TI 2005-016/4
Tinbergen Institute Discussion Paper

Large Deviations Methods and the Join-the-Shortest-Queue Model

Ad Ridder[1]
Adam Shwartz[2]

[1] *Faculty of Economics and Business Administration, Vrije Universiteit Amsterdam, and Tinbergen Institute,*
[2] *Electrical Engineering, Technion-Israel Institute of Technology.*

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Amsterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Please send questions and/or remarks of non-scientific nature to driessen@tinbergen.nl.
Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

Large deviations methods and the join-the-shortest-queue model

Ad Ridder
Vrije Universiteit Amsterdam*

Adam Shwartz
Electrical Engineering
Technion—Israel Institute of Technology†

3 Feb. 2005

Abstract

We develop a methodology for studying "large deviations type" questions. Our approach does not require that the large deviations principle holds, and is thus applicable to a larg class of systems. We study a system of queues with exponential servers, which share an arrival stream. Arrivals are routed to the (weighted) shortest queue. It is not known whether the large deviations principle holds for this system. Using the tools developed here we derive large deviations type estimates for the most likely behavior, the most likely path to overflow and the probability of overflow. The analysis applies to any finite number of queues. We show via a counterexample that this sytem may exhibit unexpected behavior.

1 Introduction

The theory of large deviation is an important tool in the analysis of performace of computer communications networks. Much effort is devoted to the development of a theory of large deviations which will apply to a large

*Work of the first author was performed in part while visiting the Technion.

†Work of the second author was performed in part while visiting the Vrije Universiteit, Amsterdam, and was supported in part by Fund for the promotion of research at the Technion.

class of models, namely to a large class of queueing systems, and providing a useful expression for the rate function so that the theory can be applied. However, this proves to be a difficult task. For example, the relatively broad study [5] does not cover our simple system, while the general methods of identifying the rate function [1] does not cover any priority systems, and requires checking rather complex hypothesis. See also their references for related studies. Sources of relevant theory include [4, 5, 6, 10, 16, 18, 23]. For a different point of view, bypassing the sample-path arguments, see [9] and references therein. At the other extreme, there are attempts to apply the theory to specific systems and questions. This is done by establishing first the validity of the large deviations principle for the specific system under investigation, and then calculating the probabilities and "optimal paths" of interest: see for example [2, 3, 8, 15, 22] and the applications in [16]. In some papers, the authors assume the large deviations principle to apply, and draw conclusions: see for example [11, 12, 13, 17]. For some cases, direct arguments allow to bypass the theory: see e.g. [20, 19] and the references in [9].

The "Join the shortest queue" (JSQ) system consists of K queues with exponential servers. Arrivals are routed to the "shortest weighted" queue. Previous work on this system covers the two dimensional system. For the symmetric system (where service rates at the queues are equal and both weightes equal 1), the analysis in [16, Chapter 15.10] establishes the large deviations principle, and computes the most likely path as well as the probability of overflow. A variation of this system includes dedicated arrival streams for each queue. This variant is analyzed in [1], where the large deviations principle is established. This is not a generalization of the JSQ, since all analyses of this variant system do not allow the dedicated arrivals to have zero rate. Turner [21] studies the two dimensional variant system with equal weights, as well as with weights that correspond to joining the shortest expected waiting time. Using the large deviations principle and ingenuous coupling arguments, the most likely path to overflow and the probability of overflow are computed. McDonald et. al [7, 14] use a more detailed analysis to study the variant JSQ with two queues and derive a more precise description of the distribution on the way to overflow.

The purpose of this paper is to show that some basic techniques of large deviations can be useful even if it is not known whether the large deviations principle holds for the system of interest. We do this through the analysis of a simple queueing system: Join the Shortest Queue. Our analysis covers the original JSQ system with any finite number of queues and any weights. Our major tools are coupling arguments, and large deviations results for the Poisson process and the M/M/1 queue. In Section 2 we give a formal description of the model. We then restrict our attention to the two dimensional system, so as to expose the ideas in a relatively simple setting. In Section 3 we describe the most likely behavior of this system, starting at any point. We give a detailed description of the two dimensional system, and then extend the result to the general case. The concepts of overflow and most likely paths to overflow are given in Section 4, as well as some technical results. Then in Section 5 we describe in detail these optimal paths. We show that, starting with an empty system, overflow always occurs by following the "weighted diagonal" at constant speed, possibly lingering at 0 before starting the excursion. For this case, we give an exact expression for the rate function. However, even in the two-dimensional case we show that, if the starting point is not the empty state, then it is possible that overflow occurs by emptying one queue (but not the other!) and then proceeding towards the "weighted diagonal." We conclude in Section 6 with an indication of extending the overflow problem to higher dimensions.

2 The Join the Shortest Queue model

The two dimensional Join the Shorter Queue (JSQ) system consists of two infinite queues, each with its own server. Service times are exponential, with parameter μ_1 and μ_2 respectively. The total service rate is $\mu := \mu_1 + \mu_2$. There are Poisson arrivals with rate λ. Let x_i denote the number of customers in queue i. An arrival is routed to one of the queues according to a control policy of the following type. The control is characterised by two positive numbers r_1, r_2 such that whenever $x_1/r_1 \leq x_2/r_2$, the arriving customer is routed to queue 1, and otherwise to queue 2. To describe the control geometrically we imagine the line $r \stackrel{\triangle}{=} (r_1, r_2) \cdot t$, $t \in \mathbb{R}_+$. When the state $x = (x_1, x_2)$ of the system is on or above the line r, arrivals join queue 1, and when the state is below the line, arrivals join queue 2. Let us call this line the control diagonal, or simply the diagonal. In accordance with the traditional JSQ terminology we say that arrivals join the shorter queue, although they actually join the "shortest weighted" queue.
Generalizing to more than two queues, service rates are $\mu_i,\ i = 1, \ldots, K$, and the control is determined by the (strictly) positive numbers $r_i,\ i = 1, \ldots, K$. If the vector of queue sizes is $x = (x_1, \ldots, x_K)$ then an arrival is routed to the queue with the smallest value of x_i/r_i. We assume a fixed rule to break ties. We let i_x denote the queue to which an arrival will be sent, if the queue sizes are given by x. The particular choice is of no consequence to our analysis: we only assume that $i_x = i_{\alpha x}$ for all $\alpha > 0$, that is, the choice depends only on the relative sizes of the queues. The case $r_i = r_j$ corresponds to the tra-

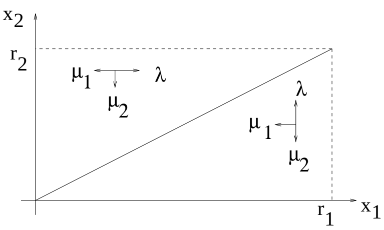

Figure 1: Jump directions and rates in the JSQ system

ditional JSQ and then r is indeed the diagonal. Another special JSQ control is obtained by setting $r_i = \mu_i$: in this case the arriving customers join the queue with the smallest expected waiting time.

Denote by e_i the unit vector in direction $i = 1, \ldots, K$, that is, $e_1 = (1, 0, \ldots, 0)$ etc. We also denote $e_{-i} = -e_i$. Vectors are viewed as row vectors. We use the notation $|x| = x_1 + \cdots + x_K$. We use x to denote points in the state space $\mathbb{Z}_+^K$, the vectors with positive integer entries. Here $x = (x_1, \ldots, x_K)$ is the vector of queue sizes. However, since we are dealing with scaled processes, we shall use the space $\mathbb{R}_+^K$, the positive K-dimensional orthant. Our scaled process is denoted by z_n: this is the process were all jumps are of size $1/n$, and all rates are multiplied by n. As before, $z_{1,n} = z_n \cdot e_1$ is the first coordinate of the process z_n.

To summarize our conventions, we describe the JSQ process in terms of jump directions and jump rates [16]. They are

direction	rate	Event
e_1	$\lambda_1(x) = \lambda \mathbf{1}[x_1/r_1 \text{ is smallest}]$	Arrival to queue 1
e_{-1}	$\lambda_{-1}(x) = \mu_1$	Departure from queue 1
$\vdots$	$\vdots$	$\vdots$
e_K	$\lambda_K(x) = \lambda \mathbf{1}[x_K/r_K \text{ is smallest}]$	Arrival to queue K
e_{-K}	$\lambda_{-K}(x) = \mu_K$	Departure from queue K .

We are interested in the most likely behavior of this system, and in its transient overflow probabilities. That is, given a starting point and a time T, we are interested in the probability that at time nT the process reaches a given point nx (or actually, reaches a small neighborhood of that point which, in our case, turns out to be the same) for $n \to \infty$. More specifically, we study the large deviations limit of this probability, and we find the so-called optimal behavior of the system until overflow.

In our analysis we couple the process to several other processes, which we now define. Fix a point x and recall that i_x denotes the queue to which an arrival will be routed. The *local process* z_n^l is obtained by routing all arrivals to queue i_x, regardless of changes in queue sizes. The *reduced process* z_n^r is a scaled one dimensional M/M/1 process with arrival rate $n\lambda$, service rate $n\mu = n\mu_1 + \cdots + n\mu_K$, jumps of size $1/n$ and starting point $|x|$. Note that the reduced process is stable (in the sense that it reaches 0 with probability one) if and only if $\mu > \lambda$. As we shall show, this is the stability condition for the JSQ system as well. Finally, the scaled *component process* z_n^c is the $K+1$-dimensional process were coordinate i is a (scaled) Poisson process with rate $n\mu_i$, the $K+1$st coordinate is a (scaled) Poisson process with rate $n\lambda$ and all jumps are of size $1/n$. All four processes are coupled by using the same Poisson processes as the arrival and (potential) departure processes. Thus, every arrival entails a jump of size $1/n$ in (a component of) all four processes, etc.

