# The reflected fBm process on a convex polyhedron as limit for the $X-$model with On/Off sources under heavy-traffic

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

June 25, 2017

*Abstract:* We consider a $X-$model with fluid queues that can be approximated under heavy-traffic conditions by a two-dimensional reflected fractional Brownian motion (rfBm). Specifically, we prove a heavy-traffic limit theorem for this single-server two-station model in which each server helps the other when free, with feedback allowed and a non-deterministic arrival process generated by a large enough number of heavy-tailed On/Off sources, say $N$. Scaling conveniently by a factor $r$ and by $N$, and letting $N$ and $r$ approach infinity (in this order), we prove that the scaled (total) workload process converges under heavy-traffic to a rfBm process on a convex polyhedron.

*Key-Words:* reflected fractional Brownian motion; convex polyhedron; On/Off sources; heavy-traffic limit; Skorokhod problem; $X-$model.

## 1 Introduction

It is well known from the seminal paper of Taqqu et al. [11] (see Theorem 1 there) that the aggregate cumulative arrival to a network, generated by the superposition of many (say $N$) heavy-tailed On/Off sources, conveniently scaled by a factor $r$ and by $N$, converges when $N$ and $r$ go to infinity, to a fractional Brownian motion (fBm) process. These limits should be treated with care, because if they are taken in the reverse order, the convergence is to an $\alpha-$stable Lévy process rather than a fBm. As fBm is a self-similar process with long-range dependent increment, this makes the "heavy-tailed On/Off sources" be a very reasonable assumption for modeling of broadband network traffic, which is known to show long-range dependence and self-similar patterns.

Different works so far have addressed the issue of transfer of the convergence of the aggregate cumulative arrival process to workload, in a heavy-traffic environment, the fBm reflected appropriately being the limit process. Indeed, Debicki and Mandjes [1] prove that the scaled workload process converges to the fBm, reflected appropriately to be non-negative, for a fluid model with only one station. Generalization to the multi-station setting has been given in [2], where the limit rfBm process lives in the positive orthant. It is also possible to find more sophisticated queueing models with heavy-tailed On/Off sources for which the scaled workload pro-

cess converges under heavy-traffic to a rfBm process living in a convex polyhedron different from the positive orthant. For instance, [3] considers a flow-level model for packet-switched telecommunications networks handling elastic flows with concurrent occupancy of resources, in which digital objects are transferred at a rate determined by capacity allocation on each route and the capacity of each node is dynamically allocated to the routes passing by it through a weighted proportional fair sharing policy. Looking for models of different nature but with the same type of limit process, a $N-$model has been considered in [4]. In this model, which is a single-server two-station polling system with fluid queues, server 2 is flexible in the sense that processes its own fluid and when free, it helps server 1, but the reciprocal is not permissible. Generalization of this model to the case in which feedback is allowed, has been studied in [5].

Polling systems are a special class of queueing systems where a single server visits a set of queues in some order, and 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. For a more detailed explanation, see the introduction of [4].

In this paper we consider a different polling system, consisting of a network composed of two single-server workstations that process continuous fluid, with an infinite-capacity buffer each one, and feedback allowed, the $X-$model portrayed schematically in Figure 1. In this model, each server can help the other when free, that is, servers are cross-trained. This kind of models have been considered in several works. For example, Perry and Whitt [9] study an ordinary differential equation which is the deterministic fluid approximation for an overloaded $X$ call-center model with two customer classes and two service pools, which is proved in [10] by the same authors to arise as the many-server heavy-traffic fluid limit of a properly scaled sequence of overloaded Markovian $X$ models under the fixed-queue-ratio-with-thresholds (FQR-T) control.

In the $X-$model 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. Whenever one station becomes empty, say station 2, while there is (class-1) fluid awaiting at the other station, 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. The situation is perfectly symmetrical between the two servers, that is, previous explanation also applies swapping server roles. We assume that no server can be idle while there is fluid awaiting for processing in any of the stations (nonidling policy). 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 either server) a proportion $p_{22}$ of class-2 fluid needs to be reprocessed by server 2, and the rest goes outside. For this process we prove a heavy-traffic limit, which shows that the scaled total workload process converges to a rfBm process on a convex polyhedron (see Figure 2).

