Centrum voor Wiskunde en Informatica

**REPORT***RAPPORT*

PNA

Probability, Networks and Algorithms

***Probability, Networks and Algorithms***

PNA

# A note on the delay distribution in GPS

P.M.D. Lieshout, M.R.H. Mandjes

**REPORT PNA-R0704 APRIL 2007**

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation 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

# A note on the delay distribution in GPS

ABSTRACT
In this note a two-class Generalized Processor Sharing (GPS) system is considered. We analyze the probability that the virtual delay of a particular class exceeds some threshold. We apply Schilder's theorem to calculate the logarithmic many-sources asymptotics of this probability in the important case of Gaussian inputs.

*2000 Mathematics Subject Classification:* 60G15;60K25
*Keywords and Phrases:* Generalized processor sharing; Gaussian traffic; Delay probabilities; Sample-path large deviations; Many-sources asymptotics
*Note*: This research has been funded by the Dutch Bsik/BRICKS (Basic Research in Informatics for Creating the Knowledge Society) project.

# A note on the delay distribution in GPS

P. Lieshout[1,*] and M. Mandjes[1,2,3]

[1]CWI
P.O. Box 94079, 1090 GB Amsterdam, the Netherlands

[2]Kortweg-de Vries Institute
University of Amsterdam
Plantage Muidergracht 24, 1018 TV Amsterdam, the Netherlands

[3]EURANDOM
P.O. Box 513, 5600 MB Eindhoven, the Netherlands

2nd April 2007

**Abstract**

In this note a two-class Generalized Processor Sharing (GPS) system is considered. We analyze the probability that the virtual delay of a particular class exceeds some threshold. We apply Schilder's theorem to calculate the logarithmic many-sources asymptotics of this probability in the important case of Gaussian inputs.

*Keywords:* Generalized processor sharing; Gaussian traffic; Delay probabilities; Sample-path large deviations; Many-sources asymptotics

*Corresponding author. Email addresses: lieshout@cwi.nl (P. Lieshout), mmandjes@science.uva.nl (M. Mandjes). This research has been funded by the Dutch BSIK/BRICKS (Basic Research in Informatics for Creating the Knowledge Society) project.

## 1 Introduction

The communication networks of the nearby future are expected to support a wide range of heterogeneous services. Here one can think of traditional applications, such as data, video and voice, but also of sophisticated multimedia applications, such as gaming and remote surgery. These services may have different traffic characteristics, but in addition may also have different Quality-of-Service (QoS) requirements. The integration of heterogeneous traffic flows thus raises the need for service differentiation.

A service discipline that is capable of supporting different QoS-levels, is Generalized Processor Sharing (GPS). The GPS discipline assigns weights to the traffic classes, and these weights determine the guaranteed minimum service rates for the classes. In case some class does not fully use its minimum rate, then the excess rate becomes available to the other classes, also shared according to these weights.

Since the exact analysis of GPS queues is often intractable, most of the work on GPS has focused on various bounds and asymptotic approximations of statistical performance guarantees, such as loss probabilities, delay characteristics, and workload distributions. In particular, in the literature one has mainly focused on the (asymptotics of the) loss probability of a particular class, see [3, 4, 7, 12, 16, 17]. Besides losses (due to buffer overflow), also delay strongly determines the QoS as perceived by customers. Particularly for real-time applications, the delay is only allowed to exceed some predefined threshold with extremely small probability. Hence, the (exponential) decay rate of the delay probability is an important performance measure. In literature, however, hardly any results are available on this decay rate. Paschalidis [14] focused on a two-class GPS system, in a discrete-time setting, in which the input traffic was assumed to be short-range dependent, and derived logarithmic asymptotics of the probability that the delay exceeds some large value.

The contribution of this note is twofold. In the first place we derive bounds on the delay probability in a two-class GPS system with general input processes, assuming that the inputs have stationary increments. Secondly, we then consider the situation of $n$ input processes of both classes, scale the link capacity with $n$ as well, and let $n$ grow large. This so-called 'many-sources regime' is motivated by the fact that, particularly in the core of the network, the queues need to serve a large number of flows at the same time. In this many-sources framework, we apply Schilder's sample-path large deviations theorem to calculate the decay rates of these bounds in the important case of Gaussian inputs, which cover both short-range and long-range dependent traffic. We note that this work is related to [10], where the authors derive (lower bounds on) the decay rate of the overflow probability in a two-class GPS systems; for other related work, see [1, 8, 11]. We show that there exist two closed intervals of GPS weight values in which the bounds are tight: one containing the special case that class 1 has priority, and the other containing the case that class 2 has priority. For the remaining middle interval, we derive bounds on the decay rate. In the special case of Brownian inputs we obtain transparent closed-form expressions.

The remainder of this note is organized as follows. In Section 2 we describe the two-class GPS model, and derive bounds on the delay probability. In Section 3 we specialize to Gaussian traffic in a many-sources setting: using Schilder's theorem and the bounds mentioned above, we derive (bounds on) the corresponding decay rate.

## 2 Bounds on the virtual delay probability

In this note we consider a two-class GPS system, served with rate $c$. Each class has its own queue, and is assigned a weight $\phi_i \geq 0$, $i = 1, 2$. Without of loss of generality it is assumed that $\phi_1 + \phi_2 = 1$. The weight $\phi_i$ determines the guaranteed minimum rate $\phi_i c$ for class $i$. If a class does not fully use this minimum rate, then the excess capacity becomes available for the other class. Note that GPS is a work-conserving scheduling discipline, i.e., if at least one of the queues is non-empty, then the server always works at full speed.

