Limited Choice and Locality Considerations for Load Balancing

Yu-Tong He
Department of Computing and Software
McMaster University
1280 Main Street West, Hamilton, ON L8S 4L7, Canada
hey3@mcmaster.ca

Douglas G. Down
Department of Computing and Software
McMaster University
1280 Main Street West, Hamilton, ON L8S 4L7, Canada
downd@mcmaster.ca

revised March 12, 2008

Abstract

This paper considers the problem of routing Poisson arrivals to N parallel servers under the condition that the system is heavily loaded. We propose a scheme in which a proportion of arrivals are routed randomly, while the others are routed to one of two neighbouring queues using load information. We show that this scheme, which exploits a limited amount of load information and takes into account locality considerations, achieves performance close to that of a routing policy which requires complete load information. In addition, we show that this scheme has a diffusion scaled queue length process that is the same as if all of the servers were pooled with a single queue (in other words, no routing decision need be made). Our insights provide an additional option in load balancing, complementing the related work on the power of two choices.

1 Introduction

There has been a significant amount of work over the past several years on the "power of two choices" for various load balancing problems (see Mitzenmacher et al. [6] for an overview of existing results and implications). In our work, we concentrate on the problem of dynamically assigning tasks to servers, where tasks arrive sequentially and must be routed to a server on arrival. If we consider the case where all servers are identical, and we are interested in minimizing a task's mean waiting time, then it seems reasonable that we would like to assign an arriving task to the least loaded server, i.e., to join the shortest queue (JSQ). Winston [11] gives optimality properties for such a policy if the service times are exponentially distributed and the local service discipline is first-come-first-serve (FCFS). Weber [9] gives similar results for service times with non-decreasing hazard rates. Unfortunately, such a policy may not be scalable, as gathering complete load information of all servers may be very expensive due to such issues as message passing overhead. On the other hand, no system state information would be required if arriving tasks were simply randomly assigned to a server. While this latter option may be attractive from an information gathering viewpoint, there is usually an unacceptable gap between random assignment and using full system information to make routing decisions.

In [5], Mitzenmacher proposed a load balancing algorithm, which is somewhere between complete random routing and JSQ routing. (We denote it as JSQ-d/N.) Suppose a system consists of N identical servers (service times exponentially distributed with mean one, FCFS discipline) and a Poisson arrival process with rate $N\lambda$. If each arrival chooses d servers independently and uniformly at random from the N servers and joins the one with the shorter queue, then the limiting behaviour of such choice leads to exponential improvements in the expected waiting time in the system for any $d \geq 2$ over $d = 1$. To be specific, let $T_d(\lambda)$ denote the expected time an arrival spends in the limiting system ($N \to \infty$) for $d \geq 2$, then the asymptotic improvement as the system approaches unity load is

$$\lim_{\lambda \to 1} \frac{T_d(\lambda)}{\log T_1(\lambda)} = \frac{1}{\log d}, \tag{1.1}$$

where $T_1(\lambda)$ is the expected waiting time for an $M/M/1$ queue with arrival rate λ and service rate 1. For example, when the system is in heavy traffic, $T_2(1) \approx 3 \log T_1(1)$ implies that JSQ-$2/N$ will achieve exponential improvement over complete random routing. This remarkable result falls under his "power of two choices" label.

In the concluding remarks of [5], the problem of dealing with locality is suggested as an interesting issue, i.e. rather than randomly choosing queues, one may want to constrain the choices of which queues may be chosen together. Insight into this problem is given in Byers et al. [1]. They showed that when n items are placed at n servers with d choices per item, when nearest neighbours are chosen, the maximum number of items assigned to a server is a constant factor larger than a system where the d choices are made randomly.

In this work, we explore similar issues in the context of server load balancing, where Poisson arrivals occur to a number of parallel servers (we begin with the servers being identical, later examining the heterogeneous case). We propose a policy that provides arrivals with either no choice or the choice of two neighbouring queues, for which they join one of the two queues with the shorter expected waiting time. We show that this policy yields a diffusion scaled queue length that is the same as one in which arrivals join a single queue in front of all of the servers and are served in FCFS order. This type of behaviour has been termed *complete resource pooling* (coined in Harrison and López [4]). We are able to apply recent work by Stolyar [8] that gives conditions for complete resource pooling to hold for a class of models that has ours as a special case. These conditions are relatively straightforward to check for our proposed policy. We then proceed to give evidence that other policies (including that in [5]) should behave in a similar manner. All of these policies share the feature that a fraction of the incoming workload may be shifted from any queue to any other queue in the system, which we believe is the mechanism that leads to the significant performance improvement.

We believe that this paper makes several contributions. The first is the one just mentioned, as it suggests that in designing good routing policies, it is key to allow incoming workload to be freely shifted between queues. Furthermore, the fact that the choice may be severely limited (a proportion of arrivals may have no choice) demonstrates (at least in heavy traffic) that not all of the arriving workload needs to be capable of being shifted. In other words, if a small proportion of arrivals can be routed using load information while the remaining arrivals are randomly routed without any load information, the system performance is close to that achieved by a load balancing policy which requires complete load information. This complements the insights given in [1, 5], in particular those in [1]. The framework that we employ also allows us to address generally distributed service times. Finally, we also examine heterogeneous servers.

The organization of the paper is as follows. Section 2 gives the model in detail and describes the diffusion approximation method that we use to analyze routing policies. Section 3 shows that for a system with identical servers, our proposed policy has a diffusion scaled queue length process that is identical to that for a system where no routing decision need be made. Section 4 extends the main results to the case of heterogeneous servers. Section 5 provides a discussion of how our main results should apply to other policies (including that in [5]). Section 6 provides a few numerical results and Section 7 provides some final thoughts. For the paper to be self-contained, Appendix A includes some existing results that will be used in our study.

2 Model

Define a finite set $\mathcal{J} = \{1, ..., N\}$ $(N \geq 2)$. The base system that we study has N parallel single-server queues. Let $\{v_{j,m} : m \geq 1\}$ be a sequence of independent and identically distributed (i.i.d.) random variables, which are formed by the service times at queue j, $j \in J$. We assume

$$\mathbf{E}[v_{j,m}] = \mu^{-1}, \quad \mathbf{var}[v_{j,m}] = \beta^2, \quad \forall j \in J,$$

for all $m \geq 1$. Also, for all $j \in \mathcal{J}$, the sequences $\{v_{j,m}\}$ are assumed mutually independent. Service at each queue is First Come First Served (FCFS). The single arrival stream of tasks follows a Poisson process with finite rate $N\lambda$. A task must be assigned to one of the servers immediately upon arrival. Our performance goal is to minimize a task's mean waiting time, or equivalently by Little's law, to minimize the mean total number of tasks in the system.

We propose a routing policy as follows. With probability p/N $(0 \leq p < 1)$, an arrival is randomly routed to one of the N identical queues. We call such arrivals *dedicated* arrivals. With probability $(1-p)/(N-1)$, an arrival is routed to one of two neighbouring queues j and $j+1$, $j \in \{1, ..., N-1\}$, which has the shorter queue length. We call these *flexible* arrivals.

Like JSQ-2/N (i.e., Mitzenmacher's Two Choices [5]), our policy is more scalable than JSQ, as the dedicated arrivals need no state information, while the flexible ones need information on the state of only two servers. A large value of the flexibility level p implies a small amount of state information is required in making routing decisions. Unlike JSQ-2/N, our policy addresses the problems where (1) not all arrivals need to be flexible for dynamic choice, (2) even for flexible arrivals, not all can afford complete random choice of all servers: the choice may be very limited due to locality constraints or personal preference of the arrivals. Therefore, our proposed policy provides another option to JSQ-2/N, in the case where a more constrained routing choice for arrivals is attractive.

To analyze the behaviour of our policy and compare it with other policies, we adopt the diffusion approximation method. Rather than give a long list of references on this method, we refer to the monograph of Chen and Yao [2] for an overview. The main idea is to establish the diffusion limit for the queueing process of interest (e.g., the queue length process or the waiting time process). To identify the diffusion limit, a sequence of queueing systems is considered. The motivation comes from the fact that the queue length process has a stationary distribution only when the system load ρ is strictly less than one, while its diffusion limit is zero when $\rho < 1$. So the queueing system we are interested in is assumed to be an element in a sequence of systems whose traffic intensities approach one. Once the limit is obtained, it can be used to approximate the queue length process of a stable system by appropriate scaling. Therefore, we can compare diffusion scaled queue length processes for different routing schemes. In contrast with the techniques in [5], we will study a fixed, finite set of queues, as the load on the system approaches unity. We will show that the diffusion scaled total queue length is the same as that of an $M/G/N$ queue in heavy traffic. The limiting process is independent of the flexibility level p, which means even a small proportion of flexible arrivals should yield significant improvement in system performance in heavy traffic.

