A heavy-traffic limit of a two-station fluid model with heavy-tailed On/Off sources, feedback and flexible servers

ROSARIO DELGADO
Universitat Autònoma de Barcelona
Department of Mathematics
Edifici C, Av. de l'Eix Central, s/n. Campus de la UAB
08193 Cerdanyola del Vallès
SPAIN
delgado@mat.uab.cat

Abstract: We consider a two-station fluid model that can be approximated under heavy-traffic by a reflected fractional Browian motion (rfBm) process on a convex polyhedron. Specifically, we prove a heavy-traffic limit theorem for a two single-server workstation fluid model with feedback and flexible servers. Flexibility here means that one of the servers is capable to help the other. The non-deterministic arrival process is generated by a large enough number of heavy-tailed On/Off sources, N. We introduce the adequate definition of *workload*, and scaling conveniently by a factor r and by N, and letting N and r approach infinity (in this order), we prove that the scaled workload process converges to a rfBm on a convex polyhedron.

Key–Words: fluid network; flexible servers; reflected fractional Brownian motion; convex polyhedron; On/Off sources; workload process; heavy-traffic limit; Skorokhod problem.

1 Introduction

The aim of this paper is to investigate the asymptotic behavior of a fluid model under heavy-traffic. The model consists of a network composed of two single-server workstations that process continuous fluid, with an infinite-capacity buffer at each one. The model, which is portrayed schematically in Figure 1, allows feedback and one of the servers to help the other.

There are two fluid classes, and class-j fluid ($j = 1, 2$) is primarily assigned to server j, which works at station j. We assume that fluid is processed in a first-in-first-out (FIFO) basis within each class. The sense in which one of the servers can help the other is the following: whenever station 2 becomes empty while class-1 fluid is awaiting at station 1, a floodgate opens and fluid begins to be transferred to station 2 so that while the situation persists, class-1 fluid is simultaneously processed by both servers (possibly at different speeds). We assume that there is no travel delay (setup time). The situation continues until either the amount of class-1 fluid in the system runs out, in which case both servers are at rest thereafter until new fluid arrive, or class-2 fluid reaches station 2 from outside, whichever happens first. In the latter case, the fluid transfer immediately ceases (the floodgate closes) and server 2 starts processing of class-2 fluid, while class-1 fluid processing continues by server 1. Is in this sense that we say that server 2 is flexible, since it gives support to the other, although the converse is not allowed. We assume that server 1 cannot be idle if there is class-1 fluid awaiting for processing at station 1, while server 2 cannot be idle if there is fluid awaiting at any of the two stations (nonidling policy).

Models with flexible servers have been used in real-life systems, including service centers, production systems, computer networks with rescheduling of jobs, parallel computing systems where processors have overlapping capabilities, and manufacturing applications in which machines may have differing primary functions and some overlapping secondary ones. See for instance [6], [7] and references therein.

Moreover, feedback is allowed: after processing of class-1 fluid (by either server), a proportion p_{11} needs reprocessing and is sent back to station 1, while the rest goes outside the network. Similarly, after processing (by server 2), a proportion p_{22} of class-2 fluid needs to be reprocessed by server 2, while a proportion p_{21} needs reprocessing but as class-1 fluid, and consequently switches class and is directed to station 1; the rest goes outside the network.

Actually, this paper presents a hybrid between the fluid model with feedback introduced in [1], with two workstations, and the cascade fluid model of [5] in which server 2 is flexible in agreement wit our definition of flexibility. Our objective is to explore the implications on the asymptotic behavior under heavy

traffic of allowing flexibility in the fluid model with feedback of [1], as well as that of allowing feedback in the cascade fluid model of [5].

We assume that for each fluid class, the process of external arrivals is a non-deterministic aggregated cumulative process generated by a large enough number of heavy tailed On/Off sources, N. This assumption relies on the presence of long-range dependence and self-similar traffic pattern in modern high-speed network traffic, and the fact that one simple physical explanation for this phenomenon is the superposition of many heavy-tailed On/Off sources (see [10], [1]).

We consider a double sequence of fluid models indexed by r (a parameter of change of scale) and N, the number of On/Off sources, whose traffic intensities tend to 1 in some sense as r and N go to infinity (*heavy-traffic condition*), and we prove a limit theorem for the two-dimensional workload process. Indeed, in Theorem 9 we prove that after adequate scaling, the workload process converges to a two-dimensional *reflected fractional Brownian motion* (rfBm) process living in a convex polyhedron (which is not the positive orthant).

The organization of the paper is as follows. In Section 2 we set up notation and preliminary definitions. Section 3 is devoted to the introduction of the model, the processes used to measure its performance and the heavy-traffic condition. In Section 4 our main result is stated and proved, and in the Appendix we present an *Invariant Principle*, which is a key ingredient in the proof of the heavy-traffic limit theorem.

2 Notations and preliminaries

Vectors will be column vectors and v^T means the transpose of a vector (or a matrix) v. By $diag(v)$ we denote the diagonal matrix with diagonal elements the components of vector v (in the same order). Inequalities for vectors must be understood in the componentwise sense. For any fixed $d \geq 1$, the identity matrix of dimension d is denoted by I_d. For any $d \times m$ matrix $A = (a_{ij})_{i=1,\ldots,d,\, j=1,\ldots,m}$, let $|A| \stackrel{\text{def}}{=} \max_{1\leq j\leq m} \big(\sum_{i=1}^{d} |a_{ij}|\big)$ (where $|x|$ denotes the absolute value of $x \in \mathbb{R}$), and $det(A)$ denotes the determinant of A if $d = m$. We will say that a sequence of $d\times m$ matrices $\{A^n\}_n$ converges to a $d\times m$ matrix A if $|A^n - A| \to 0$ as n tends to $+\infty$ (this convergence is equivalent to the convergence in the component-wise sense), and we will denote it simply $\lim_{n\to+\infty} A^n = A$ or $A^n \to A$. The inner product of a couple of vectors $u,\, v \in \mathbb{R}^d$ is $\langle u, v\rangle = \sum_{i=1}^{d} u_i\, v_i$.

Let $\mathcal{C}^d$ be the space of continuous functions ω from $[0, +\infty)$ to $\mathbb{R}^d$, with the topology of the uniform convergence on compact time intervals, and $\mathcal{D}^d$ the space of continuous on the right with limits on the left functions, endowed with the usual Skorokhod $\mathcal{J}_1-$topology. All stochastic processes in this paper will be assumed to have paths in $\mathcal{D}^d$, for some $d \geq 1$.

A sequence of stochastic processes $\{X^n\}_{n\geq 1}$ is said to be *tight* if the induced measures on $\mathcal{D}^d$ form a tight sequence (that is, the sequence of induced measures is weakly relatively compact in the space of probability measures on $\mathcal{D}^d$).

We will use $\mathcal{D} - \lim$ to denote the *convergence in distribution* on $\mathcal{C}^d$ or $\mathcal{D}^d$ (or *weak convergence*). That is, we write $\mathcal{D} - \lim_{n\to+\infty} X^n = X$ if the sequence of probability measures induced in $\mathcal{D}^d$ by $\{X^n\}_n$ converges weakly to that induced by X .

