Large Deviation Properties of Constant Rate Data Streams Sharing a Buffer with Variable Rate Cross Traffic (Invited Paper)

Kurt Majewski
Siemens AG, CT PP 7
80200 München, Germany
Kurt.Majewski@siemens.com

ABSTRACT

We consider a constant rate data stream which shares a buffer with a variable rate data stream. A first come first serve service discipline is applied at the buffer. After service at the first buffer the variable rate traffic leaves the system, whereas the constant rate traffic is sent to a second buffer. Both buffers provide non-idling service at constant rates and infinite waiting rooms. We model the behavior of the queue lengths as a function of the cumulative variable rate cross traffic arrivals. Under the assumption that the random variable rate cross traffic satisfies an appropriate sample path large deviation principle, we deduce a sample path large deviation principle for the induced queue length processes.

This allows us to investigate logarithmic large deviation asymptotics for the tail probabilities of the steady-state queue length distribution at the second buffer. We show that these asymptotics can be obtained as the solution of a two-dimensional minimization problem. We explicitly calculate rates and associated minimizing paths when the variable rate cross traffic consists of an increasing number of superimposed exponential on-off sources and compare them to related large buffer asymptotics for a single on-off source as cross traffic.

These results partially extend those of Ramanan and Dupuis [19] to more general rate functions. Also they complement our work [13] in which we investigated moderate deviations of this queueing network in critical loading.

Key words: many sources asymptotics, large buffer asymptotics, Markov-modulated fluid sources, on-off sources, queueing network, fluid model, first come first serve
Subject classification: primary 60F10; secondary 60K25, 90B15, 68M20

1. INTRODUCTION

Large deviations is a theory about the asymptotic decay of probabilities on a logarithmic scale. The decay rate is characterized as the minimum of a *rate function* on a related event [4]. The minimizer gives insight into the most likely behavior leading to the rare event [3].

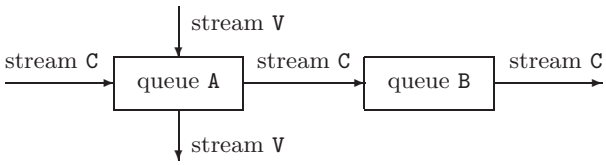

Figure 1: Queueing network with two nodes.

In this work we investigate large deviations of the queueing network with two buffers depicted in figure 1. A constant rate data stream C has to share the first buffer A with a random variable rate data stream V before being sent to the second buffer B. The steady-state queue length distribution of second buffer B is a measure for the cell delay variation induced on the constant rate traffic C by sharing the first buffer A with the random cross traffic V. We assume that both buffers provide non-idling service at constant rates and infinite waiting rooms. The first buffer A uses a first come first serve service discipline.

Large deviation properties of such a queueing network have already been investigated in the seminal work of Ramanan and Dupuis [19]. In contrast to our work they don't restrict the longitudinal traffic C to a constant rate stream. On the other hand they restrict attention to rate functions which are defined as the integral of a convex function of the derivative of a given sample path similar as in display (19) below. This form yields piecewise linear minimizers, a property which is heavily used in their proofs. But there are examples of rate functions which don't possess this property, for instance, the normalized superimposed cumulative traffic generated by an increasing number of Markov modulated fluid sources [16, 18].

In order to establish this partial generalization, we present a functional fluid model for the behavior of the queue lengths on the time interval $\mathbb{R}$. The use of the doubly infinite time interval $\mathbb{R}$ simplifies the direct examination of stationary cases. The functional relation between the cumulative cross traffic arrivals and induced queue lengths is continuous. Hence a large deviation principle for the sample paths of queue length processes can be derived from an appropriate large deviation principle for the cross traffic processes through an applica-

ValueTools 2008, October 21–23, 2008, Athens, GREECE

tion of the contraction principle. As main result of this work we show that in this situation the variational problem associated with the large deviation rate of the steady-state queue length distributions at the second buffer can be reduced to a two-dimensional minimization problem.

We then show that the prerequisites of these results are satisfied for *many sources* (see [23]) and *large buffer* (compare [22]) asymptotics with Markov modulated fluid cross traffic. In the special case of superimposed or scaled on-off sources we explicitly calculate and visualize large deviation rates and minimizing paths. Related *moderate* deviations in critical loading for the network have been obtained in [13, 15]. For many sources large deviation asymptotics of another queuing network with on-off traffic see [17].

An overview of this work is as follows: We recall necessary basic definitions and facts in Section 2. We present a functional fluid model for the network behavior in Section 3. This leads to a large deviation principle for queue lengths processes and induced stationary distributions in Section 4. Many sources and large buffer asymptotics of Markov modulated fluid cross traffic are shown to satisfy the required prerequisites in Section 5. Numerical examples for large deviation rates and minimizing paths with on-off cross traffic are presented in Section 6. Appendix A contains the proof of the reduction to a two dimensional minimization problem.

2. PRELIMINARIES

We recall the following elements of large deviations theory for which [4] is a standard reference. A $[0,\infty]$-valued lower semicontinuous function on a topological space is called *rate function*. A rate function is *good* if it has compact level sets. A sequence $(X_k)_{k\in\mathbb{N}}$ of random elements with values in a measurable space $(E,\mathcal{E})$ satisfies a *large deviation principle* with rate function I in the topology $\mathcal{T}$ on E if for every measurable set $A\in\mathcal{E}$

$$\limsup_{k\to\infty}\frac{1}{k}\log P(X_k\in A) \;\le\; -\inf_{x\in A^c} I(x), \quad (1)$$

$$\text{and}\quad \liminf_{k\to\infty}\frac{1}{k}\log P(X_k\in A) \;\ge\; -\inf_{x\in A^o} I(x), \quad (2)$$

where A^c (resp. A^o) is the closure (resp. interior) of A in the topology $\mathcal{T}$, and P is the underlying probability measure.

This work deals with large deviation principles for the sample paths of random processes. The underlying function spaces are introduced next. We let $\mathcal{C}$ be the set of functions $\mathbf{c}\colon\ \mathbb{R}\to\mathbb{R}$ which are continuous, and possess finite limits

$$\underline{\mathbf{c}} \;:=\; \lim_{t\to-\infty}\frac{\mathbf{c}(t)}{t}, \quad (3)$$

$$\overline{\mathbf{c}} \;:=\; \lim_{t\to\infty}\frac{\mathbf{c}(t)}{t}.$$

With $\mathcal{I}$ we denote the subset non-decreasing functions in $\mathcal{C}$. We provide these function spaces with the σ-algebras generated by the family of one-dimensional projections and topology induced by the norm $\|\cdot\|$ given by

$$\|\mathbf{c}\| \;:=\; \sup_{t\in\mathbb{R}}\frac{|\mathbf{c}(t)|}{1+|t|}. \quad (4)$$

Product function spaces are equipped with corresponding product σ-algebras and topologies.

With $\mathbf{id}\colon\ \mathbb{R}\to\mathbb{R}, t\mapsto t$ we denote the identity map of $\mathbb{R}$.

For $\tau\in\mathbb{R}$ we define the time-shift $\Theta_\tau\colon\mathcal{C}\to\mathcal{C}$ by

$$(\Theta_\tau\mathbf{d})(t) \;:=\; \mathbf{d}(t+\tau)$$

for every $\mathbf{d}\in\mathcal{C}$ and $t\in\mathbb{R}$. We say that a process $\mathbf{X}$ with sample paths in $\mathcal{C}$ is *stationary* (resp. has *stationary increments*), if the distribution of $\Theta_\tau\mathbf{x}$ (resp. $\Theta_\tau\mathbf{x}-\mathbf{x}(\tau)$) is the same for every $\tau\in\mathbb{R}$.

We recall that the composition $\circ\colon(\mathcal{C}\times\mathcal{I})\to\mathcal{C}, (\mathbf{c},\mathbf{d})\mapsto\mathbf{c}\circ\mathbf{d}$ is continuous and satisfies $\underline{\mathbf{c}\circ\mathbf{d}}=\underline{\mathbf{c}}\,\underline{\mathbf{d}}$ (see Lemma B.1 in [9]). Similarly, the inverse $^{-1}$ is a continuous map of the subset of strictly increasing functions in $\mathcal{I}$ onto itself [9]. On the subset $\mathcal{C}_{\mathbf{sup}}:=\{\mathbf{c}\in\mathcal{C}\colon\ \underline{\mathbf{c}}>0\}$ we consider the "running supremum" $\mathbf{sup}\colon\ \mathcal{C}_{\mathbf{sup}}\to\mathcal{I}$ defined for $\mathbf{d}\in\mathcal{C}_{\mathbf{sup}}$ and $t\in\mathbb{R}$ by

$$(\mathbf{sup}\,\mathbf{d})(t) \;:=\; \sup_{\tau\in\,]-\infty,t]}\mathbf{d}(\tau).$$

The $\mathbf{sup}$-mapping is continuous (see [5]), and satisfies $\underline{\mathbf{sup}\,\mathbf{d}}=\underline{\mathbf{d}}$ and $\overline{\mathbf{sup}\,\mathbf{d}}=\max\{0,\overline{\mathbf{d}}\}$. (For a finite subset $\mathcal{K}$ of $\mathbb{R}$ we let $\max\mathcal{K}$ denote the maximum of its elements.)