3 Main Results

We consider a sequence of the identical server systems indexed by n. For the n-th system, the arrivals follow a Poisson process with rate $N\lambda^{(n)}$; service times have mean μ^{-1} and variance β^2. Assume that the following conditions hold

$$\lim_{n\to\infty} \lambda^{(n)} \quad = \quad \lambda, \tag{3.1}$$

and

$$\sup_{n\geq 1, j\in J} \mathbf{E}\left[\left(v_{j,1}^{(n)}\right)^{2+\epsilon}\right] < \infty \tag{3.2}$$

for some $\epsilon > 0$ (i.e., the first and second moments of the service times are assumed finite). In addition, let $\tilde{\mu} = N\mu$, the heavy traffic condition

$$\lim_{n\to\infty} \sqrt{n}\Big(N\lambda^{(n)} - \tilde{\mu}\Big) = c < \infty \tag{3.3}$$

is assumed to be true for some finite constant c. Let $\rho = N\lambda/\tilde{\mu}$ be the system load. Since $c > -\infty$ corresponds to $\rho = 1$, then by (3.1) and (3.3), we let the system go to heavy traffic by increasing the arrival rate while fixing the processing time distribution.

In our main result, we will consider three different routing policies, each operating on the same sequence of systems. These are our proposed policy, JSQ, and one in which there is a single queue and no routing (so the result is an M/G/N system). Let the total queue length processes for these policies be given by $Q_B^{(n)}(t)$, $Q_{JSQ}^{(n)}(t)$, and $Q_M^{(n)}(t)$, respectively (we assume that these are all equal to zero at time zero). We form the diffusion scaled queue length processes as follows:

$$\hat{Q}_B^{(n)}(t) = \frac{1}{\sqrt{n}} Q_B^{(n)}(nt), \tag{3.4}$$

with $\hat{Q}_{JSQ}^{(n)}(t)$ and $\hat{Q}_M^{(n)}(t)$ defined in the same manner.

To state our main results, we need a few more definitions. Let $\xrightarrow{w}$ denote weak convergence (or convergence in distribution for processes in the standard Skorohod space of right continuous functions with left hand limits, see Appendix A.1) and RBM(θ, σ^2) denote a reflected Brownian motion with drift θ and variance σ^2. The following is our main result:

Theorem 3.1.

(i) The diffusion scaled total queue length process of the base system, $\hat{Q}_B^{(n)}(t)$, converges weakly to a one-dimensional reflected Brownian motion $\hat{Q}_B$ which is independent of p, $p \in [0,1)$. That is $\hat{Q}_B^{(n)}(t) \xrightarrow{w} \hat{Q}_B = RBM\left(c, N\lambda\left(1+\mu^2\beta^2\right)\right)$, as $n \to \infty$.

(ii) For JSQ, $\hat{Q}_{JSQ}^{(n)} \xrightarrow{w} \hat{Q}_B$.

(iii) For M/G/N, $\hat{Q}_M^{(n)}(t) \xrightarrow{w} \hat{Q}_B$.

Before we prove the theorem we comment on its implications. First, the theorem justifies the claim that for a stable system with load close to one, the distribution of the scaled queue length process is close to that of the RBM. Therefore, given the system load $\rho < 1$, we can choose the index $n = 1/\sqrt{1-\rho}$ and obtain the approximation of the unscaled queue length process

$$Q_B(t) \approx \hat{Q}_B\left(t(1-\rho)^2\right)/(1-\rho).$$

The mean of the distribution of $\hat{Q}_B(t)$ is close to $N\lambda\left(1+\mu^2\beta^2\right)/2|c|$, which is the mean of the stationary distribution of the RBM.

Next, it is noted that if two unscaled sequences of processes result in the same diffusion limit for their corresponding diffusion scaled processes, the difference between them is of the order $o(\sqrt{n})$. Thus, we are unable to capture, for example the fact that there could be a constant difference between the unscaled sequences of processes. However, it does suggest that the performance of the systems should be relatively close (in particular, for high loads). Theorem 3.1 part (i) implies a small amount of flexibility should give close to the performance improvement given by 100 percent flexibility. Part (ii) implies that using our proposed policy should have performance close to that of JSQ under heavy loads. Part (iii) shows that our proposed policy should have performance close to that of a system where no routing decision is required (such a system clearly provides a lower bound on achievable performance).

Proof of Theorem 3.1.

(i) For the base system considered, define the sets $\mathcal{I}_1 = \{1, ..., N\}$, $\mathcal{I}_2 = \{N+1, ..., 2N-1\}$, $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$. The arrivals can be viewed as consisting of $|\mathcal{I}|$ types, each type i having arrival rate

$$\lambda_i = \begin{cases} p\lambda, & \text{if } i \in \mathcal{I}_1 \text{, the dedicated types,} \\ \frac{N\lambda(1-p)}{N-1}, & \text{if } i \in \mathcal{I}_2 \text{ , the flexible types.} \end{cases} \tag{3.5}$$

A graph $\mathcal{G}$ is constructed which has nodes being arrival type i ($i \in \mathcal{I}$) and queue j ($j \in \mathcal{J}$), and arcs (ij) being the routing activities. To represent $\mathcal{G}$, we use a matrix $\Phi = (\phi_{i,j})_{|\mathcal{I}|\times|\mathcal{J}|}$ with non-negative elements, where $\phi_{i,j}$ is the average rate at which server j's time is allocated to serve type i customers, in the long run (so the total utilization of server j is $\rho_j = \sum_{i\in\mathcal{I}} \phi_{i,j}$). For the base system, we have

$$\Phi = \left[\begin{array}{c} \Phi_1 \\ \hline \Phi_2 \end{array}\right] = \left[\begin{array}{ccccc} \phi_{1,1} & & & & \\ & & \ddots & & \\ & & & & \phi_{N,N} \\ \hline \phi_{N+1,1} & \phi_{N+1,2} & & & \\ & \phi_{N+2,2} & \ddots & & \\ & & \ddots & \phi_{2N-2,N-1} & \\ & & & \phi_{2N-1,N-1} & \phi_{2N-1,N} \end{array}\right]_{(2N-1)\times N} \tag{3.6}$$

where the diagonal matrix Φ_1 represents the routing structure of the dedicated arrivals; the bi-diagonal matrix Φ_2, the flexible arrivals. Then the linear system

$$\sum_{j\in\mathcal{J}} \mu\phi_{i,j} = \lambda_i, \quad \forall i \in \mathcal{I} \quad \text{and} \quad \sum_{i\in\mathcal{I}} \phi_{i,j} = 1, \quad \forall j \in \mathcal{J}, \tag{3.7}$$

has the unique solution

$$\phi_{i,j} = \begin{cases} p, & i \in \mathcal{I}_1, j = i, \\ (1-p)\frac{N-j}{N-1}, & i \in \mathcal{I}_2, j = i(\text{mod } N), \\ (1-p)\frac{j-1}{N-1}, & i \in \mathcal{I}_2, j = i(\text{mod } N) + 1. \end{cases} \tag{3.8}$$

When $0 \le p < 1$, $\phi_{i,j} > 0$ for $i \in \mathcal{I}_2$ and $\phi_{i,j} \ge 0$ for $i \in \mathcal{I}_1$. The uniqueness of Φ satisfying (3.7) implies $\mathcal{G}$ is a connected tree (i.e., all of the queues are connected through the positive elements in Φ_2).

Therefore, from Theorem A.1 in Appendix A.2, the so-called complete resource pooling (CRP) condition holds for the vector $\vec{\lambda} = (\lambda_i)_{1\times|\mathcal{I}|}$ with components λ_i defined in (3.5). When the CRP condition is satisfied, in the heavy traffic limit the parallel queue system effectively forms a single pool of service capacity and the state space of the system information collapses into one dimension, typically making the system much easier to analyze.

Define

$$b_i = \begin{cases} \frac{cp}{N}, & \text{if } i \in \mathcal{I}_1, \\ \frac{c(1-p)}{N-1}, & \text{if } i \in \mathcal{I}_2, \end{cases} \tag{3.9}$$

so that $\sum_{i\in\mathcal{I}} b_i = c$. From (3.5) and (3.7), we have $\lambda = \mu$. Then, using (3.3), we have

$$\lim_{n\to\infty} \sqrt{n}\left(\lambda_i^{(n)} - \lambda_i\right) = b_i. \tag{3.10}$$

Let ξ_j be the workload contribution of server j and ν_i be the workload contribution of type i arrivals. (For the precise definitions of the vectors ξ and ν, see Appendix A.2. We do not go into detail here, as we will see shortly that these values do not appear in our final expressions.) In addition, let $Q_{B,i,j}^{(n)}(t)$ be the number of type i arrivals at queue j at time t and $\hat{Q}_{B,i,j}^{(n)}(t) = Q_{B,i,j}^{(n)}(nt)/\sqrt{n}$. Given (3.1), (3.2) and (3.10), a direct application of Theorem A.2 (in Appendix A.2) yields

$$\sum_{j\in J} \xi_j \sum_{i\in I} \mu^{-1} \hat{Q}_{B,i,j}^{(n)}(t) \xrightarrow{w} \text{RBM}(\theta, \sigma^2), \quad \text{as } n\to\infty,$$

where

$$\theta = \sum_{i\in I} \nu_i b_i, \quad \sigma^2 = \sum_{i\in I} \nu_i^2 \left[\lambda_i + \sum_{j\in J} \mu\phi_{i,j}(\mu\beta)^2\right]. \tag{3.11}$$

From (A.2) (in Appendix A.2), we have

$$\xi_j = \max_i \mu\nu_i, \quad \nu_i = \min_j \xi_j/\mu,$$

which implies that $\forall i \in I$, ν_i is equal to some constant a (which, for our purpose, is not necessary to calculate) and $\forall j \in J$, $\xi_j = a\mu$. This, with (3.7), implies that

$$\hat{Q}_B^{(n)}(t) \xrightarrow{w} \text{RBM}\left(c, N\lambda\left(1+\mu^2\beta^2\right)\right), \quad \text{as } n\to\infty,$$

which is clearly independent of p.