The sequence of processes $\{X^n\}_n$ is called $\mathcal{C}-$*tight* if it is tight, and if each weak limit point, obtained as a weak limit along a subsequence, almost surely has sample paths in $\mathcal{C}^d$.

Reflected fractional Brownian motion (rfBm) is a stochastic process that has been widely used in the context of heavy-traffic limit theorems when the arrival processes are generated by a large number of heavy-tailed On/Off sources. See for instance [1]-[3], in which the rfBm lives in the positive orthant, and [4], [5], in which lives in a convex polyhedron with constant directions of reflection along each face. We reproduce here this last definition for the sake of completeness.

Definition 1 (convex polyhedron) *A convex polyhedron S on $\mathbb{R}^d$ can be defined algebraically as the set of solutions to a system of linear inequalities:*

$$\begin{aligned} S \stackrel{\text{def}}{=} &\{x \in \mathbb{R}^d : \langle v^\ell, x\rangle \geq 0 \quad \textit{for all } \ell = 1, \ldots, d\} \\ =&\{x \in \mathbb{R}^d : \Upsilon\, x \geq 0\} \end{aligned}$$

where $v^1, \ldots, v^d \in \mathbb{R}^d$, Υ being the $d \times d$ matrix whose row vectors are $v^1, \ldots, v^d$. That is, $S = \bigcap_{\ell=1}^{d} G_\ell$ where $G_\ell = \{x \in \mathbb{R}^d : \langle v^\ell, x\rangle \geq 0\}$. The boundary of S is $\partial S = \cup_{\ell=1}^{d} F_\ell$, where $F_\ell = \{x \in S : \langle v^\ell, x\rangle = 0\}$, $\ell = 1, \ldots, d$, are the boundary faces of S.

We write $S(\Upsilon)$ to emphasize that the convex polyhedron is determined by the matrix Υ. It is assumed that the interior of $S(\Upsilon)$ is not empty and the set $\{v^1, \ldots, v^d\}$ is minimal. Then, $n^\ell = \frac{v^\ell}{||v^\ell||}$ is the inward unit normal to F_ℓ that points into the interior of S.

Associated to the convex polyhedron $S(\Upsilon)$ we introduce the *directions of reflection*, which are constant along each face, as the column vectors of a $d \times d$ matrix R, which are denoted by $u^1, \ldots, u^d$.

Definition 2 (rfBm on a convex polyhedron) *Let $S(\Upsilon)$ be a $d-$dimensional convex polyhedron as in Definition 1, with associated $d\times d$ matrix of directions of reflection R. A reflected fractional Brownian motion on $S(\Upsilon)$ associated with data $(x,\, H,\, \theta,\, \Gamma,\, R)$, where $x \in S(\Upsilon)$, $H \in (0,\, 1)$, $\theta \in \mathbb{R}^d$ and Γ is a $d\times d$ positive definite matrix, is a $d-$dimensional process $W = \{W(t) = (W_1(t), \ldots, W_d(t))^T,\, t \geq 0\}$ such that*
(i) W has continuous paths and $W(t) \in S(\Upsilon)$ for all $t \geq 0$ a.s.,
(ii) $W = X + RV$ a.s., with X and V two $d-$ dimensional processes defined on the same probability space and verifying:
(iii) X is a fractional Brownian motion (fBm) process with associated data $(x,\, H,\, \theta,\, \Gamma)$, that is, it is a continuous Gaussian process starting from x, with mean function $E\big(X(t)\big) = x + \theta\, t$ for any $t \geq 0$, and with covariance function given by $Cov\big(X(t), X(s)\big) = E\Big(\big(X(t) - (x+\theta\, t)\big)\big(X(s) - (x + \theta s)\big)^T\Big) = \Gamma_H(s,t)\Gamma$ if $t,\, s \geq 0$, where $\Gamma_H(s,t) = \dfrac{1}{2}\left(\, t^{2H} + s^{2H} - |t-s|^{2H}\,\right)$, and
(iv) V has continuous and non-decreasing paths, and for each $\ell = 1, \ldots, d$, a.s., $V_\ell(0) = 0$ and $V_\ell(t) = \int_0^t 1_{\{W(s)\in F_\ell\}}\, dV_\ell(s)$ for all $t \geq 0$ (that is, V_ℓ can only increase when W is on the boundary face F_ℓ).
If conditions (i), (ii) and (iv) are met, we say that the pair $(W,\, V)$ is a solution of the Skorokhod Problem associated to X on the convex polyhedron $S(\Upsilon)$ with associated matrix of directions of reflection R.

Remark 3 *Strong existence and uniqueness of the solution of a Skorokhod problem can be ensured if the column vectors of R are linearly independent, and matrix $\Psi = \Upsilon R$ verifies that the entries off the diagonal are non-negative and the following condition (the generalized Harrison-Reiman condition), holds:*

(**HR**) *The matrix Θ obtained from $\Psi - I_d$ by replacing its entries by their absolute values, has spectral radius strictly less than 1.*

(See Remark 1 [4] for a detailed justification.)

Loosely speaking, the rfBm process starts in the interior of S and behaves like a fBm being constrained to remain within S by reflection on the boundary. Vector u^ℓ, gives the direction of the reflection at the boundary face F_ℓ, and v^ℓ its intensity. On the intersection of two or more faces, the direction of reflection is given by a linear combination of the corresponding reflection vectors.

3 The two-station fluid model with flexible servers

This section provides a detailed exposition of the model and makes preparations for the heavy-traffic limit theorem. For facilitate access to the topics, the subsections are rendered as self-contained as possible.

3.1 Introducing the model

The basic features of the model have been explained in the Introduction. Now we go deeper into it. We assume $0 \leq p_{11} < 1$ and $0 \leq p_{22} + p_{21} < 1$, and

$$P \stackrel{\text{def}}{=} \begin{pmatrix} p_{11} & 0 \\ p_{21} & p_{22} \end{pmatrix}$$

is the (sub-stochastic) *"flow"* or *"routing"* matrix of the fluid model.

As in [1]-[5], we assume that for each station $j = 1, 2$, there are N i.i.d. external sources sending fluid to it, and that each source can be On or 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^{on} and f^{off} be the probability density functions corresponding to the lengths of On and Off-periods, which are non-negative and heavy-tailed. Therefore, their (positive) expected values are

$$\mu^{\text{on}} \stackrel{\text{def}}{=} \int_0^{+\infty} u f^{\text{on}}(u)du,\; \mu^{\text{off}} \stackrel{\text{def}}{=} \int_0^{+\infty} u f^{\text{off}}(u)du.$$

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

$$\begin{aligned} \int_x^{+\infty} f^{\text{on}}(u)\, du &\sim x^{-\beta^{\text{on}}}\, L^{\text{on}}(x), \qquad (1)\\ \int_x^{+\infty} f^{\text{off}}(u)\, du &\sim x^{-\beta^{\text{off}}}\, L^{\text{off}}(x)\,, \end{aligned}$$

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

In what follows, we use subindex j to denote the quantities related to class-j fluid, $j = 1, 2$, and subindex 12 specifically to that quantities related to processing of class-1 fluid by server 2. We define the cumulative external class-j fluid arrived up to time t (by the N sources) at station j by:

