

Large Deviations for Join the Shorter Queue

S. R. E. Turner

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

Abstract. We consider large deviations results for a network of two queues in which some customers may join the shorter queue at arrival time. The queues may be $\cdot/M/\infty$ or $\cdot/M/1$ queues. We consider the case in which the routeable customers join the queue with the shorter expected waiting time, as well as that in which they join the queue with the fewer customers. This paper is partly a survey of existing results, and partly new material. One distinctive feature of our presentation is the phrasing of the results in terms of resource pooling.

1 Introduction

The subject of large deviations is several decades old, but has undergone considerable growth in recent years, fuelled both by theoretical advances and by its applicability to engineering problems. Failure events in many electronic systems are now extremely rare, and can be conveniently modelled using large deviations theory [6, 18].

In this paper we consider a Join the Shorter Queue model in the large deviations limit. The model was chosen as one of the simplest models to involve a routing decision. However, even though it is a very simple model, the theory has not been available to analyse it until recently.

The model is as follows. There are two queues, which may be $\cdot/M/1$ or $\cdot/M/\infty$ queues. There are three streams of customers arriving at these queues: one stream, called a *dedicated stream*, arrives at each queue and must be served by that queue; the other, called the *discretionary stream*, may be routed at arrival time to either of the queues. The queueing discipline is FIFO. This model is applicable to a wide range of queueing applications. For example, one can think of a pair of call

2000 *Mathematics Subject Classification.* Primary 60F10, 60K25.

The author was supported first by EPSRC studentship 93002211, and subsequently by Sidney Sussex College, Cambridge. The work was completed while the author was visiting the University of Ottawa, partly supported by NSERC grant A4551.

centres with customers allocated geographically, or of two buffers inside a switch in a network.

We shall make the following mathematical assumptions on the model. The arrival streams are independent Poisson streams, with rates λ_1 and λ_2 for the dedicated streams and ν for the discretionary stream. All service times are independent of each other and of the arrivals processes, and are exponentially distributed. In the case of $\cdot/M/1$ queues, we shall assume that the service times at the two queues have means μ_1 and μ_2 . In the case of $\cdot/M/\infty$ queues it is more natural, in view of the canonical application to telephone networks, to assume that all service times have mean 1. We shall define $\lambda_+ = \lambda_1 + \lambda_2 + \nu$, $\mu_+ = \mu_1 + \mu_2$, $\rho_1 = \lambda_1/\mu_1$, $\rho_2 = \lambda_2/\mu_2$, and $\rho_+ = \lambda_+/\mu_+$.

In the case of $\cdot/M/1$ queues, if $\mu_1 = \mu_2$ the discretionary customers will join the shorter queue at their arrival time. If $\mu_1 \neq \mu_2$, there are two possible generalisations of this policy. The first is to continue to join the shorter queue, that is the queue with the fewer customers. This is the natural policy if arriving customers do not know the service rates of the two queues. However, if arriving customers know those service rates then the natural policy is to join the queue with the smaller expected waiting time. We refer to these two policies as *Join the Shorter Queue (JSQ)* and *Join the Smaller Expected Waiting Time (JSEW)* respectively. (A third policy, in which the service requirement of each customer is known in advance and each arriving customer joins the queue with the smaller *actual* waiting time, is analysed in [16] in this volume.)

As one of the simplest models involving a routing decision, this model has a long history, although most of the analysis has been in the heavy traffic limit, that is as $\lambda_+ \rightarrow \mu_+$, rather than in a large deviations régime. Foschini & Salz [11] and Reiman [17] showed that in the heavy traffic limit with $\mu_1 = \mu_2$, provided that in the limit $\nu > |\lambda_1 - \lambda_2|$, then the queue lengths, suitably scaled, remain equal for all time. Reiman calls this *state space collapse*, because in the limit the queue length process is confined to a low-dimensional subspace of the whole space. It implies the *resource pooling* property, that the total number of customers in the system behaves as if the system were a single queue with the total arrival rate λ_+ and the total service rate μ_+ . Resource pooling is an important property of networks in general: it means that no part of the network will become overloaded unless the whole system cannot cope with the total demand, an event which we could not do anything about without increasing capacity or decreasing demand. In terms of the performance of this particular system, the average delay for a customer is halved, compared with a policy of random routing.

Kelly & Laws [14] obtain the same resource pooling consequences without state space collapse by using a threshold routing rule, in which routing is random unless the length of either queue falls below a certain threshold. This shows that the key feature which enables resource pooling is that neither server is idle while there is still work to do in the system. Turner [21] observes partial resource pooling in the borderline case, in which in the limit $\nu = |\lambda_1 - \lambda_2|$. Kelly & Laws [14] also provide a review of the work on resource pooling in the heavy traffic limit, with many references. Harrison & Van Mieghem [12, 13] have recently united the many observations of state space collapse by explaining the phenomenon in algebraic terms, in terms of the network data (see also Bramson [5]).

There are few papers looking at join the shorter queue models in the large deviations limit. The reason is partly the scarcity of large deviations theory concerning Markov processes with jump rates which vary discontinuously. But recently several authors, most notably Dupuis & Ellis [7, 8, 9], have started to develop such a theory, which is of considerable engineering importance.

Alanyali & Hajek [2] considered the $\cdot/M/\infty$ version of our model in the case that both queues have the same capacity. They use theory which they themselves developed [1] based on earlier work of Blinovskii & Dobrushin [4]. We shall describe and use this theory in Section 3.

There are certain régimes of interest in the JSQ version of the $\cdot/M/1$ model which are not present in the $\cdot/M/\infty$ model considered by Alanyali & Hajek. Some of these régimes have been identified by Foley & McDonald [10]. While their results imply large deviations results, they obtain them through a different approach, namely the method of h -transforms using results developed by McDonald [15]. This approach yields stronger results, but is not immediately applicable to the JSEW policy or the $\cdot/M/\infty$ model. Therefore it is valuable to obtain their results again through large deviations techniques.