The organization of the paper is as follows. In Section 2 we set up notation and preliminaries. Section 3 is devoted to the introduction of the $X-$model, the processes used to measure its performance, the convex polyhedron, the heavy-traffic

condition and scaling, while the heavy-traffic limit is proved in Section 4. In the Appendix we recall an Invariant Principle proved in [4], which is a key ingredient in the proof of the limit result.

## 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. The identity matrix (of any dimension) is denoted by $I$. 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 $\mathscr{C}^d$ be the space of continuous functions $\omega$ from $[0, +\infty)$ to $\mathbb{R}^d$, with the topology of the uniform convergence on compact time intervals, and $\mathscr{D}^d$ the space of continuous on the right with limits on the left functions, endowed with the usual Skorokhod $\mathscr{J}_1-$topology. All stochastic processes in this paper will be assumed to have paths in $\mathscr{D}^d$, 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 $\mathscr{D}^d$ form a tight sequence (that is, the sequence of induced measures is weakly relatively compact in the space of probability measures on $\mathscr{D}^d$).

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

**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:

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

$v^1, \ldots, v^d \in \mathbb{R}^d$, $\Upsilon$ being the associated $d \times d$ matrix whose row vectors are $v^1, \ldots, v^d$. Its boundary is $\partial S = \cup_{\ell=1}^d F_\ell$, with $F_\ell = \{x \in S : \langle v^\ell, x\rangle = 0\}$ the boundary faces.

It is assumed that the interior of $S$ is not empty and the set $\{v^1, \ldots, v^d\}$ is minimal. Associated to $S$ 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 $S$ 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 $S$ associated with data $(x, H, \theta, \Gamma, R)$, where $x \in S, H \in (0, 1), \theta \in \mathbb{R}^d$ and $\Gamma$ 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 $W(t) \in S$ 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

$$\begin{aligned} &Cov\big(X(t), X(s)\big) = \\ &E\Big(\big(X(t) - (x + \theta t)\big)\big(X(s) - (x + \theta s)\big)^T\Big) = \\ &\Gamma_H(s,t)\Gamma \quad \text{if } t, s \geq 0, \quad \text{with} \\ &\Gamma_H(s,t) = \frac{1}{2}(t^{2H} + s^{2H} - |t-s|^{2H}), \end{aligned}$$

(iv) $V$ has continuous and non-decreasing paths, and for $\ell = 1, \ldots, d$ a.s., $V_\ell(0) = 0$ and $V_\ell(t) = \int\limits_0^t 1_{\{W(s) \in \mathrm{F}_\ell\}} \, dV_\ell(s)$ for all $t \geq 0$ (only increases if $W$ is on the boundary face $F_\ell$).
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 $S$ with associated matrix of directions of reflection $R$.

**Remark 1.** 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 nonpositive and the following condition (the *generalized Harrison-Reiman condition*), holds (see Remark 1 [3] for details):

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

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 $u^\ell$, gives the direction of the reflection at the boundary face $F_\ell$, and $v^\ell$ 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 $X-$model

The basic features of the $X-$model have been explained in the introduction (see scheme in Figure 1). Now we go deeper into it. For facilitate access to the topics, it is rendered as self-contained as possible. Assume $0 \leq p_{11}, p_{22} < 1$, and introduce

$$P \stackrel{\text{def}}{=} \begin{pmatrix} p_{11} & 0 \\ & p_{22} \end{pmatrix}, \quad \text{and}$$
$$Q \stackrel{\text{def}}{=} (I - P^T)^{-1} = \begin{pmatrix} q_{11} & 0 \\ 0 & q_{22} \end{pmatrix}$$

with $q_{jj} = \frac{1}{1-p_{jj}} \geq 1$, $j = 1, 2$, which is well defined since matrix $P$ has spectral radius less than one. $P$ is the (sub-stochastic) *"flow"* or *"routing"* matrix of the fluid model. (If $p_{11} = p_{22} = 0$ we recover the $X-$model without feedback.)

