



Critical Thresholds for Dynamic Routing in Queueing Networks

YIH-CHOUNG TEH*
Department of Statistics, University of Oxford, Oxford, UK

yc@teh.org.uk

AMY R. WARD
Department of Industrial and Systems Engineering, Georgia Institute of Technology, USA

amy@isye.gatech.edu

Received 26 July 2000; Revised 10 May 2002

Abstract. This paper studies dynamic routing in a parallel server queueing network with a single Poisson arrival process and two servers with exponential processing times of different rates. Each customer must be routed at the time of arrival to one of the two queues in the network. We establish that this system operating under a threshold policy can be well approximated by a one-dimensional reflected Brownian motion when the arrival rate to the network is close to the processing capacity of the two servers. As the heavy traffic limit is approached, thresholds which grow at a logarithmic rate are critical in determining the behavior of the limiting system. We provide necessary and sufficient conditions on the growth rate of the threshold for (i) approximation of the network by a reflected Brownian motion (ii) positive recurrence of the limiting Brownian diffusion and (iii) asymptotic optimality of the threshold policy.

Keywords: threshold strategies, heavy traffic, dynamic routing, resource pooling, Brownian network models, queueing networks

1. Introduction and model description

Dynamic routing plays an important role in the control of queueing networks and can lead to dramatic improvements in system performance. There is, however, a trade-off between the cost of keeping state information current and the performance advantage a dynamic routing policy can offer over a static one. For applications in which state information is costly or difficult to obtain, an effective solution is to develop dynamic routing policies that only require partial state information yet still offer sizable performance improvements over any static routing policy.

As a concrete example, consider a manufacturing system with outsourcing. The plant manager knows the current number of orders backlogged at his own facility; however, he generally does not know the backlog at the company where he would outsource jobs. Therefore, he must base his outsourcing decision solely on the state of his own order queue. Such a system could be modeled as a parallel server queueing network with one dedicated arrival stream (representing jobs arriving to the outsourcing facility)

* Supported by EPSRC Award Ref. 96005297 and by St. Anne's College, Oxford.

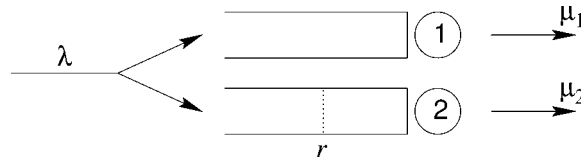


Figure 1. Two queues in parallel with a single discretionary arrival stream and unequal service rates $\mu_1 > \mu_2$. Arrivals are routed to queue 1 unless queue 2 is below the threshold r .

and one discretionary arrival stream (modeling jobs that must be routed either to the outsourcing facility or kept in-house).

In this paper, we study threshold routing policies for the aforementioned parallel-server queueing model with a single discretionary arrival stream, as illustrated in figure 1; for early work see [10,20]. (At the end of section 4, we indicate how our results can be extended to include a dedicated arrival stream to queue 1 as well.) We assume customers arrive according to a Poisson process of rate λ and that processing times are exponential with means μ_1^{-1} and μ_2^{-1} , respectively. All inter-arrival and service times are mutually independent. Our simple threshold policy routes customers to queue 1 unless queue 2 is shorter than an associated threshold r , in which case arrivals are routed to queue 2. We assume $\mu_2 < \lambda$ so that server 2 is unable to process all arrivals without the help of server 1. When $\mu_1 > \mu_2$, our routing policy has the interpretation of routing customers to the faster server unless the slower server is in danger of idling. The major contribution of this paper is the establishment of necessary and sufficient conditions for a threshold routing policy under which (i) the network can be approximated by a one-dimensional reflected Brownian motion (ii) the limiting Brownian diffusion is positive recurrent, and (iii) the threshold routing policy is *asymptotically optimal* and the system exhibits *complete resource pooling* in the heavy traffic limit.

In words, resource pooling implies that the network behaves like a single queue with the same total arrival and service rate as the entire network. For the network models of Harrison [12] (and for our model), this results in a reduction in the dimension of the limiting diffusion process (for example, see [7,14,15,17,19,25,30] and more recently [2,16,32,33,37]). In this case, an appropriate threshold value directs the majority of customers to queue 1 so that queue 1 behaves like a reflected Brownian motion (RBM) while queue 2 is kept as short as possible. Consequently, queue 2 does not affect the limiting Brownian model.

Our work is inspired by previous authors' studies of variants of the model considered here. First suppose that $\mu_1 = \mu_2$ so that processing rates are identical. In this case, when service times have increasing likelihood ratio, the optimal policy is for customers to join the shorter queue (JSQ); see [23,38]. In the case that $\mu_1 \neq \mu_2$, the natural analog of JSQ is *shortest expected delay routing* (SDR), which is studied in [34]. Surprisingly, even for Poisson arrivals and exponential service times, SDR is not always optimal in terms of minimizing the long-run average delay per customer; see [36]. However, Foschini [6] has shown that SDR is asymptotically optimal and results in complete resource pooling in the heavy traffic limit. Laws [25]

later showed that SDR also achieves resource pooling for a more general class of networks.

We offer threshold routing strategies that perform as well as SDR in the heavy traffic limit. To explain the intuition for the threshold routing policy, consider a model with two parallel servers working at the same rate, one discretionary arrival stream, and two dedicated arrival streams – one to each server. Kelly and Laws [19] argued that JSQ performs well in this symmetric model not because queue lengths are held equal, but instead because it ensures that both servers are busy when there is substantial work in the system. They propose a threshold routing strategy for this symmetric model and argue heuristically that it is asymptotically optimal in the heavy traffic limit. In particular, they consider a sequence of queueing systems, indexed by n , in which the total arrival rate approaches the total processing rate as $n \rightarrow \infty$. They conjecture that as long as the threshold is greater than a specified constant times $\log n$, the asymptotic optimality of the threshold routing strategy holds. The threshold routing strategy we propose is the interpretation of their strategy in the asymmetric case.

The importance of thresholds which grow at least as fast as a constant times $\log n$ has become apparent both in the case of discrete review policies [13,14,16,27,28,31] and continuous review policies [2,37]. Of these models, the models of Harrison [14] and Bell and Williams [2] are most similar to the model considered here. Their papers consider a network with two parallel servers and dedicated arrival streams to each of these servers' queues. One server can only process jobs from his queue while the other "super-server" can process jobs from both queues. Both authors consider threshold policies in which the "super-server" processes a job from his own queue if the regular server's queue is below a certain value and processes a job from the regular server's queue otherwise. Harrison establishes conditions under which this policy is asymptotically optimal in a discrete review setting while Bell and Williams extend this to a continuous review setting. Both authors conclude that a sufficient condition for the asymptotic optimality of a threshold policy is that the threshold grows as a constant times $\log n$. Our main contribution is to provide not only sufficient but also necessary conditions on the threshold growth rate under the assumption that jobs must be routed at the time of arrival. Furthermore, we provide the constant c_0 (which remains unspecified in [2]) such that if $c > c_0$ and the threshold grows as $c \log n$, then threshold routing is asymptotically optimal; see equation (1). We also explicitly characterize the effects of having different second order growth terms; see theorem 2, corollary 1, and theorem 3. Finally, our criterion for asymptotic optimality, which assumes identical linear holding costs for each queue, is pathwise and is as given in [14], which is stronger than that of Bell and Williams.

