IIAS-STP-00-13

Large Deviations in Some Queueing Systems

N. D. Vvedenskaya E. A. Pechersky* Yu. M. Suhov

Logarithmic asymptotics of probabilities of large delays are derived for the "last come–first served" system and system with priorities. Trajectories that determine the mean dynamics of arrival flow under the condition of large delay are described.

1 Introduction

We investigate probabilities of rear events in several queueing systems. More precisely, we are interested in the probability of large delay. One-server systems, which obey various service rules, are considered. This paper is mainly devoted to two rules: (1) the "last come–first served" rule and (2) rule with priorities. It is assumed that all systems under consideration are in a stationary regime. Our results have the following form. For a random variable ω equal to the delay in a system, the limit

$$\lim_{x\to\infty}\frac{1}{x}\ln\Pr(\omega > ax) \tag{1}$$

is found under the conditions of a Poisson arrival flow and the existence of exponential moments of message lengths.

Such problems belong to the theory of large deviations (see [1, 2]). Our method is based on the large deviation principle from [3]. In this framework, a Poisson process which describes an arrival flow is constructed. In terms of this process, the event under consideration, namely, a large delay, is represented. After that, the infimum of the rate function on this event is found, whose

*The first two authors supported in part by the Russian Foundation for Basic Research, project nos. 96-01-10020 and 99-01-00003.

value is equal to the value of the limit (1). Such approach permits us also to identify the trajectory on which the infimum is achieved. We call this trajectory the *optimal* one (see Theorems 3 and 4). In accordance with the large deviation principle, one can conclude that under the condition of large delay a trajectory of the Poisson process is close to the optimal one with large probability.

Apparently, results that give the limit can be derived from the known exact solutions for the systems under consideration (see [4]). But such an approach does not seem to be simpler (see, for example, the asymptotic analysis in [5], from which one can derive the asymptotics of the delay probability in a "first come–first served" system). Also, as was already mentioned, our approach permits us to identify the conditional mean dynamics of arrival flow under the condition of large delay, which is a more difficult problem in the case where exact solutions are used.

2 Main Results

As usual, the arrival flow is described by a marked Poisson point flow

$$\Xi = (t_i, \xi_i)_{i=-\infty}^{\infty}. \tag{2}$$

Below, assumptions imposed on Ξ are presented:

A1. The sequence of random variables (ξ_i) with values in $\mathbb{R}_+$ is formed by identically distributed independent random variables. There exists $\theta_+ > 0$ such that

$$\varphi(\theta) = \mathbf{E}\exp\{\theta\xi_1\} < \infty \quad \text{for} \quad \theta < \theta_+. \tag{3}$$

A2. The sequences (t_i) and (ξ_i) are independent.

A3. The sequence of random variables (t_i) forms a stationary Poisson point process on $\mathbb{R}$, where $t_i < t_{i+1}$ and $t_{-1} < 0 \le t_0$. The rate of this process is denoted by λ.

We assume that queueing systems under considerations have stationary regimes. The following condition guarantees this property.

A4. $\lambda\varphi'(0) < 1$.

Finally, we assume that

A5. $\lim\limits_{\theta\uparrow\theta_+} \varphi(\theta) = \infty$.

The main functional to be investigated here is the message delay in a system with a queue. It is convenient to define this functional not in terms of the process Ξ but in terms of the generalized Poisson process

$$\sigma(t) = \begin{cases} \sum\limits_{i:\,0<t_i\le t} \xi_i & \text{if } t > 0, \\ -\sum\limits_{i:\,t<t_i\le 0} \xi_i & \text{if } t \le 0. \end{cases} \tag{4}$$

It is assumed that the sum of a zero number of summands is equal to 0. Observe that the process $\sigma(t)$ is defined on the whole axis $\mathbb{R}$. Below, the word 'generalized' is omitted, and all processes considered are called Poisson processes.

A delay of the so-called virtual message is investigated. A virtual message has no length and is associated with some nonrandom time instant (most often, this is the instant 0). It is assumed that at this instant a virtual message arrives and waits to be served obeying the service rule. Being of zero length, a virtual message does not affect delays of real messages nor occupy any buffer space. For example, in a system that obeys the "first come–first served" (FCFS) rule, the delay of a virtual message, which arrives at instant t, is

$$\nu(t) = \sup_{\tilde{t}<t} \left\{ \sigma(t) - \sigma(\tilde{t}) - (t - \tilde{t}) \right\}. \tag{5}$$

By condition **A4**, $\nu(t)$ is finite with probability one.

To define a virtual-delay process in a system with the "last come–first served" (LCFS) rule, the process $\delta(\cdot)$ is used. Its value at instant t equals the time until the service end of a message processed at t. The precise definition of the process $\delta(\cdot)$ and some of its properties are given in Section 3. The virtual-message delay in an LCFS system at instant t is

$$\omega(t) = \inf \left\{ \hat{t} \ge 0 : \delta(t) + \sigma(t + \hat{t}) - \sigma(t) - \hat{t} < 0 \right\}. \tag{6}$$

This delay is the sum of the time $\delta(t)$ and service time of messages that arrive after t and are processed without interruption.

In the following theorem, we present a logarithmic asymptotics of the probability that the delay $\omega = \omega(0)$ of a virtual message that arrive at instant 0 is large.

Theorem 1 *For any $a > 0$,*

$$\lim_{x\to\infty} \frac{1}{x} \ln \Pr(\omega \ge ax) = -a \left(\theta^1 - \lambda[\varphi(\theta^1) - 1] \right), \tag{7}$$

where θ^1 is the solution of the equation

$$\lambda\varphi'(\theta) = 1. \tag{8}$$

Observe that, by **A5**, equation (8) always possesses a solution.