3. FUNCTIONAL NETWORK MODEL

We consider the queueing system depicted in figure 1. It possesses two queueing nodes `A` and `B` with infinite waiting room, constant service rates and fifo service discipline. It is populated with customers of two traffic classes `C` and `V`. Traffic of class `C` ("constant rate") must queue up for service at node `A` and subsequently at node `B`. Traffic of type `V` ("variable rate") must visit only node `A` before leaving the system. We are interested in the impact of sharing queue `A` with the random cross-traffic on the constant rate traffic. We assess this impact through large deviation asymptotics for queue lengths of queue `B`.

With a function $\mathbf{x}\in\mathcal{I}$ we can model the cumulative number of arrivals of class `V` customers at node `A` as a function of time. Concerning the constant rate traffic we fix an arrival rate $\alpha>0$ and use the function $\alpha\mathbf{id}$ to model the cumulative arrivals of the `C` traffic, thus implementing the "constant rate" property. In view of the assumed continuity of the cumulative input streams the following model falls into the category of fluid queueing models.

We let $\sigma_{\mathtt{A}}>0$, $\sigma_{\mathtt{B}}>0$ be fixed constant service rate at the first, resp. second queue. We must assume

$$\begin{aligned}\alpha &< \sigma_{\mathtt{B}},\\ \underline{\mathbf{x}} &< \mu:=\sigma_{\mathtt{A}}-\alpha,\end{aligned} \quad (5)$$

where μ is the reduced service rate of buffer `A` when traffic `C` would immediately be served. These conditions make sure that the queues are not overloaded during the negative time interval.

We set

$$\mathcal{I}_\mu \;:=\; \{\mathbf{c}\in\mathcal{I}\colon\underline{\mathbf{c}}<\mu\}.$$

The behavior of the total queue lengths at the queue `A` under the cross traffic $\mathbf{x}\in\mathcal{I}_\mu$ and the constant rate traffic $\alpha\mathbf{id}$ can be modelled through the mapping $\mathbf{Q}_{\mathtt{A}}\colon\ \mathcal{I}_\mu\to\mathcal{C}$ by setting

$$\begin{aligned}\mathbf{Q}_{\mathtt{A}}(\mathbf{x}) &:= \mathbf{x}+\alpha\mathbf{id}-\sigma_{\mathtt{A}}\mathbf{id}+\mathbf{sup}\,(\sigma_{\mathtt{A}}\mathbf{id}-\mathbf{x}-\alpha\mathbf{id})\\ &= \mathbf{x}-\mu\mathbf{id}+\mathbf{sup}(\mu\mathbf{id}-\mathbf{x}).\end{aligned} \quad (6)$$

This standard construction of the queue length behavior of a non-idling queueing node on the entire time interval $\mathbb{R}$

has its root in Loynes's lemma [1, 8] and is also used in equation (9).

Next we let $\mathbf{T}_{\mathtt{A}}(\mathbf{x})(t)$ be the time at which the customers which depart from the first node under the first in first out service discipline at time t, arrived in the system. This function satisfies

$$(\mathbf{x}+\alpha\mathbf{id})\circ\mathbf{T}_{\mathtt{A}}(\mathbf{x}) \quad = \quad \mathbf{x}+\alpha\mathbf{id}-\mathbf{Q}_{\mathtt{A}}(\mathbf{x}).$$

Since the function $\mathbf{x}+\alpha\mathbf{id}$ is strictly increasing, this specification of the mapping $\mathbf{T}_{\mathtt{A}}\colon \mathcal{I}_\mu \to \mathcal{I}$ is equivalent to the more explicit definition

$$\mathbf{T}_{\mathtt{A}}(\mathbf{x}) \quad = \quad (\mathbf{x}+\alpha\mathbf{id})^{-1}\circ(\mathbf{x}+\alpha\mathbf{id}-\mathbf{Q}_{\mathtt{A}}(\mathbf{x}))\,.$$

In particular, the cumulative departures of constant rate customers at queue one are given by

$$\mathbf{D}_{\mathtt{A},\mathtt{C}}(\mathbf{x}) \quad := \quad \alpha\mathbf{T}_{\mathtt{A}}(\mathbf{x}). \tag{7}$$

and the queue length induced by constant rate customers at queue $\mathtt{A}$ by

$$\mathbf{Q}_{\mathtt{A},\mathtt{C}}(\mathbf{x}) \quad := \quad \alpha\mathbf{id}-\mathbf{D}_{\mathtt{A},\mathtt{C}}(\mathbf{x}).$$

For later use we note the equation

$$\mathbf{Q}_{\mathtt{A}}(\mathbf{x})\circ\mathbf{T}_{\mathtt{A}}(\mathbf{x}) \quad = \quad \sigma_{\mathtt{A}}\mathbf{id}-\sigma_{\mathtt{A}}\mathbf{T}_{\mathtt{A}}(\mathbf{x}), \tag{8}$$

which is satisfied because for every $t\in\mathbb{R}$ queue $\mathtt{A}$ is fully loaded in the time interval $[\mathbf{T}_{\mathtt{A}}(\mathbf{x})(t),t]$ and thus produces $\sigma_{\mathtt{A}}t-\sigma_{\mathtt{A}}\mathbf{T}_{\mathtt{A}}(\mathbf{x})(t)$ departures which matches the number of customers which reside in the queue at time $\mathbf{T}_{\mathtt{A}}(\mathbf{x})(t)$.

The function $\mathbf{D}_{\mathtt{A},\mathtt{C}}(\mathbf{x})$ describes the cumulative arrivals of constant rate customers at the second queueing node. These arrivals induce the queue length behavior (compare equation (6))

$$\mathbf{Q}_{\mathtt{B}}(\mathbf{x}) \quad := \quad \mathbf{D}_{\mathtt{A}}(\mathbf{x})-\sigma_{\mathtt{B}}\mathbf{id}+\sup\left(\sigma_{\mathtt{B}}\mathbf{id}-\mathbf{D}_{\mathtt{A},\mathtt{C}}(\mathbf{x})\right). \tag{9}$$

See [11, 12] for a more general version of this construction.

4. LARGE DEVIATIONS OF THE QUEUES

Throughout this section we let $(\mathbf{X}_k)$ be a sequence of random cross traffic processes with sample paths in $\mathcal{I}$. We assume

$$\underline{\mathbf{X}_k} \quad < \quad \mu$$

for every $k\in\mathbb{N}$ with probability 1, in order to make sure that the queueing model of Section 3 can be applied (compare condition (5)).

Furthermore we assume that the sequence $(\mathbf{X}_k)_{k\in\mathbb{N}}$ satisfies a large deviation principle with a good rate function I satisfying

$$\mathcal{I}_I \quad := \quad \{\mathbf{x}\in\mathcal{I}\colon\ I(\mathbf{x})<\infty\}\subset\mathcal{I}_\mu.$$

Hence large deviation principles for the sequence $(\mathbf{X}_k)_{k\in\mathbb{N}}$ established on the space $\mathcal{I}$ are also valid on $\mathcal{I}_\mu$; see Lemma 4.1.5(b) in [4]. In Section 5 we present examples of cross traffic processes which satisfy these assumptions.

4.1 Sample path large deviation principle

In this section we state a large deviation principle for the queue lengths processes and motivate the importance of minimizers of the rate function.

THEOREM 4.1. *The sequence*

$$\left(\mathbf{X}_k,\mathbf{Q}_{\mathtt{A},\mathtt{V}}(\mathbf{X}_k),\mathbf{Q}_{\mathtt{A},\mathtt{C}}(\mathbf{X}_k),\mathbf{Q}_{\mathtt{B}}(\mathbf{X}_k)\right)_{k\in\mathbb{N}} \tag{10}$$

satisfies a sample path large deviation principle with good rate function J on $\mathcal{I}\times\mathcal{C}\times\mathcal{C}\times\mathcal{C}$ given by

$$J\left(\mathbf{x},\mathbf{q}_{\mathtt{A},\mathtt{V}},\mathbf{q}_{\mathtt{A},\mathtt{C}},\mathbf{q}_{\mathtt{B}}\right) \quad := \quad \inf_{\substack{\mathbf{x}\in\mathcal{I}_I,\\ \mathbf{q}_{\mathtt{A},\mathtt{V}}=\mathbf{Q}_{\mathtt{A},\mathtt{V}}(\mathbf{x}),\\ \mathbf{q}_{\mathtt{A},\mathtt{C}}=\mathbf{Q}_{\mathtt{A},\mathtt{C}}(\mathbf{x}),\\ \mathbf{q}_{\mathtt{B}}=\mathbf{Q}_{\mathtt{B}}(\mathbf{x})}} I(\mathbf{x}).$$

PROOF. The statement of the theorem is a direct consequence of the continuity of the mappings which appear in the definition of $\mathbf{Q}_{\mathtt{A},\mathtt{V}}$, $\mathbf{Q}_{\mathtt{A},\mathtt{C}}$, and $\mathbf{Q}_{\mathtt{B}}$ and the contraction principle [4]. □

This sample path large deviation principle implies the following convergence of conditional distributions to minimizing paths of the rate function; compare corollary 1 in [14].

