Queueing Syst (2013) 74:41–63
DOI 10.1007/s11134-012-9320-z

Heavy-traffic limit for a feed-forward fluid model with heterogeneous heavy-tailed On/Off sources

Rosario Delgado

Received: 6 October 2011 / Revised: 10 May 2012 / Published online: 30 June 2012

Abstract We consider a multi-station fluid model with arrivals generated by a large number of non-homogeneous heavy-tailed On/Off sources. If the model is feed-forward in the sense that fluid cannot flow from one station to other with lighter tail distributions, we prove that under heavy-traffic, the scaled workload converges in distribution to a reflected fractional Brownian motion process with a multi-dimensional Hurst parameter. As an application, we analyze the impact of having independent streams with variable parameters in high-speed telecommunication networks, on the asymptotic behavior of the maximum queue length.

Keywords Fluid model · Heavy tails · Heavy traffic · Reflected fractional Brownian motion · Heterogeneous On/Off sources · Workload process · Maximum queue length · Skorokhod problem

Mathematics Subject Classification (2000) 60K25 · 60F05 · 60G15 · 60G18 · 60G22

1 Introduction

In the last decade of the twentieth century, researches became aware of the presence of long-range dependence and self-similarity in modern high-speed network traffic, especially in Internet traffic. As the fractional Brownian motion process with Hurst parameter $H > 1/2$ has the properties of long-range dependence and self-similarity, it has been used since then to built different models for these complex networks.

R. Delgado (✉)
Departament de Matemàtiques, Universitat Autònoma de Barcelona, Edifici C Campus de la UAB, 08193 Bellaterra (Cerdanyola del Vallès), Barcelona, Spain
e-mail: delgado@mat.uab.cat

From the well-known work of Taqqu, Willinger, and Sherman [12], it has been generally accepted that one simple physical explanation for the observed phenomenon of the long-range dependence and self-similarity, consists in the superposition of many On/Off sources with strictly alternating On- and Off-periods whose lengths are heavy-tailed distributed. The fact 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 by a factor of r and $\sqrt{N}$, converges in some sense as first N tends to infinity and then r tends to infinity, to a *fractional Brownian motion* process, was proved in Theorem 1 of the paper of Taqqu, Willinger, and Sherman [12], where the authors show the relationship between the parameter describing the heaviness of the tails and the Hurst parameter of the fractional Brownian motion, which measures its degree of self-similarity.

Subsequent work has shown that the convergence of the aggregate cumulative arrival process to a fractional Brownian motion carries over to the stationary buffer content process in the heavy traffic scenario: the scaled workload process has been proved to converge to a fractional Brownian motion process but reflected appropriately to be non-negative, both in the case of single-station fluid models and in the multi-station environment (see Debicki and Mandjes [3], and Delgado [4], respectively).

More specifically, in Delgado [4] a non-deterministic fluid model which consists of d stations with a single server that processes fluid in the arrival order, an infinite buffer at each station and possible feedback routing is considered. The process of external arrivals was taken to be a non-deterministic aggregated cumulative process generated by a large enough number of homogeneous heavy tailed On/Off sources. For each station, there are a large number of sources sending fluid to it, and these sources can be On (sending fluid to the station at a constant traffic rate) or they can be Off. The lengths of the On- and Off-periods are heavy-tailed with tail decay as a power function, assumed to be the same for the d stations (but not necessarily the same for the On- and the Off-periods) in Delgado [4].

The present paper has a twofold motivation. First, one might wonder if a generalization of the heavy-traffic limit theorem of Delgado [4] could be proved if sources were heterogeneous, that is, if the power functions determining the decay of the tails for the On- and the Off-periods, were allowed to vary from one to another station. A motivation for that is the fact that some networks with independent streams with variable parameters appear in applied research. For instance, in Fitzpatrick, Murphy, and Murphy [8], a transport layer handover mechanism for Voice over Internet Protocol (VoIP) using the Stream Control Transmission Protocol (SCTP), which operates in "heterogeneous transmission rate networks" is considered. Indeed, this mechanism is shown to operate in WLAN networks where each node can communicate with the Access Point (AP) at different transmission rates.

We see in this paper that, in fact, it is possible: In Theorem 1, we prove that after adequate scaling and under heavy traffic conditions, the immediate workload process converges to a *d-dimensional reflected fractional Brownian motion* process on the positive orthant $\mathbb{R}^d_+$ with drift $\theta \in \mathbb{R}^d$, *completely-$\mathscr{S}$* reflection matrix R defined from the flow matrix associated with the fluid model, and multi-dimensional Hurst parameter $H = (H_1, \ldots, H_d)^T \in (\frac{1}{2}, 1)^d$.

That is, we extend the heavy-traffic limit given by Theorem 1 of Delgado [4] to this more general setting. The proof follows similar ideas but presents some interesting differences. On one hand, heterogeneity of the sources forces us to impose some restrictions, the most important of which, denoted by (**HP**), is a *feed-forward* condition in the sense that fluid cannot flow from one station to another with strictly lighter tail distributions. On the other hand, we introduce here a *heavy traffic* condition which generalizes that used in Delgado [4]: fixed r, the traffic intensity tends to 1 as $N \to +\infty$ in the sense that the difference multiplied by $\sqrt{N}$ converges to $-\hat{\gamma}^r \leq 0$ (assumed to be 0 in Delgado [4]), where vector $\hat{\gamma}^r$ converges to 0 as $r \to +\infty$ in the sense that if multiplied by some fixed power of r, it converges to some $\gamma \geq 0$. Because it seems interesting, we highlight both the expression in terms of the flow matrix P of the reflection matrix associated with the multidimensional reflected fractional Brownian motion appearing in Theorem 1, R, and also that the drift vector turns out to be $\theta = -R\gamma$. And last but not least, a major difference is that the proof of Theorem 1 will rely heavily on Theorem 7.2.5 of Whitt [13], which represents an improvement of Theorem 1 [12] in two ways: first, it establishes that the limit as $N \to +\infty$ is actually in distribution and not only in the sense of the convergence of the finite dimensional distributions, and secondly, because it fills a gap in the proof of the convergence in distribution as $r \to +\infty$ given in Theorem 1 [12].

In Corollary 1, we prove a Functional Weak Law of Large Numbers (FWLLN) for the total amount of fluid arriving to the stations (including both feedback flow and external input), and also for the total amount of leaving fluid from the stations (to other stations or outside the system). This result, which generalizes Theorem 2 of Delgado [4], justifies the interpretation of the solution to the limiting traffic equation as the long run fluid rate into and out of any station.

On the other hand, as second motivation we have considered the question of the impact of having independent streams with variable parameters in high-speed telecommunication networks, on the asymptotic behavior of the maximum queue length. Asymptotics of a single-server queue fed by a fractional Brownian motion process has been considered by different authors. Among them, we mention Zeevi and Glynn, which in Zeevi and Glynn [15] considered the behavior of the maximum queue length over the interval $[0, t]$ as $t \to +\infty$. More specifically, they showed that under heavy traffic, this maximum grows like t^H, $H \in (1/2, 1)$ being the Hurst parameter of the driftless fractional Brownian process that feeds the queue, whereas if the queue is stable, the maximum grows like $(\log t)^{\frac{1}{2(1-H)}}$. In Delgado [5], a generalization of the result under heavy traffic to the multidimensional and non-zero drift setting was considered. For that, it was necessary to overcome certain difficulties arising from the lack of an explicit expression of the pushing process associated to the multidimensional fractional Brownian motion process.

Moreover, in Duncan and Jin [6], a stable fluid single-server queue with an input that is the aggregation of independent driftless fractional Brownian motions, which is a generalization of the model introduced in Zeevi and Glynn [15], is considered, and its maximum queue length over $[0, t]$, $M(t)$, is proved to grow like $(\log t)^{\frac{1}{2(1-H^+)}}$, where H^+ is the largest Hurst parameter of the aggregated fractional Brownian motions. Stimulus for the introduction of this model, as the authors explain in the Introduction of their article, is that firstly, from a practical point of view, a fractional

Brownian queueing model is an approximation of Internet traffic and can produce meaningful results for queueing performance. For example, estimations of overflow probabilities $P(M(t) > b)$ are important in practice for the admission control in network systems. On the other hand, it has been observed that the Hurst parameter estimated in network does not remain constant in practice, which has suggested the aggregation of independent fractional Brownian motions with (possibly) different Hurst parameters.

Section 5 is devoted to the study of the asymptotic behavior of the maximum queue length up to time t, as $t \to +\infty$ and under heavy traffic conditions, for a queue fed by many heterogeneous heavy-tailed On/Off sources. Results of this section, which are an application of the heavy-traffic limit result given in Theorem 1, generalize those of Delgado [5] to the case of a multi-dimensional Hurst parameter, and at the same time, they extend the results of Duncan and Jin [6] to a multi-station network with not necessarily zero drift in a heavy-traffic environment. In particular, we highlight the fact that the asymptotic behavior of the maximum fluid in queue in the heavy traffic regime obtained in Corollary 2 depends not only on the Hurst parameter but also on the drift vector θ as well as on the mean service rate at each station. In particular, we show that in the driftless case $\theta = 0$, as far as the fluctuations are concerned, the station with the highest Hurst component (heaviest tails) ultimately dominates upper bound as $t \to +\infty$, while it is the station with the lowest Hurst component (lightest tails) which dominates lower bound. If $\theta \neq 0$, we observe a different behavior on the fluctuations.