It is well known (see, for example, [3, Theorem 3]) that

$$\lim_{x\to\infty} \frac{1}{x} \ln \Pr(\nu(0) \geq ax) = -a\theta^*, \tag{9}$$

where θ^* is the positive root of the equation

$$\theta = \lambda[\varphi(\theta) - 1], \tag{10}$$

which exists by **A4** and **A5**. The expression in (9) presents the asymptotics of the large-delay probability in an FCFS system. It is easy to understand that $\theta^* > \theta^1$ and, moreover,

$$\theta^* > \theta^1 - \lambda\big[\varphi(\theta^1) - 1\big]. \tag{11}$$

Therefore, the large-delay probability in an FCFS system is less than in an LCFS system. The mean delay values in these systems are equal but the delay variance in an LCFS system is larger than in an FCFS system (see [6, 7]). Relation (11) indicates that from the large-deviation point of view, an LCFS system is also worse than an FCFS one.

The following result gives a detailed description of the large-delay probability for a virtual message of the lowest priority in a system with priorities. It suffices to consider a system with two types of messages, where one type has a priority. One may assume that there are two queues with messages of only one type in each queue. If, by the end of processing a message, the queue of messages of the higher priority is nonempty, then the next message of this queue starts its service obeying the FCFS rule. In the opposite case, a message from the lower priority queue starts its service obeying the FCFS rule (of course, if this queue is nonempty). Thus, two independent flows $\Xi_1 = (t_i^1, \xi_i^1)$ and $\Xi_2 = (t_i^2, \xi_i^2)$ with different priorities arrive at the system. The cumulative arrival flow $\Xi = (t_i, \xi_i) = \Xi_1 \vee \Xi_2$ is the "superposition" of two flows. The easiest way to define this notion is to pass to the corresponding Poisson processes $\sigma_1(t)$ and $\sigma_2(t)$ that are defined by the flows Ξ_1 and Ξ_2 according to (4). Thus, the process $\sigma(t)$ that corresponds to the cumulative flow Ξ is $\sigma(t) = \sigma_1(t) + \sigma_2(t)$; i.e., it is also Poisson flow. Let λ_1 and λ_2 be the flow rates of Ξ_1 and Ξ_2 respectively, and let $\theta_+^j > 0$, $j = 1, 2$, be such that $\varphi_j(\theta) = \mathbf{E}e^{\theta\xi_1^j} < \infty$ if $\theta < \theta_+^j$. Then the flow rate of Ξ is equal to $\lambda_1 + \lambda_2$, and the Laplace transform of a jump ξ_1 of $\sigma(t)$ is

$$\varphi(\theta) = \frac{\lambda_1}{\lambda_1 + \lambda_2}\varphi_1(\theta) + \frac{\lambda_2}{\lambda_1 + \lambda_2}\varphi_2(\theta) \tag{12}$$

if $\theta < \theta_+ = \min\{\theta_+^1, \theta_+^2\}$. We assume that conditions **A1**–**A5** are fulfilled for each of the flows Ξ_j, $j = 1, 2$. Therefore, they are fulfilled for the cumulative flow Ξ. Assume that messages from Ξ_2 have the lower priority. Then the delay of virtual message from this flow that enters the system at time instant t is

$$\nu_2(t) = \inf\left\{\hat{t} \geq 0 : \nu(t) + \sigma_1\big(t + \hat{t}\big) - \sigma_1(t - 0) - \hat{t} \leq 0\right\}; \tag{13}$$

here, $\nu(t)$ is defined by formula (5), where $\sigma(t)$ is now the cumulative flow. One can see from this formula that the delay of a virtual message with the lower priority that arrives at instant t is a sum of two quantities: the workload accumulated at time t and the workload that is brought by the first-priority messages that enter the system after t until the service end. As in Theorem 1, the delay $\nu_2 = \nu_2(0)$ is investigated at time instant 0.

Theorem 2 *For any $a > 0$,*

$$\lim_{x\to\infty} \frac{1}{x} \ln \Pr(\nu^2 \geq ax) = -a\Big(\hat{\theta} - \lambda_1 \left[\varphi_1(\hat{\theta}) - 1\right]\Big), \tag{14}$$

where $\hat{\theta} = \min\{\theta^*, \theta_1^1\}$, θ_1^1 is the solution of the equation

$$\lambda_1\varphi_1'(\theta) = 1, \tag{15}$$

and θ^* is the solution of equation (10) for $\lambda = \lambda_1 + \lambda_2$ and φ defined by (12).

By **A4**–**A5**, equations (15) and (10) possess solutions. Note that, as $\lambda_2 \to 0$ and $\lambda_1 \to \lambda$, the right-hand side of (14) tends to the right-hand side of (9).