In this paper, we shall collect the results of Alanyali & Hajek, and Foley & McDonald, as well as some new results of our own. We shall then be able to compare the behaviour of the $\cdot/M/1$ and $\cdot/M/\infty$ models, as well the JSEW and JSQ policies. One distinctive feature of our presentation will be the phrasing of the results in terms of resource pooling, a concept more commonly associated with the heavy traffic limit. This continues the program begun in [19] (see also [20, 21]), where some of these results appeared in a preliminary form.

The outline of the paper is as follows. In Sections 2 and 3 we give results for the JSEW and JSQ policies respectively. In Section 4 we describe how the behaviour of $\cdot/M/\infty$ queues compares with that of $\cdot/M/1$ queues. Finally in Section 5 we comment on the extension to more than two queues.

We shall assume knowledge of key concepts from the theory of large deviations for processes. The reader is referred to Shwartz & Weiss [18] for a text book on this subject. In making large deviations statements, we shall for conciseness assume implicitly that the queue sizes have undergone the usual large deviations scaling, in which jumps are n times as fast and $1/n$ times as large as normal, and n tends to infinity. In other words, if the unscaled queue sizes are given by $Q(t)$, then we set $X^n(t) = Q(nt)/n$ and let n tend to infinity. By the *large deviations path* we shall mean the path which minimises the large deviations rate over a given set of paths. If the large deviations path is unique then, conditional on the path of X^n being in the given set, it lies in any small tube about the large deviations path with a probability approaching 1 as the limit is taken.

2 Join the smaller expected waiting time

In this section we shall examine the system under the Join the Shorter Expected Waiting Time policy. Recall that under this policy, we assume that an arriving discretionary customer knows the service rates of the two queues, and joins the queue with the smaller expected waiting time, that is, the queue with the smaller value of X_i/μ_i .

The main difficulty in finding the large deviations rate for an event of interest in any Join the Shorter Queue type of model is knowing whether a large deviations

principle even holds for the system, given the discontinuous jump rates. However, recent theoretical advances by Dupuis & Ellis [7, 8, 9] have now answered this question in the affirmative for all the models in which we shall be interested. Given the existence of a large deviations principle, we shall then be able to find the large deviations rate for an overflow event by simple comparisons with known large deviations rates for the single queue.

Alanyali & Hajek [2] have proved similar results for the $M/M/\infty$ system using a more general theory. We shall review their work in Section 3, but in this section we shall use a different method which emphasises the formulation in terms of resource pooling.

First recall the following results for a single $M/M/1$ queue receiving arrivals at rate λ and serving at rate $\mu > \lambda$ (see for example [6, 18]). Under the usual large deviations scaling, the local large deviations rate for travel at velocity y is

$$L_{\lambda,\mu}(y) = y \log \left(\frac{y + \sqrt{y^2 + 4\lambda\mu}}{2\lambda} \right) + \lambda + \mu - \sqrt{y^2 + 4\lambda\mu}. \quad (2.1)$$

For sufficiently large fixed T , the large deviations rate to go from occupancy k to occupancy $C > k$ in time at most T is then

$$\begin{aligned} I_{\lambda,\mu}(k, C) &:= (C - k) \inf\{L_{\lambda,\mu}(y)/y : y > 0\} \\ &= (C - k) \log \left(\frac{\mu}{\lambda} \right). \end{aligned} \quad (2.2)$$

This is also the large deviations rate to overflow level C before returning to 0, but we shall use the fixed T formulation because it fits in with the theoretical framework of Dupuis & Ellis [8], and also because it has more engineering significance.

We shall make the following assumption on the parameters of the model.

Assumption 2.1 There exists a $0 < \beta < 1$ such that

$$\frac{(\lambda_1 + \beta\nu)}{\mu_1} = \frac{(\lambda_2 + (1 - \beta)\nu)}{\mu_2}.$$

There are other equivalent formulations of this condition. This formulation allows us to state the condition in words as follows: that the traffic intensities at the two queues can be equalised by a *proper* allocation of the discretionary traffic, meaning an allocation which does not send all the discretionary traffic to just one of the queues. Without this assumption, the expected waiting times could not be kept equal by any routing scheme, so the two queues would effectively act independently during an overflow.

We shall first find the large deviations rate for the event that the total occupancy of the system exceeds some constant μ_+C by some large fixed time T . (For fixed C , any sufficiently large T will give the same answer: the factor μ_+ is just for notational convenience). As we shall see later, this calculation is also an important step towards calculating the large deviations rate for other overflow events.

Theorem 2.2 *Under Assumption 2.1, the large deviations rate to reach the set $X_1 + X_2 \geq \mu_+C$ by time T is $I_{\lambda_+,\mu_+}(0, \mu_+C)$, and the large deviations path travels along the diagonal $X_1/\mu_1 = X_2/\mu_2$.*

Proof First note that we can couple our system with a single $M/M/1$ queue with arrival rate λ_+ and departure rate μ_+ . We shall call this single queue the *pooled system*. Under the coupling, the arrival times remain the same in both

systems, and the departure times are the same except that a departure can never occur from an empty queue: thus the total occupancy in our system is always at least as much as the total occupancy in the pooled system. This immediately shows that the large deviations rate for an overflow from total occupancy μ_+k to μ_+C is at most $I_{\lambda_+, \mu_+}(\mu_+k, \mu_+C)$: with $k = 0$, this is the amount in the statement of the theorem.