(ii) The JSQ policy is a special case of Stolyar's MinDrift(Q) policy (see Appendix A.3). Let $\xi_j = 1/J$, then Theorem A.2 yields the result. We note that by applying Theorem 3.1 from Zhang and Hsu [12], the same result can be obtained.

(iii) A direct application of Theorem 5 in [7] yields the result. □

At this point, we make the observation that the routing structure given by (3.8) suggests that the intuition behind Theorem 3.1 is that congestion at a particular queue can be alleviated by shifting a sufficient amount of incoming workload from that queue to other queues. The fact that the tree structure for the routing of flexible arrivals allows a fraction of the incoming workload to be shifted anywhere in the system leads to complete resource pooling, i.e. all of the queues are "connected" through the tree structure. What may be a bit surprising is that an arbitrarily small amount of flexibility is enough to achieve this.

4 Extensions

We can extend our policy to cope with heterogeneous servers. This will require a non-uniform choice for the routing probabilities. Let $\{v_{j,m} : m \geq 1\}$ be a sequence of i.i.d. random variables formed by the service times at queue j and assume

$$\mathbf{E}[v_{j,1}] = \mu_j^{-1}, \quad \mathbf{var}[v_{j,1}] = \beta_j^2.$$

Also, for all $j \in \mathcal{J}$, the sequences $\{v_{j,m}\}$ are assumed mutually independent. The service discipline at each queue is still FCFS. The single arrival stream remains a Poisson process with rate $N\lambda$.

It is known from Appendix A.3 that the "join the shortest expected waiting time" (JSEW) policy, i.e., route an arrival of type i to queue j satisfying

$$j \in \arg\min_{j\in\mathcal{J}} \frac{\sum_i Q_{ij}(t)}{\mu_j},$$

exhibits complete resource pooling for appropriate routing structures. Here $Q_{ij}(t)$ is the number of type i arrivals in queue j at time t and $\mathcal{J}$ is the set of servers that can serve type i arrivals. Thus, for this section we will use JSEW routing (note that JSEW routing is simply JSQ if the servers are identical).

We modify our policy as follows. Let $\tilde{\mu} = \sum_{j=1}^{N} \mu_j$. With probability $p\mu_j/\tilde{\mu}$ $(0 \le p < 1)$, a (dedicated) arrival is routed to one of the N queues. With probability $(1-p)(\mu_j + \varepsilon_j)/\tilde{\mu}$ (with constant ε_j), a (flexible) arrival is routed to join one of two neighbouring queues, j and $j+1$, $j \in \{1, ..., N-1\}$, which has the shorter expected waiting time. That is, the routing probability is proportional to the mean service rate of the neighbour on the left.

To determine the routing probability for the flexible arrivals, the constants $\{\varepsilon_j\}$ are chosen as

$$\varepsilon_j = \begin{cases} \varepsilon \in \left(0, \min\left(\bar{\varepsilon}, \frac{\mu_N+\mu_{N-1}}{N-2}\right)\right), & j \in \{1, ..., N-2\}, \\ \mu_N - (N-2)\varepsilon, & j = N-1, \end{cases} \tag{4.1}$$

with

$$\bar{\varepsilon} = \min_{j\in\{2,...,N-1\}} \frac{\mu_j}{j-1}. \tag{4.2}$$

Now consider a sequence of the heterogeneous server systems indexed by n. For the n-th system, the arrivals follow a Poisson process with rate $N\lambda^{(n)}$; the service times at the j-th server have mean μ_j^{-1} and variance β_j^2. Conditions (3.1)–(3.3) are still assumed to be true. For the same sequence of systems, we consider three different routing policies: our modified policy, JSEW, and an M/G/N^h queue (i.e. a system with a single queue for N heterogeneous servers, service being FCFS). Let the total queue length processes for these policies be given by $Q_B^{(n)}(t)$, $Q_{JSEW}^{(n)}(t)$, and $Q_M^{(n)}(t)$, respectively. For each of these, we again need their diffusion scaled counterparts defined in the same manner as (3.4). Then we have the following result:

Theorem 4.1.

(i) The diffusion scaled total queue length process, $\hat{Q}_B^{(n)}(t)$, converges weakly to a one-dimensional reflected Brownian motion $\hat{Q}_B$ which is independent of both $p \in [0, 1)$ and ε_j, $j \in \{1, \ldots, N-1\}$. That is $\hat{Q}_B^{(n)}(t) \xrightarrow{w} \hat{Q}_B = RBM\left(c, N\lambda + \sum_{j\in\mathcal{J}} \mu_j^3\beta_j^2\right)$, as $n \to \infty$.

(ii) For JSEW, $\hat{Q}_{JSEW}^{(n)}(t) \xrightarrow{w} \hat{Q}_B$.

(iii) For M/G/N^h, $\hat{Q}_M^{(n)}(t) \xrightarrow{w} \hat{Q}_B$.

Theorem 4.1 says that our policy should have heavy traffic performance close to that in which no routing decision is made, so the suggested choice of routing probabilities should lead to good performance in general.

Proof of Theorem 4.1.

(i) We use the same sets $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}$ defined in the proof of Theorem 3.1 and define the arrival rate vector $\vec{\lambda} = (\lambda_i)_{1\times|\mathcal{I}|}$ with components

$$\lambda_i = \begin{cases} p\mu_j, & i \in \mathcal{I}_1, j = i, \\ (1-p)(\mu_j + \varepsilon_j), & i \in \mathcal{I}_2, j = i(\text{mod } N), \end{cases} \tag{4.3}$$

where ε_j is given in (4.1).

The routing structure matrix Φ is the same as (3.6). Then the linear system (3.7) has the unique solution

$$\phi_{i,j} = \begin{cases} p, & i \in \mathcal{I}_1, j = i, \\ \frac{1-p}{\mu_j} \cdot \left(\mu_j - \sum_{k=1}^{j-1} \varepsilon_k\right), & i \in \mathcal{I}_2,\, j = i(\text{mod } N), \\ \frac{1-p}{\mu_j} \cdot \sum_{k=1}^{j-1} \varepsilon_k, & i \in \mathcal{I}_2,\, j = i(\text{mod } N) + 1. \end{cases} \tag{4.4}$$

Given $0 \le p < 1$ and (4.1), we have $\phi_{i,j} > 0$ for $i \in \mathcal{I}_2$. Again, the uniqueness of Ψ implies that the associated graph $\mathcal{G}$ is a connected tree. So from Theorem A.1, the CRP condition holds for the arrival rate vector defined in (4.3).

Let

$$b_i = \begin{cases} cp\mu_j/\tilde{\mu}, & i \in \mathcal{I}_1,\, j = i \\ c(\mu_j + \varepsilon_j)/\tilde{\mu}, & i \in \mathcal{I}_2,\, j = i(\text{mod } N), \end{cases}$$

so that $\sum_{i\in I} b_i = c$. From (4.3) and (3.7), we have $\tilde{\mu} = \sum_{i\in\mathcal{I}} \lambda_i = N\lambda$. Then, using (3.3), we have

$$\lim_{n\to\infty} \sqrt{n}\left(\lambda_i^{(n)} - \lambda_i\right) = b_i. \tag{4.5}$$

Let $\hat{Q}_{B,j}^{(n)}(t)$ be the diffusion scaled queue length process at the j-th server, Theorem (A.2) yields that

$$\sum_{j\in\mathcal{J}} (\xi_j \mu_j^{-1}) \hat{Q}_{B,j}^{(n)}(t) \xrightarrow{w} \text{RBM}(\theta, \sigma^2), \quad \text{as } n \to \infty,$$

where

$$\theta = \sum_{i\in\mathcal{I}} \nu_i b_i, \qquad \sigma^2 = \sum_{i\in\mathcal{I}} \nu_i^2 \left[\lambda_i + \sum_{j\in\mathcal{J}} \mu_j \phi_{i,j} (\mu_j \beta_j)^2\right].$$

Again from (A.2), we have

$$\xi_j = \max_i \mu_j \nu_i, \quad \nu_i = \min_j \xi_j/\mu_j,$$

which implies that ν_i is equal to some constant a and $\xi_j = a\mu_j$. This, with (3.7), implies that

$$\hat{Q}_B^{(n)}(t) \xrightarrow{w} \text{RBM}\left(c, N\lambda + \sum_{j\in J} \mu_j^3 \beta_j^2\right), \quad \text{as } n \to \infty. \tag{4.6}$$

(ii) JSEW is also a special case of the MinDrift(Q) policy (see Appendix A.3). Let $\xi_j = \mu_j/\tilde{\mu}$, then a direct application of Theorem (A.2) yields the result.