3 Most likely behavior

In this section we describe the most likely behavior of the process that starts at a point x. We provide a detailed analysis of the two-dimensional system. The generalization to K dimensions is given in Section 6. The behavior described below is the path that the process follows, with probability nearly equal 1, in the sense of Kurtz's Theorem [16, Theorem 5.3].

Definition 1 *Fix $T > 0$. We call z_∞ the most likely behavior for a sequence z_n of processes over $[0, T]$ if the following hold. Given ε small enough, there is a constant $C_1 > 0$ and a function $C_2(\varepsilon) > 0$ so that*

$$\mathbb{P}_x\left(\sup_{0\le t\le T} |z_n(t) - z_\infty(t)| \ge \varepsilon\right) \le C_1 e^{-nC_2(\varepsilon)} \quad \textit{all } n > 0. \tag{1}$$

Note that the definition implies that, necessarily, $z_\infty(0) = x$.

3.1 Most likely behavior in dimension 2

If $x_1 > 0$, $x_2 > 0$ and x is not on the "diagonal," then at least locally one coordinate of the process is a Poisson process, and the other is the difference of two Poisson processes. Thus Kurtz's theorem applies [16], and we have

Lemma 1 *For x above r and off the boundary $0 < x_1/r_1 < x_2/r_2$ define*

$$z_\infty(t) = x + (\lambda - \mu_1)e_1 \cdot t - \mu_2 e_2 \cdot t\,. \tag{2}$$

Let T_x be the first time $z_\infty(t)$ hits either the line $r(t)$ or the vertical line $(0, y)$. Then the most likely path until time T_x is to follow z_∞. Moreover, C_2

is quadratic near $\varepsilon = 0$. Finally, for any given $\alpha > 0$, the estimate (1) *is uniform in $\{x : T_x > \alpha\}$. The analogue conclusion holds if x is below r.*

Proof. This is a direct application of [16, Theorem 5.3]. Consider the local process: it is identical to the JSQ process over the region of interest, and the local process satisfies the conditions of [16, Theorem 5.3]. ■

Note that, depending on the parameters, the most likely behavior may drive the process closer to r: in this case, depending on x, the process may hit r first or the vertical line first. The behavior depends both on the parameters and on the starting point. In general, if we start above r then, since arrivals all join queue 1, $z_{2,\infty}$ can only decrease, so that T_x is finite. However, if the system is stable then (if we start above r) the slope of $z_\infty(t)$ is in the cone between $(e_1, -e_2)$ clockwise to $(-e_1, 0)$. More precisely,

Corollary 2 *Suppose x is above r and off the boundary. Then the most likely path moves down with rate μ_2.*
(i) *If $\lambda > \mu_1$ then the most likely path moves to the right. If $\lambda = \mu_1$ then it moves straight down. In both cases it moves closer to the line r.*
(ii) *If $\mu_1 > \lambda$ and $\mu_2/(\mu_1 - \lambda) \geq r_2/r_1$ then the most likely path moves to the left. If the last inequality is strict, it draws closer to r. If it is an equality, it moves parallel to r.*
(iii) *If $\mu_1 > \lambda$ and $\mu_2/(\mu_1 - \lambda) < r_2/r_1$ then the most likely path moves to the left, and away from r.*
Suppose now x is below r and off the boundary. Then the most likely path moves to the left with rate μ_1.
(iv) *If $\lambda > \mu_2$ then the most likely path moves up. If $\lambda = \mu_2$ then it moves straight to the left. In both cases it moves closer to the line r.*
(v) *If $\mu_2 > \lambda$ and $\mu_1/(\mu_2 - \lambda) \geq r_1/r_2$ then the most likely path moves down. If the last inequality is strict, it draws closer to r. If it is an equality, it moves parallel to r.*
(vi) *If $\mu_2 > \lambda$ and $\mu_1/(\mu_2 - \lambda) < r_1/r_2$ then the most likely path moves down, and away from r.*
(vii) *There is no combination of coefficients so that the most likely path moves away from r both when below and when above the line r.*

Proof. This is a direct consequence of Lemma 1. Using (2), (i)-(vi) are established by elementary geometry. The last conclusion follows from (iii) and (vi) by simple algebra. ■

Lemma 3 *Suppose the starting point $x \neq 0$ is on r and that the condition of Corollary 2(iii) holds. Let $T_x \triangleq x_1(\mu_1 - \lambda)^{-1}$. Fix $T < T_x$ and define z_∞*

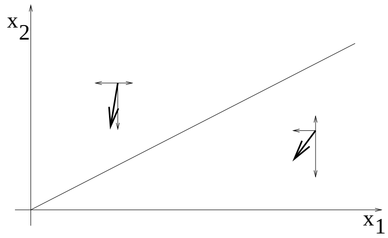

Figure 2: Drift condition (vi)

through (2)*. Then z_∞ is the most likely path until time T, C_2 is quadratic near $\varepsilon = 0$ and for any $\alpha > 0$, this estimate is uniform in $\{x : T_x > \alpha\}$.*

Proof. Let z_n^l denote the local process where arrivals join queue 1. If we couple the two queueing systems (that is, drive them with the same arrival and departure processes) then, at least until one process hits the boundary, $z_{1,n}(t) \leq z_{1,n}^l(t)$ and $z_{2,n}(t) \geq z_{2,n}^l(t)$. This holds since some arrivals to queue 1 for the process z_n^l are sent to queue 2 in the JSQ system (when below r). However, z_n^l satisfies the conditions of Kurtz theorem and so (1) holds. This implies that the scaled JSQ process moves away from r. But once it does, its increments are exactly those of z_n^l, and another application of Kurtz theorem establishes the Lemma. ■

Lemma 4 *Suppose the starting point $x \neq 0$ is on r. Suppose further that neither of the conditions of Corollary 2(iii) or (vi) hold. Let*

$$z_\infty^r(t) = x_1 + x_2 + (\lambda - \mu_1 - \mu_2) \cdot t\,.$$

If the system is stable, define T_x as the first time $z_\infty^r(t) = 0$. Otherwise, choose an arbitrary finite T_x. Fix $T < T_x$ and define z_∞ through

$$z_{2,\infty}(t)/r_2 = z_{1,\infty}(t)/r_1 \quad \textit{and} \quad z_{1,\infty}(t) + z_{2,\infty}(t) = z_\infty^r(t)\,. \tag{3}$$

Then the most likely path until time T is z_∞. Moreover, C_2 is quadratic near $\varepsilon = 0$. Finally, for any $\alpha > 0$, the estimate (1) *is uniform in $\{x : T_x > \alpha\}$.*

Proof. The proof uses standard ideas which will be elaborated for the large deviations calculation, and so will only be described briefly here. First note that these two conditions define a unique path. By coupling the JSQ process with the reduced process we immediately have that the second condition of (3) must hold: that is, in the sense of Kurtz theorem, the scaled process must satisfy this condition. Consider now an initial short period of time.

Since the component process satisfies the conditions of Kurtz theorem, we know that the process (with high probability) cannot move far. If it stays near r, we repeat the argument, to show that it stays near r until T. Once it moves away from r, we can apply Corollary 2 to conclude that it will move in the direction of r. Thus the process stays near r until T (in the sense of Kurtz theorem). Thus the first condition of (3) holds as well. ∎

If x is on the boundary, there are two cases.

Lemma 5 *Suppose $x_1 = 0$ and $x_2 > 0$.*
(i) *If $\mu_1 \geq \lambda$ then* (1) *holds with $z_\infty(t) = x - \mu_2 e_2 \cdot t$ for $T < T_x = \frac{x_2}{\mu_2}$.*
(ii) *If $\mu_1 < \lambda$ then* (1) *holds where $z_\infty(t)$ is given by* (2)*, until z_∞ meets r.*

Proof. Over the region of interest, the JSQ process is identical to the local process. Until the process reaches either r or the point 0, the two coordinates are statistically independent. The first coordinate is an M/M/1 queue, starting at 0. In case (i) it is stable, and so it stays near 0. In case (ii) it is not stable, so it follows its drift: see [16, Chapter 11]. The second coordinate is a Poisson process, so it follows its mean drift. ∎

Note that we have now estalished our stability claim: the JSQ system is stable if $\mu_1 + \mu_2 > \lambda$. We conclude the analysis of the most likely path by describing the behavior starting at 0.

Lemma 6 *Suppose $x = 0$. If the JSQ is stable than the most likely behavior is to stay near* 0*. If it is unstable, than the most likely behavior is to follow r, so that $z_{1,\infty}(t) + z_{2,\infty}(t) = (\lambda - \mu_1 - \mu_2) \cdot t$.*

Proof. If the JSQ is stable then whenever it moves away from 0, the previous results show that it must move towards 0.

In the unstable case, the reduced system follows a straight line away from 0 with speed at least $\lambda - \mu_1 + \mu_2 > 0$. This will be precisely the speed if we are away from the boundaries. But as soon as we move away from 0, our previous results imply that the process must move towards r, and in particular away from the boundaries. Therefor the process moves along r with speed $\lambda - \mu_1 - \mu_2$. ∎

3.2 Most likely behavior in higher dimensions

The most likely behavior can be analyzed much in the same way as the two dimensional system, and so we only comment on this.