Also note that if the large deviations path does not touch the axes $X_1 = 0$ or $X_2 = 0$ during the overflow then equality holds in the previous paragraph, because the total occupancy of our system, and the transition rates of the total occupancy, are then the same as those of the pooled system throughout the overflow. We shall now prove that the path does indeed not touch the axes.

Say that a path is of *class* k if the last point which it touches along the diagonal is (μ_1k, μ_2k) : formally, if

$$X_1(\sup\{t : X_1(t)/\mu_1 = X_2(t)/\mu_2\}) = \mu_1k.$$

Suppose that the large deviations path to reach $X_1 + X_2 \geq \mu_+C$ is of class $k < C$. Furthermore, suppose that it reaches $X_1 + X_2 \geq \mu_+C$ at the point $(a, \mu_+C - a)$, with $a > \mu_1C$. The key observation is that after the process has left the diagonal, all the discretionary traffic is routed to the second queue. Therefore during this part of the overflow, the queues are independent $M/M/1$ queues, and we can use the results on the single queue to calculate the large deviations behaviour for this tail portion of the overflow path.

It is clear that reaching $(a, \mu_+C - a)$ is a lower probability event than the first queue reaching a and the second queue reaching $\mu_+C - a$, not necessarily at the same time. Therefore the large deviations rate for reaching $(a, \mu_+C - a)$ from (μ_1k, μ_2k) satisfies

$$\begin{aligned} I &\geq I_{\lambda_1, \mu_1}(\mu_1k, a) + I_{\lambda_2 + \nu, \mu_2}(\mu_2k, \mu_+C - a) \\ &= (a - \mu_1k) \log\left(\frac{\mu_1}{\lambda_1}\right) + (\mu_+C - a - \mu_2k) \log\left(\frac{\mu_2}{\lambda_2 + \nu}\right). \end{aligned} \quad (2.3)$$

The derivative of (2.3) with respect to a is always positive by Assumption 2.1, so substituting $a = \mu_1C$ into (2.3), we get

$$\begin{aligned} I &\geq \mu_1(C - k) \log\left(\frac{\mu_1}{\lambda_1}\right) + \mu_2(C - k) \log\left(\frac{\mu_2}{\lambda_2 + \nu}\right) \\ &> \mu_+(C - k) \log\left(\frac{\mu_+}{\lambda_+}\right) \\ &= I_{\lambda_+, \mu_+}(\mu_+k, \mu_+C), \end{aligned} \quad (2.4)$$

where (2.4) follows from the concavity of the log function. But by the pooling argument, we have already shown that the large deviations rate for the overflow event is at most this amount. Thus this path cannot be the large deviations path for the overflow.

A parallel proof applies for a path of class $k < C$ reaching a point $(a, \mu_+C - a)$, with $a < \mu_1C$. Thus we conclude that the overflow path to $X_1 + X_2 \geq \mu_+C$ must be of class C . Since this applies for all lower C too, the path must travel up the diagonal. Finally, by the observation in the second paragraph of this proof, the large deviations rate must be as in the statement of the theorem. \square

Intuitively, Theorem 2.2 says that provided Assumption 2.1 holds, the large deviations rate to reach the set $X_1 + X_2 \geq \mu_+ C$ is the same as the bound implied by the pooling calculation. This means that an overflow does not occur unless the whole system cannot cope with the total amount of traffic arriving, an event which we could not hope to do anything about even if we could reroute customers while they were queueing. In this case we say that the system is *pooled*.

Note that the proof of this theorem did not use any ‘high level’ knowledge about the expected form of the overflow path. For example, in the next section we shall use Lemma 5.16 of Shwartz & Weiss [18] to show that the overflow path must be piecewise linear. But in the above proof, we only used simple bounds on the large deviations rate based on the known result for overflows of a single queue.

In many applications, it is more interesting to know the large deviations rate for the event that either queue becomes large, rather than that the total of the queue lengths becomes large. For example, this would be the overflow event of interest if we were concerned about overflowing some finite buffers, or about not letting any customer’s expected waiting time become too long. We have the following theorem.

Theorem 2.3 *The large deviations rate to reach the set $\{X_1 \geq \mu_1 C$ or $X_2 \geq \mu_2 C\}$ by time T is*

$$\begin{aligned} I_{\lambda_+, \mu_+}(0, \mu_+ C) & \text{ if } \nu > \text{ a certain } \nu_0 \\ I_{\lambda_1, \mu_1}(0, \mu_1 C) & \text{ if } \nu < \nu_0 \text{ and } -\mu_1 \log \rho_1 < -\mu_2 \log \rho_2 \\ I_{\lambda_2, \mu_2}(0, \mu_2 C) & \text{ if } \nu < \nu_0 \text{ and } -\mu_1 \log \rho_1 > -\mu_2 \log \rho_2. \end{aligned}$$

In the first case, the large deviations path travels along the diagonal, and in the other cases it travels along the axis corresponding to the lower value of $-\mu_i \log \rho_i$.

Proof Suppose that the large deviations path is of class k , in the sense defined above. Then the overflow can be divided into two stages: first to $(\mu_1 k, \mu_2 k)$, and then to either $X_1 \geq \mu_1 C$ or $X_2 \geq \mu_2 C$. By the Markov property, the large deviations rate for the whole path is the sum of the large deviations rates for the two stages. For the first stage, the cheapest path is to pass along the diagonal, by Theorem 2.2. During the second stage, the two queues are independent. Furthermore, the non-overflowing queue can at that time follow its most likely behaviour, corresponding to a large deviations rate of 0. The overflowing queue is the one which has lower large deviations rate to overflow from the point $(\mu_1 k, \mu_2 k)$, that is the one with lower

$$(C - k)\mu_i \log \left(\frac{\mu_i}{\lambda_i} \right). \quad (2.5)$$

Suppose that it is the i th queue which minimises (2.5). Then the whole overflow has rate

$$\mu_+ k \log \left(\frac{\mu_+}{\lambda_+} \right) + (C - k)\mu_i \log \left(\frac{\mu_i}{\lambda_i} \right). \quad (2.6)$$

Now consider minimising this expression over $k \in [0, C]$. Since the expression is linear in k , the minimum is at 0 or C according to the sign of

$$\mu_+ \log \left(\frac{\mu_+}{\lambda_+} \right) - \mu_i \log \left(\frac{\mu_i}{\lambda_i} \right).$$

If ν is sufficiently large, then this derivative is negative and the optimal path is the one of class C , travelling along the diagonal. If ν is small, then the optimal path

is of class 0, travelling along the i th axis. This gives the rates in the statement of the theorem.

