A reflected fBm limit for fluid models with ON/OFF sources under heavy traffic ⋆

Rosario Delgado

Departament de Matemàtiques. Universitat Autònoma de Barcelona. Edifici C-Campus de la UAB. 08193 Bellaterra (Cerdanyola del Vallès) SPAIN

Abstract

We consider a family of non-deterministic fluid models that can be approximated under heavy traffic conditions by a multidimensional reflected fractional Brownian motion (rfBm). Specifically, we prove a heavy traffic limit theorem for multi-station fluid models with feedback and non-deterministic arrival process generated by a large enough number of heavy tailed ON/OFF sources, say N. Scaling in time by a factor r and in state space conveniently, and letting N and r approach infinity (in this order) we prove that the scaled *immediate workload process* converges in some sense to a rfBm.

Key words: reflected fractional Brownian motion, fluid model, queueing network, workload process, on-off sources, heavy traffic

⋆ Partially supported by projects BFM2003-00261, MTM2004-01175 and the research group 2001SGR-00174 (CIRIT)

Email address: delgado@mat.uab.es (Rosario Delgado).

1 Introduction

The presence of long-range dependence in broadband network traffic as well as that of self-similar traffic patterns in modern high-speed network traffic lead naturally to the question of finding adequate traffic models for these situations. One simple physical explanation for this kind of phenomenon consists on the superposition of many ON/OFF sources with strictly alternating ON- and OFF-periods and whose ON- or OFF-periods lengths have high variability (that is, exhibit the *Noah Effect*), as can be seen in [7]. There it is proved that in that scenario aggregate network traffic can be self-similar or long-range dependent (exhibits the *Joseph Effect*): in Theorem 1 of [7] the authors prove that the superposition of N ON/OFF sources generates an aggregate cumulative arrival process that conveniently scaled in time by a factor r and in state space, converges in some sense, as N goes to infinity and after that, as r goes to infinity, to a fractional Brownian motion (fBm) (these limits should be treated with care, because if they are taken in the reverse order, the convergence is to an $\alpha-$stable Lévy process rather than a fBm). What is more, they relate the parameter that describes the intensity of the *Noah Effect* (that means, the heaviness of the tail of the distribution of lengths) of the ON- and/or OFF-periods, with the Hurst parameter of the fBm, that is a measure of its degree of self-similarity (or *Joseph Effect*).

By considering the question of predicting the performance experienced by a superposition of heavy-tailed ON/OFF sources multiplexed at a buffered resource, in Debicki and Mandjes ([3]) is considered the following question: does the convergence of the aggregate cumulative arrival process to the fBm given by Theorem 1 of [7] carry over to the stationary buffer content process? They

give a positive answer to it in a heavy-traffic environment, by showing that the scaled workload process converges to the fBm, reflected appropriately to be non-negative, for fluid models with only one station. One may ask whether this remains true in a multi-station environment, for a fluid model with feedback, that is the scenario considered in our paper. That is the question that motivates this paper.

To be more specific, we consider non-deterministic fluid models with entries like in the model of Debicki and Mandjes but with a structure similar to that introduced by Harrison in [5] as the deterministic fluid analog of multiclass queueing networks with feedback. We assume that our system has J stations with a single server and an infinite buffer at each one, feedback and FIFO (first-in-first-out) discipline. We suppose (and this gives the difference with the model considered by Harrison) that the process of external arrivals is a non-deterministic aggregated cumulative process generated by a large enough number of heavy tailed ON/OFF sources. We prove in Theorem 1 that after adequate scaling, the immediate workload process converges to a $J-$dimensional *reflected fractional Brownian motion* process, that is, we extend the result in [3] to our more general setting. A key ingredient in the proof is the *invariant principle* given by Williams in [8] for the reflected Brownian motion process, that can also be applied to the reflected fractional Brownian motion process.

We also prove a Functional Weak Law of Large Numbers (FWLLN) (see Theorem 2) for two processes defined as the total amount of fluid arriving to the stations (including both feedback flow and external input), and the total amount of leaving fluid from the stations (to other stations or outside the system), up to any time. This result justifies the interpretation of parameter λ introduced

in (6) (the solution of the limiting traffic equation) as the long run fluid rate into and out of stations.

The paper is organized as follows. In Section 2 we set up notation, definitions and some terminology. The fluid model considered in our work is introduced in Section 3.1; associated performance processes are considered in Section 3.2, where it is proved a result that establishes a useful relationship between them (Lemma 1). In Section 3.3 we introduce scaled processes, and Section 4 presents the main result, Theorem 1, that establishes the convergence, under heavy traffic, of the scaled workload process to a rfBm in some sense. Finally, Section 5 is devoted to the Functional Weak Law of Large Numbers (FWLLN), given by Theorem 2.

2 Preliminaries, notations and definitions

We will denote by I_d the $d-$dimensional identity matrix. Vectors will be column vectors unless indicated otherwise and v^T means the transpose of a vector (or a matrix) v. Given $v = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, hereafter we will denote by $diag(v)$ the $d \times d$ diagonal matrix with diagonal elements $v_1, \ldots, v_d$. For a $d \times d'$ matrix $A = (a_{ij})_{i=1,\ldots,d,\, j=1,\ldots,d'}$, let $|A| = \max\limits_{1\le i\le d} \Big(\sum\limits_{j=1}^{d'} |a_{ij}| \Big)$.

Let $\mathcal{C}^d$ be the space of continuous functions $\omega\colon [0, \infty) \to \mathbb{R}^d$, with the topology of the uniform convergence on compact time intervals. For any $T \ge 0$ and $\omega \in \mathcal{C}^d$, we define

$$||\omega(\cdot)||_T \stackrel{\text{def}}{=} \sup_{t\in[0,T]} |\omega(t)| = \sup_{t\in[0,T]} \Big(\max_{1\le \ell\le d} |\omega_\ell(t)| \Big).$$

We will say that $\omega^r \to \omega$ as $r \to \infty$ in $\mathcal{C}^d$ (*uniformly on compacts*) if for any

$T \geq 0$, $||\omega^r(\cdot) - \omega(\cdot)||_T \to 0$, and we will denote it by $\lim_{r\to\infty} \omega^r = \omega$.

