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 July 31, 2006

Abstract

We consider the problem of routing Poisson arrivals to N parallel servers under the condition that the system is heavily loaded. We show that a scheme which limits arrivals to one or two choices and takes into account locality considerations 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). We compare our insights with 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. Winston [11] gives optimality properties for such a policy if the service times are exponentially distributed, while 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 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 the following load balancing algorithm. In a system with N identical servers (service times exponentially distributed, mean one) and a Poisson arrival process with rate $N\lambda$, if an arrival randomly (with equal probabilities) chooses two of the N servers and joins the queue of the server with the shortest queue length, then there is an exponential improvement in the mean waiting time. Furthermore, increasing the number of choices for an arrival results in only a constant improvement over two choices. This remarkable result fall 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). In order to compare policies, we compare diffusion scaled queue length processes for different routing schemes. This is a technique that is now standard in the queueing community. Rather than give a long list of references, we refer to the monograph of Chen and Yao [2] for an overview. We propose a policy that provides arrivals with either no choice (with probability p) or the choice of two neighbouring queues (with probability $1-p$), for which they join the queue with the shortest 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 First Come First Served (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 that (at least in heavy traffic), that not all of the arriving workload needs to be capable of being shifted. This complements the insights given in [1, 5], in particular that in [1]. The suggestion is that routing policies that satisfy this property will have similar performance. Unfortunately, we cannot differentiate between policies on a finer level (further comments on this are given in the Conclusion). On the other hand, we believe that identifying this mechanism provides a reasonable first cut in terms of classifying policies. The framework that we employ also allows us to show that these insights extend to non-exponentially distributed service times and heterogeneous servers.

The organization of the paper is as follows. Section 2 gives the model in detail and also describes our proposed policy plus other policies that are used as a basis for comparison. 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.

2 Model

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

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

Also, for all $j \in J$, the sequences $\{v_{j,m}\}$ are assumed mutually independent. Service at each queue is First Come First Served (FCFS). The single arrival stream 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 an arrival's mean waiting time, or equivalently by Little's law, to minimize the mean total number of jobs in the system. We propose to route arrivals as follows. With probability $\frac{p}{N}$, an arrival is randomly routed to queue j, $j \in J$. We call such arrivals *dedicated* arrivals. With probability $\frac{1-p}{N-1}$, an arrival is routed to the shorter of queues j and $j+1$, $j \in \{1, ..., N-1\}$. We call these *flexible* arrivals.

We will also have reason to study two related policies: join the shortest of all N queues (JSQ) and Mitzenmacher's Two Choices [5]. The latter studied a "supermarket" model, where both inter-arrival times and service times are exponentially distributed with means $(N\lambda)^{-1}$ and 1 respectively, each arrival chooses d servers independently and uniformly at random from the N identical servers and joins the one with the shorter queue (service at each queue is FCFS). We will henceforth denote this policy as JSQ-d/N. It has been shown that the limiting behaviour of this policy leads to exponential improvements in the expected time an arrival spends in the system for any $d \geq 2$ over $d = 1$. Specifically, let $T_d(\lambda)$ denote the expected time a customer spends in the limiting system ($N \to \infty$) for $d \geq 2$, then the asymptotic improvement as the system approaches unity load is

$$\lim_{\lambda \to 1} \frac{T_d(\lambda)}{\log T_1(\lambda)} = \frac{1}{\log d}, \tag{2.1}$$

where $T_1(\lambda)$ is the expected waiting time for an $M/M/1$ queue with arrival rate λ and service rate 1.

In contrast with the techniques in [5], we will study a fixed, finite set of queues and analyze the behaviour as the load on the system approaches unity. We will show that the diffusion scaled total queue length in heavy traffic is the same as that of an $M/G/N$ queue. The limiting process is independent of the probability parameter p, which means even a small proportion of flexible arrivals should yield significant improvement in system performance in heavy traffic.

We end this section with a few comments on our proposed policy. As for JSQ-2/N, our policy is more scalable than JSQ, as a subset of arrivals needs no state information, while the remainder needs information on the state of two servers. What we will see is that our proposed policy should provide another option to JSQ-2/N (we are not suggesting that it will perform better, in fact the discussion in Section 5 suggests that JSQ-2/N should outperform our policy), in the case where a more constrained routing choice for arrivals is attractive.

