On the reflected fractional Brownian motion process on the positive orthant: asymptotics for a maximum with application to queueing networks

ROSARIO DELGADO,* *Universitat Autònoma de Barcelona*

Abstract

Let W be a $J-$dim. reflected fractional Brownian motion process (rfBm) on the positive orthant $\mathbb{R}^J_+$, with drift $\theta \in \mathbb{R}^J$ and Hurst parameter $H \in (0,\, 1)$, and let $a \in \mathbb{R}^J_+$, $a \neq 0$, be a vector of weights. We define $M(t) = \max_{0\leq s\leq t} a^T W(s)$ and prove that $M(t)$ grows like t if $\mu = a^T \theta > 0$, in the sense that its increase is smaller than that of any function growing faster than t, and if a restriction on the weights holds, it is also bigger than that of any function growing slower than t. We obtain similar results with t^H instead of t in the drift-less case ($\theta = 0$). If $\mu < 0$ we prove that the increase of $M(t)$ is smaller than that of any function growing faster than t, and also that $(\log t)^{\frac{1}{2(1-H)}}$ is a lower bound for $M(t)$. Motivation for this study is that rfBm appears as the workload limit associated to a fluid queueing network fed by a big number of heavy-tailed On/Off sources under *heavy traffic* and *state space collapse*; in this scenario, $M(t)$ can be interpreted as the maximum amount of fluid in queue at the network over the interval $[0,\, t]$, which turns out to be an interesting performance process to describe the congestion of the queueing system.

Keywords: reflected fractional Brownian motion, maximum, workload, fluid queue, heavy traffic, state space collapse

AMS 2000 Subject Classification: Primary 60G70

Secondary 60G15;60G18;60K25

* Postal address: Departament de Matemàtiques. Universitat Autònoma de Barcelona. Edifici C- Campus de la UAB. 08193 Bellaterra (Cerdanyola del Vallès)- Barcelona, Spain

* Email address: delgado@mat.uab.cat

* Partially supported by: project MEC-FEDER ref. MTM2006-06427

1 Introduction

Up to now fractional Brownian motion (fBm) seems to be a very generally accepted theoretical model for traffic in modern broadband and high-speed networks, which usually exhibits *long-range dependence* and *self-similarity*. Because queueing performance of this type of traffic is substantially different from that of traditional short-range dependent traffic, it is a challenge to analyze it.

By focusing on the asymptotic behavior of a single-server queue fed by a fBm, tail probabilities of the steady-state workload process, which is a one-dimensional fBm process reflected conveniently to be non-negative, have been studied by Norros [12] and Duffield and O'Connell [5]. Their results were improved later by giving exact tail asymptotics by Massoulie and Simonian [11] and Hüsler and Piterbarg [8], [9]. Following the same line of research, Zeevi and Glynn [15] give more results for a deeper understanding of this kind of single queue, related to the behavior of the maximum workload (or fluid in queue, equivalently) over the interval $[0,\, t]$ when $t \to +\infty$. In particular, they show that under heavy traffic, this maximum behaves like t^H, where $H \in (1/2,\, 1)$ is the Hurst parameter of the drift-less fBm process that feeds the queue, while if the queue is stable, the maximum grows like $(\log t)^{\frac{1}{2(1-H)}}$. Asymptotics for such a queue have also been considered in Duncan et al. [6], where applications to ATM broadband connections are studied.

Is it still true that the maximum fluid in queue over the interval $[0,\, t]$ grows in some sense like t^H when $t \to +\infty$, in the drift-less case, if instead of a one-dimensional we consider a multi-dimensional reflected fractional Brownian motion (rfBm) process W on the positive orthant, with drift $\theta \in \mathbb{R}^J$, reflection matrix R and Hurst parameter H ? And what happens in the case of non-zero drift θ ? The present work has been motivated by these questions.

If we define

$$M(t) \stackrel{\text{def}}{=} \max_{0 \le s \le t} \sum_{j=1}^{J} a_j\, W_j(s) = \max_{0 \le s \le t} a^T\, W(s)\,, \tag{1.1}$$

with $a = (a_1, \ldots, a_J)^T \ge 0$, $a \ne 0$, it happens that $M(t)$ can be interpreted as the total maximum fluid in queue over the interval $[0,\, t]$ for a multi-class multi-server fluid queueing network (we give more details below). In Section 3 we prove that for arbitrary $H \in (0,\, 1)$, the increase of $M(t)$ as $t \to +\infty$, is closer to that of t if $\mu \stackrel{\text{def}}{=} a^T\, \theta > 0$, in the sense that it is smaller than that of any function growing faster than t (Theorem 1), and that if a restriction

on the weights a holds, we prove the tightness of this result in the sense that the increase of $M(t)$ is bigger than that of any function growing slower than t (Theorem 2). This restriction is denoted by **(HaR)**. In the drift-less case $\theta = 0$ we obtain similar results but with t^H instead of t. Moreover, we also consider the case $\mu = a^T \theta < 0$, which includes but is not restricted to, the negative drift case $\theta < 0$, and Theorems 1 and 2 give asymptotic lower and upper bounds for $M(t)$: $(\log t)^{\frac{1}{2(1-H)}}$ and any function growing faster than t, respectively.

Asymptotics for the maximum of stochastic processes in general is an important research area with a long tradition in literature, especially for the case of Gaussian processes. That notwithstanding, due to the difficulty given by the non-Markovian character of the fBm process if $H \neq \frac{1}{2}$, and by the lack of Gaussianity of the reflected process, there is only a few sets of results about the (one-dimensional) rfBm process. See for instance the aforementioned works on the workload process associated to a single server queue. Up to our knowledge, ours is the first attempt to obtain some results when dealing with a multi-dimensional rfBm process, with or without a (linear) drift. In order to carry on that, it seemed natural to try out the *oscillation inequality* for $R-$regularizations or solutions of the Skorokhod problem given in [2] to take advantage of the Gaussian character of the fBm process, and this is what, in fact, we have done.

The study of the maximum of a one-dimensional rfBm process with Hurst parameter $H > 1/2$ in Zeevi and Glynn [15] is motivated by the fact that this process appears as the workload (equivalently, as the fluid in queue) process of a single-server queue fed by a fBm process. Here we study a maximum associated to a multi-dimensional rfBm process motivated by application to queueing networks. Indeed, whenever $H > 1/2$ this maximum appears as the total fluid limit in queue for a multi-class multi-server fluid network which generalizes that analyzed in [15] and that can be described briefly as follows: the network can process K fluid classes by using J stations which have a single server and an infinite buffer at each one, with $K \geq J \geq 1$. Each server can process one or more fluid classes but each fluid class can be processed at only one station. By following [13] we assume that the process of external arrivals is generated by a large enough number of heavy-tailed ON/OFF sources. Moreover *feedback* is allowed and a FIFO (first-in-first-out) and *non-idling* service discipline is used. Under some assumptions mainly including *heavy traffic* (or asymptotical criticality) and *state space collapse*, it is proved in [4] that the limits of the ($J-$dim.) workload and ($K-$dim.) fluid in queue processes, conveniently scaled, exist, and that the workload limit

is a rfBm process on the positive orthant S with drift vector $\theta = 0$, some Hurst parameter $H \in (1/2, 1)$ and some reflection matrix R. The results from [4] will be quoted in more detail in Section 4.1. *State space collapse* condition establishes that workload and fluid in queue processes are related by means of a *"lifting"* matrix Δ, which depends on the service discipline and allows the recovery of fluid of each class in queue from the workload at the station at which that class is processed, even if two or more fluid classes are processed at the same station; if Z and W denote respectively the limit of the fluid in queue and the workload processes, then *state space collapse* says that $Z = \Delta W$. Let us define

$$M(t) \stackrel{\text{def}}{=} \max_{0 \le s \le t} \sum_{k=1}^{K} Z_k(s) \tag{1.2}$$

to be the maximum (total) amount of fluid in queue in the system (by summing the fluid in queue over all fluid classes), on the interval $[0, t]$. Therefore, $M(t)$ can be rewritten as (1.1) for some weights $a_j > 0$ which depend on the lifting matrix Δ, and W turns out to be a rfBm process.