We will use the following notations for diferent types of convergence:

$\mathcal{D}-\lim$ for the *convergence in distribution* on $\mathcal{C}^d$, $\mathrm{P}-\lim$ for the *convergence in probability (uniformly on compacts)*, and $\tilde{\lim}$ for the *convergence of the finite-dimensional distributions (in law)*. The convergence in probability is the following sense: we say that a family $\{X^r\}_r$ of random elements on $\mathcal{C}^d$ converges in probability to the random element X if for any $T > 0$ and for any $\varepsilon > 0$,

$$\lim_{r\to\infty} P\left(||X^r(\cdot) - X(\cdot)||_T \geq \varepsilon\right) = 0\,.$$

If X is a constant-valued random element (that is, an element of $\mathcal{C}^d$), the convergence in probability is equivalent to the convergence in the distribution sense.

Fractional Brownian motion (fBm for short), that is a stochastic process depending on a parameter $H \in (0,1)$, was first introduced in a celebrated paper by Mandelbrot and Van Ness ([6]) (case $H = 1/2$ corresponds to the Brownian motion process). Two fundamental properties justify the general interest on it from the modelling point of view: fBm is a self-similar process and it has long-range dependent increments, that are positively correlated if $1/2 < H < 1$ (the most frequently encountered in modelling). For the sake of completeness we give here its definition in the multidimensional case:

Definition 1.(*fBm*) A stochastic process $B^H = \{B^H(t) = (B_1^H(t), \ldots, B_J^H(t)), t \geq 0\}$, defined on some probability space, is called a J-dimensional 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, 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 Γ is a $J \times J$ positive semi-definite matrix and

$$\Gamma_H(s,t) = \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right).$$

For short, we will say that B^H is a $J-$dimensional fBm *with associated data* (x, H, θ, Γ).

Now we introduce a process that, loosely speaking, behaves like a fBm starting and being forced to live in the positive orthant $S = \mathbb{R}^J_+$. It is called *reflected fractional Brownian motion* (rfBm for short). Although it can be found in the literature, at least in the one-dimensional case (see [3] for instance), we have not found its general definition in the multi-dimensional case and this is the reason why we give it here and treat the question of its existence.

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

(i) W has continuous paths and $W(t) \in S = \mathbb{R}^J_+$ for all $t \geq 0$ a.s.,

(ii) $W = X + R\,Y$ a.s., with X and Y two $J-$dimensional processes defined on (Ω, $\mathcal{F}$, P), verifying:

(iii) X is a 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, Y_j can only increase when W is on face $F_j = \{y \in S = \mathbb{R}^J_+ : y_j = 0\}$).

We also say that the pair (W, Y) is a *$R-$regularization* of X, that (W, Y) is a solution of the *$R-$regularization problem* of X or that it is a solution of the *multidimensional Skorokhod problem* associated to X.

To get an idea, rfBm starts in the interior of S and behaves like a fBm until it touches the boundary of S, formed by faces F_j. Therefore, it is instantaneously "reflected", by avoiding the exit of S. For each j, the jth column of the *reflection matrix* R gives the direction of the reflection on face F_j, and component Y_j of process Y gives its intensity. Figure 1 shows the connection between the reflection angles on the edges and the reflection matrix R, for the case $J = 2$.

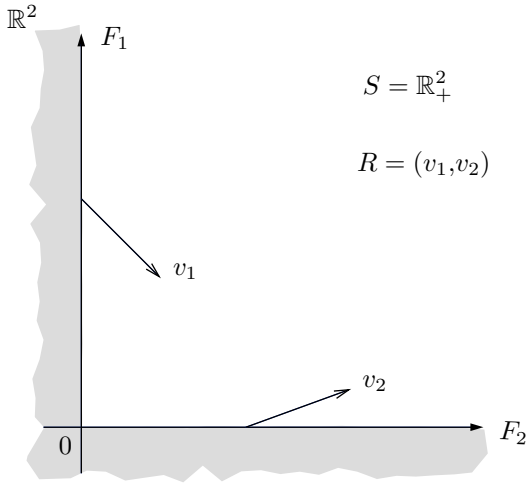

Fig. 1.

Remark 1. In the one-dimensional case, the existence of such a process is assured by [4] (see Theorem I.1.2 there) if $R > 0$. For the $J-$dimensional case, given a general process X on $(\Omega, \mathcal{F}, P)$, starting from x and with continuous

paths, and given a $J \times J$ matrix R, in order to ensure the existence of a pair (W, Y) verifying (i), (ii) and (iv) we will impose that matrix R be a *Completely-$\mathcal{S}$* matrix. Theorem 2 of [1] shows that the *completely-$\mathcal{S}$* property of matrix R is sufficient (and in some cases also necessary) for the existence of the $R-$*regularization* of X, although in the subsequent remark it is pointed out that if X is adapted to some filtration, the authors can not prove that process Y be also adapted to it. Nevertheless, Proposition 4.2 of [8] shows that under a stronger assumption on R, that we will denote by **(HR)**, this problem overcomes.

(HR) Assumption on matrix R:

$$\begin{aligned} &R \text{ can be expressed as } I_J + \Theta, \quad \text{with } \Theta \text{ a } J \times J \text{ matrix such} \\ &\text{that } |\Theta|, \text{that is the matrix obtained from } \Theta \text{ by replacing all the} \\ &\text{entries in } \Theta \text{ by their absolute values, has spectral radius less than } 1\,. \end{aligned} \tag{1}$$