Borderline cases have a continuity of paths with the same large deviations rate. They can take either behaviour in the large deviations limit, or some intermediate behaviour. \square

If the overflow passes along the diagonal, then the large deviations rate is the same as that for the pooled system. Again, an overflow does not then occur unless the whole system cannot cope with the total amount of traffic arriving. If, on the other hand, ν is too small, then there is not enough flexibility in routing. A busy period at one queue cannot be offloaded sufficiently to the other queue, and the one queue overflows with the other queue still idle, so that the process travels along the axis. This also occurs if ν is so small that Assumption 2.1 does not hold.

3 Join the shorter queue

In this section, we shall first describe the theory of Alanyali & Hajek [1]. We shall then use it to analyse the JSQ policy for $\cdot/M/1$ queues.

Let $A^+ = \{x \in \mathbb{R}^d : x_1 > 0\}$, $A^- = \{x \in \mathbb{R}^d : x_1 < 0\}$ and $A^0 = \{x \in \mathbb{R}^d : x_1 = 0\}$. Say that a Markov process on \mathbb{R}^d evolves according to a measure ν if when it is in state x it next jumps after time exponentially distributed with parameter $\nu(x, \mathbb{R}^d)$, and the jump size is in $A \subseteq \mathbb{R}^d$ with probability $\nu(x, A)/\nu(x, \mathbb{R}^d)$.

Now suppose that we are given a Markov process Q on \mathbb{R}^d which has jump rates which are continuous in each half of the state space in the following sense. From $x \in A^-$, Q evolves according to the measure $\nu^-(x, \cdot)$, and from $x \in A^+ \cup A^0$, Q evolves according to the measure $\nu^+(x, \cdot)$. Each of ν^+ and ν^- is continuous in the sense that for all $\epsilon > 0$ there exists a $\delta > 0$ such that $|x - x'| < \delta$ implies $d\nu(x)/d\nu(x') \leq (1 + \epsilon)$. We also require the communicability condition, that for each $x \in A^-$, $\nu^-(x, A^+) > 0$, and for each $x \in A^+ \cup A^0$, $\nu^+(x, A^-) > 0$: in other words, that positive jumps are possible from all negative positions and *vice versa*. Technical conditions on the boundedness and tail decay of ν^+ and ν^- are also required. Then the following result holds.

For $\sim \in \{+, -\}$, define

$$\begin{aligned} M^\sim(x, \zeta) &= \int_{\mathbb{R}^d} (e^{z\zeta} - 1) \nu^\sim(x, dz) \\ \Lambda^\sim(x, y) &= \sup_{\zeta \in \mathbb{R}^d} \{y\zeta - M^\sim(x, \zeta)\} \\ \Lambda^0(x, y) &= \inf\{\beta\Lambda^+(x, y^+) + (1 - \beta)\Lambda^-(x, y^-) : 0 \leq \beta \leq 1, \\ &\quad y^+ \in \overline{A^-}, y^- \in \overline{A^+}, \beta y^+ + (1 - \beta)y^- = y\}. \end{aligned} \quad (3.1)$$

Then in the usual large deviations scaling, the sequence X^n satisfies a large deviations principle in the space of continuous functions, with each absolutely continuous path $\phi(t)$ having large deviations rate

$$\Gamma(\phi) = \int_0^T \Lambda(\phi_t, \dot{\phi}_t) dt \quad (3.2)$$

where $\Lambda(x, y)$ equals $\Lambda^+(x, y)$, $\Lambda^-(x, y)$ or $\Lambda^0(x, y)$ according as x is in A^+ , A^- or A^0 respectively.

These expressions are easily understood. Λ^+ and Λ^- are the usual local large deviations rate functions for the measures ν^+ and ν^- . They apply while the process

stays wholly within the positive or negative half of the space. Λ^0 contributes to the rate function (3.2) when the process moves along the boundary between the two halves of the space. However, it is only in the large deviations limit that the process actually stays on this boundary. The finite- n processes merely stay close to it, spending some time on one side of the boundary and some time on the other side. Equation (3.1) encapsulates this. Intuitively, it says that the process spends a proportion β of its time in the positive half of the space, moving at an average velocity of y^+ , and a proportion $(1 - \beta)$ of its time in the negative half, moving at an average velocity of y^- . The large deviations rate for travel along the boundary at a velocity y corresponds to the cheapest way to assign these quantities, conditional on reverting towards the boundary from each half, and on the net velocity being y .