The necessary conditions we provide contrast with recent work of Limic and Williams [26] and illustrate the effect dedicated arrival traffic has on the performance of threshold policies. Limic and Williams study the model of Kelly and Laws [19] described earlier. For this network, any growing threshold is enough to obtain full resource pooling. Kurtz and Turner [24] have a complimentary result showing that even a fixed threshold value for this model is enough to obtain a sort of partial resource pooling. Intuitively, the dedicated arrival stream to queue 2 means that the process's state space is

not bounded by r and its excursions away from the boundary are sufficiently long that any growing threshold results in resource pooling. In contrast, the removal of the additional arrival stream, as considered here, forces the queue-length process at queue 2 to live in the strip $[0, r]$ so that the threshold must grow at a logarithmic rate to keep the process away from the boundary and the server from idling. Even more surprising in light of the aforementioned work, this logarithmic growth rate is necessary not merely for asymptotic optimality but even to ensure a RBM is the correct approximation for our model. (See theorem 2.)

Consider now a sequence of networks (as depicted in figure 1) indexed by n with arrival rate $\lambda(n)$ and service rates $\mu_1(n)$ and $\mu_2(n)$ which tend to the finite limits λ , μ_1 and μ_2 respectively as $n \rightarrow \infty$. Additionally, let the threshold $r(n)$ satisfy

$$r(n) \geq \frac{1 + \varepsilon}{2 \log(\lambda/\mu_2)} \log n \quad (1)$$

for all but finitely many n and for some $\varepsilon > 0$. This condition is analogous to that of Kelly and Laws [19] for the case that $\mu_1 = \mu_2$. Define $Q_k^n(t)$ to be the length of queue k ($k = 1, 2$) at time t and $\tilde{Q}_k^n(t) = Q_k^n(nt)/\sqrt{n}$ to be the scaled queue length processes. Under the above threshold policy, with the heavy traffic condition

$$c(n) = \sqrt{n}(\lambda(n) - \mu_1(n) - \mu_2(n)) \rightarrow c \in (-\infty, 0) \quad (2)$$

as $n \rightarrow \infty$, our finer results imply that, provided $r(n) = o(\sqrt{n})$, condition (1) is sufficient for both the scaled process measuring the total number of customers in the network, $\tilde{Q}_1^n + \tilde{Q}_2^n$, to converge weakly to a one-dimensional RBM and for the threshold routing policy to be asymptotically optimal in the heavy traffic limit. (See theorem 3.)

The remainder of this paper is organized as follows. Section 2 prepares for later results by establishing that the arrival process to queue 1 is a renewal process and calculating the mean and variance of the inter-arrival times. In section 3, we establish that the appropriate diffusion approximation for this network is a one-dimensional RBM. In section 4, we provide necessary and sufficient conditions for the threshold routing policy to be asymptotically optimal in the heavy traffic limit. We conclude in section 5 by discussing how to choose the appropriate threshold value for a given system and evaluating the performance of the threshold policy by simulation.

2. Preliminaries

For the two-queue network under threshold routing with an appropriate choice of the threshold $r(n)$, we aim to prove that the scaled queue-length processes, \tilde{Q}_1^n and \tilde{Q}_2^n , converge weakly to a RBM process and to the zero process respectively. The second task is straightforward (see theorem 1). To establish the first, it is useful to show that arrivals to queue 1 form a renewal process with analytically tractable formulae for the mean and variance of the inter-arrival times. Since the queue-length process at queue 1 can be represented in terms of a reflection mapping, we can then rely on the functional

central limit theorem for renewal processes and the continuous mapping principle to establish the desired weak convergence of \tilde{Q}_1^n .

In this section, we consider the arrival process to queue 1. For clarity of notation, we suppress the dependence on n . Under the threshold strategy, an arrival at time t is routed to queue 1 if and only if $Q_2(t) = r$. Notice the arrival process to queue 1 depends on the state of queue 2 but not on the state of queue 1. $Q_2(\cdot)$ is a finite birth–death process on $\{0, 1, \dots, r - 1, r\}$ with birth rates λ and death rates μ_2 . Suppose an arrival to queue 1 occurs at time $t = 0$ and define $A^1(r)$ to be the time of the next arrival to queue 1 for a given threshold r . It is beneficial to understand the structure of $A^1(r)$.

Define $\{E_i: i \geq 1\}$ to be a sequence of i.i.d. exponential random variables with rate $\lambda + \mu_2$. Since we have assumed an arrival to queue 1 occurs at time 0, under our routing assumptions, it is necessary that $Q_2(0) = r$. The next event of interest will occur after an amount of time equal to E_1 and will be either a system arrival or a departure from queue 2. (We do not consider departures from queue 1 because the state of queue 1 does not affect routing decisions.) This first case occurs with probability $p = \lambda/(\lambda + \mu_2)$ and the second with probability $1 - p$. If this event is a system arrival, that arrival is routed to queue 1 and so $A^1(r) = E_1$. Otherwise, a queue 2 departure results in a transition to $Q_2 = r - 1$. The time taken to first return to $Q_2 = r$ from $Q_2 = r - 1$ has the same distribution as $A^1(r - 1)$, the time between arrivals to queue 1 for the network with threshold $r - 1$. It follows that the third event, which occurs at time $E_1 + A^1(r - 1) + E_2$, will again be a queue 1 arrival with probability p and a queue 2 departure with probability $1 - p$. Let $M(r)$ be the number of excursions to $Q_2 = r - 1$ of duration $E_1 + A^1(r - 1)$ before returning to $Q_2 = r$. Then $A^1(r)$ can be expressed recursively in terms of $A^1(r - 1)$ as follows:

$$A^1(r) = \sum_{i=0}^{M(r)} (E_i + A_i^1(r - 1)) + E_{M(r)+1}, \tag{3}$$

where $\{A_i^1(r - 1): i \geq 1\}$ is a sequence of i.i.d. random variables having the same distribution as $A^1(r - 1)$ and $P(M(r) = m) = (1 - p)^m p$ for $m \in \{0, 1, 2, \dots\}$. In the case that $r = 1$, note that $A^1(0)$ follows an exponential distribution with rate λ since a queue 2 departure results in an empty queue. We exploit this representation of $A^1(r)$ to compute the mean and variance of the queue 1 inter-arrival times.

Proposition 1. For the threshold r , arrivals to queue 1 form a renewal process. The inter-arrival times, $\{A_i^1(r): i \geq 1\}$, have mean and variance

$$E[A_1^1(r)] = \frac{1 - \delta^{r+1}}{\lambda - \mu_2}, \quad \text{var}(A_1^1(r)) = \frac{(\lambda + \mu_2)(1 - \delta^{2r+2})}{(\lambda - \mu_2)^3} - \frac{4(r + 1)\delta^{r+1}}{(\lambda - \mu_2)^2},$$

where $\delta = \mu_2/\lambda$.

Proof. First recognize arrivals to queue 1 form a Markov-modulated Poisson process, with the queue-length process for queue 2 acting as the modulating Markov chain. The

arrival rate to queue 1 is positive in exactly one state of the modulating chain – the state r . For all other states of the modulating chain, the arrival rate to queue 1 is zero. This is a sufficient condition for the arrival process to queue 1 to be renewal; see [5,21], or [29].