We focus on the delay experienced by a packet ('fluid molecule') of a particular class, say class 1, having arrived at an arbitrary point in time, the so-called *virtual delay*. We assume that the system is stable, such that the delay is bounded almost surely. Also, without loss of generality we assume that the packet arrives at time 0. We denote the delay experienced by this packet by $D_1 \equiv D_1(0)$. Clearly,

$$p(d) := \mathbb{P}(D_1 > d) = \mathbb{P}(Q_1 > B_1(0, d)), \tag{1}$$

where $Q_i \equiv Q_i(0)$ is the steady-state queue length of class $i$, and $B_i(s, t)$ is the amount of service received by class $i$ in the interval $(s, t]$. Likewise, let $X_i(s, t)$ be the amount of traffic generated by class $i$ in the interval $(s, t]$. Throughout this note we assume that $X_1(s, t)$ is independent of $X_2(s, t)$, i.e., $\mathbb{C}\mathrm{ov}(X_1(s, t), X_2(s, t)) = 0$, $s \leq t$.

In this section we derive bounds on $p(d)$, which apply to all input processes that have stationary increments; stationarity of the increments means that the distribution of $X_i(s, s + t)$ does not depend on $s$, but just on the interval length $t$. We will use these bounds in Section 3 to derive (bounds on) the exponential decay rate of $p(d)$ in the many-sources setting.

To derive a lower bound on $p(d)$, we need to find an upper bound on $B_1(0, d)$, as follows from (1). As $B_1(0, d) \leq cd - B_2(0, d)$, this is equivalent to finding a lower bound on $B_2(0, d)$. Now, we have to distinguish between two scenarios: $(i)$ queue 2 is continuously backlogged in the interval $(0, d]$ and $(ii)$ queue 2 is empty at some time in $(0, d]$. In case $(i)$ we have that $B_2(0, d) = \phi_2 cd$, because the second class receives at least its guaranteed service rate in the interval $(0, d]$, and class 1 is continuously backlogged by definition (otherwise it cannot experience a delay of $d$), thus claiming at least its guaranteed rate in the interval $(0, d]$. In case $(ii)$ we need a different approach to derive a lower bound on $B_2(0, d)$. Let $z$ denote the last time in $(0, d]$ that the second queue was empty, that is $z := \max\{v \in (0, d] : Q_2(v) = 0\}$. This yields

$$B_2(0, d) = B_2(0, z) + B_2(z, d) = Q_2 + X_2(0, z) + \phi_2 c(d - z) \geq \inf_{u \in (0,d]} \{X_2(0, u) + \phi_2 c(d - u)\}. \tag{2}$$

Note that the right-hand side of (2) will not exceed $\phi_2 cd$. That is, it is also a lower bound on $B_2(0, d)$ in case $(i)$. Therefore, we find the following upper bound:

$$B_1(0, d) \leq cd - B_2(0, d) \leq cd - \inf_{u \in (0,d]} \{X_2(0, u) + \phi_2 c(d - u)\}.$$

Hence, we obtain

$$p(d) \geq \mathbb{P}\left(Q_1 > cd - \inf_{u \in (0,d]} \{X_2(0, u) + \phi_2 c(d - u)\}\right). \tag{3}$$

So far no explicit expressions have been found for the steady-state queue length distribution of a particular class in a GPS system. In other words: we do not know the distribution of $Q_1$, which makes the lower bound (3) not very useful; we would rather like to have a bound that is in terms of the input processes $X_1$ and $X_2$ only. Using that, for some $b \geq 0$,

$$\mathbb{P}(Q_1 > b) = \mathbb{P}\left(\bigcup_{x \geq 0} \{Q_1 + Q_2 > x + b, Q_2 \leq x\}\right),$$

we find that (3) can be rewritten as

$$p(d) \geq \mathbb{P}\left(\bigcup_{x \geq 0} \left\{Q_1 + Q_2 > x + cd - \inf_{u \in (0,d]} \{X_2(0,u) + \phi_2 c(d-u)\}, Q_2 \leq x\right\}\right). \tag{4}$$

But now observe that $Q_1 + Q_2$ is the steady-state queue length of the total queue, and hence, due to the work-conserving nature of GPS, Reich's identity [15] implies that

$$Q_1 + Q_2 = \sup_{t \geq 0} \{X_1(-t,0) + X_2(-t,0) - ct\}. \tag{5}$$

Also, again by Reich's identity,

$$Q_2 = \sup_{s \geq 0} \{X_2(-s,0) - B_2(-s,0)\}. \tag{6}$$

The negative of the optimizing $t$ ($s$), denoted by $t^*$ ($s^*$), can be interpreted as the beginning of the busy period of the total (second) queue containing time 0. Clearly, this entails that $s^* \leq t^*$. Now, using (5) and (6), the lower bound (4) can be expressed as

$$p(d) \geq \mathbb{P}\left(\exists x \geq 0, t \geq 0 : \forall s \in [0,t] : \forall u \in (0,d] : \begin{array}{l} X_1(-t,0) + X_2(-t,u) > x + ct + \phi_1 cd + \phi_2 cu; \\ X_2(-s,0) \leq x + B_2(-s,0) \end{array}\right). \tag{7}$$

From (7) we conclude that, in order to find a lower bound on $p(d)$ that only depends on the input processes $X_1$ and $X_2$, we have to find a lower bound on $B_2(-s,0)$.