3 Main Results

We consider a sequence of systems indexed by n, where the n-th system has arrival rate $N\lambda^{(n)}$, service times with mean $\left(\mu^{-1}\right)^{(n)}$ and variance $\left(\beta^2\right)^{(n)}$. Assume that the following conditions hold

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

$$\mu^{(n)} \equiv \mu, \quad \left(\beta^2\right)^{(n)} \equiv \beta^2, \tag{3.2}$$

and

$$\sup_{n\geq 1, j\in J} \mathbf{E}\left[\left(v_{j,1}^{(n)}\right)^{2+\epsilon}\right] < \infty \quad \text{for some } \epsilon > 0. \tag{3.3}$$

In other words, we fix the processing time distribution and let the system go to heavy traffic by increasing the arrival rate. Let $\tilde{\mu} = N\mu$, the heavy traffic condition

$$\lim_{n\to\infty} \sqrt{n}\Big(N\lambda^{(n)} - \tilde{\mu}\Big) = c \tag{3.4}$$

for some finite constant c is assumed to be true.

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:

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

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

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

Before we prove the theorem we comment on its implications. Note that if two unscaled sequences of processes result in the same diffusion limit for their corresponding diffusion scaled processes, it means that 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 $I_1 = \{1, ..., N\}$, $I_2 = \{N+1, ..., 2N-1\}$, $I = I_1 \cup I_2$. The arrivals can be viewed as consisting of $|I|$ types, each type i having arrival rate

$$\lambda_i = \begin{cases} p\lambda, & \text{if } i \in I_1 \text{, the dedicated types,} \\ \frac{N\lambda(1-p)}{N-1}, & \text{if } i \in I_2 \text{ , the flexible types.} \end{cases} \tag{3.5}$$

We construct a corresponding graph $\mathcal{G}$ which has nodes being arrival type i ($i \in I$) and queue j ($j \in J$), and arcs (ij) being the routing activities. To represent $\mathcal{G}$, we use a matrix $\Phi = (\phi_{i,j})_{I \times J}$ with non-negative elements, where $(\mu\phi_{i,j})$ is the average rate at which type i arrivals are routed to queue j. For the base system, we have

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

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

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

has the unique solution

$$\phi_{i,j} = \begin{cases} p, & i \in I_1, j = i, \\ (1-p)\frac{N-j}{N-1}, & i \in I_2, j = i(\text{mod } N), \\ (1-p)\frac{j-1}{N-1}, & i \in I_2, j = i(\text{mod } N)+1. \end{cases} \tag{3.8}$$

When $0 \leq p < 1$, $\phi_{i,j} > 0$ for $i \in I_2$ and $\phi_{i,j} \geq 0$ for $i \in I_1$. 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) and all of the arcs (ij) are basic activities (see Lemma $2(iii)$ in [8]). The routing activities are called *basic* because the associated $\phi_{i,j}$'s taking strictly positive values will keep all of the servers fully utilized, given the arrival rate vector. On the other hand, any non-basic routing activities taking strictly positive values will necessarily cause the capacity of at least one server to be exceeded.

Therefore, from Lemma 3 (iii) in [8] (and also Corollary 5.4 in [10]), the so-called complete resource pooling (CRP) condition holds for the vector $[\lambda_1, ..., \lambda_{2N-1}]$, where λ_i, $i \in I$, is 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

$$b_i = \begin{cases} \frac{cp}{N}, & \text{if } i \in I_1, \\ \frac{c(1-p)}{N-1}, & \text{if } i \in I_2. \end{cases} \tag{3.9}$$

It is easy to see that $\sum_{i\in I} b_i = c$. From (3.5) and (3.7), we have $\lambda = \mu$. Then, using (3.4), we have

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

Let ξ_j quantify the contribution of server j to the workload on the server side, i.e., the weight of the unfinished processing time of all types of arrivals at queue j. Let ν_i quantify the contribution of type i arrivals to the total customer workload. (For the precise definitions of ξ and ν, see [8]. 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 $Q^{(n)}_{B,i,j}(t)$ be the number of type i arrivals at queue j at time t and $\hat{Q}^{(n)}_{B,i,j}(t) = Q^{(n)}_{B,i,j}(nt)/\sqrt{n}$. Applying Theorem $2(i)$ in [8], we have

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

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 Lemma $3(ii)$ in [8], 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

$$\hat{Q}^{(n)}_B(t) \xrightarrow{w} \hat{Q}_B, \quad \text{as } n\to\infty,$$

where

$$\hat{Q}_B = \text{RBM}\left(c, N\lambda\left(1+\mu^2\beta^2\right)\right),$$

which is clearly independent of p.