The mean and variance of $A_1^1(r)$ can be calculated explicitly from the recurrence structure shown in (3). Define $m_r = E[A_1^1(r)]$ and $\delta = \mu_2/\lambda$. Then, the recursion

$$m_r = E \left[\sum_{i=0}^{M(r)} (E_i + A_i^1(r-1)) + E_{M+1} \right] = \frac{\mu_2}{\lambda} \left(\frac{1}{\lambda + \mu_2} + m_{r-1} \right) + \frac{1}{\lambda + \mu_2},$$

is valid, which implies

$$m_r + \delta m_{r-1} = \frac{1}{\lambda}.$$

The solution to this first-order inhomogeneous difference equation having initial condition $m_0 = E[A_1^1(0)] = 1/\lambda$ is $m_r = (1 - \delta^{r+1})/(\lambda - \mu_2)$, as the proposition states.

Similarly, define $v_r = \text{var}(A_1^1(r))$. Then, using the representation for $A_1^1(r)$ given in (3) and the conditional variance formula $\text{var}(A_1^1(r)) = E[\text{var}(A_1^1(r) | M(r))] + \text{var}(E[A_1^1(r) | M(r)])$, we find:

$$v_r = \frac{\mu_2}{\lambda} \left(\frac{1}{(\lambda + \mu_2)^2} + v_{r-1} \right) + \frac{1}{(\lambda + \mu_2)^2} + \frac{\mu_2(\lambda + \mu_2)}{\lambda^2} \left(\frac{1}{\lambda + \mu_2} + m_{r-1} \right)^2.$$

Substituting $m_{r-1} = (1 - \delta^r)/(\lambda - \mu_2)$ results in the first-order inhomogeneous difference equation

$$v_r - \delta v_{r-1} = \frac{(1 + \delta)(1 + \delta^{2r+1}) - 4\delta^{r+1}}{\lambda^2(1 - \delta)^2}, \tag{4}$$

which, together with the initial condition $v_0 = 1/\lambda^2$, has the solution stated in the proposition. \square

Remark 1. To determine the mean of the inter-arrival times to queue 1, we could have also exploited known results for Markov-modulated Poisson processes. Specifically, let \widehat{Q} be the infinitesimal generator for the birth–death process Q_2 (the modulating chain) and let $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$ give the arrival rates to queue 1 dependent upon the state of queue 2. Here, $\lambda_i = 0$ if $0 \leq i < r$ and $\lambda_i = \lambda$ if $i = r$. Then, from the formula given in section 2.4 of Fischer and Meier-Hellstern [5], we have

$$E[A_1^1(r)] = \lambda((\Lambda - \widehat{Q})^{-2})_{rr}. \tag{5}$$

Although formula (5) involves matrix inversion, it is straightforward since we need only know the values in the r th row and r th column of the matrix $(\Lambda - \widehat{Q})^{-1}$. As \widehat{Q} has a tri-diagonal structure, it is not hard to compute these values in terms of r . They are: $((\Lambda - \widehat{Q})^{-1})_{jr} = \lambda^{-1}$ and $((\Lambda - \widehat{Q})^{-1})_{rj} = \lambda^{-1}\delta^{r-j}$. Substituting into (5) gives the expression for $E[A_1^1(r)]$ stated in proposition 1. However, the computation involved in the formula for $E[A_1^1(r)^2] = \lambda((\Lambda - \widehat{Q})^{-3})_{rr}$ (which can again be found in [5]) is not

straightforward because we must now invert the entire matrix and find formulae for each element in the matrix in terms of r .

3. Diffusion approximations for the system

Assuming the threshold r does not grow too fast, the scaled queue-length process \tilde{Q}_2^n vanishes in the limit. Consequently, the system can be approximated by the one-dimensional limit process resulting from the weak convergence of \tilde{Q}_1^n . We first establish the weak convergence of \tilde{Q}_2^n to the zero process and then establish the weak convergence of \tilde{Q}_1^n to a RBM process. Here and throughout, weak convergence, which we denote by \Rightarrow , is in the topology of weak convergence on $D[0, \infty)$; see, for example, [3] for a discussion of this convergence concept.

Theorem 1. Assume $Q_2^n(0) \leq r(n)$ for all n . Then, if $r(n) = o(\sqrt{n})$, $\tilde{Q}_2^n \Rightarrow 0$ in $D[0, \infty)$ as $n \rightarrow \infty$.

Proof. Under the threshold routing strategy, an arrival to the n th system occurring at time t is routed to queue 1 if $Q_2(t) = r(n)$. Therefore, with probability 1,

$$\sqrt{n} \sup_{0 \leq t \leq T} \tilde{Q}_2^n(t) = \sup_{0 \leq t \leq nT} Q_2^n(t) \leq r(n)$$

for any fixed $T \in [0, \infty)$. The stated conclusion follows. □

The simplicity of the proof of theorem 1 stems from the fact that the queue length at queue 2 never exceeds the threshold r . Although a similar result holds for the model considered by Bell and Williams [2] operating under a threshold policy (see theorem 5.2 of their paper), the proof requires large deviations estimates for renewal processes. In their model, there is a dedicated arrival stream to both a “super-server” that can process jobs from either queue and a regular server. Thus, a threshold routing policy does not imply that one queue length is always held below its threshold value and so a more sophisticated methodology is required.

To establish the desired weak convergence for queue 1, there must exist $c_1 \in (-\infty, \infty)$ such that

$$c_1(n) = \sqrt{n}(\lambda_1(n) - \mu_1(n)) \rightarrow c_1 \tag{6}$$

where $\lambda_1(n) = (E[A_1^1(r(n))])^{-1}$ is the mean rate of arrivals to queue 1. Appropriate assumptions on the limiting behavior of the threshold $r(n)$ guarantee that c_1 exists and is finite. The following lemma establishes the appropriate growth conditions for desired values of c_1 and is useful both in establishing the weak convergence of \tilde{Q}_1^n and in characterizing the behavior of the limiting RBM process. The growth condition on the threshold required for asymptotic optimality (theorem 3) is stronger than the growth condition required for the combined scaled queue-length processes to weakly converge

to a positive recurrent RBM (corollary 1). This in turn is stronger than the growth condition required for the combined queue-length process to weakly converge to a possibly transient RBM (theorem 2). The three necessary and sufficient conditions on the growth of the threshold $r(n)$ share the dominant term

$$\frac{1}{2 \log \delta(n)^{-1}} \log n \quad (7)$$

but differ in the second order term. To understand the effect of the second-order term, realize that $c_1 = c$ if and only if there exists a sequence $b(n) \in \mathfrak{R}^+$ such that $b(n) \rightarrow 0$ and

$$r(n) \geq \frac{1}{2 \log \delta(n)^{-1}} \log n + \frac{1}{\log \delta(n)^{-1}} \log b(n)^{-1} \quad (8)$$

for all but finitely many n . (This statement is equivalent to part (c) of lemma 1.) Under the necessary and sufficient conditions for asymptotic optimality, the second order term diverges to infinity whereas under those for either convergence to RBM or convergence to a positive recurrent RBM, the second order term converges to a finite constant. Provided the sequence $b(n)$ in (8) is such that $\log b(n)^{-1} = o(\log n)$ (for example, $b(n) = 1/\log n$) then term (7) dominates and, as stated in the introduction, condition (1) is a sufficient condition for the threshold routing policy to be asymptotically optimal.

Recall from section 2 that $\delta(n) = \mu_2(n)/\lambda(n)$.

Lemma 1. Assume the heavy traffic condition (2) is satisfied and also assume $\lim_{n \rightarrow \infty} \sqrt{n} \delta(n)^{r(n)}$ either exists or is infinite.