**Lemma 2.1** *$p(d)$ is lower bounded by*

$$\mathbb{P}\left(\exists x \geq 0, t \geq 0 : \forall s \in [0,t] : \forall u \in (0,d] : \begin{array}{l} X_1(-t,0) + X_2(-t,u) > x + ct + \phi_1 cd + \phi_2 cu; \\ X_2(-s,0) \leq x + \phi_2 cs \end{array}\right).$$

*Proof:* Since $-s^*$ denotes the beginning of the the busy period, queue 2 is continuously backlogged in the interval $(-s^*, 0]$, and therefore $B_2(-s^*,0) \geq \phi_2 cs^*$. This implies that the right-hand side of (7) is lower bounded by the stated, and therefore also $p(d)$. □

Likewise, to derive an upper bound on $p(d)$ we need to find a lower bound on $B_1(0,d)$. A first lower bound on $B_1(0,d)$ is clearly given by $B_1(0,d) \geq cd - Q_2 - X_2(0,d)$. This is a direct implication of the fact that, in an interval $(0,d]$, a queue never claims more than the queue length at time 0, increased by the amount of traffic arriving at this queue in $(0,d]$.

**Lemma 2.2** *$p(d)$ is upper bounded by*

$$\mathbb{P}\left(\exists t \geq 0 : X_1(-t,0) + X_2(-t,d) > ct + cd\right).$$

*Proof:* Since $B_1(0,d) \geq cd - Q_2 - X_2(0,d)$, we have

$$p(d) \leq \mathbb{P}(Q_1 > cd - Q_2 - X_2(0,d)) = \mathbb{P}(Q_1 + Q_2 > cd - X_2(0,d)).$$

Using (5), it is easily seen that the right-hand side is equivalent to the stated. □

Class 1 can only experience a delay of $d$ if class 1 is continuously backlogged in the interval $(0,d]$. This implies that $B_1(0,d) \geq \phi_1 cd$, from which we deduce the following second upper bound.

**Lemma 2.3** *$p(d)$ is upper bounded by*

$$\mathbb{P}\left(\exists x \geq 0, t \geq 0 : \forall s \in [0,t] : \exists v \in [0,s] : \begin{array}{c} X_1(-t,0) + X_2(-t,0) > x + ct + \phi_1 cd; \\ X_1(-s,-v) + X_2(-s,0) \leq x + cs - \phi_1 cv \end{array}\right).$$

*Proof:* Since $B_1(0,d) \geq \phi_1 cd$, we have that $p(d) \leq \mathbb{P}(Q_1 > \phi_1 cd)$. In Section 3 of [10] it is shown that $\mathbb{P}(Q_1 > \phi_1 cd)$ is upper bounded by the stated. □

Notice the similarity between the lower bound of Lemma 2.1 and the upper bound of Lemma 2.3.

# 3 Decay rate of the virtual delay probability

In this section we derive (bounds on) the decay rate of the virtual delay probability in case of Gaussian inputs. We consider a many-sources setting, where the link capacity is scaled proportionally to the number of sources. In the special case of Brownian inputs we obtain closed-form expressions.

## 3.1 Gaussian input traffic

Let class $i$ consist of a superposition of $n$, $n \in \mathbb{N}$, i.i.d. flows (or: sources), $i = 1,2$; the analysis can easily be extended to the case of an unequal number of sources, see Remark 2.2 in [10]. Let the service capacity be $nc$. A class-$i$ flow behaves as a Gaussian process with stationary increments $\{A_i(t), t \in \mathbb{R}\}$, with $A_i(t) \equiv 0$. Also, let the mean traffic rate and variance function of a single class-$i$ flow be denoted by $\mu_i > 0$ and $v_i(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$, respectively, $i = 1,2$. This mean rate and variance curve fully characterize the probabilistic behavior of the flow $A_i(\cdot)$. To guarantee stability we assume that $\mu_1 + \mu_2 \leq c$. With $A_i(s,t) := A_i(t) - A_i(s)$ denoting the amount of traffic generated by a single flow of type $i$ in the interval $(s,t]$, $A_i(s,t)$ has a Normal distribution with $\mathbb{E}A_i(s,t) = \mu_i \cdot (t-s)$ and $\mathbb{V}\text{ar}A_i(s,t) = v_i(t-s)$; recall that the assumption of stationary increments entails that the law of $A_i(s,t)$ only depends on the *length* of the interval

$(s,t]$. We also introduce the centered process $\overline{A}_i(t) := A_i(t) - \mu_i t$; we write $\overline{A}_{i,j}(t)$ when we refer to the $j$-th flow of class $i$, $j = 1, ..., n$. It is well-known that the covariance function $\Gamma_i(s,t)$ can be written as

$$\Gamma_i(s,t) := \mathbb{C}\mathrm{ov}\left(A_i(s), A_i(t)\right) = \mathbb{C}\mathrm{ov}\left(\overline{A}_i(s), \overline{A}_i(t)\right) = \frac{1}{2}\left(v_i(s) + v_i(t) - v_i(t-s)\right), \tag{8}$$

for all $0 \leq s \leq t$. We impose the following (weak) assumptions on $v_i(\cdot)$, $i = 1, 2$.

**Assumption 3.1** *For $i = 1, 2$,*
A1 $v_i(\cdot) \in C_1([0,\infty))$.
A2 For some $\alpha < 2$ it holds that $v_i(t)/t^\alpha \to 0$, as $t \to \infty$.
A3 $v_i(\cdot)$ is non-decreasing.

### 3.2 Large deviations

In this subsection we recall two key large-deviations theorems, which are needed in the analysis of the next subsection.