(ii) In the case of the JSQ policy, the corresponding routing structure is represented by

$$\Phi_{JSQ} = [\phi_{1,1},...,\phi_{1,N}],$$

which has a unique solution to satisfy (3.7), i.e.,

$$\phi_{1,j} = 1, \quad \forall j\in J.$$

Again, from Lemma 3 (iii) in [8], the CRP condition holds for the arrival rate vector which has a single element $\lambda_{JSQ} = N\lambda$ in this case.

Using (3.4), (3.10) is written as

$$\lim_{n\to\infty}\sqrt{n}\Big(N\lambda^{(n)} - \lambda_{JSQ}\Big) = c.$$

By applying Theorem $2(i)$ in [8], we have

$$\hat{Q}^{(n)}_{JSQ}(t) \xrightarrow{w} \text{RBM}\left(c, N\lambda\left(1+\mu^2\beta^2\right)\right), \quad \text{as } n\to\infty. \tag{3.12}$$

We see that the right hand side of (3.12) is $\hat{Q}_B$. 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

$$\hat{Q}_M^{(n)}(t) \xrightarrow{w} \text{RBM}\left(c, N\lambda\left(1+\mu^2\beta^2\right)\right), \quad \text{as } n \to \infty. \tag{3.13}$$

We see that the right hand side of (3.13) is $\hat{Q}_B$. □

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 $j \in J$, 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 Section 4.4 of [8] that when the holding cost functions have the form $C_j(\zeta) = \gamma_j\zeta^2$ and the service rates only depend on the server, i.e., $\mu_{ij} = \mu_j$, the "join the shortest expected waiting time (JSEW)" policy, i.e., route an arrival of type i to queue j, where

$$j \in \arg\min_{j\in J'} \frac{\sum_i Q_{ij}(t)}{\mu_j},$$

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 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 $\tilde{\mu} = \sum_{j\in J}\mu_j$. For ease of exposition, assume that all of the arrivals are flexible. We will assign the arrivals such that an arrival is of type j with probability $(\mu_j + \varepsilon_j)/\tilde{\mu}$, where ε_j will be determined below. Type j arrivals are routed to join the shorter expected waiting time of queues j and $j+1$, $j \in \{1, ..., N-1\}$. That is, the routing probability is proportional to the mean service rate of the neighbour on the left.

Let

$$\bar{\varepsilon} = \min_{j\in\{2,\ldots,N-1\}} \frac{\mu_j}{j-1}. \tag{4.1}$$

The constants $\{\varepsilon_j\}$ are chosen as

$$\varepsilon_j = \begin{cases} \varepsilon \in \left(0, \min\left(\bar{\varepsilon}, \frac{\mu_N+\mu_{N-1}}{N-2}\right)\right), & j \in \{1, ..., N-2\}, \\ \mu_N - (N-2)\varepsilon, & j = N-1. \end{cases} \tag{4.2}$$

Now consider a sequence of systems indexed by n, where the n-th system has arrival rate $N\lambda^{(n)}$, the service times at the j-th server have mean $\left(\mu_j^{-1}\right)^{(n)}$ and variance $\left(\beta_j^2\right)^{(n)}$. We assume that the following conditions hold

$$\begin{aligned} \lim_{n\to\infty} \lambda^{(n)} &= \lambda, && (4.3) \\ \mu_j^{(n)} &\equiv \mu_j, \quad \left(\beta_j^2\right)^{(n)} \equiv \beta_j^2, && (4.4) \end{aligned}$$

and

$$\sup_{n\geq 1, j\in J} \mathbf{E}\left[\left(v_{j,1}^{(n)}\right)^{2+\epsilon}\right] < \infty \quad \text{for some } \epsilon > 0. \tag{4.5}$$

In addition, the heavy traffic condition

$$\lim_{n\to\infty} \sqrt{n}\left(N\lambda^{(n)} - \tilde{\mu}\right) = c \tag{4.6}$$

for some finite constant c is assumed to be true.

Let $Q_B^{(n)}(t)$ be the total queue length at time t. We will also need $Q_M^{(n)}(t)$, the total queue length at time t in an M/G/N^h queue (i.e. a system with a single queue for N heterogeneous servers, service being FCFS). For each of these we again need their diffusion scaled counterparts, i.e.

