Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

Large deviations for complex buffer architectures: the short-range dependent case

M.R.H. Mandjes

REPORT PNA-E0413 JULY 2004

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

Large deviations for complex buffer architectures: the short-range dependent case

ABSTRACT
This paper considers Gaussian flows multiplexed in a queueing network, where the underlying correlation structure is assumed to be short-range dependent. Whereas previous work mainly focused on the FIFO setting, this paper addresses overflow characteristics of more complex buffer architectures. We subsequently analyze the tandem queue, a priority system, and generalized processor sharing. In a many-sources setting, we explicitly compute the exponential decay rate of the overflow probability. Our study relies on large-deviations arguments, e.g., Schilder's theorem.

2000 Mathematics Subject Classification: 60K25
Keywords and Phrases: sample-path large deviations; Gaussian traffic; Schilder's theorem; short-range dependence; tandem queue; priority queue; generalized processor sharing; c ommunication networks; differentiated services

Large deviations for complex buffer architectures: the short-range dependent case

Michel Mandjes *

Abstract

This paper considers Gaussian flows multiplexed in a queueing network, where the underlying correlation structure is assumed to be short-range dependent. Whereas previous work mainly focused on the FIFO setting, this paper addresses overflow characteristics of more complex buffer architectures. We subsequently analyze the tandem queue, a priority system, and generalized processor sharing. In a many-sources setting, we explicitly compute the exponential decay rate of the overflow probability. Our study relies on large-deviations arguments, e.g., Schilder's theorem.

Key words: sample-path large deviations – Gaussian traffic – Schilder's theorem – short-range dependence – tandem queue – priority queue – generalized processor sharing – communication networks – differentiated services

*M. Mandjes is with CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands, and Korteweg-de Vries Institute, University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands. Email: michel@cwi.nl

1 Introduction

Over the past two decades, a significant research effort has been devoted to the large-deviations analysis of queues. It has culminated in a wealth of valuable contributions to the understanding of the occurrence of rare events (such as buffer overflow) in queues. Exact computation of the overflow probability is usually a demanding task, thus motivating the search for accurate approximations and asymptotics. Large-deviations analysis usually provides a rough (logarithmic) characterization of the overflow probability (in terms of an exponential decay rate), but also insight into the system's 'path' from 'average behavior' to the rare event.
In particular, the celebrated *many-sources* scaling, introduced in a seminal paper by Weiss [22], has provided a rich framework for obtaining large-deviations results. In a many-sources setting, one considers a queueing system fed by the superposition of n i.i.d. traffic sources, with the service rates and buffer thresholds scaled with n as well. In the setting of a single first-in-first-out (FIFO) queue, under very mild conditions on the source behavior, it is possible to calculate the *exponential* decay of the probability $p_n(b, c)$ that the queue (fed by n sources, and emptied at a deterministic rate nc) exceeds level nb, see, e.g., [5, 6].
Although single-class single-node FIFO queues serve as a useful baseline model and provide valuable insight, they clearly have serious limitations. First of all, traffic streams usually traverse *concatenations* of hops (rather than just a single node). Secondly, networks increasingly support a wide variety of traffic types, with each of them having its own specific (stochastic) characteristics and Quality-of-Service requirements in terms of packet delay, loss, and throughput metrics. In order to deal with the heterogeneity in traffic types, networks will typically rely on discriminatory scheduling mechanisms to distinguish between streams of the various classes, such as *priority scheduling* mechanisms, or the more advanced *Generalized Processor Sharing* (GPS) discipline, cf. [20, 21]. Thus, a fundamental understanding of the large-deviations behavior of stochastic networks with non-FIFO scheduling is expected to play a crucial role in providing end-to-end Quality-of-Service in multi-class networks. However, only few large-deviations results are known for these more complex buffer architectures.
As indicated above, each type of traffic has its own stochastic properties, often summarized by the *correlation structure*. One commonly distinguishes between *short-range dependent* input (with just a mild correlation) and long-range dependent input (in which correlations decay relatively slowly). It is noted that Gaussian models cover both, see for instance [1, 11].

Contribution. In this paper we focus on some of the above described 'complex buffer architectures'. More specifically, we will derive the many-sources asymptotics of buffer overflow in the tandem queue, the priority queue, and the queue operating under (two-class) GPS. We focus on the case of short-range dependent inputs; this class of inputs covers for instance the Gaussian counterpart [1] of the celebrated AMS model [2] (i.e., a superposition of exponential

on-off sources), which is a standard model for coded voice. In this way, this work complements other papers that predominantly focuses on long-range dependent inputs [13, 14].
In this paper we apply results from our previous work [11, 12], which applied to general Gaussian sources. There we found a lower bound on the decay rate, which was tight under a specific explicit condition. However, both the computation of the lower bound, and the verification of the tightness condition usually require non-neglible numerical computations. In the present paper, by restricting ourselves to the class of short-range dependent traffic, the main benefit is that we can explicitly characterize the decay rate (in terms of the model parameters); hence there is no need for any numerical computation.

Literature, related work. Queues with Gaussian inputs were extensively discussed in a series of articles by Mannersalo and Norros [1, 15, 16, 17]. As far as the more complex buffer architectures are concerned, the articles by Mannersalo and Norros mainly focus on heuristic approximations for the decay rate of the overflow probabilities. Mandjes and Van Uitert show in [11, 12] that these heuristics are typically close, but that there is a gap with the exact outcome; as indicated above, they find bounds on the decay rate, and derive conditions under which these bounds are tight.
The asymptotics for short-range dependent traffic feeding into a single FIFO queue, as presented in, e.g., Botvich and Duffield [5, Th. 3], turn out to be relevant for our paper – this will be discussed in detail in Section 2.3. Our work is also related to Wischik's results on sample-path large deviations for the single queue and priority system [23]. Zhang [24] focuses on large buffer asymptotics in a (discrete-time) GPS system with short-range dependent inputs; in Section 6 we comment on the relation with our (many-sources asymptotics) results for GPS.

This paper is organized as follows. Section 2 presents preliminaries. In Section 3 the tandem queue is analyzed, whereas Section 4 deals with the priority queue, and Section 5 with the GPS system. Section 6 gives a discussion of the results, and some concluding remarks.

2 Preliminaries

This paper is on rare events for queues with complex buffer architectures, fed by many short-range dependent Gaussian inputs – in this section we present the necessary prerequisites. The first subsection introduces Gaussian sources, and defines the notion of short-range dependence used in this paper. The second subsection recapitulates the framework for analyzing rare events in the many-sources setting: Cramér's theorem, and (the generalized version of) Schilder's theorem. The last subsection revisits the single FIFO-queue, and indicates what type of results we wish to derive for the more complex buffer architectures (tandem, priority, GPS).

2.1 Gaussian processes

We consider n sources, behaving as i.i.d. Gaussian processes with stationary increments. Define $A_i(s,t)$ as the amount of traffic generated by the ith source in $(s,t]$, with $s < t$ and $s,t \in \mathbb{R}$. Denote by $A(s,t)$ the generic random variable corresponding to a single source. The Gaussian sources are characterized by their *mean rate* $\mu \geq 0$ and their *variance function* $v(\cdot)$. More precisely, for all s,t with $s < t$, $\mathbb{E}A(s,t) = \mu(t-s)$ (with μ smaller than c_2) and $\mathbb{V}\mathrm{ar}A(s,t) = v(t-s)$. We also define the *centered process* $\bar{A}(\cdot)$ by putting $\bar{A}(t) := A(0,t) - \mu t$. In the sequel we often use the bivariate normal distribution of $(A(0,s), A(0,t))$; we define $\Gamma(s,t) := \mathbb{C}\mathrm{ov}(A(0,s), A(0,t))$. It holds that

$$\Gamma(s,t) = \frac{1}{2}(v(t) + v(s) - v(t-s)). \tag{1}$$

Notice that the possibility of *negative traffic* is not explicitly ruled out, as opposed to 'classical' input processes, such as (compound) Poisson input or on-off sources. For tandem queues (in which the output of the first queue feeds into a second queue), it is noticed in [11] that, by choosing an appropriate representation for the queue length of the second queue, negative queue lengths can be easily avoided. For priority systems, [17] explains in detail how to circumvent the problem of negative queue lengths (a discrete-time version of the priority discipline is introduced, in which negative traffic can annihilate queued traffic). GPS queues can be dealt with similarly.

Assumption 2.1 *We assume that*
(A1) $v(\cdot)$ *is continuous, differentiable on* $(0,\infty)$*;*
(A2) $v(\cdot)$ *is strictly increasing;*
(A3) *for some* $\alpha < 2$ *it holds that* $v(t)t^{-\alpha} \to 0$ *as* $t \to \infty$*.*

Definition 2.2 *Consider a Gaussian source with stationary increments.*

(i) Brownian motion: *The source* BM(λ, μ) *has mean input rate* μ *and variance function* $v(t) = \lambda t$*, for* $t \geq 0$*.*

(ii) Asymptotically linear variance: *The source* ALV(κ, λ, μ) *has mean input rate* μ*, and a variance function* $v(\cdot)$ *satisfying*

$$\lim_{t\to\infty} v(t) - \lambda t = \kappa.$$

Example I: Ornstein-Uhlenbeck input. Following Section 4.3 of [5], we represent an (integrated) Ornstein-Uhlenbeck input process by choosing $v(t) = t - 1 + e^{-t}$. This corresponds to an ALV process with $\kappa = -1$ and $\lambda = 1$.

Example II: M/G/∞ inputs with Pareto job sizes. A single M/G/∞ source consists of jobs that arrive according to a Poisson process of rate $\bar{\lambda}$. They stay in the system during some holding time, that is distributed as a random variable D (with $\mathbb{E}D < \infty$). During this holding time, any jobs generates a constant traffic stream at a rate of, say, 1. We assume $\mathbb{P}(D > t) = (t+1)^{-\alpha}$, i.e., D has a Pareto tail. Take $\alpha > 1$; then $\mathbb{E}D = (\alpha - 1)^{-1} < \infty$.
We now consider the Gaussian input process that has the seam mean input rate and variance function. The mean input rate is trivially $\mu := \bar{\lambda}\,\mathbb{E}D$ per unit time, whereas the variance function reads (assume for ease that $\alpha \notin \{2, 3\}$)

$$v(t) = \nu \cdot \left(1 - (t+1)^{3-\alpha} + (3-\alpha)t\right) \quad \text{with } \nu := \frac{2\bar{\lambda}}{(3-\alpha)(2-\alpha)(\alpha-1)},$$

see [10]. Importantly, if $\alpha \in (1, 2)$ the traffic process has essentially long-range dependent properties, as $v(t)$ is superlinear; if $\alpha > 3$, the process is ALV with $\kappa = \nu$ and $\lambda = \nu(3-\alpha)$. The intermediate case $\alpha \in (2, 3)$ will be commented on in Section 6.