COROLLARY 4.2. *If the measurable set $\mathcal{Q}\in\mathcal{I}\times\mathcal{C}\times\mathcal{C}\times\mathcal{C}$ satisfies*

$$\begin{aligned}&\inf_{(\mathbf{x},\mathbf{q}_{\mathtt{A},\mathtt{V}},\mathbf{q}_{\mathtt{A},\mathtt{C}},\mathbf{q}_{\mathtt{B}})\in\mathcal{Q}^c} J\left(\mathbf{x},\mathbf{q}_{\mathtt{A},\mathtt{V}},\mathbf{q}_{\mathtt{A},\mathtt{C}},\mathbf{q}_{\mathtt{B}}\right)\\ =\ &\inf_{(\mathbf{x},\mathbf{q}_{\mathtt{A},\mathtt{V}},\mathbf{q}_{\mathtt{A},\mathtt{C}},\mathbf{q}_{\mathtt{B}})\in\mathcal{Q}^o} J\left(\mathbf{x},\mathbf{q}_{\mathtt{A},\mathtt{V}},\mathbf{q}_{\mathtt{A},\mathtt{C}},\mathbf{q}_{\mathtt{B}}\right)<\infty,\end{aligned}$$

and there is a unique element $\left(\mathbf{x}^*,\mathbf{q}^*_{\mathtt{A},\mathtt{V}},\mathbf{q}^*_{\mathtt{A},\mathtt{C}},\mathbf{q}^*_{\mathtt{B}}\right)\in\mathcal{Q}^c$ which attains the infimum in the last display, the distribution of the process $\left(\mathbf{X}_k,\mathbf{Q}_{\mathtt{A},\mathtt{V}}(\mathbf{X}_k),\mathbf{Q}_{\mathtt{A},\mathtt{C}}(\mathbf{X}_k),\mathbf{Q}_{\mathtt{B}}(\mathbf{X}_k)\right)$ conditioned to the event $\mathcal{Q}$ converges to the Dirac measure of the path $\left(\mathbf{x}^*,\mathbf{q}^*_{\mathtt{A},\mathtt{V}},\mathbf{q}^*_{\mathtt{A},\mathtt{C}},\mathbf{q}^*_{\mathtt{B}}\right)$ in distribution as $k\to\infty$. Furthermore, $\mathbf{q}^*_{\mathtt{A},\mathtt{V}}=\mathbf{Q}_{\mathtt{A},\mathtt{V}}(\mathbf{x}^*)$, $\mathbf{q}^*_{\mathtt{A},\mathtt{C}}=\mathbf{Q}_{\mathtt{A},\mathtt{C}}(\mathbf{x}^*)$, and $\mathbf{q}^*_{\mathtt{B}}=\mathbf{Q}_{\mathtt{B}}(\mathbf{x}^*)$.

Hence minimizing paths of the rate function J can characterize the asymptotically most likely conditional behavior which leads to an untypical queue lengths behavior at buffer $\mathtt{A}$ under the sequence of cross traffic arrival processes $(\mathbf{X}_k)_{k\in\mathbb{N}}$.

4.2 Large deviations for the second buffer

In this section we derive large deviation asymptotics for the sequence $(B_k)_{k\in\mathbb{N}}$ of queue length distributions of buffer $\mathtt{B}$ at time 0. These random variables are defined by

$$B_k \quad := \quad \mathbf{Q}_{\mathtt{B}}(\mathbf{X}_k)(0).$$

We will need the following two assumptions on the rate function I of the cross traffic processes: firstly, $I(\mathbf{x})=\infty$ whenever $\mathbf{x}(0)\neq 0$; secondly, the rate function I is invariant under time shifts, i.e. for every $\tau\in\mathbb{R}$ and $\mathbf{x}\in\mathcal{I}$

$$I(\Theta_\tau\mathbf{x}-\mathbf{x}(\tau)+\mathbf{x}(0)) \quad = \quad I(\mathbf{x}). \tag{11}$$

Clearly, this latter assumption is satisfied whenever $\mathbf{X}_k$ has stationary increments for every $k\in\mathbb{N}$. We note that if $\mathbf{X}_k$ is a process with stationary increments, the associated queue length processes are stationary, which means that the distribution of the triple $(\Theta_\tau\mathbf{Q}_{\mathtt{A},\mathtt{C}}(\mathbf{X}_k),\Theta_\tau\mathbf{Q}_{\mathtt{A},\mathtt{V}}(\mathbf{X}_k),\Theta_\tau\mathbf{Q}_{\mathtt{B}}(\mathbf{X}_k))$ does not-depend on $\tau\in\mathbb{R}$.

For $t_1<t_2<0$ we let $I_{t_1,t_2}\colon\ \mathbb{R}^2\to\mathbb{R}_+\cup\{\infty\}$ be the good rate function defined by

$$I_{t_1,t_2}(x_1,x_2) \quad := \quad \inf_{\substack{\mathbf{x}\in\mathcal{I}_I,\\ \mathbf{x}(t_1)=x_1,\\ \mathbf{x}(t_2)=x_2}} I(\mathbf{x}). \tag{12}$$

The following theorem shows that large deviation asymptotics for the (steady-state) queue length distribution at

node B can be obtained by solving two-dimensional minimization problems defined in terms of this rate function I_{t_1,t_2}. Its proof is relegated to Appendix A. For its formulation we introduce the parameters

$$\begin{aligned}\phi &:= \frac{\sigma_A}{\sigma_B},\\ \psi &:= \frac{(\sigma_A-\sigma_B)\alpha}{\sigma_B}.\end{aligned}$$

THEOREM 4.3. *For every $\delta > 0$*

$$\begin{aligned}&\limsup_{k\to\infty}\frac{1}{k}\log P\left(B_k \geq \delta\right)\\ &\quad\leq -\inf_{\substack{t_1<t_2<0,\\ x_1\leq \mu t_1,\\ x_2\geq \phi\delta+\psi t_2}} I_{t_1,t_2}(x_1,x_2)\\ &\quad= -\inf_{t_1<t_2<0} I_{t_1,t_2}(\mu t_1,\phi\delta+\psi t_2), \qquad (13)\\ &\liminf_{k\to\infty}\frac{1}{k}\log P\left(B_k \geq \delta\right)\\ &\quad\geq -\inf_{\substack{t_1<t_2<0,\\ x_1=\mu t_1,\\ x_2>\phi\delta+\psi t_2}} I_{t_1,t_2}(x_1,x_2). \qquad (14)\end{aligned}$$

If there are unique values $t_1^* < t_2^* < 0$ satisfying

$$\begin{aligned}&I_{t_1^*,t_2^*}(\mu t_1^*,\phi\delta+\psi t_2^*)\\ &\quad= \inf_{\substack{t_1<t_2<0,\\ x_1<\mu t_1,\\ x_2>\phi\delta+\psi t_2}} I_{t_1,t_2}(x_1,x_2) < \infty, \qquad (15)\end{aligned}$$

and there is a unique path $\mathbf{x}^* \in \mathcal{I}_I$ satisfying $\mathbf{x}^*(t_1^*) = \mu t_1^*$, $\mathbf{x}^*(t_2^*) = \phi\delta + \psi t_2^*$, and

$$I(\mathbf{x}^*) = \inf_{\substack{\mathbf{x}\in\mathcal{I}_I,\\ \mathbf{x}(t_1^*)=\mu t_1^*,\\ \mathbf{x}(t_2^*)=\phi\delta+\psi t_2^*}} I(\mathbf{x}),$$

the distribution of the process $\left(\mathbf{X}_k,\ \mathbf{Q}_{A,V}(\mathbf{X}_k),\ \mathbf{Q}_{A,C}(\mathbf{X}_k),\ \mathbf{Q}_B(\mathbf{X}_k)\right)$ conditioned to the event

$$\left\{(\mathbf{x},\mathbf{q}_{A,V},\mathbf{q}_{A,C},\mathbf{q}_B)\in\mathcal{I}\times\mathcal{C}\times\mathcal{C}\times\mathcal{C}\colon\ \mathbf{q}^B(0)\geq\delta\right\},$$

converges to the Dirac measure of the path $\left(\mathbf{x}^*,\ \mathbf{Q}_{A,V}(\mathbf{x}^*),\ \mathbf{Q}_{A,C}(\mathbf{x}^*),\ \mathbf{Q}_B(\mathbf{x}^*)\right)$ in distribution as $k\to\infty$.

This theorem is very similar to Theorem 1 in [13] obtained for moderate deviation asymptotics of the same queueing system in critical loading. It partially extends Theorem 8.1 of [19]: It uses less assumptions concerning the rate function, but it is restricted to non-random longitudinal traffic. Only Problem 3 "buffer B exploits buffer A" of [19] is relevant under this restriction.

5. MARKOV MODULATED TRAFFIC

The results of Section 4 call for a sequence of cross traffic arrival processes satisfying a sample path large deviation principle. In this section we present two well known examples of such sequences. In example M ("many sources asymptotics" [23]) we consider the averaged superposition of the fluid produced by an increasing number of sources with Markov modulated fluid production rate. We contrast this example with example L ("large buffer asymptotics"), obtained from a single Markov modulated fluid source by scaling time and value.

We let $G \in \mathbb{R}^{n\times n}$ be the transition matrix belonging to a positive recurrent continuous-time Markov chain with states $\{1,\dots,n\}$. As usual we assume $Ge = 0$ where $e \in \mathbb{R}_+^n$ is the vector with ones in every component. We let $\mathbf{M} = (\mathbf{M}(t))_{t\in\mathbb{R}}$ be a finite-state stationary continuous-time Markov chain associated with G. We denote its one-dimensional marginal distribution with $\pi \in \mathbb{R}_+^n$, that is, π is the non-negative vector satisfying $e^T\pi = 1$ and $\pi^T G = 0$. Here v^T denotes the transpose of a vector v and the standard vector-vector, vector-matrix and matrix-matrix multiplications are used.