(i) First note that if the JSQ system is unstable, namely $\sum_{i=1}^{K} \mu_i < \lambda$ then, by the previous arguments, the most likely behavior is to approach the line $r(t)$ and then follow that line, where the reduced system satisfies

$$z_\infty^r(t) = x^r + \left(\lambda - \sum_{i=1}^{K} \mu_i\right) \cdot t\,.$$

(ii) Suppose that some queues are empty, and some are not. Without loss of generality, assume that $x_i = 0,\ i = 1, \ldots, i_x$ and $x_i > 0,\ i > i_x$. Suppose moreover that the subsystem consisting of i_x queues is unstable, that is $\sum_{i=1}^{i_x} \mu_i < \lambda$. Then queues $1, \ldots, i_x$ are statistically independent of the rest of the queues until one of the other queues reaches a value of $z_{i,n}(t)/r_i$ that is equal or lower than that of one of the first i_x queues. Therefore, for our local (short-time) analysis, we need not consider those queues: their behavior is simply to follow their drift $-\mu_i e_i$. We therefor ignore those queues, which amounts to setting $K = i_x$. Now by definition, the reduced system is unstable, and moreover, it is coupled to our system whenever all queues are non empty. But by definition, none of the ratios $z_{i,n}(t)/r_i$ will be larger than 1 unless all are at least 1. Thus we see that the total size of the queues $1, \ldots, i_x$ increases at the rate $\lambda - \sum_{i=1}^{i_x} \mu_i$ of the reduced system, and as analyzed before the increase is towards r.
(iii) If the subsystem is stable, than the same reasoning shows that these queues will remain empty. The queues we ignored recieve no arrivals, and therefore grow smaller. As another queue empties, the stable system which now consists of one more queue is obviously stable (more departures with the same arrival rate), and so eventually all queues will empty.
(iv) Suppose now that there are no empty queues, but that a subset of the queues have nearly the same product x_i/r_i, and further, that this product is smaller than for other queues. Again, without loss of generality, assume $x_i/r_i < x_j/r_j$ for all $i \le i_x < j$. As before, we can consider the first group of queues, ignoring the rest. If the subsystem is unstable, that is $\sum_{i=1}^{i_x} \mu_i < \lambda$ then, as before, the most likely path is for this set of queues to grow along r, until $z_{i,n}(t)/r_i \ge z_{j,n}(t)/r_j$ for some $j > i_x$. This will eventually happen, as the other queues all decrease while the first queues all grow. At that point, if with the additional queue the system is still unstable, the present analysis holds, with new growth rates.
(v) Finally, if the subsystem is stable, it may decrease along r. To find out if this is the case, consider the queue (among the first i_x queues), say j, with the smallest ratio of μ_i/r_i. This queue is our candidate to empty slower than the rest of the queues, thus staying "above the diagonal" r. To find out if this happens, we compute the rate at which the remaining queues empty.

This is done by considering the reduced system with $i_x - 1$ queues, then calculating the rate at which, say, $z_{1,n}(t)/r_1$ decreases, and comparing this to μ_j/r_j. If the latter is larger, then we know that queue j empties as fast as the remaining group, and the most likely behavior is for these i_x queues to decrease together along r. However, if μ_j/r_j is smaller, than indeed it will stay larger even though no arrivals join queue j. We then repeat the procedure, to check if another queue should be excluded.
Once this procedure is complete, The remaining group of queues empty along r, until they are all empty. This concludes the last case in our analysis.

We note that, unlike other systems, here the boudaries do not hamper our analysis. This is the case since, for our purposes, the coordinates decouple except at 0 and on r. In the first case, a one-dimensional analysis provides the picture. In the second case, the boundaries play no role, and the analysis near r can be completed.

4 Most likely path to overflow: dimension 2

We are interested in the probability and the most likely path to buffer overflow. We provide an explicit formula for the case that the system starts empty. These issues require some explanation, and as in Section 3, we give the details for the two-dimensial system. We do this mainly since the geometry is much easier to follow. In Section 6 we shall extend the arguments to higher dimensions.

By buffer overflow we mean the following. Let $C = nc$ be some high level for queue sizes. Buffer overflow occurs when the system reaches any state where one or both queues exceed the high level: $\{(x_1, x_2) : x_1 \geq nc$ or $x_2 \geq nc\}$. Unless we start very close to the set of overflow states, this set is hit at the unique system state $x = (r_1, r_2)n\beta$ where the diagonal crosses the boundary of the overflow set. For example, if $r_1 > r_2$ then $\beta = c/r_1$. From now on we assume explicitly that the JSQ system is stable, for otherwise the event of reaching overflow is not rare, and the most likely behavior brings the process to the mentioned overflow state (with probability equal nearly 1, as demonstrated in the previous section).

The concept of most likely (or optimal) path is related to the usual scaling of the system and taking large deviations limits, as in [16]. The scaled JSQ process z_n is defined in Section 2. This is a piecewise constant process. It is often more convenient to work with a process with continuous paths, which is obtained by linearly interpolating the original process between jump points. We do not distinguish between the piecewise-constant jump process and its piecewise-linear version, since the two are exponentially equivalent [16].

A path ϕ is a continuous function $[0, T] \to \mathbb{R}^2_+$ mapping a finite time interval into the state space of the scaled JSQ process. The open ε-ball around a path ϕ is the collection of paths

$$B_\varepsilon(\phi) = \left\{ \psi : \sup_{0 \le t \le T} |\psi(t) - \phi(t)| < \varepsilon \right\}.$$

Although we neither assume nor prove that the large deviations principle holds for our system, we shall make use of the following definitions.

Definition 2 *A sequence of processes $\{y_n\}$ satisfies the large deviations principle with rate function $I(\cdot)$ if I is non negative, lower semicontinuous and*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_x (y_n \in C) \le - \inf_{\phi \in C} I(\phi) \quad \textit{for all closed } C, \textit{ and}$$
$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_x (y_n \in G) \ge - \inf_{\phi \in G} I(\phi) \quad \textit{for all open } G.$$

We call ℓ the local rate function if for any absolutely continuous ϕ on $[t_0, t_1]$,

$$I(\phi) = \int_{t_0}^{t_1} \ell(\phi(t), \dot{\phi}(t)\, dt), \tag{4}$$

and $I(\phi) = \infty$ if ϕ is not absolutely continuous.

Let x^0 and x^T be two points and ϕ be a path with $\phi(0) = x^0$, and $\phi(T) = x^T$. We are concerned with the probability that z_n stays close to ϕ, given that $\lim_{n\to\infty} z_n(0) = x^0 = \phi(0)$. We denote this by $\mathbb{P}_{x^0}(z_n \in B_\varepsilon(\phi))$.

Definition 3 *We say that a path ϕ is more likely than the path ψ if*

$$\lim_{\varepsilon\downarrow 0} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_{x^0} (z_n \in B_\varepsilon(\psi)) \le \lim_{\varepsilon\downarrow 0} \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_{x^0} (z_n \in B_\varepsilon(\phi)).$$

We call ϕ the optimal path *from x^0 to x^T if this holds for all $\psi \neq \phi$, where of course $\phi(0) = \psi(0) = x^0$ and $\phi(T) = \psi(T) = x^T$.*

Note that this relation is transitive, that is, if ϕ is more likely than ψ and ψ is more likely than η then ϕ is more likely than η.

We would like to know what these optimal paths look like, given any starting and ending point. The most interesting paths are those leading to overflow. We have already indicated that overflow is reached via the diagonal point $(r_1, r_2)\beta$, hence the optimal path to overflow always ends (at time T) at this point. Furthermore, since the first queue cannot grow below the diagonal and the second queue cannot grow above the diagonal, the only way to reach the overflow point $(r_1, r_2)\beta$ is by reaching the diagonal and then following it in an upward direction (both coordinates increase). A path or a piece of path where both coordinates increase (strictly!) is called *increasing*. We can split our search of optimal path into two:

- Given a starting point x^0 and a time T, how is the diagonal reached, and in particular where and when does the optimal path start to increase along the diagonal?
- Given a starting point on the diagonal and a time T, what is the speed of the optimal path that increases along the diagonal?

We start with the second question. So, consider paths which start on the diagonal, $r^0 := (r_1, r_2)\alpha$ for some $0 < \alpha < \beta$, increase along the diagonal, and end at the overflow point $r^T := (r_1, r_2)\beta$ at time T. Notice that we restrict—for the moment—the starting point to avoid the point 0. The reason is that if you stay near the diagonal starting at r^0, then both servers are kept busy. This makes it easier to treat the process. We claim that paths with constant speed are the most likely.

Lemma 7 *Let ψ be path from r^0 to r^T, increasing along the diagonal. Let*

$$\phi(t) := r^0 + (r_1, r_2)\frac{\beta - \alpha}{T}t, \quad 0 \le t \le T.$$

Then ϕ is more likely than ψ.

Proof. Recall $|x| := x_1 + x_2$. We show that the probability for the JSQ process to stay near ϕ is (asymptoticaly) the same as the probability for the reduced process to stay near $|\phi|$ and, in addition, it is larger than the probability to stay near any other path ψ that stays on the diagonal.