2.2 Large deviations

The analysis in the next sections relies on a sample-path large deviations principle (LDP) for (centered) Gaussian processes. This subsection is devoted to a brief description of the main theorem in this field, (the generalized version of) *Schilder's theorem* [4]. However, we start by recalling (the multivariate version of) the well-known *Cramér's theorem*, see [7, Thm. 2.2.30]. We use the standard notation $\langle \cdot, \cdot \rangle$ for the inner product: $\langle a, b \rangle := a^{\mathrm{T}} b = \sum_{i=1}^{d} a_i b_i$.

Theorem 2.3 [Multivariate Cramér] *Let $X_i \in \mathbb{R}^d$ be i.i.d. d-dimensional random vectors, distributed as a random vector X with moment-generating function $\mathbb{E}e^{\langle \theta, X \rangle} < \infty$ (for any $\theta \in \mathbb{R}^d$). Then $n^{-1} \sum_{i=1}^{n} X_i$ satisfies the following LDP:*
(a) For any closed set $F \subset \mathbb{R}^d$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \in F\right) \leq -\inf_{x \in F} \Lambda(x);$$

(b) For any open set $G \subset \mathbb{R}^d$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \in G\right) \geq -\inf_{x \in G} \Lambda(x),$$

where the large deviations rate function $\Lambda(\cdot)$ is given by

$$\Lambda(x) := \sup_{\theta \in \mathbb{R}^d} \left(\langle \theta, x \rangle - \log \mathbb{E}e^{\langle \theta, X \rangle}\right), \tag{2}$$

Remark 2.4 Consider the specific case that X has a multivariate Normal distribution with mean vector μ and $(d\times d)$ non-singular covariance matrix Σ. Using $\log \mathbb{E}e^{\langle\theta,X\rangle} = \langle\theta,\mu\rangle + \frac{1}{2}\theta^{\mathrm{T}}\Sigma\theta$ it is not hard to derive that, with $(x-\mu)^{\mathrm{T}} \equiv (x_1-\mu_1,\ldots,x_d-\mu_d)$,

$$\theta^\star = \Sigma^{-1}(x-\mu) \quad \text{and} \quad \Lambda(x) = \frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu), \tag{3}$$

where $\theta^\star$ optimizes (2); it is well-known that $\Lambda(\cdot)$ is convex. $\diamond$

Lemma 2.5 *Let $(X_i, Y_i) \in \mathbb{R}^2$ i.i.d. bivariate normal random variables, with mean vector $(\mu_X, \mu_Y)^{\mathrm{T}}$, and two-dimensional covariance matrix*

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho(X,Y) \\ \rho(X,Y) & \sigma_Y^2 \end{pmatrix}.$$

Fix $a > \mu_X$ and $b > \mu_Y$. Then

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq a, \frac{1}{n}\sum_{i=1}^n Y_i \leq b \right) = -\frac{1}{2}\frac{(a-\mu_X)^2}{\sigma_X^2},$$

if

$$\mathbb{E}(Y \mid X = a) = \mu_Y + \frac{\rho(X,Y)}{\sigma_X^2}(a-\mu_X) < b;$$

otherwise

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq a, \frac{1}{n}\sum_{i=1}^n Y_i \leq b \right) = -\frac{1}{2}(a-\mu_X, b-\mu_Y)^{\mathrm{T}}\Sigma^{-1}\begin{pmatrix} a-\mu_X \\ b-\mu_Y \end{pmatrix}.$$

Proof. Using the above remark, we have to minimize $\Lambda(x,y)$ over all $x \geq a$ and $y \leq b$. Two cases may occur, as illustrated in Figure 1. The crucial difference between the graphs is that in the left figure, the contour that touches the line $x = a$ has a y-value lower than b, whereas in the right figure the opposite is the case. This is formalized as follows. Let y_0 solve

$$\left.\frac{\partial\Lambda(x,y)}{\partial y}\right|_{x=a} = 0, \qquad \text{i.e., } \; y_0 = \mu_Y + \frac{\rho(X,Y)}{\sigma_X^2}(a-\mu_X).$$

Hence, the left panel shows that if $y_0 \leq b$, then the optimum is attained at some point (x,y) in $\{a\}\times[0,b)$, whereas, according to the right panel, $y_0 > b$ implies an optimum in $(y,z) = (a,b)$. If $y_0 \leq b$, $\Lambda(a,y_0) = (a-\mu_X)^2/2\sigma_X^2$, then indeed independent of b. $\square$

We now sketch the framework of Schilder's sample-path LDP, as established in [4], see also [8]. We restrict ourselves to the aspects that are relevant in the present study; for more details we refer to [1, 11, 15]. Consider the n i.i.d. centered Gaussian processes $\bar{A}_i(\cdot) := \{\bar{A}_i(t), t \in \mathbb{R}\}$

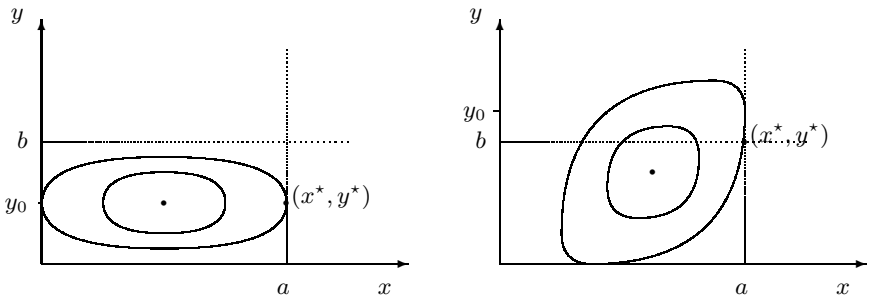

Figure 1: Contour lines of the (two-dimensional) rate function.

with stationary increments, and covariance $\mathbb{C}\mathrm{ov}\{\bar{A}_i(s), \bar{A}_i(t)\}$, which obviously equals $\Gamma(s,t)$ defined in Section 2.1. Define the path space Ω as

$$\Omega := \left\{ \omega : \mathbb{R} \to \mathbb{R},\ \text{continuous},\ \omega(0) = 0,\ \lim_{t\to\infty} \frac{\omega(t)}{1+t} = \lim_{t\to-\infty} \frac{\omega(t)}{1+t} = 0 \right\},$$

which is a separable Banach space by imposing a specific norm, as explained in [15]. Next we introduce and define the *reproducing kernel Hilbert space* $R \subseteq \Omega$ – see [3] for a more detailed account – with the property that its elements are roughly as smooth as the covariance function $\Gamma(s,\cdot)$. We start from a 'smaller' space S, defined by

$$S := \left\{ \omega : \mathbb{R} \to \mathbb{R},\ \omega(\cdot) = \sum_{i=1}^{n} a_i \Gamma(s_i, \cdot),\ \ a_i, s_i \in \mathbb{R}, n \in \mathbb{N} \right\}.$$

The inner product on this space S is, for $\omega_a, \omega_b \in S$, defined as

$$\langle \omega_a, \omega_b \rangle_R \ := \ \left\langle \sum_{i=1}^{n} a_i \Gamma(s_i,\cdot), \sum_{j=1}^{n} b_j \Gamma(s_j,\cdot) \right\rangle_R \ = \ \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \Gamma(s_i, s_j).$$

The closure of S under this norm is defined as the space R. With the norm defined by $||\omega||_R := \sqrt{\langle \omega, \omega \rangle}$, we define the rate function of the sample-path LDP:

$$I(\omega) := \begin{cases} \frac{1}{2}||\omega||_R^2 & \text{if } \omega \in R; \\ \infty & \text{otherwise.} \end{cases}$$

Under the above assumptions, e.g., (A1) and (A3), the following sample-path LDP holds.

Theorem 2.6 [Generalized Schilder] $n^{-1}\sum_{i=1}^{n} \bar{A}_i(\cdot)$ *satisfies the following LDP:*
(a) For any closed set $F \subset \Omega$*,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} \bar{A}_i(\cdot) \in F \right) \leq - \inf_{\omega \in F} I(\omega);$$

(b) For any open set $G \subset \Omega$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \bar{A}_i(\cdot) \in G\right) \geq -\inf_{\omega\in G} I(\omega).$$

The following application of 'Schilder' was proven in [1].

Lemma 2.7 *Let a, t be positive, and define $F_{a,t} := \{f \mid f(t) \geq a\}$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \bar{A}_i(\cdot) \in F_{a,t}\right) = -\frac{1}{2}\frac{a^2}{v(t)};$$

the optimizing path $f^\star(\cdot)$ is, for $r \in \mathbb{R}$, given by

$$f^\star(r) = \frac{\Gamma(r,t)}{v(t)} \cdot a.$$

Remark 2.8 For centered BM, the optimizing path $f^\star(\cdot) \in F_{a,t}$ from Lemma 2.7, equals 0 for negative r, grows linearly with slope a/t for $r \in [0,t]$, and remains at level a for $r \geq t$. $\diamond$

2.3 Results for single FIFO queue

This subsection recalls the results for overflow in the single FIFO queue. We consider n i.i.d. Gaussian sources feeding into a queue with link rate nc. We focus on the probability that the (stationary) buffer content Q_n exceeds level nb. This probability decays exponentially in n; in particular, we see that the decay rate is linear in the buffer size for BM sources, and asymptotically linear for ALV sources.