Specifically, in proof of Proposition 4.2 of [8] it is shown that **(HR)** (condition (II) there) ensures the existence of a continuous (path-to-path) mapping π from the space of continuous $J-$dimensional paths $x(\cdot)$ starting in S into the space $\mathcal{C}^J \times \mathcal{C}^J$ of continuous paths $(w, y)(\cdot)$ living in $\mathbb{R}^J \times \mathbb{R}^J$, such that for each $x(\cdot)$, $(w, y)(\cdot) = \pi(x(\cdot))$ satisfies conditions (i), (ii) and (iv) of Definition 2 with $(w, x, y)(\cdot)$ instead of $(W, X, Y)(\cdot)$ and the a.s. omitted. For each $x(\cdot)$, $(w, y)(\cdot) = \pi(x(\cdot))$ is the unique pair in $\mathcal{C}^J \times \mathcal{C}^J$ with these properties, and the values of $(w, y)(\cdot) = \pi(x(\cdot))$ on $[0,\, t]$ depend only on the values of $x(\cdot)$ on $[0, t]$, for each $t \geq 0$. Then, as $(W, Y) = \pi(X)$ a.s., we have that if X is adapted to some filtration $\{\mathcal{F}_t,\, t \geq 0\}$, therefore (W, Y) is adapted to filtration $\{\mathcal{G}_t,\, t \geq 0\}$, with $\mathcal{G}_t = \mathcal{F}_t \vee \mathcal{N}$, where $\mathcal{N}$ denotes the collection of $P-$null sets in $\mathcal{F}$ (due to the "a.s"), and **(HR)** is a sufficient condition for

strong pathwise uniqueness of the solution of the $R-$*regularization* problem of X.

3 The fluid model

3.1 Introducing the model

We consider a network composed by J stations with a single server that processes continuous fluid, and an infinite buffer, at each one.

By following the ideas of [7] for a single station, first of all suppose that for any station j, there is only one external source sending fluid to it, and that the source can be ON or OFF. This source generates a stationary binary time series $\{U_j(t),\, t \geq 0\}$ where $U_j(t) = 1$ means that at time t the source is ON (and it is sending fluid to station j, at a traffic rate say $\alpha_j > 0$), and $U_j(t) = 0$ means that it is OFF. We supose that, independently of j, the lengths of the ON-periods are i.i.d., those of the OFF-periods are i.i.d., and the lengths of ON- and OFF-periods are independent. The ON- and OFF-periods lengths may have different distributions.

Let f_1 and f_2 be the probability density functions corresponding to the lengths of ON and OFF-periods, respectively, that are non-negatives and heavy-tailed. Therefore, their expected values and variances (all of them positive) are

$$\tilde{\mu}_i = \int_0^\infty u\, f_i(u)\, du \quad \text{and} \quad \sigma_i^2 = \int_0^\infty (u - \tilde{\mu}_i)^2\, f_i(u)\, du\,, \quad i = 1,\, 2\,.$$

Assume that as $x \to \infty$ we have

$$\int_x^\infty f_1(u)\, du \sim x^{-\beta_1}\, L_1(x) \quad \text{and} \quad \int_x^\infty f_2(u)\, du \sim x^{-\beta_2}\, L_2(x)\,,$$

with $1 < \beta_1, \beta_2 < 2$ and L_1, L_2 positive slowly varying functions at infinity. Note that $\tilde{\mu}_1$ and $\tilde{\mu}_2$ are always finite but variances σ_1^2 and σ_2^2 are infinite.

Suppose now that for each station j there are N i.i.d. sources, each one with its own binary time series $\{U_j^{(n)}(t),\, t \geq 0\}$, $n = 1, \ldots, N$, on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and that they are all independent. If all sources where ON, fluid would arrive at station j at deterministic rate $\alpha_j^N > 0$ (that we assume depending on N), and the cumulative *external fluid traffic* up to time t would be deterministic and equal to $\alpha_j^N\, t$ (this is the case for the fluid model introduced by Harrison in [5]). Instead, we define the cumulative external fluid arrived up to time t (by the N sources) at station j in this way:

$$E_j^N(t) \stackrel{\text{def}}{=} \alpha_j^N \int_0^t \frac{1}{N} \Big(\sum_{n=1}^{N} U_j^{(n)}(u) \Big)\, du \tag{2}$$

The $J-$dimensional (non-deterministic) *aggregated cumulative external fluid traffic* process, defined on $(\Omega, \mathcal{F}, \mathbb{P})$, is $E^N = \{E^N(t) = \Big(E_1^N(t), \ldots,\, E_J^N(t)\Big)^T$ $t \geq 0\}$, where the component processes are all independent. We suppose, for the sake of simplicity, that at time $t = 0$ there is no accumulated fluid at the network (that is, $E^N(0) = 0$). Let $\alpha^N = (\alpha_1^N, \ldots,\, \alpha_J^N)^T$.

Although other disciplines are possible, we assume that fluid at each server is processed in a first-in-first-out (FIFO) basis. When fluid arrives at station j and the server is busy, it must wait for service at its buffer, that we suppose without restriction of capacity. We consider that our service discipline is a *non-idling* (or *work-conserving*) policy, that means that a server is never idle when there are fluid waiting to be processed at its station.

Suppose that server at station j (server j for short) processes fluid at a constant rate $\mu_j > 0$ (independent of N) if that station were never idle. Let

$m_j = 1/\mu_j$ be the *mean service rate* for station j, $m = (m_1, \ldots, m_J)^T$, $\mu = (\mu_1, \ldots, \mu_J)^T$ and $M = diag(m)$.

Let $P_{j\ell}$ be the proportion of fluid that leaving station j goes next to station ℓ. We assume that for each j, $\sum_{\ell=1}^{J} P_{j\ell} \leq 1$ and $1 - \sum_{\ell=1}^{J} P_{j\ell} \geq 0$ is the proportion of fluid that leaving station j goes outside the network. Thus, $P = (P_{j\ell})_{j,\ell=1}^{J}$ is a sub-stochastic matrix. It is called the *"flow"* or *"routing" matrix of the network*, and it is assumed to have spectral radius less than one. Hence, $Q \stackrel{\text{def}}{=} \left(I_J - P^T\right)^{-1}$ is well defined. Figure 2 shows the flow into and out of the system, and between stations (feedback) for the particular case $J = 3$.

