

## Comparing Load Balancing Algorithms for Distributed Queueing Networks

**D. R. McDonald**

Department of Mathematics and Statistics  
University of Ottawa  
585 King Edward Avenue  
Ottawa, Ontario, K1N 6N5 Canada  
dmdsg@omid.mathstat.uottawa.ca

**S. R. E. Turner**

Statistical Laboratory  
University of Cambridge  
Wilberforce Road  
Cambridge, CB3 0WB England  
sret1@cam.ac.uk

**Abstract.** We consider a network serving a patchwork of overlapping regions where jobs from a local region are assigned to a collection of local servers. Copies of these jobs are simultaneously queued at all the local servers. When a copy of the job begins service at one of the servers it is removed from the other queues. The system is equivalent to one in which the exact service requirement of each job is known at arrival time, and each job joins the local queue with the shortest waiting time.

We describe how the amount of work in the network becomes large in the simple case of two servers, with one arrival stream for each server and a third, routeable arrival stream. If the proportion of routeable jobs is large enough then the waiting times at the servers become large in tandem when the total workload becomes large, thus delaying overload as long as possible. The fact that this resource pooling can be attained with a local routing policy not dependent on the state of the network has engineering significance for load sharing among distributed call centres.

We also compare this 'join the shorter actual waiting time' policy (JSAW) with a join the shorter expected waiting time policy and join the shorter queue. For some overflow events, we find that the performance of all three policies is roughly the same in the sense that the probability of an overflow has the same exponential decay rate under any policy. Although the JSAW policy is the best, in these cases its probability of overflow is only the lowest by a subexponential factor. However, for other overflow events we find that the JSAW policy is substantially better.

---

2000 *Mathematics Subject Classification.* Primary 60K25; Secondary 60K20.

The authors were supported in part by NSERC grant A4551. SRET was also supported by Sidney Sussex College, Cambridge.

## 1 Introduction

Consider a network of call centres, arranged geographically, with the following queueing discipline. Each customer has one or more local call centres, according to his location. When he requires service, he queues simultaneously at *all* of his local call centres. When he begins service at one of the call centres, his presence in the other queues is deleted. The room available for queueing is infinite, and all queues are FIFO (except for deletions).

It is easy to see that this system is identical, in terms of the time at which each customer is served, to one in which every arriving customer joins the queue among those local to him with the shortest actual waiting time. We call the latter policy ‘join the shorter actual waiting time’ or JSAW. The simultaneous queueing at all local call centres is just a device to allow this policy to be implemented without knowing the service requirements of customers in advance.

The JSAW policy is also of interest in its own right. Intuitively, it seems that it should be advantageous to base routing decisions on the actual waiting time in each queue, if this is known, rather than just on the queue length. In order to examine this intuition, we shall compare JSAW with two other policies. In ‘join the shorter expected waiting time,’ or JSEW, each arriving customer joins the queue for which the number of customers divided by the service rate is smallest. In ‘join the shorter queue,’ or JSQ, it is assumed that the service rates of the queues are not known, and each customer simply joins the queue with the smallest number of customers.

We shall see that JSAW sometimes has comparable overflow probability to JSEW and JSQ, but is sometimes far superior, depending on the exact overflow event under consideration. Furthermore, in some ways, JSAW can be implemented with less knowledge of the state of the network than JSEW or JSQ. JSEW requires knowledge of the relative service rates of each queue, which may well be unrealistic in practice, whereas the ‘simultaneous queueing’ implementation of JSAW requires no such knowledge. And even JSQ requires up-to-date knowledge of the current queue lengths at the time of each arrival, which simultaneous queueing again does not need. It is true that JSAW does require communication between servers whenever a customer begins service, in order to delete that customer’s presence in the other queues. On the other hand, JSAW has the additional advantage of providing automatic recovery when a server fails since customers at the failed server are already queued elsewhere. Moreover recovery under JSAW automatically maintains sequencing. For these reasons, we expect that simultaneous queueing may find application in call centres and in other similar models such as DNS servers.

The outline of this paper is as follows. In the next section, we shall define our model more exactly. In Section 3 we shall examine the model in a large deviations limit, using the methods of Turner [7]. In Section 4 we shall use the  $h$ -transform method of Foley & McDonald [4] and McDonald [5]. A subsidiary aim of this paper is to compare these two approaches. We shall see that the former approach is somewhat simpler, but that when the latter is applicable, it yields more powerful results. Finally, in Section 5 we shall report the results of some simulations and draw some final conclusions.

## 2 The model

The model which we shall examine is that described in Section 1 in the case that there are only two call centres, and only one server at each centre. Thus there

are three types of customer: those which may only receive service from the first server, those which may only receive service from the second server, and those which may be served by either server. Let the arrivals of these three customer types be independent Poisson processes of rates  $\lambda_1$ ,  $\lambda_2$  and  $\gamma$  respectively. Each customer brings an amount of work which is distributed as  $\text{Exp}(1)$ , independent of any other quantity in the model. The servers serve at constant rates  $\mu_1$  and  $\mu_2$  respectively. The possible queueing disciplines — JSAW, JSEW and JSQ — have already been described in Section 1. For definiteness, we shall assume that in the case of a tie, routeable customers are routed to the second queue.

It is not unreasonable to model customer arrivals to a call centre or DNS server over a moderate period of time by a Poisson process. It is however unlikely that an exponential service time is appropriate for all models, and this assumption is introduced primarily for simplicity of analysis. We could assume more generally that each customer brings an amount of work with a short-tailed distribution with mean 1. The large deviations theory to analyse this generalization is not yet available, but we shall make a few comments about how to analyse it using the method of  $h$ -transforms at the end of Section 4.1.

We need to introduce some notation. First, let  $\lambda_+ = \lambda_1 + \lambda_2 + \gamma$ ,  $\mu_+ = \mu_1 + \mu_2$ ,  $\rho_1 = \lambda_1/\mu_1$ ,  $\rho_2 = \lambda_2/\mu_2$ , and  $\rho_+ = \lambda_+/\mu_+$ . Then let  $Q(t) = (Q_1(t), Q_2(t))$  denote the queue sizes at time  $t$ ,  $L(t) = (L_1(t), L_2(t))$  the workload remaining in each queue, and  $S(t) = (S_1(t), S_2(t)) = (L_1(t)/\mu_1, L_2(t)/\mu_2)$  the times required to complete that work.

It is clear that the following condition is necessary for the stability of the system under any routing policy.

**Condition 2.1**  $\rho_1 < 1$ ,  $\rho_2 < 1$  and  $\rho_+ < 1$ .

For JSQ, Condition 2.1 is shown to be a sufficient condition for the stability of the queue sizes in Foley & McDonald [4]. In Section 4.3 we show that under JSEW,  $Q(t)$  is stable if Condition 2.1 holds. In Section 4.1 we show that under JSAW,  $S(t)$  is stable if Condition 2.1 holds. Consequently the system is stable for all three policies if and only if Condition 2.1 holds. We shall assume Condition 2.1 throughout this paper.

We shall also sometimes assume one or both of the following additional conditions.

**Condition 2.2**  $\rho_+ > \max\{\rho_1, \rho_2\}$ .

**Condition 2.3**  $\gamma > |\rho_+^2(\mu_2 - \mu_1) + (\lambda_1 - \lambda_2)|$ .

Informally, these conditions will ensure that  $\gamma$  is sufficiently large to be able to balance the two queues. In the case that  $\mu_1 = \mu_2$ , they both reduce to the natural condition that  $\gamma > |\lambda_1 - \lambda_2|$ . Otherwise, Condition 2.3 is a stronger condition than Condition 2.2. An intuitive explanation of these conditions can be found in [7].

If we let  $\vec{X}(t) = (X_1(t), X_2(t), \dots, X_{Q_1(t)}(t))$  be the remaining service times for the customers in the first queue ( $X_1(t)$  being that customer currently in service) and similarly  $\vec{Y}(t)$  for the customers in the second queue, then the process

$$M(t) = (Q_1(t), Q_2(t), \vec{X}(t), \vec{Y}(t)) \quad (2.1)$$

is Markovian, whichever queueing discipline is being used. Let  $\pi_Q$ ,  $\pi_E$  and  $\pi_A$  be the stationary distributions of  $M(t)$  under the policies JSQ, JSEW and JSAW respectively.

There are several possible ways to measure the overload of the system. For example, if queueing space was costly, one might consider the system to be overloaded if  $Q_1(t)$  and  $Q_2(t)$  became large. But we shall concentrate in this paper on overflow events concerning the total amount of work in the system,  $L_1(t) + L_2(t)$ , and the maximum waiting time,  $\max\{S_1(t), S_2(t)\}$ . None of these measures is right or wrong in itself, and indeed many others are possible. Different ones may be more or less relevant for different networks.

### 3 Large deviations

In this section, we shall consider the system using the large deviations methods of Turner [7] in this volume. A more detailed exposition can be found there. We shall assume throughout this section that the process has been scaled in the usual large deviations way, by dividing the jump sizes by  $n$  and multiplying the speed of the process by  $n$ . Variables with superscript  $n$ 's will represent this scaled process.

Let  $C$  be a fixed constant, and consider the event that the total scaled workload of the system reaches  $\mu_+C$  in its first excursion from 0 (the  $\mu_+$  is just for notational convenience). By the results of Dupuis & Ellis [2], the scaled workload process obeys a large deviations principle, and so there is a large deviations rate corresponding to this event. (In fact [2] only directly proves a large deviations principle for the process over a finite scaled time interval  $[0, T]$ , but the large deviations rate is the same for all sufficiently large  $T$ , and thus by Lemma C.4 of Shwartz & Weiss [6] applies also for the time interval  $[0, \infty)$ .)

Now consider the pooled system consisting of a single  $M/M/1$  queue with arrival rate  $\lambda_+$  and service rate  $\mu_+$ . The large deviations rate for this queue to reach scaled workload  $\mu_+C$  during an excursion is  $\mu_+C(1 - \rho_+)$  (see, for example, Section VI.9(e) of Feller [3]). Now the JSAW system can be coupled with this system in such a way that the total workload of the JSAW system is always greater than that of the pooled system. Thus the large deviations rate for the JSAW system to reach scaled workload  $\mu_+C$  is at most  $\mu_+C(1 - \rho_+)$ .

Now consider a path to overflow which leaves the diagonal  $L_1^n/\mu_1 = L_2^n/\mu_2$  at  $(\mu_1k, \mu_2k)$ , and then overflows the line  $L_1^n + L_2^n = \mu_+C$  at  $(a, \mu_+C - a)$ , where  $a > \mu_1C$ . The tail portion of this path, after it has left the diagonal, has large deviations rate at least

$$(a - \mu_1k) \left(1 - \frac{\lambda_1}{\mu_1}\right) + (\mu_+C - a - \mu_2k) \left(1 - \frac{\lambda_2 + \gamma}{\mu_2}\right). \quad (3.1)$$

But by the assumption that  $\rho_+ > \max\{\rho_1, \rho_2\}$ , it follows that  $(\lambda_2 + \gamma)/\mu_2 > \lambda_1/\mu_1$ , and thus that the derivative with respect to  $a$  of expression (3.1) is strictly positive. Thus the large deviations rate of the tail portion of the path is strictly greater than the value of (3.1) with  $a = \mu_1C$ , that is,

$$\begin{aligned} & \mu_1(C - k) \left(1 - \frac{\lambda_1}{\mu_1}\right) + \mu_2(C - k) \left(1 - \frac{\lambda_2 + \gamma}{\mu_2}\right) \\ &= \mu_+(C - k)(1 - \rho_+), \end{aligned}$$

and by the comparison with the pooled system, this is a contradiction. So, exactly as in [7], we have shown enough to prove the following theorem.