We let $\xi \in \mathbb{R}_+^n$ be a vector of fluid production rates. When the Markov chain $\mathbf{M}$ is in state $i \in \{1,\dots,n\}$ the source produces fluid with rate ξ_i. Hence the definition

$$\mathbf{Z}(t) := \int_0^t \xi_{\mathbf{M}(t)}\ dt$$

defines the cumulative fluid produced up to time t. This amount is normalized to be 0 at time 0. We note that $\mathbf{Z}$ is non-decreasing, has stationary increments, and satisfies

$$\underline{\mathbf{Z}} = \overline{\mathbf{Z}} = \pi^T\xi \quad \text{a.s.} \qquad (16)$$

In the following we will assume (compare (5))

$$\pi^T\xi < \mu. \qquad (17)$$

5.1 Many sources asymptotics

We let $(\mathbf{Z}_k)_{k\in\mathbb{N}}$ be sequence of independent copies of the process $\mathbf{Z}$ and define

$$\left(\mathbf{X}_k^M\right)_{k\in\mathbb{N}} := \left(\frac{1}{k}\sum_{i=1}^k \mathbf{Z}_i\right)_{k\in\mathbb{N}}.$$

Because of equation (16) and condition (17), for every $k \in \mathbb{N}$

$$\underline{\mathbf{X}_k^M} = \overline{\mathbf{X}_k^M} = \pi^T\xi < \mu$$

with probability 1. Hence the mappings $\mathbf{Q}_{A,V}$, $\mathbf{Q}_{A,C}$ and $\mathbf{Q}_B$ can be applied to almost all sample paths of the processes $\mathbf{X}_k^M$.

We let $\mathcal{C}^*$ be the topological dual of $\mathcal{C}$. The logarithmic moment generating function $\Lambda\colon \mathcal{C}^* \to \mathbb{R}_+$ is given by

$$\Lambda(\mathbf{y}) := \log E\left(\exp\mathbf{y}(\mathbf{Z})\right).$$

The rate function $I^M\colon\ \mathcal{I} \to \mathbb{R}_+\cup\{\infty\}$ is defined as the Fenchel-Legendre-transform of Λ^M, that is,

$$I^M(\mathbf{x}) := \sup_{\mathbf{y}\in\mathcal{C}^*}\left(\mathbf{y}(\mathbf{x}) - \Lambda(\mathbf{y})\right).$$

THEOREM 5.1. *The sequence $(\mathbf{X}_k^M)_{k\in\mathbb{N}}$ satisfies a large deviation principle with good rate function I^M.*

PROOF. See [2] and compare [17]. □

For $t_1 < t_2 < 0$ and $y_1, y_2 \in \mathbb{R}$ we define

$$\Lambda_{t_1,t_2}(y_1,y_2) := \log E\left(\exp\left(y_1\mathbf{Z}(t_1) + y_2\mathbf{Z}(t_2)\right)\right).$$

For $t, y \in \mathbb{R}$ we let $H(t,y) \in \mathbb{R}^{n\times n}$ be the matrix

$$H(t,y) := \exp(t(G + y\ \mathrm{diag}(\xi)),$$

where $\mathrm{diag}(\xi)$ is the diagonal matrix having the components of the vector ξ in its diagonal and the exponential of a matrix is defined as usual; see Section 6.3 in [7].

LEMMA 5.2. *For every $t_1 < t_2 < 0$ and $y_1, y_2 \in \mathbb{R}$*

$$\begin{aligned}&\Lambda_{t_1,t_2}(y_1,y_2)\\ &\quad= \log\left(\pi^T H(t_2-t_1,-y_1)H(-t_2,-y_1-y_2)e\right). \qquad (18)\end{aligned}$$

PROOF. One can first use the stationary increments property of the process $\mathbf{Z}$ to obtain

$$\begin{aligned}\exp\Lambda_{t_1,t_2}(y_1,y_2) &= E\exp\left(-y_1(\mathbf{Z}(t_2)-\mathbf{Z}(t_1))-(y_1+y_2)(\mathbf{Z}(0)-\mathbf{Z}(t_2))\right)\\ &= E\exp\left(-y_1(\mathbf{Z}(t_2-t_1)-\mathbf{Z}(0))\right.\\ &\qquad\left.-(y_1+y_2)(\mathbf{Z}(-t_1)-\mathbf{Z}(t_2-t_1))\right).\end{aligned}$$

Since for every $i,j\le n$, $y\in\mathbb{R}$ and $t>0$ (see Appendix in [6])

$$E(1_{\{j\}}(\mathbf{M}(t))\exp(y\mathbf{Z}(t))\mid\mathbf{M}(t)=i) = H(t,y)_{i,j},$$

this implies formula (18) via the Markov property of the process $\mathbf{M}$ with standard conditioning arguments. Here 1_S is the indicator function of the set S. □

With the help of this two-dimensional logarithmic moment generating function, the rate function $I^{\mathtt{M}}_{t_1,t_2}$, defined by changing I to $I^{\mathtt{M}}$ in equation (12), admits a more explicit representation.

LEMMA 5.3. *For every $t_1<t_2<0$ and $x_1,x_2\in\mathbb{R}$*

$$I^{\mathtt{M}}_{t_1,t_2}(x_1,x_2) = \sup_{y_1,y_2\in\mathbb{R}}\left(x_1y_1+x_2y_2-\Lambda_{t_1,t_2}(y_1,y_2)\right).$$

This supremum is infinite if $x_1>x_2$ or $x_2>0$ or $(x_2-x_1)/(t_2-t_1)>\max_{i=1}^n\xi_i$ or $x_2/t_2>\max_{i=1}^n\xi_i$.

PROOF. The contraction principle implies that $I^{\mathtt{M}}_{t_1,t_2}$ is the rate function in a large deviation principle for the sequence $(\mathbf{X}^{\mathtt{M}}_k(t_1),\mathbf{X}^{\mathtt{M}}_k(t_2))_{k\in\mathbb{N}}$. By Cramér's Theorem, the Fenchel-Legendre transform of Λ_{t_1,t_2} is also a rate function for this sequence. Hence the identity of the lemma is a consequence of the uniqueness of rate functions in large deviation principles; compare Lemma 4.1.1 in [4]. The last statement of the lemma is clear from the fact, that the events $\{\mathbf{X}^{\mathtt{M}}_k(t_1)>\mathbf{X}^{\mathtt{M}}_k(t_2)\}$, $\{\mathbf{X}^{\mathtt{M}}_k(t_2)>0\}$, $\{\mathbf{X}^{\mathtt{M}}_k(t_2)-\mathbf{X}^{\mathtt{M}}_k(t_1)>(t_2-t_1)\max_{i=1}^n\xi_i\}$, and $\{\mathbf{X}^{\mathtt{M}}_k(t_2)>t_2\max_{i=1}^n\xi_i\}$ have probability 0 for every $k\in\mathbb{N}$. □

Certain minimizing paths of the rate function $I^{\mathtt{M}}$ can be expressed via derivatives.

LEMMA 5.4. *If $t_1<t_2<0$ and $x_1,x_2,y_1,y_2\in\mathbb{R}$ satisfy*

$$I^{\mathtt{M}}_{t_1,t_2}(x_1,x_2)+\Lambda_{t_1,t_2}(y_1,y_2) = x_1y_1+x_2y_2,$$

then the path $\mathbf{x}^*$ defined for $t=t_1$ by $\mathbf{x}^*(t):=x_1$, for $t=t_2$ by $\mathbf{x}^*:=x_2$, for $t<t_2$ by

$$\mathbf{x}^*(t) := x_1-\frac{d}{dy}\log\left(\pi^TH(t-t_1,y)\right.\\ \left.H(t_2-t_1,-y_1)H(-t_2,-y_1-y_2)e\right)\big|_{y=0},$$

for $t_1<t<t_2$ by

$$\mathbf{x}^*(t) := x_1+\frac{d}{dy}\log\left(\pi^TH(t-t_1,y-y_1)\right.\\ \left.H(t_2-t,-y_1)H(-t_2,-y_1-y_2))e\right)\big|_{y=0},$$

for $t_2<t<0$ by

$$\mathbf{x}^*(t) := x_2+\frac{d}{dy}\log\left(\pi^TH(t_2-t_1,-y_1)\right.\\ \left.H(t-t_2,y-y_1-y_2)H(-t,-y_1-y_2)\gamma_2\right)\big|_{y=0},$$

and for $t>0$ by

$$\mathbf{x}^*(t) := \frac{d}{dy}\log\left(\pi^TH(t_2-t_1,-y_1)\right.\\ \left.H(-t_2,-y_1-y_2)H(t,y)e\right)\big|_{y=0},$$

is the unique path satisfying $\mathbf{x}^*(t_1)=x_1$, $\mathbf{x}^*(t_2)=x_2$ and $I^{\mathtt{M}}(\mathbf{x}^*)=I^{\mathtt{M}}_{t_1,t_2}(x_1,x_2)$.

PROOF. Since $I^{\mathtt{M}}_{t_1,t_2}(x_1,x_2)<\infty$ and $I^{\mathtt{M}}$ has compact level sets, the definition of $I^{\mathtt{M}}_{t_1,t_2}$ implies the existence of a path $\mathbf{x}^*$ satisfying the last conditions of the last line of the statement. In order to complete the proof of the lemma it remains to verify that $\mathbf{x}^*(t)$ satisfies the equations given in the lemma for an arbitrary $t\in\mathbb{R}$.