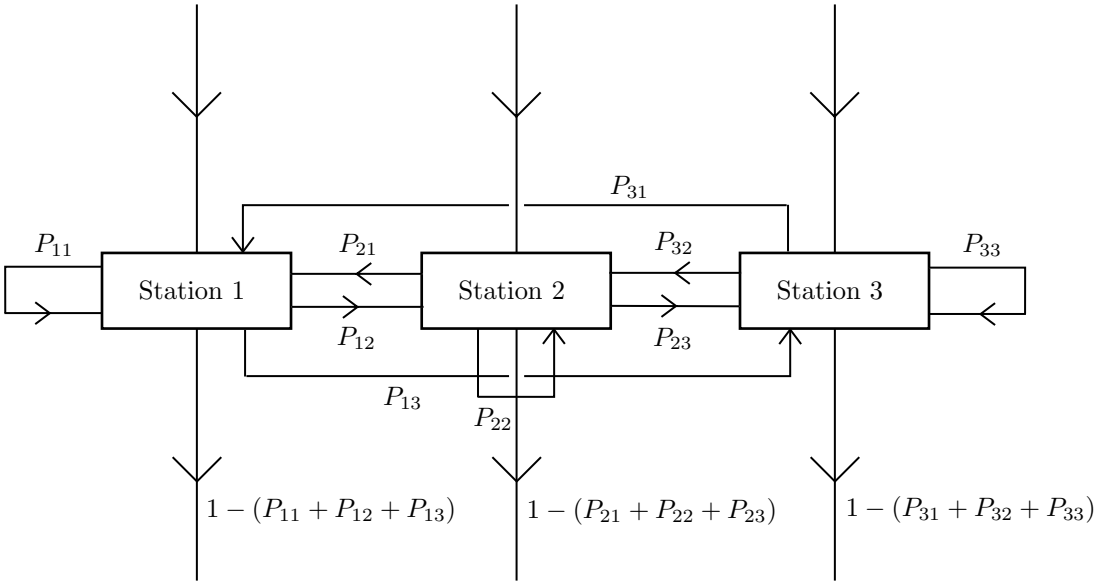

Fig. 2.

We define λ^N to be the unique $J-$dimensional vector solution to the *traffic equation*

$$\lambda^N \stackrel{\text{def}}{=} \alpha^N \frac{\tilde{\mu}_1}{\tilde{\mu}_1 + \tilde{\mu}_2} + P^T \lambda^N \quad (\text{that is} \quad \lambda^N = Q\, \alpha^N \frac{\tilde{\mu}_1}{\tilde{\mu}_1 + \tilde{\mu}_2}). \tag{3}$$

We note that for any j, λ_j^N can be interpreted as the long run fluid rate into and out of station j. The technical justification for that can be seen in

Theorem 2, Section 5.

We also define the *fluid traffic intensity* for station j as

$$\rho_j^N \stackrel{\text{def}}{=} m_j\,\lambda_j^N \quad \Big(\text{in matricial form, } \rho^N = M\,\lambda^N\Big)\,.$$

The main result of this work will be proved under a *heavy traffic condition*, that establishes that the total load imposed on each service station tends to be equal to its capacity, that is, its traffic intensity tends to be equal to 1, in the following sense:

$$\textbf{(HT)} \qquad \lim_{N\to\infty} \sqrt{N}\,(\rho^N - e) = 0 \quad \text{where } e = 1 \in \mathbb{R}^J\,. \tag{4}$$

Note that under the previous condition we deduce the existence of

$$\lim_{N\to\infty} \alpha^N = \alpha\,, \quad \text{with} \quad \alpha = \frac{\tilde{\mu}_1 + \tilde{\mu}_2}{\tilde{\mu}_1}\,Q^{-1}\,M^{-1}\,e\;(>0) \tag{5}$$

that is the limiting value for the external arrival rate needed to achieve the maximum capacity of the system. We can also deduce, from the definition of λ^N, the existence of

$$\lim_{N\to\infty} \lambda^N = \lambda \quad \text{with} \quad \lambda = M^{-1}\,e\,. \tag{6}$$

3.2 Performance processes

Two descriptive ($J-$dimensional) processes will be used to measure the performance of the queueing network:

The immediate *workload process* W^N, defined by: $W_j^N(t)$ denotes the amount of time required for server j to complete processing of all fluids in queue (or being served) at station j at time t. We assume that $W^N(0) = 0\,.$

The *cumulative idle-time process* Y^N, defined by: $Y_j^N(t)$ is the cumulative amount of time that server j has been idle in the time interval $[0,\, t]$, that is,

$$Y_j^N(t) \stackrel{\text{def}}{=} \int_0^t 1_{\{W_j^N(s)=0\}}\, ds\,. \tag{7}$$

Immediate workload process measures the congestion and delay in the network, while *idle-time* process measures utilization of resources.

Apart from them, there are other interesting ($J-$dimensional) processes to be considered in the fluid model, as processes A^N and D^N, that will appear in the proof of Lemma 1, and that are also interesting by themselves. In Section 5 we will obtain a Functional Weak Law of Large Numbers for them. Definition is as follows: $A_j^N(t)$ is the total fluid arriving to station j up to time t, including both feedback flow and external input, and $D_j^N(t)$ is the total amount of fluid departing station j (both being routing to other station or leaving the network), up to time t. We assume $A^N(0) = D^N(0) = 0$.

Our objective now is to show how aggregated cumulative external fluid traffic, workload and cumulative idle-time processes are related by means of the following result.

Lemma 1. We have that

$$W^N(t) = R\,M\,Q\,E^N(t) - R\,e\,t + R\,Y^N(t)\,, \tag{8}$$

where

$$R \stackrel{\text{def}}{=} (I_J + M\,Q\,P^T\,M^{-1})^{-1}\,. \tag{9}$$

Note that matrix R is well defined because $I_J + M\,Q\,P^T\,M^{-1} = M\,Q\,M^{-1}$ has

inverse, that is $M\,Q^{-1}\,M^{-1} = M\,(I_J - P^T)\,M^{-1}$, and therefore

$$R = I_J - M\,P^T\,M^{-1}\,. \tag{10}$$

Moreover, R verifies condition **(HR)**: by (10) we have that $R = I_J + \Theta$ with $\Theta = -M\,P^T\,M^{-1}$, and $|\Theta|$ has the same spectral radius as P, that is assumed to be less than 1.

Proof of Lemma 1:

First of all note that

$$W^N(t) = M\,A^N(t) - e\,t + Y^N(t)\,. \tag{11}$$

Expression (11) is justified by the definition of $W_j^N(t)$ as the amount of time required for server at station j to complete processing of all fluid buffered or being served at that station at time t, that equals to $M\,A_j^N(t)$, the cumulative total amount of time required for server at station j to complete processing of fluid arrived to that station up to time t, minus the amount of time, $t - Y_j^N(t)$, that that server has been busy (working) up to time t.