The organization of the rest of the paper is as follows: definitions, notations, and terminology are introduced in Sect. 2, while Sect. 3 is devoted to the introduction of the fluid model with which we deal, including the statement of the heavy traffic assumption and a technical lemma. In Sect. 4, we state and prove our heavy-traffic limit (Theorem 1) as well as the FWLLN in Corollary 1. The last section is a technical Appendix.

2 Notations and preliminaries

We will denote the identity matrix by I (regardless of its dimension). Vectors will be column vectors and v^T means the transpose of a vector (or a matrix) v. By $\mathrm{diag}(v)$ 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 $d \geq 1$, the d-dimensional positive orthant is $\mathbb{R}^d_+ = \{v = (v_1, \ldots, v_d)^T \in \mathbb{R}^d : v_i \geq 0 \ \forall i = 1, \ldots, d \ \text{(i.e. } v \geq 0)\}$. For any $d \times d'$ matrix $A = (a_{ij})_{i=1,\ldots,d, j=1,\ldots,d'}$, let $|A| \stackrel{\text{def}}{=} \max_{1 \leq j \leq d'} (\sum_{i=1}^{d} |a_{ij}|)$ (where $|x|$ denotes the absolute value of $x \in \mathbb{R}$). In particular, for any $v \in \mathbb{R}^d$, $|v| \stackrel{\text{def}}{=} \sum_{i=1}^{d} |v_i|$. For a non-negative real number x, $[x]$ denotes the maximum integer less or equal to x (the integer part of x).

Let $\mathscr{D}^d$ be the space of all right-continuous $\mathbb{R}^d$-valued functions with left limits defined on $\mathbb{R}_+ = [0, +\infty)$. Let $\mathscr{C}^d$ denote the subspace of continuous functions with the topology of the uniform convergence on compact time intervals. In $\mathscr{D}^d$, we consider the standard Skorokhod metric J_1, which relativized to $\mathscr{C}^d$ coincides with the

topology of the uniform convergence over compacts. For each $t \geq 0$ and $f \in \mathscr{C}^d$, we define the *norm* and the *oscillation* of f on $[0,t]$ by

$$\|f(\cdot)\|_t \stackrel{\text{def}}{=} \max_{0\leq s\leq t}\left(\sum_{\ell=1}^{d}|f_\ell(s)|\right) \quad \text{and}$$

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

respectively. Note that $\operatorname{Osc}(f(\cdot),[0,t]) \leq 2\|f(\cdot)\|_t$ and also that if $f(0)=0$ and $f(s)\in\mathbb{R}^d_+$ for all $s\geq 0$, then $\operatorname{Osc}(f(\cdot),[0,t]) \geq \|f(\cdot)\|_t$.

We will use the following notations for different types of convergence: D-lim denotes the *convergence in distribution* (on $\mathscr{D}^d$ or $\mathscr{C}^d$), while P-lim denotes the *convergence in probability (uniformly on compacts)*, which has the following meaning: we say that a family $\{X^r\}_r$ of random elements on $\mathscr{C}^d$ converges in probability to the random element X if for any $T>0$ and for any $\varepsilon>0$,

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

If X is a deterministic element of $\mathscr{C}^d$ the convergence in probability is equivalent to the convergence in distribution.

Let $\stackrel{\text{fdd}}{=}$ denote the equality of the finite-dimensional distributions between stochastic processes. Φ stands for the standard Gaussian distribution function, that is,

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}}\,dy \quad \text{for any } x\in\mathbb{R}.$$

Although multi-dimensional fractional Brownian motion has appeared previously in the literature (see Biagini et al. [2]), we give here its definition for the sake of completeness and in order to fix notation. As seen in the definition, what characterizes this process and distinguishes it from that introduced in Definition 1 of Delgado [4] is that each component process may have a different Hurst parameter.

Definition 1 (The multi-dimensional fractional Brownian motion) A vector valued stochastic process $B^H=\{B^H(t)=(B_1^H(t),\ldots,B_d^H(t))^T,\ t\geq 0\}$, defined on some probability space, is called a *d-dimensional fractional Brownian motion* of parameter $H=(H_1,\ldots,H_d)^T\in(0,1)^d$, starting from $x\in\mathbb{R}^d$ and with *drift vector* $\theta\in\mathbb{R}^d$, if it is a continuous Gaussian process with $E(B^H(t))=x+\theta t$ for any $t\geq 0$, and with covariance function given by: for any $s,t\geq 0$,

$$\begin{aligned}\operatorname{Cov}\big(B^H(t),B^H(s)\big) &= E\Big(\big(B^H(t)-(x+\theta t)\big)\big(B^H(s)-(x+\theta s)\big)^T\Big)\\ &= \operatorname{diag}\left(\frac{\sigma_1^2}{2}\big(t^{2H_1}+s^{2H_1}-|t-s|^{2H_1}\big),\ldots,\frac{\sigma_d^2}{2}\big(t^{2H_d}+s^{2H_d}-|t-s|^{2H_d}\big)\right),\end{aligned}$$

where $\sigma_i^2=E(B_i^H(1)-E(B_i^H(1)))^2>0$ for any $i=1,\ldots,d$. H is called the (multi-dimensional) *Hurst parameter* of the process.

Note that by definition, the component processes $B_1^H, \ldots, B_d^H$ are independent one-dimensional fractional Brownian motions of respective Hurst parameters $H_1, \ldots, H_d$. For short, we will say that B^H is a d-dimensional fBm *with associated data* (x, θ, H, σ^2), where $\sigma^2 = (\sigma_1^2, \ldots, \sigma_d^2)^T > 0$.

Remark 1 If B^H is a d-dimensional fBm *with associated data* $(x = 0, \theta = 0, H, \sigma^2)$, then B^H is *vector self-similar* in the sense of Lavancier, Philippe, and Surgailis [9], that is, for any $\lambda > 0$,

$$\big(B_1^H(\lambda t), \ldots, B_d^H(\lambda t)\big) \stackrel{\text{fdd}}{=} \big(\lambda^{H_1} B_1^H(t), \ldots, \lambda^{H_d} B_d^H(t)\big).$$

Analogously to what is done in Definition 2 of Delgado [4], we now introduce a process which behaves like a multi-dimensional fractional Brownian motion in the interior of $\mathbb{R}_+^d$ and is confined to this orthant by instantaneous "reflection" at the boundary.

Definition 2 (The multi-dimensional reflected fractional Brownian motion) A *d-dim. reflected fractional Brownian motion* on the positive orthant $\mathbb{R}_+^d$ with associated data

$$\big(x, \theta, H, \sigma^2, R\big),$$

where $x \in \mathbb{R}_+^d$, $\theta \in \mathbb{R}^d$, $H = (H_1, \ldots, H_d)^T \in (0, 1)^d$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_d^2) > 0$ and R is a d-dimensional *completely-$\mathcal{S}$* matrix, is a d-dimensional process $W = \{W(t), t \geq 0\}$ defined on some probability space such that

(i) W has continuous paths and $W(t) \in \mathbb{R}_+^d$ for all $t \geq 0$, a.s.,
(ii) $W = X + RY$ a.s., with X and Y two d-dimensional processes defined on the same probability space and verifying:
(iii) X is a d-dimensional fBm with associated data (x, θ, H, σ^2),
(iv) Y has continuous and non-decreasing paths, and for each $j = 1, \ldots, d$, 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 $F_j = \{y \in \mathbb{R}_+^d : y_j = 0\}$).

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 multi-dimensional Skorokhod problem* associated with X. Note that by definition, $W(0) = X(0) = x$. The *driftless* case corresponds to $\theta = 0$. Process Y is called the *pushing* process, and matrix R the *reflection* matrix. If the triplet W, X, and Y verifies (i), (ii), and (iv) of Definition 2, we say that $W = X + RY$ is a *Skorokhod decomposition*.

Remark 2 For each j, the direction of the reflection on face F_j is given by the jth column of the *reflection matrix* R. The *completely-$\mathcal{S}$* property of matrix R is sufficient for the existence of the *R-regularization* of X, as can be seen in Theorem 2 of Bernard and El Kharroubi [1]. But as is pointed out in the remark following that result, this property cannot ensure the adaptedness of process Y to any filtration to which X is adapted. Nevertheless, Proposition 4.2 of Williams [14] shows that under a stronger assumption on R, henceforth denoted by (**HR**), this issue is solved, where this assumption is

(**HR**) R can be expressed as $I + \Theta$, with Θ a $d \times d$ matrix such that $\langle \Theta \rangle$ has spectral radius strictly less than 1

(given a matrix A, $\langle A \rangle$ stands for the matrix obtained from A by replacing all its entries by their absolute values). (**HR**) is a sufficient condition for strong pathwise uniqueness, as can be seen in the proof of Proposition 4.2 [14, p. 23], where this assumption is denoted *condition* (II). Interested readers may alternatively consult Sect. 2 of Delgado [4] for more details.

3 The fluid model

We consider a network composed of d stations, where each station consists of a single server that processes continuous fluid, and an infinite buffer. We follow the model introduced in Delgado [4] but the main difference here is that the On/Off sources of different stations are allowed to be non-homogeneous. More specifically, we suppose first 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 constant rate), and $U_j(t) = 0$ means that it is Off. We suppose that the lengths of the On-periods are independent, those of the Off-periods are independent, and the lengths of On- and Off-periods are independent of each other.