Recall the definition of the coupled reduced process z_n^r (an M/M/1 queue with arrival rate λ and service rate $\mu = \mu_1 + \mu_2$). For $\varepsilon < \min\{r_1^0, r_2^0\}$, since ψ is increasing and $|\psi(0)| = |r^0| \neq 0$, we have that $|\psi(t)| \ge \varepsilon > 0$ and

$$\{z_n \in B_\varepsilon(\psi)\} \subset \{z_n^r \in B_\varepsilon(|\psi|)\} \subset \{z_n^r > 0\} \,. \tag{5}$$

The inclusions (5) hold since if the JSQ process stays near ψ then the coupled reduced process satisfies these conditions. However, for the reduced process we have an LDP with a convex local rate function. Hence straight lines are more likely than other increasing paths [16, Lemma 5.16]. Here there is only one straight line from $|r^0|$ to $|r^T|$, viz. $|\phi|$. Hence, we have established

$$\begin{aligned}
&\lim_{\varepsilon\downarrow 0}\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}_{r^0}\left(z_n\in B_\varepsilon(\psi)\right)\\
&\qquad\le \lim_{\varepsilon\downarrow 0}\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}_{|r^0|}\left(z_n^r\in B_\varepsilon(|\psi|)\right)\\
&\qquad\le \lim_{\varepsilon\downarrow 0}\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}_{|r^0|}\left(z_n^r\in B_\varepsilon(|\phi|)\right)\\
&\qquad= \lim_{\varepsilon\downarrow 0}\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}_{|r^0|}\left(z_n^r\in B_\varepsilon(|\phi|)\right)\,. \qquad (6)
\end{aligned}$$

The last equality holds since the M/M/1 process satisfies the LDP.

We now claim that, in fact, both departure processes as well as the arrival process are nearly straight lines. This follows from [16, Lemma 7.25] as follows. Consider the augmented M/M/1 queue where we look at a 4-dimensional process, consisting of the 3 independent Poisson processes of the component process z^c, and the M/M/1 process whose arrival process is the Poisson λ process, and whose service process consists of (the superposition of) the other two independent processes. This M/M/1 process has exactly the desired distribution, and by [16, Lemma 7.25], each of the coordinate processes follows a straight line (note that in our case, the rates and hence ℓ in [16, Eq. 7.25] do not depend on x, and r is linear so that y there is fixed, so that $\theta(s)$ and $\lambda_j(r(s))$ do not depend on s, and so wr^i there is a straight line). But if arrivals and departures follow a straight line, and the M/M/1 queue follows $|\phi|$, then by the definition of the JSQ, it follows ϕ. Therefore, we have for all $\varepsilon>0$

$$\begin{aligned}
&\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}_{|r^0|}\left(z_n^r\in B_\varepsilon(|\phi|)\right)\\
&\qquad\le \liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}_{r^0}\left(z_n\in B_\varepsilon(\phi)\right)\,. \qquad (7)
\end{aligned}$$

Take $\varepsilon\downarrow 0$ in (7) and use (6) to conclude that ϕ is more likely than ψ. ∎

We now extend Lemma 7 to include the starting point 0.

Lemma 8 *Let $r^0=(0,0), r^T=(r_1,r_2)\beta$. Define $\phi(t):=(r_1,r_2)\beta t/T$, $0\le t\le T$. Let ψ be any other path from 0 to r^T on $[0,T]$ which increases along the diagonal. Then ϕ is more likely than ψ.*

Proof. The difficulty here is that if one of the queues is empty but the other is not, then z_n does not behave like z_n^r. Let $t(\varepsilon)$ be the first time that $\psi_i(t) > \varepsilon$ for all i. Then after $t(\varepsilon)$, if z_n is in $B_\varepsilon(\psi)$ then none of the queues empties again. So let ϕ_ε be the direct path from $\psi(t(\varepsilon))$ to r^T. We can now repeat the arguments of Lemma 7, with two minor changes. First, here the starting point of $z^r(t(\varepsilon))$ is $|\psi(t(\varepsilon))|$ which is not necessarily equal to $|z_n(t(\varepsilon))|$: however their distance for the paths of z_n that stay in $B_\varepsilon(\psi)$ is at most ε. As $\varepsilon \to 0$, this does not change the proof. Finally,

$$\begin{aligned}\mathbb{P}_0\left(z_n \in B_\varepsilon(\psi)\right) \leq & \,\mathbb{P}_0\left(z_n \in B_\varepsilon(\psi), t \geq t(\varepsilon)\right) \\ \leq & \,\mathbb{P}_0\left(z_n^r \in B_\varepsilon(|\phi|), t \geq t(\varepsilon)\right)\end{aligned}$$

so that the argument of (6) applies once we note that

$$\left|\mathbb{P}_0\left(z_n^r \in B_\varepsilon(|\phi|)\right) - \mathbb{P}_0\left(z_n^r \in B_\varepsilon(|\phi|) \text{ for } t \geq t(\varepsilon)\right)\right| \leq e^{-\eta(\varepsilon)n}$$

where $\eta(\varepsilon) \to 0$ as $\varepsilon \downarrow 0$. With these changes the previous proof applies. ∎

Remark The fact that paths with constant speed are most likely among all increasing paths along the diagonal also holds for decreasing paths. ∎

Note that Lemma 8 compares ϕ only to other strictly increasing paths. We show below that for large T the most likely path stays near 0 and then follows ϕ.

We call a straight line path with constant speed a *direct path.* From the two preceeding Lemmas we deduce

Corollary 9 *Let ϕ be a direct path, increasing along the diagonal. Then*

$$I(\phi) := -\lim_{\varepsilon\downarrow 0}\lim_{n\to\infty}\frac{1}{n}\log \mathbb{P}_{x^0}\left(z_n \in B_\varepsilon(\phi)\right) \tag{8}$$

exists. Let T be the duration of the path ϕ, and $\ell(\cdot)$ the local rate function of the (non empty) M(λ)/M(μ)/1 queue [16, Expression (7.17)]. Then

$$I(\phi) = J(T; |\phi(T) - \phi(0)|) := T\ell\left(\tfrac{|\phi(T)-\phi(0)|}{T}\right), \tag{9}$$

where I is given in (4) *and J is strictly convex in T. If ϕ is decreasing than the result holds with a minus sign for $|\phi(T) - \phi(0)|$ in* (9).

Proof. That the limit exists follows directly from the proof of Lemma 7 (and of Lemma 8 in case the starting point is 0), particularly by inequalities (6) and (7). Also these inequalities say that the limit equals the Large Deviations cost rate going from $|\phi(0)|$ to $|\phi(T)|$ in T time units by the M(λ)/M(μ)/1 queue. The cost function is as stated in (9). Convexity follows by differentiation since ℓ is strictly convex. ■

Remark A similar result (limit exists and exact expression) holds for any direct path ϕ which does not cross the diagonal, for the following reason. Consider a (vector) process where the coordinates are statistically independent. Suppose an equation of the type (8) holds for each coordinate seperately, with rate $I_i(\phi_i)$ for the ith coordinate. Then it is easy to show that (8) also holds for the full process, and $I(\phi) = \sum I_i(\phi_i)$. More explicitely,

$$\begin{aligned}\left\{z_{1,n} \in B_{\varepsilon/2}(\phi_1) \text{ and } z_{2,n} \in B_{\varepsilon/2}(\phi_2)\right\} &\subset \{z_n \in B_\varepsilon(\phi)\} \\ &\subset \{z_{1,n} \in B_\varepsilon(\phi_1) \text{ and } z_{2,n} \in B_\varepsilon(\phi_2)\}\end{aligned}$$

so that the desired limit exists, and the rate functions add. Now a direct path that does not cross the diagonal is of one of the following types.

- Except possibly for the starting and/or ending point, the path lies entirely in one of the two interiors of the JSQ, i.e., either below or above the diagonal, and away from the boundaries. In these areas the system is equivalent to the local process, for which one coordinate is a Poisson process and the other an independent M/M/1 queue, so that we have an LDP (end points are easily dealt with in case they lie on a boundary, cf. proof of Lemma 8).

- The path goes along one of the boundaries. Here again the process is equivalent to the local process, comprising of a Poisson and an independent M/M/1 queue. Since the LDP holds for each of the independent coordinates, it also holds for the process, and the rate function is simply the sum of the two rates, one for each coordinate. ■

Remark We call the expression $I(\phi)$ the *rate* or the *cost* of the path ϕ. (8) holds for any direct path in the quadrant which does not cross any boundary. Clearly for increasing paths (both coordinates grow) other than along the diagonal, the cost is ∞. ■

Remark We have not yet settled the optimal paths! Lemmas 7 and 8 say only that direct paths along the diagonal are the most likely increasing paths. The optimal path may decrease from the start until some time epoch and then increase. This was our first question earlier in this section. ■

4.1 Some technical results

To continue our analysis we need some technical results. In particular, we need a concatenation property of direct paths which we formulate and prove for a particular case. Other cases are treated similarly. Let ϕ_1 be a direct path from an interior point x^0 to a diagonal point r^0 in t_1 time units, and let ϕ_2 be the direct path along the diagonal from r^0 to r^1 in t_2 time units. The concatenation $\phi := \phi_1 \to \phi_2$ of these paths is $\phi(t) = \phi_1(t)$ on $[0, t_1]$, $\quad \phi(t) = \phi_2(t - t_1)$ on $[t_1, t_1 + t_2]$.

Lemma 10 *(i) The limit* $I(\phi) = -\lim_{\varepsilon \downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_{x^0}\left(z_n \in B_\varepsilon(\phi)\right)$ *exists.*
(ii) $I(\phi) = I(\phi_1) + I(\phi_2)$.