(iii) A direct application of Theorem 5 in [7] yields the result. □

5 Rings and Supermarkets

Motivated by the observation at the end of Section 3, if the mechanism for good performance is that a sufficient proportion of incoming workload can be shifted from any server to any other server through the routing structure, then there are two natural choices that should intuitively lead to better performance (while still keeping the number of choices to be at most two). One of these is to extend our policy so that there is an additional stream of flexible arrivals that is allowed to join the shorter of queues N and 1. This would allow incoming workload to be shifted bidirectionally, rather than unidirectionally. We will call such a routing structure a "ring" structure, as opposed to the "tree" structure of our original policy. Another obvious choice is the JSQ-2/N policy, as it can spread incoming workload over many different queues, so it seems reasonable that it would also have better performance (which would be consistent with observations in [1] for the assignment of a fixed number of tasks). Like the tree structure, the ring structure provides an option to JSQ-2/N, where dynamic routing choice is more constrained. Unfortunately, as seen below, the CRP condition does not hold for both the ring structure and JSQ-2/N, but we suggest a means to make a comparison. As the JSQ-2/N policy has only been defined in the homogeneous servers case, for the rest of this section we stick to that setting.

5.1 Ring Routing Structure

The difficulty in analyzing the ring structure is that the corresponding arrival rate vector $[\lambda_1, ..., \lambda_{2N}]$, which has elements

$$\lambda_i = \begin{cases} p\lambda, & \text{if } i \in \{1, ..., N\}, \\ (1-p)\lambda, & \text{if } i \in \{N+1, ..., 2N\}, \end{cases} \tag{5.1}$$

does not satisfy the CRP condition, because the routing structure matrix

$$\Phi = \left[\begin{array}{cccc} \phi_{1,1} & & & \\ & \ddots & & \\ & & & \phi_{N,N} \\ \hline \phi_{N+1,1} & \phi_{N+1,2} & & \\ & \ddots & \ddots & \\ & & \phi_{2N-1,N-1} & \phi_{2N-1,N} \\ \phi_{2N,1} & & & \phi_{2N,N} \end{array}\right] \tag{5.2}$$

has a cycle in the corresponding graph. This means that there are multiple solutions of the linear system

$$\begin{aligned} \phi_{N+1,1} + \phi_{N+1,2} &= \lambda_{N+1}/\mu \\ &\vdots \\ \phi_{2N,N} + \phi_{2N,1} &= \lambda_{2N}/\mu \\ \phi_{2N,1} + \phi_{N+1,1} &= 1 - \lambda_1/\mu \\ &\vdots \\ \phi_{2N-1,N} + \phi_{2N,N} &= 1 - \lambda_N/\mu, \end{aligned} \tag{5.3}$$

which is of the same form as (3.7). For $0 < |\epsilon| < \min_{(i,j)} \phi_{i,j}$, define a perturbed matrix Φ' which has elements

$$\phi'_{i,j} = \begin{cases} \phi_{i,j} - \epsilon, & \text{if } i = N + j, \\ \phi_{i,j} + \epsilon, & \text{if } i (\text{mod } N) = j - 1. \end{cases}$$

Then (5.3) becomes

$$\begin{aligned} (\phi_{N+1,1} - \epsilon) + (\phi_{N+1,2} + \epsilon) &= \lambda_{N+1}/\mu \\ &\vdots \\ (\phi_{2N,N} - \epsilon) + (\phi_{2N,1} + \epsilon) &= \lambda_{2N}/\mu \\ (\phi_{2N,1} + \epsilon) + (\phi_{N+1,1} - \epsilon) &= 1 - \lambda_1/\mu \\ &\vdots \\ (\phi_{2N-1,N} + \epsilon) + (\phi_{2N,N} - \epsilon) &= 1 - \lambda_N/\mu, \end{aligned} \tag{5.4}$$

which means that if Φ is a solution of (5.3), we can perturb the $\phi_{i,j}$'s along the arcs of the cycle so as to produce a matrix $\Phi' \neq \Phi$, such that Φ' also satisfies (5.3). From Theorem A.1, the CRP condition does not hold in this case, so we cannot directly make conclusions similar to Theorem 3.1.

However, as we shall see below, this does not imply that the ring routing structure will have performance worse than the tree routing structure. (Remember that our intuition suggests that it should be better.) To see this, we modify the ring structure so that the flexible arrivals which join the shorter of queues 1 and N instead join each of those two queues with equal probability. Then the arrival rate vector $[\tilde{\lambda}_1, ..., \tilde{\lambda}_{2N-1}]$ for this modified system has elements

$$\tilde{\lambda}_i = \begin{cases} \frac{1+p}{2}\lambda, & \text{if } i \in \tilde{\mathcal{I}}_1, \\ p\lambda, & \text{if } i \in \tilde{\mathcal{I}}_2, \\ (1-p)\lambda, & \text{if } i \in \tilde{\mathcal{I}}_3, \end{cases} \tag{5.5}$$

where the sets $\tilde{\mathcal{I}}_1 = \{1, N\}$, $\tilde{\mathcal{I}}_2 = \{2, ..., N-1\}$, $\tilde{\mathcal{I}}_3 = \{N+1, ..., 2N-1\}$ and $\tilde{\mathcal{I}}_1 \cup \tilde{\mathcal{I}}_2 \cup \tilde{\mathcal{I}}_3 = \mathcal{I}$.

The corresponding routing structure matrix $\tilde{\Phi}$ has the same form as (3.6) and as a result the linear system (3.7) has the unique solution

$$\tilde{\phi}_{i,j} = \begin{cases} \frac{1+p}{2}, & i \in \tilde{\mathcal{I}}_1, j = i, \\ p, & i \in \tilde{\mathcal{I}}_2, j = i, \\ \frac{1-p}{2}, & i \in \tilde{\mathcal{I}}_3. \end{cases} \tag{5.6}$$

When $0 \leq p < 1$, $\tilde{\phi}_{i,j} > 0$ for $i \in \tilde{\mathcal{I}}_1 \cup \tilde{\mathcal{I}}_3$ and $\tilde{\phi}_{i,j} \geq 0$ for $i \in \tilde{\mathcal{I}}_2$. From Theorem A.1, the CRP condition holds for the arrival rate vector defined in (5.5).

Consider a sequence of these modified systems where conditions (3.1)– (3.3) are assumed true. Let

$$b_i = \begin{cases} \frac{c(1+p)}{2N}, & \text{if } i \in \tilde{\mathcal{I}}_1, \\ \frac{cp}{N}, & \text{if } i \in \tilde{\mathcal{I}}_2, \\ \frac{c(1-p)}{N}, & \text{if } i \in \tilde{\mathcal{I}}_3, \end{cases} \tag{5.7}$$

and as before, we see $\sum_{i \in \mathcal{I}} b_i = c$. Thus, using (3.3), (3.10) is satisfied.

Let $\tilde{Q}_{R,j}(t)$ be the number of tasks in queue j at time t. By applying Theorem $3.1(i)$, we have

$$\frac{1}{\sqrt{n}} \sum_{j \in \mathcal{J}} \tilde{Q}^{(n)}_{R,j}(nt) \xrightarrow{w} \text{RBM}\left(c, N\lambda\left(1 + \mu^2\beta^2\right)\right), \quad \text{as } n \to \infty. \tag{5.8}$$

This means that the system in which the modified version of the ring routing structure is applied, has the total queue length process achieving the same RBM limit as that of the original tree model.

At this point, we suggest a relationship $L_R \leq \tilde{L}_R$, where L_R and $\tilde{L}_R$ are the mean numbers in the system with the original and modified ring structures, respectively. While we are unable to provide a proof, the intuition is that if at the two end queues, queue 1 has a much higher workload than queue N, the original ring structure enables the incoming workload to be shifted directly from queue 1 to queue N, while the modified structure only allows sequential shifting through queues 2 to $N-1$. Apparently, congestion is alleviated more quickly in the original ring structure than the modified one. It would be useful to prove this relationship. This, together with (5.8) and Theorem $3.1(iii)$ suggest that the ring routing structure will also perform very well.