**Theorem 3.1** *If Conditions 2.1 and 2.2 hold, then in the large deviations limit the JSAW system reaches total scaled workload  $\mu_+C$  by travelling along the diagonal, with large deviations rate  $\mu_+C(1 - \rho_+)$ .*

It may be that we are worried about the waiting time for any customer becoming large. We can also consider such overflow events by this method. The event of interest is that  $L_1^n \geq \mu_1C$  or  $L_2^n \geq \mu_2C$  during an excursion. Conditioning on the last point,  $(\mu_1k, \mu_2k)$ , which the overflow path touches along the diagonal, the large deviations rate for this is

$$\inf_{i,k} \{ \mu_+k(1 - \rho_+) + \mu_i(C - k)(1 - \rho_i) \}. \quad (3.2)$$

This expression is linear in  $k$ , so the infimum is always attained at  $k = 0$  or  $k = C$ . It is then easy to solve (3.2) to obtain the following theorem.

**Theorem 3.2** *If Conditions 2.1 and 2.2 hold, then the large deviations path for the JSAW system to reach scaled waiting time  $C$  in either queue is along the diagonal if  $\gamma > \max\{\mu_1 - \lambda_1, \mu_2 - \lambda_2\}$ , with large deviations rate  $\mu_+C(1 - \rho_+)$ ; otherwise along the axis corresponding to the smaller value of  $\mu_i(1 - \rho_i)$ , with large deviations rate  $\mu_iC(1 - \rho_i)$ .*

The JSEW and JSQ policies have already been analysed in [7]. So it is next natural to want to compare the JSAW policy with those policies, to see if one has an advantage in terms of large deviations rates of overflow. However such a comparison is not immediate, because the state spaces and overflow events normally used for these policies are not the same. In the JSAW model, the state space is the actual waiting times in each queue and the overflow events concern the *actual* waiting times becoming large. In the JSEW and JSQ models, the state space is the two queue lengths, and the overflow events concern the the queues becoming long (which is equivalent to the *expected* waiting times becoming large).

To put this another way, for the comparison one wants to look at the large deviations rate for the actual waiting time becoming large in the JSEW or JSQ system. But the actual waiting time becoming large is not an event in those policies' natural state space. One can of course extend the state space to include the actual waiting times of each customer, as at equation (2.1), but there does not seem to be any way to obtain a large deviations principle for this extended process.

This problem is not confined to our model. Most large deviations results for overflows in queueing systems consider the queues becoming large. But in many networks, what the operator of the network is actually concerned about is whether waiting times become large. The queues becoming large is just used as a substitute for this because it is easier to analyse. But one cannot immediately translate from one result to the other, because during an overflow of waiting times, the waiting times of the customers already queueing are not typical. The waiting time of an arriving customer becomes large both by the number of customers queueing becoming large and by them having larger service requirements than typical customers.

However, we are able to deduce an upper bound on the large deviations rate for the actual waiting time becoming large in the JSEW and JSQ systems from the large deviations principle for the queues becoming large. We are grateful to Dr Y. Git of the University of Cambridge for observing this result.

**Theorem 3.3** *If Conditions 2.1 and 2.2 hold, then the large deviations rate for either the JSEW or the JSQ system to reach total scaled workload  $\mu_+C$  is at most  $\mu_+C(1 - \rho_+)$ .*

**Proof** Let  $q > 0$  be fixed. Let  $L^{\max}$  be the maximum total unscaled workload during a busy period, and if the maximum total unscaled queue length  $Q^{\max}$  during that busy period is at least  $nq$ , let  $L^\dagger$  be the total unscaled workload the first time that the queue length reaches at least  $nq$ . Then it is immediate that

$$P(L^{\max} > nl) \geq P(Q^{\max} > nq)P(L^\dagger > nl). \quad (3.3)$$

But the service requirements of the customers who are waiting for service at any time are independent of the number of customers in the queue. Therefore  $L^\dagger$  is the sum of  $nq$  i.i.d. typical workloads, which are  $\text{Exp}(1)$  random variables. So by Cramér's theorem,  $L^\dagger/nq$  obeys a large deviations principle with rate function  $I_M(t) = t - 1 - \log t$ . Furthermore, assuming Condition 2.2, we know from [7] that under JSEW or JSQ, the maximum total scaled queue length also obeys a large deviations principle with rate function  $I_Q(q) = -q \log \rho_+$ . (Since the number of customers in service at any time is at most 2, this large deviations principle is the same whether or not we included those currently being served.) We can now apply  $(\liminf \frac{1}{n} \log)$  to each side of equation (3.3) to obtain

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P(L^{\max} > nl) \geq (-I_Q(q) - qI_M(l/q)\chi\{l/q \geq 1\}),$$

where  $\chi$  is the indicator function. Next we can take the supremum over  $q > 0$ . Since  $I_Q(q)$  is increasing in  $q$ , the supremum is never attained at  $q > l$ . Thus we obtain

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P(L^{\max} > nl) \geq -\inf_{q>0} \{I_Q(q) + qI_M(l/q)\}. \quad (3.4)$$

We now wish to solve (3.4). Using the values of  $I_Q$  and  $I_M$  above, we find that the large deviations rate for the event that the total scaled workload exceeds  $l$  is at most

$$\inf_{q>0} \{-q \log \rho_+ + l - q - q \log l + q \log q\}. \quad (3.5)$$

Setting the derivative of this expression with respect to  $q$  equal to 0,

$$-\log \rho_+ - 1 - \log l + 1 + \log q = 0,$$

which is solved by  $q = l\rho_+$ . Then substituting back into (3.5), the large deviations rate is at most

$$\begin{aligned} & -l\rho_+ \log \rho_+ + l - l\rho_+ - l\rho_+ \log l + l\rho_+ \log(l\rho_+) \\ & = l(1 - \rho_+). \end{aligned}$$

With  $l = \mu_+C$ , this is the rate in the statement of the theorem.  $\square$

Note that the large deviations rate in Theorem 3.1 is the same as that in Theorem 3.3. In other words, on the large deviations scale the JSAW policy is at least as good as the JSEW and JSQ policies, as we would expect. One might conjecture that in fact Theorem 3.3 holds with equality. This amounts to saying that the probability of a large workload during an excursion is dominated by the probability of a large workload at the time the queue is longest (and that the greatest queue length is that found in the proof of the theorem). One would then be able to say that the JSAW, JSEW and JSQ policies were equally good in large

deviations scaling, for this overflow event. In the next section, we shall prove this result by the method of  $h$ -transforms used in McDonald [5] and we shall observe it in simulations in Section 5.

We can also use a similar method to obtain a bound on the large deviations rate for the overflow event of Theorem 3.2 under JSEW and JSQ. We shall just sketch the proof. The idea is to condition on the maximum total queue length during a busy period, as in Theorem 3.3, but then to consider the probability that just one of the two queues achieved a large waiting time at that time.

For the rest of this section, we shall assume without loss of generality that  $\mu_1 \leq \mu_2$ . Then under the JSEW policy, assuming Condition 2.2, we know from [7] that the large deviations rate to reach total scaled queue length  $\mu_+q$  is  $-\mu_+q \log \rho_+$ , and moreover that this is achieved by reaching individual scaled queue lengths  $(\mu_1q, \mu_2q)$ . The large deviations rate for those customers in the first queue at that time to have total scaled waiting time at least  $C$  is then

$$\mu_1C - \mu_1q - \mu_1q \log(\mu_1C) + \mu_1q \log(\mu_1q).$$

So the total large deviations rate to reach scaled waiting time  $C$  in this way is, in an expression parallel to (3.5),

$$\inf_{q>0} \{-\mu_+q \log \rho_+ + \mu_1C - \mu_1q - \mu_1q \log(\mu_1C) + \mu_1q \log(\mu_1q)\}. \quad (3.6)$$

This is minimised by  $q = C\rho_+^{\mu_+/\mu_1}$ , giving a rate of  $\mu_1C(1 - \rho_+^{\mu_+/\mu_1})$ .

The calculation for the JSQ case is similar, provided Condition 2.3 holds. In that case, again from [7], the large deviations rate to reach scaled queue lengths  $(\mu_1q, \mu_1q)$  is  $-2\mu_1q \log \rho_+$ . So (3.6) becomes

$$\inf_{q>0} \{-2\mu_1q \log \rho_+ + \mu_1C - \mu_1q - \mu_1q \log(\mu_1C) + \mu_1q \log(\mu_1q)\},$$

which is minimised by  $q = C\rho_+^2$ , giving an overall rate of  $\mu_1C(1 - \rho_+^2)$ . (If Condition 2.3 does not hold, then a given total occupancy is not achieved by keeping the two scaled queues equal. This makes the minimisation harder to write down, but in principle it can still be solved by the same methods.)

We state our conclusions as a theorem.

**Theorem 3.4** *If Conditions 2.1 and 2.2 hold, then the large deviations rate for the JSEW system to reach scaled waiting time  $C$  in either queue is at most  $\mu_1C(1 - \rho_+^{\mu_+/\mu_1})$ . If Condition 2.3 also holds, then the large deviations rate for the JSQ system to reach scaled waiting time  $C$  in either queue is at most  $\mu_1C(1 - \rho_+^2)$ .*

Even when Conditions 2.1–2.3 hold, there may still be a cheaper path to overflow than that implied by Theorem 3.4. For example, travel along the axis has a lower large deviations rate than the one in the theorem for some parameters. The point of this theorem is not to find the best possible rate, but to prove that for some parameters, the large deviations rate for the JSEW and JSQ systems is strictly lower than that for JSAW, given in Theorem 3.2. Thus for some overflow events, such as this one, JSAW can have a lower probability of overflow than JSEW or JSQ by an exponential factor. This is in contrast to the remarks under Theorem 3.3. We shall also observe this in simulations in Section 5.

#### 4 The $h$ -transform method

In this section we shall analyse the model again, this time using the methods of Foley & McDonald [4] and McDonald [5].

The proof of the following theorem is given in Section 4.1.

**Theorem 4.1** *If Conditions 2.1 and 2.2 hold, then under JSAW*

$$P_{\pi_A}(L_1(t) + L_2(t) \in dl, S_1(t) - S_2(t) \in dz) \sim \frac{f_A}{\tilde{d}_1} e^{-(1-\rho_+)t} dl \cdot \varphi(z) dz \quad (4.1)$$

$$P_{\pi_A}(Q_1(t) + Q_2(t) = q) \sim \frac{f_A}{\tilde{d}_1} \rho_+^q \quad (4.2)$$

as  $\ell \rightarrow \infty$  and  $q \rightarrow \infty$ , where the convergence is in total variation,  $f_A$  is defined at (4.21), and the density  $\varphi$  is defined by

$$\varphi(z) = \begin{cases} \beta \alpha_+ e^{-\alpha_+ z} & \text{if } z > 0 \\ (1 - \beta) \alpha_- e^{-\alpha_- |z|} & \text{if } z < 0 \end{cases}$$

where

$$\begin{aligned} \alpha_+ &= \left( \frac{\lambda_2 + \gamma}{\mu_2} - \frac{\lambda_1}{\mu_1} \right) \frac{\mu_1 \mu_2}{\mu_1 + \mu_2} \\ \alpha_- &= \left( \frac{\lambda_1 + \gamma}{\mu_1} - \frac{\lambda_2}{\mu_2} \right) \frac{\mu_1 \mu_2}{\mu_1 + \mu_2} \\ \text{and } \beta &= \frac{\alpha_-}{\gamma}. \end{aligned}$$

Hence under JSAW the waiting times are strongly pooled in the sense that the difference between  $S_1$  and  $S_2$  at overload converges to a random variable.

Denote the time when the workload exceeds  $\ell$  by  $T_\ell$ . Then

$$\mathbb{E}[T_\ell \mid S(0) = (0, 0)] \sim g_A^{-1} e^{(1-\rho_+)\ell} \quad (4.3)$$

where the constant  $g_A$  is defined at (4.22). Moreover,

$$\frac{1}{\ell} (L_1(T_\ell), L_2(T_\ell)) \rightarrow \left( \frac{\mu_1}{\mu_1 + \mu_2}, \frac{\mu_2}{\mu_1 + \mu_2} \right) \quad (4.4)$$

$$\frac{1}{\ell} (Q_1(T_\ell), Q_2(T_\ell)) \rightarrow \rho_+ \left( \frac{\mu_1}{\mu_1 + \mu_2}, \frac{\mu_2}{\mu_1 + \mu_2} \right) \quad (4.5)$$

where  $\rightarrow$  denotes weak convergence. Hence the queue sizes are weakly pooled in the sense of Foley & McDonald [4].