We shall now use this theory to analyse the $\cdot/M/1$ system under the JSQ routing rule, that is, when arriving discretionary customers do not know the service rates of the two servers, and so join the queue with the smaller number of customers. Since by independence we know the large deviations rates for paths away from the diagonal $X_1 = X_2$, it only remains to calculate the rate for paths that travel along the diagonal. For this system, there is an equivalent formulation of (3.1). Let $X_1 < X_2$ correspond to A^+ in the above formulation, and $X_1 > X_2$ to A^- . While the process is in A^+ , all the discretionary traffic is routed to the first queue, and conversely. Thus β should not only represent the proportion of time spent in A^+ , but also the proportion of the discretionary traffic routed to the first queue. So intuitively the local large deviations rate for travel along the diagonal at velocity (y, y) should be given not only by (3.1) but also by

$$\Lambda^0((x, x), (y, y)) = \inf\{L_{\lambda_1 + \beta\nu, \mu_1}(y) + L_{\lambda_2 + (1-\beta)\nu, \mu_2}(y) : 0 \leq \beta \leq 1\}, \quad (3.3)$$

where L is the local large deviations rate for the single $M/M/1$ queue, given at (2.1). In [2], Alanyali & Hajek show that this is in fact the case.

We shall start as in Section 2 by calculating the large deviations rate for the total occupancy of the system to exceed some constant, which we shall now call C , by some large time T . We claim that the only candidates as the large deviations path to reach the set $X_1 + X_2 \geq C$ are straight lines out of the origin. Lemma 5.16 of Shwartz & Weiss [18] shows that large deviations paths within any region of constant jump rates must be straight lines. (This follows from a convexity argument.) And paths which follow a straight line through one region and then a straight line through another region are ruled out by the argument used in the proof of Theorem 2.3: since the large deviations rate is linear within each region, it is always minimised by the whole path being in just one of the regions.

It is convenient to divide the candidate paths into three classes: those which travel along the diagonal, those which travel along an axis, and those which travel along some other ray. By the same coupling argument as at the beginning of the proof of Theorem 2.2, the large deviations rate for the overflow must always be at most $I_{\lambda_+, \mu_+}(0, C)$, and unless the optimal path is along an axis, it must be exactly this amount, with the total occupancy increasing at rate $\mu_+ - \lambda_+$.

The large deviations rate for the path along the diagonal is found by minimising (3.3) over $y > 0$ and $0 \leq \beta \leq 1$. This does not seem to be easy to do *a priori*: in particular, numerical minimisations show that the optimal β depends on y , so the optimisations over y and β cannot be carried out independently. However, the coupling argument and the bound on the answer in the previous paragraph allow us to predict the answer and verify it. The coupling argument shows that if the

path along the diagonal is to be the optimal path, then the optimal y must be $\hat{y} = (\mu_+ - \lambda_+)/2$. The optimal β can be deduced from a result in Foley & McDonald [10]. They look at this problem using the method of h -transforms. One of their results gives the stationary distribution of $(X_1 - X_2)$ when the region $X_1 + X_2 \geq C$ is reached, for large C . We can therefore propose a candidate β by summing the mass of their distribution over the positive and negative halves of the real line. This gives

$$\hat{\beta} = \frac{\lambda_2 + \nu - \lambda_1 + \rho_+^2(\mu_1 - \mu_2)}{2\nu}. \tag{3.4}$$

Substituting these \hat{y} and $\hat{\beta}$ into (3.3) does indeed give the rate $I_{\lambda_+, \mu_+}(0, C)$ implied by the pooled calculation. So provided that $0 \leq \hat{\beta} \leq 1$, the path along the diagonal is the optimal path. This condition can be written $\nu \geq |\rho_+^2(\mu_2 - \mu_1) + \lambda_1 - \lambda_2|$.

If a path along neither the diagonal nor an axis is optimal, it must have velocity $(c, 2\hat{y} - c)$, with c not equal to 0, \hat{y} or $2\hat{y}$. Suppose that the path travels below the diagonal. In this region the two queues are independent, so finding the optimal path of this type is equivalent to solving

$$\inf\{L_{\lambda_1, \mu_1}(c) + L_{\lambda_2 + \nu, \mu_2}(2\hat{y} - c) : \hat{y} < c < 2\hat{y}\}.$$

This problem can be solved algebraically to give the solution $c = \lambda_1 \rho_+^{-1} - \mu_1 \rho_+$, with the total rate given by the pooled calculation. This path is then the optimal one provided that $c > \hat{y}$. This condition can be written $\nu < \rho_+^2(\mu_2 - \mu_1) + \lambda_1 - \lambda_2$. The condition $c < 2\hat{y}$ is also required, but turns out to be redundant. There is a parallel calculation for paths above the diagonal.

Finally, the optimal path along the first axis has the large deviations rate $I_{\lambda_1, \mu_1}(0, C)$, corresponding to the first queue performing the single-queue overflow, and the second queue performing its most likely behaviour with a large deviations rate of 0. This will be the optimal path if and only if its rate is less than that implied by the pooling bound, in other words if $I_{\lambda_1, \mu_1}(0, C) < I_{\lambda_+, \mu_+}(0, C)$. This condition can be written $\rho_1 > \rho_+$.

We have thus proved the following theorem.

Theorem 3.1 *The large deviations rate to reach the set $X_1 + X_2 \geq C$ by time T is*

$$\begin{aligned} I_{\lambda_1, \mu_1}(0, C) & \text{ if } \rho_1 > \rho_+ \\ I_{\lambda_2, \mu_2}(0, C) & \text{ if } \rho_2 > \rho_+ \\ I_{\lambda_+, \mu_+}(0, C) & \text{ if } \rho_+ > \max\{\rho_1, \rho_2\}. \end{aligned}$$

In the first two cases, the large deviations path is along the respective axis. In the final case, the large deviations path is along the diagonal $X_1 = X_2$ if $\nu > |\rho_+^2(\mu_2 - \mu_1) + \lambda_1 - \lambda_2|$. Otherwise the large deviations path is a straight line at velocity $(\lambda_1 \rho_+^{-1} - \mu_1 \rho_+, (\lambda_2 + \nu) \rho_+^{-1} - \mu_2 \rho_+)$ if $\lambda_1 \rho_+^{-1} - \mu_1 \rho_+ > \lambda_2 \rho_+^{-1} - \mu_2 \rho_+$, or $((\lambda_1 + \nu) \rho_+^{-1} - \mu_1 \rho_+, \lambda_2 \rho_+^{-1} - \mu_2 \rho_+)$ in the opposite case.