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

where $\{U_j^{(n)}(t),\, t \geq 0\}$, $n = 1, \ldots, N$, is a family of binary time series with $U_j^{(n)}(t) = 1$ meaning that at time t the source n of station j is On (and it is sending fluid to station j at constant rate $\alpha_j^N > 0$), and $U_j^{(n)}(t) = 0$ meaning that it is Off. Let $\alpha^N = (\alpha_1^N,\, \alpha_2^N)^T$. The two component processes of the (non-deterministic) *cumulative external fluid arrival* process $E^N = \{E^N(t) = \left(E_1^N(t),\, E_2^N(t)\right)^T,\, t \geq 0\}$, are assumed to be independent. Let $\tilde{\alpha}^N \stackrel{\text{def}}{=} \alpha^N \frac{\mu^{\text{on}}}{\mu^{\text{on}}+\mu^{\text{off}}}$ and define $\lambda^N = (\lambda_1^N,\, \lambda_2^N)^T$ to be the unique two-dimensional vector solution to the *traffic equation*

$$\lambda^N = \tilde{\alpha}^N + P^T \lambda^N,$$

that is, $\lambda^N = Q\tilde{\alpha}^N$ where $Q \stackrel{\text{def}}{=} (I_2 - P^T)^{-1}$, which is well defined since matrix P has spectral radius less than one. Note that λ_j^N can be thought as the long run fluid rate into station j. Assume that $\lambda = \lim_{N\to+\infty} \lambda^N$ exists, $\lambda = (\lambda_1,\, \lambda_2)^T$. This implies that $\alpha = (\alpha_1,\, \alpha_2)^T = \lim_{N\to+\infty}(\alpha_1^N,\, \alpha_2^N)^T$ also exists.

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 the system. We will use r as a scalar parameter in time. For the $(r,\, N)$ fluid model, suppose that server 1 processes class-1 fluid at a constant rate $\mu_1^{r,N} > 0$ if station 1 were never idle, and server 2 processes class-2 fluid at constant rate $\mu_2^{r,N} > 0$ if server 2 devote all time to class-2 fluid, and processes class-1 fluid at a constant rate $\mu_{12}^{r,N} > 0$, not necessarily equal to $\mu_1^{r,N}$ nor to $\mu_2^{r,N}$, if station 2 devoted all time to this fluid class. We assume that $\lim_{N\to+\infty}(\mu_1^{r,N}, \mu_2^{r,N}, \mu_{12}^{r,N})$ exists and is positive, and does not depend on r; we denote it by $(\mu_1,\, \mu_2,\, \mu_{12})$. We also introduce the *fluid traffic intensity* $\rho^{r,N} = (\rho_1^{r,N},\, \rho_2^{r,N})^T$ by

$$\rho_1^{r,N} \stackrel{\text{def}}{=} \frac{\lambda_1^N}{\mu_1^{r,N}},\ \rho_2^{r,N} \stackrel{\text{def}}{=} \frac{\lambda_1^N - \mu_1^{r,N}}{\mu_{12}^{r,N}} + \frac{\lambda_2^N}{\mu_2^{r,N}}. \quad (3)$$

3.2 Performance processes

To measure the performance of our model we introduce some processes. Our definition of *workload process* $W^{r,N} = (W_1^{r,N},\, W_2^{r,N})^T$, which is not as trivial as it might initially seem, is adopted from [5] and agrees with the one given in [8]: $W_1^{r,N}(t)$ represents the total time of service that would be required to complete processing of the amount of class-1 fluid in the system at time t, if server 1 were required to complete its processing without future help from server 2, while $W_2^{r,N}(t)$ represents the total time of service that would be required to complete processing of all the class-1 and class-2 fluid in the system at time t, if server 2 were required to complete the processing of both without help from server 1. We assume $W^{r,N}(0) = 0$.

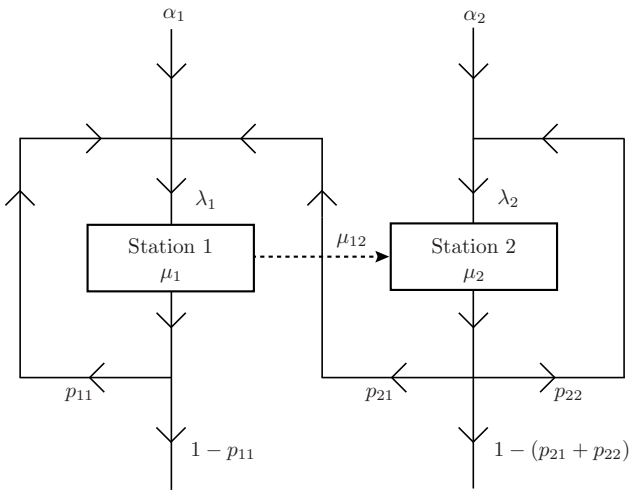

Figure 1: A two-station fluid model with feedback and flexible servers.

The *cumulative idle-time process* $Y^{r,N} = (Y_1^{r,N},\, Y_2^{r,N})^T$ is defined by: $Y_j^{r,N}(t)$ is the cumulative amount of time that server j has been idle during the time interval $[0,\, t]$, that is,

$$Y_1^{r,N}(t) \stackrel{\text{def}}{=} \int_0^t 1_{\{W_1^{r,N}(s)=0\}}\, ds,$$
$$Y_2^{r,N}(t) \stackrel{\text{def}}{=} \int_0^t 1_{\{W^{r,N}(s)=0\}}\, ds\,. \quad (4)$$

The *total service time process* $T^{r,N} = (T_1^{r,N},\, T_2^{r,N},\, T_{12}^{r,N})^T$ is defined by: $T_j^{r,N}(t)$ is the total service time devoted to class-j fluid (by server j) in the interval $[0,\, t]$, $j = 1,\, 2$, and $T_{12}^{r,N}(t)$ is the total service time devoted to class-1 by server 2 in the same time interval.

Directly related to feedback, we also introduce processes $A^{r,N} = (A_1^{r,N},\, A_2^{r,N})^T$ and $D^{r,N} = (D_1^{r,N},\, D_2^{r,N})^T$ by: $A_1^{r,N}(t)$ is the total fluid arriving at station 1 (as class-1 fluid) up to time t, including both feedback flow (from both stations) and external input. $A_2^{r,N}(t)$ is the total fluid arriving at station 2 (as class-2 fluid) up to time t, including both feedback flow (from station 2) and external input. Note that in the definition of $A_2^{r,N}$ we do not include fluid transferred from station 1 when the floodgate is open, which is class-1 fluid. $D_1^{r,N}(t)$ is the total amount of

class-1 fluid departing from station 1 or from station 2 (either by leaving the network or not) up to time t. $D_2^{r,N}(t)$ is the total amount of class-2 fluid departing from station 2 (either by leaving the network or not), up to time t. We assume $A_j^{r,N}(0) = D_j^{r,N}(0) = 0,\ j = 1, 2$.