The organization of the paper is as follows: precise definitions, notations and terminology will be introduced in the next section, where we also quote a technical result to be used in Section 3, where our results and proofs are stated: Theorems 1 and 2, which give the upper and lower bounds for $M(t)$, respectively, for arbitrary $H \in (0, 1)$. Finally, Section 4 deals with the multi-class multi-server fluid network to which we can apply our results, giving the motivation for our definition of the maximum process associated to the rfBm when $H > 1/2$, and some examples of such a network are studied.

2 Preliminaries

We will denote by I the identity matrix and by e the vector $(1, \ldots, 1)^T$ (irrespective of their dimension). Vectors will be column vectors and v^T means the transpose of a vector (or a matrix) v. By $diag(v)$ we denote the diagonal matrix with diagonal elements the components of vector v (in the same order). Inequalities for vectors must be understood in the componentwise sense. For any fixed $J \ge 1$, let S denote the positive orthant $\mathbb{R}^J_+ = \{v = (v_1, \ldots, v_J)^T \in \mathbb{R}^J \ : \ v_j \ge 0 \ \forall \ j = 1, \ldots, J \ (\text{i.e. } v \ge 0)\}$. For any $v \in \mathbb{R}^J$, we use the notation $|v| \stackrel{\text{def}}{=} \sum_{j=1}^{J} |v_j|$ (where $|x|$ denotes the absolute value of $x \in \mathbb{R}$). For a non-negative

real number x, $[x]$ denotes the maximum integer less or equal to x (the integer part of x).

P-lim denotes, as usual, the convergence in the probability sense. Φ stands for the standard Gaussian distribution function, that is, $\Phi(x) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}}\, dy$ for any $x \in \mathbb{R}$.

For any real number $t \geq 0$ and any continuous function $f\colon [0,\, \infty) \to \mathbb{R}^J$, we define

$$||f(\cdot)||_t \stackrel{\text{def}}{=} \max_{0\leq s\leq t} \Big(\sum_{\ell=1}^{J} |f_\ell(s)| \Big)\,,$$

and introduce the notion of *oscillation* of function f by

$$Osc\,\big(\, f(\cdot),\, [0,\, t]\,\big) \stackrel{\text{def}}{=} \max_{0\leq s\leq r\leq t} \big(\sum_{\ell=1}^{J} |f_\ell(r) - f_\ell(s)|\,\big)\,.$$

Note that in general

$$Osc\,\big(\, f(\cdot),\, [0,\, t]\,\big) \leq 2\,||f(\cdot)||_t\,, \qquad \text{and} \tag{2.1}$$

$$Osc\,\big(\, f(\cdot),\, [0,\, t]\,\big) \geq ||f(\cdot)||_t \;\text{ if } f(0) = 0 \text{ and } f \in S\,. \tag{2.2}$$

Loosely speaking, the reflected fractional Brownian motion (rfBm) process behaves like a fractional Brownian motion (fBm) in the interior of the first orthant S, and is confined to that orthant by instantaneous "reflection" at the boundary. Although it is known in the literature, for the convenience of the reader we set here its definition as stated in [3], in order to establish notations which will be used throughout the work. Previously we recall the definition of the fBm process.

Definition 1 (The fractional Brownian motion) A stochastic process $B^H = \{B^H(t) = (B_1^H(t), \ldots, B_J^H(t))^T,\, t \geq 0\}$, defined on some probability space, is called a (J-dimensional) *fractional Brownian motion (fBm)* of (Hurst) parameter $H \in (0,\, 1)$, starting from $x \in \mathbb{R}^J$, with *drift vector* $\theta \in \mathbb{R}^J$ and with *associated matrix* Γ (a $J \times J$ positive definite matrix), if it is a continuous Gaussian process starting from x, with $E\big(B^H(t)\big) = x + \theta\, t$ for any $t \geq 0$, and with covariance function given by

$$Cov\big(B^H(t), B^H(s)\big) = E\Big(\big(B^H(t) - (x + \theta\, t)\big)\big(B^H(s) - (x + \theta\, s)\big)^T\Big) = \Gamma_H(s,t)\,\Gamma,$$

for any $t,\, s \geq 0$, where $\Gamma_H(s,t) = \dfrac{1}{2}\,\big(\, t^{2H} + s^{2H} - |t-s|^{2H}\,\big)$. For short, we will say that B^H is a $J-$dimensional fBm *with associated data* (x, H, θ, Γ).

Definition 2 (The reflected fractional Brownian motion) A *reflected fractional Brownian motion (rfBm)* on S with associated data (x, H, θ, Γ, R), where x, $\theta \in S$, $H \in (0, 1)$ and Γ and R are $J \times J$ matrices, Γ being a positive definite one, is a $J-$dimensional process $W = \{W(t), t \geq 0\}$ defined on some probability space, with $W(t) = (W_1(t), \ldots, W_J(t))^T$, such that

(i) W has continuous paths and $W(t) \in S$ for all $t \geq 0$, a.s.,

(ii) $W = X + RY$ a.s., with X and Y two $J-$dimensional processes defined on the same probability space verifying:

(iii) X is a $J-$dimensional fBm with associated data (x, H, θ, Γ),

(iv) Y has continuous and non-decreasing paths, and for each $j = 1, \ldots, J$, a.s., $Y_j(0) = 0$ and $\int_0^\infty 1_{\{W_j(s)>0\}}\, dY_j(s) = 0$ (that means that Y_j can only increase when W is on face $F_j = \{y \in S : y_j = 0\}$).

It is also said that the pair (W, Y) is a $R-$*regularization* of X or that (W, Y) is a solution of the $R-$*regularization problem* of X. Note that by definition, $W(0) = X(0) = x$. In the *zero-delayed* case, $x = 0$, and the drift-less case corresponds to $\theta = 0$.

Note that for any $j = 1, \ldots, J$, the direction of the reflection on the jth face F_j is given by the jth column of the *reflection matrix* R. The *completely-S* property of matrix R is sufficient for the existence of the $R-$*regularization* of X, as can be seen in Theorem 2 [2]. But as is pointed out in the remark following that result, this property cannot ensure the adaptness of process Y to a filtration to which X is adapted. Nevertheless, Proposition 4.2 [14] shows that under a stronger assumption on R, that we quote below and name **(HR)**, this problem is overcome (for more details we refer the interested reader to Section 2 of [3]).

(HR) Assumption on matrix R:

$$R \text{ can be expressed as } I + \Theta, \quad \text{with } \Theta \text{ a } J \times J \text{ matrix such that } \langle\Theta\rangle \text{ has spectral radius strictly less than } 1,$$

where given a matrix A, $\langle A\rangle$ stands for the matrix obtained from A by replacing all the elements of A by their absolute values.

For the sake of completeness, we quote here a technical result that will be used in the proof of Theorem 2.

Theorem 4.3.3. in [10] (part i)

Let $\{\xi_n\}$ be a standardized stationary normal sequence with covariances $\{r_n\}$ satisfying condition $r_n \log n \to 0$. Then, for $0 \leq \tau \leq +\infty$, if $M_n = \max_{i=1,\ldots,n} \xi_i$, for any sequence of constants $\{u_n\}$ we have that

$$P\left(M_n \leq u_n\right) \to e^{-\tau} \Leftrightarrow n\left(1 - \Phi(u_n)\right) \to \tau\,. \tag{2.3}$$

3 Main results

As explained in the introduction, in this section we study the asymptotic behavior as $t \to +\infty$ of the maximum process given by formula (1.1), that is,

$$M(t) = \max_{0 \leq s \leq t} a^T\, W(s)$$

for any vector $a = (a_1, \ldots, a_J)^T \geq 0$, $a \neq 0$, W being any zero-delayed rfBm process $W = X + RY$ on S with associated data ($x = 0$, H, θ, Γ, R), where $\theta \in \mathbb{R}^J$, $H \in (0, 1)$, Γ is a positive definite $J \times J$ matrix, and R is a $J \times J$ matrix verifying assumption **(HR)**. Motivation for the study of this process when $H > 1/2$ is explained in the Introduction and more detailed in Section 4.1.