Foley & McDonald [10] refer to the paths along the diagonal, along an axis, and at an intermediate angle as the *strongly pooled* case, the *unpooled* case and the *weakly pooled* case respectively. In the pooled cases, the large deviations rate is $-C \log \rho_+$ as implied by the pooling calculation, and thus the performance of the system is, in the large deviations sense, the best one could achieve given the total arrival rate and service rate of the whole system. In the unpooled case, the large

deviations rate is the lower $-C \log \rho_1$ or $-C \log \rho_2$. In the strongly pooled case, the overflow path stays close to the diagonal, but in the weakly pooled case all the overflow traffic goes to one queue, the only interaction of the queues coming as a result of the overflow condition $X_1 + X_2 \geq C$.

Foley & McDonald show that in the strongly pooled case, the process stays close to the diagonal even in unscaled space: at the time of an overflow of a large level, the difference between the queues is close to a known constant distribution. We do not expect this to be the case for the weakly pooled case because, apart from rare interactions with the diagonal and axes, the distance of the queues from the large deviations path is a null recurrent Markov chain, so should tend to have deviations of $O(\sqrt{n})$ rather than $O(1)$.

In fact, in the formulation of the theorem we have counted as weakly pooled the case in which the overflow path is along the diagonal but $\hat{\beta}$ equals 0 or 1, because in that case too the process spends all its time on one side of the diagonal, so that all the overflow traffic goes to one queue. A more detailed analysis, as in Foley & McDonald, confirms that that case has the properties of the weakly pooled, not the strongly pooled, case. As in Theorem 2.3, we do not specify the behaviour at the boundary between the pooled and the unpooled cases, because the large deviations path is not unique at such parameters.

The theorem can alternatively be interpreted this way. With the value of $\hat{\beta}$ at (3.4),

$$(\lambda_1 + \hat{\beta}\nu)\rho_+^{-1} - \mu_1\rho_+ = (\lambda_2 + (1 - \hat{\beta})\nu)\rho_+^{-1} - \mu_2\rho_+. \quad (3.5)$$

Now for a pooled overflow, the system acts as if all the arrival rates had been multiplied by ρ_+^{-1} and all the service rates by ρ_+ . Equation (3.5) says that once the discretionary traffic has been allocated to each queue, the twisted arrival rates minus the twisted service rates are the same at each queue. So the system is strongly pooled if there is a proper allocation of the discretionary traffic which equalises these quantities. If there is no such allocation, but there is a proper allocation of the discretionary traffic which equalises the traffic intensities at the two queues, then the system is weakly pooled. Otherwise the system is unpooled.

Overflows to the set where either queue is large can be obtained easily by a calculation parallel to that of Theorem 2.3. It turns out that we just need to compare $-C \log \rho_1$, $-C \log \rho_2$ and $-(2C) \log \rho_+$. We state the result in the following way.

Theorem 3.2 *The large deviations rate to reach the set $\{X_1 \geq C\}$ or $\{X_2 \geq C\}$ by time T is*

$$\begin{aligned} I_{\lambda_+, \mu_+}(0, 2C) & \text{ if } \rho_+^2 > \max\{\rho_1, \rho_2\} \\ I_{\lambda_1, \mu_1}(0, C) & \text{ if } \rho_1 > \rho_+^2 \\ I_{\lambda_2, \mu_2}(0, C) & \text{ if } \rho_2 > \rho_+^2. \end{aligned}$$

In the first case, the large deviations path travels along the diagonal, and in the other cases it travels along one of the axes.

Note that there is no weakly pooled case in this theorem. All of the cases which were weakly pooled in Theorem 3.1, and even some of those which were strongly pooled, become unpooled in this theorem. This shows that the notion of whether a system is pooled depends not only on the parameters of the model, and the routing

rule used, but also on the particular overflow event we are considering. A system can be pooled for overflows to one set, and unpooled for overflows to a different set.

One can of course also obtain the JSEW results of Section 2 by the methods in this section. This differs from the JSQ analysis in one key respect. When minimising the parallel expression to (3.3), the optimal value of β is independent of y , and is given by

$$\hat{\beta} = \frac{\lambda_2\mu_1 - \lambda_1\mu_2 + \mu_1\nu}{(\mu_1 + \mu_2)\nu};$$

so for this problem the optimal allocation of discretionary traffic does not depend on the speed of travel along the diagonal. This value of β is the value which satisfies Assumption 2.1, in other words it equalises the traffic intensities at the two queues. In this case, therefore, the condition for strong pooling as opposed to weak pooling is that the traffic intensities can be equalised by a proper allocation of the discretionary traffic. But this is the same as the condition for pooling as opposed to no pooling, and so weak pooling does not occur in this system. This agrees with what we found in Section 2.

4 $\cdot/M/\infty$ queues

The analysis of $\cdot/M/\infty$ queues has been covered in some detail by Alanyali & Hajek [2], and we shall not repeat their work. Instead, we shall describe how this case compares with that of $\cdot/M/1$ queues.

The event of interest for $\cdot/M/\infty$ queues is that either queue exceeds a certain capacity bound, in other words that the first queue exceeds occupancy C_1 or the second exceeds C_2 . Alanyali & Hajek only consider the case that $C_1 = C_2$, but if $C_1 \neq C_2$, we shall assume the natural routing policy, that an arriving discretionary customer is routed to the queue with the smaller value of X_i/C_i . This is analogous to the Join the Smaller Expected Waiting time policy in the $\cdot/M/1$ case.