Let $A[f](s,t)$ denote the amount of traffic generated in $(s,t]$ if the sample-mean process $n^{-1}\sum_{i=1}^{n} A_i(\cdot)$ follows path f, or, in other words, $A[f](s,t) = f(t) - f(s)$. The paths leading to overflow in the FIFO setting are

$$F^{(\mathrm{f})}(b) := \{f \mid \exists t > 0 : A[f](-t,0) > b + ct\},$$

see for instance [1, 11]; here it is used that the steady-state buffer content Q_n can be written as

$$\sup_{t>0} \sum_{i=1}^{n} A_i(-t,0) - nct.$$

To find the corresponding decay rate, 'Schilder' implies that $I(f)$ needs to be minimized over all $f \in F^{(\mathrm{f})}(b)$. Addie, Mannersalo, and Norros [1] provide this decay rate, for the situation of a general variance function $v(\cdot)$:

$$K^{(\mathrm{f})}(b) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Q_n \geq nb) = -\inf_{f\in F^{(\mathrm{f})}(b)} I(f) = -\inf_{t\geq 0} \frac{(b + (c-\mu)t)^2}{2v(t)}. \tag{4}$$

The following lemma evaluates $K^{(\mathrm{f})}(b)$ in (4) for the specific situations of BM and ALV input.

Lemma 2.9 *Consider a single FIFO queue with link rate nc, fed by n i.i.d. sources.*

(i) *If the sources are* $\mathrm{BM}(\lambda, \mu)$*, then, for* $b \geq 0$*,*

$$K^{(\mathrm{f})}(b) = 2 \cdot \frac{c-\mu}{\lambda} \cdot b.$$

(ii) *If the sources are* $\mathrm{ALV}(\kappa, \lambda, \mu)$*, then*

$$\left(K^{(\mathrm{f})}(b) - 2 \cdot \frac{c-\mu}{\lambda} \cdot b \right) \longrightarrow -2\kappa \left(\frac{c-\mu}{\lambda} \right)^2, \quad b \to \infty.$$

Proof. Part (i) directly follows from computing the infimum over $t \geq 0$ in (4). Part (ii) is a consequence of [5, Thm. 3]. This theorem states the following. Let $\theta^\star$ denote the unique positive solution to

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{E} \exp(\theta \bar{A}(t)) = (c-\mu)\theta.$$

Furthermore, suppose that $-\lim_{t\to\infty} \log \mathbb{E} \exp(\theta^\star \bar{A}(t) - \theta^\star(c-\mu)t) =: \nu$ exists. Then it holds that the shape function is asymptotically linear: $K^{(\mathrm{f})}(b) - \theta^\star b \to \nu$, for $b \to \infty$.
It is not hard to check that, in our Gaussian setting, for $\mathrm{ALV}(\kappa, \lambda, \mu)$ sources,

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{E} \exp(\theta \bar{A}(t)) = \lim_{t\to\infty} \frac{1}{2t} \cdot \theta^2 v(t) = \frac{\lambda}{2} \cdot \theta^2,$$

yielding $\theta^\star = 2(c-\mu)/\lambda$. Also,

$$\begin{aligned} \lim_{t\to\infty} \log \mathbb{E} \exp(\theta^\star \bar{A}(t) - \theta^\star(c-\mu)t) &= \lim_{t\to\infty} \frac{1}{2}(\theta^\star)^2 v(t) - \theta^\star(c-\mu)t \\ &= \lim_{t\to\infty} 2 \left(\frac{c-\mu}{\lambda} \right)^2 \cdot (v(t) - \lambda t) = 2\kappa \left(\frac{c-\mu}{\lambda} \right)^2. \end{aligned}$$

This proves the stated. □

In the following alternative proof of part (ii) we explicitly use the ALV properties of the sources. We include the proof, because several proofs in the sequel of the paper are along the same lines.

Alternative proof of part (ii) of Lemma 2.9. Use expression (4), and choose $\bar{\epsilon} > 0$ arbitrarily. It is clear that, invoking (A2) and the ALV characterization, for b large enough and arbitrary M,

$$\inf_{t \leq Mb} \frac{(b + (c-\mu)t)^2}{2v(t)} \geq \frac{b^2}{2v(Mb)} \geq \frac{b^2}{2M\lambda b + \bar{\epsilon}} = O\left(\frac{1}{2M\lambda} \cdot b \right).$$

Also

$$\frac{(b + (c-\mu)t)^2}{2v(t)} \Big|_{t:=b/(c-\mu)} = O\left(2 \cdot \frac{c-\mu}{\lambda} \cdot b \right).$$

Choosing M sufficiently small, this implies that we can restrict ourselves to the infimum over t in $[Mb, \infty)$. For any $\epsilon > 0$, we can select a b sufficiently large, such that for t in this range $|v(t) - \lambda t - \kappa| < \epsilon$. Hence

$$\inf_{t \geq Mb} \frac{(b + (c-\mu)t)^2}{2v(t)} \leq \inf_{t \geq Mb} \frac{(b + (c-\mu)t)^2}{2(\lambda t + \kappa - \epsilon)}.$$

The latter optimum equals

$$2 \cdot \frac{c-\mu}{\lambda} \cdot b - 2(\kappa - \epsilon) \left(\frac{c-\mu}{\lambda} \right)^2, \quad \text{achieved for } t = \frac{b}{c-\mu} - 2 \cdot \frac{\kappa - \epsilon}{\lambda}.$$

The lower bound follows after $\epsilon \downarrow 0$ and $b \to \infty$. The upper bound is analogous, with ϵ replaced by $-\epsilon$. □

Remark 2.10 In principle, Schilder's theorem gives upper bounds for closed sets, and lower bounds for open sets. Hence, to find an explicit expression for the decay rate of some specific set, one has to show that the set under consideration, say A, is an *I-continuity set*: $\inf_{f \in A^o} I(f) = \inf_{f \in \bar{A}} I(f)$, with A^o and $\bar{A}$ the interior and closure of A, respectively; see [18]. The verification of our overflow set being I-continuous is usually straightforward but tedious. We refer to [11] for the proof of I-continuity of the set of paths corresponding to overflow in the second queue of a tandem system; the priority queue and the GPS queue can be dealt with analogously. ◇

In concrete terms, the goal of the paper is to find the counterparts of Lemma 2.9 for the 'complex buffer architectures' tandem, priority, and GPS. Our analysis of the next sections shows that this is possible, by using the bounds derived in [11, 12].

3 Buffer overflow asymptotics in the tandem queue

Consider a tandem system of queues, fed by n Gaussian sources; the output of the first queue feeds into the second queue. The queues have link speeds nc_1 and nc_2, respectively; to avoid a trivial system, we assume $c_2 < c_1$. The system is stable: a source's mean rate μ is smaller than c_2. In this section we analyze the probability that the stationary buffer content of the second queue, $Q_{2,n}$, exceeds level nb.
From [11, Lemma 2.4], we have that

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(Q_{2,n} > nb) = -K^{(\mathrm{t})}(b), \quad \text{where} \quad K^{(\mathrm{t})}(b) := \inf_{f \in F^{(\mathrm{t})}(b)} I(f);$$

the 'overflow set' $F^{(\mathrm{t})}(b)$ is defined as

$$F^{(\mathrm{t})}(b) := \{f \mid \exists t > t_0 : \forall s \in (0,t) : A[f](-t,-s) > b + c_2 t - c_1 s\},$$

with $t_0 := db$ being the 'minimal' time till overflow, starting from an empty system; $d := (c_1 - c_2)^{-1}$. The main results of this section are Theorems 3.3 and 3.4 – the former treats the BM case, whereas the latter focuses ALV input. We start by providing two lemmas, Lemmas 3.1 and 3.2; Lemma 3.1 is valid for any type of Gaussian inputs.

Lemma 3.1 *For any $b \geq 0$,*

$$K^{(\mathrm{t})}(b) \geq \inf_{t>t_0} \frac{(b + (c_2 - \mu)t)^2}{2v(t)}.$$

Proof. Notice that evidently

$$F^{(\mathrm{t})}(b) \subseteq \{f \mid \exists t > t_0 : A[f](-t, 0) > b + c_2 t\}.$$

This implies the stated immediately. □

Lemma 3.2 *Suppose the sources are* BM(λ, μ)*. If $c_1 \geq 2c_2 - \mu$, then*

$$K^{(\mathrm{t})}(b) \leq 2 \cdot \frac{c_2 - \mu}{\lambda} \cdot b.$$

If $c_1 \in (c_2, 2c_2 - \mu)$, then

$$K^{(\mathrm{t})}(b) \leq \frac{1}{2} \cdot \frac{(c_1 - \mu)^2}{(c_1 - c_2)\lambda} \cdot b.$$

Proof. (i) Suppose $c_1 \geq 2c_2 - \mu$. Then the path that generates traffic at rate $2(c_2 - \mu)$, between $-b/(c_2 - \mu)$ and 0, is in $F^{(\mathrm{t})}(b)$. (ii) Suppose $c_1 \in (c_2, 2c_2 - \mu)$. Then the path that generates traffic at rate c_1, between $-db$ and 0, is in $F^{(\mathrm{t})}(b)$. Theorem 2.6 ('Schilder') implies that the norm of any feasible path constitutes an upper bound on the decay rate. Now applying Remark 2.8, the stated follows immediately. □

The following theorems give our main results for short-range dependent traffic. It states that its precise shape depends on whether $c_1 \geq 2c_2 - \mu$, or not.

Theorem 3.3 *Suppose the sources are* BM(λ, μ)*. If $c_1 \geq 2c_2 - \mu$, then*

$$K^{(\mathrm{t})}(b) = 2 \cdot \frac{c_2 - \mu}{\lambda} \cdot b.$$

If $c_1 \in (c_2, 2c_2 - \mu)$, then

$$K^{(\mathrm{t})}(b) = \frac{1}{2} \cdot \frac{(c_1 - \mu)^2}{(c_1 - c_2)\lambda} \cdot b.$$

Proof. The result is a direct application of Lemmas 3.1 and 3.2. □

We now concentrate on the situation in which the tandem queue is fed by ALV sources. We get the following counterpart of Lemma 2.9.(ii).