As in [2]-[4], we assume that for each station there are $N$ i.i.d. external sources sending fluid to it, and that each source can be On or Off. 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 $f^{\text{on}}$ and $f^{\text{off}}$ 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 $\mu^{\text{on}} \stackrel{\text{def}}{=} \int_0^{+\infty} u f^{\text{on}}(u) du$ and $\mu^{\text{off}} \stackrel{\text{def}}{=} \int_0^{+\infty} u f^{\text{off}}(u) du$. Assume that as $x \to +\infty$,

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

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 $\mu^{\text{on}}$ and $\mu^{\text{off}}$ are finite while variances are not.

In what follows, we use subindex $j$ to denote the quantities related to class-$j$ fluid, $j = 1, 2$, subindex 12 specifically to that quantities related to processing of class-1 fluid by server 2, and 21 for processing of class-2 fluid by server 1. Define the cumulative external class-$j$ fluid arrived up to time $t$ (by the $N$ sources) at station $j$ by:

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

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 $E^N = \{E^N(t) = \big(E_1^N(t), E_2^N(t)\big)^T, t \geq 0\}$, 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, $\lambda^N = Q\tilde{\alpha}^N$. Note that $\lambda_j^N$ is 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.

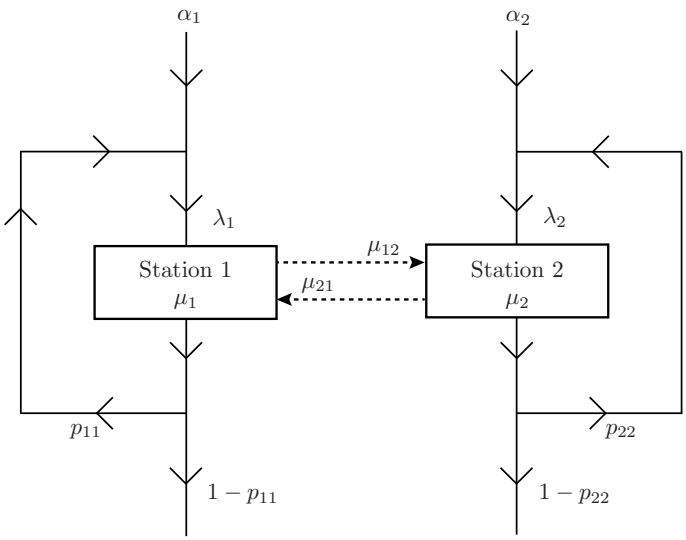


Figure 1: $X-$model with feedback

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. 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 class-2 fluid at constant rate $\mu_{21}^{r,N}>0$, not necessarily equal to $\mu_1^{r,N}$ nor to $\mu_2^{r,N}$, if server 1 devoted all time to this fluid class. Symmetrically for server 2, that processes class-2 fluid at constant rate $\mu_2^{r,N}>0$ and class-1 fluid at a constant rate $\mu_{12}^{r,N}>0$. We assume that $\lim_{N\to+\infty}(\mu_1^{r,N},\mu_2^{r,N},\mu_{12}^{r,N},\mu_{21}^{r,N})$ exists, is $>0$ and does not depend on $r$; we denote it by $(\mu_1,\mu_2,\mu_{12},\mu_{21})$. Let us introduce the *fluid traffic intensity* $\rho^{r,N}=(\rho_1^{r,N},\rho_2^{r,N})^T$ by

$$\rho_1^{r,N}\stackrel{\text{def}}{=}\frac{\lambda_2^N-\mu_2^{r,N}}{\mu_{21}^{r,N}}+\frac{\lambda_1^N}{\mu_1^{r,N}},$$
$$\rho_2^{r,N}\stackrel{\text{def}}{=}\frac{\lambda_1^N-\mu_1^{r,N}}{\mu_{12}^{r,N}}+\frac{\lambda_2^N}{\mu_2^{r,N}}. \qquad (3)$$

Stability of the $X-$model has been considered in [6], where it is proved that traffic intensity $<1$ is a sufficient condition for the stability under certain conditions.