Remark 1 It the drift-less case $\theta = 0$, the distribution of the maximum $M(t)$ can be obtained directly by self-similarity, as is done in Proposition 2 in [15] in the one-dimensional case. The argument is as follows: if (W, Y) is a $R-$regularization of X, and for an arbitrary process A we define a new process $\tilde{A}$ by $\tilde{A}(\cdot) \stackrel{\text{def}}{=} A(r\,\cdot)/r^H$, then $(\tilde{W}, \tilde{Y})$ is a $R-$regularization of process $\tilde{X}$. We take into account that when $\theta = 0$, $X \stackrel{\mathcal{D}}{=} \tilde{X}$, and as a consequence of the uniqueness of the solution of the Skorokhod problem under assumption **(HR)**, $W \stackrel{\mathcal{D}}{=} \tilde{W}$ (see the discussion after formula (1) in [3]). Thus

$$M(t) = \max_{0 \leq s \leq t} a^T\, W(s) \stackrel{\mathcal{D}}{=} t^H\, M(1)\,,$$

which immediately implies Theorems 1 and 2 for the case $\theta = 0$. Our results are established in a rather general setting: we only need assumption **(HR)** for Theorem 1, and an additional assumption denoted by **(HaR)** for Theorem 2. The obtention of t^H as lower bound in Theorem 2 in the drift-less case needs assumption **(HaR)** but is done in fact for the more general case $\mu = 0$. We point out the difficulty of working in the multi-dimensional setting,

due to the lack of a explicit expression for the pushing process Y in terms of the free process X, which does exist in the one-dimensional case and facilitates work, making assumption **(HaR)** unnecessary. In the examples of Section 4 we find restrictions on the parameters of the model to ensure **(HR)** in such cases in which it is not trivially accomplished.

Theorem 1 below asserts that when $t \to +\infty$, $M(t)$ cannot grow too much, in the sense that $M(t)$ grows less than any function $f(t)$ growing faster than t^H or t, depending on if the drift vector θ equals zero or not, respectively, independently of the weights a. Thus, Theorem 1 gives an asymptotic upper bound for the maximum.

Theorem 1 (Asymptotic upper bound) *Under assumption* **(HR)**,

$$\text{P-}\lim_{t\to+\infty} \frac{M(t)}{f(t)} = 0$$

for any positive real function f such that

$$\begin{cases} \lim\limits_{t\to+\infty} \dfrac{f(t)}{t^H} = +\infty & \text{if} \quad \theta = 0 \\ \lim\limits_{t\to+\infty} \dfrac{f(t)}{t} = +\infty & \text{otherwise.} \end{cases}$$

Furthermore, this convergence to zero in the probability sense is, in fact, convergence in L^p for any $p \geq 1$.

Proof: We split the proof into two main steps, corresponding to the convergence in the probability and in the L^p sense, respectively.

Step 1: To prove the convergence in the probability sense, the main idea is to apply the *oscillation inequality* for $R-$regularizations or solutions of the Skorokhod problem given by Bernard and el Kharroubi [2], which provides a way to relate lower boundness condition on $M(t)$ expressed in terms of the rfBm process W, with a similar condition for the maximum of process X, which is a more tractable process since X is a fBm process (in particular, it is Gaussian).

Indeed, for any $t \geq 0$, we have that

$$M(t) = \max_{0\leq s\leq t} a^T W(s) \leq |a|\, ||W(\cdot)||_t \,. \tag{3.1}$$

Taking into account that (W, Y) is a $R-$regularization of X, and that R is a *Completely-S* matrix (consequence of **(HR)**), we can apply the *oscillation inequality* given in Lemma 1 [2],

to obtain that a positive constant only depending on R (and not on t), say K'_R, exists such that for any $t > 0$,

$$Osc\big(W(\cdot),\,[0,\,t]\big) \le K'_R\, Osc\big(X(\cdot),\,[0,\,t]\big)\,.$$

Note that the fact that constant K'_R does not depend on t is a main ingredient for the proof, and this can be deduced by induction from the proof of Lemma 1 in [2]. A generalization of this result, Theorem 5.1. in [14], claims explicitly that the constant only depends on R, and this can be checked by following the induction steps of its proof.

By (2.1) and (2.2), by using that $W(0) = 0$ and $W \ge 0$, it follows with $K_R = 2\,K'_R$ that

$$||W(\cdot)||_t \le K_R\,||X(\cdot)||_t \le K_R \sum_{j=1}^{J} \max_{0\le s\le t} |X_j(s)|\,.$$

Combining this inequality with (3.1) we obtain that

$$M(t) \le K_{R,a} \sum_{j=1}^{J} \max_{0\le s\le t} |X_j(s)| \quad \text{for any } \; t \ge 0\,, \tag{3.2}$$

with some positive constant $K_{R,a}$ (only depending on R and a).

We are now in position to show that for any $\varepsilon > 0$ and any function f as in the statement of the theorem,

$$\lim_{t\to+\infty} P\Big(\frac{M(t)}{f(t)} \ge \varepsilon\Big) = 0\,. \tag{3.3}$$

In order to get this limit it is convenient to find an adequate upper bound for the probability, which is obtained from (3.2) in this way:

$$P\Big(\frac{M(t)}{f(t)} \ge \varepsilon\Big) = P\Big(M(t) \ge \varepsilon\, f(t)\Big) \le P\Big(\sum_{j=1}^{J} \max_{0\le s\le t} |X_j(s)| \ge \frac{\varepsilon}{K_{R,a}}\, f(t)\Big) \le$$

$$P\Big(\bigcup_{j=1}^{J} \{\max_{0\le s\le t} |X_j(s)| \ge \frac{\varepsilon}{K_{R,a}\,J} f(t)\}\Big) \le \sum_{j=1}^{J} P\Big(\max_{0\le s\le t} |X_j(s)| \ge \frac{\varepsilon}{K_{R,a}\,J} f(t)\Big). \tag{3.4}$$

The proof of (3.3) will be completed therefore by showing that

$$\lim_{t\to+\infty} P\Big(\max_{0\le s\le t} |X_j(s)| \ge \frac{\varepsilon}{K_{R,a}\,J}\, f(t)\Big) = 0 \quad \text{for any} \quad j = 1, \ldots, J\,. \tag{3.5}$$

In order to get (3.5), it is convenient to consider the random variable $X_j(s) \sim N(\theta_j\, s,\, s^{2H}\,\Gamma_{jj})$ for any fixed j and s, and define

$$\Psi_j(s) \stackrel{\text{def}}{=} X_j(s) - \theta_j\, s \sim N(0,\, s^{2H}\,\Gamma_{jj})\,.$$

Since $\Psi_j(r\,t)$ has the same law as $t^H\,\Psi_j(r)$ (self-similarity property), we can bound $\max\limits_{0\le s\le t}|X_j(s)|$ in law by $t^H \max\limits_{0\le r\le 1}|\Psi_j(r)| + \theta_j\,t$, and therefore

$$P\Big(\max_{0\le s\le t}|X_j(s)| \ge \frac{\varepsilon}{K_{R,a}\,J}\,f(t)\Big) \le P\Big(\max_{0\le r\le 1}|\Psi_j(r)| \ge \lambda_j(\varepsilon,\,t)\Big) \tag{3.6}$$

with

$$\lambda_j(\varepsilon,\,t) = \frac{\varepsilon}{K_{R,a}\,J}\,\frac{f(t)}{t^H} - |\theta_j|\,t^{1-H} = t^{1-H}\Big(\frac{\varepsilon}{K_{R,a}\,J}\,\frac{f(t)}{t} - |\theta_j|\Big), \tag{3.7}$$

which increases to $+\infty$ when $t\to+\infty$ for any fixed $\varepsilon>0$, by the assumptions on f.

Consequently, (3.5) follows if we prove

$$\lim_{t\to+\infty} P\Big(\max_{0\le r\le 1}|\Psi_j(r)| \ge \lambda_j(\varepsilon,\,t)\Big) = 0\,, \tag{3.8}$$

but this immediately follows after noting that $\max\limits_{0\le r\le 1}|\Psi_j(r)|$ is a finite random variable by continuity of $\Psi_j(\cdot)$, since $\lim\limits_{t\to+\infty}\lambda_j(\varepsilon,\,t) = +\infty$.