**Theorem 3.2** *Let $(X,Y) \sim Norm(0,\Sigma)$, for a non-degenerate 2-dimensional covariance-matrix $\Sigma$. Then,*

$(i)$ $\quad -\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq x\right) = \frac{1}{2}x^2/(\Sigma_{11})^2$;
$(ii)$ $\quad -\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq x, \frac{1}{n}\sum_{i=1}^n Y_i \geq y\right) = \inf_{a\geq x} \inf_{b\geq y} \Lambda(a,b)$,

*where $\Lambda(a,b) := \frac{1}{2}(a\ b)\Sigma^{-1}(a\ b)^{\mathrm{T}}$.*

We continue with a brief description of the framework of Schilder's sample-path LDP (see [2], and also Thm. 1.3.27 of [6] for a more detailed treatment). Then the path space is $\Omega := \Omega_1 \times \Omega_2$, with

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

We note that in [1] it was pointed out that $\overline{A}_i(\cdot)$ can be realized on $\Omega_i$, $i = 1, 2$. Now one can construct a reproducing kernel Hilbert space $R_i \subseteq \Omega_i$, consisting of elements that are roughly as smooth as the covariance function $\Gamma_i(s,\cdot)$; for details, see [1, 6, 10]. Suppose that the $\omega_i(\cdot)$ are linear combinations of covariance functions: $\omega_i(\cdot) = \sum_{j=1}^{m_i} a_{ij}\Gamma(s_{ij},\cdot)$, with $a_{ij}$, $s_{ij} \in \mathbb{R}$, $j = 1, ..., m_i$, $m_i \in \mathbb{N}$, $i = 1, 2$. Then we can define the rate function:

$$I(\omega) = I(\omega_1, \omega_2) := \frac{1}{2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_1} a_{1i}a_{1j}\Gamma_1(s_{1i}, s_{1j}) + \frac{1}{2}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} a_{2i}a_{2j}\Gamma_2(s_{2i}, s_{2j}); \tag{9}$$

this definition can be extended to $\omega(\cdot)$ in $R := R_1 \times R_2$ that are no linear combinations of the covariance functions, see [1, 2, 6].

**Theorem 3.3** [**Schilder**] *For centered Gaussian inputs the following sample-path large deviations principle (LDP) holds, under Assumptions A1 and A2:*
*(a) For any closed set $F \subset \Omega$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\left(\frac{1}{n}\sum_{j=1}^{n} \overline{A}_{1,j}(\cdot), \frac{1}{n}\sum_{j=1}^{n} \overline{A}_{2,j}(\cdot)\right) \in F\right) \leq -\inf_{\omega\in F} I(\omega);$$

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

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\left(\frac{1}{n}\sum_{j=1}^{n} \overline{A}_{1,j}(\cdot), \frac{1}{n}\sum_{j=1}^{n} \overline{A}_{2,j}(\cdot)\right) \in G\right) \geq -\inf_{\omega\in G} I(\omega).$$

*Remark:* Theorem 3.3 shows that the LDP consists of an upper and lower bound, which apply to closed and open sets, respectively. It can be verified that in the present paper we have open sets $G$, that are such that

$$\inf_{\omega\in G} I(\omega) = \inf_{\omega\in\overline{G}} I(\omega),$$

where $\overline{G}$ is the closure of $G$. The way to prove this is to show that an arbitrarily chosen path in $\overline{G}$ can be approximated by a path in $G$. This proof is completely analogously to [13] and Appendix A of [9].

### 3.3 Decay rate

In this subsection we derive (bounds on) the decay rate corresponding to the virtual delay probability

$$p_n(d) := \mathbb{P}(Q_{1,n} > B_{1,n}(0,d)), \quad n\to\infty,$$

where $Q_{1,n} \equiv Q_{1,n}(0)$ is the steady-state class-1 queue length and $B_{1,n}(0,d)$ is the amount of service received in the interval $(0,d]$ by class 1, in a system with $n$ class-$i$ inputs, $i=1,2$, that has service capacity $nc$.

From Theorem 3.3 and the remark in Section 3.2, it follows that

$$J(d) := -\lim_{n\to\infty} \frac{1}{n} \log p_n(d) = \inf_{f\in L} I(f) = \inf_{f\in\overline{L}} I(f),$$

where the open (closed) set $L$ ($\overline{L}$) consists of all paths $(f_1, f_2)$ that give a delay larger (larger or equal) than $d$. The path in $L$ (and likewise in $\overline{L}$) that minimizes the decay rate, $f^* = (f_1^*, f_2^*)$, is also known as the so-called *most probable path* (MPP). Informally speaking, given that the rare event occurs, with overwhelming probability a delay of $d$ is achieved by a path 'close to' the MPP, cf. [1].

Recall that in Section 2 we derived bounds on $p(d)$. In this subsection we will exploit these bounds, to derive (lower bounds on) $J(d)$. Note that the decay rates of the upper (lower) bounds on $p(d)$ are lower (upper) bounds on $J(d)$ for all $\phi_2 \in [0,1]$.

### 3.3.1 Class 2 in overload

We first focus on the regime $\phi_2 \in [0, \mu_2/c]$, i.e., class 2 in overload, and we derive an exact expression for the decay rate of $p_n(d)$. Recall that $B_1(d) \geq \phi_1 cd$ in order to have a delay of $d$. This yields

$$p_n(d) \leq \mathbb{P}(Q_{1,n} \geq n\phi_1 cd) \leq \mathbb{P}(Q_{1,n}^{n\phi_1 c} \geq n\phi_1 cd),$$

where $Q_{1,n}^a \equiv Q_{1,n}^a(0)$ denotes the stationary workload of queue 1 if it is served in isolation at constant rate $a$.

