Analysis of Randomized Join-The-Shortest-Queue (JSQ) Schemes in Large Heterogeneous Processor Sharing Systems

Arpan Mukhopadhyay and Ravi R. Mazumdar, *Fellow, IEEE*

Abstract

In this paper we investigate the stability and performance of randomized dynamic routing schemes for jobs based on the Join-the-Shortest Queue (JSQ) criterion to a heterogeneous system of many parallel servers. In particular we consider servers that use processor sharing but with different server rates and jobs are routed to the server with the smallest occupancy among a finite number of randomly selected servers. We focus on the case of two servers that is often referred to as a Power-of-Two scheme. We first show that in the heterogeneous setting there can be a loss in the stability region over the homogeneous case and thus such randomized schemes need not outperform static randomized schemes in terms of mean delay in opposition to the homogeneous case of equal server speeds where the stability region is maximal and coincides with that of static randomized routing. We explicitly characterize the stationary distributions of the server occupancies and show that the tail distribution of the server occupancy has a super-exponential behavior as in the homogeneous case as the number of servers goes to infinity. To overcome the stability issue, we show that it is possible to combine static state-independent scheme with randomized JSQ scheme that allows us to recover the maximal stability region combined with the benefits of JSQ and such a scheme is preferable in terms of average delay. The techniques are based on a mean field analysis where we show that the stationary distributions coincide with those obtained under asymptotic independence of the servers and moreover the stationary distributions are insensitive to the job size distribution.

Index Terms

Load balancing, Processor sharing, Power-of-two, Mean Field Approach, Asymptotic independence, Insensitivity

I. Introduction

A central problem in a multi-server resource sharing system is to decide which server an incoming job will be assigned to in order to obtain optimum performance, typically the low average response time. The problem becomes

The authors are with the department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada. e-mail: arpan.mukhopadhyay@uwaterloo.ca, mazum@ece.uwaterloo.ca

more challenging when the number of servers in the system is large and the servers have different service rates. Such systems are frequently encountered in large web server farms that accommodate a large number of front end servers of various service rates to process incoming job requests [1]. The load balancing scheme used plays a key role in determining the mean sojourn time of jobs in such systems. Since web applications such as online search, social networking are extremely delay sensitive, a small increase in the average response time of jobs may cause significant loss of revenue and users [2]. Therefore, the main objective of efficient load balancing is to reduce the mean sojourn times of jobs in the system. Another desirable property of a routing scheme should be its *robustness* to heterogeneity of job sizes of which a typical statistical behavior is insensitivity to job size distributions.

The join-the-shortest-queue (JSQ) scheme, commonly used in small web server farms [1], [3], assigns 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.

However, a major drawback of the JSQ scheme, when applied to a system consisting of a large number of servers, is that it requires the state information of all the servers in the system to make job assignment decisions. The use of dynamic randomized algorithms is one way to avoid requiring information about all server occupancies. It has been shown that randomized load balancing schemes based on sampling only a few servers provides most of the reduction in mean sojourn times associated with JSQ [5]. Indeed as argued in [5]–[7], most of the gains in average sojourn time reduction are obtained when selecting 2 servers at random referred to as the Power-of-Two rule. This is also referred to as the SQ(2) scheme.

The literature has treated the SQ(2) scheme for a system of identical FCFS servers assuming exponential job length distribution. The exact analysis is still difficult because of the coupling between the servers. However, as the work in [5], [7], [8] has shown, when the number of servers goes to infinity any finite collections of servers can be viewed as independent. This is termed as *asymptotic independence* or *propagation of chaos*. With this insight, it was shown that in the large system limit the stationary tail distribution of the number of remaining jobs at each server decays doubly exponentially as compared to the exponential decay in case of the optimal state independent scheme in which job assignments are done independently of the states of the servers. Consequently, the SQ(2) scheme results in an exponential reduction in the mean sojourn time of jobs as compared to the optimal state independent scheme.

The analysis of the SQ(2) scheme was further generalized to include general job length distributions in [9], [10]. However, the analyses in [5]–[7], [9], [10] were restricted to the homogeneous case where the servers in the system are identical in terms of the server speed.

In this paper, we first analyze the performance of the classical SQ(2) scheme with uniform sampling for a large system of PS servers with heterogeneous service rates for which there are no available results. The first issue that arises is the issue of the stability region for such systems or the maximum load that the system can support and still yield finite average sojourn times. In particular, we show that the stability region is strictly smaller than the maximal stability region obtained by restricting the normalized arrival rate below the average capacity of the system. Thus,

it is possible that the average sojourn time behavior can be worse than static randomized routing schemes. We then provide a detailed analysis of the heterogeneous system and provide a complete characterization of the stationary distribution when it exists. We show that under the SQ(2) scheme the system is asymptotically insensitive to the type of job length distributions. To overcome the reduction in stability we show that a scheme that combines static randomized routing to a server class, i.e., sampling with a bias and SQ(2) with uniform sampling within servers of the same class, allows us to recover the maximal stability region. We show that this hybrid scheme retains the gains of the SQ(2) scheme. This scheme is therefore always superior in the sense of smaller average sojourn time over static randomized routing schemes.

The techniques are based on a mean field approach that extends the methodology used in [5], [11] for FCFS queues with exponential job lengths to the heterogeneous PS scenario. We show uniqueness of the solution under stability. Furthermore in the asymptotic limit the stationary distribution of the server occupancies also coincides with that obtained by assuming asymptotic independence for any finite subset of the servers [8], [9].

The organization of the paper is as follows. In Section II, we present the system model and provide a description of the load balancing schemes studied in this paper. Section III presents detailed analyses of the load balancing schemes. In Section IV, numerical results are presented to compare the different schemes and validate the theoretical analyses. The paper is finally concluded in Section V.

II. System Model

We consider a system consisting of N parallel processor sharing (PS) 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, $\sum_{j=1}^{M} \gamma_j = 1$.

Jobs arrive according to a Poisson process with rate $N\lambda$. Each job requires a random amount of work and the job sizes are independent and identically distributed, with a finite mean $\frac{1}{\mu}$. 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 according to a randomized load balancing scheme. We now discuss the load balancing schemes considered in this paper.

A. Scheme 1: Optimal state independent scheme

As a baseline, we consider a scheme that assigns an incoming job to a server with a fixed probability, independent of the current state of the servers in the system. We denote by p_j, $j \in \mathcal{J}$, the probability that an arrival is assigned to one of the servers having capacity C_j. The probabilities p_j, $j \in \mathcal{J}$, are chosen in such a way that the mean sojourn time of the jobs is minimized. Clearly, in this scheme, no communication is required between the job dispatcher and the servers as the job assignment decisions are made independently of the state of the servers.

B. Scheme 2: The SQ(2) scheme

In this scheme, a subset of two servers is selected from the set of N servers uniformly at random at each arrival instant. The arriving job is then assigned to the server having the least number of unfinished jobs among the two chosen servers. In case of a tie, the job is assigned to any one of the two servers with equal probability $\frac{1}{2}$. [1]

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

This scheme combines the state independent scheme with the 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 with 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.

III. Analysis

In this section, we present the analysis of the load balancing schemes described in the previous section. Since Scheme 1 is a special case of the more general class of load balancing schemes analyzed in [12], we only state the main analytical results for Scheme 1 in Sec. III-A without giving the proofs. These results are used later to compare the different load balancing schemes considered in this paper. The detailed analyses of the SQ(2) scheme and the hybrid SQ(2) scheme are provided in Section III-B and Section III-C, respectively.