Theorem 3.4 *Suppose the sources are* $\mathrm{ALV}(\kappa,\lambda,\mu)$. *If* $c_1 \geq 2c_2-\mu$, *then*

$$\left(K^{(\mathrm{t})}(b) - 2\cdot\frac{c_2-\mu}{\lambda}\cdot b\right) \longrightarrow -2\kappa\left(\frac{c_2-\mu}{\lambda}\right)^2 \quad \textit{as } b\to\infty.$$

If $c_1 \in (c_2, 2c_2-\mu)$, *then*

$$\left(K^{(\mathrm{t})}(b) - \frac{1}{2}\cdot\frac{(c_1-\mu)^2}{(c_1-c_2)\lambda}\cdot b\right) \longrightarrow -\frac{\kappa}{2}\left(\frac{c_1-\mu}{\lambda}\right)^2 \quad \textit{as } b\to\infty.$$

Proof. Our proof consists of a lower bound and an upper bound. The lower bound follows directly from Lemma 3.1, whereas the upper bound is a matter of finding the norm of a feasible path.

Lower bound. Choose $\epsilon>0$ arbitrarily. Take b sufficiently large, such that for all t larger than db we have that $|v(t)-\lambda t-\kappa|\leq\epsilon$. Applying the generic lower bound of Lemma 3.1, we find

$$K^{(\mathrm{t})}(b) \geq \inf_{t\geq db}\frac{(b+(c_2-\mu)t)^2}{2v(t)} \geq \inf_{t\geq db}\frac{(b+(c_2-\mu)t)^2}{2(\lambda t+\kappa+\epsilon)}.$$

If the latter infimum would be over all positive t – rather than all t larger than $t_0=db$ – it would be attained at

$$t^\star(b) := \frac{b}{c_2-\mu} - 2\left(\frac{\kappa+\epsilon}{\lambda}\right);$$

it is also observed that, for $b\to\infty$, the condition $t^\star(b)>db$ reads $c_1>2c_2-\mu$.
From the above we conclude that for case $t^\star(b)>db$ the lower bound

$$K^{(\mathrm{t})}(b) \geq \frac{(b+(c_2-\mu)t^\star(b))^2}{2(\lambda t^\star(b)+\kappa+\epsilon)} = 2\cdot\frac{c_2-\mu}{\lambda}\cdot\left(b-(c_2-\mu)\cdot\frac{\kappa+\epsilon}{\lambda}\right)$$

applies, whereas for $t^\star(b)<db$ we have

$$\begin{aligned} K^{(\mathrm{t})}(b) &\geq \frac{(b+(c_2-\mu)db)^2}{2(\lambda db+\kappa+\epsilon)} \\ &= \frac{b^2(1+(c_2-\mu)d)^2}{2(\lambda db+\kappa+\epsilon)} = \frac{1}{2}\cdot\frac{(c_1-\mu)^2}{(c_1-c_2)\lambda}\cdot b - \frac{\kappa+\epsilon}{2}\left(\frac{c_1-\mu}{\lambda}\right)^2 + O\left(\frac{1}{b}\right). \end{aligned}$$

Now let $\epsilon\downarrow 0$ and $b\to\infty$, and we get the desired lower bound.

Upper bound. Choose $\epsilon>0$ arbitrarily. Define

$$s^\star_\epsilon(b) = \max\left\{d_\epsilon b,\ \frac{b}{c_2-\mu}-\frac{2\kappa}{\lambda}\right\},\quad \text{with}\quad d_\epsilon := \frac{1+\epsilon}{(c_1-\mu)-(1+\epsilon)(c_2-\mu)}. \tag{5}$$

Observe that, in particular,

$$(1+\epsilon)(b+(c_2-\mu)s^\star_\epsilon(b)) \leq (c_1-\mu)s^\star_\epsilon(b). \tag{6}$$

Define the path

$$f(r) := -\frac{\Gamma(-s_\epsilon^\star(b), r)}{v(s_\epsilon^\star(b))} \cdot (b + (c_2 - \mu)s_\epsilon^\star(b)) - \mu r.$$

First we show that this path is feasible (for b large), i.e., $f \in F^{(\mathrm{t})}(b)$. Obviously, we have that $A[f](-s_\epsilon^\star(b), 0) = b + c_2 s_\epsilon^\star(b)$. Left to prove is that $A[f](-s, 0) < c_1 s$ for all $s \in (0, s_\epsilon^\star(b))$. With straightforward calculations and applying (1), it turns out that this is equivalent to requiring, for all $\gamma \in (0, 1)$,

$$\left(1 + \frac{v(\gamma s_\epsilon^\star(b))}{v(s_\epsilon^\star(b))} - \frac{v((1-\gamma)s_\epsilon^\star(b))}{v(s_\epsilon^\star(b))}\right)(b + (c_2 - \mu)s_\epsilon^\star(b)) < 2(c_1 - \mu)\gamma s_\epsilon^\star(b). \tag{7}$$

Fixing $\gamma \in (0, 1)$, the definition of ALV (in conjunction with $s_\epsilon^\star(b) \to \infty$ as $b \to \infty$) implies that there exists a $b_0(\gamma) \geq 0$, such that for any $b \geq b_0(\gamma)$, we have that

$$\left(1 + \frac{v(\gamma s_\epsilon^\star(b))}{v(s_\epsilon^\star(b))} - \frac{v((1-\gamma)s_\epsilon^\star(b))}{v(s_\epsilon^\star(b))}\right) \leq 2(1 + \epsilon)\gamma.$$

Observe that $b_0(0)$ and $b_0(1)$ could be chosen finite numbers, and that we can choose a function $b_0(\cdot)$ that is continuous on $[0, 1]$, cf. (A1). But then requirement (7) is true for all $b \geq \max_{\gamma \in [0,1]} b_0(\gamma)$ (which is finite, as any continuous function attains a maximum on a finite interval), due to (6). We thus conclude that our path is feasible.

Recall that, because of Theorem 2.6, the norm of any feasible path constitutes an upper bound on the decay rate. We now compute (an upper bound to) the norm of our path $f(\cdot) \in F^{(\mathrm{t})}(b)$. Take $\delta > 0$ arbitrarily. Choose b sufficiently large that $|v(s_\epsilon^\star(b)) - \lambda s_\epsilon^\star(b) - \kappa| \leq \delta$ (which is possible due to the definition of ALV). Because of the feasibility of the path, and applying Lemma 2.7, we obviously have

$$K^{(\mathrm{t})}(b) \leq \frac{(b + (c_2 - \mu)s_\epsilon^\star(b))^2}{2v(s_\epsilon^\star(b))} \leq \frac{(b + (c_2 - \mu)s_\epsilon^\star(b))^2}{2(\lambda s_\epsilon^\star(b) + \kappa - \delta)} \tag{8}$$

For b large, it is clear that for $c_1 - \mu > 2(c_2 - \mu)(1 + \epsilon)$ the maximum in (5) is attained by the first argument between the brackets, i.e., $d_\epsilon b$. In this case the upper bound (8) becomes

$$K^{(\mathrm{t})}(b) \leq \frac{1}{2} \cdot \frac{(c_1 - \mu)^2}{((c_1 - \mu) - (1 + \epsilon)(c_2 - \mu))\lambda} \cdot b - \frac{\kappa - \delta}{2}\left(\frac{c_1 - \mu}{\lambda(1 + \epsilon)}\right)^2 + O\left(\frac{1}{b}\right).$$

On the other hand, for $c_1 - \mu \leq 2(c_2 - \mu)(1 + \epsilon)$,

$$K^{(\mathrm{t})}(b) \leq 2 \cdot \frac{c_2 - \mu}{\lambda} \cdot \left(b - (c_2 - \mu) \cdot \frac{\kappa - \delta}{\lambda}\right).$$

Now let $\epsilon, \delta \downarrow 0$ and $b \to \infty$, and we have established the stated. □

4 Priority queue

In a priority queue, a link of capacity nc is considered, fed by traffic of two classes, each with its own queue. Traffic of class 1 does not 'see' class 2 at all, and consequently we know how the *high-priority queue* $Q_{h,n}$ behaves, see Lemma 2.9. A more challenging task is the characterization of overflow in the low priority queue.

We let the system be fed by n i.i.d. high-priority (hp) sources, and an equal number of i.i.d. low-priority (lp) sources; both classes are independent. We assume that both hp and lp sources are Gaussian, with mean rates by μ_h and μ_ℓ, and variance functions by $v_h(\cdot)$ and $v_\ell(\cdot)$, respectively; also $\mu := \mu_h + \mu_\ell$ and $v(\cdot) := v_h(\cdot) + v_\ell(\cdot)$. (Notice that this setting also covers the case that the number of sources of both classes are *not* equal. Assume for instance that there are $n\alpha$ lp sources. Multiplying μ_ℓ and $v_\ell(\cdot)$ by α and applying the fact that the Normal distribution is infinitely divisible, we arrive at n i.i.d. sources.) We obviously assume $\mu < c$.

We use the 'two-dimensional Schilder' framework, as described in [15]. There a large deviations rate function $I(\cdot)$ is introduced, with two-dimensional argument $f(\cdot) \equiv (f_h(\cdot), f_\ell(\cdot))$. Let $A_h[f_h]$ and $A_\ell[f_\ell]$ be defined similarly as before.

Our analysis relies on the following expression for the decay rate of overflow in the lp queue.

Lemma 4.1 *The decay rate of overflow in the lp queue is given by*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Q_{\ell,n} > nb) = -K^{(\mathrm{p})}(b), \quad \textit{where} \quad K^{(\mathrm{p})}(b) := \inf_{f \in F^{(\mathrm{p})}(b)} I(f).$$

Here the 'overflow set' is defined as

$$F^{(\mathrm{p})}(b) := \left\{ f \;\middle|\; \exists t, x > 0 : \forall s \in (0,\infty) : \begin{array}{l} A_h[f_h](-t,0) + A_\ell[f_\ell](-t,0) > b + ct + x, \\ A_h[f_h](-s,0) \le cs + x \end{array} \right\}.$$

Proof. To make sure that the lp queue exceeds nb, there must be an $x > 0$ such that the total queue exceeds $nb + nx$, whereas the hp queue remains below nx. Now the stated follows from

$$Q_{h,n} + Q_{\ell,n} = \sup_{t>0} \left(\sum_{i=1}^{n} (A_{h,i}(-t,0) + A_{\ell,i}(-t,0)) - nct \right);$$

$$Q_{h,n} = \sup_{s>0} \left(\sum_{i=1}^{n} A_{h,i}(-s,0) - ncs \right),$$

where $A_{h,i}(\cdot)$ ($A_{\ell,i}(\cdot)$, respectively) corresponds to the traffic stream generated by the ith hp (lp) source. □

Notice that the 'overflow set' of the above lemma is slightly different from the one we used in [11, Section 5]. The next lemma applies the above characterization to derive a lower bound on the decay rate.