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

with $\hat{Q}_M^{(n)}(t)$ defined in the same manner. We then 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 ε_j, $j = 1, \ldots, N-1$. That is $\hat{Q}_B^{(n)}(t) \xrightarrow{w} \hat{Q}_B =$ RBM $\left(c, N\lambda + \sum_{j\in J}\mu_j^3\beta_j^2\right)$, as $n \to \infty$.

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

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 set I_2 defined in the base system in Section 3 and see the arrivals as consisting of $|I_2|$ types. Each type i has arrival rate

$$\lambda_i = \mu_j + \varepsilon_j, \qquad i \in I_2, j = i(\text{mod } N), \tag{4.7}$$

where ε_j is given in (4.2).

The routing structure matrix Φ is the same as Φ_2 given in (3.6). Then the linear system

$$\sum_{j\in J}\mu_j\phi_{i,j}=\lambda_i,\quad \forall i\in I_2 \quad \text{and} \quad \sum_{i\in I_2}\phi_{i,j}=1,\quad \forall j\in J, \tag{4.8}$$

has the unique solution

$$\phi_{i,j}=\begin{cases}\frac{1}{\mu_j}\cdot\left(\mu_j-\sum_{k=1}^{j-1}\varepsilon_k\right), & i\in I_2, j=i(\text{mod } N),\\ \frac{1}{\mu_j}\cdot\sum_{k=1}^{j-1}\varepsilon_k, & i\in I_2, j=i(\text{mod } N)+1.\end{cases} \tag{4.9}$$

Since (4.1) and (4.2) hold, then $\phi_{i,j}>0$ for $i\in I_2$. From Lemma 3 (iii) in [8], the CRP condition holds for the arrival rate vector defined in (4.7).

Let

$$b_i=\frac{c(\mu_j+\varepsilon_j)}{\tilde{\mu}},\qquad i\in I_2, j=i(\text{mod } N).$$

It is easy to see that $\sum_{i\in I}b_i=c$. From (4.7) and (4.8), we have $\tilde{\mu}=\sum_{i\in I_2}\lambda_i=N\lambda$. Then, using (4.6), we have

$$\lim_{n\to\infty}\sqrt{n}\left(\lambda_i^{(n)}-\lambda_i\right)=b_i. \tag{4.10}$$

Applying Theorem $2(i)$ in [8], if $Q_{B,j}^{(n)}(t)$ is the queue length at the jth server at time t and its corresponding diffusion scaled version is $\hat{Q}_{B,j}^{(n)}(t)=Q_{B,j}^{(n)}(nt)/\sqrt{n}$,

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

where

$$\theta=\sum_{i\in I_2}\nu_i b_i,\qquad \sigma^2=\sum_{i\in I_2}\nu_i^2\left[\lambda_i+\sum_{j\in J}\mu_j\phi_{i,j}(\mu_j\beta_j)^2\right].$$

Again from Lemma $3(ii)$ in [8], 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 (4.8), implies that

$$\hat{Q}_B^{(n)}(t)\xrightarrow{w}\text{RBM}\left(c,N\lambda+\sum_{j\in J}\mu_j^3\beta_j^2\right),\quad \text{as } n\to\infty. \tag{4.11}$$

(ii) A direct application of Theorem 5 in [7] yields

$$\hat{Q}_M^{(n)}(t)\xrightarrow{w}\text{RBM}\left(c,N\lambda+\sum_{j\in J}\mu_j^3\beta_j^2\right),\quad \text{as } n\to\infty. \tag{4.12}$$

We see that the right hand side of (4.12) is the same as that of (4.11). □

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). Unfortunately, as seen below, the CRP condition does not hold for either of these, 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

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

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

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

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 Lemma $2(iii)$ in [8], 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

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

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

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

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

When $0 \leq p < 1$, $\tilde{\phi}_{i,j} > 0$ for $i \in \tilde{I}_1 \cup \tilde{I}_3$ and $\tilde{\phi}_{i,j} \geq 0$ for $i \in \tilde{I}_2$. From Lemma 3 (iii) in [8], the CRP condition holds for the arrival rate vector defined in (5.5).