These processes are related by means of the equalities:

$$W_1^{r,N}(t) = \frac{A_1^N(t)}{\mu_1^{r,N}} - \big(T_1^{r,N}(t) + \frac{\mu_{12}^{r,N}}{\mu_1^{r,N}} T_{12}^{r,N}(t)\big), \tag{5}$$

$$W_2^{r,N}(t) = \frac{A_2^N(t)}{\mu_2^{r,N}} - T_2^{r,N}(t) + \frac{\mu_1^{r,N}}{\mu_{12}^{r,N}} W_1^{r,N}(t), \tag{6}$$

$$Y_1^{r,N}(t) = t - T_1^{r,N}(t), \tag{7}$$

$$Y_2^{r,N}(t) = t - \big(T_2^{r,N}(t) + T_{12}^{r,N}(t)\big). \tag{8}$$

The interpretation of (5) is that $A_1^N(t)/\mu_1^{r,N}$ is the amount of time required by server 1 to process all the class-1 fluid arrived up to time t to station 1, while $T_1^{r,N}(t) + \frac{\mu_{12}^{r,N}}{\mu_1^{r,N}} T_{12}^{r,N}(t)$ represents the part of this time yet consumed at instant t, by server 1, which is $T_1^{r,N}(t)$, and by server 2, which is $T_{12}^{r,N}(t)$ conveniently rescaled since the service time for class-1 fluid is different when processed by server 1 or by server 2. With respect to (6), $A_2^N(t)/\mu_2^{r,N} - T_2^{r,N}(t)$, is the amount of time required by server 2 to process all the class-2 fluid arrived to station 2 up to time t minus the amount of time devoted to this processing. By the other part, we add $W_1^{r,N}(t)$ conveniently rescaled representing the amount of time required by server 2 to process all the class-1 fluid at buffer 1 at time t. Formulae (7) and (8) are self-explanatory: the length of the interval $[0,\ t]$ can be split into two parts: idle time $Y_j^{r,N}(t)$, and working time ($T_1^{r,N}(t)$ for server 1, and $T_2^{r,N}(t) + T_{12}^{r,N}(t)$ for server 2).

As in [5], we introduce the following notation: $\widetilde{W}_2^{r,N}(t)$ is the portion of the workload $W_2^{r,N}$ that is exclusively due to class-2 fluid, that is,

$$\widetilde{W}_2^{r,N}(t) \stackrel{\text{def}}{=} \frac{A_2^N(t)}{\mu_2^{r,N}} - T_2^{r,N}(t)\,.$$

Then, by (6) it follows that

$$W_2^{r,N}(t) = \widetilde{W}_2^{r,N}(t) + \frac{\mu_1^{r,N}}{\mu_{12}^{r,N}} W_1^{r,N}(t)\,. \tag{9}$$

By (7) and (4),

$$T_1^{r,N}(t) = \int_0^t 1_{\{W_1^{r,N}(s)>0\}} ds,$$

and by definition of $\widetilde{W}_2^{r,N}$, $T_2^{r,N}$ and $T_{12}^{r,N}$, we can write

$$T_2^{r,N}(t) = \int_0^t 1_{\{\widetilde{W}_2^{r,N}(s)>0\}}\, ds\,,$$

$$T_{12}^{r,N}(t) = \int_0^t 1_{\{W_1^{r,N}(s)>0,\, \widetilde{W}_2^{r,N}(s)=0\}}\, ds\,, \tag{10}$$

using (8), (4) and that

$$W^{r,N} = 0 \Longleftrightarrow W_1^{r,N} = \widetilde{W}_2^{r,N} = 0\,.$$

Finally, we introduce the process $V^{r,N} = (V_1^{r,N},\, V_2^{r,N})^T$ as the function of $Y^{r,N}$ and $T^{r,N}$ given by:

$$V_1^{r,N}(t) \stackrel{\text{def}}{=} Y_1^{r,N}(t) + \frac{\mu_{12}^{r,N}}{\mu_1^{r,N}}\, Y_2^{r,N}(t), \tag{11}$$

$$V_2^{r,N}(t) \stackrel{\text{def}}{=} Y_2^{r,N}(t) + T_{12}^{r,N}(t)\,. \tag{12}$$

We are interested in express workload process in terms of the external arrival process $E^{r,N}$ and process $V^{r,N}$. From this relation, that will be proved in Lemma 5, we will deduce in Lemma 7 the Skorokhod representation needed to prove the heavy-traffic limit theorem. We begin with the expression of process $A^{r,N}$ in terms of $E^{r,N}$ and $W^{r,N}$, which is set in the next lemma.

Lemma 4 *Processes $A^{r,N}$, $E^{r,N}$ and $W^{r,N}$ are related by means of the following identity:*

$$A^{r,N}(t) = Q\, E^{r,N}(t) - QP^T (M^{r,N})^{-1} W^{r,N}(t) \tag{13}$$

where

$$M^{r,N} = \begin{pmatrix} \frac{1}{\mu_1^{r,N}} & 0 \\ \frac{1}{\mu_{12}^{r,N}} & \frac{1}{\mu_2^{r,N}} \end{pmatrix}.$$

Proof: By definition of process $A^{r,N}$,

$$A^{r,N}(t) = E^{r,N}(t) + P^T D^{r,N}(t) \tag{14}$$

since $P^T D^{r,N}(t)$ is the total amount of fluid arriving from *feedback*. By the other way, according to definition of process $D^{r,N}$, $D_1^{r,N}(t) = A_1^{r,N}(t) - \mu_1^{r,N} W_1^{r,N}(t)$ and $D_2^{r,N}(t) = A_2^{r,N}(t) - \mu_2^{r,N} \widetilde{W}_2^{r,N}(t)$, which in matricial form can be expressed as

$$D^{r,N}(t) = A^{r,N}(t) - (M^{r,N})^{-1} W^{r,N} \tag{15}$$

from (9) and the fact that

$$(M^{r,N})^{-1} = \begin{pmatrix} \mu_1^{r,N} & 0 \\ -\frac{\mu_1^{r,N}\mu_2^{r,N}}{\mu_{12}^{r,N}} & \mu_2^{r,N} \end{pmatrix}.$$

Substituting (15) into (14) yields $A^{r,N}(t) = E^{r,N}(t)+P^T A^{r,N}(t)-P^T(M^{r,N})^{-1}W^{r,N}(t)$, which establishes the identity of the lemma on account of the definition of matrix Q. □

Lemma 5 *Processes $W^{r,N}$, $E^{r,N}$ and $V^{r,N}$ verify the following relation:*

$$W^{r,N}(t) = M^{r,N}E^{r,N}(t) - C^{r,N}\delta^{r,N}t + R^{r,N}V^{r,N}(t) \tag{16}$$

where

$$C^{r,N} = M^{r,N}\,Q^{-1}\,(M^{r,N})^{-1},\ \delta^{r,N} = (1,\ 1+\frac{\mu_1^{r,N}}{\mu_{12}^{r,N}})^T$$

and $R^{r,N} = C^{r,N}B^{r,N}$, with

$$B^{r,N} = \begin{pmatrix} 1 & -\frac{\mu_{12}^{r,N}}{\mu_1^{r,N}} \\ \frac{\mu_1^{r,N}}{\mu_{12}^{r,N}} & 0 \end{pmatrix}.$$

Proof: From (5), (7), (11) and (12) we can rewrite $W_1^{r,N}$ as

$$W_1^{r,N}(t) = \frac{A_1^{r,N}(t)}{\mu_1^{r,N}} - t + V_1^{r,N}(t) - \frac{\mu_{12}^{r,N}}{\mu_1^{r,N}}V_2^{r,N}(t). \tag{17}$$