Lemma 4.2 *Let* $S^{(\mathrm{p})}(b,t,x) := \{(a_h, a_\ell) \mid a_h + a_\ell > b + ct + x, a_h \le ct + x\}$. *Then*

$$K^{(\mathrm{p})}(b) \ge \inf_{t,x>0} k^{(\mathrm{p})}(b,t,x),$$

with

$$k^{(\mathrm{p})}(b,t,x) := \frac{1}{2} \inf_{(a_h,a_\ell)\in S^{(\mathrm{p})}(b,x,t)} \left(\frac{(a_h - \mu_h t)^2}{v_h(t)} + \frac{(a_\ell - \mu_\ell t)^2}{v_\ell(t)} \right).$$

Also, if $(b - \mu_\ell t)v_h(t) \le ((c - \mu_h)t + x)v_\ell(t)$,

$$k^{(\mathrm{p})}(b,t,x) = k_1^{(\mathrm{p})}(b,t,x) := \frac{1}{2} \cdot \frac{(b + (c-\mu)t + x)^2}{v(t)},$$

whereas otherwise

$$k^{(\mathrm{p})}(b,t,x) = k_2^{(\mathrm{p})}(b,t,x) := \frac{1}{2} \cdot \left(\frac{((c - \mu_h)t + x)^2}{v_h(t)} + \frac{(b - \mu_\ell t)^2}{v_\ell(t)} \right).$$

Proof. First it is noted that the following trivial inclusion holds:

$$F^{(\mathrm{p})}(b) \subseteq \left\{ f \;\middle|\; \exists t, x > 0 : \begin{array}{l} A_h[f_h](-t,0) + A_\ell[f_\ell](-t,0) > b + ct + x, \\ A_h[f_h](-t,0) \le ct + x \end{array} \right\}.$$

This immediately yields the lower bound $\inf_{x,t>0} k^{(\mathrm{p})}(b,t,x)$.
The explicit expression for $k^{(\mathrm{p})}(b,t,x)$ is derived as follows. Applying Lemma 2.5, it is readily derived that the first constraint in $S^{(\mathrm{p})}(b,t,x)$, i.e., $a_h + a_\ell > b + ct + x$ is always tight. The second is tight if

$$\mu_h t + \frac{v_h(t)}{v(t)}(b + (c-\mu)t + x) > ct + x.$$

If it is tight, the infimum is achieved at $(a_h, a_\ell) = (ct + x, b)$, otherwise at

$$(a_h, a_\ell) = \left(\mu_h t + \frac{v_h(t)}{v(t)}(b + (c-\mu)t + x), \mu_\ell t + \frac{v_\ell(t)}{v(t)}(b + (c-\mu)t + x) \right).$$

Direct calculations yield the stated. □

Introduce the following notation: $\mu := \mu_h + \mu_\ell$, $\kappa := \kappa_h + \kappa_\ell$, and $\lambda := \lambda_h + \lambda_\ell$.

Lemma 4.3 *Suppose that the high-priority sources are* $\mathrm{BM}(\lambda_h, \mu_h)$ *and that the low-priority sources are* $\mathrm{BM}(\lambda_\ell, \mu_\ell)$. *Then*

$$\inf_{t,x>0} k_1^{(\mathrm{p})}(b,t,x) = 2 \cdot \frac{c-\mu}{\lambda} \cdot b;$$

$$\inf_{t,x>0} k_2^{(\mathrm{p})}(b,t,x) = \frac{\xi - \mu_\ell}{\lambda_\ell} \cdot b, \;\; \textit{where } \xi \equiv \xi(\lambda_h, \mu_h, \lambda_\ell, \mu_\ell) := \sqrt{\mu_\ell^2 + \frac{\lambda_\ell}{\lambda_h}(c - \mu_h)^2}.$$

Proof. This is a matter of standard computations. In the first minimization, it turns out that $x_1^\star = 0$, $t_1^\star = b/(c-\mu)$, whereas in the second

$$x_2^\star = 0, \quad t_2^\star = \sqrt{\frac{\lambda_h}{\lambda_h \mu_\ell^2 + \lambda_\ell (c-\mu_h)^2}} \cdot b.$$

The desired is obtained by inserting these into the objective function. □

Theorem 4.4 *Suppose that the high-priority sources are* $\mathrm{BM}(\lambda_h, \mu_h)$ *and that the low-priority sources are* $\mathrm{BM}(\lambda_\ell, \mu_\ell)$. *If* $\lambda_h(c-\mu_h-2\mu_\ell) \le \lambda_\ell(c-\mu_h)$, *then*

$$K^{(\mathrm{p})}(b) = 2 \cdot \frac{c-\mu}{\lambda} \cdot b.$$

If $\lambda_h(c-\mu_h-2\mu_\ell) > \lambda_\ell(c-\mu_h)$, *then*

$$K^{(\mathrm{p})}(b) = \frac{\xi - \mu_\ell}{\lambda_\ell} \cdot b.$$

Proof. *Lower bound.* We use Lemma 4.2. Define three sets:

$$T_1 := \{(t,x) \in \mathbb{R}_+^2 \mid (b-\mu_\ell t) v_h(t) \le ((c-\mu_h)t + x) v_\ell(t)\};$$

T_2 with the '$\le$'-sign replaced by '$\ge$', and $\bar{T}$ with the '$\le$'-sign replaced by '='. Notice that $k_1^{(\mathrm{p})}(b,\cdot,\cdot)$ and $k_2^{(\mathrm{p})}(b,\cdot,\cdot)$ coincide for (t,x) in $\bar{T}$. Let $t_i^\star, x_i^\star$ $(i=1,2)$ be defined as in the proof of Lemma 4.3.
First consider the infimum of $k^{(\mathrm{p})}(b,t,x)$ over $(t,x) \in T_1$. Clearly the optimum is in $T_1 \setminus \bar{T}$ iff $(b-\mu_\ell t_1^\star) v_h(t_1^\star) < ((c-\mu_h)t_1^\star + x_1^\star) v_\ell(t_1^\star)$, or, equivalently,

$$\lambda_h(c-\mu_h-2\mu_\ell) < \lambda_\ell(c-\mu_h); \tag{9}$$

otherwise the optimum over T_1 is attained in $\bar{T}$.
Then consider the infimum of $k^{(\mathrm{p})}(b,t,x)$ over $(t,x) \in T_2$. Now the optimum is in $T_2 \setminus \bar{T}$ iff

$$(b-\mu_\ell t_2^\star) v_h(t_2^\star) > ((c-\mu_h)t_2^\star + x_2^\star) v_\ell(t_2^\star),$$

and otherwise at the boundary $\bar{T}$. More tedious calculations yields that this condition is equivalent to

$$\lambda_h(c-\mu_h-2\mu_\ell) > \lambda_\ell(c-\mu_h). \tag{10}$$

Because both conditions (9) and (10) are mutually exclusive, this proves the lower bound.

Upper bound. The upper bound is just a matter of computing the norms of paths in $F^{(\mathrm{p})}(b)$, just as in the tandem case. □

Remark 4.5 It is straightforward, but tedious, to check that both expressions for $K^{(\mathrm{p})}(b)$ (from Theorem 4.4) coincide if $\lambda_h(c-\mu_h-2\mu_\ell)=\lambda_\ell(c-\mu_h)$. During these computations, we also find that then $\xi(\lambda_h-\lambda_\ell)=\mu_\ell\lambda$. $\diamond$

We now turn to the situation of ALV sources.

Lemma 4.6 *Suppose that the high-priority sources are* $\mathrm{ALV}(\kappa_h,\lambda_h,\mu_h)$ *and that the low-priority sources are* $\mathrm{ALV}(\kappa_\ell,\lambda_\ell,\mu_\ell)$. *Then*

$$\lim_{b\to\infty}\left(\inf_{x,t>0} k_1^{(\mathrm{p})}(b,t,x)-2\cdot\frac{c-\mu}{\lambda}\cdot b\right)=-2\kappa\left(\frac{c-\mu}{\lambda}\right)^2.$$

Proof. This is equivalent to Lemma 2.9. $\square$

Lemma 4.7 *Suppose that the high-priority sources are* $\mathrm{ALV}(\kappa_h,\lambda_h,\mu_h)$ *and that the low-priority sources are* $\mathrm{ALV}(\kappa_\ell,\lambda_\ell,\mu_\ell)$. *Then*

$$\lim_{b\to\infty}\left(\inf_{x,t>0} k_2^{(\mathrm{p})}(b,t,x)-\frac{\xi-\mu_\ell}{\lambda_\ell}\cdot b\right)=-\frac{(c-\mu_h)^2}{\lambda_h}\left(\frac{\kappa_h}{2\lambda_h}-\Phi\right)+\Phi^2\xi\cdot\frac{\mu_\ell}{2\kappa_\ell},$$

with

$$\Phi:=-\frac{\kappa_\ell}{\lambda_\ell}\left(1-\frac{\mu_\ell}{\xi}\right).$$

Proof. First observe that, for any given value of b, t, the infimum of $k_2^{(\mathrm{p})}(b,t,x)$ over x is attained in 0. Therefore consider $\inf_{t\geq 0} k_2^{(\mathrm{p})}(b,t,0)$. We borrow the argument of the alternative proof of part (ii) of Lemma 2.9.

- We first show that we can restrict ourselves to $t\geq Mb$. Choosing $M<\mu_\ell^{-1}$, and $\bar\epsilon>0$ arbitrarily, and using (A2) and the ALV properties, we notice that for b large

$$\inf_{t<Mb} k_2^{(\mathrm{p})}(b,t,0)\geq\frac{(1-\mu_\ell M)^2}{2}\cdot\frac{b^2}{M\lambda_\ell b+\kappa_\ell+\bar\epsilon}=O\left(\frac{(1-\mu_\ell M)^2}{2M\lambda_\ell}\cdot b\right). \tag{11}$$

  Also, it is not hard to verify that for $b\to\infty$

$$k_2^{(\mathrm{p})}\left(b,b\cdot\sqrt{\frac{\lambda_h}{\lambda_\ell(c-\mu_h)^2+\lambda_h\mu_\ell^2}},0\right)=O\left(\frac{\xi-\mu_\ell}{\lambda_\ell}\cdot b\right). \tag{12}$$

  Choosing $M>0$ sufficiently small, (11) majorizes (12), and hence we can restrict ourselves in $\inf_t k_2^{(\mathrm{p})}(b,t,0)$ to $t\geq Mb$.

