf{g})}{\partial g_n^{(j)}}$, $\frac{\partial^2 f(\mathbf{g})}{\partial g_n^{(j)^2}}$, and $\frac{\partial^2 f(\mathbf{g})}{\partial g_n^{(j)} \partial g_{n'}^{(j')}}$ exist for all j, j', n, n', and are uniformly bounded in modulus by some constant, we have

$$N\gamma_j \left(f(\mathbf{g} + \frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{g})\right) \to \frac{\partial f(\mathbf{g})}{\partial g_n^{(j)}} \quad (27)$$

$$N\gamma_j \left(f(\mathbf{g} - \frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{g})\right) \to -\frac{\partial f(\mathbf{g})}{\partial g_n^{(j)}} \quad (28)$$

uniformly in $\mathbf{g}$ from $\bar{\mathcal{U}}^M$ as $N \to \infty$.

REFERENCES

[1] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg, "Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services," *Performance Evaluation*, vol. 68, no. 11, pp. 1056–1071, 2011.

[2] E. Schurman and J. Brutlag, "The user and business impact on server delays, additional bytes and http chunking in web search," in *O'Reilly Velocity Web Performance and Operations Conference*, Jun. 2009.

[3] K. Salchow, "Load balancing 101: Nuts and bolts," in *White Paper, F5 Networks, Inc.*, 2007.

[4] V. Gupta, M. H. Balter, K. Sigman, and W. Whitt, "Analysis of join-the-shortest-queue routing for web server farms," *Performance Evaluation*, vol. 64, no. 9-12, pp. 1062–1081, 2007.

[5] 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.

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

[7] M. Bramson, Y. Lu, and B. Prabhakar, "Randomized load balancing with general service time distributions," in *Proceedings of ACM SIGMETRICS*, pp. 275–286, 2010.

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

[9] 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.

[10] A. Mukhopadhyay and R. R. Mazumdar, "Analysis of load balancing in large heterogeneous processor sharing systems." arXiv:1311.5806 [cs.DC].

[11] E. Altman, U. Ayesta, and B. J. Prabhu, "Load balancing in processor sharing systems," *Telecommunication Systems*, vol. 47, no. 1-2, pp. 35–48, 2008.

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

[13] C. Graham, "Chaoticity on path space for a queueing network with selection of shortest queue among several," *Journal of Applied Probability*, vol. 37, no. 1, pp. 198–211, 2000.

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

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

[16] M. Bramson, "Stability of join the shortest queue networks," *Annals of Applied Probability*, vol. 21, no. 4, pp. 1568–1625, 2011.