Here we just consider the case $t>0$. We define $x_3:=\mathbf{x}^*(t)$ and

$$\Lambda_{t_1,t_2,t}(y_1,y_2,y) := \log E\left(\exp(y_1\mathbf{Z}(t_1)+y_2\mathbf{Z}(t_2)+y\mathbf{Z}(t))\right).$$

The properties of $\mathbf{x}^*$ imply (compare Lemmas 5.2 and 5.3)

$$\begin{aligned}I^{\mathtt{M}}(\mathbf{x}^*) &= \sup_{y_1,y_2,y_3\in\mathbb{R}}\left(\sum_{i=1}^3x_iy_i-\Lambda_{t_1,t_2,t}(y_1,y_2,y_3)\right)\\ &\le \sup_{y_1,y_2\in\mathbb{R}}\left(\sum_{i=1}^2x_iy_i-\Lambda_{t_1,t_2,t}(y_1,y_2,0)\right)\\ &= \sup_{y_1,y_2\in\mathbb{R}}\left(\sum_{i=1}^2x_iy_i-\Lambda_{t_1,t_2}(y_1,y_2)\right)\\ &= I^{\mathtt{M}}(x_1,x_2)=I^{\mathtt{M}}(\mathbf{x}^*).\end{aligned}$$

Since $\Lambda_{t_1,t_2,t}$ is convex (compare Example 2.16 in [20]) and differentiable, this implies the equation in the last display of the lemma. □

In particular, prerequisites of Theorem 4.1, Corollary 4.2 and Theorem 4.3 are satisfied when $I=I^{\mathtt{M}}$ and $\mathbf{X}_k=\mathbf{X}^{\mathtt{M}}_k$ for every $k\in\mathbb{N}$. This setup is called example $e=\mathtt{M}$.

5.2 Large buffer asymptotics

In order to compare the numerical findings below with the results of [19], we also consider the following sequence $(\mathbf{X}^{\mathtt{L}}_k)_{k\in\mathbb{N}}$ of *scaled* Markov modulated fluid processes defined for $k\in\mathbb{N}$ and $t\in\mathbb{R}$ by

$$\mathbf{X}^{\mathtt{L}}_k(t) := \frac{1}{k}\mathbf{Z}(kt).$$

Because of equation (16) and condition (17) for every $k\in\mathbb{N}$

$$\underline{\mathbf{X}^{\mathtt{L}}_k}=\overline{\mathbf{X}^{\mathtt{L}}_k}=\pi^T\xi<\mu \quad\text{a.s.}$$

Hence the mappings $\mathbf{Q}_{\mathtt{A},\mathtt{V}}$, $\mathbf{Q}_{\mathtt{A},\mathtt{C}}$ and $\mathbf{Q}_{\mathtt{B}}$ can be applied to almost all sample paths of the processes $\mathbf{X}^{\mathtt{L}}_k$.

The appropriate rate function $I^{\mathtt{L}}\colon\ \mathcal{I}\to\mathbb{R}_+\cup\{\infty\}$ for this sequence is given by

$$I^{\mathtt{L}}(\mathbf{x}) := \int_{\mathbb{R}}f(\dot{\mathbf{x}}(t))\,dt, \tag{19}$$

if $\mathbf{x}$ is absolutely continuous and $\mathbf{x}(0)=0$, and $I^{\mathtt{L}}(\mathbf{x})=\infty$, otherwise. The function $f\colon\ \mathbb{R}_+\to\mathbb{R}_+\cup\{\infty\}$ is given as

$$f(r) := \inf_{\substack{x\in\mathbb{R}^n_+,\\ e^Tx=1,\\ \xi^Tx=r}}\sup_{u>0}\left(-\sum_{i=1}^n\sum_{j=1}^n\frac{x_iG_{i,j}u_j}{u_i}\right). \tag{20}$$

THEOREM 5.5. *The sequence* $(\mathbf{X}_k^{\mathtt{L}})_{k\in\mathbb{N}}$ *satisfies a large deviation principle with good rate function* $I^{\mathtt{L}}$.

PROOF. See references in [19]. The "local" rate function f is calculated in Exercise 4.2.28 of [4]. The strengthening to the topology considered here can be done with [10]. □

The rate function $I^{\mathtt{L}}_{t_1,t_2}$, defined by replacing I with $I^{\mathtt{L}}$ in equation (12), admits a more explicit representation:

LEMMA 5.6. *For every* $t_1 < t_2 < 0$ *and* $x_1, x_2 \in \mathbb{R}$

$$I^{\mathtt{L}}_{t_1,t_2}(x_1,x_2) \quad = \quad (t_2-t_1) f\left(\frac{x_2-x_1}{t_2-t_1}\right) - t_2 f\left(\frac{x_2}{t_2}\right).$$

In particular, $I^{\mathtt{L}}_{t_1,t_2}(x_1,x_2) < \infty$, *if and only if* $x_1 \le x_2 \le 0$, $(x_2-x_1)(t_2-t_1) \le \max_{i=1}^n \xi_i$ *and* $x_2/t_2 \le \max_{i=1}^n \xi_i$.

Furthermore, we can specify minimizing paths.

LEMMA 5.7. *If* $t_1 < t_2 < 0$, $x_1 \le x_2 \le 0$ *and* $I^{\mathtt{L}}_{t_1,t_2}(x_1,x_2) < \infty$, *the function* $\mathbf{x}^* \in \mathcal{I}$ *given by*

$$\mathbf{x}^*(t) \quad := \quad \begin{cases} x_1 - (t_1 - t)\pi^T\xi & \text{if } t < t_1, \\ \frac{t-t_1}{t_2-t_1}x_2 + \frac{t_2-t}{t_2-t_1}x_1 & \text{if } t_1 \le t < t_2, \\ \frac{t}{-t_2}x_2 & \text{if } t_2 \le t < 0, \\ t\pi^T\xi & \text{if } t \ge 0, \end{cases}$$

is the unique function satisfying $\mathbf{x}^*(t_1) = x_1$, $\mathbf{x}^*(t_2) = x_2$ *and* $I^{\mathtt{L}}(\mathbf{x}^*) = I^{\mathtt{L}}_{t_1,t_2}(x_1,x_2)$.

We omit the detailed proofs of these two lemmas which could be based on Jensen's inequality.

The prerequisites of Theorem 4.1, Corollary 4.2 and Theorem 4.3 are satisfied when $I = I^{\mathtt{L}}$, $I_{t_1,t_2} = I^{\mathtt{L}}_{t_1,t_2}$ and $\mathbf{X}_k = \mathbf{X}_k^{\mathtt{L}}$ for every $k \in \mathbb{N}$. This setup is called example $e = \mathtt{L}$.

6. ON-OFF CROSS TRAFFIC

In order to derive explicit numerical results we further specializing the cross traffic to exponential on-off traffic. This can be achieved by choosing $\mathbf{M}$ as Markov chain with two states $n = 2$ and by assuming that the fluid arrival rate in state 1 ("off") is $\xi_1 = 0$ and in state 2 ("on") is $\xi_2 > 0$. Hence the transition matrix G can be written as

$$G \quad = \quad \begin{pmatrix} -g_2 & g_2 \\ g_1 & -g_1 \end{pmatrix}.$$

with $g_1 > 0$ being the transition rate from "on" to "off", $g_2 > 0$ being the transition rate from "off" to "on". The steady-state distribution of $\mathbf{Z}$ is given by

$$\pi \quad = \quad \frac{1}{g_2+g_1}\begin{pmatrix} g_1 \\ g_2 \end{pmatrix}.$$

In this setting the process $\mathbf{Z}$ behaves like an on-off fluid source with fluid production rate ξ_2 during on phases. The on phase durations (= lengths of maximal time intervals during which $\mathbf{M} = 2$) form a sequence of independent and exponentially distributed random variables with mean g_2. Similarly, the off phase durations (= lengths of maximal time intervals during which $\mathbf{M} = 1$) form a sequence of independent and exponentially distributed random variables with mean g_1. These two sequences are independent.

In order for our network model to be well defined we have to assume condition (17), which bounds the mean fluid production rate of this source through

$$\xi_2 g_2/(g_1+g_2) \quad < \quad \mu.$$

In the remainder of this section we explain how to calculate the two-dimensional logarithmic moment generating function Λ_{t_1,t_2} and the local rate function f in this on-off source case. These expressions simplify the numerical calculations of the two-dimensional rate functions $I^{\mathtt{M}}_{t_1,t_2}$ and $I^{\mathtt{L}}_{t_1,t_2}$.

For $y \in \mathbb{R}$ we define the values $\eta_+(y), \eta_-(y) \in \mathbb{R}$ by

$$\eta^{\pm}(y) \quad := \quad \frac{y\xi_2}{2} - \frac{g_1+g_2}{2} \pm \frac{1}{2}\sqrt{(y\xi_2+g_2-g_1)^2 + 4g_1g_2}.$$

These values are the eigenvalues of the matrix $G + y\mathrm{diag}(\xi)$. Therefore (compare Exercise 2 in Section 9.3 of [7])

$$\begin{aligned} H(t,y) \quad = \quad & \frac{\exp(t\eta_+(y))}{\eta_+(y)-\eta_-(y)}\left(G + \mathrm{diag}(y\xi - \eta_-(y)e)\right) \\ & - \frac{\exp(t\eta_-(y))}{\eta_+(y)-\eta_-(y)}\left(G + \mathrm{diag}(y\xi - \eta_+(y)e)\right). \end{aligned}$$

This representation is also the subject of Lemma 2.1 in [17].