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 [12], that there exists probabilities p_j, $j \in \mathcal{J}$, for which the system is stable under Scheme 1 if and only if the following condition holds:

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

It was also shown in [12] that, under the above stability condition, the routing probabilities p_j, $j \in \mathcal{J}$ can be chosen such that the mean sojourn time of jobs in the system is minimized. The mean sojourn time minimization problem, formulated as a convex optimization problem, was solved in Theorem 1 of [12]. It was found that the index set $\mathcal{J}_{\text{opt}} = \{1, 2, \ldots, j^*\} \subseteq \mathcal{J}$ of server capacities and the loads $\boldsymbol{\rho}^* = \{\rho_1^*, \rho_2^*, \ldots, \rho_M^*\}$ in the optimal state independent scheme are given by

[1]The analysis of the SQ(2) scheme can be readily generalized to the SQ(d) scheme where an incoming job is assigned to the least loaded server among d randomly chosen ones at the cost of more notation and complication. However, since the SQ(2) scheme provides most of the improvements, we do not pursue it in this paper.

$$j^* = \sup\left\{ j \in \mathcal{J} : \frac{1}{\sqrt{C_j}} < \frac{\sum_{i=1}^{j} \gamma_i \sqrt{C_i}}{\sum_{i=1}^{j} \gamma_i C_i - \frac{\lambda}{\mu}} \right\}. \tag{2}$$

$$\rho_i^* = \begin{cases} 1 - \sqrt{\frac{1}{C_i}} \frac{\sum_{k=1}^{j^*} \gamma_k C_k - \frac{\lambda}{\mu}}{\sum_{k=1}^{j^*} \gamma_k \sqrt{C_k}}, & \text{if } i \in \mathcal{J}_{\text{opt}} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Here, we have assumed that the the server capacities are ordered as $C_1 \geq C_2 \geq \ldots \geq C_M$. The optimal routing probabilities p_j^*, $j \in \mathcal{J}$, can be computed from (3) by using the relations $\rho_j^* = \frac{p_j^* \lambda}{\gamma_j \mu C_j}$.

B. Scheme 2: The SQ(2) scheme

In the SQ(2) scheme, the job assignments are done based on the instantaneous states of two randomly selected servers in the system. Therefore, 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 the stationary distribution very difficult for a finite value of N. However, the mean field approach outlined in [5], [11] or the propagation of chaos arguments used in [8]–[10] 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 simulation results that such asymptotic analysis accurately captures the behaviour of a large but finite system of servers.

In this paper we say that a Markov process is stable if it is positive Harris recurrent. We now characterize the stability region of the system described in Scheme 2.

Let N^* denote the smallest positive integer (> 2) such that $\gamma_j N^*$ is a positive integer for all $j \in \mathcal{J}$. Now, let Λ_k, for $k \in \mathbb{N}$, 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 Λ_k, $k \in \mathbb{N}$ defined as above, and Λ as given in (1), we have

$$\Lambda \supseteq \Lambda_1 \supseteq \Lambda_2 \supseteq \ldots \tag{4}$$

Furthermore, if $\Lambda_\infty = \cap_{k=1}^{\infty} \Lambda_k$, then Λ_∞ is given by

$$\Lambda_\infty = \left\{ 0 \leq \lambda < \mu \min_{\mathcal{I} \subseteq \mathcal{J}} \left\{ \frac{\left(\sum_{j \in \mathcal{I}} \gamma_j C_j \right)}{\left(\sum_{j \in \mathcal{I}} \gamma_j \right)^2} \right\} \right\} \tag{5}$$

Proof: The proof is given in Appendix A. ■

Remark *1:* From (4), it is clear that for any finite value of N, the stability region under Scheme 2 is a subset of that under Scheme 1. Further, the stability region under Scheme 2 decreases as N increases keeping the proportions γ_j, $j \in \mathcal{J}$, fixed. Hence Λ_∞ denotes the region where the system is stable for all N. We then show that in this region the mean field has a unique, globally asymptotically stable equilibrium point in the space of empirical tail measures that are summable.

Remark *2:* Under the notation $\nu_j = \lambda/\mu C_j$, it is easy to see that $\lambda < \mu \min_{\mathcal{I}\subseteq\mathcal{J}} \left\{ \frac{\left(\sum_{j\in\mathcal{I}} \gamma_j C_j\right)}{\left(\sum_{j\in\mathcal{I}} \gamma_j\right)^2} \right\}$ in (5) can be equivalently expressed as

$$\sum_{j\in\mathcal{I}} \frac{\gamma_j}{\nu_j} > \left(\sum_{j\in\mathcal{I}} \gamma_j\right)^2 \text{ for all } \mathcal{I} \subseteq \mathcal{J}. \tag{6}$$

1) Mean Field Analysis: Assuming exponential job length distribution (with mean $1/\mu$), we now characterize the stationary distribution of the system under the SQ(2) scheme as $N \to \infty$. To do so we extend the mean field approach of [5], [11] from the homogeneous scenario to the heterogeneous scenario.

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. From the Poisson arrival and exponential job size assumptions, 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:

$$\bar{\mathcal{U}}_N^{(j)} = \{g = (g_n, n \in \mathbb{Z}_+) : g_0 = 1, g_n \ge g_{n+1} \ge 0, N\gamma_j g_n \in \mathbb{N}\ \forall n \in \mathbb{Z}_+\}. \tag{7}$$

We generalize the space $\bar{\mathcal{U}}_N^{(j)}$ to the space $\bar{\mathcal{U}}$ by removing the last constraint in its definition (7). Hence, the space $\bar{\mathcal{U}}$ is defined as follows:

$$\bar{\mathcal{U}} = \{g = (g_n, n \in \mathbb{Z}_+) : g_0 = 1, g_n \ge g_{n+1} \ge 0\ \forall n \in \mathbb{Z}_+\}. \tag{8}$$

This space will be required to study the limiting properties of the process $\mathbf{x}_N(t)$ as $N \to \infty$.

We seek to show the weak convergence of the process $\mathbf{x}_N(t)$ as $N \to \infty$ to the deterministic process $\mathbf{u}(t) = \left\{ u_n^{(j)}(t), n \in \mathbb{Z}_+, j \in \mathcal{J} \right\}$, governed by the following system of differential equations that represents the mean field:

$$\mathbf{u}(0) = \mathbf{g}, \tag{9}$$

$$\dot{\mathbf{u}}(t) = \mathbf{h}(\mathbf{u}(t)), \tag{10}$$

where $\mathbf{g} \in \bar{\mathcal{U}}^M$, $\mathbf{h}(\mathbf{u}) = \left\{ h_n^{(j)}(\mathbf{u}), n \in \mathbb{Z}_+, j \in \mathcal{J} \right\}$, and for $j \in \mathcal{J}$

$$h_0^{(j)}(\mathbf{u}) = 0, \tag{11}$$

$$h_n^{(j)}(\mathbf{u}) = \lambda \left(u_{n-1}^{(j)} - u_n^{(j)} \right) \sum_{i=1}^{M} \gamma_i \left(u_{n-1}^{(i)} + u_n^{(i)} \right) - \mu C_j \left(u_n^{(j)} - u_{n+1}^{(j)} \right) \tag{12}$$

for all $n \ge 1$.

More specifically, we 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 the process $\mathbf{x}_N$ converges weakly to the deterministic process $\mathbf{u}$ given by the solution

of (9)-(12). We further show that under condition (6), the system (9)-(12) has a unique, globally asymptotically stable equilibrium point $\mathbf{P} = \left\{P_k^{(j)}, k \in \mathbb{Z}_+, 1 \leq j \leq M\right\}$, obtained by solving the equation $\dot{\mathbf{u}}(t) = \mathbf{h}(\mathbf{u}(t)) = 0$, in the space of empirical measures having finite mean.

In the following proposition, we summarize some important properties of the equilibrium point $\mathbf{P}$ of the system (9)-(12).