(a) The constant c_1 is finite if and only if there exists $b \in (0, \infty)$ such that

$$r(n) \geq \frac{1}{2 \log \delta(n)^{-1}} \log n + \frac{1}{\log \delta(n)^{-1}} \log b^{-1} \quad (9)$$

for all but finitely many n .

(b) The constant c_1 is strictly less than zero if and only if there exists $b \in (0, -\lambda c/\mu_1 \mu_2)$ such that (9) is satisfied for all but finitely many n .

(c) The constant c_1 equals c if and only if for all $b \in (0, \infty)$, (9) is satisfied for all but finitely many n , where c is as given in (2).

Proof. By proposition 1, $c_1(n)$ can be expressed as follows:

$$c_1(n) = \sqrt{n} \left(\frac{\lambda(n) - \mu_2(n)}{1 - \delta(n)^{r(n)+1}} - \mu_1(n) \right). \quad (10)$$

Algebraic manipulations of (10) together with the substitution $c(n) = \sqrt{n}(\lambda(n) - \mu_1(n) - \mu_2(n))$ then yield:

$$c_1(n) = \frac{c(n)}{1 - \delta(n)^{r(n)+1}} + \frac{\mu_1(n)\delta(n)}{1 - \delta(n)^{r(n)+1}} \sqrt{n} \delta(n)^{r(n)}. \quad (11)$$

Assuming any of conditions (a)–(c) hold, $r(n) \rightarrow \infty$ as $n \rightarrow \infty$. Therefore, since $\delta(n) < 1 - \varepsilon$ for some $\varepsilon > 0$ for all but finitely many n , the first term in (11) converges to c and the second term converges to $(\mu_1\mu_2/\lambda)x \geq 0$ as $n \rightarrow \infty$ where $x = \lim_{n \rightarrow \infty} \sqrt{n}\delta(n)^{r(n)}$ so that c_1 is bounded below by c . Equation (9) is equivalent to

$$\sqrt{n}\delta(n)^{r(n)} \leq b \tag{12}$$

for all but finitely many n . Part (a) then follows since x is finite if and only if there exists $b \in (0, \infty)$ such that (12) holds. Part (b) follows since $x \in [0, -\lambda c/\mu_1\mu_2]$ if and only if $b \in (0, -\lambda c/\mu_1\mu_2)$. Part (c) follows since $x = 0$ if and only if, for all $b \in (0, \infty)$, (12) is satisfied for all but finitely many n . \square

Let $E^n(t)$ be the number of customers that arrive to queue 1 in the n th system during the time window $[0, t]$ when customer inter-arrival times are i.i.d. with rate $\lambda_1(n) = E[A^1(r(n))]^{-1}$ and variance $var(A^1(r(n)))$ as given in proposition 1. Let N be a Poisson process with rate 1 and let $B_1^n(t) = \int_0^t \mathbb{1}_{\{Q_1^n(s) > 0\}} ds$ be the cumulative busy time of server 1 during $[0, t]$ so that $N(\mu_1(n)B_1^n(t))$ represents the number of departures from queue 1 during $[0, t]$. The pathwise construction of the queue-length process at queue 1 is:

$$Q_1^n(t) = Q_1^n(0) + E^n(t) - N(\mu_1(n)B_1^n(t)).$$

Let $I_1^n(t) = \int_0^t \mathbb{1}_{\{Q_1^n(s) = 0\}} ds = t - B_1^n(t)$ be the idle time of server 1 during $[0, t]$. Define

$$\begin{aligned} X^n(t) &= Q_1^n(0) + E^n(t) - \lambda_1(n)t \\ &\quad - (N(\mu_1(n)B_1^n(t)) - \mu_1(n)B_1^n(t)) + t(\lambda_1(n) - \mu_1(n)), \end{aligned} \tag{13}$$

and

$$Y^n(t) = \mu_1(n)I_1^n(t)$$

so that $Q_1^n(t) = X^n(t) + Y^n(t)$. Since $Q_1^n(t) \geq 0$, $Y^n(0) = 0$, $dY^n(t) \geq 0$, and $Q_1^n(t) dY^n(t) = 0$ for all $t \geq 0$, given X^n for which $X^n(0) \geq 0$, we can express Q_1^n and Y^n in terms of X^n as follows:

$$Q_1^n = \phi(X^n) = X^n + \Psi(X^n) \tag{14}$$

and

$$Y^n = \Psi(X^n) = \sup_{0 \leq s \leq t} [-X^n(s)]^+ \tag{15}$$

where (ϕ, Ψ) is the one-sided reflection mapping or one-sided regulator; see theorem 6.1 of [4]. This representation of Q_1^n is the key to establishing the main result of this section, a theorem providing necessary and sufficient conditions for the weak convergence of the scaled queue-length and idleness processes to a RBM and a multiple of the local time at the origin of that RBM.

To set the stage for this theorem, let $X^R = (X^R(t): t \geq 0)$ be a RBM with drift c_1 as specified in (6) and variance $\sigma^2 = \lambda + \mu_1 + \mu_2$ so that X^R is the solution to the stochastic differential equation

$$dX^R(t) = c_1 dt + \sigma dB(t) + dL(t) \tag{16}$$

subject to $X^R(0) = x \geq 0$. Here, $L = (L(t): t \geq 0)$ is the local time process and is the minimal nondecreasing process which makes $X^R(t) \geq 0$ for $t \geq 0$ and increases only when X^R is zero. Finally, let $\tilde{I}_1^n(t) = (1/\sqrt{n})I_1^n(nt)$.

Theorem 2 (RBM convergence). Suppose the conditions for lemma 1 are satisfied and $(1/\sqrt{n})Q_1^n(0) \Rightarrow X^R(0) = x$, where X^R is a RBM described by (16). Let $I(t) = \mu_1 L(t)$ for $t \geq 0$. Then,

$$(\tilde{Q}_1^n, \tilde{I}_1^n) \Rightarrow (X^R, I) \tag{17}$$

as $n \rightarrow \infty$ in $D[0, \infty)$ if and only if there exists $b \in (0, \infty)$ such that the threshold $r(n)$ satisfies (9) for all but finitely many n .

Proof. Assume there exists $b \in (0, \infty)$ such that (9) is satisfied for all but finitely many n and define the scaled and centered processes

$$\begin{aligned} \tilde{E}^n(t) &= \frac{1}{\sqrt{n}}E^n(nt) - \sqrt{n}\lambda_1(n)t, \\ \tilde{N}^n(t) &= \frac{1}{\sqrt{n}}N(n\mu_1(n)t) - \sqrt{n}\mu_1(n)t. \end{aligned}$$

Define the scaled process $\tilde{X}^n(t) = (1/\sqrt{n})X^n(nt)$ for $t \geq 0$, where X^n is as defined in equation (13). Since $r(n)$ satisfies (12) for all but finitely many n , by proposition 1, $\lambda_1(n) \rightarrow \lambda - \mu_2$ and $\text{var}(A_1^1(r(n))) \rightarrow (\lambda + \mu_2)/(\lambda - \mu_2)^3$ as $n \rightarrow \infty$. Therefore, the functional central limit theorem for renewal processes guarantees

$$\tilde{E}^n \Rightarrow \text{BM}_0(0, \lambda + \mu_2) \quad \text{and} \quad \tilde{N}^n \Rightarrow \text{BM}_0(0, \mu_1).$$

where $\text{BM}_x(\mu, \sigma^2)$ denotes a Brownian motion with drift μ and variance σ^2 starting from initial position x . By part (a) of lemma 1,

$$\sqrt{n}(\lambda_1(n) - \mu_1(n)) \rightarrow c_1 < \infty,$$

which, by employing the random time-change theorem, (see chapter 17 of [3]), implies

$$\tilde{X}^n \Rightarrow \text{BM}_x(c_1, \lambda + \mu_1 + \mu_2)$$

once we establish $B_1^n(nt)/n \rightarrow t$ uniformly on compact sets (u.o.c.).