Process $\delta(\cdot)$

To construct the process $\delta(\cdot)$, we divide the time axis into two regions: the region $\mathbf{R}^{(0)}$, where the server is idle; and its complement $\mathbf{R}^{(1)}$, the busy region. In an FCFS system,

$$\mathbf{R}^{(0)} = \{t : \nu(t) = 0\}.$$

It is well known that busy and idle periods are the same for all systems with conservative service.[1] Therefore, the set $\mathbf{R}^{(0)}$ is also the set of idle periods of an LCFS system. Since trajectories of the process $\sigma(t)$ are monotone step functions, it is easily seen that $\mathbf{R}^{(0)}$ is the union of half-open intervals of the $[u, v)$ type. Note that by **A4** all intervals in $\mathbf{R}^{(0)}$, as well as in $\mathbf{R}^{(1)}$, are finite with probability one. Intervals that form the set $\mathbf{R}^{(1)}$ are called busy intervals. Thus,

$$\mathbf{R}^{(1)} = \bigcup_{j=-\infty}^{\infty} I_j^1,$$

where I_j^1 is a busy interval. The set of message arrival times $(t_i) \subset \mathbf{R}^{(1)}$. Therefore, the whole sequence (t_i) is split into finite subsets $F_j = \{t_{k_j}, t_{k_j+1}, \dots, t_{k_{j+1}-1}\}$, which include those and only those points of (t_i) that belong to some interval I_j^1. We rearrange the points of this set to put them into an order in which corresponding messages are processed by the LCFS system. Let the rearranged sequence be $(s_{k_j}, s_{k_j+1}, \dots, s_{k_{j+1}-1})$. It is clear that $s_{k_j} = t_{k_j}$ is the arrival time of the first message of the group that forms the busy period I_j^1. The server processes this message for some time ξ_{k_j} during the interval $B_j^0 = (s_{k_j}, s_{k_j} + \xi_{k_j}]$. During this interval, new messages arrive at the system. Let them arrive at instants $F_j^0 = \{t_{k_j+1}, \dots, t_{k_j+r_0}\}$. If F_j^0 is empty, then $t_{k_j+1} > t_{k_j} + \xi_{k_j}$ and $t_{k_j} + \xi_{k_j}$ is the right end of the busy interval I_j^1; thus, F_j consists of one point t_{k_j}. In the case $F_j^0 \neq \emptyset$, the next message to be served is the one that arrived at instant $t_{k_j+r_0}$, that is, $s_{k_j+1} = t_{k_j+r_0}$. Let $\tilde{F}_j^1$ be the set of arrival times $\{t_{k_j}, \dots, t_{k_j+r_0}, t_{k_j+r_0+1}, \dots, t_{k_j+r_0+r_1}\}$ of all messages that arrived at the system during the time interval $B_j^1 = (s_{k_j}, s_{k_j} + \xi_{k_j} + \xi_{k_j+r_0}]$. Introduce a set $F_j^1 = \tilde{F}_j^1 \setminus \{t_{k_j+r_0}\}$ of arrival times of messages that arrived during the interval B_j^1 and are not processed during this interval. Note that r_1 can be equal to 0. After the service end of a message that arrived at instant $t_{k_j+r_0}$, the next message to be served is a message with the last arrival time among all arrival times from F_j^1. Therefore, we have

$$s_{k_j+2} = \begin{cases} t_{k_j+r_0+r_1} & \text{if } r_1 \neq 0, \\ t_{k_j+r_0-1} & \text{if } r_1 = 0. \end{cases}$$

Continuing this construction, we get a rearranged sequence $(s_{k_j}, \dots, s_{k_{j+1}-1})$. Let $\eta_1^j, \dots, \eta_{k_{j+1}-k_j}^j$ be the lengths of messages that arrived at instants

[1] A service is conservative if the server is idle in the case where there is no work only.

$s_{k_j}, \ldots, s_{k_{j+1}-1}$ respectively. By definition,

$$\delta(t) = \begin{cases} 0 & \text{if } t \in \mathbf{R}^{(0)}, \\ t_j^+ + \sum\limits_{k=1}^{r} \eta_k^j - t & \text{if } t \in [t_j^-, t_j^+) \subset \mathbf{R}^{(1)} \text{ and } t_j^- + \sum\limits_{k=1}^{r-1} \eta_k^j \le t < t_j^- + \sum\limits_{k=1}^{r} \eta_k^j. \end{cases}$$

An important relation is given by the following lemma.

Lemma 1 *With probability one, $\forall t \in \mathbb{R}$,*

$$\delta(t) \le \nu(t). \tag{16}$$

A proof is required for the case where $t \in \mathbf{R}^{(1)}$ only. First of all, note that for any $t \in [t_j^-, t_j^+)$, we have $\nu(t) = \sigma(t) - \sigma(t_j^- - 0) - (t - t_j^-)$. Let a message of length η be processed at instant $t \in [t_j^-, t_j^+)$, and let its processing start at $\hat{t} \le t$. Recall that $\delta(t) = \eta - (t - \hat{t})$. Before the instant $\hat{t}$, the server "non-stop" processed messages that had arrived before $\hat{t}$. This set of messages does not include the message of length η whose service started at instant $\hat{t}$. Therefore,

$$\sigma(\hat{t}) - \sigma(t_j^- - 0) - (t - \hat{t}) - \eta \ge 0. \tag{17}$$

Using (17) and the monotonicity of $\sigma(t)$, one gets

$$\begin{aligned} \nu(t) &= \left[\sigma(t) - \sigma(\hat{t}) + \eta - (t - \hat{t})\right] + \left[\sigma(\hat{t}) - \sigma(t_j^- - 0) - (\hat{t} - t_j^-) - \eta\right] \\ &\ge \eta - (t - \hat{t}) = \delta(t). \end{aligned}$$

3 Reduction to the Large Deviation Principle

Proofs of the theorems formulated above are based on the large deviation principle. We describe this reduction for a system with priorities, which appears in Theorem 2. The proof of Theorem 1 uses the same ideas and some estimates presented in Section 4.