We can write

$$A^N(t) = E^N(t) + F^N(t)$$

where $F_\ell^N(t) = \sum_{j=1}^{J} P_{j\ell}\,D_j^N(t)$ is the total amount of fluid that arrives from *feedback* to station ℓ (due to the fraction of the amount $D_j^N(t)$ of fluid that leaving station j is next routed to station ℓ, summed over the totality of stations). That is,

$$A^N(t) = E^N(t) + P^T\,D^N(t)\,. \tag{12}$$

It is immediate to realize that

$$D^N(t) = A^N(t) - M^{-1}\, W^N(t)\,, \tag{13}$$

that is, the total amount of fluid leaving station j up to time t is the total amount of fluid arriving station j up to time t minus the fluid in queue or being processed at that station at time t, that is $\mu_j\, W_j^N(t)$, by definition of the workload process. By substituting this expression into (12) we have

$$A^N(t) = E^N(t) + P^T\,(A^N(t) - M^{-1}\, W^N(t))\,,$$

and taking into account the definition of matrix Q,

$$A^N(t) = Q\,(E^N(t) - P^T\, M^{-1}\, W^N(t))\,,$$

that replaced into (11) gives

$$W^N(t) = M\, Q\, E^N(t) - M\, Q\, P^T\, M^{-1}\, W^N(t) - e\, t + Y^N(t)\,,$$

and then obtain (8) just by using the definition of matrix R given by (9). $\square$

3.3 Scaled processes

In order to define the *scaled processes* (in space, by a factor $\sqrt{N}$ and in time by a factor r) associated to the fluid model, we must introduce previously some notation used in [7]. For any $j = 1, 2,$ set $a_j = \frac{\Gamma(2-\beta_j)}{(\beta_j - 1)}$. The normalization factors used below depend on whether b, defined by $b \stackrel{\text{def}}{=} \lim_{t\to\infty} t^{\beta_2-\beta_1}\, \frac{L_1(t)}{L_2(t)}$, is finite, zero, or infinite. If $0 < b < \infty$ (implying $\beta_1 = \beta_2$ and $b = \lim\limits_{t\to\infty} \dfrac{L_1(t)}{L_2(t)}$), set $\beta_{min} = \beta_1 = \beta_2$, $L = L_2$ and

$$\sigma^2_{lim} \stackrel{\text{def}}{=} \frac{2\left(\tilde{\mu}_2^2\, a_1\, b + \tilde{\mu}_1^2\, a_2\right)}{\left(\tilde{\mu}_1 + \tilde{\mu}_2\right)^3 \Gamma(4 - \beta_{min})}$$

(factor σ^2_{lim} does not appear in scaling, but will play an important role in the sequel). If, on the other hand, $b = 0$ or $b = \infty$, set $L = L_{min}$ and

$$\sigma^2_{lim} \stackrel{\text{def}}{=} \frac{2\,\tilde{\mu}^2_{max}\,a_{min}}{\left(\tilde{\mu}_1 + \tilde{\mu}_2\right)^3 \Gamma(4 - \beta_{min})}$$

where *min* is the index 1 if $b = \infty$ and 2 if $b = 0$; *max* denoting the other index.

In either case, $\beta_{min} \in (1,\, 2)$. Let we define $H \stackrel{\text{def}}{=} \frac{3-\beta_{min}}{2}$. Therefore, $H \in (\frac{1}{2},\, 1)$.

Now we can introduce the *scaled processes* associated to the fluid model. We will use a hat to denote them:

$$\hat{W}^{r,N}(t) \stackrel{\text{def}}{=} \sqrt{N}\,\frac{W^N(r\,t)}{r^H\,L^{1/2}(r)} \tag{14}$$

$$\hat{E}^{r,N}(t) \stackrel{\text{def}}{=} \sqrt{N}\,\frac{E^N(r\,t) - \alpha^N\,r\,t\,\frac{\tilde{\mu}_1}{\tilde{\mu}_1+\tilde{\mu}_2}}{r^H\,L^{1/2}(r)} \tag{15}$$

$$\hat{Y}^{r,N}(t) \stackrel{\text{def}}{=} \sqrt{N}\,\frac{Y^N(r\,t)}{r^H\,L^{1/2}(r)}\,, \tag{16}$$

and we can see that they are related by means of

$$\hat{W}^{r,N}(t) = \hat{X}^{r,N}(t) + R\,\hat{Y}^{r,N}(t)\,, \tag{17}$$

with

$$\hat{X}^{r,N}(t) = R\,M\,Q\,\hat{E}^{r,N}(t) + \frac{\sqrt{N}}{r^H\,L^{1/2}(r)}\,R\,(\rho^N - e)\,r\,t\,. \tag{18}$$

Proof of formula (17):

By definition (see (14)) and (8), we can write

$$\hat{W}^{r,N}(t) = \frac{\sqrt{N}}{r^H\,L^{1/2}(r)}\left(R\,M\,Q\,E^N(r\,t) - R\,e\,r\,t + R\,Y^N(r\,t)\right),$$

that can be expressed as the sum of the following three factors:

$$\frac{\sqrt{N}}{r^H\, L^{1/2}(r)}\, R\, M\, Q\left(E^N(r\,t) - \alpha^N\, r\, t\, \frac{\tilde{\mu}_1}{\tilde{\mu}_1+\tilde{\mu}_2}\right) = R\, M\, Q\, \hat{E}^{r,N}(t)\,,$$

$$\frac{\sqrt{N}}{r^H\, L^{1/2}(r)}\, R\left(M\, Q\, \alpha^N\, r\, t\, \frac{\tilde{\mu}_1}{\tilde{\mu}_1+\tilde{\mu}_2} - e\, r\, t\right) = \frac{\sqrt{N}}{r^H\, L^{1/2}(r)}\, R\,(\rho^N - e)\, r\, t\,,$$

and $R\,\hat{Y}^{r,N}(t)$. $\square$

Remark 2. Note that processes appearing in expression (17) verify: $\hat{W}^{r,N}$ has continuous paths; for any $t \geq 0$, a.s. $\hat{W}^{r,N}(t) \in S = \mathbb{R}_+^J$; $\hat{Y}^{r,N}$ has continuous and non-decreasing paths, and for each j, a.s. $\hat{Y}_j^{r,N}(0) = 0$ and

$$\int_0^\infty \hat{W}_j^{r,N}(s)\, d\hat{Y}_j^{r,N}(s) = 0 \quad \Big(\text{equivalently,} \quad \int_0^\infty 1_{\{\hat{W}_j^{r,N}(s)>0\}}\, d\hat{Y}_j^{r,N}(s) = 0\Big)\,.$$

4 The main result

Our goal now is to prove that the scaled workload process, $\hat{W}^{r,N}$, converges to a rfBm in some sense, as N and r increase. This is established in the following result:

Theorem 1.