Substituting (5) into (6), and then combining (7) with (8), yields

$$\begin{aligned} W_2^{r,N}(t) = &\frac{A_1^{r,N}(t)}{\mu_{12}^{r,N}} + \frac{A_2^{r,N}(t)}{\mu_2^{r,N}} \\ &- (1+\frac{\mu_1^{r,N}}{\mu_{12}^{r,N}})t + \frac{\mu_1^{r,N}}{\mu_{12}^{r,N}}Y_1^{r,N}(t) + Y_2^{r,N}(t), \end{aligned}$$

that applying (11) can be rewritten as

$$\begin{aligned} W_2^{r,N}(t) = &\frac{A_1^{r,N}(t)}{\mu_{12}^{r,N}} + \frac{A_2^{r,N}(t)}{\mu_2^{r,N}} \\ &- (1+\frac{\mu_1^{r,N}}{\mu_{12}^{r,N}})t + \frac{\mu_1^{r,N}}{\mu_{12}^{r,N}}V_1^{r,N}(t). \end{aligned} \tag{18}$$

We can express (17) and (18) in matricial form as

$$W^{r,N}(t) = M^{r,N}A^{r,N}(t) - \delta^{r,N}t + B^{r,N}V^{r,N}(t). \tag{19}$$

Substituting (13) into (19) we can assert that

$$\begin{aligned} W^{r,N}(t) = &M^{r,N}QE^{r,N}(t) \\ &- M^{r,N}QP^T(M^{r,N})^{-1}W^{r,N}(t) \\ &- \delta^{r,N}t + B^{r,N}V^{r,N}(t), \end{aligned}$$

that is,

$$\begin{aligned} &(I_2 + M^{r,N}QP^T(M^{r,N})^{-1})W^{r,N}(t) \\ &= M^{r,N}QE^{r,N}(t) - \delta^{r,N}t + B^{r,N}V^{r,N}(t). \end{aligned} \tag{20}$$

It is straightforward to see that $I_2 + M^{r,N}QP^T(M^{r,N})^{-1} = M^{r,N}Q(M^{r,N})^{-1}$, whose inverse is $C^{r,N}$. It follows immediately from (20) that $W^{r,N}(t) = C^{r,N}M^{r,N}QE^{r,N}(t) - C^{r,N}\delta^{r,N}t + C^{r,N}B^{r,N}V^{r,N}(t)$, which gives the desired result since $C^{r,N}B^{r,N} = R^{r,N}$ by definition and $C^{r,N}M^{r,N}Q = M^{r,N}$. □

Remark 6 *We clearly see the existence of the following limits:*

$$\begin{aligned} M &= \lim_{N\to+\infty} M^{r,N} = \begin{pmatrix} \frac{1}{\mu_1} & 0 \\ \frac{1}{\mu_{12}} & \frac{1}{\mu_2} \end{pmatrix}, \\ B &= \lim_{N\to+\infty} B^{r,N} = \begin{pmatrix} 1 & -\frac{\mu_{12}}{\mu_1} \\ \frac{\mu_1}{\mu_{12}} & 0 \end{pmatrix}, \\ C &= \lim_{N\to+\infty} C^{r,N} = MQ^{-1}M^{-1}, \\ R &= \lim_{N\to+\infty} R^{r,N} = CB \\ &= \begin{pmatrix} 1-p_{11} & -\frac{\mu_{12}}{\mu_1}(1-p_{11}) - \frac{\mu_2}{\mu_1}p_{21} \\ \frac{\mu_1}{\mu_{12}}(1-p_{11}) & p_{11}-p_{22}-\frac{\mu_2}{\mu_{12}}p_{21} \end{pmatrix}. \end{aligned} \tag{21}$$

3.3 Sequence of convex polyhedra in $\mathbb{R}^2$

Let us define

$$\begin{aligned} G_1 &\stackrel{\text{def}}{=} \{(x,y) \in \mathbb{R}^2 \,:\, x \geq 0\}, \\ G_2 &\stackrel{\text{def}}{=} \{(x,y) \in \mathbb{R}^2 : y \geq \frac{\mu_1}{\mu_{12}}\,x\}, \end{aligned}$$

and $S \stackrel{\text{def}}{=} G_1 \cap G_2$. Then, S is a convex polyhedron in $\mathbb{R}^2$ determined by matrix

$$\Upsilon \ = \begin{pmatrix} 1 & 0 \\ -\frac{\mu_1}{\mu_{12}} & 1 \end{pmatrix}, \tag{22}$$

the set of row vectors $\{v^1,\, v^2\}$ with $v^1 = (1,0)^T$ and $v^2 = (-\frac{\mu_1}{\mu_{12}}, 1)^T$, is minimal, the boundary faces are

$$\begin{aligned} F_1 &= \{(x,y) \in S \,:\, x = 0\}, \\ F_2 &= \{(x,y) \in S \,:\, y = \frac{\mu_1}{\mu_{12}}\,x\}, \end{aligned}$$

and the boundary of S is $\partial S = F_1 \cup F_2$ (Figure 2).

We also introduce a sequence of convex polyhedra: for the (r,N) model, the corresponding convex polyhedron $S^{r,N}$ is introduced analogously to S by replacing μ_1 and μ_{12} by $\mu_1^{r,N}$ and $\mu_{12}^{r,N}$, respectively.

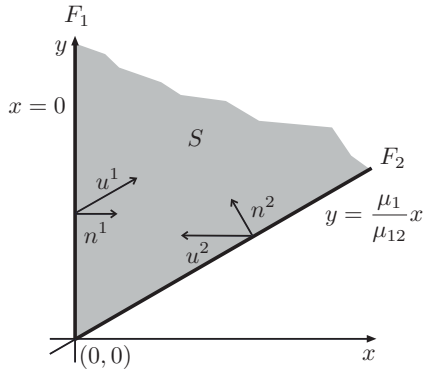

Figure 2: The convex polyhedron.

We will use superscript r, N to refer to items associated to $S^{r,N}$. For instance, $S^{r,N} = S^{r,N}(\Upsilon^{r,N})$, where

$$\Upsilon^{r,N} = \begin{pmatrix} 1 & 0 \\ -\frac{\mu_1^{r,N}}{\mu_{12}^{r,N}} & 1 \end{pmatrix}. \tag{23}$$

(Note that $\lim_{N\to+\infty} \Upsilon^{r,N} = \Upsilon$.)

We wish to stress that the key technical difficulty of our main result (Theorem 9) stems from the fact that the faces of the convex polyhedron $S(\Upsilon^{r,N})$, associated to the (r, N) model, do depend on r and N.