Step 2: L^p-convergence follows by checking the usual sufficient condition for the uniform integrability:

$$\sup_t\, E\Big(\big(\frac{M(t)}{f(t)}\big)^{p+1}\Big) < +\infty \quad \text{for any } p\ge 1\,. \tag{3.9}$$

We now proceed to show (3.9). By (3.4) and (3.6) we have that for any $y>0$,

$$P\Big(\frac{M(t)}{f(t)}\ge y\Big) \le \sum_{j=1}^{J} P\Big(\max_{0\le r\le 1}|\Psi_j(r)| \ge \lambda_j(y,\,t)\Big)$$

with $\lambda_j(y,\,t)$ given by (3.7) (note that for any fixed t, $\lim\limits_{y\to+\infty}\lambda_j(y,\,t)=+\infty$, and that for any fixed y, $\lim\limits_{t\to+\infty}\lambda_j(y,\,t)=+\infty$).

We can bound the expectation in (3.9) by using that $\lambda_j(y,\,t)$ can be expressed as $A_j(t)\,y - B_j(t)$, with $A_j(t) = \dfrac{1}{K_{R,a}\,J}\,\dfrac{f(t)}{t^H}$ and $B_j(t) = |\theta_j|\,t^{1-H}$, in this way:

$$\begin{aligned}
E\Big(\big(\frac{M(t)}{f(t)}\big)^{p+1}\Big) &= \int_0^{+\infty}(p+1)\,y^p\,P\Big(\frac{M(t)}{f(t)}\ge y\Big)\,dy \\
&\le \sum_{j=1}^{J}\int_0^{+\infty}(p+1)\,y^p\,P\Big(\max_{0\le r\le 1}|\Psi_j(r)| \ge A_j(t)\,y - B_j(t)\Big)\,dy \\
&= \sum_{j=1}^{J} E\Big(\Big(\frac{\max_{0\le r\le 1}|\Psi_j(r)| + B_j(t)}{A_j(t)}\Big)^{p+1}\Big) \\
&= \sum_{j=1}^{J}\frac{1}{(A_j(t))^{p+1}}\,E\Big(\Big(\max_{0\le r\le 1}|\Psi_j(r)| + B_j(t)\Big)^{p+1}\Big)\,.
\end{aligned} \tag{3.10}$$

If $\theta_j = 0$, $B_j(t) = 0$ for any t and the expectation in the corresponding summand of expression (3.10) is bounded (by Borell's isoperimetric inequality; a version of this result can be found in Theorem 21, pag. 43 [1]). Moreover, on account on the assumptions on function f $\lim_{t\to+\infty} A_j(t) = +\infty$ and this summand converges to zero.

Otherwise, if $\theta_j \neq 0$, $\lim_{t\to+\infty} B_j(t) = +\infty$ and we can express the corresponding summand in (3.10) as

$$\Big(\frac{B_j(t)}{A_j(t)}\Big)^{p+1} E\Big(\Big(\frac{\max_{0\le r\le 1}|\Psi_j(r)|}{B_j(t)} + 1\Big)^{p+1}\Big),$$

where the expectation is bounded again, and then it converges to zero because

$$\lim_{t\to+\infty} \frac{B_j(t)}{A_j(t)} = \lim_{t\to+\infty} |\theta_j|\, K_{R,a}\, J\, \frac{t}{f(t)} = 0\,. \quad \square$$

Remark 2 The next result shows the tightness in cases $\theta = 0$ and $\theta > 0$, of the approximation of the increase of $M(t)$ as $t \to +\infty$ given by Theorem 1 in the following sense: we will see that $M(t)$ grows more than any function $g(t)$ than can be chosen growing only a little bit slower than t^H or t, depending on if $\theta = 0$ or $\theta > 0$, respectively, if the vector of weights a verifies the following restriction involving the reflection matrix R:

(**HaR**)	$R^T a \ge 0$

This assumption, as we will see in Section 4.2, can be easily checked in some interesting examples.

Remark 3 Assumption **(HR)** implies that R is a *Completely-$\mathcal{S}$* matrix, which is equivalent to saying that R is *strictly semi-monotone*. This last property means that for each principal sub-matrix $\tilde{R}$ of R, the system

$$\tilde{R}\, x \le 0 \quad \text{and} \quad x \ge 0$$

has the unique solution $x = 0$. In particular, this implies that $R^T a$ cannot be ≤ 0 since $a \ge 0$ but $a \neq 0$. Note that we need to impose **(HaR)**, which is a more restrictive condition in some sense (unless $J = 1$, in which case they are equivalent), but only for the vector of weights a.

In the sequel, in addition to μ, already introduced in Section 1 as notation for $a^T\,\theta$, we will use the notation $\sigma^2 \stackrel{\text{def}}{=} a^T\,\Gamma\, a\,(>0)$. Note that $\theta = 0$ (respectively $< 0, > 0$) implies $\mu = 0$ (respectively $< 0, > 0$), but that the converses are not true.

Theorem 2 (Asymptotic lower bound) *Under assumptions* **(HR)** *and* **(HaR)**,

a) If $\mu \geq 0$, *then*

$$\text{P-}\lim_{t\to+\infty} \frac{M(t)}{g(t)} = +\infty$$

for any positive real function g *such that*

$$\begin{cases} \lim\limits_{t\to+\infty} \dfrac{g(t)}{t^H} = 0 \quad \textit{if} \quad \mu = 0 \\ \lim\limits_{t\to+\infty} \dfrac{g(t)}{t} = 0 \quad \textit{if} \quad \mu > 0\,. \end{cases}$$

b) If $\mu < 0$, *then*

$$\lim_{t\to+\infty} P\Big(\frac{M(t)}{(\log t)^{\frac{1}{2(1-H)}}} \geq C \Big) = 1 \quad \textit{for any} \quad 0 < C < \Big(\frac{\sigma^2}{2\,(-\mu)^{2H}} \Big)^{\frac{1}{2(1-H)}}\,.$$

Proof: We will prove that

$$\lim_{t\to+\infty} P\Big(\frac{M(t)}{g(t)} \geq C \Big) = 1 \tag{3.11}$$

for any arbitrary $C > 0$ and g verifying the aforementioned assumptions if $\mu \geq 0$, in part a), which implies that $\dfrac{M(t)}{g(t)}$ is unbounded in probability, and for any $0 < C < \Big(\frac{\sigma^2}{2\,(-\mu)^{2H}} \Big)^{\frac{1}{2(1-H)}}$ and $g(t) = (\log t)^{\frac{1}{2(1-H)}}$ if $\mu < 0$, in part b). The proof falls into several steps.

Step 1: Fix $t > 0$. We claim that for any $\delta \in (0,\, t)$,

$$M(t) \geq \max_{k=1,\ldots,[\frac{t}{\delta}]} V_k(\delta) \tag{3.12}$$

$$\text{with} \quad V_k(\delta) \stackrel{\text{def}}{=} a^T \big(X(k\delta) - X((k-1)\delta) \big)\,. \tag{3.13}$$

Indeed, since process Y has non-decreasing paths, it follows that if $0 \leq s \leq t$, $Y(t) - Y(s) \geq 0$ and therefore $a^T\, R\big(Y(t) - Y(s)\big) \geq 0$ by assumption **(HaR)**, which implies $-a^T\, RY(s) \geq -a^T\, RY(t)$, and therefore,

$$\inf_{0\leq s\leq t} \big(-a^T\, RY(s) \big) \geq -a^T\, RY(t)\,. \tag{3.14}$$

Taking into account that

$$0 \leq a^T\, W(s) = a^T\, X(s) + a^T\, RY(s) \Rightarrow a^T\, X(s) \geq -a^T\, RY(s)\,,$$

accordingly to (3.14) we have that $\inf_{0\le s\le t} a^T X(s) \ge -a^T R Y(t)$, which implies (by replacing s by r and t by s) $a^T R Y(s) \ge -\inf_{0\le r\le s} a^T X(r)$. On account of the above inequality, we have

$$\begin{aligned} a^T W(s) = a^T X(s) + a^T R Y(s) &\ge a^T X(s) - \inf_{0\le r\le s} a^T X(r) \\ &\text{(and with } r = s-\tilde\delta, \text{ for any } \tilde\delta \in (0, s]\,) \\ &\ge a^T X(s) - a^T X(s-\tilde\delta) = a^T \big(X(s) - X(s-\tilde\delta)\big)\,. \end{aligned} \tag{3.15}$$

Fix now $\delta \in (0,\, t)$. Since $M(t) = \max_{0\le s\le t} a^T W(s) \ge \max_{k=1,\dots,[\frac{t}{\delta}]} a^T W(k\,\delta)$, choosing $\tilde\delta = \delta \in (0,\, k\,\delta]$, inequality (3.15) with $s = k\,\delta$ shows that

$$M(t) \ge \max_{k=1,\dots,[\frac{t}{\delta}]} a^T \big(X(k\delta) - X((k-1)\delta)\big) = \max_{k=1,\dots,[\frac{t}{\delta}]} V_k(\delta)\,,$$

which is the desired inequality (3.12).