### 3.1 Performance processes

To measure the performance of our model we introduce some processes. Definition of *workload* $W^{r,N}=(W_1^{r,N},W_2^{r,N})^T$ is adopted from [4] and agrees with the one given in [7]: for $j=1,2$, the *total workload* $W_j^{r,N}(t)$ is defined as the total time of service that would be required to complete processing of the total amount of both classes of fluid in the system at time $t$, if it were to be performed by server $j$ without help from the other server. We assume $W^{r,N}(0)=0$. The *cumulative idle-time* $Y^{r,N}=(Y_1^{r,N},Y_2^{r,N})^T$ is defined by: $Y_j^{r,N}(t)$ is the cumulative amount of time that server $j$ has been idle in $[0,t]$:

$$Y_1^{r,N}(t)=Y_2^{r,N}(t)=\int_0^t 1_{\{W^{r,N}(s)=0\}}\,ds \qquad (4)$$

(note that the idle-time for each server is the same, and corresponds to the time that there is no workload, nor for server 1, no for server 2). The *total service time* $T^{r,N}=(T_1^{r,N},T_2^{r,N},T_{12}^{r,N},T_{21}^{r,N})^T$ is defined by: $T_j^{r,N}(t)$ is the total service time devoted to class-$j$ fluid (by server $j$) in the interval $[0,t]$, $j=1,2$, while $T_{12}^{r,N}(t)$ and $T_{21}^{r,N}(t)$ are respectively the total service time devoted to class-1 by

server 2, and to class-2 by server 1, in $[0,t]$. 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: $A_j^{r,N}(t)$ is the total fluid arrived at station $j$, as class-$j$ fluid, up to time $t$, including both external input and feedback flow. Note that we do not include fluid transferred from the other station when the floodgate is open. $D_j^{r,N}(t)$ is the total amount of class-$j$ fluid departing, either by leaving the network or not, from either station, up to time $t$. Assume $A^{r,N}(0)=D^{r,N}(0)=0$.

As it is done in [4], we introduce the following notation: $\widetilde{W}_j^{r,N}$ is defined as the portion of the workload $W_j^{r,N}$ exclusively due to class-$j$ fluid, that is,

$$\widetilde{W}_1^{r,N}=\frac{A_1^N}{\mu_1^{r,N}}-\big(T_1^{r,N}+\frac{\mu_{12}^{r,N}}{\mu_1^{r,N}}T_{12}^{r,N}\big), \tag{5}$$

$$\widetilde{W}_2^{r,N}=\frac{A_2^N}{\mu_2^{r,N}}-\big(T_2^{r,N}+\frac{\mu_{21}^{r,N}}{\mu_2^{r,N}}T_{21}^{r,N}\big). \tag{6}$$

The interpretation of (16) 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. Interpretation of (17) is analogous by symmetry. These processes are related by means of the equalities:

$$\begin{aligned}
W_1^{r,N}&=\widetilde{W}_1^{r,N}+\frac{\mu_2^{r,N}}{\mu_{21}^{r,N}}\widetilde{W}_2^{r,N},\\
W_2^{r,N}&=\widetilde{W}_2^{r,N}+\frac{\mu_1^{r,N}}{\mu_{12}^{r,N}}\widetilde{W}_1^{r,N},
\end{aligned} \tag{7}$$

$$\begin{aligned}
t&=Y_1^{r,N}(t)+T_1^{r,N}(t)+T_{21}^{r,N}(t),\\
t&=Y_2^{r,N}(t)+T_2^{r,N}(t)+T_{12}^{r,N}(t),
\end{aligned} \tag{8}$$

$$\begin{aligned}
T_1^{r,N}(t)&=\int_0^t 1_{\{\widetilde{W}_1^{r,N}(s)>0\}}ds,\\
T_2^{r,N}(t)&=\int_0^t 1_{\{\widetilde{W}_2^{r,N}(s)>0\}}ds,
\end{aligned} \tag{9}$$

$$\begin{aligned}
T_{12}^{r,N}(t)&=\int_0^t 1_{\{\widetilde{W}_1^{r,N}(s)>0,\widetilde{W}_2^{r,N}(s)=0\}}ds,\\
T_{21}^{r,N}(t)&=\int_0^t 1_{\{\widetilde{W}_1^{r,N}(s)=0,\widetilde{W}_2^{r,N}(s)>0\}}ds.
\end{aligned} \tag{10}$$