Proof. The issue is that in order to use our result concerning direct paths along the diagonal, we need the starting point $z_n(t_1)$ of the second segment to equal $\phi_2(0)$, and in particular to be on the diagonal. However, the fact that z_n is near ϕ_1 guarantees this only in the limit. To overcome this, recall the definition of the local process z_n^l and denote its rate function by I^l. Let ϕ_1^δ denote a continuation of ϕ_1, by a path which moves perpendicular to the diagonal (crossing the diagonal) with speed 1 and is defined until time $t_1 + \delta$. Let $A(n, \varepsilon, \delta)$ denote the event that, for some $t_1 - \delta \le t \le t_1 + \delta$, $z_n(t)$ is on the diagonal within $\varepsilon/2$ of $\phi_1(t_1)$. We couple z_n and z_n^l, so that they share the same arrivals and potential services: in this case, if z_n^l stays near ϕ_1^δ until $t_1 + \delta$ and $\delta > \varepsilon$, then necessarily z_n crosses the diagonal so that the event $A(n, \varepsilon, \delta)$ occurs. Thus, since the large deviations estimate holds separately for each direct path, we have for all $\delta > 0$ small enough

$$
\begin{aligned}
-I(\phi_1) &= \lim_{\varepsilon\downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_{x^0}\left(z_n \in B_\varepsilon(\phi_1)\right) \\
&\ge \lim_{\varepsilon\downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_{x^0}\left(z_n \in B_\varepsilon(\phi_1)\ , A(n, \varepsilon, \delta)\right) \\
&\ge \lim_{\varepsilon\downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_{x^0}\left(z_n^l \in B_{\varepsilon/2}(\phi_1^\delta)\right) \\
&= -I^l(\phi_1^\delta)\,.
\end{aligned}
$$

However, $\lim_{\delta\downarrow 0} I^l(\phi_1^\delta) = I(\phi_1)$ since I^l is continuous with respect to extensions with bounded speed, and on ϕ_1 the processes z_n and z_n^l agree.

Thus we may assume without loss of generality that z_n actually hits the diagonal, and does so within $\varepsilon/2$ of $\phi_1(t_1)$. Now $I(\phi_2)$ is continuous with respect to shifts along the diagonal as well as with respect to small time changes. Therefore, using the Markov property, we see that we may indeed assume that $z_n(t_1) = \phi_2(0)$. The result follows now by another application of the Markov property. ■

Corollary 11 *Let ϕ be a concatenation of a finite number of direct paths ϕ_i. Assume none of the paths crosses the diagonal. Then*

$$I(\phi) = -\lim_{\varepsilon \downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_{x^0}\left(z_n \in B_\varepsilon(\phi)\right) = -\sum_i I(\phi_i)\,.$$

Proof. The proof is the same as that of Lemma 10. ■

Remark. Corollary 11 can in fact be extended to the concatenation of general paths which do not cross the diagonal, provided that paths along the diagonal are direct paths.

5 Overflow paths in dimension 2

In this section we shall find and describe the optimal paths to overflow. It turns out that it is convenient to have a notation for the cost rate of direct paths.

Notation for cost rate functions
(i) The expression (9) for the cost rate of the direct path between two points r^0 and r^1 on the diagonal in T units of time is denoted by $I^r(r^0, r^1, T)$.

(ii) The expression for the cost rate of the straight line between x and y ($y > x$) in t time units in a Poisson process with rate μ_j is denoted by $I^j(x, y, t)$. It is equal to $t\ell^j\left((y-x)/t\right)$, where $\ell^j(a) = a\log(a/\mu_j) - a + \mu_j$ is the local rate of the Poisson process.

(iii) The expression for the cost rate of the straight line between x and y in t time units in an M/M/1 queue with arrival rate λ and service rate μ_j is denoted by $I^{MM1}(x, y, t; \mu_j)$. It is equal to $t\ell^{MM1}\left((y-x)/t; \mu_j\right)$, where $\ell^{MM1}(\cdot; \mu_j)$ is the local rate function of the M(λ)/M(μ_j)/1 queue [16, Expression (7.17)]

5.1 Starting point 0

First we consider paths starting at 0. When the time to overflow T is large enough, the optimal path remains at 0 and then moves straight with constant speed $\mu_1 + \mu_2 - \lambda$ to the overflow point. Otherwise, it leaves 0 immediately and moves with constant speed.

Theorem 12 *Let $T_0^* := |r^T|(\mu_1 + \mu_2 - \lambda)^{-1}$. The most likely way for the stable JSQ to reach the overflow point $r^T = (r_1, r_2)\beta$ starting at 0, is:*
(i) If $T \le T_0^*$ then the optimal path is the direct path between 0 and r^T, i.e., $\phi(t) = (r_1, r_2)\beta t/T,$, $t \in [0, T]$, with cost rate $I(\phi) = J(T, |r^T|)$.
(ii) If $T > T_0^*$ then the optimal path is to stay in 0 until $T - T_0^*$, and then proceed along the the diagonal with speed $\mu_1 + \mu_2 - \lambda$, i.e.

$$\phi(t) = (r_1, r_2)(r_1 + r_2)^{-1}(\mu_1 + \mu_2 - \lambda)(t - (T - T_0^*)), \;\; t \in [T - T_0^*, T],$$

and $\phi(t) = 0, \;\; t \in [0, T - T_0^*]$, with cost rate

$$I(\phi) = J(T_0^*, |r^T|) = (r_1^T + r_2^T)\log\left((\mu_1 + \mu_2)\lambda^{-1}\right).$$

Proof. Let ψ a path containing a detour. That is, there are $t_1 < t_2 < t_3$ with $\psi(0) = 0$, $\psi(T) = r^T$, $\psi(t_1)$ and $\psi(t_2)$ are on the diagonal and $\psi(t)$ is not on the diagonal for $t_1 < t < t_2$. Since $\min\{z_{1,n}(t)/r_1, z_{2,n}(t)/r_2\}$ can only increase on the diagonal, then if z_n is near ψ with positive probability, then necessarily $\psi(t_2)$ is below and to the left of $\psi(t_1)$. Since $\psi(T) = r^T$ and since increase only occurs on the diagonal, there must be a t_3, the smallest time after t_2 so that $\psi(t_3) = \psi(t_1)$. Define $\Delta = t_3 - t_1$ and consider

$$\overline{\psi}(t) = \begin{cases} 0 & 0 \le t \le \Delta \\ \psi(t - \Delta) & \Delta \le t \le t_3 \\ \psi(t) & t_3 \le t. \end{cases} \tag{10}$$

We now use coupling to show that the probability to stay near $\overline{\psi}$ is larger than that of ψ. To do this, consider the scaled component process z_n^c. Construct the coupled process $\overline{z}_n^c$ by interchanging its segments as in (10). Since Poisson processes are memoryless, the distribution of the two processes is the same. By this coupling,

$$\begin{aligned} \mathbb{P}_0\left(z_n \in B_\varepsilon(\psi)\right) &\le \mathbb{P}_0\left(\sup|z_n(t) - \psi(t)| < \varepsilon, \text{ for } t \in [0, t_1] \cup [t_3, T]\right) \\ &= \mathbb{P}_0\left(\sup|\overline{z}_n(t) - \overline{\psi}(t)| < \varepsilon, \text{ for } t \in [\Delta, T]\right). \end{aligned}$$

Since the JSQ system is stable,

$$\begin{aligned} &\lim_{\varepsilon \downarrow 0}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_0\left(\sup|\overline{z}_n(t) - \overline{\psi}(t)| < \varepsilon, \text{ for } t \in [\Delta, T]\right) \\ &\qquad = \lim_{\varepsilon \downarrow 0}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_0\left(\overline{z}_n \in B_\varepsilon(\overline{\psi})\right). \end{aligned}$$

Thus paths without detours are more likely. But then Corollary 9 applies, and the cost is that of the reduced M/M/1 queue, which gives (i) and (ii).∎

5.2 Starting point on the diagonal other than 0

The analysis becomes more complicated when the starting point lies on the diagonal other than 0, i.e., when $r^0 = (r_1, r_2)\alpha$ where $0 < \alpha < \beta$. In fact, we will see a variety of types of optimal paths. Consider the possibilities:

1. The path is the direct path to r^T.

2. The path decreases along the diagonal to a lower diagonal point (except 0), from where it increases with constant speed to the overflow point.

3. The path goes via a detour to a lower diagonal point (except 0), from where it increases with constant speed to the overflow point. The detour lies in the region below (or above) the diagonal and may touch the 0-boundary but never stay there.

4. The path goes to a 0-boundary, runs along this boundary, returns to a diagonal point (except 0), from where it increases with constant speed to the overflow point.

5. The path decreases along the diagonal to point 0, stays there some time and finally increases with constant speed to the overflow point.

6. The path goes via a detour to point 0, stays there some time and finally increases with constant speed to the overflow point. The detour lies in the region below (or above) the diagonal and may touch the 0-boundary but never stay there.

7. The path goes to a 0-boundary, runs along this boundary to point 0, stays in 0 some time and finally increases with constant speed to the overflow point.

First we shall show that paths of types 2, 3 and 6 are ruled out as candidates for optimal path (Lemma 14). Then we show that there exist optimal paths of all the other types. Which type actually is optimal depends on the parameters. However, a general result holds when the overflow time T is small or large.

In order to rule out the types 2, 3 and 6 we need the following observation. It says that when you want to go down from a diagonal point to a lower diagonal point (except 0) where you want to be exactly t time units later, and you are not allowed to stay along the 0-boundaries, then the cheapest way is to follow the diagonal, and not some kind of detour.

Lemma 13 *Let $r^1 = (r_1, r_2)\gamma$ and $r^2 = (r_1, r_2)\delta$ be two points on the diagonal with $0 < \delta < \gamma$. Let ϕ be the direct path from r^1 to r^2 during t time units. Let ψ be a detour, also connecting r^1 with r^2 in t time units which never runs along the 0-boundaries. Then $I(\phi) \le I(\psi)$.*

Proof. Note that by Corollary 11 and the following remark, $I(\psi)$ exists since ψ is a detour so that it touches the diagonal only at the endpoints. The proof of Lemma 10 then shows that the endpoints have no effect, and coupling to the local process we obtain the existence of the rate.