Step 2: By assumption, X is a (zero-delayed $J-$dimensional) fBm with associated data $(0,\, H,\, \theta,\, \Gamma)$. Therefore, for any $\delta \in (0,\, t)$ and any $k = 1, \dots, [\frac{t}{\delta}]$,

$$X(k\,\delta) - X((k-1)\,\delta) \sim N_J\big(\,\delta\,\theta,\, \delta^{2\,H}\,\Gamma\,\big)\,,$$

and consequently $V_k(\delta)$ defined by (3.13) is a (one-dimensional) Gaussian random variable,

$$V_k(\delta) \sim N\big(\,a^T\,\delta\,\theta,\, \delta^{2\,H}\,a^T\,\Gamma\,a\,\big) = N\big(\,\mu\,\delta,\, \sigma^2\,\delta^{2\,H}\,\big)\,.$$

We can normalize these variables by defining

$$Z_k \stackrel{\text{def}}{=} \frac{V_k(\delta) - \mu\,\delta}{\sigma\,\delta^H} \sim N(0,\,1)\,, \quad \text{for } k = 1\,, \dots,\, [\frac{t}{\delta}]\,, \tag{3.16}$$

which form a stationary sequence of standardized Gaussian random variables (called *fractional Gaussian noise*) with the property that their covariance function $\rho_Z(\ell) \stackrel{\text{def}}{=} E\,\big(\,Z_1\,Z_{1+\ell}\,\big)$ goes to zero, when ℓ grows, in the sense that $\lim_{\ell\to+\infty} \rho_Z(\ell)\,\log\ell = 0$, because $H < 1$ and

$$\rho_Z(\ell) = H\,(2H-1)\,\ell^{2H-2} + O(\ell^{2H-3})\,.$$

Consequently, we can apply (2.3) with $\tau = +\infty$: if we find non-negative functions of t, say $\delta(t)$ (with $0 < \delta(t) < t$ for t big enough) and $u(t) > 0$, such that by defining $m(t) \stackrel{\text{def}}{=} \big[\frac{t}{\delta(t)}\big]$ we have that

(i) $\lim_{t\to+\infty} m(t) = +\infty$, and

(ii) $\lim_{t\to+\infty} m(t)\,\big(1-\Phi(u(t))\big) = +\infty\,,$

therefore we will obtain

$$\lim_{t\to+\infty} P\Big(\max_{k=1,\dots,[\frac{t}{\delta(t)}]} Z_k \geq u(t)\Big) = 1\,,$$

and as a consequence and taking into account (3.16) and (3.12), we will deduce that

$$\lim_{t\to+\infty} P\Big(M(t) \geq u(t)\,\sigma\,(\delta(t))^H + \mu\,\delta(t)\Big) = 1\,. \tag{3.17}$$

In order to check the limit in (ii), we can use the usual lower bound for the tail of the standard gaussian distribution (see for instance inequalities (2.1) in [1]) to obtain

$$m(t)\,\big(1-\Phi(u(t))\big) \geq \frac{1}{\sqrt{2\,\pi}}\,\frac{m(t)}{u(t)}\,\big(1-\frac{1}{(u(t))^2}\big)\,e^{-\frac{(u(t))^2}{2}}\,.$$

Then, it is sufficient to see that

$$\lim_{t\to+\infty} \frac{m(t)}{u(t)}\,\big(1-\frac{1}{(u(t))^2}\big)\,e^{-\frac{(u(t))^2}{2}} = +\infty\,. \tag{3.18}$$

Step 3: The rest of the proof takes into account the sign of $\mu = a^T\,\theta$, so we split it into three cases:

- *Case* $\mu = 0$

  Fix an arbitrary $C > 0$ and define

$$\delta(t) \stackrel{\text{def}}{=} \frac{(C\,g(t))^{1/H}}{\Big(\sigma^2\,\log\big(\frac{t}{g(t)^{1/H}}\big)\Big)^{\frac{1}{2H}}} \quad \text{and} \quad u(t) \stackrel{\text{def}}{=} \frac{C\,g(t)}{\sigma\,(\delta(t))^H} = \Big(\log\big(\frac{t}{g(t)^{1/H}}\big)\Big)^{1/2}\,.$$

With these definitions, conditions (i) and (ii) in Step 2 are satisfied: (i) is true because $\frac{t}{(g(t))^{1/H}}\,\Big(\log\big(\frac{t}{g(t)^{1/H}}\big)\Big)^{\frac{1}{2H}}$ grows to $+\infty$ by the assumption on function g, and we can see that (ii) is also true by checking (3.18). Indeed, (3.18) is accomplished because

$$\frac{t}{g(t)^{1/H}}\,\Big(\log\big(\frac{t}{g(t)^{1/H}}\big)\Big)^{\frac{1-H}{2H}}\,\Big(1-\frac{1}{\log\big(\frac{t}{g(t)^{1/H}}\big)}\Big)\,e^{-\frac{1}{2}\,\log\big(\frac{t}{g(t)^{1/H}}\big)} =$$
$$\Big(\frac{t}{g(t)^{1/H}}\Big)^{1/2}\,\Big(\log\big(\frac{t}{g(t)^{1/H}}\big)\Big)^{\frac{1-H}{2H}}\,\Big(1-\frac{1}{\log\big(\frac{t}{g(t)^{1/H}}\big)}\Big) \longrightarrow +\infty$$

By considering that $u(t)\,\sigma\,(\delta(t))^H + \mu\,\delta(t) = u(t)\,\sigma\,(\delta(t))^H = C\,g(t)\,,$ (3.17) gives (3.11) in this case.

- *Case* $\mu > 0$

Fix again an arbitrary constant $C > 0$ and define

$$\delta(t) \stackrel{\text{def}}{=} \frac{C\, g(t)}{\mu}, \quad \text{and} \quad u(t) \stackrel{\text{def}}{=} \Big(\log\big(\frac{t}{g(t)}\big) \Big)^{1/2}.$$

We next prove that both conditions (i) and (ii) in Step 2 are satisfied with these definitions. Indeed, we see that (i) is true because $\lim_{t\to+\infty} \frac{t}{g(t)} = +\infty$ by the assumption on g, and (3.18) holds because

$$\frac{t}{g(t)} \Big(\log\big(\frac{t}{g(t)}\big) \Big)^{-1/2} \Big(1 - \frac{1}{\log\big(\frac{t}{g(t)}\big)}\Big) e^{-\frac{1}{2}\, \log\left(\frac{t}{g(t)}\right)} =$$
$$\Big(\frac{t}{g(t)}\Big)^{1/2} \Big(\log\big(\frac{t}{g(t)}\big) \Big)^{-1/2} \Big(1 - \frac{1}{\log\big(\frac{t}{g(t)}\big)}\Big) \longrightarrow +\infty$$

Taking into account that in this case $u(t)\, \sigma\, (\delta(t))^H + \mu\, \delta(t) \geq \mu\, \delta(t) = C\, g(t)$ and (3.17), we have that (3.11) is accomplished and in this way we finish the proof of part a) of the theorem.