**Lemma 3.4** *If $\phi_2 \in [0, \mu_2/c]$, then*

$$J(d) = \inf_{t\geq 0} \frac{(\phi_1 cd + (\phi_1 c - \mu_1)t)^2}{2v_1(t)}. \tag{10}$$

*Let $t^*$ be optimizer in the above equation. Then, the MPP is given by*

$$\begin{aligned}
f_1^*(r) &= \begin{cases} -\mathbb{E}(A_1(r,0)|A_1(-t^*,0) = \phi_1 c(t^*+d)) & \textit{for } r \leq 0; \\ \mathbb{E}(A_1(0,r)|A_1(-t^*,0) = \phi_1 c(t^*+d)) & \textit{for } r > 0. \end{cases} \\
f_2^*(r) &= \begin{cases} -\mathbb{E}(A_2(r,0)|A_1(-t^*,0) = \phi_1 c(t^*+d)) & \textit{for } r \leq 0; \\ \mathbb{E}(A_2(0,r)|A_1(-t^*,0) = \phi_1 c(t^*+d)) & \textit{for } r > 0. \end{cases}
\end{aligned} \tag{11}$$

*Proof:* The decay rate $J^L(d)$ of $\mathbb{P}(Q_{1,n}^{n\phi_1 c} \geq n\phi_1 cd)$ in case $\phi_2 \in [0, \mu_2/c]$ is given in Theorem 6.1 of [10]. Note that because $\mathbb{P}(Q_{1,n}^{n\phi_1 c} \geq n\phi_1 cd)$ is an upper bound on the delay probability, its decay rate $J^L(d)$ is a lower bound on $J(d)$. In addition, in Section 6 of [10] the authors also derived the MPP $\tilde{f} = (\tilde{f}_1, \tilde{f}_2)$ corresponding to $J^L(d)$ using (9). The decay rate $J^L(d)$ is given by (10) and $\tilde{f}$ is given by (11). What is left to show is that $J(d) = J^L(d)$ and $f^* = \tilde{f}$.

Finding an upper bound $J^U(d)$ on $J(d)$ is a matter of computing the rate function ('costs') of a path in $L$, i.e., a path that produces a delay of at least $d$. Let $(Y_1, Y_2)$ be bivariate Normally distributed. Now, using that the random variable $(Y_1|Y_2 = y)$, for some $y \in \mathbb{R}$, is Normally distributed with mean

$$\mathbb{E}(Y_1|Y_2 = y) = \mathbb{E}Y_1 + \frac{\mathbb{C}\mathrm{ov}(Y_1, Y_2)}{\mathbb{V}\mathrm{ar}\, Y_2}(y - \mathbb{E}Y_2), \tag{12}$$

it can be verified that

$$\tilde{f}_1(r) = \mu_1 r - \frac{(\phi_1 cd + (\phi_1 c - \mu_1)t^*)}{v_1(t^*)}\Gamma_1(-r, t^*) \quad \text{for } r \in (-t^*, 0];$$

$$\tilde{f}_2(r) = \mu_2 r \quad \text{for } r \in (-t^*, d].$$

This path is such that, at time 0, queue 1 has buffer content $\phi_1 cd$ (as $Q_1(-t^*) = 0$), and such that queue 2 continuously claims service rate $\phi_2 c$ in the interval $(0, d]$ (as $\mu_2 \geq \phi_2 c$), i.e., service rate $\phi_1 c$ is available for class 1 in the interval $(0, d]$. Hence, we conclude that path $\tilde{f}$ results in a delay of exactly $d$, i.e., $\tilde{f} \in L$, implying that $f^* = \tilde{f}$ and, using (9), $J^U(d) = J^L(d) = J(d)$. □

#### 3.3.2 Class 2 in underload

We now consider the regime $\phi_2 \in (\mu_2/c, 1]$ and we derive the decay rate $J(d)$. In the analysis below the following critical class 2 weight is of importance:

$$\phi_2^F := \frac{\mu_2}{c} + \frac{v_2'(d-r^*) + v_2'(t^*+r^*)}{2(v_1(t^*) + v_2(t^*+d))} \left( \left(1 - \frac{\mu_1+\mu_2}{c}\right) t^* + \left(1 - \frac{\mu_2}{c}\right) d \right), \tag{13}$$

where $t^*$ is minimizer of

$$\inf_{t\geq 0} \frac{(c(t+d) - \mu_1 t - \mu_2(t+d))^2}{2v_1(t) + 2v_2(t+d)}, \tag{14}$$

and where $r^*$ is maximizer of

$$\sup_{r\in(-t^*,d]} v_2'(d-r) + v_2'(t^*+r). \tag{15}$$

Note that $\phi_2^F \geq \mu_2/c$, as $v'(\cdot) \geq 0$ by Assumption A3, and possibly larger than 1. The next theorem presents the exact decay rate in case $\phi_2 \in [\phi_2^F, 1]$ (if this interval is non-empty).

**Lemma 3.5** *If $\phi_2 \in [\phi_2^F, 1]$, then*

$$J(d) = \inf_{t\geq 0} \frac{(c(t+d) - \mu_1 t - \mu_2(t+d))^2}{2v_1(t) + 2v_2(t+d)}. \tag{16}$$

*Let $t^*$ be optimizer in the above equation. Then, the MPP is given by*

$$\begin{aligned}
f_1^*(r) &= \begin{cases} -\mathbb{E}(A_1(r,0)|A_1(-t^*,0) + A_2(-t^*,d) = c(t^*+d)) & \text{for } r \leq 0; \\ \mathbb{E}(A_1(0,r)|A_1(-t^*,0) + A_2(-t^*,d) = c(t^*+d)) & \text{for } r > 0. \end{cases} \\
f_2^*(r) &= \begin{cases} -\mathbb{E}(A_2(r,0)|A_1(-t^*,0) + A_2(-t^*,d) = c(t^*+d)) & \text{for } r \leq 0; \\ \mathbb{E}(A_2(0,r)|A_1(-t^*,0) + A_2(-t^*,d) = c(t^*+d)) & \text{for } r > 0. \end{cases}
\end{aligned} \tag{17}$$

*Proof:* Lemma 2.2 gives an upper bound on the delay probability. The decay rate $J^L(d)$ of this upper bound and the corresponding MPP $\tilde{f}$, the latter obtained by using (9), are well known (see for instance [1]), and given by (16) and (17), respectively. Below we show that $J^L(d) = J(d)$, or equivalently, that $\tilde{f} \in L$ if $[\phi_2^F, 1]$ (similarly to the proof of Lemma 3.4).