Let f_j^{on} and f_j^{off} be the probability density functions corresponding to the lengths of On- and Off-periods for the source feeding station j, respectively, which are non-negative and heavy-tailed. Therefore, their (positive) expected values are

$$\tilde{\mu}_j^{\mathrm{on}} = \int_0^{+\infty} u f_j^{\mathrm{on}}(u)\, du \quad \text{and} \quad \tilde{\mu}_j^{\mathrm{off}} = \int_0^{+\infty} u f_j^{\mathrm{off}}(u)\, du.$$

Assume that as $x \to +\infty$,

$$\int_x^{+\infty} f_j^{\mathrm{on}}(u)\, du \sim x^{-\beta_j^{\mathrm{on}}} L_j^{\mathrm{on}}(x) \quad \text{and} \quad \int_x^{+\infty} f_j^{\mathrm{off}}(u)\, du \sim x^{-\beta_j^{\mathrm{off}}} L_j^{\mathrm{off}}(x), \qquad (1)$$

where $1 < \beta_j^{\mathrm{on}}, \beta_j^{\mathrm{off}} < 2$ and L_j^{on}, L_j^{off} are positive slowly varying functions at infinity such that if $\beta_j^{\mathrm{on}} = \beta_j^{\mathrm{off}}$, then $\lim_{x \to +\infty} \frac{L_j^{\mathrm{on}}(x)}{L_j^{\mathrm{off}}(x)}$ exists and belongs to $(0, +\infty)$. Note that $\tilde{\mu}_j^{\mathrm{on}}$ and $\tilde{\mu}_j^{\mathrm{off}}$ are finite while variances are not.

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, and that they are all independent. We define the cumulative external fluid arrived up to time t (by the N sources) at station j by

$$E_j^N(t) \stackrel{\mathrm{def}}{=} \alpha_j^N \int_0^t \frac{1}{N} \left(\sum_{n=1}^{N} U_j^{(n)}(u) \right) du,$$

where $\alpha_j^N > 0$ is the (possibly dependent on N) deterministic rate at which fluid would arrive at station j if all sources were On, or *external arrival rate*. The d component processes of the (non-deterministic) *cumulative external fluid arrival* process $E^N = \{E^N(t) = (E_1^N(t), \ldots, E_d^N(t))^T \ t \geq 0\}$, are all independent. We assume $E^N(0) = 0$. Let $\alpha^N = (\alpha_1^N, \ldots, \alpha_d^N)^T$.

We suppose that fluid at each server is processed in the arrival order (FIFO service discipline). 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 the service discipline is a *non-idling* (or *work-conserving*) policy that means that a server is never idle when there is fluid waiting to be processed at its station.

Let $P_{j\ell}$ be the proportion of fluid that leaving station j goes next to station ℓ. We assume that for each j, $\sum_{\ell=1}^d P_{j\ell} \leq 1$ and $1 - \sum_{\ell=1}^d P_{j\ell} \geq 0$ is the proportion of fluid that leaving station j goes outside the network. Thus, $P = (P_{j\ell})_{j,\ell=1}^d$ is a sub-stochastic matrix. It is called the "*flow*" *matrix of the fluid network*, and it is assumed to have spectral radius strictly less than one. Hence, $Q \stackrel{\text{def}}{=} (I - P^T)^{-1}$ is well defined.

Hereafter, we will denote by $\tilde{\alpha}^N$ the vector

$$\begin{aligned}\tilde{\alpha}^N &\stackrel{\text{def}}{=} \left(\alpha_1^N \frac{\tilde{\mu}_1^{\text{on}}}{\tilde{\mu}_1^{\text{on}} + \tilde{\mu}_1^{\text{off}}}, \ldots, \alpha_d^N \frac{\tilde{\mu}_d^{\text{on}}}{\tilde{\mu}_d^{\text{on}} + \tilde{\mu}_d^{\text{off}}}\right)^T \\ &= \text{diag}\left(\frac{\tilde{\mu}_1^{\text{on}}}{\tilde{\mu}_1^{\text{on}} + \tilde{\mu}_1^{\text{off}}}, \ldots, \frac{\tilde{\mu}_d^{\text{on}}}{\tilde{\mu}_d^{\text{on}} + \tilde{\mu}_d^{\text{off}}}\right)\alpha^N,\end{aligned}$$

and we can define λ^N to be the unique d-dimensional vector solution to the *traffic equation*

$$\lambda^N \stackrel{\text{def}}{=} \tilde{\alpha}^N + P^T\lambda^N \quad \left(\text{that is, } \lambda^N = Q\tilde{\alpha}^N\right).$$

We point out that λ_j^N can be interpreted as the long run fluid rate into and out of station j. Indeed, Corollary 1 in Sect. 4 gives support to this interpretation.

Two descriptive (d-dimensional) processes will be used to measure the performance of the fluid model: the immediate *workload process* W^N and the *cumulative idle-time process* Y^N. Let $W_j^N(t)$ denote the amount of time required for server j to complete processing of all the fluid in queue (or being processed) at station j at time t, and $Y_j^N(t)$ denote the cumulative amount of time that server j has been idle in the time interval $[0, t]$, that is,

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

Immediate workload process measures the congestion and delay in the network, while *idle-time* process measures utilization of resources. We assume $W^N(0) = 0$.

In addition, other processes such as A^N and D^N also deserve consideration. $A_j^N(t)$ is the total fluid arriving to station j up to time t, including both feedback flow from other stations and external input, and $D_j^N(t)$ is the total amount of fluid departing station j (both being routed to other station or leaving the network), up to time t. We

assume $A^N(0) = D^N(0) = 0$. For these processes, we will obtain in Corollary 1 a functional weak law of large numbers.

For any $r > 0$ real valued parameter, we can consider a sequence of fluid models indexed by (r, N), where N is the number of On/Off sources feeding each station. We will use r as a scalar parameter in time. For the (r, N) fluid model, suppose that server at station j processes fluid at a constant rate $\mu_j^{r,N}$ if station j were never idle (that is, $m_j^{r,N} = 1/\mu_j^{r,N}$ is the *mean service time* for station j). Let $m^{r,N} = (m_1^{r,N}, \ldots, m_d^{r,N})^T$ and $M^{r,N} = \mathrm{diag}(m^{r,N})$. We assume that $\lim_{N\to+\infty} M^{r,N}$ exists and does not depend on r; we denote it by M.

We also introduce the *fluid traffic intensity* for station j by

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

In order to define the *scaled processes* associated with the (r, N) fluid model, we have to introduce some notation by following Taqqu, Willinger, and Sherman [12] (see also Delgado [4]). For any $j = 1, \ldots, d$ set $a_j^{\text{on}} = \frac{\Gamma(2-\beta_j^{\text{on}})}{(\beta_j^{\text{on}}-1)}$ and $a_j^{\text{off}} = \frac{\Gamma(2-\beta_j^{\text{off}})}{(\beta_j^{\text{off}}-1)}$, where β_j^{on} and β_j^{off} are defined by (1). The normalization factors used below depend on b_j, defined by $b_j \stackrel{\text{def}}{=} \lim_{t\to+\infty} t^{\beta_j^{\text{off}}-\beta_j^{\text{on}}} \frac{L_j^{\text{on}}(t)}{L_j^{\text{off}}(t)}$, which exists although it could be infinite. If $0 < b_j < +\infty$ (implying $\beta_j^{\text{on}} = \beta_j^{\text{off}}$ and $b_j = \lim_{t\to+\infty} \frac{L_j^{\text{on}}(t)}{L_j^{\text{off}}(t)}$), set $\beta_j = \beta_j^{\text{on}} = \beta_j^{\text{off}}$, $L_j = L_j^{\text{off}}$ and

$$\sigma_j^{2,\lim} \stackrel{\text{def}}{=} \frac{2((\tilde{\mu}_j^{\text{off}})^2 a_j^{\text{on}} b_j + (\tilde{\mu}_j^{\text{on}})^2 a_j^{\text{off}})}{(\tilde{\mu}_j^{\text{on}} + \tilde{\mu}_j^{\text{off}})^3 \Gamma(4-\beta_j)}.$$

If, on the other hand, $b_j = +\infty$ ($\beta_j^{\text{off}} > \beta_j^{\text{on}}$), set $L_j = L_j^{\text{on}}$, $\beta_j = \beta_j^{\text{on}}$, and

$$\sigma_j^{2,\lim} \stackrel{\text{def}}{=} \frac{2(\tilde{\mu}_j^{\text{off}})^2 a_j^{\text{on}}}{(\tilde{\mu}_j^{\text{on}} + \tilde{\mu}_j^{\text{off}})^3 \Gamma(4-\beta_j)}.$$

If $b_j = 0$ ($\beta_j^{\text{off}} < \beta_j^{\text{on}}$), set $L_j = L_j^{\text{off}}$, $\beta_j = \beta_j^{\text{off}}$, and

$$\sigma_j^{2,\lim} \stackrel{\text{def}}{=} \frac{2(\tilde{\mu}_j^{\text{on}})^2 a_j^{\text{off}}}{(\tilde{\mu}_j^{\text{on}} + \tilde{\mu}_j^{\text{off}})^3 \Gamma(4-\beta_j)}.$$

In either case, $\beta_j \in (1, 2)$ for any j. Let we define $H_j \stackrel{\text{def}}{=} \frac{3-\beta_j}{2}$. Therefore, $H_j \in (\frac{1}{2}, 1)$. Let $H \stackrel{\text{def}}{=} (H_1, \ldots, H_d)^T$.