Consider a sequence of these modified systems where conditions (3.1)– (3.4) are assumed true. Let

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

and as before, we see $\sum_{i\in I} b_i = c$. Thus, using (3.4), (3.10) is satisfied.

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

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

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 that the modification that has been made would degrade the performance in the sense that if L_R is the mean number in the system for the ring routing structure and $\tilde{L}_R$ is the mean number in the system for the modified ring structure, then $L_R \le \tilde{L}_R$. While we are unable to provide a proof, the fact that we are taking an arrival stream that is making a locally optimal choice and instead making a random assignment suggests that such a relationship should be true. It would be useful to prove this. 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

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

where $I_R = \{1, ..., 2N\}$ 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

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

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

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

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 $\frac{1}{N}$, queue j is selected and with probability $\frac{1}{N-1}$, 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 $I = \{1, ..., \frac{N(N-1)}{2}\}$, so the arrivals consist of $|I|$ *flexible* types, each with rate

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

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.

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). We used the CSIM19 package to perform all of the simulations. All statistics have an accuracy no worse than 5 percent at a 95 percent confidence level.

Using N $M/M/1$ queues (each being 95% loaded) as a reference, we begin with the comparison of 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 model has a single Poisson arrival stream with rate $N\lambda$ and N identical servers, each with exponential rate μ. The performance improvements of the tree structure with different levels of flexibility are shown along with the relative differences with the total mean queue length of N $M/M/1$ 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%

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. So, while the improvements are not as dramatic as those for the diffusion limits (or the $M/M/N$ queues, see Table 2), they are quite significant.

Next, we compare the mean number in system of different models, namely the tree structure, the ring structure, JSQ-$2/N$ and $M/M/N$. All of the four models have a single Poisson arrival stream with rate $N\lambda$ and N identical servers, each with exponential rate μ. The tree and the ring structures are at 100 percent flexibility.

It is noted from Table 2 that the improvements for the tree and ring structures and JSQ-$2/N$ 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, this would suggest that all of these 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

Table 2: Total mean queue lengths vs. routing structures, i.i.d. exponential service times, 95% loaded

Model	$N\times$M/M/1	Tree	Ring	JSQ-2/N	M/M/N
λ	0.95				
μ	1				
	$N = 4$				
Total queue length	76.00	26.87	23.94	23.72	20.74
	$N = 20$				
Total queue length	380.00	115.79	89.74	72.03	33.35
	$N = 100$				
Total queue length	1900.00	469.98	446.54	328.23	104.62

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 ($N \times M/M/1$). Tables 3 and 5 show 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 Tables 4 and 6). When the proportion of flexible arrivals increases, the improvements increase at a speed smaller than that in heavy traffic. For example in Table 3, at 30 percent flexibility, the amount of improvement is less than 60 percent of that with 100 percent flexibility. Tables 4 and 6 again support the observations made in Table 2, while in moderate traffic the difference between the four models becomes smaller.

Table 3: Total mean queue lengths vs. the proportion of flexible arrivals, i.i.d. exponential service times, 70% loaded

Model	$N\times$M/M/1	Tree				
λ	0.7	0.7				
μ	1	1				
$(1-p)$	–	0.03	0.1	0.3	0.5	1
	$N = 4$					
Total queue length	9.33	8.92	8.23	7.19	6.52	5.62
Improvement	0%	4%	12%	23%	30%	40%
	$N = 20$					
Total queue length	46.67	44.67	41.39	35.52	30.95	27.35
Improvement	0%	4%	11%	24%	34%	41%
	$N = 100$					
Total queue length	233.33	223.03	207.49	177.64	159.71	135.16
Improvement	0%	4%	11%	24%	32%	42%

Fourthly, we examine the effects of changing the service time variance. Tables 7 and 8 show the results when we change the service times to have an Erlang-k distribution, with rate one and variance 0.1. Tables

Table 4: Total mean queue lengths vs. routing structures, i.i.d. exponential service times, 70% loaded

Model	$N\times$M/M/1	Tree	Ring	JSQ-2/N	M/M/N
λ	0.7				
μ	1				
	$N = 4$				
Total queue length	9.33	5.62	5.43	5.31	3.80
	$N = 20$				
Total queue length	46.67	27.35	26.87	24.89	14.22
	$N = 100$				
Total queue length	233.33	135.16	133.97	122.66	70.00

Table 5: Total mean queue lengths vs. the proportion of flexible arrivals, i.i.d. exponential service times, 85% loaded