To arrival processes Ξ_1, Ξ_2, and Ξ at a system with priorities, we assign the sequences of Poisson processes $\sigma_j^n(t) = \dfrac{\sigma_j(nt)}{n}$, $j = 1, 2$, and $\sigma^n(t) = \sigma_1^n(t) + \sigma_2^n(t) = \dfrac{\sigma(nt)}{n}$. Introduce the following functionals on these processes:

$$\begin{aligned} \nu_n(t) &= \sup_{\tilde{t} \ge 0} \left\{\sigma^n(t) - \sigma^n(t - \tilde{t}) - \tilde{t}\right\}, \\ \nu_{n,2}(t) &= \inf\left\{\tilde{t} \ge 0 \; \nu_n(t) + \sigma_1^n(t + \tilde{t}) - \sigma_1^n(t) - \tilde{t} \le 0\right\}. \end{aligned}$$

Lemma 2 *The following identity of events takes place:*

$$\big(\nu_2(0) \geq an\big) = \big(\nu_{n,2}(0) \geq a\big).$$

The proof follows from the equalities $\nu_n(t) = \dfrac{1}{n}\nu(nt)$ and $\nu_{n,2}(t) = \dfrac{1}{n}\nu_2(nt)$, which in their turn follow from the definitions.

For $x_1 > x_2$, the following inclusion takes place:

$$\big(\nu_2(t) \geq ax_1\big) \subseteq \big(\nu_2(t) \geq ax_2\big).$$

Therefore, it suffices to find the limit of the left-hand side of (14) for a natural $x = n$.

Let us introduce the space $\mathcal{X}$ of the nondecreasing-on-$\mathbb{R}$ functions with numerical values which satisfy the following conditions: If $x \in \mathcal{X}$, then

B1. $\lim\limits_{u \downarrow t} x(u) = x(t)$,

B2. $\lim\limits_{t \uparrow 0} x(t) \leq 0 \leq x(0)$.

Use a simple generalization of the topology introduced in [3, Section 2] (see also [8]). In [3], the space of trajectories defined on the semiaxis $[0, \infty)$ is considered. A generalization to the case considered here, i.e., to trajectories defined on the whole axis $\mathbb{R}$, is obvious. To give a short description of this topology, one can say that, in this topology, convergence of trajectories from $\mathcal{X}$ restricted onto a compact subset in $\mathbb{R}$ is equivalent to weak convergence of these restrictions. But this topology is stronger than that generated by the weak convergence of restrictions on compacts.

A sequence of processes $(\sigma_1^n(t), \sigma_2^n(t))$ generates a sequence of probability distributions P_n on a Cartesian product $\mathcal{X}^2$. Here $P_n \Rightarrow \delta_{(l_1,l_2)}$ as $n \to \infty$, where $\delta_{(l_1,l_2)}$ is a Dirac measure, $\Rightarrow$ denotes the weak convergence of measures on $\mathcal{X}^2$ with respect to the topology mentioned above, and l_1 and l_2 are linear functions:

$$l_1(t) = \lambda_1\varphi_1'(0)t, \qquad l_2(t) = \lambda_2\varphi_2'(0)t.$$

In terms of the space $\mathcal{X}^2$, the event $(\nu_{n,2} \geq a)$ is a set of trajectories

$$\mathcal{U}_a = \big\{(x_1, x_2) : V^2(x_1, x_2) \geq a\big\},$$

where

$$V^2(x_1, x_2) = \inf_{\hat{t}} \big\{\hat{t} : N(x_1, x_2) + x_1(\hat{t}) - x_1(0) - \hat{t} \leq 0\big\} \text{ and} N(x_1, x_2)$$

$$= \sup_{t\le 0}\{x_1(0)+x_2(0)-x_1(t)-x_2(t)+t\}.$$

Thus, $\Pr(\nu_{n,2}\ge a) = P_n(\mathcal{U}_a)$. Such reduction permits us to use the large deviation principle from [3], which implies the following inequalities:

$$-\inf_{(x_1,x_2)\in\overset{\circ}{\mathcal{U}}_a} I(x_1,x_2) \le \lim_{n\to\infty}\frac{1}{n}\ln P_n(\mathcal{U}_a) \le -\inf_{(x_1,x_2)\in\overline{\mathcal{U}}_a} I(x_1,x_2). \qquad (18)$$

In this expression, $\overset{\circ}{\mathcal{U}}_a$ and $\overline{\mathcal{U}}_a$ are the interior and closure of a set $\mathcal{U}_a$ respectively. The rate functional I has the form

$$I(x_1,x_2) = \int_{-\infty}^{\infty}\Lambda(\dot{x}_1(t),\dot{x}_2(t))dx, \qquad (19)$$

where $\dot{x}_i$ is the derivative of x_i with respect to t and

$$\Lambda(s_1,s_2) = \sup_{\theta_1,\theta_2}\Big\{s_1\theta_1 + s_2\theta_2 - \lambda_1[\varphi_1(\theta_1)-1] - \lambda_2[\varphi_2(\theta_2)-1]\Big\},$$

where $s_1, s_2 \in \mathbb{R}^1$. These formulas are valid for a pair of absolutely continuous trajectories (x_1,x_2). We do not present here the general definition of a rate function because it is not needed in this paper. One can find in [3] (see also [8] and [9]) a detailed description of the rate functional defined on trajectories along a semiaxis.

Any trajectory $(x_1,x_2)\in\mathcal{X}^2$ specifies the *macro-scale* behavior of our process. This means that the probability that a process (σ_1^n,σ_2^n) is located in a small neighborhood of (x_1,x_2) is approximately equal to $e^{-nI(x_1,x_2)}$.