- Choose b large enough, such that, for all $t\geq Mb$, both $|v_h(t)-\lambda_h t-\kappa_h|<\epsilon$ and $|v_\ell(t)-\lambda_\ell t-\kappa_\ell|<\epsilon$. Then applying Lemma A.1 and letting $\epsilon\downarrow 0$ and $b\to\infty$ yields the stated. $\square$

Theorem 4.8 *If* $\lambda_h(c-\mu_h-2\mu_\ell) \leq \lambda_\ell(c-\mu_h)$*, then*

$$\lim_{b\to\infty}\left(K^{(\mathrm{p})}(b) - 2\cdot\frac{c-\mu}{\lambda}\cdot b\right) = -2\kappa\left(\frac{c-\mu}{\lambda}\right)^2.$$

If $\lambda_h(c-\mu_h-2\mu_\ell) > \lambda_\ell(c-\mu_h)$*, then*

$$\lim_{b\to\infty}\left(K^{(\mathrm{p})}(b) - \frac{\xi-\mu_\ell}{\lambda_\ell}\cdot b\right) = -\frac{(c-\mu_h)^2}{\lambda_h}\left(\frac{\kappa_h}{2\lambda_h}-\Phi\right) + \Phi^2\xi\cdot\frac{\mu_\ell}{2\kappa_\ell}.$$

Proof. The lower bound is analogous to the lower bound in Theorem 4.4, but with Lemmas 4.6 and 4.7 replacing Lemma 4.3. The upper bound is analogous to the upper bound in the tandem case, i.e., in the proof of Theorem 3.4. □

Waiting time asymptotics

The first part of this section was devoted to buffer overflow in the lp queue, or, more precisely, the probability that the buffer content of the lp queue exceeds some predefined level. We now focus on the probability of a long delay in the lp queue. To this end, following Norros [19], consider the notion of *virtual* waiting time. The random variable $V_{\ell,n}(0)$ is defined as the time it takes to transmit a 'fluid molecule' that enters the lp queue at time 0. We define

$$L^{(\mathrm{p})}(b) := -\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}(V_{\ell,n}(0) > b).$$

Lemma 4.9

$$L^{(\mathrm{p})}(b) \geq \frac{1}{2}\sup_{u\in(0,b)}\inf_{s>0}\frac{((c-\mu_h-\mu_\ell)s+(c-\mu_h)u)^2}{v_h(s+u)+v_\ell(s)}.$$

Proof. The set of paths such that the virtual delay exceeds b equals

$$G^{(\mathrm{p})}(b) := \{f \mid \forall u\in(0,b) : \exists s>0 : A_h[f_h](-s,u) + A_\ell[f_\ell](-s,0) > c(u+s)\},$$

see [19, Section 4]. The stated follows immediately by applying Lemma 2.7. □

Theorem 4.10 *Suppose that the high-priority sources are* $\mathrm{BM}(\lambda_h,\mu_h)$*, and that the low-priority sources are* $\mathrm{BM}(\lambda_\ell,\mu_\ell)$*. Then*

$$L^{(\mathrm{p})}(b) = 2\left(\frac{c-\mu_h}{\lambda}\right)\left(\frac{\lambda_h\mu_\ell+\lambda_\ell(c-\mu_h)}{\lambda}\right)b.$$

Suppose that the high-priority sources are $\mathrm{ALV}(\kappa_h,\lambda_h,\mu_h)$*, and that the low-priority sources are* $\mathrm{ALV}(\kappa_\ell,\lambda_\ell,\mu_\ell)$*. Then*

$$\left(L^{(\mathrm{p})}(b) - 2\left(\frac{c-\mu_h}{\lambda}\right)\left(\frac{\lambda_h\mu_\ell+\lambda_\ell(c-\mu_h)}{\lambda}\right)b\right) \longrightarrow -2\kappa\left(\frac{c-\mu}{\lambda}\right)^2.$$

Proof. The lower bounds follow directly from Lemma 4.9, as follows. For BM traffic the optimizing $s^\star(u)$, for given u, equals

$$s^\star(u) = \left(\frac{c - \mu_h}{c - \mu_h - \mu_\ell} - 2 \cdot \frac{\lambda_h}{\lambda} \right) \cdot u.$$

The resulting function of u increases on $[0, b]$, so the optimum is attained at $u^\star = b$. For ALV traffic we get

$$s^\star(u) - \left(\frac{c - \mu_h}{c - \mu_h - \mu_\ell} - 2 \cdot \frac{\lambda_h}{\lambda} \right) \cdot u \to -2 \cdot \frac{\kappa}{\lambda},$$

also leading to $u^\star = b$. Now straightforward algebra (use Lemma A.1!) gives the lower bound. The upper bound is a matter of computing the norms of feasible paths, as before. Unlike the *overflow* asymptotics of the tandem and priority queue, there is just a single regime, which makes the analysis somewhat simpler. □

Remark 4.11 We remark that in case of BM the most likely path is such that (i) between time epochs $-s^\star(b)$ and b any hp source generates traffic at a constant rate $c - \mu_\ell$, whereas (ii) between $-s^\star(b)$ and 0 any lp source transmits at rate $c - \mu_h$. Notice that, due to the stability constraint $\mu_h + \mu_\ell < c$, these rates exceed the mean rates of the sources, i.e., μ_h and μ_ℓ, respectively. Outside the intervals indicated above the sources obey their mean rates. ◇

5 Generalized processor sharing

In this section we consider a system where traffic is served according to a generalized processor sharing (GPS) mechanism, consisting of two queues sharing a link of capacity nc. We assume the system to be fed by traffic from two classes, where class i uses queue i, for $i = 1, 2$. It is assumed that both classes consist of n flows (but, again, due to the infinite divisibility of the normal distribution, this is not a restriction).

A weight $\phi_i \geq 0$ is assigned to class i and, without loss of generality, assume that these add up to 1, i.e., $\phi_1 + \phi_2 = 1$. The GPS mechanism then works as follows. Class i receives service at rate $n\phi_i c$ when both classes are backlogged. Because class i gets at least service at rate $n\phi_i c$ when it has backlog, we will refer to it as the *guaranteed rate* of class i. If one of the classes has no backlog and is transmitting at a rate less than or equal to its guaranteed rate, then this class is served at its transmission rate, while the other class receives the remaining service capacity. If both classes are sending at rates less than their guaranteed rates, then they are both served at their sending rate, and some service capacity is left unused. We assume that the buffer sizes of both queues are infinitely large.

Without loss of generality, we focus on the workload of the first queue. The goal here is to analyze the decay rate of the probability that the stationary workload exceeds a threshold nb.

Hence, denoting by $Q_{i,n} \equiv Q_{i,n}(0)$ the stationary workload in the ith GPS queue at time 0, the probability of our interest is $\mathbb{P}(Q_{1,n} \geq nb)$.
In [12, Lemma 3.2] it is shown that the 'overflow set' is bounded from above by

$$\overline{F^{(\mathrm{g})}}(b) := \left\{ f \;\middle|\; \begin{array}{l} \exists t, x > 0 : \forall s \in (0,t) : \exists u \in (0,s) : \\ \quad A_1[f_1](-t,0) + A_2[f_2](-t,0) > b + ct + x, \\ \quad A_1[f_1](-s,-u) + A_2[f_2](-s,0) \leq cs - \phi_1 cu + x \end{array} \right\},$$

we have that

$$K^{(\mathrm{g})}(b) := -\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Q_{1,n} \geq nb) \geq \inf_{f \in \overline{F^{(\mathrm{g})}}(b)} I(f).$$

In this paper we concentrate on BM input – the case of ALV is harder, and we comment on it later (Section 6). The input of class i consists of n sources of the type $\mathrm{BM}(\lambda_i, \mu_i)$. Define also $\mu := \mu_1 + \mu_2$ and $\lambda := \lambda_1 + \lambda_2$, and the 'reduced rates': $\bar{c} := c - \mu$ and $\bar{c}_i = c\phi_i - \mu_i$ (for $i = 1, 2$). The system is stable: $\mu < c$.
We start by presenting an introductory lemma.

Lemma 5.1 *Suppose that the class-1 sources are* $\mathrm{BM}(\lambda_1, \mu_1)$ *and that the class-2 sources are* $\mathrm{BM}(\lambda_2, \mu_2)$. *Let* $S^{(\mathrm{g})}(b,x,u,t) := \{(a_1,a_2) \mid a_1 > b + ct + x, a_2 \leq ct - c\phi_1 u + x\}$. *Then*

$$K^{(\mathrm{g})}(b) \geq \inf_{x,t>0, u\in(0,t)} k^{(\mathrm{g})}(b,x,u,t),$$

where $k^{(\mathrm{g})}(b,x,u,t) := \frac{1}{2} \inf_{(a_1,a_2)\in S^{(\mathrm{g})}(b,x,u,t)} H(a_1,a_2)$, *with*

$$H(a_1,a_2) := \left(\frac{(a_1 - \mu t)^2}{\lambda_1 u} - \frac{2(a_1 - \mu t)(a_2 - \mu t + \mu_1 u)}{\lambda_1 u} + \frac{(a_2 - \mu t + \mu_1 u)^2}{\lambda_1 u} \frac{\lambda t}{\lambda t - \lambda_1 u} \right).$$

Also, if $\lambda t(b + \bar{c}_1 u) \leq \lambda_1 u(b + \bar{c}t + x)$, *then*

$$k^{(\mathrm{g})}(b,x,u,t) = k_1^{(\mathrm{g})}(b,x,u,t) := \frac{1}{2} \cdot \frac{(b + \bar{c}t + x)^2}{\lambda t},$$

whereas otherwise

$$k^{(\mathrm{g})}(b,x,u,t) = k_2^{(\mathrm{g})}(b,x,u,t) := \frac{1}{2} \cdot \left(\frac{(\bar{c}t - \bar{c}_1 u + x)^2}{\lambda t - \lambda_1 u} + \frac{(b + \bar{c}_1 u)^2}{\lambda_1 u} \right).$$