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 $j \in \mathcal{J}$,

$$P_{k+1}^{(j)} = \nu_j \left[\gamma_j \left(P_k^{(j)}\right)^2 + P_k^{(j)} \left(\sum_{\substack{i=1 \\ i \neq j}}^{M} \gamma_i P_k^{(i)}\right) + \sum_{\substack{i=1 \\ i \neq j}}^{M} \sum_{l=k}^{\infty} \gamma_i \left(P_{l+1}^{(i)} P_l^{(j)} - P_l^{(i)} P_{l+1}^{(j)}\right)\right]. \tag{13}$$

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

$$\sum_{j=1}^{M} \frac{\gamma_j}{\nu_j} P_{k+1}^{(j)} = \left(\sum_{j=1}^{M} \gamma_j P_k^{(j)}\right)^2 \tag{14}$$

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 (9)-(12), we need to 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 following space of tail distributions on $\mathbb{Z}_+$ that has finite first moment.

$$\mathcal{U} = \{g = (g_n, n \in \mathbb{Z}_+) : g_0 = 1, g_n \geq g_{n+1} \geq 0 \ \forall n \in \mathbb{Z}_+, \sum_{n=0}^{\infty} g_n < \infty\}. \tag{15}$$

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

$$\|u\| = \sup_{1 \leq j \leq M} \sup_{n \in \mathbb{Z}_+} \frac{\left|u_n^{(j)}\right|}{n+1}. \tag{16}$$

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 (9)-(12) and its equilibrium point $\mathbf{P}$. To emphasize the dependence of the solution of the system (9)-(12) 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 (9)-(12) 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 (6), there exists a unique equilibrium point or fixed point $\mathbf{P}$ of the system (9)-(12) in the space $\mathcal{U}^M$. Therefore, $\mathbf{P}$ satisfies the properties stated in Proposition 2.

iii) Under condition (6),

$$\lim_{t\to\infty} \mathbf{u}(t,\mathbf{g}) = \mathbf{P} \text{ for all } \mathbf{g} \in \mathcal{U}^M. \tag{17}$$

Proof: The proof is given in Appendix C. ■

Having established the existence and uniqueness of the equilibrium point of the system (9)-(12), 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 $N \to \infty$ [13].

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 $\mathbf{e}(n,j) = \left(e_k^{(i)}\right)_{k\in\mathbb{Z}_+, i\in\mathcal{J}}$ 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

$$\mathbf{A}_N f(\mathbf{g}) = \lambda N \sum_{n\geq 1} \sum_{j=1}^M \sum_{i=1}^M \gamma_i\gamma_j \left[g_{n-1}^{(j)} - g_n^{(j)}\right] \times \left[g_{n-1}^{(i)} + g_n^{(i)}\right] \left[f(\mathbf{g} + \frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{g})\right]$$
$$+ \mu N \sum_{n\geq 1} \sum_{j=1}^M \gamma_j C_j \left[g_n^{(j)} - g_{n+1}^{(j)}\right] \left[f(\mathbf{g} - \frac{\mathbf{e}(n,j)}{N\gamma_j}) - f(\mathbf{g})\right]. \tag{18}$$

Proof: The proof follows by noting that the transition rate from the state $\mathbf{g}$ to the state $\mathbf{g} - \mathbf{e}(n,j)/N\gamma_j$, where $n \geq 1$, 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 $n \geq 1$, is given by $\lambda N \left[g_{n-1}^{(j)} - g_n^{(j)}\right] \sum_{i=1}^M \gamma_i\gamma_j \left[g_{n-1}^{(i)} + g_n^{(i)}\right]$. ■

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 $\mathbf{T}_N(t)f = \exp\left(t\mathbf{A}_N\right) f$. 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 $t \geq 0$,

$$\lim_{N\to\infty} \sup_{\mathbf{g}\in\bar{\mathcal{U}}^M} |\mathbf{T}_N(t)f(\mathbf{g}) - f(\mathbf{u}(t,\mathbf{g}))| = 0 \tag{19}$$

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

Proof: The proof follows from the smoothness assumptions on $f : \bar{\mathcal{U}}^M \to \mathbb{R}$. We omit the technical details. ■

From Theorem 2.11 of Chapter 4 of [13], the above proposition implies that $\mathbf{x}_N \Rightarrow \mathbf{u}$ as $N \to \infty$, where $\Rightarrow$ denotes weak convergence. This implies the weaker result that $\mathbf{x}_N(t) \Rightarrow \mathbf{u}(t)$ for each $t \geq 0$. It also implies that any limit point of the sequence of invariant measures $\{\pi_N\}$ of the processes $\{\mathbf{x}_N\}$ is an invariant measure of the map $\mathbf{g} \mapsto \mathbf{u}(t,\mathbf{g})$. We now show that, under condition (6), there is at most one such limit point which is given by the Dirac measure concentrated at the equilibrium point $\mathbf{P} \in \mathcal{U}^M$ of the system (9)-(12).

Proposition 5: Under the condition (6), 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, $\pi_N \to \delta_P$ weakly as $N \to \infty$, where $\delta_{\mathbf{P}}$ is as defined in Proposition 3, i.e.,

$$\lim_{N\to\infty} \mathbb{E}_{\pi_N} f(\mathbf{g}) = f(\mathbf{P}) \tag{20}$$

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 the arguments in Theorem 4.(ii) of [11] *mutatis mutandis.* ■

Remark *4:* The above results establish that the following interchange holds:

$$\lim_{t\to\infty}\lim_{N\to\infty} \mathbf{x}_N(t) = \lim_{N\to\infty}\lim_{t\to\infty} \mathbf{x}_N(t) = \mathbf{P}, \tag{21}$$

where the limits are in the sense of weak convergence.

Due to exchangeability of states among servers having the same capacity, the above interchange of limits also implies that in the limit as $N \to \infty$ the servers in the system evolve independently of each other [8]. More specifically, the tail distribution of number of pending jobs at time time $t \geq 0$ at a server of capacity C_j in the limiting system is given by $\left\{u_n^{(j)}(t, \mathbf{g}), n \geq 0\right\}$, independent of any other server in the system, where $\left\{g_k^{(i)}, k \geq 0\right\}$, for $i \in \mathcal{J}$, is the initial tail distribution of server occupancies at any type i server. Further, the stationary tail distribution of server occupancies for a type j server in the limiting system is also independent of all other servers and is given by $\left\{P_n^{(j)}, n \geq 0\right\}$. This property is formally known as *propagation of chaos*.

2) Insensitivity: So far, we have assumed that the job lengths are i.i.d exponential random variables. We now show that the stationary distribution of the mean field coincides with the stationary distribution obtained when the queues are independent at equilibrium. This will imply that stationary distribution of server occupancies in the limiting system is insensitive to the job length distribution and only depends on their means.

Proposition 6: Assume that condition (6) and asymptotic independence of queues in equilibrium in the mean field limit, i.e., for any finite set B of servers,

$$\Pi^{(B)} = \bigotimes_{n\in B} \pi^{(n)} \tag{22}$$

where $\pi^{(n)}$ and $\Pi^{(B)}$ denote the marginals of Π for the n^{th} server and for the servers in set B, respectively.

Then, in equilibrium, the arrival process of jobs at any given server in the limiting system becomes a state dependent Poisson process whose intensity is given by:

$$\lambda_k = \lambda \sum_{i=1}^{M} \gamma_i \left(P_k^{(i)} + P_{k+1}^{(i)}\right), \tag{23}$$

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. Moreover, 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 (13).

Furthermore, the stationary distribution of a server depends on the job size only through its mean, or the queues are *insensitive*.

Proof: Consider any particular server (say server 1) in the system. Consider the arrivals that have server 1 as one of its two possible destinations. These arrivals constitute the *potential arrival process* at the server. The probability that the server is selected as a potential destination server for a new arrival is $\left(1-\frac{\binom{N-1}{2}}{\binom{N}{2}}\right)=\frac{2}{N}$. Thus, from Poisson thinning, the potential arrival process to a server is a Poisson process with rate $\frac{2}{N}\times N\lambda=2\lambda$.

Next, we consider the arrivals that actually join server 1. These arrivals constitute the actual arrival process at the server. For finite N, this process is not Poisson since a potential arrival to server 1 actually joins server 1 depending on the number of users present at the other possible destination server. However, as $N\to\infty$, due to the asymptotic independence property stated in (22), the numbers of jobs present at these two servers become independent of each other. As a result, in equilibrium the actual arrival process converges to a state dependent Poisson process as $N\to\infty$.