We can compare the above results with those in Foley & McDonald [4] for the join the shorter queue policy. Proofs are given in Section 4.3.

**Theorem 4.2** *If Conditions 2.1, 2.2 and 2.3 hold, then under JSQ, as  $\ell \rightarrow \infty$  and  $q \rightarrow \infty$ ,*

$$P_{\pi_Q}(L_1(t) + L_2(t) \in dl, Q_1(t) - Q_2(t) = k) \sim \frac{f_Q}{\tilde{d}_1} e^{-(1-\rho_+)t} dl \cdot \varphi_Q(k) \quad (4.6)$$

$$\lim_{q \rightarrow \infty} P_{\pi_Q}(Q_1(t) + Q_2(t) = q) \sim \frac{f_Q}{\mu_1 + \mu_2 - (\lambda_1 + \lambda_2 + \gamma)} \rho_+^q \quad (4.7)$$

where the convergence is in total variation and where  $\tilde{d}_1 = \mu_+ - \lambda_+$ ,  $f_Q$  is defined at (4.21) and the density  $\varphi_Q$  is defined by

$$\varphi_Q(k) = \begin{cases} \varphi(0) \frac{\rho_+^{-1}(\lambda_1 + \gamma/2) + \rho_+ \mu_2}{\rho_+^{-1}(\lambda_2 + \gamma) + \rho_+ \mu_1} \left( \frac{\rho_+^{-1} \lambda_1 + \rho_+ \mu_2}{\rho_+^{-1}(\lambda_2 + \gamma) + \rho_+ \mu_1} \right)^{k-1} & \text{if } k > 0 \\ \varphi(0) \frac{\rho_+^{-1}(\lambda_2 + \gamma/2) + \rho_+ \mu_1}{\rho_+^{-1}(\lambda_1 + \gamma) + \rho_+ \mu_2} \left( \frac{\rho_+^{-1} \lambda_2 + \rho_+ \mu_1}{\rho_+^{-1}(\lambda_1 + \gamma) + \rho_+ \mu_2} \right)^{|k|-1} & \text{if } k < 0 \end{cases}$$

where  $\varphi_Q(0) \equiv \varphi(0)$  is defined to be

$$\left( \frac{\rho_+^{-1}(\lambda_1 + \gamma/2) + \rho_+ \mu_2}{\rho_+^{-1}(\lambda_2 + \gamma) + \rho_+ \mu_1 - (\rho_+^{-1} \lambda_1 + \rho_+ \mu_2)} + \frac{\rho_+^{-1}(\lambda_2 + \gamma/2) + \rho_+ \mu_1}{\rho_+^{-1}(\lambda_1 + \gamma) + \rho_+ \mu_2 - (\rho_+^{-1} \lambda_2 + \rho_+ \mu_1)} + 1 \right)^{-1}.$$

If the time when the workload exceeds  $\ell$  is again denoted by  $T_\ell$ , then under JSQ

$$\mathbb{E}[T_\ell \mid S(0) = (0, 0)] \sim g_Q^{-1} e^{(1-\rho_+)\ell} \quad (4.8)$$

where the constant  $g_Q$  is defined at (4.22). Moreover,

$$\frac{1}{\ell} L_1(T_\ell) \rightarrow \frac{1}{2} \quad \text{and} \quad \frac{1}{\ell} L_2(T_\ell) \rightarrow \frac{1}{2} \quad (4.9)$$

$$\frac{1}{\ell} Q_1(T_\ell) \rightarrow \frac{\rho_+}{2} \quad \text{and} \quad \frac{1}{\ell} Q_2(T_\ell) \rightarrow \frac{\rho_+}{2} \quad (4.10)$$

where  $\rightarrow$  denotes weak convergence again.

In Section 4.3 we also sketch the corresponding result for JSEW.

**Theorem 4.3** *If Conditions 2.1 and 2.2 hold, then under JSEW,*

$$P_{\pi_E} \left( L_1(t) + L_2(t) \in d\ell, \frac{Q_1(t)}{\mu_1} - \frac{Q_2(t)}{\mu_2} \in dz \right) \sim \frac{f_E}{d_1} e^{-(1-\rho_+)\ell} d\ell \cdot \varphi(z) dz \quad (4.11)$$

$$P_{\pi_E}(Q_1(t) + Q_2(t) = q) \sim \frac{f_E}{d_1} \rho_+^q \quad (4.12)$$

as  $\ell \rightarrow \infty$ , where the convergence is in total variation and where  $f_E$  is defined at (4.21). The density  $\varphi$  is defined on the additive subgroup  $\{m/\mu_1 + n/\mu_2\}$  where  $m$  and  $n$  are integers.

If the time when the workload exceeds  $\ell$  is again denoted by  $T_\ell$ , then under JSEW

$$\mathbb{E}[T_\ell \mid S(0) = (0, 0)] \sim g_E^{-1} e^{(1-\rho_+)\ell} \quad (4.13)$$

where the constant  $g_E$  is defined in (4.22). Moreover, as for JSAW,

$$\frac{1}{\ell} (L_1(T_\ell), L_2(T_\ell)) \rightarrow \left( \frac{\mu_1}{\mu_1 + \mu_2}, \frac{\mu_2}{\mu_1 + \mu_2} \right) \quad (4.14)$$

$$\frac{1}{\ell} (Q_1(T_\ell), Q_2(T_\ell)) \rightarrow \rho_+ \left( \frac{\mu_1}{\mu_1 + \mu_2}, \frac{\mu_2}{\mu_1 + \mu_2} \right) \quad (4.15)$$

where  $\rightarrow$  denotes weak convergence. Hence the queue sizes are weakly pooled in the sense of Foley & McDonald [4].

If Conditions 2.1–2.3 hold, all three policies are roughly equivalent for overflows of the total workload and total queue size. In particular, all policies give the same rough asymptotic probability of a large deviation of the total workload to level  $\ell$ , namely  $\exp(-(1-\rho_+)\ell)$ , and of a large deviation of the total queue size to level  $q$ , namely  $\rho_+^q$ . When the rare event of interest is a large deviation of the total workload, then if Conditions 2.1 and 2.2 hold, JSAW strongly pools the waiting times while the queue lengths are weakly pooled. JSEW has the weaker property

of strongly pooling the expected waiting times when a large deviation of the total workload occurs. Under JSQ the queue lengths are strongly pooled when a large deviation of the total workload occurs but only if Condition 2.3 holds. In this sense JSAW and JSEW are more robust policies than JSQ because under those policies we obtain strong pooling without Condition 2.3.

**4.1 Exact asymptotics of JSAW.** In this subsection we use an extension of the  $h$ -transform method developed in McDonald [5] and Foley & McDonald [4] to prove Theorem 4.1. Here is an overview of the method. We wish to calculate the probability of large deviations trajectories of a Markov chain  $W$  (here a process with two components; one the workload and the other the difference of the waiting times at the two queues) to a set  $F$  (here,  $F$  represents states where the workload exceeds  $\ell$ ). We find a boundary  $\Delta$  (here, the states where either queue is empty) and a function  $h$  such that  $h$  is harmonic for  $W$  outside  $\Delta$  and simple on  $F$  (here,  $h$  is just a function of the workload). Then we can construct the  $h$ -transformed process  $\mathcal{W}^\infty$  on  $\Delta^c$ . The ratio of the likelihood of a path from  $w_0 \in \Delta$  to  $w_f \in F$  without hitting  $\Delta$  under  $W$  to the likelihood of that path under  $\mathcal{W}^\infty$  is  $h(w_0)/h(w_f)$ . Hence, if this path is likely under  $\mathcal{W}^\infty$  then  $1/h(w_f)$  measures just how unlikely the path is under  $W$  and in fact gives the asymptotics of the steady state probability of  $F$  as  $F$  becomes rare. This trajectory is weighted by the factor  $\pi(w_0)h(w_0)$ , and Condition 6 in McDonald [5], checked below, requires the total weight  $\int_{w_0 \in \Delta} \pi(dw_0)h(w_0)$  to be finite.

We start by defining the generator  $G$  of the Markov process  $S(t)$ .  $S(t)$  is a Markov process in  $[0, \infty) \times [0, \infty)$  whose two components decrease linearly with time between jumps. At a point  $(x, y)$  where  $x \geq y$  the waiting time of the first queue exceeds the second so the second queue receives the discretionary customers. Consequently the waiting time of the second queue increases by jumps of exponential size with parameter  $\mu_2$  at a rate  $\lambda_2 + \gamma$  while the waiting time of the first queue increases by jumps of exponential size with parameter  $\mu_1$  at a rate  $\lambda_1$ . Hence if  $f$  is a smooth function on  $[0, \infty) \times [0, \infty)$  and  $x \geq y$  then

$$\begin{aligned} Gf(x, y) &= -\chi\{x > 0\} \frac{\partial f}{\partial x}(x, y) - \chi\{y > 0\} \frac{\partial f}{\partial y}(x, y) \\ &\quad + \lambda_1 \int_0^\infty [f(x+s, y) - f(x, y)] \mu_1 \exp(-\mu_1 s) ds \\ &\quad + (\lambda_2 + \gamma) \int_0^\infty [f(x, y+s) - f(x, y)] \mu_2 \exp(-\mu_2 s) ds, \end{aligned}$$

where  $\chi$  is the indicator function. Similarly if  $x < y$ ,

$$\begin{aligned} Gf(x, y) &= -\chi\{x > 0\} \frac{\partial f}{\partial x}(x, y) - \chi\{y > 0\} \frac{\partial f}{\partial y}(x, y) \\ &\quad + (\lambda_1 + \gamma) \int_0^\infty [f(x+s, y) - f(x, y)] \mu_1 \exp(-\mu_1 s) ds \\ &\quad + \lambda_2 \int_0^\infty [f(x, y+s) - f(x, y)] \mu_2 \exp(-\mu_2 s) ds. \end{aligned}$$

We can easily check that  $S(t)$  is stable under the join the shorter waiting time policy. It suffices to consider the Liapounov function  $V(x, y) := \mu_1 x^2 + \mu_2 y^2$ . If

$x \geq y \geq 0$ ,

$$\begin{aligned} GV(x, y) &= -2\mu_1 x - 2\mu_2 y + \lambda_1 \mu_1 \int_0^\infty [(x+s)^2 - x^2] \mu_1 \exp(-\mu_1 s) ds \\ &\quad + (\lambda_2 + \gamma) \mu_2 \int_0^\infty [(y+s)^2 - y^2] \mu_2 \exp(-\mu_2 s) ds \\ &= -2\mu_1 x - 2\mu_2 y + \lambda_1 (2x + 2/\mu_1) + (\lambda_2 + \gamma) (2y + 2/\mu_2) \\ &= -2x(\mu_1 - \lambda_1) - 2y(\mu_2 - (\lambda_2 + \gamma)) + (2\lambda_1/\mu_1 + 2(\lambda_2 + \gamma)/\mu_2) \end{aligned}$$

Hence  $GV(x, y)$  is bounded above by

$$\begin{cases} -2x(\mu_1 - \lambda_1) + (2\lambda_1/\mu_1 + 2(\lambda_2 + \gamma)/\mu_2) & \text{if } \mu_2 > (\lambda_2 + \gamma) \\ -2x(\mu_1 + \mu_2 - (\lambda_1 + \lambda_2 + \gamma)) + (2\lambda_1/\mu_1 + 2(\lambda_2 + \gamma)/\mu_2) & \text{if } \mu_2 \leq (\lambda_2 + \gamma). \end{cases}$$

In either of these cases we have  $GV(x, y) \leq -1$  for  $x$  sufficiently large by Condition 2.1. We can do a similar calculation for  $x < y$ , so we conclude that  $GV(x, y) \leq -1$  except for a finite region near  $(0, 0)$ . It follows that  $S(t)$  is stable.