Under heavy traffic (condition **(HT)**), we have that there exist the limits

$$\hat{\hat{W}}^r = \lim_{N\to\infty}{}^{\sim}\, \hat{W}^{r,N} \quad \text{and} \quad W = \mathcal{D}\text{-}\lim_{r\to\infty} \hat{\hat{W}}^r\,,$$

and that W is a drift-less rfBm on $S = \mathbb{R}_+^J$ with associated data

$$(x = 0,\, H = \frac{3-\beta_{min}}{2},\, \theta = 0,\, \Gamma,\, R)\,,$$

where $\Gamma = \sigma_{lim}^2\, R\, M\, Q\, diag(\alpha)^2\, Q^T\, M\, R^T$, with α given by (5), and R the matrix introduced in Lemma 1.

Proof of Theorem 1.

Let us first note that $\hat{E}_j^{r,N}(t)$, defined by (15) and (2), can be written in the following way:

$$\hat{E}_j^{r,N}(t) = \alpha_j^N \, \frac{\int_0^{rt} \varphi_j^N(u)\, du}{r^H \, L^{1/2}(r)}\,, \quad \text{where} \quad \varphi_j^N(u) \stackrel{\text{def}}{=} \frac{1}{N^{1/2}} \sum_{n=1}^{N} \left(U_j^{(n)}(u) - \frac{\tilde{\mu}_1}{\tilde{\mu}_1 + \tilde{\mu}_2} \right).$$

The convergence of $\hat{E}^{r,N}$ to a process B^H, that is $J-$dimensional drift-less fractional Brownian motion with associated data ($x = 0$, $H = \frac{3-\beta_{min}}{2}$, $\theta = 0$, $\Gamma = \sigma_{lim}^2 \, diag(\alpha)^2$) , is proved as in Taqqu et al. ([7]). This convergence is in the sense that there exist the limit

$$\hat{\tilde{E}}^r = \lim_{N\to\infty}\tilde{} \, \hat{E}^{r,N} = \alpha^T \, \frac{\int_0^{r\cdot} G(u)\, du}{r^H \, L^{1/2}(r)}\,, \tag{19}$$

with $\{G(t),\, t \geq 0\}$ some $J-$dimensional drift-less gaussian and stationary process, and that

$$\mathcal{D}- \lim_{r\to\infty} \hat{\hat{E}}^r = B^H\,.$$

Combining (18), heavy traffic condition **(HT)**, and the *continuous mapping theorem* (see Corollary 1 of Theorem 5.10 in [2]), we can assert that there exists $\hat{\hat{X}}^r = \lim_{N\to\infty}\tilde{} \, \hat{X}^{r,N}$, with

$$\hat{\hat{X}}^r = R\, M\, Q\, \hat{\hat{E}}^r \tag{20}$$

and that there also exists $\mathcal{D}- \lim_{r\to\infty} \hat{\hat{X}}^r = X\,,$ with $X = R\, M\, Q\, B^H\,,$ that is a $J-$dimensional fBm with associated data $\Big(x = 0,\, H = \frac{3-\beta_{min}}{2},\, \theta = 0,\, \Gamma\Big)$, Γ being the matrix given by

$$\Gamma = \sigma_{lim}^2 \, R\, M\, Q\, diag(\alpha)^2\, Q^T\, M\; R^T\,.$$

We proceed now to show the corresponding convergence for processes $\hat{W}^{r,N}$

and $\hat{Y}^{r,N}$, by taking into account that matrix R verifies condition **(HR)**, as is pointed out in Lemma 1.(20) makes it obvious that $\hat{\hat{X}}^r$ has continuous paths, because $\hat{\hat{E}}^r$ does (which is clear from (19)). Comments made on last paragraph of Remark 1 show that, in this situation, there exists a unique strong pathwise solution of the $R-$regularization problem of $\hat{\hat{X}}^r$, that coincides with

$$\left(\tilde{\lim_{N\to\infty}} \hat{W}^{r,N}, \tilde{\lim_{N\to\infty}} \hat{Y}^{r,N} \right), \quad \text{by} \quad (17)\,.$$

If we denote $\tilde{\lim_{N\to\infty}} \hat{Y}^{r,N}$ by $\hat{\hat{Y}}^r$ and $\tilde{\lim_{N\to\infty}} \hat{W}^{r,N}$ by $\hat{\hat{W}}^r$, we have that the unique solution of the $R-$regularization problem of $\hat{\hat{X}}$ is $(\hat{\hat{W}}^r, \hat{\hat{Y}}^r)$, and then

$$\hat{\hat{W}}^r = \hat{\hat{X}}^r + R\,\hat{\hat{Y}}^r\,. \tag{21}$$

This fact implies that $\hat{\hat{W}}^r$, $\hat{\hat{X}}^r$ and $\hat{\hat{Y}}^r$ verify hypotheses of the *invariant principle* of Theorem 4.1 in [8], taking into account that $\mathcal{D}-\lim_{r\to\infty} \hat{\hat{X}}^r = X$, and that R is a *Completely-*$\mathcal{S}$ matrix, because it verifies **(HR)**. We have that $\left\{\left(\hat{\hat{W}}^r, \hat{\hat{X}}^r, \hat{\hat{Y}}^r\right)\right\}_r$ inherits tightness from sequence $\left\{\hat{\hat{X}}^r\right\}_r$ and consequently, by assumption **(HR)** on R (see Corollary 4.3 of [8]), there exists $\mathcal{D}-\lim_{r\to\infty} \left(\hat{\hat{W}}^r, \hat{\hat{X}}^r, \hat{\hat{Y}}^r\right) = (W, X, Y)$, where $W = X + R\,Y$ and conditions of Definition 2 are satisfied. Therefore, W is a rfBm on $S = \mathbb{R}^J_+$ with associated data $(x = 0,\, H = \frac{3-\beta_{min}}{2},\, \theta = 0,\, \Gamma,\, R)$.