Now we can introduce the *scaled processes* associated with the (r, N) fluid model. We will use a hat to denote them. For any $j = 1, \ldots, d$,

$$\hat{W}_j^{r,N}(t) \stackrel{\text{def}}{=} \sqrt{N}\frac{W_j^N(rt)}{r^{H_j} L_j^{1/2}(r)}$$

$$\hat{E}_j^{r,N}(t) \stackrel{\text{def}}{=} \sqrt{N}\frac{E_j^N(rt) - \tilde{\alpha}_j^N rt}{r^{H_j} L_j^{1/2}(r)}$$

$$\hat{Y}_j^{r,N}(t) \stackrel{\text{def}}{=} \sqrt{N}\frac{Y_j^N(rt)}{r^{H_j} L_j^{1/2}(r)}$$

or, in matricial form,

$$\hat{W}^{r,N}(t) = \left(L^{r,H}\right)^{-1}\sqrt{N}W^N(rt)$$
$$\hat{E}^{r,N}(t) = \left(L^{r,H}\right)^{-1}\sqrt{N}\left(E^N(rt) - \tilde{\alpha}^N rt\right)$$
$$\hat{Y}^{r,N}(t) = \left(L^{r,H}\right)^{-1}\sqrt{N}Y^N(rt),$$

where

$$L^{r,H} \stackrel{\text{def}}{=} \operatorname{diag}\left(r^{H_1}L_1^{1/2}(r), \ldots, r^{H_d}L_d^{1/2}(r)\right). \tag{2}$$

The next lemma generalizes the Skorokhod decomposition given by formula (17) of Delgado [4] to our setting, and will be used in the proof of Theorem 1 below.

Lemma 1 *The scaled processes are related by means of*

$$\hat{W}^{r,N}(t) = \hat{X}^{r,N}(t) + \hat{R}^{r,N}\hat{Y}^{r,N}(t), \tag{3}$$

with

$$\hat{X}^{r,N}(t) \stackrel{\text{def}}{=} M^{r,N}\hat{E}^{r,N}(t) + \left(L^{r,H}\right)^{-1}R^{r,N}\sqrt{N}\left(\rho^{r,N} - e\right)rt, \tag{4}$$

$R^{r,N}$ *being a square matrix defined by*

$$R^{r,N} \stackrel{\text{def}}{=} M^{r,N}Q^{-1}\left(M^{r,N}\right)^{-1} = I - M^{r,N}P^T\left(M^{r,N}\right)^{-1},$$

and

$$\hat{R}^{r,N} \stackrel{\text{def}}{=} \hat{\Lambda}^{r,N}Q^{-1}\left(\hat{\Lambda}^{r,N}\right)^{-1} = I - \hat{\Lambda}^{r,N}P^T\left(\hat{\Lambda}^{r,N}\right)^{-1}$$
$$\text{with } \hat{\Lambda}^{r,N} = \left(L^{r,H}\right)^{-1}M^{r,N}. \tag{5}$$

Proof Analogously to Lemma 1 of Delgado [4] we can obtain that for the (r, N) fluid model,

$$W^N(rt) = M^{r,N}E^N(rt) - R^{r,N}ert + R^{r,N}Y^N(rt).$$

Indeed, by using the notations of this paper, relation (8) in Lemma 1 [4] can be written as $W^N(t) = R^{r,N}M^{r,N}QE^N(t) - R^{r,N}et + R^{r,N}Y^N(t)$, and taking into account that $R^{r,N}M^{r,N}Q = (M^{r,N}Q^{-1}(M^{r,N})^{-1})M^{r,N}Q = M^{r,N}$, we obtain the desired expression simply by replacing t by rt.

Therefore, we can write for each $j = 1, \ldots, d$,

$$\hat{W}_j^{r,N}(t) = m_j^{r,N}\hat{E}_j^{r,N}(t) + \frac{\sqrt{N}}{r^{H_j}L_j^{1/2}(r)}\left(R^{r,N}\left(\rho^{r,N} - e\right)\right)_j rt$$

$$+ \frac{\sqrt{N}}{r^{H_j}L_j^{1/2}(r)}\left(R^{r,N}Y^N(rt)\right)_j$$

because $M^{r,N}\tilde{\alpha}^N = R^{r,N}\rho^{r,N}$. Then expression (3) is proved, with $\hat{X}^{r,N}(t)$ given by (4) and

$$\hat{R}^{r,N} = \left(L^{r,H}\right)^{-1}R^{r,N}L^{r,H} = \hat{\Lambda}^{r,N}Q^{-1}\left(\hat{\Lambda}^{r,N}\right)^{-1} = I - \hat{\Lambda}^{r,N}P^T\left(\hat{\Lambda}^{r,N}\right)^{-1}. \quad \square$$

Remark 3 Taking the limit as $N \to +\infty$ in (5), we can introduce

$$\hat{\hat{R}}^r \stackrel{\text{def}}{=} \lim_{N\to+\infty}\hat{R}^{r,N} = \hat{\hat{\Lambda}}^r Q^{-1}\left(\hat{\hat{\Lambda}}^r\right)^{-1} = I - \hat{\hat{\Lambda}}^r P^T\left(\hat{\hat{\Lambda}}^r\right)^{-1} \quad \text{with } \hat{\hat{\Lambda}}^r \stackrel{\text{def}}{=} \left(L^{r,H}\right)^{-1}M,$$

and matrix $\hat{\hat{R}}^r$ verifies (**HR**) since $\hat{\hat{\Lambda}}^r P^T(\hat{\hat{\Lambda}}^r)^{-1}$ has the same spectral radius as P, which is assumed to be strictly less than 1.

Remark 4 Note that processes appearing in expression (3) verify: $\hat{W}^{r,N}$ has continuous paths; for any $t \geq 0$, a.s. $\hat{W}^{r,N}(t) \in \mathbb{R}_+^d$; $\hat{Y}^{r,N}$ has continuous and non-decreasing paths, and for each j, a.s. $\hat{Y}_j^{r,N}(0) = 0$ and

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

This shows that (3) turns out to be a Skorokhod decomposition.

4 The heavy-traffic limit

Our goal now is to prove that the scaled workload process $\hat{W}^{r,N}$ converges to a d-dimensional reflected fractional Brownian motion process in distribution, when N first and then r, tend to infinity in this order, under heavy traffic. This result generalizes Theorem 1 of Delgado [4]. *Heavy traffic condition* establishes that the total load imposed on each service station tends to the value of its capacity, that is, its traffic intensity tends to be equal to 1, in the following sense:

$$\textbf{(HT)} \qquad \lim_{N\to+\infty}\sqrt{N}\left(\rho^{r,N} - e\right) = -\hat{\gamma}^r,$$

where $e = (1, \ldots, 1)^T \in \mathbb{R}^d$, for some $\hat{\gamma}^r = (\hat{\gamma}_1^r, \ldots, \hat{\gamma}_d^r)^T \geq 0$ such that it converges to zero as $r \to +\infty$ in the sense that a vector $\gamma = (\gamma_1, \ldots, \gamma_d)^T \geq 0$ exists such that

$$\lim_{r\to+\infty}\frac{r^{1-H_j}}{L_j^{1/2}(r)}\hat{\gamma}_j^r = \gamma_j \quad \text{for any } j = 1, \ldots, d. \tag{6}$$

Note that from (**HT**) and the fact that $\lambda^N = (M^{r,N})^{-1}\rho^{r,N}$ we deduce the existence of $\lambda \stackrel{\text{def}}{=} \lim_{N\to+\infty}\lambda^N$ and also that $M = \operatorname{diag}(\lambda_1,\ldots,\lambda_d)^{-1}$ (or $\lambda = M^{-1}e$). Moreover, since $\lambda^N = Q\tilde{\alpha}^N$ and $\lim_{N\to+\infty}\tilde{\alpha}^N = Q^{-1}M^{-1}e$, the limit external arrival rate needed to achieve the maximum capacity of the system is

$$\lim_{N\to+\infty}\alpha^N = \alpha = \operatorname{diag}\left(\frac{\tilde{\mu}_1^{\text{on}}}{\tilde{\mu}_1^{\text{on}}+\tilde{\mu}_1^{\text{off}}},\ldots,\frac{\tilde{\mu}_d^{\text{on}}}{\tilde{\mu}_d^{\text{on}}+\tilde{\mu}_d^{\text{off}}}\right)^{-1} Q^{-1}M^{-1}e(>0).$$

Remark 5 Heavy traffic condition (**HT**) generalizes that introduced in Delgado [4] in the sense that $\hat{\gamma}^r$ was taken there to be identically zero. Motivation for this generalization is what is named "*thin control*" in the literature (see Lee and Weerasinghe [11]), which typically consists of processing rates of the form

$$\mu_j^{r,N} = \lambda_j^N\left(1+\frac{1}{\sqrt{N}}\hat{\gamma}_j^r\right)$$

with $\hat{\gamma}_j^r$ satisfying (6). If this is the case, (**HT**) necessarily holds.

Before stating our result, we must introduce two more assumptions, namely

$$(\textbf{HL}) \quad \text{if } H_i = H_j, \text{ then there exists } \ell_{ij} = \lim_{t\to+\infty}\frac{L_j(t)}{L_i(t)} \in (0,+\infty)$$

and

$$(\textbf{HP}) \quad P_{ij} = 0 \quad \text{if } H_j < H_i, i,j = 1,\ldots,d.$$

(Assumption (**HL**) is technical while (**HP**), which is vacuous if $H_i = H_j$ for all $i, j = 1,\ldots,d$, forces the system to be feed-forward in the sense that fluid cannot flow from one station to other with lighter tail distributions.)