By the functional strong law of large numbers,

$$\frac{E^n(nt)}{n} - \lambda(n)t \rightarrow 0$$

u.o.c. and

$$\frac{N(\mu_1(n)B_1^n(nt))}{n} - \mu_1(n)\frac{B_1^n(nt)}{n} \rightarrow 0$$

u.o.c. since $0 \leq B_1^n(nt)/n \leq t$ for all $t \in \mathfrak{R}$. Since $\lambda_1(n) - \mu_1(n) \rightarrow \lambda - \mu_1 - \mu_2 = 0$, $X^n(nt)/n \rightarrow 0$ u.o.c.

The reflection mapping is known to be continuous in $D[0, \infty)$ (see theorem 6.1 in [4]), and therefore, the weak convergence of $(\tilde{Q}_1^n, \tilde{I}_1^n)$ follows by the continuous mapping theorem.

In the case that there does not exist a finite b such that (9) holds for all but finitely many n , $c_1 = \infty$, and so the weak convergence in (17) cannot hold. \square

Corollary 1 (Stability of the limiting system). Again suppose the conditions for lemma 1 are satisfied and $(1/\sqrt{n})Q_1^n(0) \Rightarrow X^R(0) = x$. Then, the scaled queue-length process \tilde{Q}_1^n weakly converges to a positive recurrent RBM if and only if there exists $b \in (0, -\lambda c/\mu_1\mu_2)$ such that the threshold $r(n)$ satisfies (9) for all but finitely many n .

The proof of corollary 1 follows from part (b) of lemma 1 since a RBM process is positive recurrent if and only if it has negative drift.

4. Asymptotic optimality

A “good” routing policy is one which in some sense minimizes the combined number of customers in queues 1 and 2. In particular, the process tracking the total number of customers in the network, $Q_1(t) + Q_2(t)$, should resemble the queue length process of a system that pools the processing power of the two servers. For our model, the appropriate comparison system is a M/M/1 queue operating under a non-idling policy with the same Poisson arrival stream of rate λ having a single server with exponential service times of rate $\mu_1 + \mu_2$. In contrast with our setting, the server in the pooled system (the M/M/1 system) can never idle when there is work in the system. Notice in our network that if several jobs are held in queue 1 but queue 2 is empty, server 2 idles, and vice versa. Therefore, assuming our network and the pooled system experience identical inter-arrival and service time sequences, the queue-length process in the pooled system, which we denote by $Q_P(t)$, lower bounds the total number of customers in our network; i.e.,

$$Q_P(t) \leq Q_1(t) + Q_2(t) \tag{18}$$

under any routing policy. Let $\tilde{Q}_P^n(t) = (1/\sqrt{n})Q_P(nt)$ be the scaled queue-length process in the pooled system. Our aim in this section is to specify a routing policy under which the distributions of \tilde{Q}_P and $\tilde{Q}_1 + \tilde{Q}_2$ are in some sense “close,” thereby providing support for the argument that we have found a good routing policy.

Assume holding costs are continuously incurred at a rate of h dollars per hour for each job that remains in the network, regardless of the buffer where it resides. Define

$$\xi^n(t) = \int_0^t h(Q_1^n(s) + Q_2^n(s)) ds$$

to be the cumulative cost process in the n th system and

$$\tilde{\xi}^n(t) = n^{-3/2}\xi^n(nt)$$

to be the scaled cumulative cost process. Also define

$$\begin{aligned}\xi_p^n(t) &= \int_0^t hQ_p^n(s) ds, \\ \tilde{\xi}_p^n(t) &= n^{-3/2}\xi_p^n(nt)\end{aligned}$$

to be the cumulative and scaled cumulative cost processes for the n th pooled system. Applying theorem 3.5 of [30] (first proved in [18]), we have

$$\tilde{Q}_P \Rightarrow X^R,$$

where X^R is a RBM with drift c and variance $\lambda + \mu_1 + \mu_2$. Therefore, by the continuous mapping theorem,

$$\tilde{\xi}_P^n \Rightarrow X^{R,*},$$

where $X^{R,*}(t) = \int_0^t hX^R(s) ds$. Under any routing policy, (18) guarantees

$$\limsup_{n \rightarrow \infty} P(\tilde{\xi}^n(t) \leq x) \leq P(X^{R,*}(t) \leq x)$$

for each fixed $t > 0$ and $x > 0$. We call a routing policy asymptotically optimal if

$$\lim_{n \rightarrow \infty} P(\tilde{\xi}^n(t) \leq x) = P(X^{R,*}(t) \leq x) \quad (19)$$

for each fixed $t > 0$ and $x > 0$. In words, we require an asymptotically optimal policy to minimize the scaled cumulative cost incurred up to any time t with probability 1 in the heavy traffic limit. The following theorem states necessary and sufficient conditions on the growth rate of the threshold $r(n)$ for the threshold routing policy to be asymptotically optimal.

Theorem 3 (Asymptotic optimality). Assume the conditions for lemma 1 are satisfied. Also suppose $(1/\sqrt{n})Q_1^n(0) \Rightarrow X^R(0)$ and $r(n) = o(\sqrt{n})$. Then, the threshold routing policy is asymptotically optimal if and only if, for all $b \in (0, \infty)$, the threshold $r(n)$ satisfies (9) for all but finitely many n .

Proof. Theorem 1 guarantees $\tilde{Q}_2^n \Rightarrow 0$ in $D[0, \infty)$ as $n \rightarrow \infty$. Part (c) of lemma 1 together with theorem 2 establishes $\tilde{Q}_1^n \Rightarrow X^R$ if and only if for all $b \in (0, \infty)$ the threshold $r(n)$ satisfies (9) for all but finitely many n . Therefore, by the continuous

mapping theorem, $\tilde{Q}_1^n + \tilde{Q}_2^n \Rightarrow X^R$ if and only if the threshold $r(n)$ satisfies the aforementioned conditions. (It is well known that addition preserves convergence when both limit processes have continuous paths; see [3] or [35].) When $\tilde{Q}_1^n + \tilde{Q}_2^n \Rightarrow X^R$ as $n \rightarrow \infty$, $\tilde{\xi}^n \Rightarrow X^{R,*}$ so that the conditions for asymptotic optimality given in (19) are satisfied. Otherwise, either the threshold $r(n)$ satisfies part (a) of lemma 1 with $c_1 > c$ or it does not. If it does, by theorem 2, $\tilde{Q}_1^n + \tilde{Q}_2^n \Rightarrow Y^R$, where Y^R is a RBM with drift c_1 so that (19) does not hold. If it does not, then $\tilde{Q}_1^n + \tilde{Q}_2^n$ does not converge to a limiting diffusion since $c_1 = \infty$ and, again, (19) is not satisfied. \square

The key feature of an asymptotically optimal routing policy is that it ensures both servers are busy when there is substantial work in the system. Theorem 3 supports the arguments of Kelly and Laws [19] that the stronger condition of keeping expected delays for each queue equal is not necessary for asymptotic optimality – a simple threshold policy suffices. The next theorem establishes that when the network operates under an asymptotically optimal threshold routing policy, the scaled idleness process at server 2 vanishes in the heavy traffic limit. In other words, server 2 is almost always busy when there is work in the system.