We mention that Theorem 4.1 in [8] gives the convergence in the distributional sense on $\mathcal{D}^J$, the space of functions from $[0,\infty)$ to $\mathbb{R}^J$ that are right continuous and have finite left hand limits, with the Skorokhod topology. Our convergence is taken in the distributional sense on $\mathcal{C}^J$, and is implied by the convergence on $\mathcal{D}^J$ because the Skorokhod topology relativized to $\mathcal{C}^J$ coincides with the uniform topology over compacts. $\square$

5 Functional Weak Law of Large Numbers (FWLLN) for processes A^N and D^N

In this section we will prove a Functional Weak Law of Large Numbers for processes A^N and D^N introduced in Section 3.2, that are the total amount of fluid arriving and departing stations up to any time, respectively. This result gives support to the interpretation of λ, that is the solution of the limiting traffic equation $\lambda = Q\,\alpha\dfrac{\tilde{\mu}_1}{\tilde{\mu}_1+\tilde{\mu}_2}$ (as can be seen by combining (5) with (6)), as the long run fluid rate into and out of the system.

Let us first introduce the associated *scaled processes*

$$\hat{A}^{r,N}(t) \stackrel{\text{def}}{=} \sqrt{N}\;\frac{A^N(r\,t)-\lambda^N\,r\,t}{r} \tag{22}$$

$$\hat{D}^{r,N}(t) \stackrel{\text{def}}{=} \sqrt{N}\;\frac{D^N(r\,t)-\lambda^N\,r\,t}{r} \tag{23}$$

Theorem 2. *(FWLLN for processes A^N and D^N)*

Under heavy traffic (condition **(HT)**), we have that there exist the limits

$$\hat{\hat{A}}^r = \lim_{N\to\infty}\tilde{}\,\hat{A}^{r,N} \quad\text{and}\quad \hat{\hat{D}}^r = \lim_{N\to\infty}\tilde{}\,\hat{D}^{r,N}\,,$$

and

$$\mathcal{D}-\lim_{r\to\infty}\hat{\hat{A}}^r = \mathcal{D}-\lim_{r\to\infty}\hat{\hat{D}}^r = 0$$

Proof of Theorem 2:

The proof falls naturally into two parts. We first justify the existence of $\hat{\hat{A}}^r$ and $\hat{\hat{D}}^r$:

By (22), (23), (13) and (14), we have that

$$\begin{aligned}\hat{A}^{r,N}(\cdot) - \hat{D}^{r,N}(\cdot) =& \frac{\sqrt{N}}{r}\Big(A^N(r\,\cdot) - D^N(r\,\cdot)\Big) = \frac{\sqrt{N}}{r}\, M^{-1}\, W^N(r\,\cdot) \\ =& \frac{r^H\, L^{1/2}(r)}{r}\, M^{-1}\, \hat{W}^{r,N}(\cdot)\,.\end{aligned}$$

Then, by Theorem 1, there exists

$$\tilde{\lim_{N\to\infty}}\Big(\hat{A}^{r,N} - \hat{D}^{r,N}\Big) = \frac{r^H\, L^{1/2}(r)}{r}\, M^{-1}\, \hat{\hat{W}}^r\,. \tag{24}$$

Now, by using (12), we have that

$$\begin{aligned}\frac{\sqrt{N}}{r}A^N(r\,t) =& \frac{\sqrt{N}}{r}\Big(E^N(r\,t) + P^T\, D^N(r\,t)\Big) \\ =& \frac{\sqrt{N}}{r}\Big(E^N(r\,t) + P^T\,\Big(D^N(r\,t) - A^N(r\,t)\Big) + P^T\, A^N(r\,t)\Big)\,,\end{aligned}$$

and therefore

$$\frac{\sqrt{N}}{r}A^N(r\,t) = Q\,\frac{\sqrt{N}}{r}\,E^N(r\,t) + Q\,\frac{\sqrt{N}}{r}\,P^T\,\Big(D^N(r\,t) - A^N(r\,t)\Big)\,.$$

It follows, by applying (15), that

$$\begin{aligned}\frac{\sqrt{N}}{r}A^N(r\,t) =& Q\,\Big(\frac{r^H\, L^{1/2}(r)}{r}\,\hat{E}^{r,N}(t) + \sqrt{N}\,\alpha^N\, t\,\frac{\tilde{\mu}_1}{\tilde{\mu}_1+\tilde{\mu}_2}\Big) + \\ &+ Q\,\frac{\sqrt{N}}{r}\,P^T\,\Big(D^N(r\,t) - A^N(r\,t)\Big)\,,\end{aligned}$$

and by combining this expression with (22), (23) and (3), we get

$$\hat{A}^{r,N} = \frac{r^H\, L^{1/2}(r)}{r}\, Q\,\hat{E}^{r,N} + Q\,P^T\,\Big(\hat{D}^{r,N} - \hat{A}^{r,N}\Big)\,.$$

Therefore, we can conclude from (24) and Theorem 1 again, that there exists

$$\hat{\hat{A}}^r = \tilde{\lim_{N\to\infty}}\,\hat{A}^{r,N} = \frac{r^H\, L^{1/2}(r)}{r}\, Q\,\Big(\hat{\hat{E}}^r - P^T\, M^{-1}\,\hat{\hat{W}}^r\Big)\,, \tag{25}$$

and combining (24) with (25) we deduce the existence of

$$\hat{\hat{D}}^r = \tilde{\lim_{N\to\infty}}\,\hat{D}^{r,N} = \frac{r^H\, L^{1/2}(r)}{r}\,\Big(Q\,\hat{\hat{E}}^r - \Big(I_J + Q\,P^T\Big)\, M^{-1}\,\hat{\hat{W}}^r\Big)\,. \tag{26}$$