Consider the potential arrivals at a server when the number of users present at the server is k. This arrival actually joins the server either with probability $\frac{1}{2}$ or with probability 1 depending on whether the number unfinished jobs at the other possible destination server is exactly k or greater than k, respectively. Since a server having capacity C_j is chosen with probability γ_j, the total probability that the potential arrival joins the server at state k is $\sum_{j=1}^{M}\gamma_j\left(0.5\left(P_k^{(j)}-P_{k+1}^{(j)}\right)+P_{k+1}^{(j)}\right)=0.5\sum_{j=1}^{M}\gamma_j\left(P_k^{(j)}+P_{k+1}^{(j)}\right)$. Therefore, the rate at which arrivals occur at stake k is given by $2\lambda\times 0.5\sum_{j=1}^{M}\gamma_j\left(P_k^{(j)}+P_{k+1}^{(j)}\right)$. This simplifies to (23).

Since processor sharing is a symmetric service discipline, it follows from Theorems 3.10 and 3.14 of [14] and also from Theorem 4.2 of [4] that the detailed balance equations hold for state dependent Poisson arrivals. Therefore, we have for $k\in\mathbb{Z}_+$ and $j\in\mathcal{J}$ that

$$P_{k+1}^{(j)}-P_{k+2}^{(j)}=\frac{\lambda_k}{\mu C_j}(P_k^{(j)}-P_{k+1}^{(j)}). \tag{24}$$

Substituting the value of λ_k from (23) into (24) and upon further simplification we get (13). ■

Remark *5:* Thus, we have shown that under the assumption of asymptotic independence of the servers in equilibrium, the stationary distribution of server occupancies coincides with that of the mean field. From the uniqueness of the solution we can conclude that asymptotic independence should hold also for heterogeneous systems. A direct proof of the asymptotic independence is, however, extremely difficult. The proof remains an open problem even for homogeneous systems and any local service discipline [9]. In recent work [15] propagation of chaos has been established for FCFS systems under general service time distributions.

Remark *6:* The long run probability that a user joins a server with capacity C_j is given by $\frac{N\gamma_j\bar{\lambda}^{(j)}}{N\lambda}$, where, $\bar{\lambda}^{(j)}=\sum_{k=0}^{\infty}\lambda_k\left(P_k^{(j)}-P_{k+1}^{(j)}\right)$ denotes the average arrival rate to a server having capacity C_j. From (23) and (13), we obtain that $\frac{\gamma_j\bar{\lambda}^{(j)}}{\lambda}=\frac{\gamma_jP_1^{(j)}}{\nu_j}$ for each $j\in\mathcal{J}$. Thus, the long run probability that a user joins a server with capacity C_j is $\frac{\gamma_jP_1^{(j)}}{\nu_j}$.

Proposition 7: The mean sojourn time, $\bar{T}$, of a job in the heterogeneous system under Scheme 2 is given by

$$\bar{T} = \frac{1}{\lambda} \sum_{j=1}^{M} \sum_{k=1}^{\infty} \gamma_j P_k^{(j)}, \tag{25}$$

where $P_k^{(j)}$, $k \in \mathbb{Z}_+$ 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 $\sum_{k=1}^{\infty} P_k^{(j)}$. Let the average arrival rate at the server be denoted by $\bar{\lambda}^{(j)}$. Thus, applying Little's formula we have $\bar{T}_j = \frac{\sum_{k=1}^{\infty} P_k^{(j)}}{\bar{\lambda}^{(j)}}$

As discussed in Remark 6, 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 $\bar{T} = \sum_{j=1}^{M} \frac{\gamma_j \bar{\lambda}^{(j)}}{\lambda} \bar{T}_j = \frac{1}{\lambda} \sum_{j=1}^{M} \sum_{k=1}^{\infty} \gamma_j P_k^{(j)}$. ■

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

We saw that the classical SQ(2) scheme can have smaller stability region than Scheme 1. We now show that it is possible to recover the stability region of Scheme 1 by using the hybrid SQ(2) scheme.

In the hybrid SQ(2) scheme, for each $j \in \mathcal{J}$, a service rate $C_j \in \mathcal{C}$ is selected for a new arrival with a probability p_j. Hence, the aggregate Poisson arrival rate to the set of $N\gamma_j$ servers, each having capacity C_j, is $p_j N\lambda$. The system may, therefore, be viewed as being composed of M parallel homogeneous subsystems each working under the classical SQ(2) scheme. The 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 [5], [6], [10], [16], we know that 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, $j \in \mathcal{J}$ for which 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 the hybrid SQ(2) scheme if and only if $\lambda \in \Lambda$.

Proof: Let us assume that (1) 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 (1) 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 (1) is necessary. ■

Remark *7:* We have seen that with the hybrid SQ(2) scheme it is possible to recover the stability region as defined in (1). The intuition behind the loss of stability region under the SQ(2) scheme is related to the fact that under uniform sampling, depending on the proportions of fast and slow servers, one could frequently choose slower servers even when they are heavily loaded and there are faster servers available with less congestion. Clearly, a biased sampling of the servers is one way to avoid this. The hybrid SQ(2) scheme provides the optimal way of choosing the bias.

Henceforth we will assume that (1) holds. We proceed to find the vector $\mathbf{p}^* = \left\{p_j^*, j \in \mathcal{J}\right\}$ or equivalently the vector $\boldsymbol{\rho}^* = \left\{\rho_j^*, j \in \mathcal{J}\right\}$ that minimizes the mean sojourn time of jobs in the limiting system under the hybrid SQ(2) scheme. Similar to the SQ(2) scheme, it can be shown that the mean sojourn time of jobs in the limiting system under the hybrid SQ(2) scheme is given by $\bar{T} = \frac{1}{\lambda} \sum_{j=1}^{M} \sum_{k=1}^{\infty} \gamma_j P_k^{(j)}$, where $P_k^{(j)}$, $j \in \mathcal{J}$ and $k \in \mathbb{Z}_+$,

denotes the stationary probability that a server with capacity C_j in the limiting system has atleast k unfinished jobs under the hybrid SQ(2) scheme. From the results of [5], [6], [10] it is known that for each $j \in \mathcal{J}$ and $k \in \mathbb{Z}_+$ we have $P_k^{(j)} = \rho_j^{2^k-1}$.

Therefore, the overall mean sojourn time of jobs is given by $\bar{T}(\boldsymbol{\rho}) = \frac{1}{\lambda}\sum_{j=1}^{M}\gamma_j \sum_{k=1}^{\infty}\rho_j^{2^k-1}$. We now formulate the mean sojourn time minimization problem in terms of the loads ρ_j, $j \in \mathcal{J}$, as follows:

$$\begin{aligned} \underset{\boldsymbol{\rho}}{\text{Minimize}} \quad & \frac{1}{\lambda}\sum_{j\in\mathcal{J}}\gamma_j\sum_{k=1}^{\infty}\rho_j^{2^k-1} \\ \text{subject to} \quad & 0 \le \rho_j < 1, \text{for all } j \in \mathcal{J} \\ & \sum_{j\in\mathcal{J}}\gamma_j C_j \rho_j = \frac{\lambda}{\mu}. \end{aligned} \tag{26}$$

To characterize the solution of the convex problem defined in (26), we assume without loss of generality that the server capacities are ordered as follows:

$$C_1 \ge C_2 \ge \ldots \ge C_M \tag{27}$$

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 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 hybrid SQ(2) scheme is then given by $\mathcal{J}_{\text{opt}} = \{1, 2, \ldots, j^*\}$, where j^* is given by

$$j^* = \sup\left\{ j \in \mathcal{J} : \frac{1}{C_j} < \Psi_j(\lambda) \right\}. \tag{28}$$

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

$$\rho_i^* = \begin{cases} \Phi(\Psi_{j*}(\lambda)C_i), & \text{if } i \in \mathcal{J}_{\text{opt}} \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