In order to convert this problem into the Markov additive framework of McDonald [5] we make a change of variable,  $v = \mu_1 x + \mu_2 y$  and  $z = x - y$ . In these new coordinates we have an equivalent process

$$W(t) := (\tilde{W}_1(t), \hat{W}_2(t)) \equiv (\mu_1 S_1(t) + \mu_2 S_2(t), S_1(t) - S_2(t))$$

defined on  $\mathcal{D} := \{(v, z) = (\mu_1 x + \mu_2 y, x - y) : x, y \geq 0\}$ . Let the boundary of this domain be

$$\Delta = \Delta_1 \cup \Delta_2 \text{ where } \Delta_1 := \{(\mu_2 y, -y), y \geq 0\}, \Delta_2 := \{(\mu_1 x, x), x \geq 0\}.$$

The associated generator  $L$  is

$$\begin{aligned} Lg(v, z) &= -(\mu_1 \chi\{(v, z) \in \mathcal{D} \setminus \Delta_1\} + \mu_2 \chi\{(v, z) \in \mathcal{D} \setminus \Delta_2\}) \frac{\partial g}{\partial v}(v, z) \\ &\quad + \lambda_1 \int_0^\infty [g(v + \mu_1 s, z + s) - g(v, z)] \mu_1 \exp(-\mu_1 s) ds \\ &\quad + (\lambda_2 + \gamma) \int_0^\infty [f(v + \mu_2 s, z - s) - f(v, z)] \mu_2 \exp(-\mu_2 s) ds \end{aligned}$$

if  $z \geq 0$ , where  $g$  is any smooth function. Similarly if  $z < 0$ ,

$$\begin{aligned} Lg(v, z) &= -(\mu_1 \chi\{(v, z) \in \mathcal{D} \setminus \Delta_1\} + \mu_2 \chi\{(v, z) \in \mathcal{D} \setminus \Delta_2\}) \frac{\partial g}{\partial v}(v, z) \\ &\quad + (\lambda_1 + \gamma) \int_0^\infty [g(v + \mu_1 s, z + s) - g(v, z)] \mu_1 \exp(-\mu_1 s) ds \\ &\quad + \lambda_2 \int_0^\infty [f(v + \mu_2 s, z - s) - f(v, z)] \mu_2 \exp(-\mu_2 s) ds. \end{aligned}$$

$W$  is a Markov process on  $\mathcal{D}$  with a stationary distribution which by an abuse of notation we continue to denote by  $\pi_A$ . The overflow set  $F_\ell$  becomes the set  $\{(v, z) : v \geq \ell\}$ . To apply the theory in McDonald [5] we need to consider the *free* process on  $(-\infty, \infty) \times (-\infty, \infty)$  with generator  $L^\infty$  extended from  $L$  by removing the boundary  $\Delta$ . It is easy to verify that  $h(v, z) = \exp((1 - \rho_+)v)$  is a harmonic for this generator, that is, a solution to  $L^\infty h(v, z) = 0$ .

For  $z > 0$  and any smooth function  $g$  we now calculate the generator of the twisted or  $h$ -transformed process. Free quantities have a superscript of  $\infty$  while twisted quantities are denoted with calligraphic letters as in McDonald [5].

$$\begin{aligned}
\mathcal{L}^\infty g(v, z) &:= \frac{1}{h(v, z)} L^\infty [h(v, z) \cdot g(v, z)] \\
&= e^{-(1-\rho_+)v} \left( -(\mu_1 + \mu_2) \frac{\partial}{\partial v} \left( e^{(1-\rho_+)v} g(v, z) \right) \right) \\
&\quad + e^{-(1-\rho_+)v} \lambda_1 \times \\
&\quad \int_0^\infty \left[ e^{(1-\rho_+)(v+\mu_1 s)} g(v + \mu_1 s, z + s) - e^{(1-\rho_+)v} g(v, z) \right] \mu_1 \exp(-\mu_1 s) ds \\
&\quad + e^{-(1-\rho_+)v} (\lambda_2 + \gamma) \times \\
&\quad \int_0^\infty \left[ e^{(1-\rho_+)(v+\mu_2 s)} g(v + \mu_2 s, z - s) - e^{(1-\rho_+)v} g(v, z) \right] \mu_2 \exp(-\mu_2 s) ds \\
&= -(\mu_1 + \mu_2) \frac{\partial}{\partial v} g(v, z) \\
&\quad + \lambda_1 \rho_+^{-1} \int_0^\infty [g(v + \mu_1 s, z + s) - g(v, z)] \mu_1 \rho_+ \exp(-\mu_1 \rho_+ s) ds \\
&\quad + (\lambda_2 + \gamma) \rho_+^{-1} \int_0^\infty [g(v + \mu_2 s, z - s) - g(v, z)] \mu_2 \rho_+ \exp(-\mu_2 \rho_+ s) ds.
\end{aligned}$$

The final equality follows by adding and subtracting  $e^{(1-\rho_+)(v+\mu_1 s)} g(v, z)$  to the first integral and  $e^{(1-\rho_+)(v+\mu_2 s)} g(v, z)$  to the second integral in order to pull out  $g(v, z) L^\infty h(v, z) = 0$ .

Similarly for  $z < 0$

$$\begin{aligned}
\mathcal{L}^\infty g(v, z) &:= -(\mu_1 + \mu_2) \frac{\partial}{\partial v} g(v, z) \\
&\quad + (\lambda_1 + \gamma) \rho_+^{-1} \int_0^\infty [g(v + \mu_1 s, z + s) - g(v, z)] \mu_1 \rho_+ \exp(-\mu_1 \rho_+ s) ds \\
&\quad + \lambda_2 \rho_+^{-1} \int_0^\infty [g(v + \mu_1 s, z - s) - g(v, z)] \mu_2 \rho_+ \exp(-\mu_2 \rho_+ s) ds.
\end{aligned}$$

We see that the twisted process  $\mathcal{W}^\infty$  is a Markov additive process whose Markovian part  $\hat{\mathcal{W}}^\infty$  corresponds to the difference between waiting times of two queues having exponential service times with parameters  $\mu_1 \rho_+$  and  $\mu_2 \rho_+$  respectively. The additive part  $\hat{\mathcal{W}}^\infty$  corresponds to the total workload in the system. Dedicated jobs arrive at the two queues with twisted rates  $\lambda_1 \rho_+^{-1}$  and  $\lambda_2 \rho_+^{-1}$  respectively while discretionary jobs arrive at the twisted rate  $\gamma \rho_+^{-1}$ . We note that this twisted system is precisely that discovered in Foley & McDonald [4].

We now calculate the steady state of  $\hat{\mathcal{W}}^\infty$ . We make an Ansatz and suppose that the density of the steady state is given by

$$\varphi(z) = \begin{cases} \beta \alpha_+ e^{-\alpha_+ z} & \text{if } z > 0, \\ (1 - \beta) \alpha_- e^{-\alpha_- |z|} & \text{if } z < 0. \end{cases} \quad (4.16)$$

We ignore the additive component of  $\mathcal{W}^\infty$  and match the flow into and out of an infinitesimal neighbourhood of  $z > 0$ . This gives

$$\begin{aligned} & \rho_+^{-1}(\lambda_1 + \lambda_2 + \gamma)\varphi(z) \\ &= \lambda_1 \rho_+^{-1} \int_0^z \mu_1 \rho_+ \exp(-\mu_1 \rho_+(z-u))\varphi(u) du \\ & \quad + (\lambda_2 + \gamma) \rho_+^{-1} \int_z^\infty \mu_2 \rho_+ \exp(-\mu_2 \rho_+(u-z))\varphi(u) du \\ & \quad + (\lambda_1 + \gamma) \rho_+^{-1} \int_0^\infty \mu_1 \rho_+ \exp(-\mu_1 \rho_+(z+u))\varphi(-u) du. \end{aligned}$$

Substituting in (4.16), we get

$$\begin{aligned} & \rho_+^{-1}(\lambda_1 + \lambda_2 + \gamma)\beta\alpha_+ e^{-\alpha_+ z} \\ &= \lambda_1 \mu_1 e^{-\mu_1 \rho_+ z} \int_0^z \exp(u(\mu_1 \rho_+ - \alpha_+))\beta\alpha_+ du \\ & \quad + (\lambda_2 + \gamma) \mu_2 e^{\mu_2 \rho_+ z} \int_z^\infty \exp(-u(\mu_2 \rho_+ + \alpha_+))\beta\alpha_+ du \\ & \quad + (\lambda_1 + \gamma) \mu_1 e^{-\mu_1 \rho_+ z} \int_0^\infty \exp(-u(\mu_1 \rho_+ + \alpha_-))(1-\beta)\alpha_- du \end{aligned}$$

or

$$\begin{aligned} & \rho_+^{-1}(\lambda_1 + \lambda_2 + \gamma)\beta\alpha_+ e^{-\alpha_+ z} \\ &= \lambda_1 \mu_1 e^{-\mu_1 \rho_+ z} \beta\alpha_+ \frac{1}{\mu_1 \rho_+ - \alpha_+} \left( e^{(\mu_1 \rho_+ - \alpha_+)z} - 1 \right) \\ & \quad + (\lambda_2 + \gamma) \mu_2 \beta\alpha_+ e^{\mu_2 \rho_+ z} \frac{1}{\mu_2 \rho_+ + \alpha_+} e^{-(\mu_2 \rho_+ + \alpha_+)z} \\ & \quad + (\lambda_1 + \gamma) \mu_1 (1-\beta)\alpha_- e^{-\mu_1 \rho_+ z} \frac{1}{\mu_1 \rho_+ + \alpha_-} \\ &= \frac{\lambda_1 \mu_1 \beta\alpha_+}{\mu_1 \rho_+ - \alpha_+} (e^{-z\alpha_+} - e^{-\mu_1 \rho_+^{-1} z}) + \frac{(\lambda_2 + \gamma) \mu_2 \beta\alpha_+}{\mu_2 \rho_+ + \alpha_+} e^{-z\alpha_+} \\ & \quad + \frac{(\lambda_1 + \gamma) \mu_1 (1-\beta)\alpha_-}{\mu_1 \rho_+ + \alpha_-} e^{-\mu_1 \rho_+ z}. \end{aligned}$$

The only way this equation can hold for all  $z > 0$  is for the terms in  $\exp(-\alpha_+ z)$  to balance and the terms in  $\exp(-\mu_1 \rho_+ z)$  to cancel. This means that

$$\rho_+^{-1}(\lambda_1 + \lambda_2 + \gamma)\beta\alpha_+ = \frac{\lambda_1 \mu_1 \beta\alpha_+}{\mu_1 \rho_+ - \alpha_+} + \frac{(\lambda_2 + \gamma) \mu_2 \beta\alpha_+}{\mu_2 \rho_+ + \alpha_+} \quad (4.17)$$

$$\frac{\lambda_1 \mu_1 \beta\alpha_+}{\mu_1 \rho_+ - \alpha_+} = \frac{(\lambda_1 + \gamma) \mu_1 (1-\beta)\alpha_-}{\mu_1 \rho_+ + \alpha_-}. \quad (4.18)$$

And by a similar calculation for  $z < 0$ ,

$$\rho_+^{-1}(\lambda_1 + \lambda_2 + \gamma)(1-\beta)\alpha_- = \frac{(\lambda_1 + \gamma) \mu_1 (1-\beta)\alpha_-}{\mu_1 \rho_+ + \alpha_-} + \frac{\lambda_2 \mu_2 (1-\beta)\alpha_-}{\mu_2 \rho_+ - \alpha_-} \quad (4.19)$$

$$\frac{\lambda_2 \mu_2 (1-\beta)\alpha_-}{\mu_2 \rho_+ - \alpha_-} = \frac{(\lambda_2 + \gamma) \mu_2 \beta\alpha_+}{\mu_2 \rho_+ + \alpha_+}. \quad (4.20)$$

By inspection, the solution to (4.17)–(4.20) is that given in Theorem 4.1. We remark that we require Conditions 2.1 and 2.2 in order to have a steady state. We also note that the mean drift of the twisted process in the strongly pooled case is

$$\tilde{d}_1 = \mu_1 + \mu_2 - (\lambda_1 + \lambda_2 + \gamma) > 0.$$

As in McDonald [5], we define  $f_A$ ,  $f_E$  and  $f_Q$  as follows:

$$f_i := \int_{(x,y) \in \Delta} \pi_i(dx, dy) h(x, y) P_{(x,y)}(\tilde{\mathcal{W}}_1^\infty \rightarrow \infty \text{ without hitting } \Delta). \quad (4.21)$$

where  $i$  represents  $A$ ,  $E$  or  $Q$  according as JSAW, JSEW or JSQ respectively is being used.  $f_i$  is nonzero because the mean drift is positive. Similarly  $g_A$ ,  $g_E$  and  $g_Q$  are defined by

$$g_i := f_i \cdot \int_v e^{-(1-\rho_+)v} \mu(dv), \quad (4.22)$$

where  $\mu$  is the steady state of the excess ladder height process  $\tilde{\mathcal{W}}^\infty(\mathcal{T}_\ell) - \ell$ . In practice,  $g$  can be estimated using importance sampling. Since  $\pi$  and the twisted process  $\mathcal{W}^\infty$  differ according to the routing policy used,  $f$  and  $g$  do too.