Theorem 1 (The heavy-traffic limit) *Under heavy traffic condition* (**HT**) *and assumptions* (**HL**) *and* (**HP**), *the following limits exist*:

$$\hat{\hat{W}}^r = \underset{N\to+\infty}{\text{D-lim}}\ \hat{W}^{r,N} \quad \left(\text{in } \mathscr{D}^d\right) \quad \text{and} \quad W = \underset{r\to+\infty}{\text{D-lim}}\ \hat{\hat{W}}^r \quad \left(\text{in } \mathscr{C}^d\right),$$

and W is a d-dimensional reflected fractional Brownian motion process on $\mathbb{R}_+^d$ with associated data

$$\left(x = 0, \theta = -R\gamma, H = (H_1,\ldots,H_d)^T, \sigma^2, R\right),$$

where $H \in (\frac{1}{2},1)^d$, $\gamma \geq 0$, $\sigma^2 = M^2\operatorname{diag}(\alpha)^2\sigma^{2,\lim}$ and $R = I - \tilde{P}$, $\tilde{P}$ being the d-dimensional matrix defined by

$$\tilde{P}_{ij} = \begin{cases} P_{ji}\frac{m_i}{m_j}\ell_{ij}^{1/2} & \text{if } H_i = H_j, \\ 0 & \text{otherwise.} \end{cases}$$

Proof The proof follows the same ideas as that of Theorem 1 of Delgado [4], uses the *invariance principle* of Williams [14] as a basic ingredient, and relies heavily on Theorem 7.2.5 of Whitt [13], from which the weak convergence of process $\hat{E}^{r,N}$, first as $N \to +\infty$, and then as $r \to +\infty$, is obtained (while Theorem 1 [12] only provides the corresponding convergence of the finite dimensional distributions in the first case).

First, by Lemma 1 $(\hat{W}^{r,N}, \hat{Y}^{r,N})$ is the solution of the Skorokhod problem associated with $\hat{X}^{r,N}$ with reflection matrix $\hat{R}^{r,N}$. In order to use the invariance principle of Williams (actually, Corollary 4.3 [14], as stated here in the Appendix for the convenience of the reader) for processes $\hat{W}^{r,N}$, $\hat{X}^{r,N}$ and $\hat{Y}^{r,N}$, we will use the fact that $\lim_{N\to+\infty} \hat{R}^{r,N} = \hat{\hat{R}}^r$ satisfies assumption (**HR**) as discussed in Remark 3, and also the weak convergence of $\hat{X}^{r,N}$ as $N \to +\infty$, which is a consequence of Theorem 1 [12] and Theorem 7.2.5 [13]. Indeed, for any $j = 1, \ldots, d$,

$$\begin{aligned}\hat{E}_j^{r,N}(t) &= \sqrt{N}\frac{E_j^N(rt) - \tilde{\alpha}_j^N rt}{r^{H_j} L_j^{1/2}(r)} \\ &= \frac{\alpha_j^N}{r^{H_j} L_j^{1/2}(r)} \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \left(\int_0^{rt} U_j^{(n)}(u)\,du - \frac{\tilde{\mu}_j^{\text{on}}}{\tilde{\mu}_j^{\text{on}} + \tilde{\mu}_j^{\text{off}}} rt \right)\end{aligned}$$

and using both results we have that in $\mathscr{D}^d$ there exists the limit

$$\hat{\hat{E}}^r = \underset{N\to+\infty}{\text{D-lim}}\, \hat{E}^{r,N}, \tag{7}$$

which has paths in $\mathscr{C}^d$, and in $\mathscr{C}^d$

$$\underset{r\to+\infty}{\text{D-lim}}\, \hat{\hat{E}}^r = B^H, \tag{8}$$

B^H being a d-dimensional fractional Brownian motion with associated data $(x = 0, \theta = 0, H, \text{diag}(\alpha)^2\sigma^{2,\text{lim}})$.

Now we apply Lemma 1 and combining (4), heavy traffic condition (**HT**), and the *continuous mapping theorem*, we obtain that there exists $\hat{\hat{X}}^r = \text{D-lim}_{N\to+\infty} \hat{X}^{r,N}$, with

$$\hat{\hat{X}}^r(t) = M\hat{\hat{E}}^r(t) + \hat{\hat{R}}^r \left(L^{r,H}\right)^{-1}\left(-\hat{\gamma}^r\right)rt, \tag{9}$$

that implies that paths of $\hat{\hat{X}}^r$ are continuous. Indeed, (9) is justified because

$$\left(L^{r,H}\right)^{-1} R^{r,N}\sqrt{N}\left(\rho^{r,N} - e\right) = \left(L^{r,H}\right)^{-1} M^{r,N} Q^{-1}\left(M^{r,N}\right)^{-1}\sqrt{N}\left(\rho^{r,N} - e\right),$$

which converges as $N \to +\infty$ to

$$\left(L^{r,H}\right)^{-1} M Q^{-1} M^{-1}\left(-\hat{\gamma}^r\right) = \hat{\hat{\Lambda}}^r Q^{-1}\left(\hat{\hat{\Lambda}}^r\right)^{-1}\left(L^{r,H}\right)^{-1}\left(-\hat{\gamma}^r\right) = \hat{\hat{R}}^r\left(L^{r,H}\right)^{-1}\left(-\hat{\gamma}^r\right).$$

Then, by Corollary 4.3 [14] (see the Appendix), we have that in $\mathscr{D}^d$ there exists

$$\mathsf{D}\text{-}\lim_{N\to+\infty}\left(\hat{W}^{r,N},\hat{X}^{r,N},\hat{Y}^{r,N}\right)=\left(\hat{\hat{W}}^{r},\hat{\hat{X}}^{r},\hat{\hat{Y}}^{r}\right)$$

and the limit satisfies the conditions (i), (ii), and (iv) of Definition 2, that is,

$$\hat{\hat{W}}^{r}=\hat{\hat{X}}^{r}+\hat{\hat{R}}^{r}\hat{\hat{Y}}^{r}$$

is a Skorokhod decomposition.

To reapply this result now to processes $\hat{\hat{W}}^r$, $\hat{\hat{X}}^r$, and $\hat{\hat{Y}}^r$, we must check the weak convergence, as $r\to+\infty$, of $\hat{\hat{X}}^r$. Indeed, this is a consequence of (9), (8), (6) and the *continuous mapping theorem*, that imply the existence of $\mathsf{D}\text{-}\lim_{r\to+\infty}\hat{\hat{X}}^r=X$, with $X=MB^H-R\gamma t$, which is a d-dimensional fractional Brownian motion process with associated data $(x=0,\theta=-R\gamma,H,\sigma^2)$, where

$$\gamma\geq 0,\qquad \sigma^2=M^2\,\mathrm{diag}(\alpha)^2\sigma^{2,\lim}\quad\text{and}\quad R=I-\tilde{P}.$$

Note that $R=\lim_{r\to+\infty}\hat{\hat{R}}^r$, by assumptions (**HL**) and (**HP**), and also that R satisfies (**HR**). We can see this by using that the spectral radius of $\tilde{P}$ is strictly less than one if and only if the limit as $k\to+\infty$ of powers $\tilde{P}^k$ equals zero. But this can be easily checked since

$$\tilde{P}=\lim_{r\to+\infty}\hat{\hat{\Lambda}}^r P^T\left(\hat{\hat{\Lambda}}^r\right)^{-1}\quad\text{and}\quad\left(\hat{\hat{\Lambda}}^r P^T\left(\hat{\hat{\Lambda}}^r\right)^{-1}\right)^k=\hat{\hat{\Lambda}}^r\left(P^T\right)^k\left(\hat{\hat{\Lambda}}^r\right)^{-1},$$

which implies that $\lim_{k\to+\infty}\tilde{P}^k=\lim_{k\to+\infty}\lim_{r\to+\infty}\hat{\hat{\Lambda}}^r(P^T)^k(\hat{\hat{\Lambda}}^r)^{-1}$, and if we denote by $a_{ij}(r,k)$ the elements of matrix $\hat{\hat{\Lambda}}^r(P^T)^k(\hat{\hat{\Lambda}}^r)^{-1}$, we can readily see that

$$a_{ij}(r,k)=\begin{cases}0 & \text{if } H_j>H_i,\\ r^{H_j-H_i}\dfrac{L_j^{1/2}(r)}{L_i^{1/2}(r)}\dfrac{m_i}{m_j}P_{ij}^{(k)} & \text{if } H_j\leq H_i,\end{cases}$$

denoting by $P_{ij}^{(k)}$ the elements of matrix $(P^T)^k$. As a consequence, we have that, for any fixed k, there exists $\lim_{r\to+\infty}\hat{\hat{\Lambda}}^r(P^T)^k(\hat{\hat{\Lambda}}^r)^{-1}$ since

$$\lim_{r\to+\infty}a_{ij}(r,k)=\begin{cases}0 & \text{if } H_j\neq H_i,\\ P_{ii}^{(k)} & \text{if } i=j,\\ \ell_{ij}^{1/2}\dfrac{m_i}{m_j}P_{ij}^{(k)} & \text{if } H_i=H_j, i\neq j\end{cases}$$

and there also exists $\lim_{k\to+\infty}\hat{\hat{\Lambda}}^r(P^T)^k(\hat{\hat{\Lambda}}^r)^{-1}=0$ uniformly in r, because there exists a constant $C>0$ such that for any i,j,r,k, $a_{ij}(r,k)<CP_{ij}^{(k)}$, and $\lim_{k\to+\infty}(P^T)^k=0$ because the spectral radius of P^T (the same as P) is strictly

less than one by hypothesis. Therefore,

$$\lim_{k\to+\infty} \tilde{P}^k = \lim_{r\to+\infty} \lim_{k\to+\infty} \hat{\hat{\Lambda}}^r \big(P^T\big)^k \big(\hat{\hat{\Lambda}}^r\big)^{-1} = 0.$$