Proof: The Lagrangian associated with problem (26) is given by

$$L(\boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\zeta}, \theta) = \sum_{j=1}^{M}\gamma_j\sum_{k=1}^{\infty}\rho_i^{2^k-1} + \sum_{j=1}^{M}\nu_j(0-\rho_j) + \sum_{j=1}^{M}\zeta_j\left(\rho_j - 1\right) + \theta\left(\sum_{j=1}^{M}\gamma_j C_j\rho_j - \frac{\lambda}{\mu}\right), \tag{30}$$

where $\boldsymbol{\nu} \ge \mathbf{0}$, $\boldsymbol{\zeta} \ge \mathbf{0}$, and $\theta \in \mathbb{R}$. Since problem (26) is strictly convex and a feasible solution exits (due to condition (1)), by Slater's condition, strong duality is satisfied. Hence, the primal optimal solution $\boldsymbol{\rho}^*$ and the dual optimal solution $(\boldsymbol{\nu}^*, \boldsymbol{\zeta}^*, \theta^*)$ have zero duality gap if they satisfy the KKT conditions given as follows:

$$\mathbf{0} \leq \boldsymbol{\rho}^* < \mathbf{1}$$

$$\sum_{j=1}^{M} \gamma_j C_j \rho_j^* = \frac{\lambda}{\mu}$$

$$\theta^* \in \mathbb{R}, \boldsymbol{\nu}^* \geq \mathbf{0}, \boldsymbol{\zeta}^* \geq \mathbf{0}$$

$$\nu_j^* \rho_j^* = 0, \zeta_j^* (\rho_j^* - 1) = 0 \ \forall j \in \mathcal{J} \tag{31}$$

$$\gamma_j \sum_{k=1}^{\infty} \left(2^k - 1\right) \left(\rho_j^*\right)^{2^k - 2} - \theta^* \gamma_j C_j - \nu_j^* + \zeta_j^* = 0 \ \forall j \in \mathcal{J} \tag{32}$$

Since the objective function tends to infinity as $\rho_j^* \to 1$, for each $j \in \mathcal{J}$, it follows that necessarily $\boldsymbol{\rho}^* < \mathbf{1}$. Hence, from (31), $\boldsymbol{\zeta}^* = \mathbf{0}$. Since $\boldsymbol{\nu}^* \geq \mathbf{0}$,

$$\theta^* \leq \frac{1}{C_j} \sum_{k=1}^{\infty} \left(2^k - 1\right) \left(\rho_j^*\right)^{2^k - 2} \ \forall j \in \mathcal{J} \tag{33}$$

Further, by eliminating ν_j^* from (32) we obtain

$$\left(\sum_{k=1}^{\infty} \left(2^k - 1\right) \left(\rho_j^*\right)^{2^k - 2} - \theta^* C_j \right) \rho_j^* = 0 \tag{34}$$

Thus, if, for some $j \in \mathcal{J}$, $\theta^* > \frac{1}{C_j}$, then $\rho_j^* > 0$. Therefore, from (34) and from the definition of the map Φ we have $\rho_j^* = \Phi(\theta^* C_j)$. If $\theta^* \leq \frac{1}{C_j}$ for some $j \in \mathcal{J}$, then $\rho_j^* = 0$. Hence, we have

$$\rho_j^* = \begin{cases} \Phi(\theta^* C_j), & \text{if } \frac{1}{C_j} < \theta^* \\ 0, & \text{otherwise.} \end{cases} \tag{35}$$

To find θ^*, we use the equality constraint in (26). If the first j^* server capacities are used in the optimal SQ(2) scheme then

$$\sum_{j=1}^{j^*} \gamma_j C_j \Phi(\theta^* C_j) = \frac{\lambda}{\mu} \tag{36}$$

Hence by definition of the map Ψ_j,

$$\theta^* = \Psi_{j*}(\lambda), \tag{37}$$

where j^* is defined as in (28). ■

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.

Remark *8:* One drawback of the hybrid SQ(2) scheme is that the arrival rates need to be estimated to obtain the optimal sampling biases that would minimize the average delay. However, if one is only interested in maximize the stability region, then a much simpler biasing scheme exists in which the knowledge of the server speeds and

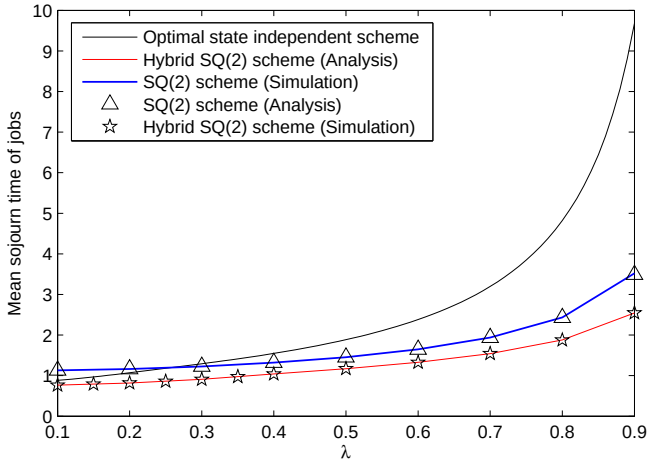


Fig. 1. Mean sojourn time jobs as a function of λ for $C_1 = 2/3$, $C_2 = 4/3$, $N = 200$ and $\gamma_1 = \gamma_2 = 1/2$

their proportions is sufficient. Indeed, it is easy to see that choosing the sampling probabilities as: $p_i = \frac{\gamma_i C_i}{\sum_{j=1}^{M} \gamma_j C_j}$ for each $i \in \mathcal{J}$ gives Λ as the stability region.

Such a sampling bias will not necessarily minimize the average delay.

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 SQ(2) and the hybrid SQ(2) schemes in predicting their performance in a finite system of servers. We set $\mu = 1$ in all our simulations. We also plot the simulation results for the SQ(5) scheme whose analysis and characterization is extremely complicated in the heterogeneous case but can be shown to be superior to the SQ(2) case by coupling arguments. But as argued by [5], [7] most of the gains are achieved (super-exponential decay of the tail distributions) by considering the SQ(2) scheme - the focus of this paper.

We first set $C_1 = 4/3$, $C_2 = 2/3$, $N = 200$ and $\gamma_1 = \gamma_2 = \frac{1}{2}$. Using conditions (1), (6) it is found that $\Lambda = \Lambda_\infty = \{0 \leq \lambda < 1\}$. 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. It is observed from the plot that the SQ(2) scheme performs better than Scheme 1 for higher values of λ and the hybrid SQ(2) scheme results in the least mean sojourn time of jobs among all the three schemes.

The performance of the SQ(2) scheme may 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, we have $\Lambda = \{0 \leq \lambda < 1\}$ and $\Lambda_\infty = \{0 \leq \lambda < 2/3\}$. Therefore, in this setting, the asymptotic stability region under the SQ(2) scheme is a strict subset of the stability region under Scheme 1 and the hybrid SQ(2) scheme. In Figure 2, we plot the average response time of jobs as a function of λ for the three

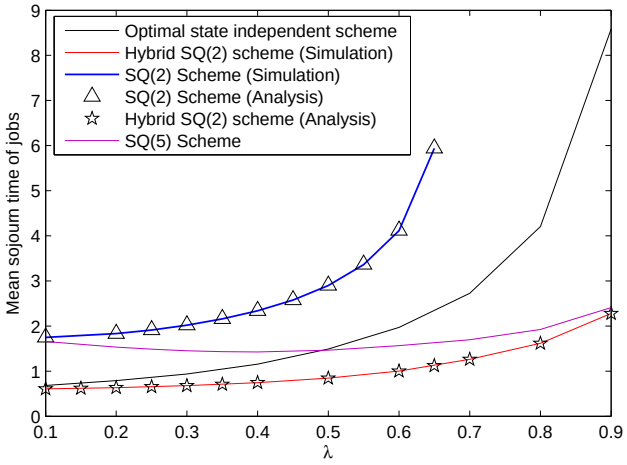


Fig. 2. Mean sojourn time jobs as a function of λ for $C_1 = 1/3$, $C_2 = 5/3$, $N = 200$, and $\gamma_1 = \gamma_2 = 1/2$

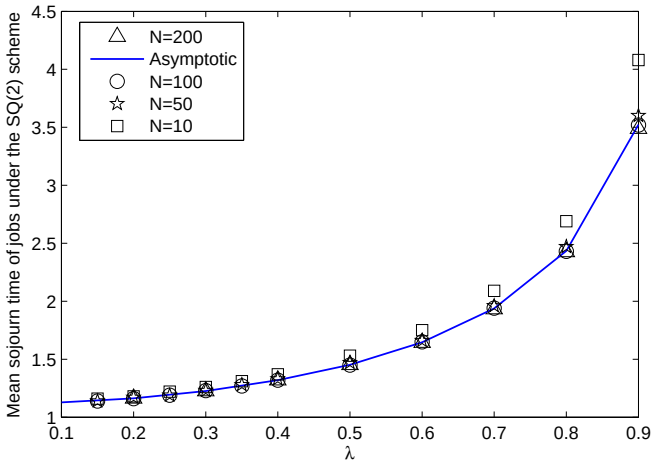


Fig. 3. Mean sojourn time jobs under the SQ(2) scheme as a function of λ for different values of N

schemes and the SQ(5) scheme. We see that the mean response time of jobs is lower in Scheme 1 than that in the SQ(2) scheme. As in the previous setting, the hybrid SQ(2) scheme outperforms Scheme 1 and the SQ(2) scheme. Furthermore, the SQ(5) scheme outperforms the SQ(2) scheme.