3.4 Scaled processes

In order to define the *scaled processes* associated with the $(r,\ N)$ model we have to introduce some notation that goes back as far as the work of Taqqu, Willinger and Sherman [10] (see also [1], [2], [5]). Set $a^{\rm on} \stackrel{\rm def}{=} \frac{\Gamma(2-\beta^{\rm on})}{(\beta^{\rm on}-1)}$ and $a^{\rm off} \stackrel{\rm def}{=} \frac{\Gamma(2-\beta^{\rm off})}{(\beta^{\rm off}-1)}$, where $\beta^{\rm on}$ and $\beta^{\rm off}$ are defined by (1). The normalization factors used below depend on b, defined by $b \stackrel{\rm def}{=} \lim_{t\to+\infty} \frac{L^{\rm on}(t)}{L^{\rm off}(t)}\, t^{\beta^{\rm off}-\beta^{\rm on}}$, which exists although it could be infinite. If $0 < b < +\infty$ (implying $\beta^{\rm on} = \beta^{\rm off}$ and $b = \lim_{t\to+\infty} \frac{L^{\rm on}(t)}{L^{\rm off}(t)}$), set $\beta \stackrel{\rm def}{=} \beta^{\rm on} = \beta^{\rm off}$, $L \stackrel{\rm def}{=} L^{\rm off}$ and

$$\sigma^{2,\lim} \stackrel{\rm def}{=} \frac{2\left((\mu^{\rm off})^2\, a^{\rm on}\, b + (\mu^{\rm on})^2\, a^{\rm off}\right)}{\left(\mu^{\rm on}+\mu^{\rm off}\right)^3 \Gamma(4-\beta)}.$$

If, on the other hand, $b = +\infty$ ($\beta^{\rm off} > \beta^{\rm on}$), set $L \stackrel{\rm def}{=} L^{\rm on}$, $\beta \stackrel{\rm def}{=} \beta^{\rm on}$ and

$$\sigma^{2,\lim} \stackrel{\rm def}{=} \frac{2\,(\mu^{\rm off})^2\, a^{\rm on}}{\left(\mu^{\rm on}+\mu^{\rm off}\right)^3 \Gamma(4-\beta)}.$$

If $b = 0$ ($\beta^{\rm off} < \beta^{\rm on}$), set $L \stackrel{\rm def}{=} L^{\rm off}$, $\beta \stackrel{\rm def}{=} \beta^{\rm off}$ and

$$\sigma^{2,\lim} \stackrel{\rm def}{=} \frac{2\,(\mu^{\rm on})^2\, a^{\rm off}}{\left(\mu^{\rm on}+\mu^{\rm off}\right)^3 \Gamma(4-\beta)}.$$

In either case, $\beta \in (1,\ 2)$. Let us define

$$H \stackrel{\rm def}{=} \frac{3-\beta}{2} \quad \left(\in (\tfrac{1}{2},\ 1)\right). \tag{24}$$

Now we can introduce the *heavy-traffic condition*, which establishes that the *fluid traffic intensity* $\rho^{r,N}$ defined by (3) tends to $e = (1,\ 1)^T$ in the following sense:

$$\textbf{(HT)} \begin{cases} \lim_{N\to+\infty} \sqrt{N}(\rho^{r,N} - e) = \widehat{\gamma}^r \in \mathbb{R}^2 \text{ and} \\ \lim_{r\to+\infty} \frac{r^{1-H}}{L^{1/2}(r)} \widehat{\gamma}^r = \gamma \in \mathbb{R}^2. \end{cases}$$

We can introduce the *scaled processes* associated with the $(r,\ N)$ fluid model and use a hat to denote them: $\widehat{W}^{r,N} = (\widehat{W}_1^{r,N}, \widehat{W}_2^{r,N})$ is defined by

$$\widehat{W}_j^{r,N}(t) \stackrel{\rm def}{=} \sqrt{N}\, \frac{W_j^{r,N}(r\,t)}{r^H\, L^{1/2}(r)}, \tag{25}$$

and similarly for the other processes except for $\widehat{E}^{r,N} = (\widehat{E}_1^{r,N}, \widehat{E}_2^{r,N})$, defined by

$$\widehat{E}_j^{r,N}(t) \stackrel{\rm def}{=} \sqrt{N}\, \frac{E_j^N(r\,t) - \tilde{\alpha}_j^N\, r\,t}{r^H\, L^{1/2}(r)}. \tag{26}$$

(where $j = 1,\ 2$). From (11) and (12) we obtain

$$\widehat{V}_1^{r,N}(t) = \widehat{Y}_1^{r,N}(t) + \frac{\mu_{12}^{r,N}}{\mu_1^{r,N}}\, \widehat{Y}_2^{r,N}(t), \tag{27}$$

$$\widehat{V}_2^{r,N} = \widehat{Y}_2^{r,N}(t) + \widehat{T}_{12}^{r,N}(t), \tag{28}$$

from (4) it follows that

$$\widehat{Y}_1^{r,N}(t) = \sqrt{N} \frac{r^{1-H}}{L^{1/2}(r)} \int_0^t 1_{\{\widehat{W}_1^{r,N}(s)=0\}} ds, \tag{29}$$

$$\widehat{Y}_2^{r,N}(t) = \sqrt{N} \frac{r^{1-H}}{L^{1/2}(r)} \int_0^t 1_{\{\widehat{W}^{r,N}(s)=0\}} ds, \tag{30}$$

and from (10) we deduce that

$$\widehat{T}_{12}^{r,N}(t) = \frac{\sqrt{N} r^{1-H}}{L^{1/2}(r)} \int_0^t 1_{\{\widehat{W}_1^{r,N}(s)>0, \widehat{\widehat{W}}_2^{r,N}(s)=0\}} ds. \tag{31}$$

The following lemma provides a Skorokhod decomposition that will prove extremely useful in the proof of Theorem 9 below.

Lemma 7 *The scaled processes are related by means of*

$$\widehat{W}^{r,N}(t) = \widehat{X}^{r,N}(t) + R^{r,N}\,\widehat{V}^{r,N}(t)\,,$$

with

$$\widehat{X}^{r,N} = M^{r,N}\widehat{E}^{r,N}(t) + \frac{\sqrt{N}r^{1-H}}{L^{1/2}(r)}C^{r,N}(\rho^{r,N} - e)t. \tag{32}$$

Proof: From (25), (16) and (26) we obtain

$$\begin{aligned}\widehat{W}^{r,N}(t) &= \sqrt{N}\,\frac{W^{r,N}(r\,t)}{r^H\,L^{1/2}(r)}\\ &= \frac{\sqrt{N}}{r^H\,L^{1/2}(r)}\big(M^{r,N}E^{r,N}(rt) - C^{r,N}\delta^{r,N}rt\\ &+ R^{r,N}V^{r,N}(rt)\big)\\ &= M^{r,N}\widehat{E}^{r,N}(rt) + \frac{\sqrt{N}r^{1-H}}{L^{1/2}(r)}M^{r,N}\tilde{\alpha}^N t\\ &- \frac{\sqrt{N}r^{1-H}}{L^{1/2}(r)}C^{r,N}\delta^{r,N}t + R^{r,N}\widehat{V}^{r,N}(rt).\end{aligned}$$

By using that $\tilde{\alpha}^N = Q^{-1}\lambda^N$, we can rewrite this expression as:

$$\begin{aligned}\widehat{W}^{r,N}(t) &= M^{r,N}\widehat{E}^{r,N}(rt)\\ &+ \frac{\sqrt{N}r^{1-H}}{L^{1/2}(r)}\big(M^{r,N}Q^{-1}\lambda^N - C^{r,N}\delta^{r,N}\big)t\\ &+ R^{r,N}\widehat{V}^{r,N}(rt),\end{aligned}$$

which is our claim, due to the fact that

$$\begin{aligned}&M^{r,N}Q^{-1}\lambda^N - C^{r,N}\delta^{r,N}\\ &= C^{r,N}\big(M^{r,N}\lambda^N - \delta^{r,N}\big) = C^{r,N}(\rho^{r,N} - e).\,\square\end{aligned}$$