With respect to (7), total workload $W_1^{r,N}$ is equal to $\widetilde{W}_1^{r,N}$ by adding $\widetilde{W}_2^{r,N}(t)$ conveniently rescaled representing the amount of time required by server 1 to process all the class-2 fluid at buffer 2 at time $t$, and analogous for $W_2^{r,N}$. On the other hand, (8) expresses that the length of the interval $[0,t]$ can be split into idle time $Y_j^{r,N}(t)$, and working time $(T_1^{r,N}(t)+T_{21}^{r,N}(t)$ for $j=1$, while $T_2^{r,N}(t)+T_{12}^{r,N}(t)$ for $j=2$). Formulae (9) and (10) are self-explanatory, taking into account the definitions of the involved processes.

We will use notation $r_1^{r,N}\stackrel{\text{def}}{=}\frac{\mu_1^{r,N}}{\mu_{21}^{r,N}}$, $r_2^{r,N}\stackrel{\text{def}}{=}\frac{\mu_2^{r,N}}{\mu_{12}^{r,N}}$ and $r_j=\lim_{N\to+\infty}r_j^{r,N}$, assumed to be independent of $r$ and positive, for $j=1,2$. From now on we will assume $r_1r_2\neq 1$, which implies that $r_1^{r,N}r_2^{r,N}\neq 1$ for all $r$ and for $N$ big enough. Indeed, to fix ideas and avoid repetition, we assume from now on that $r_1r_2>1$ (the other case is similar). Finally, we introduce the ancillary process $V^{r,N}$ as the function of $Y^{r,N}$ and $T^{r,N}$ by

$$V_1^{r,N}\stackrel{\text{def}}{=}\big(1+\frac{\mu_2^{r,N}}{\mu_{21}^{r,N}}\big)Y_1^{r,N}+\frac{\mu_{12}^{r,N}}{\mu_1^{r,N}}(r_1^{r,N}r_2^{r,N}-1)T_{12}^{r,N}, \tag{11}$$

$$V_2^{r,N}\stackrel{\text{def}}{=}\big(1+\frac{\mu_1^{r,N}}{\mu_{12}^{r,N}}\big)Y_2^{r,N}+\frac{\mu_{21}^{r,N}}{\mu_2^{r,N}}(r_1^{r,N}r_2^{r,N}-1)T_{21}^{r,N}. \tag{12}$$

We are interested in express workload process in terms of processes $E^{r,N}$ and $V^{r,N}$. From this relation, proved in Lemma 2 below, we deduce in Lemma 3 the Skorokhod decomposition. 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 1.** Processes $A^{r,N}, E^{r,N}$ and $W^{r,N}$ are related by means of:

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

where $M^{r,N} = \begin{pmatrix} \frac{1}{\mu_1^{r,N}} & \frac{1}{\mu_{21}^{r,N}} \\ \frac{1}{\mu_{12}^{r,N}} & \frac{1}{\mu_2^{r,N}} \end{pmatrix}$.

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

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

since $P^T D^{r,N}$ is the total amount of fluid arriving from *feedback*. On the other hand, according to definition of process $D^{r,N}$, $D_j^{r,N} = A_j^{r,N} - \mu_j^{r,N}\widetilde{W}_j^{r,N}$, which in matricial form can be expressed as