We still need to check Conditions 6 and 7 of McDonald [5]. Condition 6 is shown in Section 4.2, and Condition 7 is automatic. Putting all these facts together and applying Theorem 1.6 in McDonald [5] gives (4.1). Applying Theorem 1.10 in McDonald [5] gives (4.3).

Denote the queue size of the first queue at the moment of workload overload by  $Q_1(\mathcal{T}_\ell)$  and that of the second by  $Q_2(\mathcal{T}_\ell)$ . The large deviation was produced by a sequence of i.i.d. service times  $\mathcal{X}_i^1$  having mean  $(\mu_1 \rho_+)^{-1}$  for the first queue and  $\mathcal{X}_i^2$  with mean  $(\mu_2 \rho_+)^{-1}$  for the second. Hence

$$\mathcal{S}_1(\mathcal{T}_\ell) = \sum_{i=1}^{Q_1(\mathcal{T}_\ell)} \mathcal{X}_i^1 \text{ and } \mathcal{S}_2(\mathcal{T}_\ell) = \sum_{i=1}^{Q_2(\mathcal{T}_\ell)} \mathcal{X}_i^2.$$

Since  $\mathcal{S}_1(\mathcal{T}_\ell) \sim \mathcal{S}_2(\mathcal{T}_\ell)$  and

$$\frac{1}{\ell}(\mathcal{L}_1(\mathcal{T}_\ell) + \mathcal{L}_2(\mathcal{T}_\ell)) \rightarrow 1,$$

(4.4) follows immediately. It also follows that

$$Q_1(\mathcal{T}_\ell)/\ell \rightarrow \frac{\rho_+ \mu_1}{\mu_1 + \mu_2} \text{ and } Q_2(\mathcal{T}_\ell)/\ell \rightarrow \frac{\rho_+ \mu_2}{\mu_1 + \mu_2}.$$

This gives (4.5).

We now turn to equation (4.2). Redefine  $W(t) := (Q_1(t) + Q_2(t), M(t))$ . Extend the definition of  $\pi_A$  to be the steady state of  $W$ . Also extend  $\Delta$  to be those states where one of the queues is empty (so the waiting time is zero). This process may be imbedded into a Markov additive process  $W^\infty$  with additive component  $Q_1(t) + Q_2(t)$  by removing  $\Delta$ . Note that there is no  $\tilde{W}$  component. Let  $w = (q_1 + q_2, (q_1, q_2, \vec{x}, \vec{y}))$  be some state of  $W^\infty(t)$ . As in Foley & McDonald [4],  $h(w) := \rho_+^{-q}$  is seen to be harmonic for JSQ. Now remark that  $h$  is harmonic for the JSAW and JSEW also! This means the associated twist is same; that is, identical to that derived earlier in this section. We can now apply Theorem 1.6 in McDonald [5] to get (4.2) if we can check all the conditions in McDonald [5] for this new process  $W$ . Only Condition 6 is non-trivial.

Condition 6 amounts to showing that

$$\begin{aligned} \sum_{q_1 \geq 0} \rho_+^{-q_1} \pi_A(Q_1 = q_1, Q_2 = 0) &< \infty \\ \text{and } \sum_{q_2 \geq 0} \rho_+^{-q_2} \pi_A(Q_1 = 0, Q_2 = q_2) &< \infty. \end{aligned}$$

Note that the workload at the first queue  $L_1(t)$  given the number of customers  $Q_1(t) = q_1$ ,  $Q_2(t) = q_2$  in each queue is a sum of  $q_1$  independent, exponential random variables with mean 1. Hence,

$$\begin{aligned} E_{\pi_A}(\exp((1 - \rho_+)(L_1(t) + L_2(t))) \mid Q_1 = q_1, Q_2 = 0) \\ = \left( \int_{s=0}^{\infty} \exp((1 - \rho_+)\mu_1 s) \mu_1 e^{-\mu_1 s} ds \right)^{q_1} \\ = \rho_+^{-q_1}. \end{aligned}$$

Consequently,

$$\begin{aligned} \sum_{q_1 \geq 0} \rho_+^{-q_1} \pi_A(Q_1 = q_1, Q_2 = 0) \\ = E_{\pi_A}(\exp((1 - \rho_+)(L_1(t) + L_2(t))) \chi\{Q_2 = 0\}) \\ = E_{\pi_A}(\exp((1 - \rho_+)(L_1(t))) \chi\{S_2 = 0\}) \\ = \int_0^{\infty} \pi_A(S_1 \in dx, S_2 = 0) e^{(1 - \rho_+)\mu_1 x}. \end{aligned}$$

So it remains to prove that

$$\int_0^{\infty} \pi_A(S_1 \in dx, S_2 = 0) e^{(1 - \rho_+)\mu_1 x} \equiv \int_0^{\infty} \pi_A(dx, 0) e^{(1 - \rho_+)\mu_1 x} < \infty \quad (4.23)$$

and

$$\int_0^{\infty} \pi_A(S_1 = 0, S_2 \in dy) e^{(1 - \rho_+)\mu_2 y} \equiv \int_0^{\infty} \pi_A(0, dy) e^{(1 - \rho_+)\mu_2 y} < \infty. \quad (4.24)$$

Consequently we will have Condition 6 for the harmonic function of the total queue size,  $\rho_+^{-(q_1 + q_2)}$ , when we have checked Condition 6 for the harmonic function of the total load,  $\exp((1 - \rho_+)(\mu_1 x + \mu_2 y))$ , in Section 4.2.

We now comment on the extension where the arriving workloads are not exponential. If the work brought by each customer has density  $\sigma$  then a customer routed to the first queue increases the waiting time there by an amount  $x$  with density  $\mu_1 \sigma(\mu_1 x)$ . Similarly the density of the increased waiting time  $y$  caused by a new customer at the second queue is  $\mu_2 \sigma(\mu_2 y)$ . We can generalize  $L^\infty$  by replacing  $\mu_1 \exp(-\mu_1 x)$  by  $\mu_1 \sigma(\mu_1 x)$  and  $\mu_2 \exp(-\mu_2 x)$  by  $\mu_2 \sigma(\mu_2 y)$ . It is easy to check that  $h(v, z) = \exp(\theta v)$  is harmonic for this generalized  $L^\infty$  if there exists a solution  $\alpha = \theta > 0$  to the equation

$$T(\alpha) = 1 + \rho_+^{-1} \alpha. \quad (4.25)$$

where  $T(\alpha) := \int_0^{\infty} e^{\alpha v} \sigma(v) dv$ .

$T(\alpha)$  is convex and, by hypothesis, has a derivative at  $\alpha = 0$  less than  $\rho_+^{-1}$ , the derivative of the right hand side at  $\alpha = 0$ . Consequently such a  $\theta$  exists as long as  $T(\alpha) < \infty$  for  $\alpha$  in an interval from 0 up to  $\theta$ , the point of intersection of the curves given by  $T(\alpha)$  and the right hand side of (4.25). This means that  $\sigma$  must be short-tailed. This would be the case if  $T(\alpha) < \infty$  on an interval where

$0 \leq \alpha \leq 2(\rho_+^{-1} - m_1)/m_2$  where  $m_1$  and  $m_2$  are the first two moments of  $\sigma$ . (To see this, just do a second-order Taylor expansion of  $T$ ).

With this harmonic function we can calculate the twisted kernel as above. In the twisted system each customer brings a reduced amount of work having density  $e^{\theta v} \sigma(v)/(1 + \rho_+^{-1}\theta)$  while the arrival rates are increased by a factor of  $1 + \rho_+^{-1}\theta$ . We shall not pursue this generalization further, but it is clear that if the system is pooled, the probability that the workload in the system reaches a high level  $\ell$  is roughly  $\exp(-\theta\ell)$ .

**4.2 Checking condition 6 from McDonald [5].** To check Condition 6 of McDonald [5] for the harmonic function  $\exp((1 - \rho_+)(\mu_1 x + \mu_2 y))$  amounts to showing (4.23) and (4.24). Below we will construct a Liapounov function  $v(x, y)$  which has the property that, outside a region where both  $x$  and  $y$  are small,

$$Gv(x, y) \leq -\chi\{y = 0\} \exp((1 - \rho_+)\mu_1 x) - \chi\{x = 0\} \exp((1 - \rho_+)\mu_2 y). \quad (4.26)$$

Then by a continuous version of Theorem 14.3.7 in Meyn & Tweedie [8],

$$\int_x \pi_A(dx, 0) e^{(1-\rho_+)\mu_1 x} + \int_y \pi_A(0, dy) e^{(1-\rho_+)\mu_2 y} < \infty,$$

and this gives the bounds (4.23) and (4.24), and thus proves Condition 6 of McDonald [5].

Define the function

$$V_1(x, y) = \exp\left((1 - \rho_+)\sqrt{(\mu_1 x)^2 + (\mu_2 y)^2}\right) = \exp((1 - \rho_+)r),$$

where  $r = \sqrt{(\mu_1 x)^2 + (\mu_2 y)^2}$ . Define  $z_1 = \mu_1 x/r$  and  $z_2 = \mu_2 y/r$  so  $z_1^2 + z_2^2 = 1$ . We shall assume without loss of generality that  $\mu_1 \leq \mu_2$ .

Below we shall check that for sufficiently large  $r$ ,

$$GV_1(x, y) \leq \begin{cases} -c_1 \chi\{y = 0\} \exp((1 - \rho_+)\mu_1 x) & \text{for } 0 \leq y \leq (\mu_1/\mu_2)x, \\ C_1 V_1(x, y) & \text{for } (\mu_1/\mu_2)x < y \leq x, \\ -c_2 \chi\{x = 0\} \exp((1 - \rho_+)\mu_2 y) & \text{for } 0 \leq x < y. \end{cases} \quad (4.27)$$

where  $c_1$ ,  $C_1$  and  $c_2$  are all strictly positive.

Now pick  $\beta \in (0, 1)$  such that  $z_1 + z_2 > 1/\beta$  whenever  $z_1 < z_2 \leq (\mu_2/\mu_1)z_1$  and  $z_1^2 + z_2^2 = 1$ . Multiplying by  $r$ , this implies that  $\sqrt{(\mu_1 x)^2 + (\mu_2 y)^2} < \beta(\mu_1 x + \mu_2 y)$  whenever  $(\mu_1/\mu_2)x < y \leq x$ . So if we define  $V_2(x, y) = \exp((1 - \rho_+)\beta(\mu_1 x + \mu_2 y))$ , then  $V_1(x, y) < V_2(x, y)$  for  $(\mu_1/\mu_2)x < y \leq x$ .