Lemma 8 *For the (r, N) fluid model, the column vectors of matrix $R^{r,N}$ given by Lemma 5 are linearly independent and the product of matrices $\Psi^{r,N} = \Upsilon^{r,N}\,R^{r,N}$ verify that the entries outside the main diagonal are nonpositive and also condition* **(HR)** *(see Remark 3), where matrix $\Upsilon^{r,N}$ is given by* (23)*. Moreover, by taking the limit as $N \to +\infty$ (which does not depend on r) the column vectors of matrix R are linearly independent, and $\Psi = \Upsilon\,R$ verifies that the entries outside the main diagonal are nonpositive and also condition* **(HR)**, *where matrices R and Υ are given by* (21) *and* (22)*, respectively.*

The proof of this lemma is straightforward and omitted.

4 The heavy-traffic limit

Our goal now is to state that the scaled workload process $\widehat{W}^{r,N}$ converges in distribution to a two-dimensional rfBm process on the convex polyhedron $S(\Upsilon)$, when N first and then r, tend to infinity in this order, under heavy-traffic. The following result may be proved in much the same way as Theorem 1 [1] and Theorem 1 [5].

Theorem 9 (heavy-traffic limit) *Under the heavy-traffic condition* **(HT)** *the following limits exist in $\mathcal{C}^2$:*

$$\begin{aligned}\widehat{\widehat{W}}^{r} &= \mathcal{D} - lim_{N\to+\infty}\widehat{W}^{r,N}\,,\\ W &= \mathcal{D} - lim_{r\to+\infty}\widehat{\widehat{W}}^{r}\,,\end{aligned}$$

and W is a two-dimensional rfBm process on the convex polyhedron $S(\Upsilon)$ with Υ given by (22) *and associated data*

$$(x = 0,\, H,\, \theta = C\gamma,\, \Gamma,\, R)\,,$$

where $H \in (\frac{1}{2},\, 1)$ is defined by (24)*, $\gamma \in \mathbb{R}^2$ is given by condition* **(HT)**,

$$\Gamma = \sigma^{2,\lim}\begin{pmatrix} \frac{\alpha_1^2}{\mu_1^2} & \frac{\alpha_1^2}{\mu_1\,\mu_{12}}\\ \\ \frac{\alpha_1^2}{\mu_1\,\mu_{12}} & \frac{\alpha_1^2}{\mu_{12}^2} + \frac{\alpha_2^2}{\mu_2^2}\end{pmatrix} \tag{33}$$

with $\sigma^{2,\lim}$ given by Section 3.4*, and C and R are given by Remark 6.*

Proof:

Fix $r > 0$. Let us first show that Proposition 10 in the Appendix can be applied to the sequence $(\widehat{W}^{r,N},\, \widehat{X}^{r,N},\, \widehat{V}^{r,N})_N$. To see this, note that $(S^{r,N},\, R^{r,N})$ verifies conditions (A1)-(A5) [9] for any N, which is clear from Lemma 8, where $S^{r,N} = S^{r,N}(\Upsilon^{r,N})$. It remains to prove that $(\widehat{W}^{r,N},\, \widehat{X}^{r,N},\, \widehat{V}^{r,N})_N$ verifies conditions (i)-(iv) in Assumption (h) in the Appendix. Indeed,

(i) $\widehat{W}_1^{r,N}, \widehat{\widehat{W}}_2^{r,N} \geq 0$, and from (25) and (9) it follows that

$$\widehat{W}_2^{r,N}(t) = \widehat{\widehat{W}}_2^{r,N}(r\,t) + \frac{\mu_1^{r,N}}{\mu_{12}^{r,N}}\,\widehat{W}_1^{r,N}(t)\,,$$

which implies $\widehat{W}_2^{r,N}(t) \geq \frac{\mu_1^{r,N}}{\mu_{12}^{r,N}}\,\widehat{W}_1^{r,N}(t)$. This clearly forces $\widehat{W}^{r,N}(t) \in S^{r,N}$ for all $t \geq 0$.
(ii) the Skorokhod decomposition has been proved in Lemma 7.

(iii) By (27), (29) and (30) we get

$$\widehat{V}_1^{r,N}(t) = \int_0^t 1_{\{\widehat{W}^{r,N}(s)\in F_1^{r,N}\}}\, d\widehat{V}_1^{r,N}(s)$$

and by (28), (30) and (31) we analogously obtain

$$\widehat{V}_2^{r,N}(t) = \int_0^t 1_{\{\widehat{W}^{r,N}(s)\in F_2^{r,N}\}}\, d\widehat{V}_2^{r,N}(s),$$

taking into account that $F_1^{r,N} = \{(x,y) \in S^{r,N} : x = 0\}$ and $F_2^{r,N} = \{(x,y) \in S^{r,N} : y = \frac{\mu_1^{r,N}}{\mu_{12}^{r,N}}\, x\}$.

(iv) is consequence of the weak convergence of $\widehat{X}^{r,N}$ as $N \to +\infty$, which, in turn, is a consequence of Theorem 1 [10] and Theorem 7.2.5 [11]. Indeed, for any $j = 1,\, 2$, from (26) and (2) we can write

$$\widehat{E}_j^{r,N}(t) = \frac{\alpha_j^N}{r^H\, L^{1/2}(r)}\, \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \Big(\int_0^{r\,t} U_j^{(n)}(u)\, du - \frac{\mu^{\rm on}}{\mu^{\rm on}+\mu^{\rm off}}\, r\, t\Big)$$

and deduce the existence of the limit $\widehat{\widehat{E}}^{\,r} = \mathcal{D} - \lim_{N\to+\infty} \widehat{E}^{r,N}$, which has paths in $\mathcal{C}^2$, and the existence of the limit

$$\mathcal{D} - \lim_{r\to+\infty} \widehat{\widehat{E}}^{\,r} = B^H\,, \tag{34}$$

B^H being a two-dimensional fBm process with associated data ($x = 0,\, H,\, \theta = 0,\, diag(\alpha)^2\, \sigma^{2,\lim}$), which is condition (a) in Proposition 10.

Combining (32), **(HT)** and the *continuous mapping theorem*, according to the above limit $\widehat{\widehat{E}}^{\,r}$, we deduce the existence of $\widehat{\widehat{X}}^{\,r} = \mathcal{D} - \lim_{N\to+\infty} \widehat{X}^{r,N}$, which verifies that

$$\widehat{\widehat{X}}^{\,r}(t) = M\widehat{\widehat{E}}^{\,r}(t) + \frac{r^{1-H}}{L^{1/2}(r)} C\widehat{\gamma}^r t, \tag{35}$$

implying the continuity of the paths of $\widehat{\widehat{X}}^{\,r}$ and (iv).