Using (12), it can be verified that

$$\tilde{f}_1(r) = \mu_1 r - \frac{\Gamma_1(-r,t^*)}{v_1(t^*) + v_2(t^*+d)} (c(t^*+d) - \mu_1 t^* - \mu_2(t^*+d)) \quad \text{for } r \in (-t^*, 0];$$

$$\tilde{f}_2(r) = \mu_2 r - \frac{\Gamma_2(t^*, t^*+d) - \Gamma_2(t^*+r, t^*+d)}{v_1(t^*) + v_2(t^*+d)} ((c-\mu_2)(t^*+d) - \mu_1 t^*) \quad \text{for } r \in (-t^*, d];$$

$$g_2(r) := \frac{\mathrm{d}\tilde{f}_2(r)}{\mathrm{d}r} = \mu_2 + \frac{v_2'(d-r) + v_2'(t^*+r)}{2v_1(t^*) + 2v_2(t^*+d)} ((c-\mu_2)(t^*+d) - \mu_1 t^*),$$

i.e., $g_2(\cdot)$ represents the input rate of the path $\tilde{f}_2$, which is derived using (8). Note that $-t^*$ denotes the beginning of the busy period of the total queue, i.e., $Q_1(-t^*) = Q_2(-t^*) = 0$. Hence, if the input rate of class 2 is smaller than the guaranteed minimum rate $\phi_2 c$ for all $r \in (-t^*, d]$, then clearly queue 2 is empty in the interval $(-t^*, d]$. Let

$$r^* := \arg \max_{r \in (-t^*, d]} g_2(r),$$

i.e., $r^*$ is maximizer of (15). Then queue 2 is empty in the interval $(-t^*, d]$ if $\phi_2 \geq g_2(r^*)/c = \phi_2^F$. Now, note that the path $\tilde{f}$ is such that

$$Q_1 + Q_2 = Q_1 = cd - \tilde{f}_2(d) = \int_0^d (c - g_2(r))\mathrm{d}r,$$

in case $\phi_2 \in [\phi_2^F, 1]$. As class 2 only uses rate $g_2(r) \leq \phi_2 c$ in this case, this implies that rate $c - g_2(r)$ is available for the first class, $r \in (-t^*, d]$. It is not hard to see that, given $Q_1(0) = \int_0^d (c - g_2(r))\mathrm{d}r$ and service rate $c - g_2(r)$ for the first class, the experienced delay in steady state is exactly $d$. This proves that $\tilde{f} \in L$, i.e., $f^* = \tilde{f}$ and $J(d) = J^L(d)$. □

We now focus on the remaining interval of weights: $\phi_2 \in \left(\mu_2/c, \phi_2^F\right)$. We have not succeeded in finding the exact decay rate in this middle regime, but we present two lower bounds; it is noted that lower bounds on the decay rate, which correspond to upper bounds on the probability $p_n(d)$, are usually of practical interest, as typically the network has to be designed such that $p_n(d)$ is small.

Clearly, the decay rate of Lemma 3.5 is also a lower bound on $J(d)$ in case $\phi_2 \in (\mu_2/c, \phi_2^F)$, as it is independent of $\phi_2$ (see proof of Lemma 3.5). However, the corresponding path $f^*$ is not necessarily contained in $L$, and therefore it is not known whether the bound is tight.

We proceed by presenting a second lower bound on $J(d)$.

**Lemma 3.6** *$J(d)$ is lower bounded by*

$$\inf_{x \geq 0} \inf_{t \geq 0} \sup_{s \in [0,t]} \inf_{v \in [0,s]} \inf_{\substack{z_1 \geq x + ct + \phi_1 cd \\ z_2 \leq x + cs - \phi_1 cv}} \Lambda(z_1, z_2),$$

*where*

$$\Lambda(z_1, z_2) := \frac{1}{2} \begin{pmatrix} z_1 - (\mu_1 + \mu_2)t \\ z_2 - (\mu_1 + \mu_2)s + \mu_2 v \end{pmatrix}^{\mathrm{T}}$$
$$\begin{pmatrix} v_1(t) + v_2(t) & \Gamma_1(s,t) - \Gamma_1(v,t) + \Gamma_2(s,t) \\ \Gamma_1(s,t) - \Gamma_1(v,t) + \Gamma_2(s,t) & v_1(s-v) + v_2(s) \end{pmatrix}^{-1} \begin{pmatrix} z_1 - (\mu_1 + \mu_2)t \\ z_2 - (\mu_1 + \mu_2)s + \mu_2 v \end{pmatrix}.$$

*Proof:* Let the exact decay rate of the upper bound in Lemma 2.3 be denoted by $J^L(d)$. Define the set of paths

$$S^{s,t,v,x} := \left\{ \overline{f} \in \Omega \,\middle|\, \begin{array}{l} -f_1(-t) - f_2(-t) \geq x + ct + \phi_1 cd; \\ f_1(-v) - f_1(-s) - f_2(-s) \leq x + cs - \phi_1 cv \end{array} \right\},$$

where

$$S := \bigcup_{x\geq 0}\bigcup_{t\geq 0}\bigcap_{s\in[0,t]}\bigcup_{v\in[0,s]} S^{s,t,v,x},$$

and $\overline{f}(t) = (\overline{f}_1(t), \overline{f}_2(t)) := (f_1(t) - \mu_1 t, f_2(t) - \mu_2 t)$ is the centered path. Then using Lemma 2.3 and 'Schilder' (recall that Schilder's theorem relates to *centered* Gaussian inputs), we obtain that

$$J(d) \geq \inf_{f\in S} I(\overline{f}) = J^L(d) \geq \inf_{x\geq 0}\inf_{t\geq 0}\sup_{s\in[0,t]}\inf_{v\in[0,s]}\inf_{f\in S^{s,t,v,x}} I(\overline{f}).$$