Proof. First it is noted that

$$\overline{F^{(\mathrm{g})}}(b) \subseteq \left\{ f \;\middle|\; \begin{array}{l} \exists t, x > 0 : \exists u \in (0,t) : \\ \quad A_1[f_1](-t,0) + A_2[f_2](-t,0) > b + ct + x, \\ \quad A_1[f_1](-t,-u) + A_2[f_2](-t,0) \leq ct - c\phi_1 cu + x \end{array} \right\}.$$

Identifying a_1 with $A_1(-t,0)+A_2(-t,0)$, and a_2 with $A_1(-t,-u)+A_2(-t,0)$, the lower bound is derived as follows. Notice that these random variables have variances λt and $\lambda_1(t-u)+\lambda_2 t = \lambda t - \lambda_1 u$, respectively, while their covariance reads $\lambda t - \lambda_1 u$. Also

$$(a_1-\mu t, a_2-\mu t+\mu_1 u)^{\mathrm{T}} \begin{pmatrix} \lambda t & \lambda t-\lambda_1 u \\ \lambda t-\lambda_1 u & \lambda t-\lambda_1 u \end{pmatrix}^{-1} \begin{pmatrix} a_1-\mu t \\ a_2-\mu t+\mu_1 u \end{pmatrix}$$

equals $H(a_1,a_2)$. This proves the first part of the lemma. The evaluation of the maximum over $(a_1,a_2)\in S^{(\mathrm{g})}(b,x,u,t)$ is analogous to the priority case (use Lemma 2.5). □

The following lemma presents an explicit expression for the infimum of $k_1^{(\mathrm{g})}(b,t,u,x)$ (minimized over t,u,x). Its proof is standard, and we therefore omit the proof.

Lemma 5.2 *Suppose that the class-1 sources are* $\mathrm{BM}(\lambda_1,\mu_1)$ *and that the class-2 sources are* $\mathrm{BM}(\lambda_2,\mu_2)$. *Then*

$$\inf_{t,x>0,u\in(0,t)} k_1^{(\mathrm{g})}(b,t,u,x) = 2\cdot\frac{c-\mu}{\lambda}\cdot b.$$

The derivation of the infimum of $k_2^{(\mathrm{g})}(b,t,u,x)$ (minimized over t,u,x) is considerably harder. We first define

$$T(\gamma) := \bar{c}_1^2\gamma^2 + (\bar{c}-\bar{c}_1\gamma)^2\cdot\frac{\lambda_1\gamma}{\lambda-\lambda_1\gamma}; \quad t(\gamma) := \frac{1}{\sqrt{T(\gamma)}}.$$

Lemma 5.3 *Suppose that the class-1 sources are* $\mathrm{BM}(\lambda_1,\mu_1)$ *and that the class-2 sources are* $\mathrm{BM}(\lambda_2,\mu_2)$. *Then*

$$\inf_{t,x>0,u\in(0,t)} k_2^{(\mathrm{g})}(b,t,u,x) = \frac{1}{2}\inf_{\gamma\in(0,1)} U(\gamma)\cdot b,$$

with

$$U(\gamma) := \frac{(1+\bar{c}_1\gamma t(\gamma))^2}{\lambda_1\gamma t(\gamma)} + \frac{(\bar{c}-\bar{c}_1\gamma)^2 t(\gamma)}{\lambda-\lambda_1\gamma}.$$

Proof. First observe that, for any given value of b,t,u, the infimum over x is attained in 0. Therefore consider $k_2^{(\mathrm{g})}(b,t,u,0)$ to be optimized over positive t and $u\in(0,t)$. Now write $u=\gamma t$, with $\gamma\in(0,1)$. Perform the optimization over t. Straightforward calculus yields that the minimum is attained at $t=bt(\gamma)$. Inserting this yields the stated. □

The following two lemmas determine the infimum of $U(\gamma)$ over $\gamma\in(0,1)$.

Lemma 5.4 *Suppose*

$$\phi_1 \geq \phi_1^c := \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2} \cdot \left(1 - \frac{\mu}{c}\right) + \frac{\mu_1}{c}.$$

For all $\gamma \in (0,1)$*, it holds that* $\gamma t(\gamma) \leq t(1)$.

Proof. It is equivalent to check that $T(\gamma) \geq \gamma^2 T(1)$ for $\gamma \in (0,1)$, or

$$E_\ell(\gamma) := \left(\frac{\bar{c} - \bar{c}_1\gamma}{\bar{c}_2}\right)^2 \geq \frac{(\lambda - \lambda_1\gamma)\gamma}{\lambda_2} =: E_r(\gamma).$$

Due to the fact that $E_\ell(\cdot)$ and $E_r(\cdot)$ correspond to parabolas (where the former is convex and the latter is concave), it is enough to verify whether $E_\ell'(1) \leq E_r'(1)$. This yields the condition

$$-2 \cdot \frac{\bar{c}_1}{\bar{c}_2} \leq 1 - \frac{\lambda_1}{\lambda_2},$$

which is in turn equivalent to $\phi_1 \geq \phi_1^c$. □

Lemma 5.5 *If* $\phi_1 \geq \phi_1^c$*, then* $\inf_{\gamma\in(0,1)} U(\gamma) = U(1)$.

Proof. We need to check if, for all $\gamma \in (0,1)$,

$$\frac{(1 + \bar{c}_1\gamma t(\gamma))^2}{\lambda_1\gamma t(\gamma)} + \frac{(\bar{c} - \bar{c}_1\gamma)^2 t(\gamma)}{\lambda - \lambda_1\gamma} \geq \frac{(1 + \bar{c}_1 t(1))^2}{\lambda_1 t(1)} + \frac{\bar{c}_2^2 t(1)}{\lambda_2}.$$

Straightforward algebraic manipulations yield that this is equivalent to

$$\lambda_2(\lambda - \lambda_1\gamma)t(1) \;-\; (\lambda - \lambda_1\gamma)\gamma V_1(\gamma)t(\gamma) \;+\; \lambda_2 t(1) V_2(\gamma) t^2(\gamma) \geq 0, \quad \text{with}$$

$$V_1(\gamma) := \lambda_1\bar{c}_2^2 t^2(1) + \lambda_2(1 + \bar{c}_1^2 t^2(1)); \qquad V_2(\gamma) := (\lambda - \lambda_1\gamma)\bar{c}_1^2\gamma^2 + \lambda_1\gamma(\bar{c} - \bar{c}_1\gamma)^2.$$

Now it is not hard to see that $V_1(\gamma) = 2\lambda_2$ and $V_2(\gamma) = (\lambda - \lambda_1\gamma)/t^2(\gamma)$, such that it remain to verify that

$$2\lambda_2 \cdot (\lambda - \lambda_1\gamma) \cdot (t(1) - \gamma t(\gamma)) \geq 0,$$

but this holds (for $\phi_1 \geq \phi_1^c$) because of Lemma 5.4. □

With the above results we can prove our main theorem for the two-queue GPS system with Brownian inputs.

Theorem 5.6 *Suppose that the class-1 sources are* $\mathrm{BM}(\lambda_1, \mu_1)$ *and that the class-2 sources are* $\mathrm{BM}(\lambda_2, \mu_2)$*. Then, with*

$$\phi_2^c = 1 - \phi_1^c; \qquad \phi_2^o = \frac{\mu_2}{c},$$

it holds that (i) for $\phi_2 \in [0, \phi_2^o]$,

$$K^{(\mathrm{g})}(b) = 2 \cdot \frac{\phi_1 c - \mu_1}{\lambda_1} \cdot b;$$

(ii) for $\phi_2 \in [\phi_2^o, \phi_2^c]$,

$$K^{(\mathrm{g})}(b) = \frac{1}{2} \cdot U(1) \cdot b;$$

(iii) for $\phi_2 \in [\phi_2^c, 1]$,

$$K^{(\mathrm{g})}(b) = 2 \cdot \frac{c - \mu}{\lambda_1 + \lambda_2} \cdot b.$$

Proof. Case (i) follows directly from [12, Section 6]. Class 2 is in overload, and 'takes away it weight' without any effort. As a consequence, in essence, class 1 sees a queue with service rate $n\phi_1 c$.
The proof of cases (ii) and (iii) mimicks the proof of Theorem 4.4. The *lower bound* uses Lemma 5.1. Then the proof is as in the lower bound of Theorem 4.4, with the sets

$$T_1 := \{(t, u, x) \in \mathbb{R}_+^3 \mid \lambda t(b + \bar{c}_1 u) \leq \lambda_1 u(b + \bar{c}t + x)\};$$

T_2 with the '$\leq$'-sign replaced by '$\geq$', and $\bar{T}$ with the '$\leq$'-sign replaced by '='. The infima of the $k_i^{(\mathrm{g})}(b, t, u, x)$ (over t, u, x) for $i = 1, 2$ follow then from Lemmas 5.2, 5.3, and 5.5.
The *upper bound* is just a matter of verifying that the paths of the lower bound are feasible, and computing their norm. □

Example 5.7 Here we illustrate the result of the previous theorem by an example. We suppose that both types of sources correspond to Brownian motions, with $\mu_1 = 0.2$, $\mu_2 = 0.3$, $v_1(t) = 2t$, and $v_2(t) = t$. Take $c = 1$. With the buffer size of class i denoted by $B_i \equiv nb_i$, let $K_i^{(\mathrm{g})}(b_i)$ be the decay rate of class i, and $L_i^{(\mathrm{g})} := K_i^{(\mathrm{g})}(b_i)/b_i$. These $L_i^{(\mathrm{g})}$ are given in Figure 2, as a function of the weight ϕ_1.
Suppose the weight ϕ_1 (and hence implicitly also $\phi_2 = 1 - \phi_1$) has to be chosen such that the decay rate of class i is larger than δ_i (for $i = 1, 2$). Then we need to verify whether there is a $\phi_1 \in [0, 1]$ such that both $b_1 L_1^{(\mathrm{g})} \geq \delta_1$ and $b_2 L_2^{(\mathrm{g})} \geq \delta_2$. This can easily be verified from the graph below. ◇

6 Concluding remarks

In this paper we have computed, in a many-sources setting, the exponential decay rate of the overflow probability in a tandem queue, a priority system, and a system operating under a GPS scheduler. The input was assumed short-range dependent Gaussian traffic; we have