We conjecture that although the total queue length process of the ring structure does not collapse into a one-dimensional RBM in an arbitrarily long time range (since the CRP condition does not hold), a linear combination of its queue length processes collapses, at each point of time, into a one-dimensional RBM and achieves the same limit as that of the original tree model. It has been noticed that given the routing structure matrix in (5.2), there are multiple solutions of $\phi_{i,j}$'s to the linear system

$$\sum_{j \in \mathcal{J}} \mu\phi_{i,j} = \lambda_i, \quad \forall i \in \mathcal{I}_R \quad \text{and} \quad \sum_{i \in \mathcal{I}_R} \phi_{i,j} = 1, \quad \forall j \in \mathcal{J}, \tag{5.9}$$

where $\mathcal{I}_R = \{1, ..., 2N\}$ and λ_i is defined in (5.1). Suppose $\Phi = (\phi_{i,j})_{2N\times N}$ and $\Phi' = (\phi'_{i,j})_{2N\times N}$ are two of the solutions. Recalling (3.11) in the proof of Theorem 3.1(i), let

$$\begin{aligned}\sigma^2 &= \sum_{i\in\mathcal{I}_R} \nu_i^2 \left[\lambda_i + \sum_{j\in\mathcal{J}} \mu\phi_{i,j}(\mu\beta)^2\right], \\ \sigma'^2 &= \sum_{i\in\mathcal{I}_R} \nu_i^2 \left[\lambda_i + \sum_{j\in\mathcal{J}} \mu\phi'_{i,j}(\mu\beta)^2\right],\end{aligned}$$

where ν_i^2 is the workload contribution of type i arrivals. Using (5.9), we have

$$\sigma^2 = \sigma'^2 = \sum_{i\in\mathcal{I}_R} \nu_i^2\lambda_i \left(1+\mu^2\beta^2\right). \tag{5.10}$$

If at two different points in time, the system characterized by Φ and Φ' respectively followed an RBM limit and the limit was characterized by the corresponding σ^2 and σ'^2, (5.10) implies the two limits would be identical.

5.2 JSQ-$2/N$

In the supermarket model, there are N $(N > 2)$ parallel single-server queues, where the service times are i.i.d.and exponentially distributed with mean μ^{-1}. The single arrival stream follows a Poisson process with rate $N\lambda$. With probability $1/N$, queue j is selected and with probability $1/(N-1)$, queue j' is selected, $j,\ j' \in J$, $j \neq j'$. An arrival chooses to join the shorter of queues j and j'.

Define the set $\mathcal{I} = \{1, ..., N(N-1)/2\}$, so the arrivals consist of $|\mathcal{I}|$ *flexible* types, each with rate

$$\lambda_i = \frac{2\lambda}{N-1}, \qquad i \in \mathcal{I}. \tag{5.11}$$

Again, however, the arrival rate vector does not satisfy the CRP condition, because the routing structure matrix

$$\Phi = \left[\begin{array}{ccccc} \phi_{1,1} & \phi_{1,2} & & & \\ \vdots & & & \ddots & \\ \phi_{N-1,1} & & & & \phi_{N-1,N} \\ \hline & \phi_{N,2} & \phi_{N,3} & & \\ & \vdots & & \ddots & \\ & \phi_{2N-3,2} & & & \phi_{2N-3,N} \\ \hline & & \ddots & & \\ \hline & & & \phi_{\frac{N^2-N}{2},N-1} & \phi_{\frac{N^2-N}{2},N} \end{array}\right] \tag{5.12}$$

has multiple cycles in the corresponding graph. These multiple cycles reflect the mechanism by which the JSQ-$2/N$ policy shifts the workload among the queues, i.e. the workload is shifted not just to the neighbouring queues, but to all of the other queues without consideration of locality constraints. Thus our intuition suggests that the JSQ-$2/N$ policy should perform even better than the ring structure.

To see this, we can modify the JSQ-$2/N$ policy such that the modifications produce the tree routing structure, by making some of the arrivals dedicated. It follows from the facts that (1) if each flexible arrival can join only two out of N queues and the routing structure matrix for the flexible arrivals is a connected tree, then it must have the same form as Φ_2 in (3.6), i.e., there are exactly $(N-1)$ flexible types; (2) the number of dedicated types is at most N, which is equal to the number of queues.

The arrival streams that choose between queues j and $j+1$, $j = 1, \ldots, N-1$ are unchanged. For all of the other streams, we convert them to dedicated arrivals by the following manner: if a stream is choosing between queues k and ℓ ($k < \ell$, $\ell \neq k+1$), then it is modified to randomly join queues k and ℓ, with equal probability. Using the same reasoning as at the end of Section 5.1, it is reasonable that this makes the performance of the system worse. Thus there is now a dedicated arrival stream to each queue of rate $(N-2)\lambda/N$.

The routing structure matrix thus has the same form as (3.6) and as a result the linear system (3.7) has the same unique solution as (3.8), with $p = (N-2)/N$. Now we can apply Theorem 3.1 and deduce that the modified system has the same diffusion limit as an M/G/N queue. As the modifications have degraded the performance, then JSQ-2/N should also exhibit good performance under high loads (see Figures 1 and 6).

6 Simulation

In our simulation work, we try to give some idea of the performance improvement that can be achieved using different routing policies, as one backs away from heavy traffic (but still keeping the system heavily loaded). Five simulation models are compared. Each model has a single Poisson arrival stream with rate $N\lambda$ and N identical servers, each with rate $\mu = 1$. Models 1 to 4 have the same topology: servers work in parallel and each server maintains its own queue with a buffer of infinite size. But they adopt different routing structures. In Model 1, an incoming task is randomly routed to one of the identical queues, with equal probabilities. So, Model 1 is equivalent to N $M/G/1$ queues. Models 2, 3 and 4 use the tree structure, ring structure and JSQ-$2/N$, respectively. Model 5 is an $M/G/N$ queue, so no routing is needed. We compare the performance of Models 2 to 5, using Model 1 as a reference. All statistics have an accuracy no worse than 5 percent at a 95 percent confidence level.

We begin with exponential service times. Using N $M/M/1$ queues (each being 95% loaded) as a reference, we first compare the mean number in system of the tree structure (our original policy) and study the impact of the proportion of flexible arrivals, $(1-p)$. The performance improvements of the tree structure with different levels of flexibility are shown in Table 1.

Several observations can be made from Table 1. First, for the tree structure under high load, there is a significant improvement for even a very small level of flexibility (about 20 percent improvement at 3 percent flexibility, around 40 percent improvement at 10 percent flexibility). Also, at 30 percent flexibility, the amount of improvement is about 80 percent of that with 100 percent flexibility. Thirdly, at 100 percent flexibility, the improvements increase as the number of queues N increases. The first two observations are consistent with Theorem 3.1, in that (1) in heavy traffic, the performance of the tree structure should approach that of an $M/M/N$ queue, which obviously outperforms N $M/M/1$ queues, (2) such performance improvement will be independent of the flexibility level under very high load. The third observation supports the intuition behind Theorem 3.1, in that shifting incoming workload from one queue to the other queues

Table 1: Total mean queue lengths vs. the proportion of flexible arrivals, i.i.d. exponential service times, 95% loaded

Model	$N\times$M/M/1	Tree				
λ	0.95	0.95				
μ	1	1				
$(1-p)$	–	0.03	0.1	0.3	0.5	1
	$N=4$					
Total queue length	76.00	59.74	47.16	36.14	31.20	26.87
Improvement	0%	22%	38%	53%	59%	65%
	$N=20$					
Total queue length	380.00	293.34	226.70	161.46	137.55	115.79
Improvement	0%	23%	41%	58%	64%	70%
	$N=100$					
Total queue length	1900.00	1500.41	1135.10	766.11	627.97	469.98
Improvement	0%	21%	40%	60%	67%	75%

plays an important role in making performance improvement. When the system size grows, more candidate queues are available for shifting workload and each queue is less congested. Actually, it can be seen that the average queue length at each queue is shorter in a larger system.

Next, in Figure 1, we compare the performance of the tree, ring and JSQ-2/N structures with an $M/M/N$ queue, using the improvement of expected waiting time (over Model 1). The tree and the ring structures are at 100 percent flexibility. It is noted that the improvements for the three routing structures appear to be of the same order of magnitude. This is consistent with our observations in Section 5, and combined with the observations in [5] that giving each arrival two choices yields an exponential improvement (over one choice, i.e., Model 1), this would suggest that all of these routing policies are roughly equivalent in terms of giving a significant improvement. As suggested in Section 5, the JSQ-2/N policy would be preferred if implementable, an observation supported by the simulation results. Note that our results are also consistent with the observation in [1] that for the static assignment problem, using nearest neighbour policies gives only a constant degradation of performance (in terms of maximum queue length) over completely random assignment.