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

since by (7),

$$\widetilde{W}_1^{r,N} = \frac{1}{1-r_1^{r,N}r_2^{r,N}}(W_1^{r,N} - \frac{\mu_2^{r,N}}{\mu_{21}^{r,N}}W_2^{r,N}), \tag{16}$$

$$\widetilde{W}_2^{r,N} = \frac{1}{1-r_1^{r,N}r_2^{r,N}}(-\frac{\mu_1^{r,N}}{\mu_{12}^{r,N}}W_1^{r,N} + W_2^{r,N}), \tag{17}$$

and using that

$$(M^{r,N})^{-1} = \frac{1}{1-r_1^{r,N}r_2^{r,N}}\begin{pmatrix} \mu_1^{r,N} & -r_1^{r,N}\mu_2^{r,N} \\ -r_2^{r,N}\mu_1^{r,N} & \mu_2^{r,N} \end{pmatrix}.$$

Replacing (15) into (14) yields $A^{r,N} = E^{r,N} + P^T A^{r,N} - P^T(M^{r,N})^{-1}W^{r,N}$, which establishes (13) on account of the definition of matrix $Q$. ■

**Lemma 2.** Processes $W^{r,N}, E^{r,N}$ and $V^{r,N}$ verify the following relation:

$$W^{r,N}(t) = M^{r,N}E^{r,N}(t) - R^{r,N}\delta^{r,N}t + R^{r,N}V^{r,N}(t) \tag{18}$$

where

$$R^{r,N} = M^{r,N}Q^{-1}(M^{r,N})^{-1} \quad \text{and}$$
$$\delta^{r,N} = (1 + \frac{\mu_2^{r,N}}{\mu_{21}^{r,N}}, 1 + \frac{\mu_1^{r,N}}{\mu_{12}^{r,N}})^T.$$

*Proof:* From equations (16)-(10), the fact that $Y_1^{r,N} = Y_2^{r,N}$ and the definition of process $V^{r,N}$ given by (11) and (12), we can rewrite $W_1^{r,N}$ and $W_2^{r,N}$ as

$$W_1^{r,N}(t) = \frac{A_1^{r,N}(t)}{\mu_1^{r,N}} + \frac{A_2^{r,N}(t)}{\mu_{21}^{r,N}} - (1 + \frac{\mu_2^{r,N}}{\mu_{21}^{r,N}})t + V_1^{r,N}(t),$$
$$W_2^{r,N}(t) = \frac{A_1^{r,N}(t)}{\mu_{12}^{r,N}} + \frac{A_2^{r,N}(t)}{\mu_2^{r,N}} - (1 + \frac{\mu_1^{r,N}}{\mu_{12}^{r,N}})t + V_2^{r,N}(t).$$

In matricial form,

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

Finally, we get (18) by replacing (13) into (19), taking into account that

$$I + M^{r,N}QP^T(M^{r,N})^{-1} = M^{r,N}(I + QP^T)(M^{r,N})^{-1}$$
$$= M^{r,N}Q(M^{r,N})^{-1} = (R^{r,N})^{-1}. \quad ■$$

**Remark 2.** Note that clearly we obtain the existence of the following limits:

$$M = \lim_{N\to+\infty} M^{r,N} = \begin{pmatrix} \frac{1}{\mu_1} & \frac{1}{\mu_{21}} \\ \frac{1}{\mu_{12}} & \frac{1}{\mu_2} \end{pmatrix} \quad \text{and}$$
$$R = \lim_{N\to+\infty} R^{r,N} = MQ^{-1}M^{-1}$$
$$= \frac{1}{1-r_1r_2}\begin{pmatrix} \frac{1}{q_{11}} - \frac{r_1r_2}{q_{22}} & \frac{\mu_2}{\mu_{21}}(\frac{1}{q_{22}} - \frac{1}{q_{11}}) \\ \frac{\mu_1}{\mu_{12}}(\frac{1}{q_{11}} - \frac{1}{q_{22}}) & \frac{1}{q_{22}} - \frac{r_1r_2}{q_{11}} \end{pmatrix}. \tag{20}$$

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

Let us define (for $r_1r_2 > 1$) $S \stackrel{\text{def}}{=} \{(x,y) \in \mathbb{R}^2 : \frac{\mu_{21}}{\mu_2}x \le y \le \frac{\mu_1}{\mu_{12}}x\}$, which is the convex polyhedron in Figure 2 with boundary faces $F_1 = \{(x,y) \in S : y = \frac{\mu_1}{\mu_{12}}x\}$ and $F_2 = \{(x,y) \in S : y = \frac{\mu_{21}}{\mu_2}x\}$, and associated matrix

$$\Upsilon = \begin{pmatrix} q_{22} & -q_{22}\frac{\mu_{12}}{\mu_1} \\ -q_{11}\frac{\mu_{21}}{\mu_2} & q_{11} \end{pmatrix}.$$