It is easy to check that

$$GV_2(x, y) = (1 - \rho_+) \left( -(\mu_1 \chi\{x > 0\} + \mu_2 \chi\{y > 0\})\beta + \frac{\beta(\lambda_1 + \lambda_2 + \gamma)}{1 - (1 - \rho_+)\beta} \right) V_2(x, y) \quad (4.28)$$

both for  $0 \leq x \leq y$  and  $0 \leq y \leq x$ . Define

$$C = (1 - \rho_+) \left( (\mu_1 + \mu_2)\beta - \frac{\beta(\lambda_1 + \lambda_2 + \gamma)}{1 - (1 - \rho_+)\beta} \right).$$

We can then rewrite (4.28) as

$$GV_2(x, y) = ((1 - \rho_+)(\mu_1 \chi\{x = 0\} + \mu_2 \chi\{y = 0\})\beta - C) \exp((1 - \rho_+)\beta(\mu_1 x + \mu_2 y)).$$

As a function of  $\beta$ ,  $C$  has zeros at  $\beta = 0$  and  $\beta = 1$ , and negative second derivative on the interval  $[0, 1]$ . Consequently, for all  $0 < \beta < 1$ ,  $C > 0$ .

Now define  $V(x, y) = V_1(x, y) + (C_1/C)V_2(x, y)$ . It follows that if  $(\mu_1/\mu_2)x < y \leq x$  then  $GV(x, y) \leq C_1V_1(x, y) - (C_1/C)CV_2(x, y) < 0$  since  $V_1(x, y) < V_2(x, y)$  there. Clearly  $GV(x, y) \leq 0$  if  $0 < y \leq (\mu_1/\mu_2)x$  since both  $GV_1(x, y) \leq 0$  and  $GV_2(x, y) \leq 0$  there. And

$GV(x, 0) \leq -c_1 \exp((1 - \rho_+)\mu_1 x) + (C_1/C)((1 - \rho_+)\beta\mu_2 - C) \exp((1 - \rho_+)\beta\mu_1 x)$ , and since  $\beta < 1$  it follows that  $GV(x, 0) \leq -(c_1/2) \exp((1 - \rho_+)\mu_1 x)$  for  $x$  large enough. Therefore, once we have verified (4.27), we have that for  $0 \leq y \leq x$  and  $r$  large enough,

$$GV(x, y) \leq -\frac{c_1}{2} \chi\{y = 0\} \exp((1 - \rho_+)\mu_1 x).$$

Similarly, for  $0 < x < y$ ,  $GV(x, y) \leq 0$ , since both  $GV_1(x, y) \leq 0$  and  $GV_2(x, y) \leq 0$  there. Finally,

$$GV(0, y) \leq -c_2 \exp((1 - \rho_+)\mu_2 y) + (C_1/C)((1 - \rho_+)\beta\mu_1 - C) \exp((1 - \rho_+)\beta\mu_2 y).$$

Again, since  $\beta < 1$  it follows that  $GV(0, y) \leq -(c_2/2) \exp((1 - \rho_+)\mu_2 y)$  for  $y$  large enough. Therefore, we have that for  $0 \leq x \leq y$  and  $r$  large enough,

$$GV(x, y) \leq -\frac{c_2}{2} \chi\{x = 0\} \exp((1 - \rho_+)\mu_2 y).$$

The function  $v(x, y) = 2V(x, y)/\min\{c_1, c_2\}$  now satisfies (4.26) for all non-negative  $x$  and  $y$ , and  $r$  sufficiently large.

It remains to verify (4.27). First, define

$$\begin{aligned} AV_1(x, y) &:= \int_0^\infty [V_1(x + s, y) - V_1(x, y)] \mu_1 \exp(-\mu_1 s) ds \\ BV_1(x, y) &:= \int_0^\infty [V_1(x, y + s) - V_1(x, y)] \mu_2 \exp(-\mu_2 s) ds \end{aligned}$$

and note that

$$GV_1(x, y) = \begin{cases} -\chi\{x > 0\} \frac{\partial V_1}{\partial x}(x, y) - \chi\{y > 0\} \frac{\partial V_1}{\partial y}(x, y) \\ \quad + \lambda_1 AV_1(x, y) + (\lambda_2 + \gamma) BV_1(x, y) & \text{if } x \geq y \\ -\chi\{x > 0\} \frac{\partial V_1}{\partial x}(x, y) - \chi\{y > 0\} \frac{\partial V_1}{\partial y}(x, y) \\ \quad + (\lambda_1 + \gamma) AV_1(x, y) + \lambda_2 BV_1(x, y) & \text{if } x < y. \end{cases}$$

Next,

$$\begin{aligned} AV_1(x, y) &= \int_0^\infty \left( e^{(1-\rho_+)\sqrt{(\mu_1(x+s))^2 + (\mu_2 y)^2}} - e^{(1-\rho_+)\sqrt{(\mu_1 x)^2 + (\mu_2 y)^2}} \right) \mu_1 \exp(-\mu_1 s) ds \\ &= e^{(1-\rho_+)r} \int_0^\infty \left( e^{(1-\rho_+)(\sqrt{r^2 + 2\mu_1^2 x s + \mu_1^2 s^2} - r)} - 1 \right) \mu_1 \exp(-\mu_1 s) ds \\ &= e^{(1-\rho_+)r} \int_0^\infty \left( e^{E(r)} - 1 \right) \mu_1 \exp(-\mu_1 s) ds \end{aligned}$$

where

$$E(r) := (1 - \rho_+) \left( \sqrt{r^2 + 2\mu_1 z_1 s r + \mu_1^2 s^2} - r \right).$$

Now for fixed  $z_1 \leq 1$ , the derivative of  $E(r)$  with respect to  $r$  is less than or equal to 0. Also, using l'Hôpital's rule, the limit of  $E(r)$  as  $r \rightarrow \infty$  is

$$\lim_{r \rightarrow \infty} E(r) = (1 - \rho_+) z_1 \mu_1 s.$$

Next note that  $\int_0^\infty (e^{E(r)} - 1) \mu_1 \exp(-\mu_1 s) ds < \infty$ , again using the fact that  $z_1 \leq 1$ . Therefore by the dominated convergence theorem,

$$\begin{aligned} e^{-(1-\rho_+)r} AV(x, y) &\downarrow \int_0^\infty (e^{(1-\rho_+)z_1 \mu_1 s} - 1) \mu_1 \exp(-\mu_1 s) ds \\ &= \frac{(1-\rho_+)z_1}{1-(1-\rho_+)z_1} \end{aligned}$$

as  $r \rightarrow \infty$ .

Similarly

$$e^{-(1-\rho_+)r} BV_1(x, y) \downarrow \frac{(1-\rho_+)z_2}{1-(1-\rho_+)z_2}$$

as  $r \rightarrow \infty$ . Hence as  $r \rightarrow \infty$ ,  $e^{-(1-\rho_+)r} GV_1(x, y)$  decreases to

$$\begin{aligned} &\begin{cases} (1-\rho_+)(-\mu_1 z_1 - \mu_2 z_2 + \frac{\lambda_1 z_1}{1-(1-\rho_+)z_1} + \frac{(\lambda_2 + \gamma)z_2}{1-(1-\rho_+)z_2}) & \text{if } 0 \leq y \leq x \\ (1-\rho_+)(-\mu_1 z_1 - \mu_2 z_2 + \frac{(\lambda_1 + \gamma)z_1}{1-(1-\rho_+)z_1} + \frac{\lambda_2 z_2}{1-(1-\rho_+)z_2}) & \text{if } 0 \leq x < y \end{cases} \\ = &\begin{cases} (1-\rho_+)(g_1(z_1) + g_2(z_2)) & \text{if } 0 \leq y \leq x \\ (1-\rho_+)(f_1(z_1) + f_2(z_2)) & \text{if } 0 \leq x < y \end{cases} \end{aligned}$$

where

$$\begin{aligned} g_1(z) &= -\mu_1 z + \lambda_1 \frac{z}{1-(1-\rho_+)z} \\ g_2(z) &= -\mu_2 z + (\lambda_2 + \gamma) \frac{z}{1-(1-\rho_+)z} \\ f_1(z) &= -\mu_1 z + (\lambda_1 + \gamma) \frac{z}{1-(1-\rho_+)z} \\ f_2(z) &= -\mu_2 z + \lambda_2 \frac{z}{1-(1-\rho_+)z}. \end{aligned}$$

Note that

$$\begin{aligned} g_1(1) + g_2(0) &= -(\mu_1 - \lambda_1/\rho_+) < 0 \\ f_1(0) + f_2(1) &= -(\mu_2 - \lambda_2/\rho_+) < 0 \end{aligned}$$

since  $\rho_+ > \max\{\rho_1, \rho_2\}$  by hypothesis.

We now evaluate  $g_1(z_1) + g_2(z_2)$  with  $0 < z_2 \leq z_1 < 1$ . First remark that both  $g_1(z)$  and  $g_2(z)$  are convex since both have positive second derivatives.  $g_1$  has zeros at 0 and  $(1-\rho_1)/(1-\rho_+)$  while  $g_2$  has zeros at 0 and  $(1-(\lambda_2 + \gamma)/\mu_2)/(1-\rho_+)$ . Since  $0 < z_1 < 1$  it follows that  $g_1(z_1) < 0$ .

Consider the case when  $z_1 \geq (1-(\lambda_2 + \gamma)/\mu_2)/(1-\rho_+)$ . Then  $g_2(z_1) > 0$  and  $g_2(z_2) \leq g_2(z_1)$ . Consequently

$$\begin{aligned} g_1(z_1) + g_2(z_2) &\leq g_1(z_1) + g_2(z_1) \\ &= -\mu_+ z_1 + \lambda_+ \frac{z_1}{1-(1-\rho_+)z_1} \\ &= -\frac{\mu_+}{1-(1-\rho_+)z_1} (1-\rho_+)z_1(1-z_1) \\ &< 0. \end{aligned}$$

The alternative case is when  $z_1 < (1-(\lambda_2 + \gamma)/\mu_2)/(1-\rho_+)$ . By hypothesis  $z_2 \leq z_1$  so it follows that  $z_2 < (1-(\lambda_2 + \gamma)/\mu_2)/(1-\rho_+)$ . This means that  $g_2(z_2) < 0$ . Consequently  $g_1(z_1) + g_2(z_2) < 0$ .

This means that  $g_1(z_1) + g_2(z_2) < 0$  for all  $z_1^2 + z_2^2 = 1$ ,  $0 \leq z_2 \leq z_1$ . By compactness then  $g_1(z_1) + g_2(z_2) < -\delta$  on the same region for some  $\delta > 0$ . Hence for  $r$  sufficiently large,  $\exp(-(1 - \rho_+)r)GV_1(x, y) \leq -\delta/2 =: -c_1$  for  $0 \leq z_2 \leq z_1$  or equivalently for  $0 \leq y \leq (\mu_1/\mu_2)x$ . This implies the first line of (4.27).

Take any  $\epsilon > 0$ . In the region  $z_1 < z_2 \leq (\mu_2/\mu_1)z_1$  or  $(\mu_1/\mu_2)x < y \leq x$  the above argument shows that  $GV_1(x, y) \leq C_1 \exp((1 - \rho_+)r)$  where

$$C_1 := (1 + \epsilon)(1 - \rho_+) \sup_{z_1 < z_2 \leq (\mu_2/\mu_1)z_1} (g_1(z_1) + g_2(z_2)).$$

This is the second line of (4.27).

Finally, it is easy to check that  $f_1(z_1) + f_2(z_2) < 0$  when  $0 \leq z_1 < (\mu_1/\mu_2)z_2$  (i.e. when  $0 \leq x < y$ ) since we assumed  $\mu_1 \leq \mu_2$ . We follow the same argument as above by swapping  $z_1$  for  $z_2$ ,  $g_1$  for  $f_2$  and  $g_2$  for  $f_1$ . This gives the final line of (4.27).