Then, by Corollary 4.3 [14] (see the Appendix) again, there exists

$$\mathsf{D}\text{-}\lim_{r\to+\infty} \big(\hat{\hat{W}}^r, \hat{\hat{X}}^r, \hat{\hat{Y}}^r\big) = (W, X, Y),$$

where the triplet (W, X, Y) satisfies conditions (i)–(iv) of the Definition 2. Thus, $W = X + RY$ is a d-dimensional reflected fractional Brownian motion on $\mathbb{R}^d_+$ with associated data $(x = 0, \theta = -R\gamma, H, \sigma^2, R)$. □

Analogous to Theorem 2 of Delgado [4], a functional weak law of large numbers for processes A^N and D^N, which are the total amount of fluid arriving and departing stations up to any time, respectively, can be proved. This result reinforces the interpretation of λ, which is the solution of the limiting traffic equation $\lambda = Q\tilde{\alpha}$, with $\tilde{\alpha} = \operatorname{diag}(\frac{\tilde{\mu}_1^{\text{on}}}{\tilde{\mu}_1^{\text{on}}+\tilde{\mu}_1^{\text{off}}}, \ldots, \frac{\tilde{\mu}_d^{\text{on}}}{\tilde{\mu}_d^{\text{on}}+\tilde{\mu}_d^{\text{off}}})\alpha$, as the long run fluid rate into and out of the system.

Let us first introduce the associated *scaled processes*

$$\hat{A}^{r,N}(t) \stackrel{\text{def}}{=} \sqrt{N}\frac{A^N(rt) - \lambda^N rt}{r} \quad \text{and} \quad \hat{D}^{r,N}(t) \stackrel{\text{def}}{=} \sqrt{N}\frac{D^N(rt) - \lambda^N rt}{r}.$$

Corollary 1 (FWLLN for processes A^N and D^N) *Under heavy traffic condition* (**HT**) *and assumptions* (**HL**) *and* (**HP**), *there exist the limits in* $\mathscr{D}^d$

$$\hat{\hat{A}}^r = \mathsf{D}\text{-}\lim_{N\to+\infty} \hat{A}^{r,N} \quad \textit{and} \quad \hat{\hat{D}}^r = \mathsf{D}\text{-}\lim_{N\to+\infty} \hat{D}^{r,N},$$

and there also exist the limits in $\mathscr{C}^d$

$$\mathsf{D}\text{-}\lim_{r\to+\infty} \hat{\hat{A}}^r = \mathsf{D}\text{-}\lim_{r\to+\infty} \hat{\hat{D}}^r = 0.$$

Proof The proof is similar to that of Theorem 2 of Delgado [4] and, therefore, we emphasize primarily those aspects which are different. In order to justify the existence of $\hat{\hat{A}}^r$ and $\hat{\hat{D}}^r$, we consider that for any $j = 1, \ldots, d$,

$$\hat{A}_j^{r,N}(\cdot) - \hat{D}_j^{r,N}(\cdot) = \frac{\sqrt{N}}{r}\big(A_j^N(r\cdot) - D_j^N(r\cdot)\big) = \frac{r^{H_j} L_j^{1/2}(r)}{r}\big(m_j^{r,N}\big)^{-1}\hat{W}_j^{r,N}(\cdot).$$

Then, by Theorem 1, there exists in $\mathscr{D}^d$,

$$\mathsf{D}\text{-}\lim_{N\to+\infty} \big(\hat{A}^{r,N} - \hat{D}^{r,N}\big) = \frac{1}{r} L^{r,H} M^{-1} \hat{\hat{W}}^r. \tag{10}$$

We can write

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

and, therefore, we can conclude from (10) and Theorem 1 again that there exists

$$\hat{\hat{A}}^r = \underset{N\to+\infty}{\text{D-lim}}\, \hat{A}^{r,N} = \frac{1}{r} Q\big(L^{r,H} \hat{\hat{E}}^r - P^T L^{r,H} M^{-1} \hat{\hat{W}}^r\big), \tag{11}$$

and by combining (10) with (11) we deduce the existence of

$$\hat{\hat{D}}^r = \underset{N\to+\infty}{\text{D-lim}}\, \hat{D}^{r,N} = \frac{1}{r}\big(Q L^{r,H} \hat{\hat{E}}^r - \big(I + Q P^T\big) L^{r,H} M^{-1} \hat{\hat{W}}^r\big).$$

Now we show that the following limits (in $\mathscr{C}^d$) exist and are equal to zero:

$$\underset{r\to+\infty}{\text{D-lim}}\, \hat{\hat{A}}^r = \underset{r\to+\infty}{\text{D-lim}}\, \hat{\hat{D}}^r (= 0),$$

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

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

We only consider the case of $\hat{\hat{A}}^r$ (the same conclusion can be drawn for $\hat{\hat{D}}^r$): We must show that if we fix $T > 0$ and $\varepsilon > 0$, for any $\delta > 0$ there exists r_0 such that if $r \geq r_0$,

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

but this can be done since from (11) we obtain the upper bound

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

(recall the definition of $L^{r,H}$ given by (2)). □

5 Asymptotics for the maximum queue length

First of all, we consider the extension of the results of Delgado [5] to our setting, in which the Hurst parameter can vary with the station, and study the asymptotic behavior as $t \to +\infty$ of the maximum process given by formula

$$M(t) \overset{\text{def}}{=} \max_{0\leq s\leq t} \sum_{j=1}^{d} a_j W_j(s) = \max_{0\leq s\leq t} a^T W(s),$$

with $a = (a_1, \ldots, a_d)^T > 0$, W being any general d-dimensional reflected fractional Brownian motion process $W = X + RY$ on $\mathbb{R}^d_+$ with associated data

$$\big(x = 0, \theta, H = (H_1, \ldots, H_d)^T, \sigma^2, R\big),$$

where $\theta \in \mathbb{R}^d$, $H \in (0,1)^d$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_d^2)^T > 0$ and R is a $d \times d$ *completely-$\mathscr{S}$* matrix.

For the case of homogeneous Hurst parameters (that is, $H_i = H_j = H$ for any $i, j = 1, \ldots, d$), it was proved in Delgado [5] that the increase of $M(t)$ is closer to that of t if $a^T\theta > 0$, in the sense that it is smaller than that of any function growing faster than t (Theorem 3.1 of Delgado [5]), and that if a restriction (named (**HaR**)) on the weights a holds, this result is tight in the sense that the increase of $M(t)$ is bigger than that of any function growing slower than t (see Theorem 3.2 of Delgado [5]). In the driftless case $\theta = 0$ similar results were obtained but with t^H instead of t. The case $a^T\theta < 0$, which includes but is not restricted to, the negative drift case $\theta < 0$, was also considered, obtaining the result that $(\log t)^{\frac{1}{2(1-H)}}$ and any function growing faster than t, respectively, are the asymptotic lower and upper bounds for $M(t)$.

We note that, although in the results of Delgado [5] hypothesis (**HR**) on matrix R is assumed, it is indeed sufficient to have the (weaker) *completely-$\mathscr{S}$* condition. As mentioned above, there is an additional assumption on matrix R to be considered:

$$(\mathbf{HaR}) \quad R^T a \geq 0.$$

Remark 6 We assume that R is a *completely-$\mathscr{S}$* matrix, which is equivalent to saying that R is *strictly semi-monotone*. This last property means that for each principal submatrix $\tilde{R}$ of R, the system

$$\tilde{R}x \leq 0 \quad \text{and} \quad x \geq 0$$

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

In the sequel, we will use notations

$$H^+ \stackrel{\text{def}}{=} \max\{H_1, \ldots, H_d\} \quad \text{and} \quad H^- \stackrel{\text{def}}{=} \min\{H_1, \ldots, H_d\},$$

and also $\mu \stackrel{\text{def}}{=} a^T\theta$. Note that $\theta = 0$ (respectively $< 0, > 0$) implies $\mu = 0$ (respectively $< 0, > 0$), but that the converses are not true.