Theorem 4. Assume $Q_2^n(0)$ is distributed according to its stationary distribution. If the threshold $r(n)$ satisfies the condition for asymptotic optimality given in theorem 3, then

$$\tilde{I}_2^n \Rightarrow 0$$

in $D[0, \infty)$ as $n \rightarrow \infty$.

Proof. Since Q_2^n is a birth–death process, it has a known stationary distribution and so we can calculate

$$\pi_0^n = P(Q_2^n(\infty) = 0) = \frac{\delta_n^{r(n)}(1 - \mu_2(n)/\lambda(n))}{1 - \delta(n)^{r(n)+1}}$$

where the random variable $Q_2^n(\infty) = \lim_{t \rightarrow \infty} Q_2^n(t)$ has the stationary distribution for the queue length in the n th system. Since I_2^n is nondecreasing in t , we have

$$\sup_{0 \leq s \leq T} \tilde{I}_2^n(s) = \tilde{I}_2^n(T) \quad \text{a.s.}$$

so that for any $\varepsilon > 0$,

$$\begin{aligned} P\left(\sup_{0 \leq s \leq T} \tilde{I}_2^n(s) > \varepsilon\right) &\leq \varepsilon^{-1} E[\tilde{I}_2^n(T)] \\ &= \frac{\pi_0^n n T}{\varepsilon \sqrt{n}} \\ &= \varepsilon^{-1} (\sqrt{n} \delta(n)^{r(n)}) \frac{1 - \mu_2(n)/\lambda(n)}{1 - \delta(n)^{r(n)+1}} T \\ &\rightarrow 0 \end{aligned}$$

by part (c) of lemma 1. \square

Together theorems 1 and 4 show that under the threshold routing strategy both the scaled queue length and idleness processes for queue 2 vanish in the heavy traffic limit. This seems paradoxical: it is impossible for both the queue to be always empty and the server always busy. However, the heavy traffic analysis indicates that when the original system is heavily loaded, the threshold policy is capable of finely balancing the rate of arrivals to queue 2 such that the queue length is kept small compared with the faster queue 1 and its server is rarely idle. Most importantly, examining theorems 2 and 3, we see that neither server is idle in the heavy traffic limit except when there is no work in the system (since server 1 only idles when its queue is empty). Although in practice this is unachievable (since it is only true in the limit), the asymptotic optimality result implies that any good control policy should keep both servers busy unless there are very few customers in the system. The threshold routing strategy is arguably the simplest control policy that achieves this.

Our results generalize to a model that includes a dedicated arrival stream to queue 1 with general inter-arrival times of mean $\alpha(n)^{-1} \rightarrow \alpha^{-1}$ and variance $a_1(n) \rightarrow a_1$. In this case, the heavy traffic condition (2) must be modified so that

$$c(n) = \sqrt{n}(\lambda(n) + \alpha(n) - \mu_1(n) - \mu_2(n)) \rightarrow c' \in (-\infty, 0).$$

Then, the superposition of this dedicated arrival stream and the arrival stream of jobs routed from queue 1 under the threshold policy (specified in proposition 1) forms the queue 1 arrival stream. Similar arguments to those in theorems 1–3 establish

$$\tilde{Q}_1^n + \tilde{Q}_2^n \Rightarrow \text{RBM}(c', \lambda + \alpha^3 a_1 + \mu_1 + \mu_2)$$

as $n \rightarrow \infty$. The appropriate comparison or pooled system (one that lower bounds the combined queue length process) is now a queue whose input process is two renewal processes superimposed and again has a single server with exponential service times of rate $\mu_1 + \mu_2$. As before, let \tilde{Q}_p^n denote the queue length process in the pooled system. From theorem 3.5 of [30],

$$\tilde{Q}_p^n \Rightarrow \text{RBM}(c', \lambda + \alpha^3 a_1 + \mu_1 + \mu_2)$$

as $n \rightarrow \infty$, showing a threshold routing strategy satisfying the conditions for part (c) of lemma 1 is still asymptotically optimal.

5. Implementation and simulation

In the previous two sections we derived theoretical properties of the limiting system in heavy traffic, under the threshold routing strategy. We now turn our attention to practical implementation. An attraction of the threshold routing strategy is that it is straightforward to implement: if queue 2 is shorter than the given threshold, arrivals join queue 2, and otherwise they are routed to queue 1. It remains to determine the optimal threshold which minimizes average system population (or equivalently, typical customer delay) for any particular arrival rate λ and service rates μ_1, μ_2 . We consider two simple heuristic

Table 1
 Threshold value r for service rates $\mu_1 = 1.6, \mu_2 = 0.4$, and arrival rate $\lambda = 2\rho$ calculated (i) empirically by simulation (ii) from equation (20) and (iii) minimizing $\bar{q}_1 + \bar{q}_2$ in (21), (22) over r .

ρ	0.50	0.80	0.90	0.95	0.975	0.99
Simulation	1	1	2	3	4	5
Simplest heuristic	2	2	3	3	4	5
Minimization	1	1	2	3	4	5

approximations as rough and ready estimates of the optimal threshold and find that they compare favorably with the best choice of threshold determined empirically by simulation.

With no loss of generality, we consider the model with a fixed overall service rate of $\mu_1 + \mu_2 = 2$ ($\mu_1 > \mu_2$) and arrival rate $\lambda = 2\rho$, where $\rho = \lambda/(\mu_1 + \mu_2)$ is the traffic intensity. The optimal choice of the threshold r then depends on two parameters: the traffic intensity ρ and the service rate μ_2 . The best values of r were determined from simulation for values of ρ ranging from 0.5 to 0.99 and for $\mu_2 = 0.4$; see table 1. (We note that for different values of μ_2 results are similar; i.e., the performance of the methods we propose for estimating r when $\mu_2 = 0.4$ is indicative of the performance of these methods for other values of μ_2 .)

To derive approximations for the queueing system from our earlier heavy traffic limit theorems, we must first identify the appropriate values for the index n . Following Reiman [30], one approach is to restate the limit theorem in terms of $\rho \rightarrow 1$, which we can achieve by setting $n = (1 - \rho)^{-2}$. We then obtain $c = -(\mu_1 + \mu_2)$ as the drift of the approximating RBM process. The simplest approach is to adopt (1) and set

$$r = \left\lceil \frac{1 + \varepsilon}{2 \log(\lambda/\mu_2)} \log(1 - \rho)^{-2} \right\rceil \tag{20}$$

for $\lambda = (\mu_1 + \mu_2)\rho$ and ε small. To determine an appropriate value for ε , we simulated the system across a range of values μ_1, μ_2 , with $\mu_1 + \mu_2 = 2$, and selected a value that performed well across this spectrum. As shown in table 1, for $\varepsilon = 0.5$, (20) slightly overestimates the optimal value of r for $\rho < 0.9$. As explained later in this section (see figure 2), it is preferable for the threshold to be above its optimum value than below it.