In Figure 3, we plot the mean sojourn time of jobs as a function of λ for different values of the system size N. The plots are obtained for the first parameter setting where $\Lambda = \Lambda_\infty$. We observe a good match between the asymptotic analysis and the simulation results for $N = 50, 100, 200$. The simulation results deviate from the analysis for $N = 10$ where the percentage of deviation is between 5-15%. This leads us to believe that the mean-field results derived in this paper can be used to accurately predict the behavior of the schemes even for moderate number of servers.

The asymptotic insensitivity of the SQ(2) scheme. is numerically validated in Table I, where the the mean sojourn time of jobs were obtained for the parameter setting $C_1 = 4/3$, $C_2 = 2/3$, $N = 200$ and $\gamma_1 = \gamma_2 = \frac{1}{2}$. We chose the following two 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 insignificant 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 SQ(2) SCHEME

λ	Mean sojourn time $\bar{T}$	Constant	Power Law
	Theoretical	Simulation	Simulation
0.2	1.1614	1.1623	1.1620
0.3	1.2257	1.2257	1.2261
0.5	1.4547	1.4533	1.4550
0.7	1.9375	1.9377	1.9380
0.8	2.4265	2.4335	2.4330
0.9	3.5300	3.5204	3.5210

V. CONCLUSION

In this paper, we considered randomized load balancing schemes for large heterogeneous processor sharing systems. It was shown that, as in the homogeneous case, the asymptotic stationary tail distribution of loads at each server decreases doubly exponentially and is insensitive to the type of job length distribution under the SQ(2) scheme. However, unlike the homogeneous case, in the heterogeneous case, the SQ(2) scheme has a smaller stability region than the average capacity of the system. We have shown that the maximal stability region can be fully recovered by using a scheme that combines the SQ(2) scheme with the state independent scheme and provides the best mean sojourn time behaviour among all the schemes considered in the paper.

APPENDIX

A. Proof of Proposition 1

From condition (1.2) of [16] for any finite value of N, the system is stable under the SQ(2) scheme if the following condition is satisfied:

$$\max_{\mathcal{B}\subseteq\mathcal{S}} \left\{ \left(\sum_{i\in\mathcal{B}} C_{(i)} \right)^{-1} \frac{N\lambda}{\mu} \frac{\binom{|\mathcal{B}|}{2}}{\binom{N}{2}} \right\} < 1, \tag{38}$$

where $\mathcal{S} = \{1, 2, \ldots, N\}$ denotes the index set of servers, $\mathcal{B} \subseteq \mathcal{S}$ is a subset of servers of size at least 2, 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

$$\Lambda_k = \left\{ \lambda > 0 : \left(\sum_{i \in \mathcal{B}} C_{(i)} \right)^{-1} \frac{N\lambda}{\mu} \frac{\binom{|\mathcal{B}|}{2}}{\binom{N}{2}} < 1 \ \forall \ \mathcal{B} \subseteq \mathcal{S}_k \right\} \tag{39}$$

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 (39) 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 (38) then we get (1). Hence, for all $k \in \mathbb{N}$, $\Lambda \supseteq \Lambda_k$. This proves (4).

To prove (5), 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 $\frac{\left(\sum_{i\in\mathcal{I}} a_i\right)\left(\sum_{i\in\mathcal{I}} a_i - 1\right)}{\sum_{i\in\mathcal{I}} a_i C_i}$ is an increasing function in each of the variables a_i. Therefore, we have

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

The first equality follows from simplifying the expression on the L.H.S. The second inequality follows from the first since we have $\frac{N\alpha - 1}{N-1} \leq \frac{N\alpha}{N} = \alpha$ for $\alpha \leq 1$. 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 $\frac{\lambda}{\mu} \frac{\left(\sum_{i\in\mathcal{I}} \gamma_i\right)^2}{\left(\sum_{i\in\mathcal{I}} \gamma_i C_i\right)} < 1$. 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 (12), we have that, for each $l \in \mathbb{Z}_+$ and $j \in \mathcal{J}$,

$$P_{l+1}^{(j)} - P_{l+2}^{(j)} = \nu_j \left(P_l^{(j)} - P_{l+1}^{(j)} \right) \sum_{i=1}^{M} \gamma_i \left(P_l^{(i)} + P_{l+1}^{(i)} \right). \tag{40}$$

Since by hypothesis $P_l^{(j)} \to 0$ as $l \to \infty$, adding the above equations for $l \geq k$ yields (13) upon simplification.

ii) Equation (14) is a direct consequence of (13).

iii) From (14) we obtain $\frac{\gamma_j P_{k+1}^{(j)}}{\nu_j} \leq \left(\sum_{j=1}^{M} \gamma_j P_k^{(j)} \right)^2 \leq \left(\tilde{P}_k \right)^2$, 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 $\delta = \tilde{P}_k \max_{1\leq j\leq M} (\nu_j/\gamma_j)$. Since by hypothesis, for each j, $P_k^{(j)} \to 0$ as $k \to \infty$, one

can choose k sufficiently large such that $\delta < 1$. Hence, we have $\left(\max_{1\le j\le M} P_{k+1}^{(j)}\right) \le \delta \tilde{P}_k$. Similarly we have, $\left(\max_{1\le j\le M} P_{k+n}^{(j)}\right) \le \delta^{2^n-1}\tilde{P}_k$. 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 (9)-(12).

$$\mathbf{u}(0) = \mathbf{g}, \tag{41}$$

$$\dot{\mathbf{u}}(t) = \tilde{\mathbf{h}}(\mathbf{u}(t)), \tag{42}$$

where for $1 \le j \le M$,

$$\tilde{h}_0^{(j)}(\mathbf{u}) = 0, \tag{43}$$

$$\tilde{h}_n^{(j)}(\mathbf{u}) = \lambda\left[\theta(u_{n-1}^{(j)}) - \theta(u_n^{(j)})\right]_+ \sum_{i=1}^{M} \gamma_i \left[\theta(u_{n-1}^{(i)}) + \theta(u_n^{(i)})\right] - \mu C_j \left[\theta(u_n^{(j)}) - \theta(u_{n+1}^{(j)})\right]_+ \tag{44}$$

for all $n \ge 1$. Note that the right hand side of (12) and (44) 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 $h_n^{(j)}(\mathbf{u}(t)) \ge 0$ and $h_{n+1}^{(j)}(\mathbf{u}(t)) \le 0$, and if $u_n^{(j)}(t) = 0$ for some j, n, t, then $h_n^{(j)}(\mathbf{u}) \ge 0$. Hence, to prove the uniqueness of solution of (9)-(12), we need to show that the modified system (41)-(44) has a unique solution in $(\mathbb{R}^{\mathbb{Z}_+})^M$.