The first observation is that Theorem 2.2 carries over essentially unchanged. In other words, the large deviations rate to reach the line $X_1 + X_2 = k$ is the same as that for the pooled system, and the large deviations path travels along the diagonal $X_1/C_1 = X_2/C_2$. Here the pooled system is the single $M/M/\infty$ queue with the same total arrival rate and total capacity as the whole system. The proof is essentially the same as before, with just a couple of technical modifications. (See [19] for details.)

However, there are some differences as well. The large deviations path for the overflow of a single $M/M/\infty$ queue is not linear in time, nor is the large deviations rate linear in the capacity to be overflowed. This leads to two main differences when we consider the analogue to Theorem 2.3.

First, the minimisation at (2.6) no longer necessarily results in a minimum at 0 or C . This means that an overflow may be of the following form: the occupancies rise together along the diagonal for a time, and then one of the queues overflows with the other reverting to its average behaviour. We view this as a sort of *partial resource pooling* because the overflow behaviour is intermediate between the unpooled and fully pooled cases described earlier.

Secondly if $C_1 \neq C_2$, it is no longer immediate which queue is more likely to overflow, and furthermore the choice does not only depend on a function of λ_i and C_i . In terms of equation (2.5), the minimisation over $i = 1, 2$ cannot be carried out independently of k . For example, suppose that $\lambda_1 = 7$, $\lambda_2 = 11.5$, $C_1 = 10$

and $C_2 = 15$. If $\nu = 1$ then the optimal value of k is 0.8 and the second queue overflows. But if $\nu = 2$, the process travels further along the diagonal to $k = 0.91$ and then the first queue overflows. There is a balance between the second queue being busier and it having further to go to overflow, and which of these factors dominates depends on the optimal value of k , and thus on all the parameters in the model.

5 More than two queues

Consider now a queueing network with more than two queues. Customers are only served at one queue, but each arriving customer has a routing choice between some subset of the queues. We shall consider only the JSEW policy. We shall phrase our comments in terms of $\cdot/M/1$ queues, but there is an immediate analogue in terms of $\cdot/M/\infty$ queues.

This extension appears to be substantially harder than the case of two queues, and we only have a conjecture. We shall begin by reviewing previous progress, and then give our conjecture and an example of how the result could be used.

First, the work of Dupuis & Ellis [8] again shows that a large deviations principle exists for the paths of this process, but does not identify the rate function explicitly.

The theory of Alanyali & Hajek [1] does not cover this case because it only allows two regions separated by a hyperplane. We conjecture that their result extends in the obvious way to paths travelling along an edge bordering more than two regions. However, we do not expect the proof to be easy. In any case, it would be difficult to optimise over a vector of β 's in an expression parallel to (3.1) or (3.3).

In a recent paper, Atar & Dupuis [3] have avoided the optimisation over a vector of β 's by phrasing their problem in terms of the twisted jump rates in each direction instead. If the problem possesses certain structural properties, then a convexity argument shows that essentially the same twist is applied in all regions bordering an edge, thus reducing the problem from one involving a number of β 's that grows exponentially as the dimension increases to one involving only a few jump directions. However, their work currently only applies to a discontinuity at a boundary of the state space, not to an internal discontinuity as in our problem.

In Turner [19], an induction argument is used to identify the rate function for any path in the problem, provided that a certain load balancing condition, generalising Assumption 2.1, holds. The load balancing condition is a strong one, requiring the traffic intensities at any subset of queues to be able to be equalised not only when the subset receives all of the traffic which *must* use it, but also when it is given *some* of the traffic streams which *may* use it. Under this condition, any subset of queues which reaches a certain total occupancy does so by maintaining equal expected waiting times, regardless of the state of the other queues in the network. The induction argument proceeds by regarding this subset as equivalent to a single pooled resource, thus reducing the number of resources being considered. The drawback of this approach is that the load balancing condition required is unrealistically strong: for example, it immediately requires that there is a discretionary stream for each pair of queues. Furthermore, even when the large deviations rate for every path is known, it does not seem easy to enumerate the set of paths corresponding to an overflow event of interest. In the two-dimensional case given in Section 2, we were able to index all paths by the last point which the path touched along the diagonal. However in more than two dimensions such a reduction is not

possible. In order to proceed with this approach, one therefore needs to make an extra assumption on the type of path which could be an overflow path.

We now introduce our conjecture. Suppose that the overflow event of interest is analogous to that in Theorem 2.3, namely that $X_i \geq \mu_i C$ for some i by time T . Call an allocation of a discretionary traffic stream *proper* if each of the queues to which that stream may be routed receives a positive proportion of the traffic.

Now call a subset of queues *balanced* if the following conditions hold. Restrict to just the traffic streams that must use the subset in question, that is, that have no choices outside it. The subset is then balanced if first, the traffic intensities at its queues can be equalised by a proper allocation of the discretionary traffic streams, and secondly, for any partition of the subset there is at least one discretionary stream which can be routed to either sub-subset. Conceptually, a balanced subset is one within which we could hope to maintain equal expected waiting times. The conditions aim to ensure that the process restricted to the subset drifts towards the line of equal expected waiting times from any point in its state space.

We now make the following conjecture.