**4.3 Comparison with JSEW and JSQ.** We now compare the JSAW policy with JSEW and JSQ. The key observation is that the functions  $\rho_+^{q_1+q_2}$  and  $\exp((1 - \rho_+)(\mu_1 s_1 + \mu_2 s_2))$  are harmonic for all three policies (at least away from the boundaries representing empty queues).

We first apply the results in Foley & McDonald [4] to large deviations of the workload under JSQ. Redefine the process  $W(t) := (\mu_1 S_1(t) + \mu_2 S_2(t), M(t), Q_1(t) - Q_2(t))$ . Extend the definition of  $\pi_Q$  to be the steady state of  $W$ . Again extend  $\Delta$  to be those states where one of the queues is empty (so the waiting time is zero). This process may be imbedded into a Markov additive process  $W^\infty$  with additive component  $\mu_1 S_1(t) + \mu_2 S_2(t)$  by removing  $\Delta$ . The Markovian component is  $\hat{W} = Q_1(t) - Q_2(t)$ . Let  $w = (\mu_1 s_1 + \mu_2 s_2, (q_1, q_2, \vec{x}, \vec{y}), q_1 - q_2)$  be some state of  $W^\infty(t)$ . Extending the results in Section 4.1 we see that  $h(w) := \exp((1 - \rho_+)(\mu_1 s_1 + \mu_2 s_2))$  is harmonic for JSAW. Now remark that  $h$  is harmonic for JSQ also! This means that the associated twist is same; that is, identical to that derived in Section 4.1.

From Theorem 2 in Foley & McDonald [4] we see that  $\hat{W}$  has steady state  $\varphi_Q$  provided Condition 2.3 holds. In order to apply the results in McDonald [5] we need to verify Condition 6 for  $\pi_Q$  and  $h$  given above. This amounts to checking that

$$\int_0^\infty \pi_Q(S_1 \in ds_1, S_2 = 0) e^{(1 - \rho_+)\mu_1 s_1} < \infty \quad (4.29)$$

$$\text{and } \int_0^\infty \pi_Q(S_1 = 0, S_2 \in ds_2) e^{(1 - \rho_+)\mu_2 s_2} < \infty. \quad (4.30)$$

Note that under JSQ the future service times

$$(X_1(t), X_2, \dots, X_{Q_1(t)}) \text{ and } (Y_1(t), Y_2, \dots, Y_{Q_2(t)})$$

are independent of the current queue sizes  $(Q_1(t), Q_2(t))$ . Consequently,

$$\begin{aligned} & E\pi_Q(\exp((1 - \rho_+)(\mu_1 S_1(t) + \mu_2 S_2(t))) \mid Q_1(t) = q_1, Q_2(t) = 0) \\ &= \left( \int_{s=0}^\infty \exp((1 - \rho_+)\mu_1 s) \mu_1 e^{-\mu_1 s} ds \right)^{q_1} = \rho_+^{-q_1} \\ & E\pi_Q(\exp((1 - \rho_+)(\mu_1 S_1(t) + \mu_2 S_2(t))) \mid Q_1(t) = 0, Q_2(t) = q_2) \\ &= \left( \int_{s=0}^\infty \exp((1 - \rho_+)\mu_2 s) \mu_2 e^{-\mu_2 s} ds \right)^{q_2} = \rho_+^{-q_2}. \end{aligned}$$

Therefore, checking (4.29) and (4.30) is equivalent to checking

$$\sum_{q_1=0}^{\infty} \pi_Q(Q_1 = q_1, Q_2 = 0) \rho_+^{-q_1} \text{ and } \sum_{q_2=0}^{\infty} \pi_Q(Q_1 = 0, Q_2 = q_2) \rho_+^{-q_2},$$

but this was shown in Foley & McDonald [4] as long as Conditions 2.1 and 2.2 hold.

Applying Theorem 1.5 in McDonald [5] we get equation (4.6) in Theorem 4.2. (4.7) was proved in Foley & McDonald [4]. Applying Theorem 1.10 in McDonald [5] we get (4.8). Recall that the large deviation was produced by a sequence of i.i.d. service times  $\mathcal{X}_i^1$  having mean  $(\mu_1 \rho_+)^{-1}$  for the first queue and  $\mathcal{X}_i^2$  with mean  $(\mu_2 \rho_+)^{-1}$  for the second. Since  $\mathcal{Q}_1(\mathcal{T}_\ell) \sim \mathcal{Q}_2(\mathcal{T}_\ell)$  and  $(\mu_1 \mathcal{S}_1(\mathcal{T}_\ell) + \mu_2 \mathcal{S}_2(\mathcal{T}_\ell))/\ell \rightarrow 1$ , equation (4.10) follows, and from that so does (4.9).

We now sketch the results associated with JSEW. First of all, we show that stability follows from Condition 2.1. It suffices to consider the Liapounov function  $V(q_1, q_2) = q_1^2/\mu_1 + q_2^2/\mu_2$ . If  $G$  is the generator of the jump process  $(Q_1, Q_2)$  under JSEW and  $0 < q_1/\mu_1 < q_2/\mu_2$  then

$$\begin{aligned} GV(q_1, q_2) &= (\lambda_1 + \gamma) \frac{(q_1 + 1)^2 - q_1^2}{\mu_1} + \lambda_2 \frac{(q_2 + 1)^2 - q_2^2}{\mu_2} \\ &\quad + \mu_1 \frac{(q_1 - 1)^2 - q_1^2}{\mu_1} + \mu_2 \frac{(q_2 - 1)^2 - q_2^2}{\mu_2}. \end{aligned}$$

Hence

$$\begin{aligned} GV(q_1, q_2) &= 2 \left( (\lambda_1 + \gamma - \mu_1) \frac{q_1}{\mu_1} + (\lambda_2 - \mu_2) \frac{q_2}{\mu_2} \right) + \left( \frac{\lambda_1 + \gamma}{\mu_1} + \frac{\lambda_2}{\mu_2} + 2 \right) \\ &\leq 2 \frac{q_1}{\mu_1} (\lambda_1 + \lambda_2 + \gamma - \mu_1 - \mu_2) + c, \end{aligned}$$

where  $c = (\lambda_1 + \gamma)/\mu_1 + \lambda_2/\mu_2 + 2$ . For  $q_1$  large enough, this is negative by Condition 2.1. The cases when  $q_1 = 0$  or  $q_2 = 0$  or  $q_1/\mu_1 \geq q_2/\mu_2$  follow similarly.

We now sketch the proof of the results in Theorem 4.3. For more results consult Coombs-Reyes [1]. Redefine the process  $W(t) := (\mu_1 S_1(t) + \mu_2 S_2(t), M(t), Q_1(t)/\mu_1 - Q_2(t)/\mu_2)$ . Extend the definition of  $\pi_E$  to be the steady state of  $W$ . Again extend  $\Delta$  to be those states where one of the queues is empty (so the waiting time is zero). This process may be imbedded into a Markov additive process  $W^\infty$  with additive component  $\mu_1 S_1(t) + \mu_2 S_2(t)$  by removing  $\Delta$ . The Markovian component is the difference of the expected waiting times  $\hat{W} = Q_1(t)/\mu_1 - Q_2(t)/\mu_2$ . This process is rather awkward since the state space may be some subgroup of the reals.

Let  $w = (\mu_1 s_1 + \mu_2 s_2, (q_1, q_2, \vec{x}, \vec{y}), q_1/\mu_1 - q_2/\mu_2)$  be some state of  $W^\infty(t)$ . Again  $h(w) := \exp((1 - \rho_+)(\mu_1 s_1 + \mu_2 s_2))$  is harmonic for JSAW and for JSEW. This means the associated twist is same; that is, identical to that derived in Section 4.1. Hence the arrival rates are multiplied by  $\rho_+^{-1}$  and the service rates are multiplied by  $\rho_+$ .

The strong pooling of the expected waiting times will hold if  $\hat{W}$  is stable. This follows assuming Condition 2.2 (but not necessarily Condition 2.3) since the mean

drift for  $\hat{W}$  when  $q_1/\mu_1 > q_2/\mu_2$  is

$$\begin{aligned} & \rho_+^{-1} \lambda_1 \frac{1}{\mu_1} + \rho_+ \mu_2 \frac{1}{\mu_2} - \rho_+^{-1} (\lambda_2 + \gamma) \frac{1}{\mu_2} - \rho_+ \mu_1 \frac{1}{\mu_1} \\ &= \rho_+^{-1} \left( \frac{\lambda_1}{\mu_1} - \frac{\lambda_2 + \gamma}{\mu_2} \right), \end{aligned}$$

and this is negative if Condition 2.2 holds. Similarly, the mean drift for  $\hat{W}$  is positive when  $q_1/\mu_1 < q_2/\mu_2$ .

In order to apply the results in McDonald [5] we need to verify Condition 6 of that paper for  $\pi_E$  and  $h$  given above. This amounts to checking that

$$\int_0^\infty \pi_E(S_1 \in ds_1, S_2 = 0) e^{(1-\rho_+)\mu_1 s_1} < \infty \quad (4.31)$$

$$\text{and } \int_0^\infty \pi_E(S_1 = 0, S_2 \in ds_2) e^{(1-\rho_+)\mu_2 s_2} < \infty. \quad (4.32)$$

Note that under JSEW, like under JSQ, the vectors

$$(X_1(t), X_2, \dots, X_{Q_1(t)}) \text{ and } (Y_1(t), Y_2, \dots, Y_{Q_2(t)})$$

of future service times are independent of the current queue sizes  $(Q_1(t), Q_2(t))$ . Consequently,

$$\begin{aligned} & E_{\pi_E}(\exp((1-\rho_+)(\mu_1 S_1(t) + \mu_2 S_2(t))) \mid Q_1(t) = q_1, Q_2(t) = 0) \\ &= \left( \int_{s=0}^\infty \exp((1-\rho_+)\mu_1 s) \mu_1 e^{-\mu_1 s} ds \right)^{q_1} = \rho_+^{-q_1} \\ & E_{\pi_E}(\exp((1-\rho_+)(\mu_1 S_1(t) + \mu_2 S_2(t))) \mid Q_1(t) = 0, Q_2(t) = q_2) \\ &= \left( \int_{s=0}^\infty \exp((1-\rho_+)\mu_2 s) \mu_2 e^{-\mu_2 s} ds \right)^{q_2} = \rho_+^{-q_2}. \end{aligned}$$

Therefore, checking (4.31) and (4.32) is equivalent to checking

$$\sum_{q_1=0}^\infty \pi_E(Q_1 = q_1, Q_2 = 0) \rho_+^{-q_1} \text{ and } \sum_{q_2=0}^\infty \pi_E(Q_1 = 0, Q_2 = q_2) \rho_+^{-q_2},$$

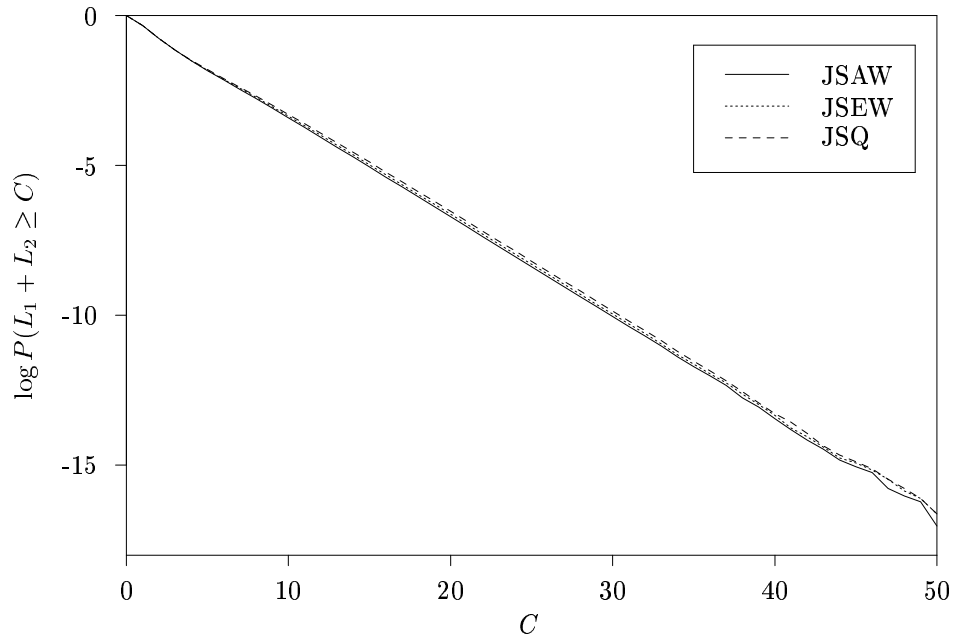
but this was shown in Coombs-Reyes [1] as long as Conditions 2.1 and 2.2 hold.