The last inequality above was given in Theorem 4.1 of [10]. We now focus on the calculation of $\inf_{f\in S^{s,t,v,x}} I(\overline{f})$ for fixed $s, t, v$ and $x$. Recognize that $\Lambda(z_1, z_2)$ is the large-deviations rate function of the bivariate random variable $(A_1(-t,0) + A_2(-t,0), A_1(-s,-v) + A_2(-s,0))$. Finally, using Theorem 3.2,

$$\inf_{f\in S^{s,t,v,x}} I(\overline{f}) = \inf_{\substack{z_1\geq x+ct+\phi_1 cd\\ z_2\leq x+cs-\phi_1 cv}} \Lambda(z_1, z_2).$$

This proves the stated. □

The following theorem summarizes Lemmas 3.4-3.6.

**Theorem 3.7** *Suppose that class-1 and class-2 sources are Gaussian inputs. Then, under Assumptions A1-A3,*

$$J(d) = \begin{cases} (i) & \inf_{t\geq 0} \frac{(\phi_1 cd + (\phi_1 c - \mu_1)t)^2}{2v_1(t)} & \text{for } \phi_2 \in [0, \mu_2/c]; \\ (iii) & \inf_{t\geq 0} \frac{(c(t+d) - \mu_1 t - \mu_2(t+d))^2}{2v_1(t) + 2v_2(t+d)} & \text{for } \phi_2 \in [\phi_2^F, 1], \end{cases}$$

*and*

$$(ii) \quad J(d) \geq \max\left\{ \inf_{t\geq 0} \frac{(c(t+d) - \mu_1 t - \mu_2(t+d))^2}{2v_1(t) + 2v_2(t+d)}, \inf_{x\geq 0}\inf_{t\geq 0}\sup_{s\in[0,t]}\inf_{v\in[0,s]}\inf_{\substack{z_1\geq x+ct+\phi_1 cd\\ z_2\leq x+cs-\phi_1 cv}} \Lambda(z_1, z_2) \right\}$$

*for $\phi_2 \in (\mu_2/c, \phi_2^F)$, where $\Lambda(z_1, z_2)$ is as in Lemma 3.6.*

### 3.4 Brownian inputs

For most Gaussian inputs that satisfy A1-A3 there does not exist a closed-form expression for (bounds on) the decay rates as presented in Theorem 3.7. In case of Brownian inputs, however, we can derive explicit expressions for (bounds on) the decay rate $J(d)$. Brownian motions can be used to approximate weakly-dependent traffic streams as suggested by the celebrated Central Limit Theorem in functional form, see also [8]. We let the variance functions be characterized through $v_i(t) = \lambda_i t$, with $\lambda_i > 0$, $i = 1, 2$.

Straightforward calculus shows that (14) is minimized for

$$t^* = \begin{cases} d\left(\frac{c-\mu_2}{c-\mu_1-\mu_2} - 2\frac{\lambda_2}{\lambda_1+\lambda_2}\right) & \text{if } \frac{\mu_2}{c} + \frac{2\lambda_2}{\lambda_1+\lambda_2}\left(1 - \frac{\mu_1+\mu_2}{c}\right) \leq 1; \\ 0 & \text{otherwise.} \end{cases}$$

Since $v_i'(t) = \lambda_i$, we obtain from (13) that

$$\phi_2^F = \min\left\{\frac{\mu_2}{c} + \frac{2\lambda_2}{\lambda_1+\lambda_2}\left(1 - \frac{\mu_1+\mu_2}{c}\right), 1\right\}$$

The following theorem characterizes the decay rate $J(d)$.

**Proposition 3.8** *Suppose that class-1 and class-2 sources are Brownian inputs. Then,*

$$J(d) = \begin{cases} (i) \quad 2\phi_1 c d \frac{\phi_1 c - \mu_1}{\lambda_1} & \text{for } \phi_2 \in [0, \mu_2/c]; \\ (iii) \quad 2d\left(\frac{c-\mu_1-\mu_2}{\lambda_1+\lambda_2}\right)\left(\frac{(c-\mu_2)\lambda_1+\mu_1\lambda_2}{\lambda_1+\lambda_2}\right) & \text{for } \phi_2 \in [\phi_2^F, 1], \end{cases}$$

*and*

$$(ii) \quad J(d) \in \left[\frac{1}{2}d\left(\frac{(\phi_1 c + (\phi_1 c - \mu_1)u^*)^2}{\lambda_1 u^*} + \frac{(\phi_2 c - \mu_2)^2}{\lambda_2}u^*\right), \frac{1}{2}d\left(\frac{(\phi_1 c + (\phi_1 c - \mu_1)u^*)^2}{\lambda_1 u^*} + \frac{(\phi_2 c - \mu_2)^2}{\lambda_2}(u^*+1)\right)\right]$$

*for $\phi_2 \in (\mu_2/c, \phi_2^F)$, with the 'critical time scale' $u^*$ given by*

$$u^* := \frac{\phi_1 c}{\sqrt{(\phi_1 c - \mu_1)^2 + (\phi_2 c - \mu_2)^2 \frac{\lambda_1}{\lambda_2}}}.$$