Conjecture 5.1 *The large deviations rate and path to reach the set $\{\exists i : X_i \geq \mu_i C\}$ by time T can be found as follows. Let $S = \{S(x) : 0 \leq x \leq C\}$ be any process taking values in the space of subsets of the queues, with $S(x)$ balanced for all x and with $S(x) \subseteq S(y)$ for all $x \geq y$. Let $0 = x_0, x_1, x_2, \dots, x_N = C$ be the points at which $S(x)$ changes. Let $\mu(x)$ be the total service rate of queues in $S(x)$, and $\lambda(x)$ be the total arrival rate at streams which must use that subset, both μ and λ being taken to be right continuous. Then the large deviations rate for the overflow is*

$$\inf_S \sum_{i=0}^{N-1} I_{\lambda(x_i), \mu(x_i)}(\mu(x_i)x_i, \mu(x_i)x_{i+1}),$$

and the large deviations path involves the queues in $S(x_i)$ rising along their line of equal expected waiting times from x_i to x_{i+1} , with the rest of the queues free.

Note that because of the linearity of the rate function for $\cdot/M/1$ queues, it will always turn out that the optimal S is constant, as when we were minimising expression (2.6). The problem then reduces to comparing $\mu(S) \log(\mu(S)/\lambda(S))$ for each balanced subset S . We phrased the conjecture in the above way because it allows the extension to $\cdot/M/\infty$ queues, where S is not necessarily constant. Indeed, Alanyali & Hajek [2] have already made a similar conjecture in the $\cdot/M/\infty$ case.

Assuming this conjecture, we can analyse many models of interest. For example, consider the following model, taken from Turner [20]. There is a set of N $\cdot/M/1$ queues, N being large, arranged on a circle. Suppose that each queue has its own dedicated stream of traffic of rate $\lambda = 0.2$, and that each pair of adjacent queues shares a discretionary stream of traffic of rate $\nu = 0.75$, with all queues serving at rate $\mu = 1$. We are interested in the event that any queue exceeds capacity C . The optimisation reduces to comparing $I_{i\lambda+(i-1)\nu, i\mu}(0, iC)$ for different values of i . In this case the optimum is at $i = 3$, so, assuming the conjecture, the way that the system overflows in the large deviations limit is for three queues to rise together, with large deviations rate $-3C \log(2.1/3)$.

On the other hand, consider the same model but composed of $\cdot/M/\infty$ queues, and with $\lambda = 100$ and $\nu = 150$. Using the $\cdot/M/\infty$ version of the conjecture, the most likely path for one queue to exceed a large capacity C involves four adjacent

queues rising together from occupancy 252.7 to 254.9, then three queues until 261.2, then two until 306.3, then one queue on its own until the capacity bound is reached.

References

- [1] Alanyali, M. and Hajek, B. *On large deviations of Markov processes with discontinuous statistics*. Ann. Appl. Probab., **8** (1998), 45–66.
- [2] Alanyali, M. and Hajek, B. *On large deviations in load sharing networks*. Ann. Appl. Probab., **8** (1998), 67–97.
- [3] Atar, R. and Dupuis, P. *Large deviations and queueing networks: methods for rate function identification*. Preprint FI-PIA1998-003, Fields Institute, Toronto, 1998.
- [4] Blinovskii, V. M. and Dobrushin, R. L. *Process level large deviations for a class of piecewise homogeneous random walks*. In *The Dynkin Festschrift: Markov processes and their applications*, Birkhauser, 1994, pp. 1–59.
- [5] Bramson, M. *State space collapse with application to heavy traffic limits for multiclass queueing networks*. Queueing Systems Theory Appl., **30** (1998), 89–148.
- [6] Bucklew, J. A. *Large deviation techniques in decision, simulation, and estimation*. John Wiley & Sons, 1990.
- [7] Dupuis, P. and Ellis, R. S. *Large deviations for Markov processes with discontinuous statistics, II: random walks*. Probab. Theory Related Fields, **91** (1992), 153–194.
- [8] 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.
- [9] Dupuis, P., Ellis, R. S. and Weiss, A. *Large deviations for Markov processes with discontinuous statistics, I: general upper bounds*. Ann. Probab., **19** (1991), 1280–1297.
- [10] Foley, R. and McDonald, D. R. *Join the shortest queue: stability and exact asymptotics*. Preprint, March 1998.
- [11] Foschini, G. J. and Salz, J. *A basic dynamic routing problem and diffusion*. IEEE Trans. Commun., **26** (1978), 320–327.
- [12] Harrison, J. M. *Brownian models of open processing networks: canonical representation of workload*. Preprint, July 1998.
- [13] Harrison, J. M. and Van Mieghem, J. A. *Dynamic control of Brownian networks: state space collapse and equivalent workload formulations*. Ann. Appl. Probab., **7** (1997), 747–771.
- [14] Kelly, F. P. and Laws, C. N. *Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling*. Queueing Systems Theory Appl., **13** (1993), 47–86.
- [15] McDonald, D. R. *Asymptotics of first passage times for random walk in a quadrant*. Ann. Appl. Probab., **9** (1999), 110–145.
- [16] McDonald, D. R. and Turner, S. R. E. *Resource Pooling in Distributed Queueing Networks*. Fields Institute Communications, **28**, 107–131.
- [17] Reiman, M. I. *Some diffusion approximations with state space collapse*. In F. Baccelli and G. Fayolle, editors, *Modelling and Performance Evaluation Methodology*, number 60 in Lecture Notes in Control and Information Sciences. INRIA, Springer-Verlag, 1984.
- [18] Shwartz, A. and Weiss, A. *Large Deviations for performance analysis: queues, communications and computing*. Chapman & Hall, 1995.
- [19] Turner, S. R. E. *Resource pooling in stochastic networks*. Ph.D. dissertation, University of Cambridge, 1996.
- [20] Turner, S. R. E. *The effect of increasing routing choice on resource pooling*. Probab. Engrg. Inform. Sci., **12** (1998), 109–124.
- [21] Turner, S. R. E. *A join the shorter queue model in heavy traffic*. To appear J. Appl. Probab., **37**, no. 1 (2000).