Thirdly, we study the effects of changing the traffic load. We let each queue be 70% and 85% loaded in the reference model ($N \times M/M/1$). Figure 2 shows that the improvements under moderate traffic load are relatively less than those under heavy traffic (so are the improvements of the $M/M/N$ queues, see Figure 3). When the proportion of flexible arrivals increases, the improvements increase at a speed smaller than that in heavy traffic. For example in Figure 2, at 30 percent flexibility, the amount of improvement is only around 60 percent of that with 100 percent flexibility. Figure 3 again supports the observations made in Figure 1, while in moderate traffic the difference between the tree, ring and JSQ-2/N structures becomes smaller. All these observations are consistent with the fact that as the system backs further off the heavy traffic conditions, the performance of the three policies goes further away from the lower bound of the achievable performance that is indicated by Theorem 3.1.

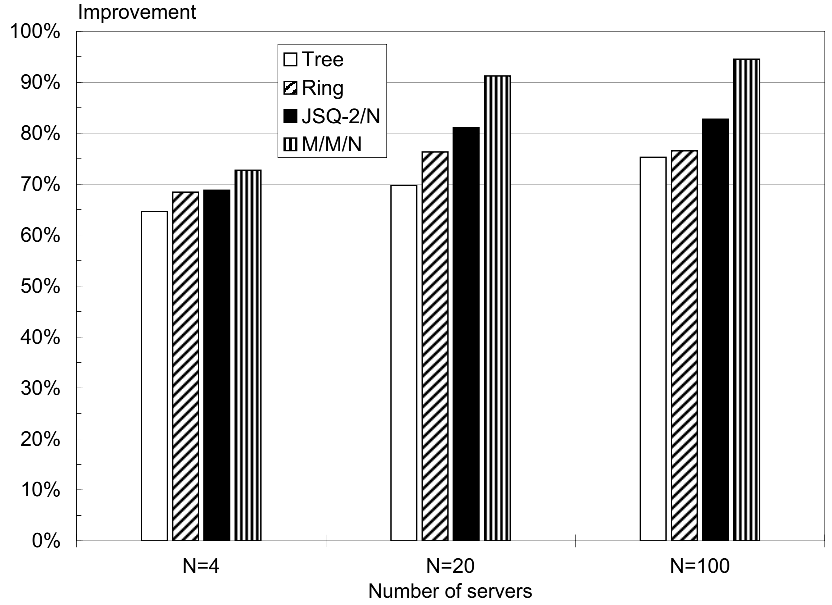


Figure 1: Improvement of expected waiting time vs. routing structures, i.i.d. exponential service times, 95% loaded

Fourthly, we examine the effects of changing the service time variance. Two more service time distributions are applied. One is an Erlang-k distribution with rate one and variance 0.1. The other is a hyper-exponential distribution with rate one and variance 10.0. The results are shown in Figures 4 and 6. In addition to the observations made for the exponential service times setting, we see that all three policies (tree, ring, JSQ-2/N) have larger improvement in systems with larger service time variance than in those with small variance. This is probably not too surprising, as it follows from the observation that when the service time variance is small, the performance is less sensitive to the policy, i.e., for small service time variance if some policy balances the load over long time scales, it is highly likely to also balance the load under shorter time scales. For example, in the extreme of constant service times, an optimal routing policy would be round robin. On the other hand, with large service time variance, load imbalances may occur over short time scales due to the variability in service times, so it becomes more desirable to be able to shift the incoming work between queues.

Fifthly, we look at three models with heterogeneous servers. Each model has 20 parallel queues. The service time distribution at queue j is exponential with rate μ_j. Let the service rate vector be $[1, 2, ..., 20]$ and the single Poisson arrival stream have rate $\tilde{\lambda} = 199.5$, so that $\tilde{\lambda}/\sum_{j=1}^{20} \mu_j = 0.95$. For the system with 20 $M/M/1$ queues, the arrivals are routed to queue j at rate $0.95\mu_j$, so the mean number at each queue is the same and the mean waiting time is calculated by the total mean number in system divided by $\tilde{\lambda}$. In Table 2, we can see that both the tree and ring routing structures yield similar improvements as those seen in the homogeneous server case. Actually, we know from Theorems 3.1 and 4.1 that in the case of exponential service times, both the homogeneous and the heterogeneous systems have the same reflected Brownian motion limit (when the CRP condition is satisfied), so this observation is not surprising.

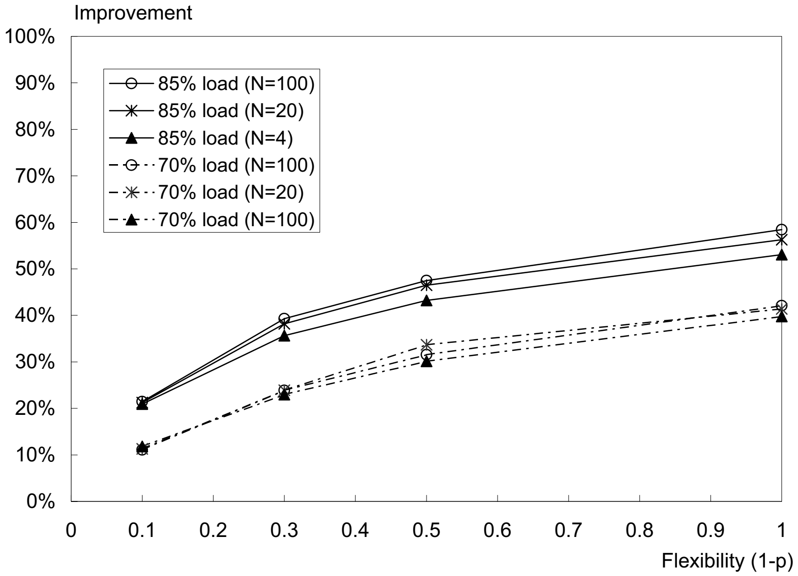


Figure 2: Improvement of total mean queue lengths vs. system load, i.i.d. exponential service times

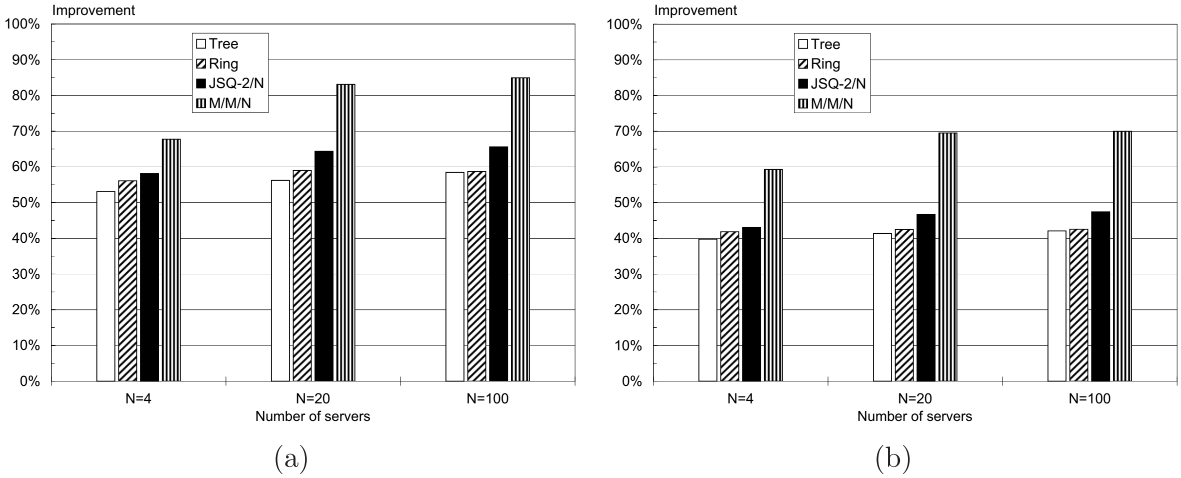
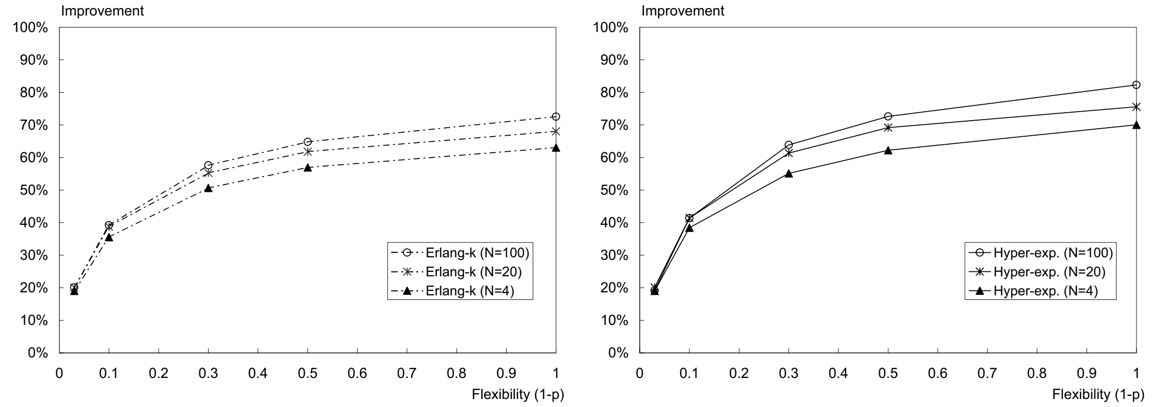
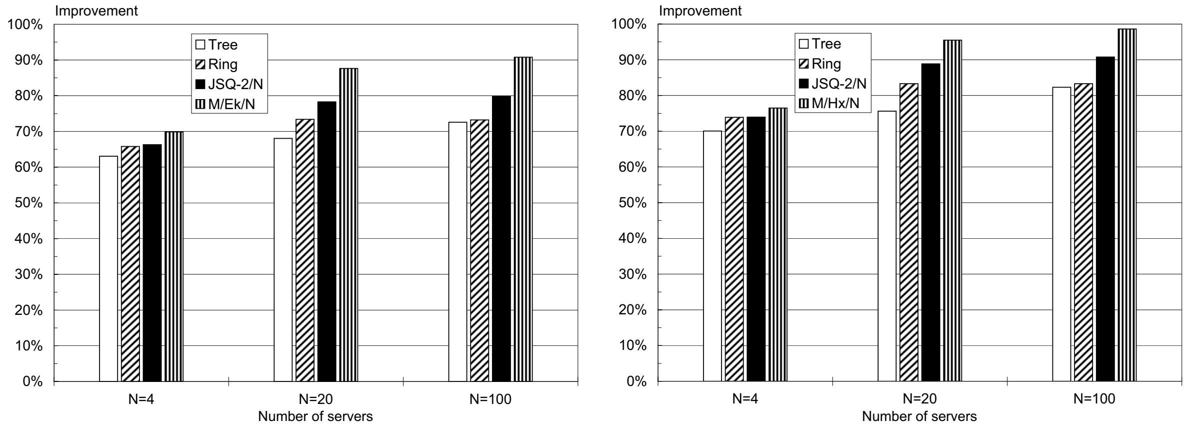