- *Case* $\mu < 0$

The last case considered corresponds to part b) of the theorem. Fix an (arbitrary by the moment) constant $C > 0$ and define

$$\delta(t) \stackrel{\text{def}}{=} \frac{C\, (\log t)^{\frac{1}{2(1-H)}}}{-\mu} \quad \text{and} \quad u(t) \stackrel{\text{def}}{=} \frac{2\, C\, (\log t)^{\frac{1}{2(1-H)}}}{\sigma\, (\delta(t))^H} = \frac{2\, C^{1-H}\, (\log t)^{1/2}\, (-\mu)^H}{\sigma}$$

Condition (i) in Step 2 is accomplished since $\frac{t}{(\log t)^{\frac{1}{2(1-H)}}} \longrightarrow +\infty$ as $t \to +\infty$, and (3.18) also holds because

$$\frac{t}{(\log t)^{\frac{1}{2(1-H)}}}\, (\log t)^{-1/2}\, (1 - \frac{1}{\log t})\, e^{-\frac{1}{2\sigma^2} 4\, C^{2(1-H)}\, \log t\, (-\mu)^{2H}} =$$
$$\frac{1}{(\log t)^{\frac{2-H}{2(1-H)}}}\, t^{1 - \frac{2\, C^{2(1-H)}\, (-\mu)^{2H}}{\sigma^2}}\, (1 - \frac{1}{\log t}) \longrightarrow +\infty$$

if and only if $0 < C < \Big(\frac{\sigma^2}{2\, (-\mu)^{2H}} \Big)^{\frac{1}{2(1-H)}}$. In this case we have that

$$u(t)\, \sigma\, (\delta(t))^H + \mu\, \delta(t) = 2\, C\, (\log t)^{\frac{1}{2(1-H)}} - C\, (\log t)^{\frac{1}{2(1-H)}} = C\, (\log t)^{\frac{1}{2(1-H)}},$$

and the argument follows similarly to the other cases by replacing $g(t)$ by $(\log t)^{\frac{1}{2(1-H)}}$, obtaining that from (3.17), (3.11) holds. $\square$

Remark 4 We can introduce the level-crossing times for process $a^T W(\cdot)$ in this way: for any $b > 0$, let

$$T(b) \stackrel{\text{def}}{=} \inf\{\, t \geq 0 \,:\, a^T W(t) \geq b \,\}\,.$$

By the following relationship between these level-crossing times and the maximum,

$$\{\, T(b) \leq t \,\} = \{\, M(t) \geq b \,\}\,,$$

we obtain from Theorems 1 and 2, analogously to Theorem 2 in [15], that as $b \to +\infty$, the growth of $T(b)$ is:

$$\begin{cases} \text{like that of } \ b \ \text{ if } \mu > 0 \\ \text{like that of } \ b^{1/H} \ \text{ if } \theta = 0\,(\Rightarrow \mu = 0) \\ \text{between that of } \ b \ \text{ and that of } \ e^{\left(b^{2(1-H)}\right)} \ \text{ if } \mu < 0\,. \end{cases}$$

4 Application to queueing networks

4.1 The multi-class multi-server fluid queueing network

In this section, by following [4] we explain in some detail the multi-class multi-server fluid queueing network already introduced in Section 1, to which we can apply the results of Section 3, and give some specific notations that will be used in the examples of Section 4.2. Let $m = (m_1, \ldots, m_K)^T > 0$ and $M = diag(m)$, $\frac{1}{m_k}$ being the processing rate for the class k fluid (thus m_k being the corresponding mean service time). The many-to-one mapping from fluid classes to stations is denoted by s; then $s(k)$ is the station at which class k fluid is processed and $s^{-1}(j)$ is the set of fluid classes processed at station j. We introduce the $J \times K$ *constituency matrix* $C = (C_{jk})$ by

$$C_{jk} \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } j = s(k) \\ 0 & \text{otherwise} \end{cases}$$

The exogenous arrival process is assumed to be generated in this way: for each fluid class, say k, there are N i.i.d. ON/OFF heavy-tailed sources, with N tending to $+\infty$; when one of them is ON, it sends class k fluid to the network at deterministic rate (depending on N) $\alpha_k^N \geq 0$. Let $\alpha^N = (\alpha_1^N, \ldots, \alpha_K^N)^T$ and $\alpha \stackrel{\text{def}}{=} \lim\limits_{N \to +\infty} \alpha^N$. Fluid is processed by each

server in a *first-in-first-out* (FIFO) discipline which is also assumed to be *non-idling* (*work-conserving*), that is, a server is never idle if there is some fluid waiting to be processed at its station. We also assume no capacity restriction in buffers.

We allow *feedback* in the system, that is, fluid can come back to revisit previously visited stations (by conveniently changing class). Let $P = (P_{k\ell})_{k,\ell=1}^{K}$ be the sub-stochastic *"fluid"* or *"routing matrix"* defined by: $P_{k\ell}$ is the proportion of class k fluid that leaving station $s(k)$ goes next to station $s(\ell)$ as class ℓ fluid, and $1 - \sum_{\ell=1}^{K} P_{k\ell} \geq 0$ is the proportion that leaves the network. It is assumed that P has spectral radius strictly less than one, and therefore, matrix $Q \stackrel{\text{def}}{=} (I - P^T)^{-1}$ is well defined.

Two main performance processes related with the multi-class fluid queueing network which measure congestion and delay in the system are the $J-$dimensional workload and the $K-$dimensional fluid in queue. Workload is defined to be the amount of time required for any server to complete processing of all fluids in queue (or being processed) at any given time, and the fluid in queue is the amount of any fluid class in the queue or being processed at any given time. For simplicity we assume to be zero the workload at time $t = 0$.

Let ν be the constant depending on the (finite) expected values of the lengths of the ON- and OFF- periods, ν_{on} and ν_{off} respectively, defined by

$$\nu \stackrel{\text{def}}{=} \frac{\nu_{\text{on}}}{\nu_{\text{on}} + \nu_{\text{off}}}\,, \tag{4.1}$$

and let λ be the unique $K-$dimensional vector solution to the *traffic equation*:

$$\lambda = \alpha\,\nu + P^T\,\lambda \quad (\text{ that is, } \lambda = Q\,\alpha\,\nu\,). \tag{4.2}$$

For any k, λ_k can be interpreted as the long run class k fluid rate into and out of station $s(k)$. As usual, the fluid traffic intensity for station j is defined by

$$\rho_j \stackrel{\text{def}}{=} \sum_{k \in s^{-1}(j)} \lambda_k\, m_k$$

and the corresponding vector is $\rho = C\,M\,\lambda$. We assume *heavy traffic*, that is, $\rho = e$. Let we introduce the $K \times J$ matrix $\Delta = (\Delta_{kj})$ defined by

$$\Delta_{kj} \stackrel{\text{def}}{=} \begin{cases} \lambda_k & \text{if } k \in s^{-1}(j)\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

We also assume *state space collapse* condition, which is a condition that relates processes $\hat{Z}$ and $\hat{W}$, obtained from the fluid in queue and the workload processes, respectively, when the number of ON/OFF sources N goes to ∞, after multiply the original processes by $\sqrt{N}$. Obviously, these two processes are related by means of

$$\hat{W} = C\,M\,\hat{Z}\,, \tag{4.3}$$

that is, for any station j,

$$\hat{W}_j = \sum_{k \in s^{-1}(j)} m_k\,\hat{Z}_k$$

since workload is the sum of the ratio between the amount of each fluid class in queue at this station by its processing rate. *State space collapse condition* is established by means of the *"lifting"* matrix Δ as follows: $\hat{Z} = \Delta\,\hat{W}$. This means that for any fluid class, the amount of fluid in queue is a fixed portion of the workload at the station at which this class is processed, that is, for any k,

$$\hat{Z}_k = \lambda_k\,\hat{W}_{s(k)}\,.$$

In other words, under *state space collapse* we can recover the fluid in queue for each fluid class from workload at the station at which that class is processed, even if at this station two or more fluid classes are processed.

Note that if $K = J$ then $C = I$ and from (4.3) we have that $\hat{Z} = M^{-1}\,\hat{W}$, that is, *state space collapse* is trivially accomplished with $\Delta = M^{-1}$ (then, $\lambda_k = \frac{1}{m_k}$ for any k).