Theorem 2

(a) (*Asymptotic upper bound for the maximum.*)

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

for any positive real function f such that

$$\begin{cases} \lim_{t \to +\infty} \frac{f(t)}{t^{H^+}} = +\infty & \text{if } \theta = 0, \\ \lim_{t \to +\infty} \frac{f(t)}{t} = +\infty & \text{if } \theta \neq 0. \end{cases}$$

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

(b) (*Asymptotic lower bound for the maximum.*)
Assume that condition (**HaR**) *holds.*

– *If $\mu \geq 0$, then*

$$\operatorname*{P-lim}_{t\to+\infty} \frac{M(t)}{g(t)} = +\infty$$

for any positive real function g such that

$$\begin{cases} \lim_{t\to+\infty} \frac{g(t)}{t^{H^-}} = 0 & \text{if } \mu = 0, \\ \lim_{t\to+\infty} \frac{g(t)}{t} = 0 & \text{if } \mu > 0. \end{cases}$$

– *If $\mu < 0$, then*

$$\lim_{t\to+\infty} P\left(\frac{M(t)}{(\log t)^{\frac{1}{2(1-H^-)}}} \geq C \right) = 1$$

$$\text{for any} \quad 0 < C < \left(\frac{(a^T\sigma)^2}{2(-\mu)^{2H^-}} \right)^{\frac{1}{2(1-H^-)}}.$$

Proof

Part (a) The proof of this part follows that of Theorem 3.1 of Delgado [5] by taking into account that for each $j = 1, \ldots, d$ and $s > 0$, the random variable $X_j(s) \sim N(\theta_j s, s^{2H_j}\sigma_j^2)$, and that

$$\lambda_j(\varepsilon, t) \overset{\text{def}}{=} \frac{\varepsilon}{K_{R,a}d} \frac{f(t)}{t^{H_j}} - |\theta_j| t^{1-H_j} = t^{1-H_j} \left(\frac{\varepsilon}{K_{R,a}d} \frac{f(t)}{t} - |\theta_j| \right),$$

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

Part (b) The proof of this part is similar to that of Theorem 3.2 of Delgado [5]. We will prove that

$$\lim_{t\to+\infty} P\left(\frac{M(t)}{g(t)} \geq C \right) = 1 \tag{12}$$

for any arbitrary $C > 0$ and g verifying the aforementioned assumptions if $\mu \geq 0$, which implies that $\frac{M(t)}{g(t)}$ is unbounded in probability, and for any

$$0 < C < \left(\frac{(a^T\sigma)^2}{2(-\mu)^{2H^-}} \right)^{\frac{1}{2(1-H^-)}}$$

and $g(t) = (\log t)^{\frac{1}{2(1-H^-)}}$ if $\mu < 0$.

Fix $t > 0$; for any $\delta \in (0, t)$, by assumption (**HaR**) we have

$$M(t) \geq \max_{k=1,\ldots,[\frac{t}{\delta}]} V_k(\delta) \tag{13}$$

with $V_k(\delta) \stackrel{\text{def}}{=} a^T (X(k\delta) - X((k-1)\delta))$ (see Step 1 in the proof of Theorem 3.2 of Delgado [5]).
As X is a d-dimensional fractional Brownian motion process with associated data $(0, \theta, H = (H_1, \ldots, H_d)^T, \sigma^2)$, we find that $V_k(\delta) \sim N(\mu\delta, \sigma_\delta^2)$ with

$$\sigma_\delta = \sum_{j=1}^{d} a_j \sigma_j \delta^{H_j}. \tag{14}$$

$V_k(\delta)$ can be normalized by

$$Z_k \stackrel{\text{def}}{=} \frac{V_k(\delta) - \mu\delta}{\sigma_\delta} \sim N(0,1), \quad \text{for } k = 1, \ldots, \left[\frac{t}{\delta}\right], \tag{15}$$

forming a stationary sequence of standardized Gaussian random variables such that

$$\lim_{\ell \to +\infty} \rho_Z(\ell) \log \ell = 0$$

(because $H_j < 1$) with their covariance function

$$\rho_Z(\ell) \stackrel{\text{def}}{=} E(Z_1 Z_{1+\ell}) = \frac{1}{\sigma_\delta^2} \sum_{j=1}^{d} a_j^2 \frac{\sigma_j^2}{2} \delta^{2H_j} \left(2H_j(2H_j - 1)\ell^{2H_j - 2} + O\left(\ell^{2H_j - 3}\right)\right).$$

Following Theorem 4.3.3 of Leadbetter, Lindgren, and Rootzén [10], if we find functions $\delta(t)$ and $u(t)$ with $0 < \delta(t) < t$ and $u(t) > 0$ and such that

(i) $\lim_{t \to +\infty} [\frac{t}{\delta(t)}] = +\infty$, and

(ii) $\lim_{t \to +\infty} \frac{t/\delta(t)}{u(t)} (1 - \frac{1}{(u(t))^2}) e^{-\frac{(u(t))^2}{2}} = +\infty$,

then we have $\lim_{t \to +\infty} P(\max_{k=1,\ldots,[\frac{t}{\delta(t)}]} Z_k \geq u(t)) = 1$. Taking into account (15) and (13), we can then deduce that

$$\lim_{t \to +\infty} P\left(M(t) \geq u(t)\sigma_{\delta(t)} + \mu\delta(t)\right) = 1, \tag{16}$$

where $\sigma_{\delta(t)}$ is obtained from (14) by replacing δ by $\delta(t)$. This finally implies (12) if we define property $\delta(t)$ and $u(t)$, and finishes the proof.
In order to define functions $\delta(t)$ and $u(t)$, we take into account the sign of $\mu = a^T\theta$ and split the definition into two cases:

- *Case* $\mu \geq 0$. Actually we only will consider here the case $\mu = 0$ because if $\mu > 0$ the proof is similar to that of Step 3 in the proof of Theorem 3.2 of Delgado [5].

Then we assume $\mu = 0$, fix an arbitrary $C > 0$, and define

$$\delta(t) \stackrel{\text{def}}{=} \frac{(Cg(t))^{1/H^-}}{(\log(\frac{t}{g(t)^{1/H^-}}))^{\frac{1}{2H^+}}} \quad \text{and} \quad u(t) \stackrel{\text{def}}{=} \frac{1}{a^T\sigma}\left(\log\left(\frac{t}{g(t)^{1/H^-}}\right)\right)^{1/2}.$$

With these definitions, conditions (i) and (ii) are easily checked. Now we can see that (16) implies (12) by considering that

$$\begin{aligned} &u(t)\sigma_{\delta(t)} + \mu\delta(t) \\ &\quad = u(t)\sigma_{\delta(t)} = u(t)\sum_{j=1}^{d} a_j\sigma_j\big(\delta(t)\big)^{H_j} \\ &\quad = \frac{1}{a^T\sigma}\left(\log\left(\frac{t}{g(t)^{1/H^-}}\right)\right)^{1/2}\left(\sum_{j=1}^{d} a_j\sigma_j\frac{(Cg(t))^{\frac{H_j}{H^-}}}{(\log(\frac{t}{g(t)^{1/H^-}}))^{\frac{H_j}{2H^+}}}\right) \\ &\quad \geq \frac{1}{a^T\sigma}\sum_{j=1}^{d} a_j\sigma_j\big(Cg(t)\big)^{\frac{H_j}{H^-}} \geq Cg(t), \end{aligned}$$

where we have used the facts that $\frac{H_j}{2H^+} \leq \frac{1}{2}$ and that $\frac{H_j}{H^-} \geq 1$.

- *Case* $\mu < 0$. Fix an (arbitrary by the moment) constant $C > 0$ and define

$$\begin{aligned} \delta(t) &\stackrel{\text{def}}{=} \frac{C(\log t)^{\frac{1}{2(1-H^-)}}}{-\mu} \quad \text{and} \\ u(t) &\stackrel{\text{def}}{=} \frac{2C(\log t)^{\frac{1}{2(1-H^-)}}}{a^T\sigma(\delta(t))^{H^-}} \\ &= \frac{2C^{1-H^-}(\log t)^{1/2}(-\mu)^{H^-}}{a^T\sigma}. \end{aligned}$$

Condition (i) is trivially satisfied and condition (ii) also holds if and only if

$$0 < C < \left(\frac{(a^T\sigma)^2}{2(-\mu)^{2H^-}}\right)^{\frac{1}{2(1-H^-)}}.$$

In this case, we have

$$\begin{aligned} u(t)\sigma_{\delta(t)} + \mu\delta(t) &= u(t)\sum_{j=1}^{d} a_j\sigma_j\big(\delta(t)\big)^{H_j} + \mu\delta(t) \geq u(t)a^T\sigma\big(\delta(t)\big)^{H^-} + \mu\delta(t) \\ &= 2C(\log t)^{\frac{1}{2(1-H^-)}} - C(\log t)^{\frac{1}{2(1-H^-)}} = C(\log t)^{\frac{1}{2(1-H^-)}}, \end{aligned}$$

and we obtain (12) from (16). □

Secondly, in Corollary 2 below, we can study, as an application, the asymptotic behavior of the total fluid in queue (or maximum queue length) under heavy traffic for the fluid model considered in this paper. For any station j, the limit distribution under heavy traffic of the fluid in queue at time t is that of $\frac{W_j(t)}{m_j}$, so we can introduce the maximum (total) amount of fluid in queue in the system on the interval $[0,t]$, by summing the fluid in queue over all stations, whose limit distribution under heavy traffic is that of

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

where W is a d-dimensional reflected fBm process on $\mathbb{R}_+^d$ with associated data $(x=0, \theta=-R\gamma, H=(H_1,\dots,H_d)^T, \sigma^2, R)$, with $H\in(\frac{1}{2},1)^d, \gamma\ge 0$, $\sigma_j^2 = m_j^2\alpha_j^2\sigma_j^{2,\lim}$, and $R=I-\tilde{P}$. Recall that $\mu = a^T\theta$ so, in this case, $\mu = (\frac{1}{m_1},\dots,\frac{1}{m_d})\theta$.