Secondly, since hypothesis (b) is accomplished by Lemma 8 and Remark 3, we can apply Proposition 10 to obtain that there exists the following limit:

$$\mathcal{D} - \lim_{N\to+\infty} \big(\widehat{W}^{r,N},\, \widehat{X}^{r,N},\, \widehat{V}^{r,N}\big) = (\widehat{\widehat{W}}^{\,r},\, \widehat{\widehat{X}}^{\,r},\, \widehat{\widehat{V}}^{\,r}),$$

and that the limit satisfies conditions (i), (ii) and (iv) of Definition 2, that is, $(\widehat{\widehat{W}}^{\,r},\, \widehat{\widehat{V}}^{\,r})$ is a solution of the Skorokhod Problem associated to $\widehat{\widehat{X}}^{\,r}$ on the convex polyhedron $S(\Upsilon)$ with associated matrix of directions of reflection R.

The repeated application of Proposition 10, in this case to the sequence $\{(\widehat{\widehat{W}}^{\,r},\, \widehat{\widehat{X}}^{\,r},\, \widehat{\widehat{V}}^{\,r})\}_r$, enables us to complete the proof. Indeed, from (35), (34), **(HT)** and the *continuous mapping theorem*, we can ensure the existence of $\mathcal{D} - \lim_{r\to+\infty} \widehat{\widehat{X}}^{\,r} = X$, with $X(t) = MB^H(t) + C\gamma t$, which is a two-dimensional fBm process with associated data $\big(x = 0,\, H,\, \theta = C\gamma,\, \Gamma\big)$, where $\Gamma = \sigma^{2,\lim} M diag(\alpha)^2 M^T$ is given by (33). Moreover, by Lemma 8 we can assert that (b) in Proposition 10 holds, by Remark 3, and from Proposition 10 it follows the existence of

$$\mathcal{D} - \lim_{r\to+\infty} \big(\widehat{\widehat{W}}^{\,r},\, \widehat{\widehat{X}}^{\,r},\, \widehat{\widehat{V}}^{\,r}\big) = (W,\, X,\, V)\,,$$

where the triplet $(W,\, X,\, V)$ satisfies conditions (i)-(iv) of the Definition 2.

Thus, $W = X + R\,V$ is a two-dimensional rfBm on the convex polyhedron $S(\Upsilon)$ with associated data $(x = 0,\, H,\, \theta = C\gamma,\, \Gamma,\, R)$, which is our claim. □

5 Appendix: The invariance principle

Kang and Williams prove in Theorem 4.3 [9] an *Invariance Principle* for Semimartingale reflecting Brownian motions (SRBMs) living in the closure of a domain with piecewise smooth boundaries. This provides sufficient conditions for a process that satisfies the definition of a SRBM except for small random perturbations in the defining conditions, to be close in distribution to an SRBM. The version of this result stated in [5] gives sufficient conditions for validating approximations involving rfBm processes on a convex polyhedron with a constant reflection vector field on each face, in such a way the approximating processes live in a sequence of convex polyhedra. For the convenience of the reader, we reproduce here the invariance principle in [5] without proof, thus making our exposition self-contained. Let $\{S^n\}_n$ denote a sequence of convex polyhedra that converges to the convex polyhedron S. The invariance principle requires the following hypothesis, which is a version of Assumption 4.1 [9]:

Assumption (h) *For each positive integer n, there are processes W^n, X^n having paths in $\mathcal{D}^d$ and V^n having paths in $\mathcal{C}^d$ defined on some probability space $(\Omega^n,\, \mathcal{F}^n,\, P^n)$ such that $X^n(0) \in S^n$ and:*

(i) *P^n-a.s., $W^n(t) \in S^n$ for all $t \geq 0$,*

(ii) *P^n-a.s., $W^n(t) = X^n(t) + R^n\, V^n(t)$ for all $t \geq 0$,*

(iii) P^n-*a.s., for each* $i = 1, \ldots, d$, $V_i^n(0) = 0$, V_i^n *is nondecreasing and* $V_i^n(t) = \int_0^t 1_{\{W^n(s) \in F_i^n\}} \, dV_i^n(s)$,

(iv) $\{X^n\}_n$ *is* $\mathcal{C}-$*tight.*

Proposition 10 (The Invariance Principle) *Suppose that* Assumption (h) *and assumptions* (A1)-(A5) [9] *hold, and also that* $\lim_{n \to +\infty} R^n = R$. *Then, the sequence* $\{(W^n, X^n, V^n)\}_n$ *is* $\mathcal{C}-$*tight and any (weak) limit point of this sequence is of the form* (W, X, V) *where* W, X *and* V *are continuous* $d-$*dimensional processes defined on some probability space* $(\Omega, \mathcal{F}, P)$, *such that conditions* (i), (ii) *and* (iv) *of Definition 2 hold,* $W(0) = X(0)$ *and* $V(0) = 0$, *that is,* (W, V) *is a solution of the Skorokhod Problem associated to* X *on the convex polyhedron* S *with associated matrix of directions of reflection* R. *If, in addition,*

(a) $\{X^n\}_n$ *converges in distribution to a* $d-$*dimensional fBm process with associated data* (x, H, θ, Γ), *and*

(b) *the Skorokhod Problem associated to* X *on the convex polyhedron* S *with associated matrix of directions of reflection* R *has a unique strong solution,*

then W *is a rfBm process on* S *with associated data* $(x, H, \theta, \Gamma, R)$.

Acknowledgements: The author is supported by Ministerio de Educación, Cultura y Deporte de España and ERDF (European Regional Development Found "A way to build Europe"), project ref. MTM2012-33937.

References:

[1] R. Delgado, A reflected fBm limit for fluid models with ON/OFF sources under heavy-traffic, *Stochastic Processes and Their Applications* **117**, 2007, pp. 188-201.

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

[3] R. Delgado, Heavy-traffic limit for a feed-forward fluid model with heterogeneous heavy-tailed On/Off sources, *Queueing Systems* **74(1)**, 2013, pp. 41-63.

[4] R. Delgado, State space collapse and heavy-traffic for a packet-switched network with On/Off sources and a fair bandwidth sharing policy, to appear in *Telecommunication Systems* (2015). http://dx.doi.org/10.1007/s11235-015-0086-6

[5] R. Delgado, A heavy-traffic limit of cascade fluid networks with heavy-tailed On/Off sources, submitted for publication (2015).

[6] R. Delgado, E. Morozov, Stability analysis of cascade networks via fluid models, *Performance Evaluation* **82**, 2014, pp. 39-54.

[7] R. Delgado, E. Morozov, Stability analysis of some networks with interacting servers. ASMTA 2014, LNCS 8499, 2014, pp. 1-15.

[8] M. Harrison, Heavy-traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review pollcies, *Ann. Appl. Probab.* **8(3)**, 1998, pp. 822-848.

[9] W. N. Kang, R. J. Williams, An invariance principle for semimartingale reflecting Brownian motions in domains with piecewise smooth boundaries, *Ann. Appl. Probab.* **17**, 2007, pp. 741-779.

[10] M. S. Taqqu, W. Willinger, R. Sherman, Proof of a fundamental result in self-similar traffic modeling, *Comput. Commun. Rev.* **27**, 1997, pp. 5-23.

[11] W. Whitt, Weak convergence theorems for priority queues: preemptive resume discipline, *Journal of Applied Probability* **8**, 1971, pp. 74-94.