In the on-off case one can also derive an explicit representation of the local rate function f defined in equation (20).

LEMMA 6.1. *We have for* $r \in [0, \xi_2]$

$$f(r) \quad = \quad \frac{\xi_2 - r}{\xi_2}g_2 + \frac{r}{\xi_2}g_1 - 2\sqrt{\frac{r(\xi_2-r)g_1g_2}{\xi_2^2}},$$

and $f(r) = \infty$, *otherwise.*

PROOF. We calculate for $x_1, x_2 \in \mathbb{R}_+$ with $x_1 + x_2 = 1$

$$\begin{aligned} & \sup_{u_1,u_2>0}\left(-\sum_{i=1}^n\sum_{j=1}^n \frac{x_iG_{i,j}u_j}{u_i}\right) \\ & \qquad = \quad x_1g_2 + x_2g_1 - 2\sqrt{x_1x_2g_1g_2} \end{aligned}$$

This implies the statement of the lemma. □

We note that the non-linear solver Ipopt [21] has been used to solve the optimization problems which arose in the numerical examples of the following two sections.

6.1 Overflow rates

In this section we display numerical examples for the values

$$I^* \quad := \quad \inf_{t_1<t_2<0} I^e_{t_1,t_2}(\mu t_1, \phi\delta + \psi t_2), \tag{21}$$

which, appear on the right hand side of display (13) in theorem 4.3. We will consider both examples $e \in \{\mathtt{M}, \mathtt{L}\}$ in the special case of on-off sources.

We introduce the parameters $\xi_{\mathrm{mean}} := \xi^T\pi$ and $\beta := g_1 + g_2$. The value ξ_{mean} is the mean traffic intensity of the on-off stream and β the sum of the transition rates. Given the maximum cross traffic rate $\xi_2 > \xi_{\mathrm{mean}}$ and $\beta > 0$ we can express the individual transition rates g_1 and g_2 as

$$\begin{aligned} g_2 \quad &= \quad \frac{\xi_{\mathrm{mean}}\beta}{\xi_2}, \\ g_1 \quad &= \quad \beta - g_2. \end{aligned}$$

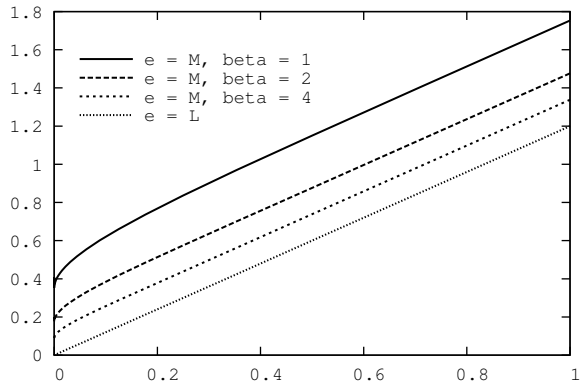

Figure 2: Large deviation rates I^*/β as functions of δ for $\beta \in \{1, 2, 4\}$ and $e = \{\mathtt{M}, \mathtt{L}\}$.

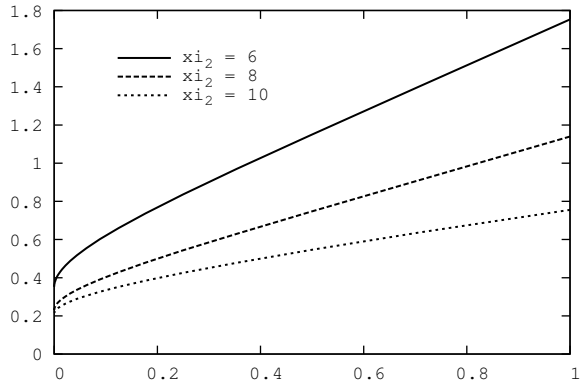

Figure 3: Large deviation rates I^* as functions of δ for $e = \mathtt{M}$ and $\xi_2 \in \{6, 8, 10\}$.

In all examples we consider the set-up $\xi_{\text{mean}} = 2$, $\sigma_{\mathtt{A}} = 4$, $\sigma_{\mathtt{B}} = 2$ and $\alpha = 1$. This implies $\mu = 3$, $\phi = 2$, and $\psi = 1$. In particular, we cannot observe strictly positive queue lengths if $\xi_2 \leq 3$. The remaining parameters are $\xi_2 > 3$, $\beta > 0$, $\delta > 0$, and $e \in \{\mathtt{M}, \mathtt{L}\}$.

In figure 2 we keep $\xi_2 = 6$ constant and display I^*/β as a function of δ for different values of β. We note that for $e = \mathtt{L}$ the value I^*/β does not depend on β such that there is only one function for this case.

The jump which appears for $e = \mathtt{M}$ at $\delta = 0$ is in sharp contrast to the linearity of the functions for $e = \mathtt{L}$. This jump corresponds to the asymptotic effort to increase the traffic intensity at node B to 1. The results of [19] are restricted to the linear case. Figure 2 suggests that the values of I^*/β for $e = \mathtt{M}$ converge to those of $e = \mathtt{L}$ as β increases to infinity and all other parameters remain fixed.

In figure 3 we show the values I^* as a function of δ for fixed parameters $e = \mathtt{M}$ and $\beta = 1$ and varying parameters $\xi_2 \in \{6, 8, 10\}$. The values seem to decrease when maximum rate ξ_2 increases.

6.2 Minimizing paths

In this section we consider the fixed parameters $\delta = 1$, $\alpha = 1$, $\sigma_{\mathtt{B}} = 2$, $\sigma_{\mathtt{A}} = 4$, $\xi_2 = 8$, $g_1 = 3\beta/4$, $g_2 = \beta/4$ which implies $\mu = 3$, and $\xi_{\text{mean}} = 2$.

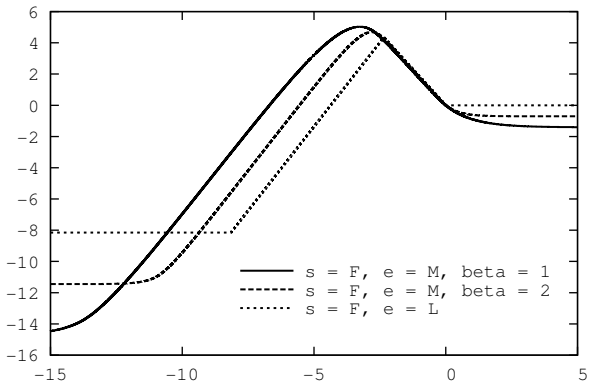

Figure 4: Differences between $\mathbf{x}^*$ and ξ_{mean}id.

The numerical solution of the two-dimensional minimization problems (21) indicates that for both examples $e \in \{\mathtt{M}, \mathtt{L}\}$ there exist unique values $t_1^* < t_2^* < 0$ such that equation (15) is satisfied. It determines the values of these minimizing time points as a by-product. By setting $x_1^* := \mu t_1^*$ and $x_2^* := \phi\delta + \psi t_2^*$ we can apply lemmas 5.4 and 5.7 in order to calculate minimizing paths associated with these time points and fixed values. (In order to apply lemma 5.4 one first has to calculate the associated dual values y_1^*, y_2^*.) Hence, in all numerical examples considered here there exists a unique path $\mathbf{x}^*$ minimizing the rate function I subject to the conditions $\mathbf{x}^*(t_1^*) = x_1^*$ and $\mathbf{x}^*(t_2^*) = x_2^*$. According to theorem 4.3, this path $\mathbf{x}^*$ characterizes the asymptotically most likely behavior of the centered and scaled cross traffic which leads to a large queue length at the second node. This asymptotically most likely conditional behavior has a period with an extraordinarily large number of arrivals in the cross traffic followed by a period with unusually few arrivals. In figure 6.2 we show the minimizing path $\mathbf{x}^*$ for each example $e = \{\mathtt{M}, \mathtt{L}\}$, and $\beta \in \{1, 2\}$. We note that the minimizing path for the example $e = \mathtt{L}$ does not depend on β, that is, it is invariant under simultaneous scalings of the transition rates g_1 and g_2.

In figure 6 we show the associated queue length behaviors $\mathbf{Q}_{\mathtt{V},\mathtt{A}}(\mathbf{x}^*)$, $\mathbf{Q}_{\mathtt{C},\mathtt{A}}(\mathbf{x}^*)$, $\mathbf{Q}_{\mathtt{B}}(\mathbf{x}^*)$ for $\beta = 1$ and the two cases $e \in \{\mathtt{M}, \mathtt{L}\}$. The case $e = \mathtt{L}$ reproduces the piecewise linear behavior called "buffer B exploits buffer A" in [19]. The other graphs contrast this known case with the new results obtained in this work.

7. REFERENCES

[1] S. Asmussen. *Applied Probability and Queues*, volume 51 of *Applications of Mathematics: Stochastic Medelling and Applied Probability*. Springer, New York, second edition, 2003.

[2] R. Bahadur and S. Zabell. Large deviations of the sample mean in general vector spaces. *The Annals of Probability*, 7(4):587–621, 1979.

[3] J. A. Bucklew. *Large Deviations Techniques in Decision, Simulation and Estimation*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, 1990.