The following theorem characterizes the behavior of the mean arrival flow under the condition that a large deviation of a delay value takes place. The theorem can be considered as an addition to Theorem 2.

Recall that θ^* and θ_1^1 are solutions of equations (10) and (15) respectively.

Theorem 3 *The infimum on the left- and right-hand sides of* (18) *is achieved on the trajectory* $(x_1^*(t),x_2^*(t))$ *whose derivatives are defined in the following way:*

If $\theta^* < \theta_1^1$, *then*

$$\dot{x}_1^*(t) = \begin{cases} \lambda_1\varphi_1'(\theta^*) & \textit{if } t \in \left(-a\dfrac{1-\lambda_1\varphi_1'(\theta^*)}{\lambda\varphi'(\theta^*)-1}, a\right) \\ \lambda_1\varphi_1'(0) & \textit{otherwise}, \end{cases}$$
$$\dot{x}_2^*(t) = \begin{cases} \lambda_2\varphi_2'(\theta^*) & \textit{if } t \in \left(-a\dfrac{1-\lambda_1\varphi_1'(\theta^*)}{\lambda\varphi'(\theta^*)-1}, 0\right) \\ \lambda_2\varphi_2'(0) & \textit{otherwise}. \end{cases} \tag{20}$$

If $\theta^* \geq \theta_1^1$, *then*

$$\dot{x}_1^*(t) = \begin{cases} 1 & \textit{if } t \in (0, a), \\ \lambda_1\varphi_1'(0) & \textit{otherwise}, \end{cases}$$
$$\dot{x}_2^*(t) = \lambda_2\varphi_2'(0). \tag{21}$$

As one can see from this theorem, in the case $\theta^* < \theta_1^1$, the optimal trajectory that describes the mean dynamics looks as follows. Entering the system at instant 0, a virtual message meets a queue. This queue needs macro-scaled time $a\big(1 - \lambda_1\varphi_1'(\theta^*)\big)$ to be processed. Thus, during this time, the virtual message stays in the system and waits while the messages of both types that have arrived before 0 are processed. During the waiting time, new messages of the first type, which have priority, enter the system. Therefore, the virtual message also waits for the time needed for processing first-type messages that arrive after instant 0 during the waiting time of the virtual message. This waiting time is equal to $a\lambda_1\varphi_1'(\theta^*)$. Note that a trajectory $x_1^*(t)$, which describes the mean flow of first-type messages under the condition of large delay have no break at 0. That is, the arrival rate of these messages is the same during both some interval before the virtual message arrival and some interval after this instant. The trajectory that describes the flow of second-type messages has a break at zero because after the virtual message arrival this flow does not affect the delay value.

In the case $\theta^* \geq \theta_1^1$, the mean dynamics of the arrival flow does not cause a long queue at the arrival time 0. The macro-scale queue is equal to zero. But new messages of the first type can arrive before the service end of a message that was processed at the virtual message arrival time 0. These messages are processed before the virtual message. Such flow dynamics is characterized by the absence of a large queue (the macro-scale queue is zero).

But the waiting time of a virtual message is large. We call such dynamics *sliding dynamics.*

Theorems 2 and 3 Our goal is to find the minimums in (18). To describe the sets $\overset{\circ}{\mathcal{U}}_a$ and $\overline{\mathcal{U}}_a$, some functionals on $\mathcal{X}^2$ are introduced. Let

$$T(x_1,x_2) = \inf\{t \geq 0 : N(x_1,x_2) + x_1(t) - x_1(0) < t\},$$
$$\hat{T}(x_1,x_2) = \inf\{t \geq 0 : N(x_1,x_2) + x_1(t) - x_1(0) \leq t\}.$$

Note that N is a continuous functional in the topology of $\mathcal{X}^2$; hence, $\{(x_1,x_2) : N(x_1,x_2) > h\}$ is an open set, and $\{(x_1,x_2) : N(x_1,x_2) \geq h\}$ is closed. The functionals T and $\hat{T}$ are discontinuous but T is semicontinuous from above and $\hat{T}$ is semicontinuous from below, i.e.,

$$\liminf_{n\to\infty} T(x_1^n, x_2^n) \leq T(x_1,x_2), \qquad \limsup_{n\to\infty} \hat{T}(x_1^n, x_2^n) \geq \hat{T}(x_1,x_2) \tag{22}$$

if $(x_1^n, x_2^n) \to (x_1,x_2)$ as $n \to \infty$ (details can be found in [10], where functionals of such kind are considered). Let

$$\mathcal{W}_h = \{(x_1,x_2) : T(x_1,x_2) > h\}, \qquad \hat{\mathcal{W}}_h = \{(x_1,x_2) : \hat{T}(x_1,x_2) > h\},$$

where $h > 0$. By (22), the set $\overline{\mathcal{W}}_h = \{(x_1,x_2) : T(x_1,x_2) \geq h\}$ is closed. Since $\mathcal{U}_a = \overline{\mathcal{W}}_a$, we have that $\mathcal{U}_a$ is also a closed set.