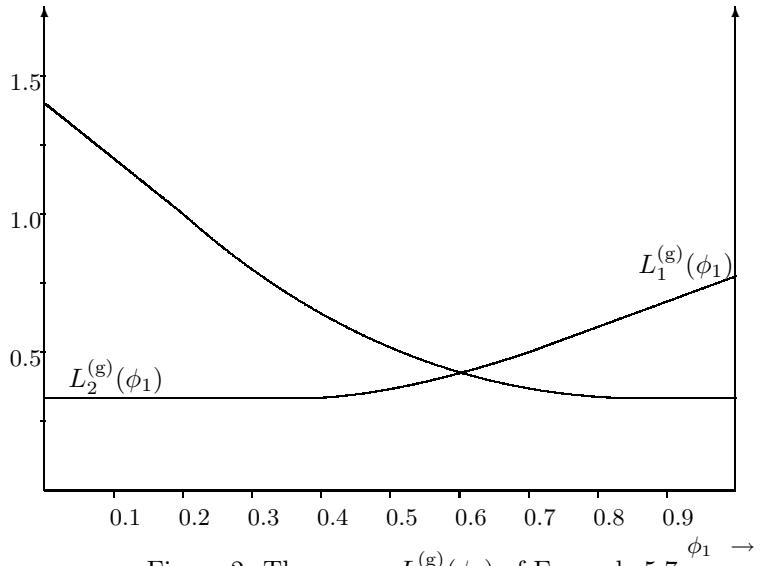

Figure 2: The curves $L_i^{(g)}(\phi_1)$ of Example 5.7.

distinguished between Brownian-motion input and input with an asymptotically linear variance function. A few remarks are in place.

- In Section 2.1, we introduced the M/G/∞ input process with Pareto jobs. It was argued that for $\alpha > 3$ the input is ALV. For $\alpha \in (2,3)$ it is true that $v(t)/t$ tends to a constant, but $v(t) - t$ does not. Hence, the process is short-range dependent, but *not* ALV.
To get an impression of the large-buffer behavior for $\alpha \in (2,3)$, we consider the FIFO queue fed by Gaussian sources with the (somewhat simpler) variance function $v(t) = t + t^\beta$, for $\beta \in (0,1)$; for ease, take $\mu = 0$. It is readily verified that the optimizing $t = t(b)$ is the inverse of

$$b(t) := \frac{ct + (2-\beta)ct^\beta}{1 + \beta t^{\beta-1}};$$

for large t, it holds that $b(t) \approx ct + 2(1-\beta)ct^\beta$, and hence also, for large b, that the optimizing t looks like $b/c - 2(1-\beta)(b/c)^\beta$. Now it can be verified that

$$\left(\inf_{t>0} \frac{(b+ct)^2}{2v(t)}\right) \approx 2bc - 2c^{2-\beta}b^\beta.$$

We see that a variance function consisting of a linear part as well as a polynomial, sublinear part leads to a decay rate function with a linear and a polynomial, sublinear part. We expect this type of behavior to carry over to the complex buffer architectures considered in this paper.

- In the GPS setting we only considered the case of BM input. In the situation with ALV input, we run into technical problems. In the counterpart of Lemma 5.1 for ALV sources, the

minimum needs to be taken over all $t \geq 0$ and $u \in (0, t)$. Because u can be chosen close to t, we expect that we have to impose regularity conditions on $v(\cdot)$ around 0, to be able to compute the minimum over t and u.

- Zhang [24] also considers behavior of GPS schedulers for short-range dependent traffic (more general than Gaussian, but discrete-time). His assumptions are in line with those in, e.g., Glynn and Whitt [9], and are of the following type.
With $A(t)$ denoting the traffic generated by a single source in an interval of length t, it is assumed that $\lim_{t\to\infty} t^{-1} \log \mathbb{E} \exp(\theta A(t))$ is finite for positive θ; for Gaussian sources this would be equivalent to requiring that $v(t)$ is at most linear. Such a framework obviously allows for instance $v(t)/t \to \lambda$. The results obtained are of the type $I(b)/b \to \theta^\star$ for $b \to \infty$, where $I(b)$ is the decay rate of overflow in the queue under consideration, and $\theta^\star$ is a positive constant. Our requirement in the variance function, i.e., $v(t) - \lambda t \to \kappa$ for ALV sources, is more demanding (in the sense that it implies that $v(\cdot)$ is at most linear), but, in return, we get more precise results: $I(b) - \theta^\star b \to \nu$.

Acknowledgment. The author is indebted to Miranda van Uitert (CWI and Vrije Universiteit, Amsterdam) for several useful comments.

References

[1] R. Addie, P. Mannersalo and I. Norros. Most probable paths and performance formulae for buffers with Gaussian input traffic. *European Transactions on Telecommunications*, 13: 183 – 196, 2002.

[2] D. Anick, D. Mitra and M. Sondhi. Stochastic theory of a data handling system with multiple sources, *Bell System Technical Journal,* 61: 1871 – 1894, 1982.

[3] R. Adler. An introduction to continuity, extrema, and related topics for general Gaussian processes. *IMS Lecture Notes-Monograph Series*, 12, 1990.

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

[5] D. Botvich and N. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems*, 20: 293 – 320, 1995.

[6] C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *Journal of Applied Probability*, 33: 886 – 903, 1996.

[7] A. Dembo and O. Zeitouni. Large deviations techniques and applications, 2nd edition. Springer Verlag, New York, USA, 1998.

[8] J.-D. Deuschel and D. Stroock. Large Deviations. Academic Press, London, 1989.

[9] P. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability*, 31A: 131 – 156, 1994.

[10] M. Mandjes, I. Saniee and A. Stolyar. Load characterization, overload prediction, and load anomaly detection for voice over IP traffic. *Proceedings 38th Allerton Conference,* Urbana-Champaign, US, 567 – 576, 2000.

[11] M. Mandjes and M. van Uitert. Sample-path large deviations for tandem and priority queues with Gaussian inputs. To appear in *Annals of Applied Probability.*

Available from: `http://db.cwi.nl/rapporten/index.php?persnr=1271`.

[12] M. Mandjes and M. van Uitert. Sample-path large deviations for generalized processor sharing queues with Gaussian inputs. To appear in *Performance Evaluation.*

Available from: `http://db.cwi.nl/rapporten/index.php?persnr=1271`.

[13] M. Mandjes, P. Mannersalo and I. Norros. Gaussian tandem queues, and resulting performance formulae. *Submitted.*

Available from: `http://db.cwi.nl/rapporten/index.php?persnr=1271`.

[14] M. Mandjes, P. Mannersalo, I. Norros and M. van Uitert. Large deviations of infinite intersections of events in Gaussian processes. *Submitted.*

Available from: `http://db.cwi.nl/rapporten/index.php?persnr=1271`.

[15] P. Mannersalo and I. Norros. Approximate formulae for Gaussian priority queues. *Proceedings ITC 17*, Salvador da Bahia, Brazil, 991 – 1002, 2001.

[16] P. Mannersalo and I. Norros. GPS schedulers and Gaussian traffic. *Proceedings IEEE Infocom*, New York, USA, 1660 – 1667, 2002.

[17] P. Mannersalo and I. Norros. A most probable path approach to queueing systems with general Gaussian input. *Computer Networks*, 40: 399 – 412, 2002.

[18] I. Norros. Busy periods of fractional Brownian storage: a large deviations approach. *Advances in Performance Analysis,* 2: 1 – 20, 1999.

[19] I. Norros. Most probable paths in Gaussian priority queues. COST257 TD(99)16.

Available from: `http://www.vtt.fi/tte/tte23/cost257/`.

[20] A. Parekh and R. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The single node case. *IEEE/ACM Transactions on Networking*, 1: 344 – 357, 1993.

[21] A. Parekh and R. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The multiple node case. *IEEE/ACM Transactions on Networking*, 2: 137 – 150, 1994.

[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. *Annals of Applied Probability*, 11: 379 – 404, 2001.

[24] Z.-L. Zhang. Large deviations and the processor sharing scheduling for a two-queue system. *Queueing Systems,* 26: 229 – 264, 1997.

Appendix

Lemma A.1 *Take $A, C \in \mathbb{R}$; $B, D > 0$; $\sigma > 0$; $\tau \in \mathbb{R}$. Then*

$$\lim_{x\to\infty}\left(\inf_{t\geq 0}\frac{(\sigma t)^2}{A+Bt}+\frac{(x-\tau t)^2}{C+Dt}\right)-\frac{2x}{D}\left(\sqrt{\tau^2+\frac{D\sigma^2}{B}}-\tau\right)=-\frac{\sigma^2}{B}\left(\frac{A}{B}-K_2\right)+\frac{K_2^2}{K_1}\cdot\frac{\tau}{C},$$

with

$$K_1 := \left(\sqrt{\tau^2+\frac{D\sigma^2}{B}}\right)^{-1}, \quad K_2 := -\frac{C}{D}\left(1-\frac{\tau}{\sqrt{\tau^2+D\sigma^2/B}}\right).$$

Proof. First fix x, and differentiate with respect to t to find the following first-order condition:

$$\sigma^2\cdot\frac{2At+Bt^2}{(A+Bt)^2}+2\left(\frac{\tau t-x}{C+Dt}\right)-D\left(\frac{\tau t-x}{C+Dt}\right)^2=0,$$

which is solved by

$$\frac{\tau t-x}{C+Dt}=\frac{1}{D}\left(\tau+\sqrt{\tau^2+D\sigma^2\cdot\frac{2At+Bt^2}{(A+Bt)^2}}\right).$$

We can equivalently express x as function of t:

$$x(t)=-\frac{C}{D}\cdot\tau+\left(\frac{C}{D}+t\right)\sqrt{\tau^2+D\sigma^2\cdot\frac{2At+Bt^2}{(A+Bt)^2}}.$$

Now it is readily checked that

$$\lim_{t\to\infty}\left(x(t)-t\sqrt{\tau^2+\frac{D\sigma^2}{B}}\right)=\frac{C}{D}\left(\sqrt{\tau^2+\frac{D\sigma^2}{B}}-\tau\right),$$

and hence $\lim_{x\to\infty} t(x) - K_1 x = K_2$. Inserting these into the objective function yields the stated. □