Model	$N\times$M/M/1	Tree				
λ	0.85	0.85				
μ	1	1				
$(1-p)$	–	0.03	0.1	0.3	0.5	1
	$N = 4$					
Total queue length	22.67	20.81	17.92	14.59	12.87	10.64
Improvement	0%	8%	21%	36%	43%	53%
	$N = 20$					
Total queue length	113.33	102.61	89.24	70.05	60.65	49.57
Improvement	0%	9%	21%	38%	46%	56%
	$N = 100$					
Total queue length	566.67	513.95	445.17	344.04	297.52	235.44
Improvement	0%	9%	21%	39%	47%	58%

9 and 10 show the results when the processing times are hyperexponentially distributed, with rate one and service time variance 10.0. 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.

Finally, we look at three models with heterogeneous service time distributions. 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 $\tilde{\lambda}/\sum_{j=1}^{20}\mu_j = 0.95$.

Table 6: Total mean queue lengths vs. routing structures, i.i.d. exponential service times, 85% loaded

Model	$N\times$M/M/1	Tree	Ring	JSQ-2/N	M/M/N
λ	0.85				
μ	1				
	$N = 4$				
Total queue length	22.67	10.64	9.95	9.50	7.31
	$N = 20$				
Total queue length	113.33	49.57	46.50	40.39	19.18
	$N = 100$				
Total queue length	566.67	235.44	234.22	194.90	85.42

Table 7: Total mean queue lengths vs. the proportion of flexible arrivals, i.i.d. Erlang-k service times

Model	$N\times$M/E_k/1	Tree				
λ	0.95	0.95				
(μ^{-1}, β^2)	$(1, 0.1)$	$(1, 0.1)$				
$(1-p)$	–	0.03	0.1	0.3	0.5	1
	$N = 4$					
Total queue length	43.52	35.24	28.05	21.47	18.74	16.08
Improvement	0%	19%	36%	51%	57%	63%
	$N = 20$					
Total queue length	217.60	173.94	133.25	97.36	83.11	69.53
Improvement	0%	20%	39%	55%	62%	68%
	$N = 100$					
Total queue length	1088.00	869.08	661.29	461.25	382.93	298.45
Improvement	0%	20%	39%	58%	65%	73%

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 11, 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.

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.

Table 8: Total mean queue lengths vs. routing structures, i.i.d. Erlang-k service times

Model	$N\times$M/E_k/1	Tree	Ring	JSQ-2/N	M/E_k/N
λ	0.95				
(μ^{-1},β^2)	$(1, 0.1)$				
	$N=4$				
Total queue length	43.52	16.08	14.89	14.67	13.12
	$N=20$				
Total queue length	217.60	69.53	57.90	47.26	26.89
	$N=100$				
Total queue length	1088.00	298.45	291.22	219.57	100.29

Table 9: Total mean queue lengths vs. the proportion of flexible arrivals, i.i.d. hyperexponential service times

Model	$N\times$M/H/1	Tree				
λ	0.95	0.95				
(μ^{-1},β^2)	$(1, 10)$	$(1, 0.1)$				
$(1-p)$	–	0.03	0.1	0.3	0.5	1
	$N=4$					
Total queue length	389.88	315.70	240.08	174.87	147.27	116.75
Improvement	0%	19%	38%	55%	62%	70%
	$N=20$					
Total queue length	1949.40	1557.53	1142.14	753.49	600.79	475.63
Improvement	0%	20%	41%	61%	69%	76%
	$N=100$					
Total queue length	9747.00	7877.65	5712.20	3520.11	2667.10	1726.48
Improvement	0%	19%	41%	64%	73%	82%

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 (2.1), 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 second author is supported by the Natural Sciences and Engineering Research Council of Canada.

Table 10: Total mean queue lengths vs. routing structures, i.i.d. hyperexponential service times

Model	$N\times$M/H/1	Tree	Ring	JSQ-2/N	M/H/N
λ	0.95				
(μ^{-1},β^2)	$(1,10)$				
	$N=4$				
Total queue length	389.88	116.75	101.87	101.62	91.67
	$N=20$				
Total queue length	1949.40	475.63	325.22	217.48	87.87
	$N=100$				
Total queue length	9747.00	1726.48	1627.27	900.23	134.41

Table 11: Expected waiting times, heterogeneous servers, exponential service times

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

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 customers 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.