Corollary 2 *Assume that the fluid model considered in the previous sections verifies* (**HL**) *and that the flow matrix verifies assumption* (**HP**) *and also that*

$$\text{for any } i=1,\dots,d, \qquad \sum_{j\in\{1,\dots,d\}:H_j=H_i} P_{ij}\ell_{ji}^{1/2} \le 1. \tag{17}$$

We have

(a) *if* $\theta=0$, *then* $M(t)$ *grows less than any function growing faster than* t^{H^+} *and more than any function growing slower than* t^{H^-},
(b) *if* $\theta\neq 0$ *and* $\mu=0$, *then* $M(t)$ *grows less than any function growing faster than* t *and more than any function growing slower than* t^{H^-},
(c) *if* $\mu>0$, *then* $M(t)$ *grows like* t,
(d) *if* $\mu<0$, *then* $M(t)$ *grows less than any function growing faster than* t *and more than any function growing slower than* $(\log t)^{\frac{1}{2(1-H^-)}}$.

The proof of Corollary 2 is immediate from Theorem 2 by taking into account that assumption (**HaR**) becomes (17) when applied to vector $a=(\frac{1}{m_1},\dots,\frac{1}{m_d})^T$ and matrix $R=I-\tilde{P}$. The fact that P is a sub-stochastic matrix does not necessarily implies (17), therefore this condition should be imposed.

Finally, we introduce the level-crossing times for fluid in queue process as usual: for any $b>0$, let

$$T(b) \stackrel{\text{def}}{=} \inf\left\{ s\ge 0 : \sum_{j=1}^{d} \frac{1}{m_j} W_j(s) \ge b \right\}.$$

Then, using that the level-crossing times and the maximum are related by means of

$$\big\{T(b)\le t\big\} = \big\{M(t)\ge b\big\},$$

by direct application of Corollary 2, we obtain the following.

Corollary 3 *Under the same assumptions of Corollary 2, we have*

(a) *if* $\theta = 0$, *then the growth of* $T(b)$ *as* $b \to +\infty$ *is between that of* b^{1/H^+} *and that of* b^{1/H^-},
(b) *if* $\theta \neq 0$ *and* $\mu = 0$, *then the growth of* $T(b)$ *as* $b \to +\infty$ *is between that of* b *and that of* b^{1/H^-},
(c) *if* $\mu > 0$, *then* $T(b)$ *grows as* b *as* $b \to +\infty$,
(d) *if* $\mu < 0$, *then the growth of* $T(b)$ *as* $b \to +\infty$ *is between that of* b *and that of* $\exp(b^{2(1-H^-)})$.

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. R. Delgado was supported by Ministerio de Educación y Ciencia de España and FEDER, project ref. MTM2009-08869.

Appendix

The *invariance principle* of Williams and its corollary, Theorem 4.1, and Corollary 4.3 [14], respectively, are stated and proved for a reflected Brownian motion, but these results, whose proofs rely heavily on the *oscillation inequality* (Theorem 5.1 [14]), do not depend on fact on the law of the process, as can be checked by following the steps of the proof. For convenience of the reader, we present the particular (simplified) version of Corollary 4.3 [14] we use in our proof:

Version of Corollary 4.3 [14] *Assume that matrix* R *satisfies* (**HR**). *For each positive integer* n, *let* R^n *be a* $d \times d$ *matrix and* W^n, X^n, Y^n *be processes with paths in* $\mathscr{D}^d$ *defined on some probability space such that*

(i)′ $W^n(0) = 0$ *and* $W^n(t) \in \mathbb{R}^d_+$ *for all* $t \geq 0$, *a.s.*,
(ii)′ $W^n = X^n + R^n Y^n$ *a.s.*,
(iii)′ X^n *converges in distribution as* $n \to +\infty$ *to some process* X *whose paths live in* $\mathscr{C}^d$,
(iv)′ Y^n *has non-decreasing paths, and for each* $j = 1, \ldots, d$, *a.s.*, $Y^n_j(0) = 0$ *and* $\int_0^{+\infty} 1_{\{W^n_j(s)>0\}}\, dY^n_j(s) = 0$,
(v)′ R^n *converges to* R *as* $n \to +\infty$.

Then there exists $\mathsf{D}\text{-}\lim_{n\to+\infty}(W^n, X^n, Y^n) = (W, X, Y)$, *where the limit* (W, X, Y) *satisfies conditions* (i), (ii) *and* (iv) *of Definition 2* (*with* $W(0) = X(0) = Y(0) = 0$), *that is*, $W = X + RY$ *is a Skorokhod decomposition*.

Brief justification First, note that we use Theorem 4.1 [14] in a particular situation in which the probability measure ν on $\mathbb{R}^d_+$ gives probability 1 to the point $0 \in \mathbb{R}^d_+$, $\alpha^n = \gamma^n = \delta^n = 0$ for all $n \geq 1$, and hypothesis (iii) of Theorem 4.1 is replaced by (iii)′. The proof of this theorem does not uses any specific property of the Brownian motion process. Indeed, the tightness of sequence $\{X^n\}_n$ is a consequence of (iii)′ and $\{(W^n, X^n, Y^n)\}_n$ inherits tightness from it (the necessary and sufficient conditions for tightness given in Corollary 3.7.4 of Ethier and Kurtz [7] are verified). Moreover, by Theorem 3.10.2 [7], any (weak) limit point of $\{(W^n, X^n, Y^n)\}_n$ has continuous paths. Let (W, X, Y) be a (weak) limit point, that is, there is a subsequence

$\{(W^{n_k}, X^{n_k}, Y^{n_k})\}_k$ such that $\text{D-lim}_{k\to+\infty}(W^{n_k}, X^{n_k}, Y^{n_k}) = (W, X, Y)$. Then the Skorokhod representation theorem (Theorem 3.1.8 [7]) is used to replace the above sequence of processes by one that is term-by-term equivalent in distribution to the original one and which a.s. converges uniformly on compact intervals. With this simplification, it is easily seen that the limit triplet (W, X, Y) inherits properties (i), (ii) and (iv) of Definition 2 from properties (i)′, (ii)′, and (iv)′, except for

$$\int_0^{+\infty} 1_{\{W_j(s)>0\}} dY_j(s) = 0,$$

whose proof is technically more complicated and can be seen on pages 21 and 22 of [14]. Now, instead of using Theorem 3.1 [14] to ensure that all (weak) limit points of $\{(W^n, X^n, Y^n)\}_n$ have the same law, as is done in the proof of Theorem 4.1 [14], we use that by assumption (**HR**) on matrix R, the law of the pair (W, Y) is unique (see Remark 2), which gives the desired result. Combining this uniqueness with tightness, it follows that the whole sequence $\{(W^n, X^n, Y^n)\}_n$ converges in distribution to a triplet (W, X, Y) which satisfies conditions (i), (ii), and (iv) of Definition 2 (with $W(0) = X(0) = Y(0) = 0$). □

References

1. Bernard, A., el Kharroubi, A.: Régulations déterministes et stochastiques dans le premier orthant de $\mathbb{R}^n$. Stoch. Stoch. Rep. **34**, 149–167 (1991)
2. Biagini, F., Hu, Y., Øksendal, B., Sulem, A.: A stochastic maximum principle for processes driven by fractional Brownian motion. Stoch. Process. Appl. **100**, 233–253 (2002)
3. Debicki, K., Mandjes, M.: Traffic with an fBm limit: convergence of the stationary workload process. Queueing Syst. **46**, 113–127 (2004)
4. Delgado, R.: A reflected fBm limit for fluid models with ON/OFF sources under heavy traffic. Stoch. Process. Appl. **117**, 188–201 (2007)
5. Delgado, R.: On the reflected fractional Brownian motion process on the positive orthant: asymptotics for a maximum with application to queueing networks. Stoch. Models **26**, 272–294 (2010)
6. Duncan, T.E., Jin, Y.: Maximum queue length of a fluid model with an aggregated fractional Brownian input. In: Markov Processes and Related Topics. IMS Collections, vol. 4, pp. 235–251 (2008)
7. Ethier, S.N., Kurtz, T.G.: Markov Processes: Characterization and Convergence. Wiley, New York (1986)
8. Fitzpatrick, J., Murphy, S., Murphy, J.: SCTP based handover mechanism for VoIP over IEEE 802.11b wireless LAN with heterogeneous transmission rates. In: Proc. of IEEE International Conference on Communications, ICC, 2006 (2006)
9. Lavancier, F., Philippe, A., Surgailis, D.: Covariance function of vector self-similar processes. Stat. Probab. Lett. **79**, 2415–2421 (2009)
10. Leadbetter, M.R., Lindgren, G., Rootzén, H.: Extremes and Related Properties of Random Sequences and Processes. Springer, Berlin (1983)
11. Lee, C., Weerasinghe, A.: Stationarity and control of a tandem fluid network with fractional brownian motion input. Adv. Appl. Probab. **43**(3), 847–874 (2011)
12. Taqqu, M.S., Willinger, W., Sherman, R.: Proof of a fundamental result in self-similar traffic modeling. Comput. Commun. Rev. **27**, 5–23 (1997)
13. Whitt, W.: Stochastic-Process Limits. An Introduction to Stochastic-Process Limits and Their Applications to Queues. Springer Series in Operations Research (2002)
14. Williams, R.J.: An invariance principle for semimartingale reflecting Brownian motions in an orthant. Queueing Syst. **30**, 5–25 (1998)
15. Zeevi, A.J., Glynn, P.W.: On the maximum workload of a queue fed by fractional Brownian motion. Ann. Appl. Probab. **10**(4), 1084–1099 (2000)