If we assume that matrix $C\,M\,Q\,\Delta$ is invertible, we can define matrix R by

$$R \stackrel{\text{def}}{=} (\,C\,M\,Q\,\Delta\,)^{-1}\,. \tag{4.4}$$

Corollary 1 [4] establishes that in this setting, under assumptions **(HR)** on matrix R, *heavy traffic* ($\rho = e$) and *state space collapse* ($\hat{Z} = \Delta\,\hat{W}$), processes $\hat{W}$ and $\hat{Z}$, conveniently scaled, converge respectively to processes W and Z (verifying that $Z = \Delta\,W$), W being a rfBm on the first orthant S with associated data ($x = 0$, H, $\theta = 0$, Γ, R), where $H \in (1/2,\,1)$.

Recall that in the Introduction we have defined the **maximum (total) amount of fluid in queue in the system** over the interval $[0,\,t]$ by (1.2). Taking into account that by *state space collapse* $Z = \Delta\,W$, we can rewrite $M(t)$ as

$$M(t) = \max_{0 \le s \le t} e^T\,\Delta\,W(s)\,,$$

and using that $e^T\,\Delta\,W = \sum_{k=1}^{K} \lambda_k\, W_{s(k)} = \sum_{j=1}^{J} \big(\sum_{k\in s^{-1}(j)} \lambda_k \big)\, W_j = \sum_{j=1}^{J} a_j\, W_j$ with

$$a_j = \sum_{k\in s^{-1}(j)} \lambda_k > 0 \quad \text{for any } j = 1, \ldots, J\,, \tag{4.5}$$

we observe that $M(t)$ can be rewritten as (1.1) and by Remark 1, $M(t) \overset{\mathcal{D}}{=} t^H\, M(1)$.

4.2 Some examples

4.2.1 The particular case $K = J = 1$: the single-class single-server network with feedback

In the one-dimensional case ($K = J = 1$), we can consider in general a rfBm process $W = X + RY$ on $S = [0, +\infty)$, with associated data ($x = 0$, H, θ, Γ, R), for any $\theta \in \mathbb{R}$, $H \in (0, 1)$, $\Gamma > 0$ and $0 < R < 2$ (R verifies **(HR)** if and only if $R \in (0, 2)$ in this case). Thus, our results apply to

$$M(t) = \max_{0\le s\le t} a\, W(s)$$

for any $a > 0$ (since condition **(HaR)** is satisfied automatically for any $R > 0$).

In particular, this situation covers the application to any single-class single-server fluid queueing network with feedback of the type considered in Section 4.1. In this case, $P = p$, p being the proportion of fluid that finishing process at the station, comes back to be reprocessed, with $0 \le p < 1$. Then, $R = 1 - p$ since $C = 1$, $M = m$, $Q = \frac{1}{1-p}$ and $\lambda = \frac{1}{m}$ by assuming *heavy traffic*. Taking into account that for $K = J$ *state space collapse* condition vanishes, we have that the workload limit of this queue is a one-dimensional rfBm process with associated data ($x = 0$, H, $\theta = 0$, Γ, $R = 1 - p$) with $H \in (\frac{1}{2}, 1)$ and some $\Gamma > 0$. The particular case of *no-feedback* (that is, $p = 0$) is the one considered in [15] (then, $R = 1$). The weight is $a = \lambda = \frac{1}{m}$ and then

$$M(t) = \max_{0\le s\le t} \frac{1}{m} W(s)$$

is the maximum total amount of fluid in queue in $[0, t]$, and $M(t) \overset{\mathcal{D}}{=} t^H\, M(1)$ by Remark 1, which could also be obtained analogously to Proposition 2 [15].

4.2.2 The single-class multi-server network ($K = J > 1$) with feedback

For this particular fluid queueing network the constituency matrix C is the identity, and if we assume *heavy traffic* $C\,M\,\lambda = e$, we have that $\lambda_k = \frac{1}{m_k}$ for all k. Consequently, $a_j = \lambda_j$ and

$$M(t) = \max_{0 \le s \le t} \sum_{j=1}^{J} \frac{1}{m_j}\, W_j(s)$$

is the maximum total amount of fluid in queue in the network over $[0,\, t]$. Matrix R is given by (4.4) and taking into account that $Q = (I - P^T)^{-1}$, we have that

$$R = (\, M\, Q\, M^{-1}\,)^{-1} = M\,(I - P^T)\,M^{-1} = I - M\,P^T\,M^{-1}\,,$$

which verifies condition **(HR)** since the spectral radius of $M\,P^T\,M^{-1}$ coincides with that of P, assumed to be strictly less than 1. Since *State space collapse* condition is also trivially accomplished in this case because $K = J$ with $\Delta = M^{-1}$, our results can be applied in this setting to $M(t)$, and we obtain that $M(t)$ grows like t^H .

Note also that assumption **(HaR)** would be trivially accomplished because

$$R^T\, a = \big(\frac{1}{m_1}\,(1 - \sum_{\ell=1}^{J} P_{1\ell}), \ldots, \frac{1}{m_j}\,(1 - \sum_{\ell=1}^{J} P_{j\ell}), \ldots, \frac{1}{m_J}\,(1 - \sum_{\ell=1}^{J} P_{J\ell})\big)^T\,,$$

which is ≥ 0 since P is a sub-stochastic matrix. In fact, this is true for any vector $a \ge 0$.

4.2.3 A tandem queue with feedback

Consider the fluid tandem queue displayed in figure 1, with two stations ($J = 2$) and three fluid classes ($K = 3$), where class 1 fluid enters the system from outside (at rate $\alpha_1^N > 0$) and it is processed by server 1. After being processed (at constant processing rate $1/m_1$) by the first server, this fluid goes into station 2 as class 3 fluid, where it is processed at constant processing rate $1/m_3$. After that, a proportion $q \in (0,\, 1]$ of fluid goes outside the network while the proportion $p = 1 - q \in [0,\, 1)$ goes back to station 1 to be reprocessed as class 2 fluid, at constant processing rate $1/m_2$, and then goes again to station 2 as class 3 fluid, and so on. This model, which is the two-stage queueing system considered in Section 6.1 of [4] (where the interested reader is addressed for details), seems adequate, for instance, in situations in which there is recycling, that is, quality control inspection is performed after first stage at the second one, and fluid that does not meet quality standards is sent back to station 1 for reprocessing. Case $p = 0$ corresponds to the *non-feedback* situation.

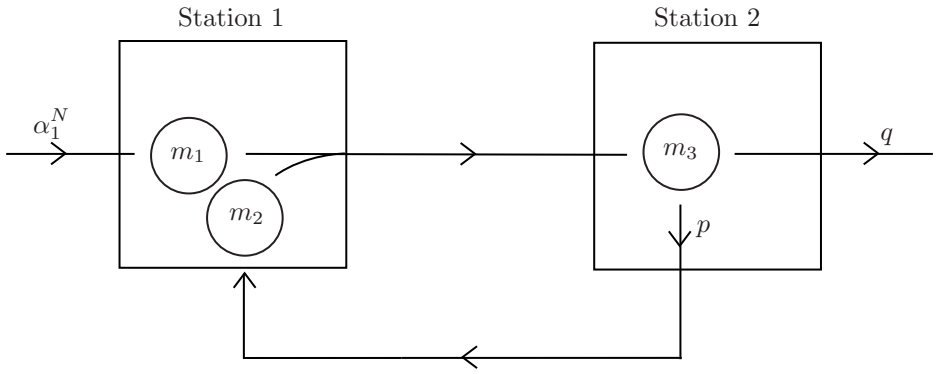

Figure 1: a tandem queue with feedback

In this case,

$$R = \begin{pmatrix} 1 & -p\,\frac{m_2}{m_3} \\ -1 & q + p\,\frac{m_2}{m_3} \end{pmatrix},$$

which verifies condition **(HR)** if in case $m_2 > m_3$ we impose $p < \frac{m_3}{2\,m_2 - m_3}$ (otherwise, **(HR)** is always verified). In particular, in the non-feedback case ($p = 0$), this condition holds automatically. By (4.5), $a = (\lambda_1 + \lambda_2,\, \lambda_3)^T = (\frac{1}{m_3},\, \frac{1}{m_3})^T$ and therefore we have that

$$M(t) = \max_{0 \le s \le t} \frac{1}{m_3} \left(W_1(s) + W_2(s) \right)$$

is the maximum total fluid in queue in the system over the interval $[0,\, t]$, and under *heavy traffic* and *state space collapse* assumptions, if for $m_2 > m_3$ we have that $p < \frac{m_3}{2\,m_2 - m_3}$, we obtain that the growth of $M(t)$ is like that of t^H.