Next, consider the set $\overset{\circ}{\mathcal{U}}_a$. Note first that $\hat{\mathcal{W}}_h$ is open. Since $\hat{T} \leq T$, we have $\hat{\mathcal{W}}_h \subseteq \mathcal{W}_h$ for $h > 0$ and therefore $\hat{\mathcal{W}}_a \subseteq \mathcal{U}_a$. Since $\hat{\mathcal{W}}_a$ is an open set, which is contained in $\mathcal{U}_a$, we have $\hat{\mathcal{W}}_a \subseteq \overset{\circ}{\mathcal{U}}_a$. Now let us calculate

$$\overset{\circ}{I} = \inf\{I(x_1,x_2) : (x_1,x_2) \in \widehat{\mathcal{W}_a}\} \tag{23}$$

and

$$\overline{I} = \inf\{I(x_1,x_2) : (x_1,x_2) \in \overline{\mathcal{W}}_a\} \tag{24}$$

and show that $\overset{\circ}{I} = \overline{I}$. To calculate (23), the set $\hat{\mathcal{W}}_a$ is represented as a continual union

$$\hat{\mathcal{W}}_a = \bigcup_{\substack{g>h>0 \\ g\geq a}} (\mathcal{V}_h \cap \mathcal{T}_g),$$

where $\mathcal{V}_h = \{(x_1,x_2) : N(x_1,x_2) = h\}$ and $\mathcal{T}_g = \{(x_1,x_2) : \hat{T}(x_1,x_2) = g\}$. Hence,

$$\overset{\circ}{I} = \inf_{\substack{g>h>0 \\ g\geq a}} J(h,g), \quad J(h,g) = \inf\left\{I(x_1,x_2) : (x_1,x_2) \in \mathcal{V}_h \cap \mathcal{T}_g\right\}. \tag{25}$$

Let us show that (25) is achieved on the trajectory (x_1^*, x_2^*) introduced in (20) and (21).

The event $\mathcal{V}_h$ means that, by instant 0, a queue accumulated in the system needs time h to be processed. This queue consists of messages of both types. The event $\mathcal{V}_h$ is determined by the behavior of the trajectory $(x_1, x_2) \in \mathcal{V}_h$ before 0. Therefore, the rate functional depends on the derivatives $(\dot{x}_1, \dot{x}_2)$ before 0 (see (19)). The event $\mathcal{T}_g$ is determined by the trajectory $(x_1, x_2) \in \mathcal{T}_g$ after 0. Therefore, the rate functional depends on the derivatives $(\dot{x}_1, \dot{x}_2)$ in the region $\mathbb{R}_+$. In fact, the rate functional does not depend on $\dot{x}_2$ in the positive region. Since the rate functional is an integral over $\mathbb{R}$, by finding its minimum for fixed h and g we can minimize the parts of this integral over $(-\infty, 0)$ and $(0, \infty)$ separately. Let (x_1^0, x_2^0) be the trajectory that minimizes (19) for fixed h and g. Then

$$\dot{x}_1^0(t) + \dot{x}_2^0(t) = \begin{cases} \lambda\varphi'(\theta^*) & \text{if } t \in \left(-\dfrac{h}{\lambda\varphi'(\theta^*) - 1}, 0\right), \\ \lambda\varphi'(0) & \text{if } t < -\dfrac{h}{\lambda\varphi'(\theta^*) - 1}. \end{cases} \tag{26}$$

On the interval $\left(-\infty, -\dfrac{h}{\lambda\varphi'(\theta^*) - 1}\right)$, this trajectory does not contribute to the rate functional.

On the positive axis,

$$\dot{x}_1^0(t) = \begin{cases} \dfrac{g-h}{g}t & \text{if } t \in (0, g), \\ \lambda_1\varphi_1'(0) & \text{if } t > g \end{cases} \tag{27}$$

(see [8, Lemma 5.5]). On the interval $(0, g)$, the second component $x_2^0(t)$ does not contribute to the value of $J(h, g)$, therefore $\dot{x}_2^0(t) = \lambda_2\varphi_2'(0)$ at this interval.

The value of $I(x_1^0, x_2^0)$ is a sum of tree summands, $I_1 + I_2 + I_3$, corresponding to three regions: $\left[-\dfrac{h}{(\lambda_1 + \lambda_2)\varphi'(\theta^*) - 1}, 0\right]$, $[0, g]$, and the complement to these intervals. The third summand I_3 is equal to zero by (26). The first summand is $I_1 = h\theta^*$ (see [3, Section 7]). The second summand is $I_2 = (g - h)\theta^0 - \lambda_1 g[\varphi_1(\theta^0) - 1]$, where θ^0 is the unique solution of the equation $1 - \dfrac{h}{g} = \lambda_1 \varphi_1'(\theta)$. Thus, we get

$$J(h, g) = h\theta^* + (g - h)\theta^0 - \lambda_1 g\big[\varphi_1(\theta^0) - 1\big]. \tag{28}$$

Note that $\theta^0 < \theta_1^1$. In the opposite case, for $h \geq 0$ the trajectory $x_1(t)$ would not intersect the line $t - h$ on $[0, \infty)$.

Let us calculate $\overset{\circ}{I}$, which is equal to the minimum of $J(h, g)$ in h and g for $g > h > 0$ and $g \geq a$. First, consider the case $\theta^* < \theta_1^1$. By the strong convexity of $\lambda_1[\varphi_1 - 1]$ and since $\theta^0 < \theta_1^1$, we get $1 > \lambda_1 \varphi_1'(\theta^*)$. The partial derivative of $J(h, g)$ with respect to h is

$$\frac{\partial J}{\partial h} = \theta^* - \theta^0. \tag{29}$$

Setting $\theta^* = \theta^0$, we obtain the optimal value

$$h^0 = g\big(1 - \lambda_1 \varphi_1'(\theta^*)\big). \tag{30}$$

Therefore,

$$J(h^0, g) = g\big(\theta^* - \lambda_1\big[\varphi_1(\theta^*) - 1\big]\big).$$

It follows from the definition of θ^* that the value in the parentheses is positive. Thus, the infimum $J(h^0, g)$ in g is achieved at the minimal possible value $g = a$.