Clearly we only have to consider a detour lying either below or above the diagonal. We take a detour below the diagonal. For small ε let $x^1 := \psi(\varepsilon)$ and $x^2 = \psi(t-\varepsilon)$. These two points lie just below the diagonal and $x^j \approx r^j$, $j = 1, 2$. Let ψ^ε be the direct path between these two points. It runs just below the diagonal (almost parallel).

The JSQ process is coupled to the local process in which all arrivals are sent to queue 2. The first coordinate of the local process is a pure death process with rate μ_1, the second coordinate is an M/M/1 queue with rates λ and μ_2. The death process connects x_1^1 with x_1^2 in $t - 2\varepsilon$ time units without becoming 0, it is similar to a pure birth process from x_1^2 to x_1^1. The optimal way is a straight line with cost rate $I^1(x_1^2, x_1^1, t-2\varepsilon)$. Similarly, the M/M/1 queue connects x_2^1 with x_2^2 in $t - 2\varepsilon$ time units where it may reach the 0-axis, but leaves it immediately. Again, among all these paths the straight line has minimal cost $I^{MM1}(x_2^1, x_2^2, t-2\varepsilon; \mu_2)$. Thus we get

$$I(\psi^\varepsilon) = I^1(x_1^2, x_1^1, t-2\varepsilon) + I^{MM1}(x_2^1, x_2^2, t-2\varepsilon; \mu_2).$$

Let η be the sum of the cost rates of the two (tiny) paths connecting r^1 with x^1 and r^2 with x^2, respectively. Then, for all ε small enough, $I(\psi) \ge I(\psi^\varepsilon) + \eta$. Letting $\varepsilon \to 0$ we get $\eta \to 0$, hence

$$\begin{aligned} I(\psi) &\ge \lim_{\varepsilon \to 0} I(\psi^\varepsilon) = I^1(r_1^2, r_1^1, t) + I^{MM1}(r_2^1, r_2^2, t; \mu_2) \\ &= I^r(r^1, r^2, t) = I(\phi). \end{aligned}$$

The equality holds by coupling the component process of the JSQ process with the birth (μ_1) process and the independent Poisson (λ) and Poisson (μ_2) processes of the M/M/1 queue. ∎

Remark
Lemma 13 holds true also when the end point $r^2 = 0$. The proof goes similarly with a slight adjustment for x^2. ∎

Lemma 14 *Paths of type 2, 3 and 6 can not be optimal paths.*

Proof. Let $r^1 = (r_1, r_2)\gamma$, $0 < \gamma < \alpha$.

- Consider a path $\psi = \psi_1 \to \psi_2$ of type 2: ψ_1 is a decreasing path from r^0 to r^1 in t time units along the diagonal, ψ_2 is the direct path from r^1 to r^T along the diagonal with constant speed in the remaining $T-t$ time. The cost rates are $I(\psi_1) \geq I^r(r^0, r^1, t)$ (constant speed is more likely along the diagonal) and $I(\psi_2) = I^r(r^1, r^T, T-t)$. Then

$$\begin{aligned} I(\psi) = I(\psi_1) + I(\psi_2) &\geq I^r(r^0, r^1, t) + I^r(r^1, r^T, T-t) \\ &\geq I^r(r^0, r^T, T) = I(\phi). \end{aligned}$$

  The latter is the cost of the direct path ϕ from r^0 to overflow. The inequality follows because in the M/M/1 queue straight lines have lower cost than other paths which do not run along the 0-axis.

- Consider a path $\psi = \psi_1 \to \psi_2$ of type 3: ψ_1 is a detour from r^0 to r^1 in t time units below (or above) the diagonal never running along the 0-boundaries, ψ_2 is the direct path from r^1 to r^T along the diagonal with constant speed in the remaining $T-t$ time. According to Lemma 13 the cost rate $I(\psi_1) \geq I^r(r^0, r^1, t)$. Again $I(\psi_2) = I^r(r^1, r^T, T-t)$. Reason as above.

- Consider a path $\psi = \psi_1 \to \psi_2 \to \psi_3$ of type 6: ψ_1 is a detour from r^0 to 0 in t_1 time units below (or above) the diagonal never running along the 0-boundaries, ψ_2 is the zero speed path at point 0 during t_2 time units, ψ_3 is the direct path from 0 to r^T along the diagonal with constant speed in the remaining $T - t_1 - t_2$ time. According to the remark below Lemma 13 the cost rate $I(\psi_1) \geq I^r(r^0, 0, t_1)$. Furthermore $I(\psi_2) = 0$ and $I(\psi_3) = I^r(0, r^T, T - t_1 - t_2)$. Similar as above we get $I(\psi) \geq I(\phi)$ where ϕ is the path that replaces the first part ψ_1 by going along the diagonal. ∎

5.3 Starting on the diagonal and small overflow time

The cost rate of the direct path in T time units is $I_d(T) := I^r(r^0, r^T, T)$. Recall that the cost rate for the M/M/1 queue is a strictly convex, unimodal function. Its minimum is easily determined:

$$T_d := \arg\min_{T>0} \; I_d(T) = \frac{|r^T - r^0|}{\mu_1 + \mu_2 - \lambda}.$$

Theorem 15 *The direct path is optimal if $T \leq T_d$.*

Notice that we only give a sufficient condition for type 1 path to be optimal. We conjecture that there is some $T_d^* \geq T_d$ such that the direct path is optimal if $T < T_d^*$ and not optimal if $T > T_d^*$.

Proof. The proof follows directly from results of the M/M/1 queue with parameters λ and $\mu_1 + \mu_2$. In the M/M/1 queue, for any $t < T_d$ the optimal path from $|r^0|$ to $|r^T|$ in t time units is the straight line, with cost $I^{MM1}(|r^0|, |r^T|, t; \mu_1 + \mu_2) = I^r(r^0, r^T, t)$. As a function of t the cost decreases on $(0, T_d)$.

Let $T < T_d$. If the direct path in the JSQ process would not be optimal, then the optimal path would start with a detour to a lower diagonal point or to point 0 before increasing along the diagonal to the overflow point. This means for the M/M/1 queue that the distance of increase would be $y > |r^T| - |r^0|$ to be covered in $t < T$ time. When you do this, then you cover $|r^T| - |r^0|$ in even less time, say $s < t$. In other words, the cost of this would-be optimal path is larger than $I^{MM1}(|r^0|, |r^T|, s; \mu_1 + \mu_2) > I^{MM1}(|r^0|, |r^T|, T; \mu_1 + \mu_2) = I^r(r^0, r^T, T) = I_d(T)$–a contradiction. ∎

5.4 Starting on the diagonal and large overflow time

In this section we let the overflow time T become large.

Theorem 16 *Let $T_0^* := (r_1^T + r_2^T)(\mu_1 + \mu_2 - \lambda)$. There is a $T^* > T_0^*$ such that when the overflow $T > T^*$, the optimal path is to follow the most likely behavior to 0, to stay there until $T - T_0^*$ and to use the remaining time to approach the overflow point with constant speed along the diagonal. The cost rate equals*

$$(r_1^T + r_2^T) \log \frac{\mu_1 + \mu_2}{\lambda}.$$

Notice that the optimal path is either type 5 or type 7. This depends on the parameters.

Proof. First we consider the case where the most likely behavior is along the diagonal as in Lemma 4. By Lemma14 we may exclude types 2,3 and 6. The three types of paths which are candidates to be optimal are:

- Type 1: direct. The cost is $I_1(T) = I^r(r^0, r^T, T)$.
- Type 4: it moves to the lower boundary, stays along the boundary for some time, returns to the diagonal to the point $r^1 = (r_1, r_2)\gamma$ at time t and then uses a direct path along the diagonal to overflow. The

cost function for such a path exists and satisfies $I_2(T) \geq I(r^0, r^1, t) + I^r(r^1, r^T, T-t)$, where $I(r^0, r^1, t)$ is the optimal cost to go during time t from r^0 to r^1.

- Type 5: it moves during t time units along the diagonal to 0, stays there t_0 time, and finally it goes straight to overflow. The cost is $I_5(T) \geq I^r(r^0, 0, t) + I^r(0, r^T, T-t-t_0)$.

The reason we do not need to consider type 7 is that, since the most likely behavior is along the diagonal, the segment along the diagonal is more likely than that of type 7, while the segment from 0 to overflow is identical.

Since the limit $\lim_{s\to\infty} I^r(x, y, s) = \infty$ for any $x, y \neq 0$ and similarly for $I(x, y, s)$ (e.g. [16, Section 11.2]), $\lim_{T\to\infty} I_1(T) = \infty$ and $\lim_{T\to\infty} I_2(T) = \infty$. The latter, because at least one of the terms in $I_2(T)$ must go to infinity. We therefore need only show that the cost for type 5 is bounded. Considering the cost of type 5 paths, the lower bound is sharp when t and t_0 are chosen correctly. When T is large enough we can choose t equal to the most likely time to arrive at 0,

$$t = T_0 := \frac{r_1^0 + r_2^0}{\mu_1 + \mu_2 - \lambda},$$

and we choose t_0 such that $T - t - t_0$ is the optimal time to get to overflow from 0,

$$T - t - t_0 = T_0^* := \frac{r_1^T + r_2^T}{\mu_1 + \mu_2 - \lambda}.$$

In other words, for all $T \geq T_0 + T_0^*$ the cost of the type 5 path is

$$I_5(T) = I^r(0, r^T, T_0^*) = (r_1^T + r_2^T) \log \frac{\mu_1 + \mu_2}{\lambda},$$

the optimal cost rate for large overflow time in the M/M/1 queue, [16, Theorem 11.3]. It remains to find the minimal $T^* \geq T_0 + T_0^*$ so that for all $T > T^*$ both $I_5(T) < I_1(T)$ and $I_5(T) < I_2(T)$. Then for all $T > T^*$ the optimal path is that both queues first empty in the most likely speed, remain empty, and finally both increase optimally.