From the expression of matrix R we can see that assumption **(HaR)** would be trivially accomplished, in particular, for any vector of weights $a = (a_1,\, a_2)^T > 0$ such that $a_1 = a_2$.

4.2.4 A parallel-server example (with $K = 4$, $J = 2$)

Let us consider now the system with two parallel servers ($J = 2$) corresponding to figure 2 below.

Server 1 processes class 1 fluid, which arrives from outside at rate $\alpha_1^N > 0$, and server 2 processes class 2 fluid, arrived from outside at rate $\alpha_2^N > 0$. After finishing processing at station 1, a proportion p_1 of fluid goes next to station 2 as class 3 fluid, and the rest $q_1 = 1 - p_1$ exits the system. After finishing processing at station 2, a proportion p_2 of fluid

(independently of its class) goes next to station 1 as class 4 fluid, and the rest $q_2 = 1 - p_2$ exits the system. Proportion p_1 is irrespective of the fluid class, 1 or 4. We have $K = 4$ fluid classes with $\alpha_3^N = \alpha_4^N = 0$. Assume that $0 \le p_1,\, p_2 \le 1$ but $p_1\, p_2 < 1$ (to ensure that the spectral radius of the routing matrix is strictly less than 1).

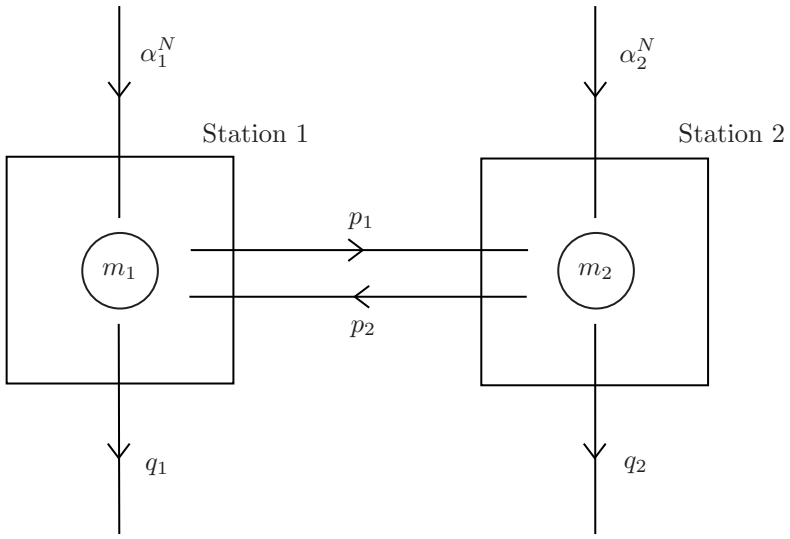

Figure 2: a two parallel-server system with feedback

The mean service rates, assumed to depend only on the server (and not on the class), are $m_1 > 0$ for server 1 and $m_2 > 0$ for server 2. Therefore, under *heavy traffic* we have that

$$\lambda = \big(\, \frac{1}{m_1} - \frac{p_2}{m_2},\, \frac{1}{m_2} - \frac{p_1}{m_1},\, \frac{p_1}{m_1},\, \frac{p_2}{m_2} \,\big)^T$$

with the following restrictions:

$$\begin{cases} 0 \le p_1 \le \frac{m_1}{m_2} & \text{if } m_1 < m_2 \\ 0 \le p_2 \le \frac{m_2}{m_1} & \text{if } m_2 < m_1 \\ p_1\, p_2 < 1 & \text{if } m_1 = m_2 \end{cases}$$

(which imply $p_1\, p_2 < 1$ in any case). Moreover, by formula (4.5),

$$a_1 = \lambda_1 + \lambda_4 = \frac{1}{m_1}\,(> 0) \quad \text{and} \quad a_2 = \lambda_2 + \lambda_3 = \frac{1}{m_2}\,(> 0)\,,$$

so $M(t) = \max\limits_{0 \le s \le t} \big(\, \frac{1}{m_1}\, W_1(s) + \frac{1}{m_2}\, W_2(s) \,\big)$ is the maximum total fluid in queue in the system

over the interval $[0,\, t]$ under *heavy traffic* and *state space collapse*. Taking into account that

$$C\,M\,Q\,\Delta = \frac{1}{1-p_1\,p_2}\begin{pmatrix} 1 & \frac{m_1\,p_2}{m_2} \\ \frac{m_2\,p_1}{m_1} & 1 \end{pmatrix} \text{ and then } R = \begin{pmatrix} 1 & -\frac{m_1\,p_2}{m_2} \\ -\frac{m_2\,p_1}{m_1} & 1 \end{pmatrix},$$

which trivially verifies assumptions **(HR)**, we obtain that $M(t)$ grows like t^H.

Assumption **(HaR)** can be easily checked since $p_1\,p_2 < 1$ and $R^T\,a = \left(\,\frac{1-p_1}{m_1},\,\frac{1-p_2}{m_2}\,\right)^T \geq 0$.

Acknowledgements

The author wishes to thank the anonymous referees for careful reading and very helpful comments that resulted in an overall improvement of the paper.

References

[1] Adler, R.J. An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. Lecture Notes-Monograph Series. **12**. Institute of Mathematical Statistics, Hayward, California, 1990.

[2] Bernard, A.; El Kharroubi, A. Régulations déterministes et stochastiques dans le premier "orthant" de $\mathbb{R}^n$. Stoch. Stoch. Rep. **1991**, *34*, 149-167.

[3] Delgado, R. A reflected fBm limit for fluid models with ON/OFF sources under heavy traffic. Stoch. Process. Appl. **2007**, *117*, 188-201.

[4] Delgado, R. State-space collapse for asymptotically critical multi-class fluid networks.Queueing Syst. **2008**, *59*, 157-184.

[5] Duffield, N. G.; O'Connell, N. Large deviations and overflow probabilities with general single-server queue, with applications. Math. Proc. Cambridge Philos. Soc. **1995**, *118*, 363-374.

[6] Duncan, T. E.; Yan, Y.; Yan, P. Exact a symptotics for a queue with fractional Brownian input and applications in ATM networks. J. Appl. Probab. **2001**, *38*, 932-945.

[7] El Karoui, N.; Chaleyat-Maurel, M. Un problème de réflexion et ses applications au temps local et aux équations différentielles stochastiques sur $\mathbb{R}$, cas continu. Société Mathématique de France, Asterisque **1978**, *52-53*, 117-144.

[8] Hüsler, J.; Piterbarg, V. Extremes of a certain class of Gaussian processes. Stoch. Process. Appl. **1999**, *83*, 257-271.

[9] Hüsler, J.; Piterbarg, V. Limit theorem for maximum of the storage process with fractional Brownian motion as input. Stoch. Process. Appl. **2004**, *114*, 231-250.

[10] Leadbetter, M. R.; Lindgren, G.; Rootzén, H. Extremes and Related Properties of Random Sequences and Processes, Springer-Verlag, 1983.

[11] Massoulie, L.; Simonian, A. Large buffer asymptotics for the queue with FBM input. J. Appl. Probab. **1999**, *36*, 894-906.

[12] Norros, I. A storage model with self-similar input. Queueing Syst. **1994**, *16*, 387-396.

[13] Taqqu, M. S.; Willinger, W.; Sherman, R. Proof of a Fundamental Result in Self-Similar Traffic Modeling. Comput. Commun. Rev. **1997**, *27*, 5-23.

[14] Williams, R. J. An invariance principle for semimartingale reflecting Brownian motions in an orthant. Queueing Syst. **1998**, *30*, 5-25.

[15] Zeevi, A. J.; Glynn, P. W. On the maximum workload of a queue fed by fractional brownian motion. Ann. Appl. Probab. **2000**, *10 (4)*, 1084-1099.