In the case $\theta^* > \theta_1^1$, the derivative (29) is positive for all $\theta^0 \leq \theta_1^1$, which gives $h^0 = 0$. Hence, $\theta^0 = \theta_1^1$ and

$$J(0, g) = g\big(\theta_1^1 - \lambda_1\big[\varphi_1(\theta_1^1) - 1\big]\big). \tag{31}$$

Arguments similar to those presented above give $g = a$. Observe that the trajectory where (31) is achieved does not belong to $\mathcal{W}_a$.

Finally, we get that

$$\overset{\circ}{I} = a\big(\hat{\theta} - \lambda_1\big[\varphi_1(\hat{\theta}) - 1\big]\big),$$

where $\hat{\theta} = \min\{\theta^*, \theta_1^1\}$.

Using similar arguments, one can find that $\overline{I} = \overset{\circ}{I}$.

Now it is easily seen that the trajectory (x_1^*, x_2^*) is optimal. Indeed, in the case $\theta_1^1 > \theta^*$ the optimal value is $g^0 = a$ and, therefore, the optimal queue is $h^0 = a\big(1 - \lambda_1\varphi_1(\theta^*)\big)$ (see (30)). Among the trajectories that, at point 0, have a queue causing the delay h^0, the optimal one is the trajectory for which the sum of the derivatives of the components on the interval $\left(-a\dfrac{1-\lambda_1\varphi_1(\theta^*)}{\lambda\varphi(\theta^*)-1}, 0\right)$ is $\dot{x}_1^*(t) + \dot{x}_2^*(t) = \lambda\varphi'(\theta^*)$. Hence, we get (20) on this interval. Substituting $g = a$ into (27), we get (20) on the interval $(0, a)$. In the case of sliding dynamics and for $\theta_1^1 < \theta^*$, the queue is such that $h^0 = 0$. Therefore, the total delay is determined by the flow of first-type messages only. It is clear that in this case we get the formula (21).

4 Proof of Theorem 1

The proof is based on Theorems 1 and 2 already proved and on two estimates. The first estimate is obtained in Lemma 1. The second one is presented in Lemma 3 (see below). Let

$$\zeta(t) = \inf\Big\{\hat{t} \ge 0 : \ \nu(t) + \sigma\big(t + \hat{t}\,\big) - \sigma(t) - \hat{t} \le 0\Big\}.$$

By inequality (16), $\zeta(t) \ge \omega(t)$ with probability one. Hence,

$$\Pr\big(\omega(0) \ge an\big) \le \Pr\big(\zeta(0) \ge an\big). \tag{32}$$

For any $c > 0$, let

$$\zeta_A(t) = \inf\Big\{\hat{t} \ge 0 : \ A + \sigma\big(t + \hat{t}\,\big) - \sigma(t) - \hat{t} \le 0\Big\}.$$

Lemma 3 *Let $t \in \mathbb{R}$ and let $c > 0$ be such that $p_c = \Pr(\delta(t) > c) > 0$. Then, for any $x > 0$,*

$$\frac{1}{p_c}\Pr\big(\omega(t) \ge x\big) \ge \Pr\big(\zeta_c(t) \ge x\big).$$

Recall that $\sigma(t)$ is a process with independent increments. Therefore, the random variable $\sigma(t + \hat{t}\,) - \sigma(t)$ does not depend on the σ-algebra F_t induced by all random variables $\sigma(s)$ with $s \le t$. The random variable $\delta(t)$ is measurable with respect to F_t. Therefore, $\delta(t)$ and $\sigma(t + \hat{t}\,) - \sigma(t)$ are independent, and hence, random variables $\delta(t)$ and $\zeta_A(t)$ are independent. The following relations between random variables are obvious:

$$\begin{aligned} A_c &= \big(\zeta_A(t) \ge x,\ \delta(t) \ge c\big) \\ &\subset \Big(\inf\Big\{\hat{t} \ge 0 : \ \delta(t) + \sigma\big(t + \hat{t}\,\big) - \sigma(t) - \hat{t} \le 0\Big\} \ge x,\ \delta(t) \ge c\Big) \subset \big(\omega(t) \ge x\big). \end{aligned}$$

By the independence indicated above,

$$\Pr(A_c) \le \Pr\big(\zeta_c(t) \ge x)\big) \Pr\big(\delta(t) \ge c)\big) \le \Pr\big(\omega(t) \ge x)\big).$$

Thus, to find the limit of $\frac{1}{n} \ln \Pr(\omega \ge an)$, it suffices to show that the limits

$$\frac{1}{n} \ln \Pr\big(\varsigma(0) \ge an\big) \tag{33}$$

and

$$\frac{1}{n} \ln \Pr\big(\zeta_c(0) \ge an\big) \tag{34}$$

are equal, where $c > 0$ is such that $\Pr(\delta(0) > c) > 0$.