The second case occurs when the most likely behavior moves away from the diagonal, condition (iii) or (vi) in Corollary 2. Before we give the details of analysis we say that the concludsion is that there is some T^* so that, for all $T > T^*$, the optimal path is for queue 2 to empty with the most likely speed faster than queue 1 until they are both empty, thereafter the queues remain empty, until they increase optimally to oveflow.

We consider condition (vi) in Corollary 2, that is, the most likely behavior is away and down. The three types of paths candidates to be optimal are Type 1, 4 and 7.

- Type 1: direct with cost.

$$I_1(T) = I^r(r^0, r^T, T).$$

- Type 4: it moves during t time units via a detour to a lower point $r^1 = (r_1, r_2)\gamma$ except 0, $0 < \gamma < \alpha$, from where it goes straight to overflow. The detour runs partly along the 0-boundary. The cost is not known precisely, because the M/M/1 coordinate of the local process falls down to 0 and later moves up, no straight line. However it is easy to derive a lower bound by canceling the queue part. Since the birth process coordinate follows a straight line, we have

$$I_4(T) \geq I^1(0, r_1^0 - r_1^1, t) + I^r(r^1, r^T, T - t).$$

- Type 7: it moves during t time units via a detour to 0, stays there t_0 time, and finally it goes straight to overflow. The detour ends in 0 by running along the 0-boundary. A lower bound is obtained by taking the cost of the last part only:

$$I_7(T) \geq I^r(0, r^T, T - t - t_0).$$

Also the local cost function of the Poisson process is convex, thus $\lim_{s\to\infty} I^1(0, x, s) = \infty$ for any $x > 0$. So again,

$$\lim_{T\to\infty} I_1(T) = \infty$$
$$\lim_{T\to\infty} I_4(T) = \infty.$$

Considering the cost of type 7 paths, the lower bound is sharp when t and t_0 are chosen correctly, similar to above. We choose t to be the time it takes for the birth process coordinate to reach 0 most likely. Notice that the M/M/1 coordinate then surely has reached 0.

$$t = T_1 := \frac{r_1^0}{\mu_1},$$

and we choose t_0 again by $T - t - t_0 = T_0^*$ as above. In other words, for all $T \geq T_1 + T_0^*$ is the cost of the type 7 path

$$I_7(T) = I^r(0, r^T, I_0^*) = (r_1^T + r_2^T) \log \frac{\mu_1 + \mu_2}{\lambda}.$$

It remains to find the minimal $T^* \geq T_1 + T_0^*$ so that for all $T > T^*$ both $I_7(T) < I_1(T)$ and $I_7(T) < I_4(T)$. ∎

5.5 An optimal path of type 4

In this subsection we slightly change notation for the starting and ending points. The starting point is $x^0 = (r_1, r_2)\alpha$ and the overflow point is $x^4 = (r_1, r_2)\beta$. We saw in the previous section that for large T the optimal path to overflow is to follow the most likely behavior to point 0, and wait there, until it can optimally proceed along the diagonal, viz. the constant speed $\mu_1 + \mu_2 - \lambda$. The cost of such a path is

$$I_0 := (x_1^4 + x_2^4) \log \frac{\mu_1 + \mu_2}{\lambda}.$$

Assume that condition (vi) in Corollary 2 holds, that is

- the most likely behavior moves away and down from the diagonal,

$$\frac{\mu_2 - \lambda}{\mu_1} \geq \frac{r_2}{r_1}. \tag{11}$$

In that case, the minimal time that a path brings you to overflow in the manner just described is with a waiting time in point 0 of 0:

$$T_0 := \frac{x_1^0}{\mu_1} + \frac{x_1^4 + x_2^4}{\mu_1 + \mu_2 - \lambda}. \tag{12}$$

This 'minimal' path follows the most likely behavior by emptying queue 2 faster than queue 1 (condition (11)) until the boundary (queue 2 is empty), then it goes along the boundary to 0 with most likely speed $-\mu_1$ emptying queue 1. On arriving in 0 it immediatly approaches the overflow point optimally along the diagonal. The total cost is I_0. We do not say that this is the optimal path to overflow in T_0 amount of time. However, we do know that any other path via 0 is less likely because that has as much cost for the 0-to-overflow part as I_0.

The direct path from x^0 to x^4 has cost (for any overflow time T)

$$I_d(T) := T\,\ell^r\left(\frac{|x^4 - x^0|}{T}\right).$$

We impose

- the condition that the direct cost when $T = T_0$ is strictly larger than the cost for large T:

$$I_d(T_0) = T_0\,\ell^r\left(\frac{|x^4 - x^0|}{T_0}\right) > (x_1^4 + x_2^4) \log \frac{\mu_1 + \mu_2}{\lambda} = I_0\,. \tag{13}$$

Because the direct cost $I_d(T)$ is continuous (in T) and condition (13), we can choose $\varepsilon > 0$ such that $I_d(T_0 - \varepsilon) > I_0$. And we choose

- a diagonal point x^3 such that

$$\frac{x_1^3}{\mu_1} + \frac{x_1^3 + x_2^3}{\mu_1 + \mu_2 - \lambda} = \varepsilon. \tag{14}$$

We shall construct a path ψ from starting point x^0 to overflow point x^4 in $T_0 - \varepsilon$ time units. When this path would go via point 0, then it would have cost higher than I_0 because there is not enough time to do it in the most likely manner just described. We let ψ be almost similar to the minimal path but let it return to the diagonal (in x^3) before it would have reached point 0 (see Figure 3). In other words, ψ is a type 4 path. We construct ψ such that it has cost $I(\psi) < I_0$. Because the cost is less than I_0 we conclude that ψ is (a) more likely than the direct path (type 1), and (b) more likely than any path via 0 (types 5 and 7): ψ or some other type 4 path must be optimal.

For the construction of ψ we introduce the boundaries points $x^1 = (x_1^1, 0)$ (where ψ hits the boundary coming from the diagonal) and $x^2 = (x_1^2, 0)$ (where ψ leaves the boundary heading for x^3 on the diagonal). We let this all happen in a way that the birth coordinate of the local process behaves with the most likely speed until point x^3 has been reached. And we let the queue coordinate behave most likely while emptying and optimally when building up. So, the time from the diagonal until the boundary equals $t_1 = x_2^0/(\mu_2 - \lambda)$. The time from the boundary until the diagonal is $t_3 = x_2^3/(\mu_2 - \lambda)$. Then we get

$$x_1^1 = x_1^0 - t_1\mu_1, \quad x_1^2 = x_1^3 + t_3\mu_1.$$

To recapitulate, path ψ concatenates the following four subpaths.

1. The straight line $\psi_1(t)$ which follows the most likely behavior from x^0 towards the boundary, until it hits the boundary in x^1 at time t_1.

2. The straight line $\psi_2(t)$ along the boundary between x^1 and x^2 taking $t_2 = (x_1^1 - x_1^2)/\mu_1$ time units.

3. The straight line $\psi_3(t)$ connecting boundary point x^2 with diagonal point x^3 in t_3 time units. Notice that, since the birth process behaves most likely, also $t_3 = (x_1^2 - x_1^3)/\mu_1$.

4. The straight line $\phi_4(t)$ from x^3 to x^4 along the diagonal with constant speed $\mu_1 + \mu_2 - \lambda$. The time required is $t_4 = (|x^4 - x^3|)/(\mu_1 + \mu_2 - \lambda)$.

A quick check shows that $t_1 + t_2 + t_3 + t_4 = T_0 - \varepsilon$ as required. The cost of ψ is

$$\begin{aligned}
I(\psi) &= I(\psi_0) + I(\psi_1) + I(\psi_2) + I(\psi_3) \\
&= 0 + 0 + x_2^3 \log \frac{\mu_2}{\lambda} + (x_1^4 + x_2^4 - x_1^3 - x_2^3) \log \frac{\mu_1 + \mu_2}{\lambda} \\
&= I_0 + x_2^3 \log \frac{\mu_2}{\lambda} - (x_1^3 + x_2^3) \log \frac{\mu_1 + \mu_2}{\lambda} \\
&< I_0.
\end{aligned}$$

Numerical Example
We provide an example to show that the imposed conditions (11) – (14) are consistent. The parameters are

$$\lambda = 1, \mu_1 = 0.25, \mu_2 = 4, r_1 = 1, r_2 = 1.$$

The starting point and overflow point are

$$x^0 = (1, 1), x^4 = (2, 2).$$

The unit of time is second.

It takes at least $T_0 = 5.23$ seconds to go most likely to 0 and optimally to overflow. The cost of the type 7 path doing this is $I_0 = 5.7877$. Let $\varepsilon = 1$ and $T = T_0 - \varepsilon = 4.23$ seconds. Then

- the cost of the type 1 path (direct) in T seconds is $I_d(4.23) = 6.2182$;
- the cost of a type 7 path (via 0) is at least $I_0 = 5.7877$.

The points x^1, x^2 on the boundary and x^3 on the diagonal in the construction of the type 4 path ψ are

$$x^1 = (0.9167, 0), x^2 = (0.2347, 0), x^3 = (0.2167, 0.2167).$$

The intermediate running times are

$$t_1 = 0.33, t_2 = 2.73, t_3 = 0.07, t_4 = 1.10,$$

which adds up to $T = 4.23$ seconds. The cost is $I(\psi) = 5.4610$.