Figure 3: Improvement of expected waiting time vs. routing structures, i.i.d. exponential service times, (a) 85% loaded, (b) 70% loaded,



(a) (b)

Figure 4: Improvement of total mean queue lengths vs. the proportion of flexible arrivals, 95% loaded, (a) i.i.d. Erlang-k service times, (b) i.i.d. hyper-exponential service times



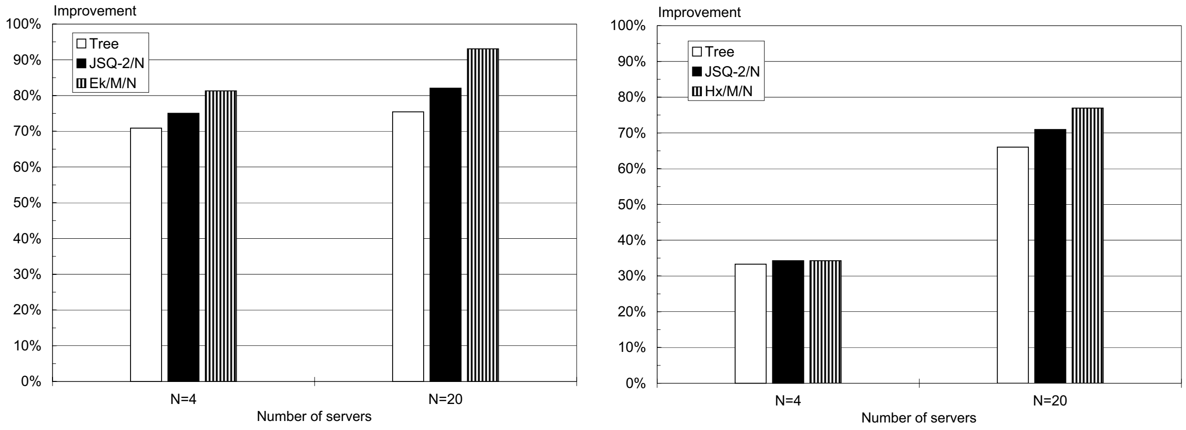
(a) (b)

Figure 5: Improvement of expected waiting time vs. routing structures, 95% loaded, (a) i.i.d. Erlang-k service times, (b) i.i.d. hyper-exponential service times

Table 2: Expected waiting times, heterogeneous servers, exponential service times

Model	N^h × M/M/1	Tree	Ring
N^h	20		
$\tilde{\lambda}$	199.5		
$(\mu_j)_{j\in\{1,\ldots,N^h\}}$	$[1, 2, ..., 20]$		
$(1-p)$	–	1	1
Expected waiting times	1.90	0.45	0.44
Improvement	0%	76%	77%

Finally, we test some cases to see whether the main results of this paper are valid even for non Poisson arrivals. Two more inter-arrival time distributions are examined, Erlang$-k$ and hyperexponential, with squared coefficients of variation 0.1 and 10, respectively. Figure 6 shows the simulation results of the performance achieved by the three routing structures in homogeneous systems, corresponding to those in Theorem 3.1. The improvement of waiting time is calculated using 100 percent random routing ($p = 1$) as a reference. It can be seen that when the system load is high, the tree (with 100 percent flexible arrivals) and JSQ$-2/N$ routing structures achieve performance close to that of the multi-server single queue, which is the lower bound of the achievable performance. This is similar to what has been seen in Figure 1, however, currently we are not able to prove if Theorem 3.1 holds for non Poisson arrivals.



(a) (b)

Figure 6: Improvement of expected waiting time vs. routing structures, 95% loaded, exponential service times, (a) i.i.d. Erlang-k inter-arrival times, (b) i.i.d. hyper-exponential inter-arrival times

7 Conclusion

Using diffusion limits we have provided an explanation for the benefits of certain limited choice routing structures for the problem of load balancing in parallel server systems. In addition to this viewpoint, we have also demonstrated that such schemes are effective for service times with general distributions, as well as heterogeneous servers. The schemes that we have suggested are competitive with that in [5], which we hope gives designers an additional option.

On the methodological side, it is interesting to note that in Section 6, even at 95 percent load, the resulting mean queue lengths are small to moderate. So, while the techniques presented here are useful for classifying policies, it may be useful to examine whether the techniques of Halfin and Whitt [3] yield limits which allow one to differentiate between various policies in finer granularity (and also give better approximations). In particular, using such limits should capture the relation (1.1), which our technique is unable to do. However, it is not clear how to adapt such techniques to a system where routing decisions must be made on arrival ([3] has a single queue and many servers).

8 Acknowledgment.

The authors thank the anonymous referees for their valuable comments.

References

[1] J.W. Byers, J. Considine and M. Mitzenmacher. Geometric generalizations of the power of two choices. *Proceedings of the Sixteenth Annual ACM Symposium on Parallelism in Algorithms and Architectures*, Barcelona, 54–63, 2004.

[2] H. Chen and D.D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*, Springer New York, 2001.

[3] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, **29**:567–588, 1981.

[4] J. M. Harrison and M. J. López. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, **33**:339–368, 1999.

[5] M. Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems*, **12**(10):1094–1104, 2001.

[6] M. Mitzenmacher, A. Richa and R. Sitaraman. The power of two choices: a survey of techniques and results. *Handbook of Randomized Computing: volume 1*, P. Pardalos, S. Rajasekaran and J. Rolim (eds.), Kluwer, 255–312, 2001.

[7] M. Reiman. Some diffusion approximations with state space collapse. *Modelling and Performance Evaluation Methodology*, F. Baccelli and G. Fayolle (eds.), Lecture Notes in Control and Information Sciences, **60**:209–240, Springer, New York, 1984.

[8] A. L. Stolyar. Optimal routing in output-queued flexible server systems. *Probability in the Engineering and Informational Sciences*, **19**:141–189, 2005.

[9] R. W. Weber. On the optimal assignment of tasks to parallel servers. *Journal of Applied Probability*, **15**:406–413, 1978.

[10] R. J. Williams. On dynamic scheduling of a parallel server system with complete resource pooling. *Fields Institute Communications*, **28**:47–71, 2000.

[11] W. Winston. Optimality of the shortest line discipline. *Journal of Applied Probability*, **14**:181–189, 1977.

[12] H. Zhang and G.-H. Hsu. Heavy traffic limit theorems for a sequence of shortest queueing systems. *Queueing Systems*, **21**:217–238, 1995.

A Mathematical Background

To make this paper self-contained, we provide the mathematical background on weak convergence and complete resource pooling.