Let us return for a while to a system with priorities. If the flow with the lower priority is small, then $\theta_1^1 < \theta^*$. For example, this can be achieved taking λ_2 small. One can see that by Theorems 2 and 3 the limit of the expression

$$\frac{1}{n} \ln \Pr(\nu^2 \ge an) \tag{35}$$

does not depend on λ_2 for small values of λ_2 such that $\theta_1^1 < \theta^*$. Therefore, for $\lambda_2 = 0$, the asymptotics is the same. Thus, we get that, as $n \to \infty$, the limit of the expression (33) coincides with that of (35) and is equal to

$$-a\big(\theta^1 - \lambda\big[\varphi(\theta^1) - 1\big]\big). \tag{36}$$

One can find the limit of the sequence (34) by using again the large deviation principle. In this case, the set of trajectories $\mathcal{S}_a \subset \mathcal{X}$ corresponding to the event under consideration is

$$\mathcal{S}_a = \Big\{x : \inf_{t \in [0,a]} \{x(t) - t\} \ge 0\Big\}.$$

Carrying on the analysis similar to that in the proofs of Theorems 2 and 3, we get that the derivative of the optimal trajectory $x(t)$ is

$$\dot{x}^*(t) = \begin{cases} 1 & \text{if } t \in [0, a], \\ \lambda\varphi'(0) & \text{otherwise.} \end{cases} \tag{37}$$

Therefore, the limit of (34) coincides with the limit of (33) and is equal to (36).

In the cause of the proof, we have found the mean dynamics of an LCFS system under the condition of large delay. This is the sliding dynamics, which was already found in a system with priorities in the case of weak low-priority flow rate. This fact is stated by Theorem 4, which is the corollary of Theorem 1.

Theorem 4 *Let* $\mathcal{T}_a = \Big\{x \in \mathcal{X} : \inf_{0 \le t \le a} \{x(t) - t)\} \ge 0\Big\}$. *The macro-scaled set of this trajectories corresponds to the event* $(\omega(0) \ge an)$,

$$\inf_{x \in \mathring{\mathcal{T}}_a} I(x) = \inf_{x \in \overline{\mathcal{T}}_a} = I(x^*),$$

where the optimal trajectory x^* *is defined in* (37).

5 Final Remarks

An analysis similar to that used in this paper can be done for a preemptive LCLF system. Obeying such a rule, a newly arrived message interrupts the processing of the previous message and starts to be processed. The processing of the interrupted message resumes at the first instant t^* when all messages that arrived after the interrupted one, but before t^*, are already processed. In such a system, a virtual message has nonzero length because in the opposite case its delay would always be zero. It is natural to assume that the length of a virtual message is random, distributed according to the arrival-flow messages, and independent of the arrival flow. The message delay $\tau(t)$ in such a system is the time between the instant of its arrival t and the instant when it is completely processed. The asymptotics of a large delay is

$$\lim_{x \to \infty} \frac{1}{x} \ln \Pr\big(\tau(0) > ax\big) = -a\big(\theta^1 - \lambda\big[\varphi(\theta^1) - 1\big]\big),$$

where θ^1 is a solution of equation (8).

The assumption **A5** is mainly technical. Under this assumption, solutions of equations (8), (15), and (10) always exist. All the analysis presented above can be carried out without this assumption. The answers will change slightly. Similar considerations for other models are carried out in [8, 10].

The authors sincerely thank the Dublin School of Theoretical Physics and its director, Professor J.T. Lewis, for the support and warm reception during the authors' visits in 1997–2000. Yu.M. Suhov thanks the IHES for their hospitality during his visits in 1997 and 1998. N.D. Vvedenskaya thanks the London Mathematical Society for the partial support and the Statistical Laboratory of the Department of Pure Mathematics and Mathematical Statistics of the Cambridge University for their reception in 1997 and 1998.

REFERENCES

[1] Deuschel, J.-D. and Stroock, D.W., *Large Deviations*, London: Academic, 1989.

[2] Dembo, A. and Zeituni, O., *Large Deviations: Technique and Applications*, Boston: Johnes & Bartlett, 1992.

[3] Dobrushin, R.L. and Pechersky, E.A., Large Deviations for Random Processes with Independent Increments on Infinite Intervals, *Probl. Peredachi Inf.*, 1998, vol. 34, no. 4, pp. 76–108 [*Probl. Inf. Trans.* (Engl. Transl.), 1998, vol. 34, no. 4, pp. 354–382].

[4] Klimov, G.P., *Stokhasticheskie sistemy obsluzhivaniya* (Stochastic Queueing Systems), Moscow: Nauka, 1966.

[5] Borovkov, A.A., *Veroyatnostnye protsessy v teorii massovogo obsluzhivaniya*, Moscow: Nauka, 1972. Engl. Transl.: *Stochastic Processes in Queueing Theory*, New York: Springer, 1976.

[6] Kingman, J.F., The Effect of Queue Discipline on Waiting Time Variance, *Proc. Cambridge Philos. Soc.*, 1962, vol. 58, pp. 163–164.

[7] Kleinrock, L., *Queueing Systems*, vol. 2: *Computer Applications*, New York: Wiley, 1976.

[8] Dobrushin, R.L. and Pechersky, E.A., Large Deviations for Tandem Queueing Systems, *J. Appl. Math. Stoch. Analysis*, 1994, vol. 7, pp. 301–330.

[9] Lynch, J. and Sethuraman, J., Large Deviations for Processes with Independent Increments, *Ann. Probab.*, 1987, vol. 15, pp. 610–627.

[10] Aspandiarov, S. and Pechersky, E.A., One Large Deviation Problem for Compound Poisson Processes in Queueing Theory, *Markov Proc. Rel. Fields*, 1997, vol. 3, pp. 333–366.

[11] Pechersky, E.A., Suhov, Yu.M., and Vvedenskaya, N.D., Large Deviations for LIFO Protocol and for Protocol with Priorities, *Proc. Int. Symp. Inf. Theory (ISIT-98)*, Cambridge, 1998, p. 222.