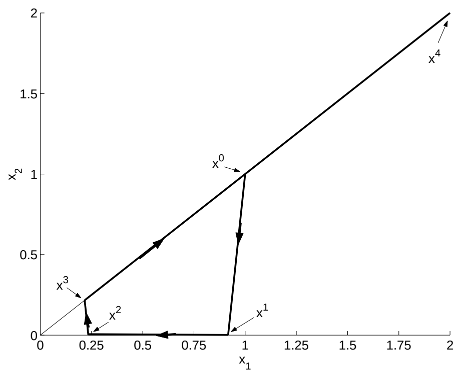

Figure 3: The type 4 path ψ.

5.6 General starting point

In this section we let the overflow path start in a point x^0 in the interior below the diagonal. As above, the overflow point is r^T to be reached T time units later. We assume that x^0 lies to the left from r^T, that is $x_1^0 < r_1^T$. The most likely behavior in x^0 is either towards the diagonal or away from the diagonal (this includes parallel). We shall sketch the possible types of optimal paths. For large T we have a generic result.

Most likely behavior away from the diagonal
Three types of paths are candidates:

1. The path connects x^0 with a diagonal point r^1 (including 0) through a straight line of duration t; arriving at r^1 it immediately continues with constant speed along the diagonal to the overflow point. During the first part, we deal with the local process with all arrivals sent to queue 2; the second part is equivalent to the reduced queue. Hence, the cost equals

$$I^1(x_1^0, r_1^1, t) + I^{MM1}(x_2^0, r_2^1, t; \mu_2) + I^r(r^1, r^T, T - t). \qquad (15)$$

2. The path connects x^0 with a diagonal point r^1 (including 0) via the boundary; arriving at r^1 it immediately continues with constant speed along the diagonal to the overflow point. This path is preferable above type 1 if (and only if) it moves to and proceeds along the boundary most likely, and goes up to the diagonal optimally (when possible in the given time T) . Because, if the time permits such an excursion, than certainly it gives the least cost for the part until point r^1. And if

time does not permit it, then the straight line of type 1 is better. Thus the cost is

$$I^{MM1}(0, r_2^1, t; \mu_2) + I^r(r^1, r^T, T - t).$$

Notice that the diagonal point r^1 and the time of arriving t are related because the first coordinate decreases with most likely speed μ_1: $r_1^1 = x_1^0 - t\mu_1$.

3. The path returns to 0, in the most likely manner, waits in 0, and prolongs along the diagonal optimally. We have seen these paths previously, the cost is

$$I^r(0, r^T, (r_1^T + r_2^T)/(\mu_1 + \mu_2 - \lambda)) = (r_1^T + r_2^T) \log \frac{\mu_1 + \mu_2}{\lambda}.$$

Similarly to Theorem 16 one can show that the type 3 paths are optimal for large T. However, there is a remarkable correspondence between optimal paths in this case of general starting point and in case of diagonal starterts. Define r^0 to be the diagonal point where the diagonal intersects with the line $x^0 + (\mu_1, \mu_2 - \lambda)t$ (from x^0 moving upwards against the most likely behavior direction). Let $t_0 = (r_1^0 - x_1^0)/\mu_1$ be the most likely time to get to x^0 from r^0. Then we claim:

- Type 2 is optimal if and only if type 4 in Section 5.2 is optimal when starting in r^0 and having overflow time $T + t_0$. (And they have the same cost.)
- Type 3 is optimal if and only if type 7 in Section 5.2 is optimal when starting in r^0 and having overflow time $T + t_0$. (And they have the same cost.)

The proof of these claims is based on the principle of optimality. Since there is only one type left, we also say

- Type 1 is optimal if and only if type 1 in Section 5.2 is optimal when starting in r^0 and having overflow time $T + t_0$.

The optimal path of this type and its cost are obtained by minimizing (15) simultaneously with respect to r^1 and t.

Most likely behavior towards the diagonal
This is easy. Two candidates of optimal paths:

1. The path connects x^0 with a diagonal point r^1 (including 0) through a straight line of duration t, and which is not the most likely path; arriving at r^1 it immediately continues with constant speed along the diagonal to the overflow point. As (15), the cost is

$$I^1(x_1^0, r_1^1, t) + I^{MM1}(x_2^0, r_2^1, t; \mu_2) + I^r(r^1, r^T, T - t).$$

2. Let r^0 be the point where the diagonal intersects the line $x^0 - (\mu_1, \mu_2 - \lambda)t$ (from x^0 following the most likely behavior). Let $t_0 = (x_1^0 - r_1^0)/\mu_1$ be the most likely time to get to x^0 from r^0. The path does this, and from r^0 it follows the optimal path to overflow in $T - t_0$ time units as found in Section 5.2. The cost is denoted as $I_{r^0}(T - t_0)$.

The best type 1 path is obtained by minimizing the cost with respect to r^1 and t. This is to be compared with the cost $I_{r^0}(T - t_0)$ of type 2 path. Clearly there is some T_2 such that when $T > T_2$ type 1 cannot be optimal.

6 Extension to higher dimensions

As we have seen in subection 3.2 the results for the most likely behavior extend naturally, and the complexity grows but is quite manageable. However, for the overflow problem the situation is more complex.

The methods of this paper apply to the higher dimensional system and they show that the rate function is the correct (asymptotic-logarithmic) value of the probabilities of overflow, even though the general large deviations principle may not hold. The main results, namely listed in sections 4 and 5 hold as stated. Consequently, the infinite-dimensional optimization problem (finding the optimal path) is reduced to a finite dimensional problem: finding the parameters of the optimal path, which is constructed from a finite collection of straight lines with fixed speed. Moreover, the rate function for each such line possesses an explicit expression, since it is the rate for a collection of independent processes, where one process is an M/M/1 queue and the rest are Poisson processes. However, it is also clear that the complexity of the problem increases very fast with the dimension of the system, since the number of possibilities increases rapidly, and complicated paths, such as in subsection 5.5 need to be considered. Fortunately, the behavior of this system is relatively simple since increase may only happen along a one-dimensional surface—the diagonal.

References

[1] Atar, R. and P. Dupuis (1999) Large deviations and queueing networks: methods for rate identification. *Stoch. Proc. Appl.* 84: 255–296.

[2] Bertsimas D., L. Paschalides and J.N. Tsitsiklis (1999) Large deviations analysis of the generalized processer sharing policy. *Queueing Systems* 32: 319–349.

[3] Chang, C-S. (1995) Sample path large deviations and intree networks. *Queueing Systems* 20: 7–36.

[4] Dembo, A. and O. Zeitouni (1996) *Large deviations techniques and applications*, Second edition. Springer Verlag.

[5] Dupuis, P. and R.E. Ellis (1995) The large deviations principle for a general class of queueing systems I. *Trans. American Math. Soc.* 347: 2689–2751.

[6] Dupuis, P. and R.E. Ellis (1995) *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley.

[7] Foley, R.D. and D. McDonald (2001) Join the shortest queue: stability and exact asymptotics. *Ann. Appl. Prob.* 11: 569–607.

[8] Ganesh, A.J. and N. O'Connell (2002) A large deviation principle with queueing applications. *Stoch. and Stoch. Reports* 73: 25–35.

[9] Ganesh, A.J., N. O'Connell and D. Wischik (2004) *Big Queues*. Springer.

[10] I. Ignatiouk-Robert (2004) Large deviations for processes with discontinuous statistics. *Ann. Prob.* to appear.

[11] Mandjes, M. (1999) Rare event analysis of the state frequencies of a large number of markov chains. *Stoch. Models* 15: 577–592.

[12] Mandjes, M. and A. Ridder (1999) Optimal trajectory in a queue fed by a large number of sources. *Queueing Systems* 31: 137–170.

[13] Mandjes, M. and A. Weiss (2003) Sample path large deviations of a multiple time-scale queueing model. Preprint.

[14] McDonald, D. (1996) Overloading parallel servers when arrivals join the shorest queue. pp. 169–196 In: Glasserman, P., K. Sigman and D. Yao (ed.) *Stochastic Networks: stability and rare events*. Springer Verlag.

[15] Puhalskii, A.A. and W. Whitt (1998) Functional large deviation principles for waiting and departure processes. *Prob. Engin. Info. Sci.* 12: 479–507.

[16] Shwartz, A. and A. Weiss (1995) *Large deviations for performance analysis: queues, communication and computing*. Chapman Hall.

[17] Shwartz, A. and A. Weiss (1999) Multiple time scales in Markovian ATM models I. Formal calculations. CC PUB 267, Elec. Engin. Technion.

[18] Shwartz, A. and A. Weiss (2005) Large Deviations with diminishing rates. to appear, *Math. of Oper. Res.*.

[19] Stolyar, A.L. (2003) Control of End-to-End Delay Tails in a Multiclass Network: LWDF Discipline Optimality. *Ann. Appl. Prob.* 13: 1151–1206.

[20] Stolyar A.L. and K. Ramanan (2001) Largest Weighted Delay First Scheduling: Large Deviations and Optimality. *Ann. Appl. Prob.* 11: 1–48.

[21] Turner, S. (2000) Large deviations for join the shorter queue. In: McDonald, D. and S. Turner (ed) *Analysis of Communication Networks: Call Centres, Traffic and Performance*. The Fields Institute.

[22] Wischik D.J. (2001) Sample path large deviations for queues with many inputs. *Ann. Appl. Prob.* 11: 379–404.

[23] Weiss, A. (1986), A new technique for analyzing large traffic systems. *Adv. Appl. Prob.* 18: 506–532.