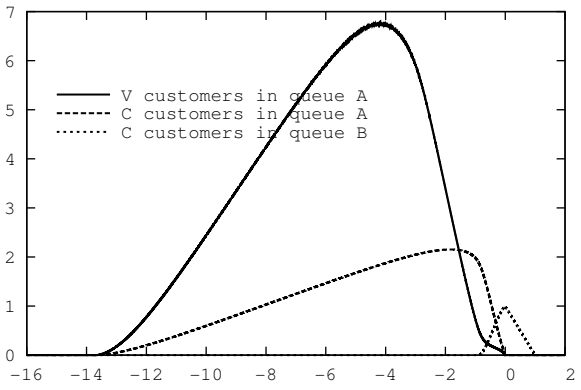

Figure 5: Queue lengths $\mathbf{Q}_{\mathtt{V},\mathtt{A}}(\mathbf{x}^*)$, $\mathbf{Q}_{\mathtt{C},\mathtt{A}}(\mathbf{x}^*)$, $\mathbf{Q}_{\mathtt{B}}(\mathbf{x}^*)$ for $e = \mathtt{M}$.

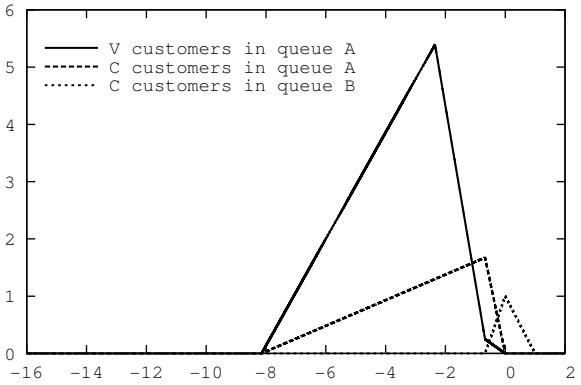

Figure 6: Queue lengths $\mathbf{Q}_{\mathtt{V},\mathtt{A}}(\mathbf{x}^*)$, $\mathbf{Q}_{\mathtt{C},\mathtt{A}}(\mathbf{x}^*)$, $\mathbf{Q}_{\mathtt{B}}(\mathbf{x}^*)$ for $e = \mathtt{L}$.

[4] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Jones and Bartlett, London, 1993.

[5] A. Ganesh and N. O'Connell. A large deviation principle with queueing applications. *Stochastics and Stochastics Reports*, 73(1-2):25–35, 2002.

[6] G. Kesidis, J. Walrand, and C.-S. Chang. Effective bandwidth for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking*, 1(4):424–427, 1993.

[7] P. Lancaster and M. Tismenetsky. *The Theory of Matrices*. Academic Press, New York, second edition, 1985.

[8] R. Loynes. The stability of a queue with non-independent inter-arrivals and service times. *Math. Proc. Cambridge Philos. Soc.*, 58:497–520, 1962.

[9] K. Majewski. Single class queueing networks with discrete and fluid customers on the time interval $\mathbb{R}$. *Queueing Systems*, 36(4):405–435, 2000.

[10] K. Majewski. Large deviations for multi-dimensional reflected fractional Brownian motion. *Stochastics and Stochastics Reports*, 75(4):233–257, 2003. Corrigendum 76(5):479, 2004.

[11] K. Majewski. Fractional Brownian heavy traffic approximations of multiclass feedforward queueing networks. *Queueing Systems*, 50(2-3):199–230, 2005.

[12] K. Majewski. Sample path large deviations for multiclass feedforward queueing networks in critical loading. *The Annals of Applied Probability*, 16(4):1893–1924, 2006.

[13] K. Majewski. Large deviation properties of constant rate data streams sharing a buffer with long-range dependent traffic in critical loading. *Advances in Applied Probability*, 39(2):407–428, 2007.

[14] K. Majewski. Minimizing large deviation paths for a family of long-range dependent processes and their fractional Brownian approximations. *Stochastic Models*, 23(1):49–77, 2007.

[15] K. Majewski. Sample path moderate deviations for the cumulative fluid produced by an increasing number of exponential on-off sources. *Queueing Systems*, 56:9–26, 2007.

[16] M. Mandjes. Rare event analysis of the state frequencies of a large number of Markov chains. *Communications in Statistics - Stochastic Models*, 15(3):577–592, 1999.

[17] M. Mandjes and P. Mannersalo. Queueing systems fed by many exponential on-off sources: an infinite-intersection approach. *Queueing Systems*, 54:5–20, 2006.

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

[19] K. Ramanan and P. Dupuis. Large deviation properties of data streams that share a buffer. *The Annals of Applied Probability*, 8(4):1070–1129, 1998.

[20] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer Verlag, Berlin, 1998.

[21] A. Wächter and L. T. Biegler. On the implmentation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program., Ser. A*, 106:25–57, 2006.

[22] A. Weiss. A new technique for analyzing large traffic systems. *Advances in Applied Probability*, 18:506–532, 1986.

[23] D. Wischik. Sample path large deviations for queues with many inputs. *The Annals of Applied Probability*, 11:379–404, 2001.

APPENDIX

A. PROOF OF THEOREM 4.3

Throughout this section we assume that the prerequisites of Theorem 4.3 are satisfied, namely those at the beginning of Sections 4 and 4.2.

The last statement of the theorem is a consequence of the statements in the text of Section 3. Hence it only remains to prove inequalities (13) and (14), which is done in the remaining two sections.

A.1 Proof of inequality (13)

Since the mapping $\mathbf{Q}_{\mathtt{B}}$ is continuous and the set $\{\mathbf{q} \in \mathcal{C} : \mathbf{q}(0) \geq \delta\}$ closed, we obtain from the upper bound (1) of

the large deviation principle of Theorem 4.1

$$\limsup_{k\to\infty} \frac{1}{k} \log P\left(B_k \geq \delta\right) \quad \leq \quad - \inf_{\substack{\mathbf{x} \in \mathcal{I}_I, \\ \mathbf{Q}_{\mathtt{B}}(\mathbf{x})(0) \geq \delta,}} I(\mathbf{x}). \quad (22)$$

When the right hand side in this display is infinite inequality (13) is trivially satisfied. In the following we can therefore assume that the right hand side in the previous display is finite.

Because the rate function I has compact level sets, there exists a path $\mathbf{a} \in \mathcal{I}_I$ satisfying $\mathbf{Q}_{\mathtt{B}}(\mathbf{a})(0) \geq \delta$ and being such that the value of $-I(\mathbf{a})$ is equal to the right hand side of inequality (22). Since this value is finite the path $\mathbf{a}$ is continuous and satisfies $\mathbf{a}(0) = 0$.

There exists a largest time $\tau_{\mathtt{B}} < 0$ at which the continuous path $\mathbf{Q}_{\mathtt{B}}(\mathbf{a})$ is zero. (Queue $\mathtt{B}$ cannot be non-empty throughout the negative time interval because the long-term arrival rate α of the constant rate traffic is strictly less than the service rate $\mu_{\mathtt{B}}$.) Due to the shift property (11) we can assume without loss of generality that $\mathbf{Q}_{\mathtt{B}}(\mathbf{a})(0) = \delta$ and $0 < \mathbf{Q}_{\mathtt{B}}(\mathbf{a})(t) < \delta$ if $\tau_{\mathtt{B}} < t < 0$.

We let $\tau_{\mathtt{A}} < 0$ be the largest time with $\mathbf{Q}_{\mathtt{A}}(\mathbf{a})(\tau_{\mathtt{A}}) = 0$. We show $\tau_{\mathtt{A}} < \tau_{\mathtt{B}}$ by contradiction: If there were a time t with $\tau_{\mathtt{B}} \leq t < 0$ with $\mathbf{Q}_{\mathtt{A}}(\mathbf{a})(t) = 0$ then $\mathbf{D}_{\mathtt{A},\mathtt{C}}(\mathbf{a})$ could increase from time t to 0 at most by $-\alpha t$. Since the service speed at buffer $\mathtt{B}$ is $\sigma_{\mathtt{B}} > \alpha$ this would imply that $Q_{\mathtt{B}}(\mathbf{a})$ at time t is strictly greater than δ which is a contradiction.

Since the queues cannot idle when their buffer is non-empty this implies

$$\mathbf{Q}_{\mathtt{A}}(\mathbf{a})(t) \quad = \quad \mathbf{a}(t) - \mathbf{a}(\tau_{\mathtt{A}}) - \mu(t - \tau_{\mathtt{A}}) \quad (23)$$

for $t \in [\tau_{\mathtt{A}}, 0]$ and

$$\mathbf{Q}_{\mathtt{B}}(\mathbf{a})(t) \quad = \quad \mathbf{D}_{\mathtt{A},\mathtt{C}}(\mathbf{a})(t) - \mathbf{D}_{\mathtt{A},\mathtt{C}}(\mathbf{a})(\tau_{\mathtt{B}}) - \sigma_{\mathtt{B}}(t - \tau_{\mathtt{B}}) \quad (24)$$

for $t \in [\tau_{\mathtt{B}}, 0]$. In particular, setting $t = 0$ the last display implies

$$\mathbf{D}_{\mathtt{A},\mathtt{C}}(\mathbf{a})(0) - \mathbf{D}_{\mathtt{A},\mathtt{C}}(\mathbf{a})(\tau_{\mathtt{B}}) \quad = \quad \delta - \sigma_{\mathtt{B}}\tau_{\mathtt{B}}. \quad (25)$$