*Proof:* Straightforward calculus shows that the optimizer of Lemma 3.4 is

$$t^* = \frac{\phi_1 c d}{\phi_1 c - \mu_1},$$

from which we obtain $(i)$. In Lemma 3.5 the optimizer is

$$t^* = d\left(\frac{c-\mu_2}{c-\mu_1-\mu_2} - 2\frac{\lambda_2}{\lambda_1+\lambda_2}\right),$$

which yields $(iii)$. The lower bound in $(ii)$ follows from Lemma 3.6, and was proved in Theorem 5.6 of [8]. The upper bound in $(ii)$ is a matter of calculating the costs of a path in $L$. Consider the following path:

$$f_1(r) = \begin{cases} -\mathbb{E}(A_1(r,0) | A_1(-t^*,0) = \phi_1 c(t^*+d)) & \text{for } r \le 0; \\ \mathbb{E}(A_1(0,r) | A_1(-t^*,0) = \phi_1 c(t^*+d)) & \text{for } r > 0. \end{cases}$$

$$f_2(r) = \begin{cases} -\mathbb{E}(A_2(r,0) | A_2(-t^*,d) = \phi_2(t^*+d)) & \text{for } r \le 0; \\ \mathbb{E}(A_2(0,r) | A_2(-t^*,d) = \phi_2 c(t^*+d)) & \text{for } r > 0. \end{cases}$$

where $t^* = u^* d$. Using (12), it can be verified that this path is such that class 1 produces traffic at constant rate $\phi_1 c(t^*+d)/t^* > \phi_1 c$ in the interval $(-t^*, 0]$ and at constant rate $\mu_1$ elsewhere, whereas class 2 produces traffic at constant rate $\phi_2 c$ in the interval $(-t^*, d]$ and at constant rate $\mu_2$ elsewhere. This obviously leads to $Q_1(0) = \phi_1 c d$ (as $Q_1(-t^*) = Q_2(-t^*) = 0$), and thus a delay of $d$, as class 2 continuously claims its guaranteed rate in the interval $(-t^*, d]$. In case of standard Brownian inputs it is well-known that (9) can be rewritten as (Thm. 5.2.3 of [5])

$$I(\omega) = \frac{1}{2}\int (\omega_1'(t))^2 \mathrm{d}t + \frac{1}{2}\int (\omega_2'(t))^2 \mathrm{d}t. \tag{18}$$

Using (18), the decay rate associated with $f$ is therefore given by $I(\tilde{f}_1, \tilde{f}_2)$, with $\tilde{f}_i(t) := (f_i(t) - \mu_i t)/\sqrt{\lambda_i}$, $i = 1, 2$ (recall that (18) relates to *standard* Brownian inputs), which is equivalent to the desired expression. □

## 4 Conclusion

In this note we analyzed the delay in a two-class GPS system. We derived bounds on the probability that the virtual delay experienced by a packet ('fluid molecule') of a particular class exceeds some threshold. Assuming Gaussian input traffic, we applied Schilder's sample-path large deviations theorem to calculate the logarithmic asymptotics of these bounds. We showed that for certain weights in the GPS system the bounds are tight. In the special case of Brownian inputs we obtained closed-form expressions.

Future research directions include the derivation of the decay rate of the *joint* overflow probability: $\mathbb{P}(Q_{1,n} > nb_1, Q_{2,n} > nb_2)$, in particular in case of Brownian inputs.

## References


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

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

[3] D. Bertsimas, I. Paschalidis, J. Tsitsiklis (1999). Large deviations analysis of the Generalized Processor Sharing policy. *Queueing Systems*, 32: 319–349.

[4] S. Borst, M. Mandjes, M. van Uitert (2003). Generalized Processor Sharing queues with light-tailed and heavy-tailed input. *IEEE/ACM Transactions on Networking*, 11: 821–834.

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

[6] J.-D. Deuschel, D. Stroock (1989). Large deviations. Academic Press, London.

[7] F. Guillemin, D. Pinchon (2004). Analysis of Generalized Processor-Sharing systems with two classes of customers and exponential services. *Journal of Applied Probability*, 41: 832-858.

[8] M. Mandjes (2005). Large deviations for complex buffer architectures: the short-range dependent case. *Stochastic Models*, 22: 99–128.

[9] M. Mandjes, M. van Uitert (2005). Sample-path large deviations for tandem and priority queues with Gaussian inputs. *Annals of Applied Probability*, 15: 1193–1226.

[10] M. Mandjes, M. van Uitert (2005). Sample-path large deviations for Generalized Processor Sharing queues with Gaussian inputs. *Performance Evaluation*, 61: 225–256.

[11] P. Mannersalo, I. Norros (2002). GPS schedulers and Gaussian traffic. In: *Proc. IEEE Infocom 2002*, New York, USA, 1660–1667.

[12] L. Massoulié (1999). Large deviations estimates for polling and weighted fair queueing service systems. *Advances in Performance Analysis*, 2: 103–128.

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

[14] I. Paschalidis (1999). Class-specific Quality-of-Service guarantees in multimedia communication networks. *Automatica* (Special issue on control methods for communication networks; V. Anantharam and J. Walrand, Eds.), 35: 1951–1969.

[15] E. Reich (1958). On the integrodifferential equation of Takács I. *Annals of Mathematical Statistics*, 29: 563–570.

[16] Z.-L. Zhang (1997). Large deviations and the generalized processor sharing scheduling for a two-queue system. *Queueing Systems*, 26: 229–245.

[17] Z.-L. Zhang (1998). Large deviations and the Generalized Processor Sharing scheduling for a multiple-queue system. *Queueing Systems*, 28: 349–376.