Applying Theorem 1.6 in McDonald [5] we get (4.11) in Theorem 4.3. (4.12) follows similarly. Applying Theorem 1.10 in McDonald [5] we get (4.13). Recall that the large deviation was produced by a sequence of i.i.d. service times  $\mathcal{X}_i^1$  having mean  $(\mu_1 \rho_+)^{-1}$  for the first queue and  $\mathcal{X}_i^2$  with mean  $(\mu_2 \rho_+)^{-1}$  for the second. Since  $Q_1(\mathcal{T}_\ell)/\mu_1 \sim Q_2(\mathcal{T}_\ell)/\mu_2$  and  $(\mu_1 S_1(\mathcal{T}_\ell) + \mu_2 S_2(\mathcal{T}_\ell))/\ell \rightarrow 1$  then by the law of large numbers equation (4.15) follows, and from that so does (4.14).

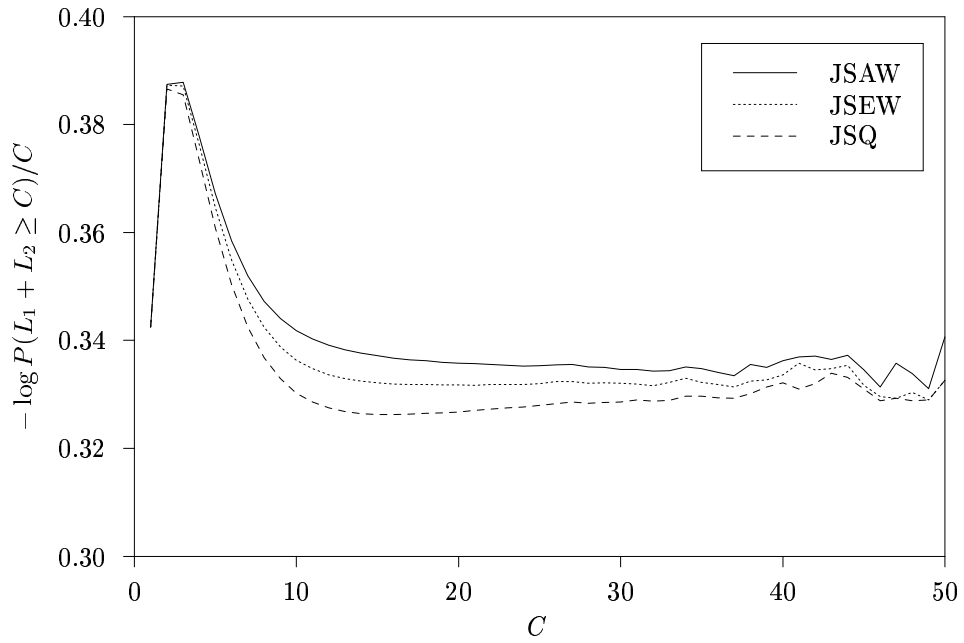
## 5 Simulations

In this section, we report the results of some simulations of the model. First,  $10^8$  coupled excursions were simulated for each of the three policies, with the following parameters:  $\lambda_1 = \lambda_2 = 0$ ,  $\gamma = 2$ ,  $\mu_1 = 1$  and  $\mu_2 = 2$ . For each excursion, the maximum workload attained was recorded. Thus this simulates the overflow event of Theorems 3.1 and 3.3 and Section 4.

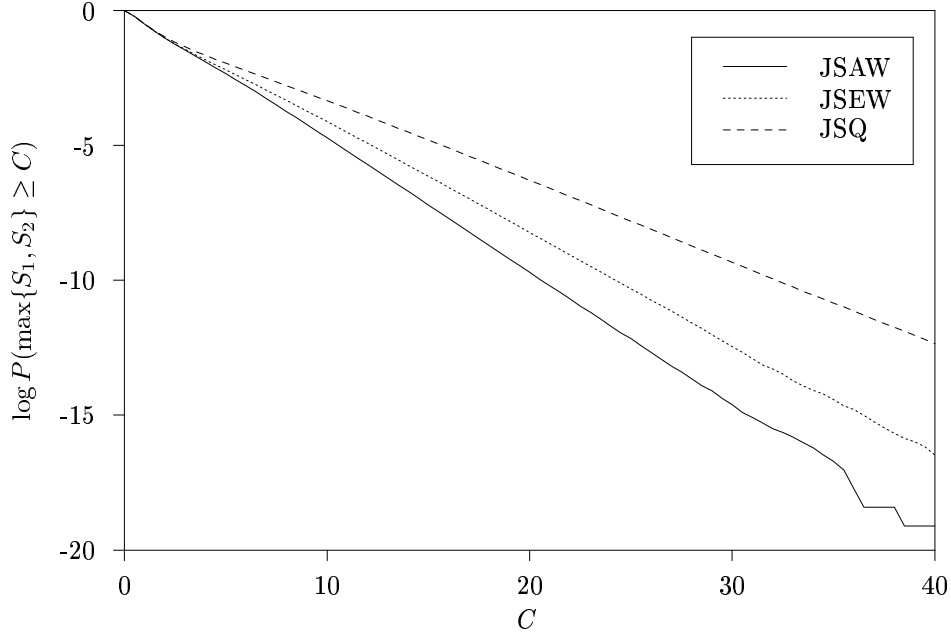
Figure 1 shows the log-probability of the event that the maximum workload exceeds each value. Figure 2 shows the implied large deviations rate. This rate is close to the predicted value of  $1/3$  for all three policies. We see from both graphs



**Figure 1** The empirical log-probability of an excessive total workload during an excursion.



**Figure 2** The empirical large deviations rate of an excessive total workload during an excursion.



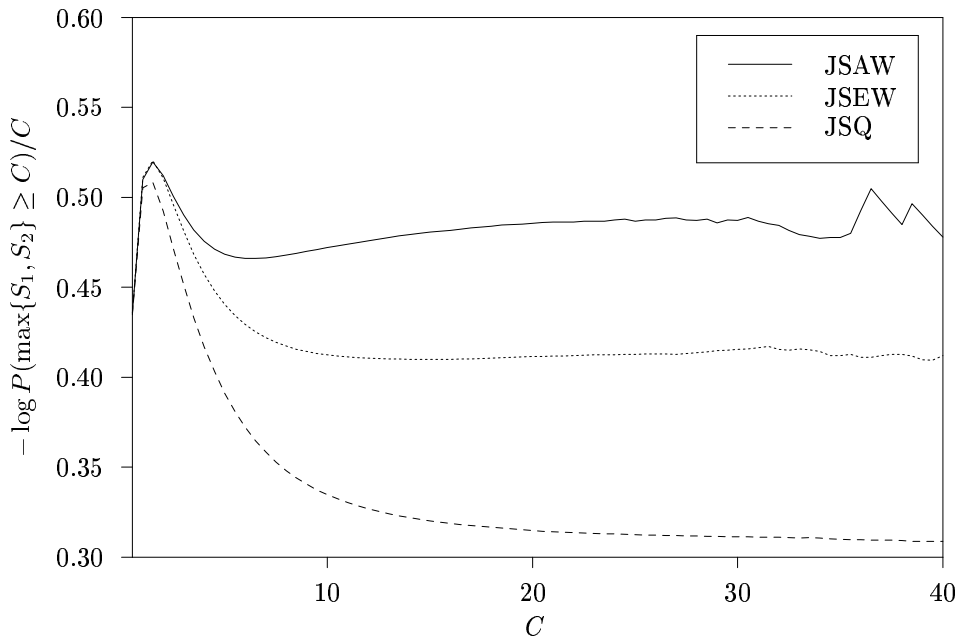
**Figure 3** The empirical log-probability of an excessive waiting time in either queue during an excursion.

that the policies all have very similar overflow probabilities, in spite of the fact that all of the traffic is routeable (and thus potentially routed differently by the three policies).

However, as we have already discovered analytically, the observation that the policies are almost identical in overflow probability is dependent on which overflow event is used. We also carried out some simulations in which the overflow event of interest was that of Theorems 3.2 and 3.4, that either waiting time became large. These results are shown in Figures 3 and 4 for the case  $\lambda_1 = \lambda_2 = 0$ ,  $\gamma = 2.5$ ,  $\mu_1 = 1$  and  $\mu_2 = 2$ , with  $2 \times 10^8$  excursions for each policy. These figures show that in this case, the JSAW policy was substantially better than JSEW, which was in turn substantially better than JSQ. And for these parameters, the results are close to the bounds given in Theorems 3.2 and 3.4, namely a large deviations rate for JSAW of  $\mu_+(1 - \rho_+) = 0.5$ , for JSEW of  $\mu_1(1 - \rho_+^{\mu_+/\mu_1}) \approx 0.42$ , and for JSQ of  $\mu_1(1 - \rho_+^2) \approx 0.31$ .

An intuitive way of understanding these results is as follows. Under any policy, the total workload in the system behaves exactly like that in the pooled system described at the beginning of Section 3, provided only that the process stays away from the axes so that neither server is ever idle. Thus any reasonably sensible routing rule will achieve essentially the same probability of the *total* workload becoming large.

However, if the overflow event is that *either* waiting time becomes large, then it is not sufficient to stay away from the axes: it now becomes important to stay close to the diagonal of equal waiting times. If the process strays from this diagonal, then the overflow set may be reached by a shorter path. The JSAW policy keeps the



**Figure 4** The empirical large deviations rate of an excessive waiting time in either queue during an excursion.

waiting times very nearly equal; the JSEW policy only keeps the expected waiting times equal; and the JSQ policy does not even do that. In the terminology of Foley & McDonald [4], the actual waiting times are strongly pooled under JSAW; weakly pooled about equality under JSEW; and weakly pooled about another ratio under JSQ.

In conclusion, we see from the analysis and the simulations that the JSAW policy sometimes has a substantial advantage over JSEW and JSQ, depending on exactly which overflow event is of interest. Furthermore, using the ‘simultaneous queueing’ algorithm of Section 1, it can be implemented without knowledge of the customers’ service requirements, the servers’ service rates, or even the current queue lengths. We therefore expect that simultaneous queueing may find application in distributed queueing networks, such as call centres.

## References

- [1] Coombs-Reyes, J. Ph.D. thesis, ISYE at Georgia Institute of Technology. Preprint, May 1999.
- [2] Dupuis, P. and Ellis, R. S. *The large deviation principle for a general class of queueing systems, I*. Trans. Amer. Math. Soc., **347** (1995), 2689–2751.
- [3] Feller, W. *An Introduction to Probability Theory and its Applications*, vol. II, 2nd ed. Wiley, 1971.
- [4] Foley, R. and McDonald, D. R. *Join the shortest queue: stability and exact asymptotics*. Submitted Ann. Appl. Probab., March 1998.
- [5] McDonald, D. R. *Asymptotics of first passage times for random walk in an orthant*. Ann. Appl. Probab., **9** (1999), 110–145.
- [6] Shwartz, A. and Weiss, A. *Large Deviations for performance analysis: queues, communications and computing*. Chapman & Hall, 1995.

- [7] Turner, S. R. E. *Large deviations for join the shorter queue*. Fields Institute Communications, **28**, 93–106.
- [8] Meyn, S. P. and Tweedie, R. L. *Markov Chains and Stochastic Stability*. Springer Verlag, 1993.