The second part of the proof consists in showing the existence of the limits

$$\mathcal{D}-\lim_{r\to\infty}\hat{\hat{A}}^r=\mathcal{D}-\lim_{r\to\infty}\hat{\hat{D}}^r\,\Big(=0\Big)\,,$$

that is equivalent to prove that for any $T>0$ and for any $\varepsilon>0$,

$$\lim_{r\to\infty}P\Big(||\hat{\hat{A}}^r(\cdot)||_T\geq\varepsilon\Big)=\lim_{r\to\infty}P\Big(||\hat{\hat{D}}^r(\cdot)||_T\geq\varepsilon\Big)=0\,.$$

We will see that this is the case for $\hat{\hat{A}}^r$ (the same conclusion can be drawn for $\hat{\hat{D}}^r$). Specifically, we will check that if we fix $T>0$ and $\varepsilon>0$, for any $\delta>0$ there exists r_0 such that if $r\geq r_0$,

$$P\left(||\hat{\hat{A}}^r(\cdot)||_T\geq\varepsilon\right)\leq\delta\,.$$

Indeed, we obtain from (25) that

$$||\hat{\hat{A}}^r(\cdot)||_T\leq\frac{r^H\,L^{1/2}(r)}{r}\left(|Q|\,||\hat{\hat{E}}^r(\cdot)||_T+|Q\,P^T\,M^{-1}|\,||\hat{\hat{W}}^r(\cdot)||_T\right),$$

and consequently

$$\begin{aligned}P\left(||\hat{\hat{A}}^r(\cdot)||_T\geq\varepsilon\right)\leq{}&P\left(\frac{r^H\,L^{1/2}(r)}{r}\,|Q|\,||\hat{\hat{E}}^r(\cdot)||_T\geq\frac{\varepsilon}{2}\right)+\\&+P\left(\frac{r^H\,L^{1/2}(r)}{r}\,|Q\,P^T\,M^{-1}|\,||\hat{\hat{W}}^r(\cdot)||_T\geq\frac{\varepsilon}{2}\right)\end{aligned}\tag{27}$$

Since $\mathcal{D}-\lim_{r\to\infty}\hat{\hat{E}}^r=B^H$ and process B^H is continuous, as we have seen in Theorem 1, by the *continuous mapping theorem* applied to $||\cdot||_T$ we have that given $\delta>0$, there exist $K_\delta>0$ and r_1 such that for any $r\geq r_1$,

$$P\left(||\hat{\hat{E}}^r(\cdot)||_T<K_\delta\right)\geq1-\frac{\delta}{2}\,.$$

Moreover, taking into account that

$$\lim_{r\to\infty}\frac{r^H\,L^{1/2}(r)}{r}=\lim_{r\to\infty}\left(\frac{L(r)}{r^{\beta_{min}-1}}\right)^{1/2}=0\tag{28}$$

(justified by the fact that $\beta_{min} - 1 \in (0,1)$ and L is a slowly varying function at infinity), we can ensure that there exists r_2 such that for any $r \geq r_2$,

$$\frac{r^H\, L^{1/2}(r)}{r} < \frac{1}{K_\delta}\,\frac{1}{|Q|}\,\frac{\varepsilon}{2}\,.$$

As a consequence, if $r \geq r_1 \vee r_2\, (= \max(r_1, r_2))$,

$$1 - \frac{\delta}{2} \leq P\left(||\hat{\hat{E}}^r(\cdot)||_T < K_\delta\right) \leq P\left(\frac{r^H\, L^{1/2}(r)}{r}\ |Q|\ ||\hat{\hat{E}}^r(\cdot)||_T < \frac{\varepsilon}{2}\right),$$

that is,

$$\forall r \geq r_1 \vee r_2\,, \quad P\left(\frac{r^H\, L^{1/2}(r)}{r}\ |Q|\ ||\hat{\hat{E}}^r(\cdot)||_T \geq \frac{\varepsilon}{2}\right) \leq \frac{\delta}{2}\,.$$

By applying again Theorem 1, we have that $\mathcal{D}\!-\lim_{r\to\infty} \hat{\hat{W}}^r = W$, with W a continuous process, and we can now proceed analogously to obtain that there exists r_3 such that

$$\forall r \geq r_3\,, \quad P\left(\frac{r^H\, L^{1/2}(r)}{r}\ |Q\, P^T\, M^{-1}|\ ||\hat{\hat{W}}^r(\cdot)||_T \geq \frac{\varepsilon}{2}\right) \leq \frac{\delta}{2}\,.$$

This finishes the proof, by (27), because we have shown that there exists r_0 (the maximum of r_1, r_2 and r_3) such that

$$\forall r \geq r_0\,, \quad P\left(||\hat{\hat{A}}^r(\cdot)||_T \geq \varepsilon\right) \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta\,. \quad \square$$

References

[1] A. Bernard, A. el Kharroubi, Régulations déterministes et stochastiques dans le premier orthant de $\mathbb{R}^n$, *Stochastics and Stochastics Reports* **34** (1991) 149-167.

[2] P. Billingsley, Convergence of Probability Measures. *Wiley*, (1968).

[3] K. Debicki, M. Mandjes, Traffic with an fBm limit: Convergence of the Stationary Workload process, *Queueing Systems* **46** (2004) 113-127.

[4] N. El Karoui, M. Chaleyat-Maurel, Un problème de réflexion et ses aplications au temps local et aux équations différentielles stochastiques sur $\mathbb{R}$. Cas continu, *Société Mathématique de France, Astérisque* **52-53** (1978) 117-144.

[5] J. M. Harrison, Balanced fluid models of multiclass queueing networks: a heavy traffic conjecture, in: Frank P. Kelly, Ruth J. Williams (Eds.), *Stochastic Networks*, IMA Volumes in Mathematics and its Applications, **71**, Springer-Verlag, New-York, 1995, pp. 1-20.

[6] B. B. Mandelbrot, J. W. Van Ness, Fractional Brownian Motions, Fractional Noises and Applications, *SIAM Review* **10 (4)** (1968) 422-437.

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

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