Another approach is to calculate the stationary distribution of queue 2, approximate the stationary distribution of queue 1, and then find the threshold value r that minimizes the total number of customers in the system. Recall that queue 2 is a finite birth–death process so that it is straightforward to calculate the mean of its stationary distribution to be:

$$\bar{q}_2 = \frac{r - (r + 1)\delta + \delta^{r+1}}{(1 - \delta^{r+1})(1 - \delta)}. \tag{21}$$

Queue 1 is a G/M/1 queue. Although theorems for obtaining the mean of this stationary distribution are well known (see, for example, [1,9] or [22]) the calculation is cum-

bersome. Therefore, we use Kingman's approximation [20] for the steady-state queue-length of a G/G/1 queue to approximate the mean number of customers in queue 1 as:

$$\bar{q}_1 = \frac{\lambda_1^2(\text{var}(A_1^1(r)) + \mu_1^{-2})}{2(1 - \lambda_1/\mu_1)} + \frac{\lambda_1}{\mu_1}. \quad (22)$$

We take $\bar{q}_1 + \bar{q}_2$ given by (21), (22) as our approximation for the total number of customers in the network. Since $\bar{q}_1 + \bar{q}_2$ is a convex function in r , we can readily find the minimizing threshold r . As evidenced in table 1, the results of this procedure agree with the results obtained via simulation.

From figure 2, we observe that the choice of threshold r has a similar effect on mean system population as do trunk reservation parameters in loss networks (see [8]): mean system population increases rapidly for values of r below the optimum, but rather more slowly for values above it. This behavior may be understood intuitively as follows. The stationary mean queue-length at queue 1 can be approximated from the RBM limit process (which has a known stationary distribution; see [11]) by reversing the normalization to obtain:

$$\bar{q}_1 \approx \sqrt{n} \frac{\lambda + \mu_1 + \mu_2}{2(\mu_1 + \mu_2)},$$

from which we expect the unscaled length of queue 1 to grow like \sqrt{n} . Since $Q_2^n(\cdot) \leq r(n)$ and the arrival rate λ exceeds the service rate μ_2 , we expect the length of queue 2 to follow the behavior of $r(n)$, which is of order $\log n$. If the choice of threshold r is too small when compared with the optimal value, server 2 spends too much time being idle and the effective rate of arrivals to queue 1 is too great, resulting in an increase in

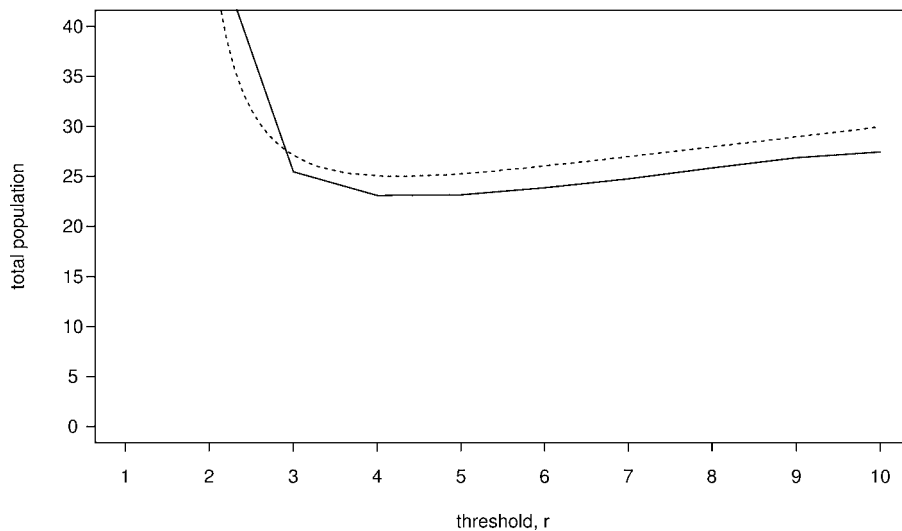


Figure 2. Mean population for system determined from simulation (solid line) and from minimizing $\bar{q}_1 + \bar{q}_2$ over r (dashed line) with traffic intensity $\rho = 0.95$ and service rates $\mu_1 = 1.3$, $\mu_2 = 0.7$.

the length of queue 1 which is of order \sqrt{n} . If r is too large, then idleness at server 2 is kept low, but queue 2 is longer than necessary, resulting in a penalty of order $\log n$. From this argument we see that system performance suffers comparably less (order $\log n$ compared to order \sqrt{n}) if the threshold is above the optimum rather than below it.

We conclude this section by comparing the threshold routing policy (using the optimal threshold r) with shortest expected delay routing (SDR) and weighted random routing (WRR). The WRR policy we consider, which minimizes mean delay among static policies, sends arrivals to queue 1 with probability $\mu_1/2$ and to queue 2 with probability $\mu_2/2$. One expects SDR to outperform our threshold policy and our threshold policy to outperform WRR, which our simulation results verify. The reason for making this comparison is to understand the advantage gained by keeping track of additional state information (both queue lengths for SDR, 1 queue length for threshold routing, nothing for WRR). For applications where it is costly or difficult to observe queue lengths, this analysis illuminates the trade-off between the cost of state information and network performance as measured by the steady-state mean number of customers in the network.

Since our simulation results indicate network performance is much more dependent on ρ than μ_2 , we only display results for one fixed value of μ_2 . For $\mu_1 = 1.6$ and $\mu_2 = 0.4$, we simulated the network under each of the three policies with ρ rang-

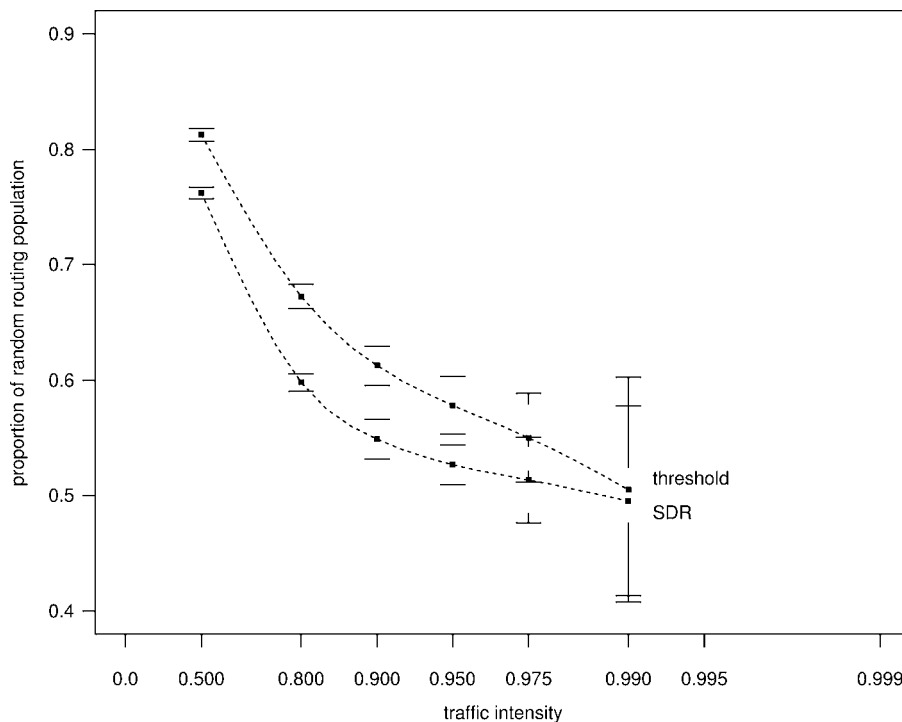


Figure 3. Comparison of simulation results for SDR and threshold routing policies for $\mu_1 = 1.6$ and $\mu_2 = 0.4$.

ing from 0.5 to 0.99. For each value of ρ , we considered approximately $10^6\rho$ arrivals with initial queue lengths set to equal the expected average under random routing. Additionally, results from the first tenth of the simulation were discarded to reduce any initial bias and 95% confidence intervals calculated from a sample of ten independent simulations. Behavior under each policy was coupled using the same arrival and service processes for each simulation run to reduce the variance of the difference in population between policies. Mean system population (with 95% confidence intervals) under SDR and the threshold policy, as a proportion of the mean population under WRR, is shown in figure 3.