Using the norm defined in (16) and the facts that $|x_+ - y_+| \le |x - y|$ for any $x, y \in \mathbb{R}$, $|a_1 b_1 - a_2 b_2| \le |a_1 - a_2| + |b_1 - b_2|$ for any $a_1, a_2, b_1, b_2 \in [0, 1]$, and $|\theta(x) - \theta(y)| \le |x - y|$ for any $x, y \in \mathbb{R}$ we obtain

$$\|\tilde{\mathbf{h}}(\mathbf{u})\| \le K_1 \tag{45}$$

$$\|\tilde{\mathbf{h}}(\mathbf{u}_1) - \tilde{\mathbf{h}}(\mathbf{u}_2)\| \le K_2 \|\mathbf{u_1} - \mathbf{u_2}\|, \tag{46}$$

where K_1 and K_2 are constants defined as $K_1 = 2\lambda + \mu(\max_{1\le j\le M} C_j)$ and $K_2 = 8\lambda + 2\mu(\max_{1\le j\le M} C_j)$. The uniqueness follows from inequalities (45) and (46) by using Picard's successive approximation technique since $\bar{\mathcal{U}}^M$ is complete under the norm defined in (16).

ii) For ease of exposition we provide a 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 (40) for all $l \in \mathbb{Z}_+$, $j = 1, 2$, then it must be an equilibrium point of the system (9)-(12). 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 proceed to prove that such $\mathbf{P}$ exists.

We 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: $P_0^{(1)}(\alpha) = P_0^{(2)}(\alpha) = 1$, $P_1^{(1)}(\alpha) = \alpha$, $P_1^{(2)}(\alpha) = \frac{\nu_2}{\gamma_2}\left(1 - \frac{\gamma_1}{\nu_1}\alpha\right)$, and for $l \geq 0$ and $j = 1, 2$ the following recursive relationship holds

$$P_{l+2}^{(j)}(\alpha) = P_{l+1}^{(j)}(\alpha) - \nu_j \left(P_l^{(j)}(\alpha) - P_{l+1}^{(j)}(\alpha)\right)\left(\sum_{i=1}^{2} \gamma_i \left(P_l^{(i)}(\alpha) + P_{l+1}^{(i)}(\alpha)\right)\right). \tag{47}$$

Note that the above relation is same as (40). We show that there exists some value of α, 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 (monotonicity will follow from non-negativity by virtue of (47)). It can be shown from the relations $\frac{\gamma_1}{\nu_1}P_1^{(1)}(\alpha) + \frac{\gamma_2}{\nu_2}P_1^{(2)}(\alpha) = 1$ and (47) that

$$\sum_{j=1}^{2} \frac{\gamma_j}{\nu_j} P_{l+1}^{(j)}(\alpha) = \left(\sum_{j=1}^{2} \gamma_j P_l^{(j)}(\alpha)\right)^2 \quad \text{for } l \geq 0 \tag{48}$$

From condition (6) and the construction above we see that for $\alpha \in \left(\max\left(0, \frac{\nu_1}{\gamma_1}\left(1 - \frac{\gamma_2}{\nu_2}\right)\right), \min\left(1, \frac{\nu_1}{\gamma_1}\right)\right)$ we have $1 = P_0^{(j)}(\alpha) > P_1^{(j)}(\alpha) > 0$ for $j = 1, 2$. By using (47) for $l = 2$ and $j = 1$, we have that $P_2^{(1)}(\alpha) < 0$ for $\alpha = 0$ and $P_2^{(1)}(\alpha) > 0$ for $\alpha = 1, \frac{\nu_1}{\gamma_1}$. Hence there must exist at least one root of $P_2^{(1)}(\alpha)$ in $\left(0, \min\left(1, \frac{\nu_1}{\gamma_1}\right)\right)$. Let the maximum of these roots be r_1. Therefore, if $\alpha \in \left(\max\left(r_1, \frac{\nu_1}{\gamma_1}\left(1 - \frac{\gamma_2}{\nu_2}\right)\right), \min\left(1, \frac{\nu_1}{\gamma_1}\right)\right)$ then $1 = P_0^{(1)}(\alpha) > P_1^{(1)}(\alpha) > P_2^{(1)}(\alpha) > 0$. Similarly, for $\alpha = r_1, \frac{\nu_1}{\gamma_1}\left(1 - \frac{\gamma_2}{\nu_2}\right)$, we have $P_2^{(2)}(\alpha) > 0$ and for $\alpha = \frac{\nu_1}{\gamma_1}$ we have $P_2^{(2)}(\alpha) < 0$. Therefore, there must exist a root of $P_2^{(2)}(\alpha)$ in $\alpha \in \left(\max\left(r_1, \frac{\nu_1}{\gamma_1}\left(1 - \frac{\gamma_2}{\nu_2}\right)\right), \frac{\nu_1}{\gamma_1}\right)$. If we denote the minimum of these roots by r_2, then for $\alpha \in \left(\max\left(r_1, \frac{\nu_1}{\gamma_1}\left(1 - \frac{\gamma_2}{\nu_2}\right)\right), \min\left(r_2, 1\right)\right)$ we get $1 = P_0^{(j)}(\alpha) > P_1^{(j)}(\alpha) > P_2^{(j)}(\alpha) > 0$ for $j = 1, 2$. Continuing in this way we can always get a range of $\alpha \in \left(\max\left(r_{2k+1}, \frac{\nu_1}{\gamma_1}\left(1 - \frac{\gamma_2}{\nu_2}\right)\right), \min\left(r_{2k+2}, 1\right)\right)$ such that $1 = P_0^{(j)}(\alpha) > P_1^{(j)}(\alpha) > P_2^{(j)}(\alpha) > \ldots > P_{k+2}^{(j)}(\alpha) > 0$ for $j = 1, 2$. Hence, there exists a value of α for which the sequences $\left\{P_l^{(1)}(\alpha), l \in \mathbb{Z}_+\right\}$ and $\left\{P_l^{(2)}(\alpha), l \in \mathbb{Z}_+\right\}$ are non-negative, monotonically decreasing sequences in $[0, 1]$ starting at 1 satisfying (40). In other words there exists α for which $\mathbf{P}(\alpha) = \left\{P_l^{(j)}(\alpha), l \in \mathbb{Z}_+, j = 1, 2\right\}$ is in $\bar{\mathcal{U}}^M$ and is an equilibrium point of the system (9)-(12). We now prove that for such $\mathbf{P}(\alpha)$, $P_l^{(j)}(\alpha) \to 0$ as $l \to \infty$ for $j = 1, 2$.

We have seen that there exists a value of α such that the sequences $\left\{P_l^{(1)}(\alpha), l \in \mathbb{Z}_+\right\}$ and $\left\{P_l^{(2)}(\alpha), l \in \mathbb{Z}_+\right\}$ are non-negative and monotonically decreasing sequences in $[0, 1]$ starting at 1. Hence, by monotone convergence theorem, both these sequences must converge in $[0, 1]$. Let $P_l^{(1)}(\alpha) \to \zeta_1 \in [0, 1]$ and $P_l^{(2)}(\alpha) \to \zeta_2 \in [0, 1]$ as $l \to \infty$. Hence, by taking limit as $l \to \infty$ on both sides of relation (48), we obtain

$$\sum_{j=1}^{2} \frac{\gamma_j}{\nu_j} \zeta_j = \left(\sum_{j=1}^{2} \gamma_j \zeta_j\right)^2 \tag{49}$$

Expressing the above equation as a quadratic equation $q(\zeta_1)$ in ζ_1 we see that that $q(0) = \gamma_1\zeta_2\left(\gamma_2\zeta_2 - \frac{1}{\nu_2}\right) < 0$ for $0 < \zeta_2 \leq 1$ since by (6) $\gamma_2\nu_2 < 1$. Further, $q(1) = \gamma_2^2\zeta_2^2 + \left(2\gamma_1\gamma_2 - \frac{\gamma_2}{\nu_2}\right)\zeta_2 + \left(\gamma_1^2 - \frac{\gamma_1}{\nu_1}\right)$. By using the stability condition (6) it can be easily shown that $q(1) < 0$ if $0 < \zeta_2 \leq 1$. Hence, either both roots or no roots of $q(\zeta_1) = 0$ must lie in $[0, 1]$. Now, since the product of the roots of $q(\zeta_1) = 0$ is $q(0)/\gamma_1^2 < 0$ for $0 < \zeta_2 \leq 1$ we

conclude that there is no root of $q(\zeta_1) = 0$ in $[0, 1]$ if $0 < \zeta_2 \leq 1$. Hence, $\zeta_2 = 0$. For $\zeta_2 = 0$, the only solution of $q(\zeta_1) = 0$ in $[0, 1]$ is $\zeta_1 = 0$. Therefore, we conclude $\zeta_1 = \zeta_2 = 0$. Therefore, there exists a value of α such that $P_l^{(1)}(\alpha), P_l^{(2)}(\alpha) \downarrow 0$ as $l \to \infty$. Thus, there exists α such that $\mathbf{P}(\alpha) \in \mathcal{U}^M$ and is an equilibrium point of (9)-(12). The uniqueness will follow from part (iii) of the proposition due to uniqueness of the limit.

iii) The proof is similar to the proof of Theorem 1.(iii) of [11] and hence is omitted to conserve space.