A.1 Weak Convergence

Let the metric space (S, m) be endowed with the Borel σ-field $\mathcal{B}(S)$ and X be a mapping from a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ to $(S, \mathcal{B}(S))$. The distribution of X is the image probability measure P induced by X on $(S, \mathcal{B}(S))$, denoted as $P(A) := \mathcal{P}(\{\omega \in \Omega : X(\omega) \in A\})$, $A \in \mathcal{B}(S)$. If S is a space of K-dimensional real-valued functions which are defined on the subinterval $[\,0, T\,]$ of the real line and are right-continuous with left limits, X is a K-dimensional stochastic process. The corresponding function space is denoted as $\mathcal{D}$. Let $\{X_n : n \geq 1\}$ be a sequence of stochastic processes, all defined on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Let P and P_n be the distributions of X and X_n, respectively. We say that P_n converges weakly to P if for every bounded and continuous function f on $\mathcal{D}$,

$$\lim_{n\to\infty} \int_{\mathcal{D}} f dP_n = \int_{\mathcal{D}} f dP.$$

In other words, X_n converges weakly to X (or X_n converges to X in distribution), denoted by $X_n \xrightarrow{w} X$, if and only if $\lim_{n\to\infty} E\,[f(X_n)] = E[f(X)]$, for every f.

A.2 Complete Resource Pooling

We follow [8] to introduce the mathematical definition of the complete resource pooling (CRP) condition.

Let $\mathcal{I} = \{1, ..., I\}$ be the set of task types and $\mathcal{J} = \{1, ..., J\}$ be the set of servers. Define a matrix $\Psi = (\psi_{i,j})_{I\times J}$, with all $\psi_{i,j} \geq 0$. Each element $\psi_{i,j}$ is the average rate at which server j's time is allocated to serve type i tasks, in the long run. So the total utilization of server j is $\rho_j = \sum_{i=1}^{I} \psi_{i,j}$. Let $\mu_{i,j}$ be the mean service rate of task type i at server j. The service capacity for type i tasks is $\kappa_i = \sum_{j=1}^{J} \mu_{i,j}\psi_{i,j}$. Given the matrix Ψ, if tasks of type i are routed upon arrival to queue j at the average rate $(\mu_{i,j}\psi_{i,j})$, then the total service capacity for type i tasks equals the mean arrival rate λ_i.

Define vectors $\lambda = [\,\lambda_1, \ldots, \lambda_I\,]$, $\kappa = [\,\kappa_1, \ldots, \kappa_I\,]$ and $\rho = [\,\rho_1, \ldots, \rho_J\,]$. The utilization region is denoted by $\mathcal{U} = \{\rho \in \mathbb{R}^J_+ : \kappa \geq \lambda\}$, where $\mathbb{R}^J_+ = \{x \in \mathbb{R}^J : x \geq 0\}$ and the vector comparison is component-wise. Let the vector $\xi^* = [\,\xi^*_1, \ldots, \xi^*_J\,]^T$ be the outer normal vector to the convex polyhedron $\mathcal{U}$ at the point $\mathbf{1} \in \mathbb{R}^J$. The inner product of vectors ρ and ξ^* is written as $\rho \cdot \xi^*$.

Theorem A.1 ([8], Lemma 3, complete resource pooling)**.** *The CRP condition for a fixed vector λ holds if and only if the following two conditions hold.*

(i) Vector $\mathbf{1} \in \mathbb{R}^J$ solves the problem

$$\begin{aligned} \min_{\rho \in \mathcal{U}} \quad & \rho \cdot \xi^* \\ \textit{s.t.} \quad & \kappa \geq \lambda. \end{aligned}$$

(ii) The matrix Ψ which solves the linear system

$$\lambda = \kappa, \quad \rho = \mathbf{1} \tag{A.1}$$

is unique.

Let the matrix Ψ^* be a solution to (A.1). A graph $\mathcal{G}$ is constructed with nodes being task types i and servers j, arcs (i,j) corresponding to a positive element $\psi^*_{i,j} > 0$. The CRP condition is equivalent to the condition that the graph $\mathcal{G}$ is a tree (Corollary 5.4 in [10]).

Define the capacity region $\mathcal{K} = \{\kappa \in \mathbb{R}^I_{++} : \rho \leq \mathbf{1}\}$, where $\mathbb{R}^I_{++} = \{x \in \mathbb{R}^I : x > 0\}$. Let the vector $\nu^* = [\,\nu^*_1, \ldots, \nu^*_I\,]^T$ be the outer normal vector to the convex polyhedron $\mathcal{K}$ at the point λ. The CRP condition also implies $\mathbf{1} \cdot \xi^* = \lambda \cdot \nu^*$. Moreover, ξ^* is related to ν^* as follows:

$$\xi^*_j = \max_i \mu_{i,j} \nu^*_i,\ j \in \mathcal{J} \quad \text{and} \quad \nu^*_i = \min_j \xi^*_j / \mu_{i,j},\ i \in \mathcal{I}. \tag{A.2}$$

The component ξ^*_j is called the workload contribution of server j; ν^*_i is the workload contribution of task type i. By workload, we mean the amount of unfinished processing time of all tasks in the system.

A.3 Reflected Brownian Motion Limit

Finally, we introduce the diffusion limit of the total weighted workload process in a system which operates with Stolyar's MinDrift(Q) routing rule [8].

Let $Q_{i,j}(t)$ denote the number of type i tasks at server j at time t, including the one in service. The Q-estimated workload at server j is $Z_j(t) = \sum_{i=1}^I \mu^{-1}_{i,j} Q_{i,j}(t)$. The total (server) workload of the system is

$$Z(t) = \sum_{j=1}^J \xi^*_j Z_j(t), \tag{A.3}$$

which is weighted by the server contributions ξ^*_j.

Assume that each server j is assigned a convex holding cost function $C_j(\cdot)$, whose first derivative $C_j^{'}(\cdot)$ is strictly increasing in its argument. The MinDrift(Q) rule routes a type i customer at arrival time t to a server j which satisfies

$$j \in \arg\min_{j\in\mathcal{J}} \frac{C_j^{'}(Z_j(t))}{\mu_{i,j}}.$$

Ties are broken arbitrarily. In the special case where $I = 1$ and $\mu_{i,j} = \mu$ for all $i \in \mathcal{I}$, $j \in \mathcal{J}$, MinDrift(Q) reduces to JSQ, if the cost function is of the form $C_j(Z_j(t)) = \gamma Z_j^2(t)$, for any positive constant γ. If $I = 1$ and $\mu_{i,j} = \mu_j$ for all $i \in \mathcal{I}$, MinDrift(Q) reduces to JSEW, if the cost function is of the form $C_j(Z_j(t)) = \mu_j Z_j^2(t)$.

Suppose there is a system equipped with the MinDrift(Q) routing rule and an arbitrary non preemptive, work-conserving local scheduling rule. Associated are the mean arrival rate vector λ which satisfies the CRP condition, the matrix Ψ^*, and the vectors ξ^* and ν^*. All of the queues are empty at the initial time $t = 0$. Consider a sequence of such systems, indexed by n. For the n-th system, the inter-arrival times of task type i have mean $\left(\lambda_i^{-1}\right)^{(n)}$ and variance $\left(\alpha_i^2\right)^{(n)}$; the service times at server j for type i arrivals have mean $\mu_{i,j}^{-1}$ and variance $\beta_{i,j}^2$. We assume that the following conditions hold

$$\lim_{n\to\infty} \left(\alpha_i^2\right)^{(n)} = \alpha_i^2 \tag{A.4}$$

and

$$\sup_{n\geq 1, i\in\mathcal{I}} \mathbf{E}\left[\left(u_{i,1}^{2+\epsilon}\right)^{(n)}\right] < \infty, \tag{A.5}$$

$$\mathbf{E}\left[v_{i,j,1}^{2+\epsilon}\right] \equiv c_{i,j}, \tag{A.6}$$

for some $\epsilon > 0$ and finite constant $c_{i,j}$. In addition the heavy traffic condition

$$\lim_{n\to\infty} \sqrt{n}\left(\lambda_i^{(n)} - \lambda_i\right) = b_i \tag{A.7}$$

for some finite constant b_i is assumed to be true for all $i \in \mathcal{I}$.

Define the scaled processes for (A.3) as $\hat{Z}^{(n)}(t) = Z^{(n)}(nt)/\sqrt{n}$.

Theorem A.2 ([8], Theorem $2(i)$). *If* (A.4)–(A.7) *hold, then as* $n \to \infty$, $\hat{Z}^{(n)}(t) \xrightarrow{w} \hat{Z} = \mathit{RBM}\left(\theta,\ \sigma^2\right)$, *where*

$$\theta = \sum_{i\in\mathcal{I}} \nu_i^* b_i, \quad \sigma^2 = \sum_{i\in\mathcal{I}} (\nu_i^*)^2 \left[\lambda_i^3 \alpha_i^2 + \sum_{j\in\mathcal{J}} \psi_{i,j}^* \mu_{i,j}^3 \beta_{i,j}^2\right].$$