First observe the significant advantage that both dynamic policies offer over WRR. Next notice that as traffic intensity increases, the discrepancy between the performance of the SDR policy and the threshold policy decreases as well, which is to be expected since both policies are asymptotically optimal in the heavy traffic limit. Furthermore, mean population under threshold routing never exceeded about 112% of the population under SDR, indicating that a simple threshold policy performs well over a range of traffic intensities. As the threshold policy only requires state information on queue 2, in applications where it is easy to observe one queue length but not both, the threshold policy provides a valuable alternative to the WRR and SDR policies.

Acknowledgements

We are grateful to Frank Kelly and Neil Laws for many valuable discussions during the early stages of this work. We would also like to thank Anton Kleywegt, Erica Plambeck, Stephen Turner, Ruth Williams, and the two anonymous referees for their many helpful comments.

References

- [1] S. Asmussen, *Applied Probability and Queues* (Wiley, New York, 1987).
- [2] S.L. Bell and R.J. Williams, Dynamic scheduling of a system with two parallel servers in heavy traffic with complete resource pooling: asymptotic optimality of a continuous review threshold policy, to appear in *Ann. Appl. Probab.* (2001).
- [3] P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968).
- [4] H. Chen and D.D. Yao, *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization* (Springer, New York, 2001).
- [5] W. Fischer and K. Meier-Hellstern, The Markov-modulated Poisson process (MMPP) cookbook, *Performance Evaluation* 18 (1992) 149–171.
- [6] G.J. Foschini, On heavy traffic diffusion analysis and dynamic routing in packet switched networks, in: *Computer Performance Measurements, Modeling and Evaluation*, eds. M. Reiser and K. Chandy (North-Holland, Amsterdam, 1977) pp. 499–514.
- [7] G.J. Foschini and J. Salz, A basic dynamic routing problem and diffusion, *IEEE Trans. Commun.* 26 (1978) 320–327.
- [8] R.J. Gibbens and F.P. Kelly, Dynamic routing in fully connected networks, *IMA J. Math. Control Inform.* 7 (1990) 77–111.

- [9] D. Gross and C. Harris, *Fundamentals of Queueing Theory* (Wiley, New York, 1985).
- [10] F.A. Haight, Two queues in parallel, *Biometrika* 45 (1958) 401–410.
- [11] J.M. Harrison, *Brownian Motion and Stochastic Flow Systems* (Wiley, New York, 1985).
- [12] J.M. Harrison, Brownian models of queueing networks with heterogeneous customer populations, in: *Stochastic Differential Systems, Stochastic Control Theory and Applications*, Vol. 10, eds. W. Fleming and P.L. Lions (Springer, New York, 1988) pp. 147–186.
- [13] J.M. Harrison, The BIGSTEP approach to flow management in stochastic processing networks, in: *Stochastic Networks: Theory and Applications*, eds. F.P. Kelly, S. Zachary and I. Ziedins, RSS Lecture Note Series, Vol. 4 (Oxford Univ. Press, Oxford, 1996) pp. 57–90.
- [14] J.M. Harrison, Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete-review policies, *Ann. Appl. Probab.* 8 (1998) 822–848.
- [15] J.M. Harrison, Brownian models of open processing networks: canonical representation of workload, *Ann. Appl. Probab.* 10 (1999) 75–103.
- [16] J.M. Harrison and M.J. Lopez, Heavy traffic resource pooling in parallel-server systems, *Queueing Systems* 33 (1999) 339–368.
- [17] J.M. Harrison and J.A. Van Mieghem, Dynamic control of Brownian networks: State space collapse and equivalent workload formulations, *Ann. Appl. Probab.* 7 (1997) 747–771.
- [18] D.L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic I, *Adv. in Appl. Probab.* 2 (1970) 150–177.
- [19] F.P. Kelly and C.N. Laws, Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling, *Queueing Systems* 13 (1993) 47–86.
- [20] J.F.C. Kingman, Two similar queues in parallel, *Ann. Math. Statist.* 32 (1961) 1314–1323.
- [21] J.F.C. Kingman, On doubly stochastic Poisson processes, *Proc. Cambridge Phil. Soc.* 60 (1964) 923–930.
- [22] L. Kleinrock, *Queueing Systems*, Vol. I: *Theory* (Wiley, New York, 1975).
- [23] G. Koole, P.D. Sparaggis and D. Towsley, Minimizing response times and queue lengths in systems of parallel queues, *J. Appl. Probab.* 36 (1999) 1185–1193.
- [24] T.G. Kurtz and S.R.E. Turner, A threshold routing rule with fixed thresholds, in progress (2002).
- [25] C.N. Laws, Resource pooling in queueing networks with dynamic routing, *Adv. in Appl. Probab.* 24 (1992) 699–726.
- [26] V. Limic and R.J. Williams, Heavy traffic analysis of two servers with threshold routing of discretionary traffic, in progress (2002).
- [27] C. Maglaras, Dynamic scheduling in multiclass queueing networks: stability under discrete-review policies, *Queueing Systems* 31 (1999) 171–206.
- [28] C. Maglaras, Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality, *Ann. Appl. Probab.* 10 (2000) 897–929.
- [29] M.F. Neuts and U. Sumita, Renewal characterization of Markov modulated Poisson processes, *J. Math. Simulation* 2 (1989) 53–70.
- [30] M.I. Reiman, Some diffusion approximations with state space collapse, in: *Modelling and Performance Evaluation Methodology*, eds. F. Baccelli and G. Fayolle, Lecture Notes in Control and Information Sciences, Vol. 60 (Springer, New York, 1983) pp. 209–240.
- [31] Y.C. Teh, Threshold routing strategies for queueing networks, D. Phil. thesis, University of Oxford (1999).
- [32] Y.C. Teh, Dynamic scheduling for queueing networks derived from discrete-review policies, in: *Analysis of Communication Networks: Call Centres, Traffic and Performance*, eds. D. McDonald and S.R.E. Turner, Fields Institute Communication Series (Amer. Math. Soc., Providence, RI, 2000).
- [33] S.R.E. Turner, A join the shorter queue model in heavy traffic, *J. Appl. Probab.* 37 (2000) 212–223.
- [34] L.M. Wein, Brownian networks with discretionary routing, *Oper. Res.* 39 (1991) 332–340.
- [35] W. Whitt, Some useful functions for functional limit theorems, *Math. Oper. Res.* 5 (1980) 67–85.

- [36] W. Whitt, Deciding which queue to join: Some counterexamples, *Oper. Res.* 34 (1986) 55–62.
- [37] R.J. Williams, On dynamic scheduling of a parallel server system with complete resource pooling in heavy traffic, in: *Analysis of Communication Networks: Call Centres, Traffic and Performance*, eds. D. McDonald and S.R.E. Turner, Fields Institute Communication Series (Amer. Math. Soc., Providence, RI, 2000).
- [38] W. Winston, Optimality of the shortest line discipline, *J. Appl. Probab.* 14 (1977) 181–189.