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] 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.

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

[7] 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.

[8] 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.

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

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

[12] 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.

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

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

[15] M. Aghajani and K. Ramanan, "Hydrodynamic limits of randomized load balancing networks." preprint, Dept. of Applied Mathematics, Brown University, 2014.

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

Arpan Mukhopadhyay received his Bachelors of Engineering (B.E) degree in Electronics and Telecommunication Engineering from Jadavpur University, Culcutta, India in 2009, and his Master of Engineering (M.E) degree in Telecommunication from Indian Institute of Science, Bangalore, India in 2011.

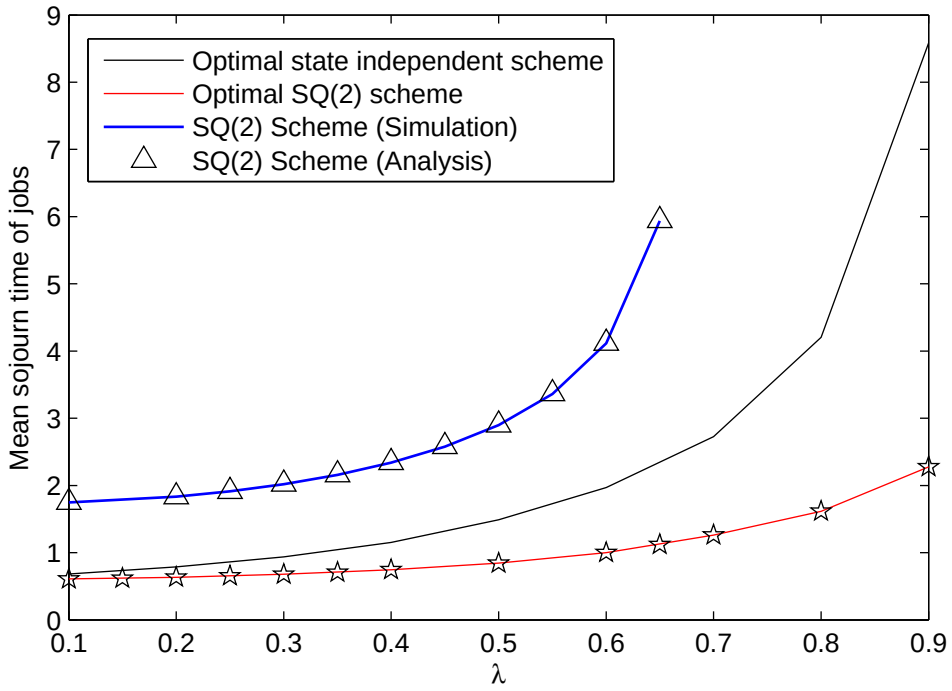
He is currently pursuing his Ph.D in Electrical and Computer Engineering at the University of Waterloo, Canada. His current areas of research are broadly stochastic modeling and analysis of large distributed networks, queuing theory, and network resource allocation and optimization algorithms.

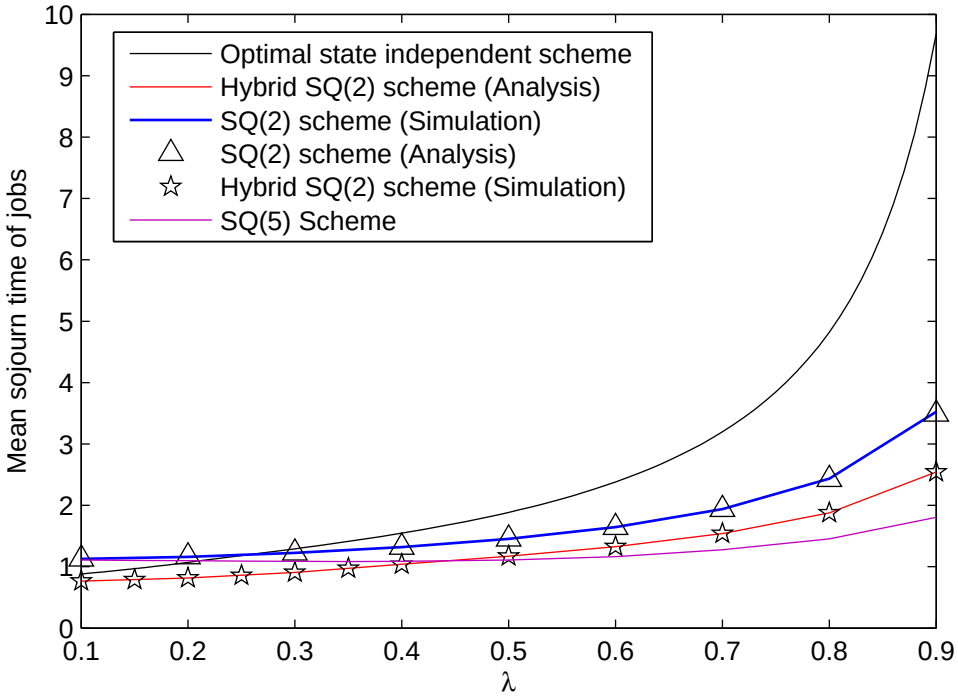
Ravi Mazumdar (F'05) was born in April 1955 in Bangalore, India. He obtained the B.Tech. in Electrical Engineering from the Indian Institute of Technology, Bombay, India in 1977, the M.Sc. DIC in Control Systems from Imperial College, London, U.K. in 1978 and the Ph.D. in Systems Science from the University of California, Los Angeles, USA in 1983.

He is currently a University Research Chair Professor of Electrical and Computer Engineering at the University of Waterloo, Waterloo, Canada and an Adjunct Professor of ECE at Purdue University. He has served on the faculties of Columbia University (NY, USA) and INRS-Telecommunications (Montreal, Canada) . He held a Chair in Operational Research and Stochastic Systems in the Dept. of Mathematics at the University of Essex (Colchester, UK), and from 1999-2005 was Professor of ECE at Purdue University (West Lafayette, USA). He has held visiting positions and sabbatical leaves at UCLA, the University of Twente (Netherlands), the Indian Institute of Science (Bangalore), the Ecole Nationale Superieure des Telecommunications (Paris), INRIA (Paris) and the University of California, Berkeley.

He is a Fellow of the IEEE and the Royal Statistical Society. He is a member of the working groups WG6.3 and 7.1 of the IFIP and a member of SIAM and the IMS. He is a recipient of the IEEE INFOCOM 2006 Best Paper Award and was runner-up for the Best Paper at INFOCOM 1998,

His research interests are in applied probability, stochastic analysis, optimization, and game theory with applications to network science, traffic engineering, filtering theory, and wireless systems.


Optimal state independent scheme
Optimal SQ(2) scheme
SQ(2) Scheme (Simulation)
SQ(2) Scheme (Analysis)
Mean sojourn time of jobs
λ
0
1
2
3
4
5
6
7
8
9
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9


Optimal state independent scheme
Hybrid SQ(2) scheme (Analysis)
SQ(2) scheme (Simulation)
SQ(2) scheme (Analysis)
Hybrid SQ(2) scheme (Simulation)
SQ(5) Scheme
Mean sojourn time of jobs
λ
0
1
2
3
4
5
6
7
8
9
10
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9