If we define

$$\begin{aligned} \theta_0 \quad &:= \quad \mathbf{T}_{\mathtt{A}}(\mathbf{a})(0), \\ \theta_1 \quad &:= \quad \mathbf{T}_{\mathtt{A}}(\mathbf{a})(\tau_{\mathtt{B}}). \end{aligned}$$

then, displays (7) and (25) imply

$$\begin{aligned} \tau_{\mathtt{B}} \quad &= \quad \frac{1}{\sigma_{\mathtt{B}}} \left(\delta - \mathbf{D}_{\mathtt{A},\mathtt{C}}(\mathbf{a})(0) + \mathbf{D}_{\mathtt{A},\mathtt{C}}(\mathbf{a})(\tau_{\mathtt{B}})\right) \\ &= \quad \frac{\delta - \alpha(\theta_0 - \theta_1)}{\sigma_{\mathtt{B}}}. \end{aligned}$$

Since $\mathbf{Q}_{\mathtt{A}}(\mathbf{a}) \circ \mathbf{T}_{\mathtt{A}}(\mathbf{a}) = \sigma_{\mathtt{A}}(\mathbf{id} - \mathbf{T}_{\mathtt{A}}(\mathbf{a}))$ (compare equation (8)) we obtain from (23) and (24)

$$\begin{aligned} \mathbf{a}(\theta_1) - \mathbf{a}(\tau_{\mathtt{A}}) - \mu(\theta_1 - \tau_{\mathtt{A}}) = \mathbf{Q}_{\mathtt{A}}(\mathbf{a})(\theta_1) \quad &= \quad \sigma_{\mathtt{A}}(\tau_{\mathtt{B}} - \theta_1), \\ \mathbf{a}(\theta_0) - \mathbf{a}(\tau_{\mathtt{A}}) - \mu(\theta_0 - \tau_{\mathtt{A}}) = \mathbf{Q}_{\mathtt{A}}(\mathbf{a})(\theta_0) \quad &= \quad -\sigma_{\mathtt{A}}\theta_0. \end{aligned}$$

Taking differences and rearranging terms this yields

$$\begin{aligned} \mathbf{a}(\theta_1) - \mathbf{a}(\theta_0) \quad &= \quad \mu(\theta_1 - \theta_0) + \sigma_{\mathtt{A}}(\tau_{\mathtt{B}} - \theta_1 + \theta_0) \\ &= \quad \frac{\sigma_{\mathtt{A}}}{\sigma_{\mathtt{B}}}\delta + \left(\frac{\sigma_{\mathtt{A}}}{\sigma_{\mathtt{B}}} - 1\right)\alpha(\theta_1 - \theta_0). \end{aligned}$$

Hence the path $\mathbf{x}^* := \Theta_{\theta_0}\mathbf{a} - \mathbf{a}(\theta_0)$ satisfies $I(\mathbf{x}^*) = I(\mathbf{a})$ and, with $t_1 := \tau_{\mathtt{A}} - \theta_0$ and $t_2 := \theta_1 - \theta_0$,

$$\begin{aligned} \mathbf{x}^*(t_1) \quad &= \quad \mathbf{a}(\tau_{\mathtt{A}}) - \mathbf{a}(\theta_0) \\ &= \quad \sigma_{\mathtt{A}}\theta_0 - \mu(\theta_0 - \tau_{\mathtt{A}}) \leq \mu t_1, \\ \mathbf{x}^*(t_2) \quad &= \quad \mathbf{a}(\theta_1) - \mathbf{a}(\theta_0) = \phi\delta + \psi t_2. \end{aligned}$$

In particular, we established

$$\begin{aligned} &\limsup_{k\to\infty} \frac{1}{k} \log P(B_k \geq \delta) \leq -I(\mathbf{a}) \\ &\quad = \quad -I(\mathbf{x}^*) \leq - \inf_{\substack{t_1 < t_2 < 0, \\ x_1 \leq \mu t_1, \\ x_2 \geq \phi\delta + \psi t_2}} I^{t_1,t_2}(x_1, x_2), \end{aligned}$$

which completes the proof of the inequality in (13).

In order to prove the right hand side equality in display (13), we just have to show "$\leq$". Hence we can assume without loss of generality that the second last term is finite. For every $\epsilon > 0$ we can therefore find values $t_1' < t_2' < 0$, $x_1' \leq \mu t_1'$, $x_2' \geq \phi\delta + \psi t_2'$ and a path $\mathbf{a} \in \mathcal{I}_I$ such that $\mathbf{a}(t_1') = x_1'$, $\mathbf{a}(t_2') = x_2'$ and

$$\begin{aligned} I(\mathbf{a}) \quad &= \quad I_{t_1',t_2'}(x_1', x_2') \\ &\leq \quad \epsilon + \inf_{\substack{t_1 < t_2 < 0, \\ x_1 \leq \mu t_1, \\ y \geq \phi\delta + \psi t_2}} I_{t_1,t_2}(x_1, x_2) < \infty. \end{aligned}$$

In particular, $\mathbf{a}$ is continuous. Since $\phi > 0$ and $\psi < \mu$, there exist values $t_1'' < t_2'' < 0$ such that $\mathbf{a}(t_1'') = \mu t_1''$ and $\mathbf{a}(t_2'') = \phi\delta + \psi t_2''$. This implies

$$\begin{aligned} &I_{t_1'',t_2''}(\mu t_1'', \phi\delta + \psi t_2'') \\ &\quad \leq \quad I(\mathbf{a}) \leq \epsilon + \inf_{\substack{t_1 < t_2 < 0, \\ x_1 \leq \mu t_1, \\ x_2 \geq \phi\delta + \psi t_2}} I_{t_1,t_2}(x_1, x_2). \end{aligned}$$

Letting ϵ tend to 0, we therefore recover the equation in display (13).

A.2 Proof of inequality (14)

In order to show inequality (14) we can assume without loss of generality that its right hand side is finite. For $\epsilon > 0$ we can therefore find values $t_1' < t_2' < 0$, $x_1' = \mu t_1'$, and $x_2' > \phi\delta + \psi t_2'$ such that

$$I_{t_1',t_2'}(x_1', x_2') \quad \leq \quad \epsilon + \inf_{\substack{t_1 < t_2 < 0, \\ x_1 = \mu t_1, \\ x_2 > \phi\delta + \psi t_2}} I_{t_1,t_2}(x_1, x_2) < \infty.$$

Since the function I has compact level sets we can find a path $\mathbf{a} \in \mathcal{I}_I$ such that $I(\mathbf{a})$ equals the value of the left hand side in this display and the equations $\mathbf{a}(t_1') = x_1'$ and $\mathbf{a}(t_2') = x_2'$ are satisfied. In particular, the function $\mathbf{a}$ is continuous and satisfies $\mathbf{a}(0) = 0$. Without loss of generality, we can additionally assume

$$\mathbf{a}(t) \quad > \quad \mu t$$

for every t with $t_1' < t < 0$. (Otherwise use property (11) to modify $\mathbf{a}$ in this sense.) This last inequality implies that queue $\mathtt{A}$ is non-empty during the time interval $]t_1, 0]$.

The $-\alpha t_2'$ type $\mathtt{C}$ customers which arrive in the time interval $[t_2', 0]$ leave the first queue during the time interval $[t_2' + \mathbf{Q}_{\mathtt{A}}(\mathbf{a})(t_2')/\sigma_{\mathtt{A}}, \mathbf{Q}_{\mathtt{A}}(\mathbf{a})(0)/\sigma_{\mathtt{A}}]$. Since

$$\begin{aligned} \mathbf{Q}_{\mathtt{A}}(\mathbf{a})(0) - \mathbf{Q}_{\mathtt{A}}(\mathbf{a})(t_2') \quad &= \quad -\mathbf{a}(t_2') - \alpha t_2' + \sigma_{\mathtt{A}} t_2' \\ &= \quad -x_2' + \mu t_2', \end{aligned}$$

at most

$$\sigma_{\mathtt{B}}\left(-t_2' - \frac{x_2' - \mu t_2'}{\sigma_{\mathtt{A}}}\right) \quad < \quad \delta - \alpha t_2'$$

customers can be served during this time interval by buffer B. Hence more than δ type C customers remain unserved in buffer B at time $\tau := \mathbf{Q}_{\mathtt{A}}(\mathbf{a})(0)/\sigma_{\mathtt{A}}$. Through the definition $\mathbf{x}^* := \Theta_\tau ba - \mathbf{a}(\tau)$ we therefore obtain a function which satisfies $I(\mathbf{x}^*) = I(\mathbf{a})$ and

$$\mathbf{Q}_{\mathtt{B}}(\mathbf{x}^*)(0) \quad = \quad \mathbf{Q}_{\mathtt{B}}(\mathbf{a})(\tau) > \delta.$$

Since the mapping $\mathbf{Q}_{\mathtt{B}}$ is continuous we conclude that there is an environment of $\mathbf{x}^*$ such that all elements $\mathbf{x}$ in this environment satisfy $\mathbf{Q}_{\mathtt{B}}(\mathbf{x})(0) > \delta$. With the help of the lower bound (2) of the large deviation principle for the sequence (10) this yields

$$\begin{aligned}\liminf_{k\to\infty} \frac{1}{k} \log P(B_s \geq \delta) \\ \geq \quad -I(\mathbf{x}^*) \geq -\epsilon - \inf_{\substack{t_1 < t_2 < 0,\\ x_1 < \mu t_1,\\ x_2 > \phi\delta + \psi t_2}} I_{t_1,t_2}(x_1, x_2).\end{aligned}$$

Letting ϵ decrease to zero we obtain inequality (14) and completed the proof.