We also introduce a sequence of convex polyhedra indexed by $(r,N)$ by replacing $\mu_1, \mu_2, \mu_{12}$ and

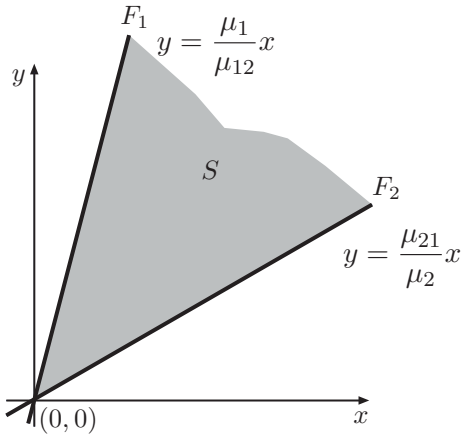


Figure 2: The convex polyhedron for the $X-$model (case $r_1\, r_2 > 1$).

$\mu_{21}$ by $\mu_1^{r,N}, \mu_2^{r,N}, \mu_{12}^{r,N}$ and $\mu_{21}^{r,N}$ respectively, and we will use the superscript $r,N$.

We wish to stress that the key technical difficulty of our main result (Theorem 1) stems from the fact that the faces of the convex polyhedron associated to the $(r,N)$ model actually do depend on $r$ and $N$.

### 3.3 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 [11] (see also [2], [4]). 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 $\ell$, defined by $\ell \stackrel{\rm def}{=} \lim_{t\to+\infty} \frac{L^{\rm on}(t)}{L^{\rm off}(t)}\, t^{\beta^{\rm off}-\beta^{\rm on}}$, which exists although it could be infinite. If $0<\ell<+\infty$ (implying $\beta^{\rm on}=\beta^{\rm off}$ and $\ell = \lim_{t\to+\infty} \dfrac{L^{\rm on}(t)}{L^{\rm off}(t)}$), set $\beta \stackrel{\rm def}{=} \beta^{\rm on} = \beta^{\rm off}$, $L \stackrel{\rm def}{=} L^{\rm off}$ and $\sigma^{2,\lim} \stackrel{\rm def}{=} \dfrac{2\left((\mu^{\rm off})^2 a^{\rm on}\ell + (\mu^{\rm on})^2 a^{\rm off}\right)}{\left(\mu^{\rm on}+\mu^{\rm off}\right)^3 \Gamma(4-\beta)}$. If, on the other hand, $\ell=+\infty$ ($\beta^{\rm off} > \beta^{\rm on}$), set $L \stackrel{\rm def}{=} L^{\rm on}$, $\beta \stackrel{\rm def}{=} \beta^{\rm on}$ and $\sigma^{2,\lim} \stackrel{\rm def}{=} \dfrac{2(\mu^{\rm off})^2 a^{\rm on}}{\left(\mu^{\rm on}+\mu^{\rm off}\right)^3 \Gamma(4-\beta)}$. If $\ell=0$ ($\beta^{\rm off} < \beta^{\rm on}$), set $L \stackrel{\rm def}{=} L^{\rm off}$, $\beta \stackrel{\rm def}{=} \beta^{\rm off}$ and $\sigma^{2,\lim} \stackrel{\rm def}{=} \dfrac{2(\mu^{\rm on})^2 a^{\rm off}}{\left(\mu^{\rm on}+\mu^{\rm off}\right)^3 \Gamma(4-\beta)}$. In either case, $\beta \in (1,2)$. Let us define

$$H \stackrel{\rm def}{=} \frac{3-\beta}{2} \quad \left(\in (\frac{1}{2},1)\right). \tag{21}$$

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: there exist $(\widehat{\gamma}^r)_r$ and $\gamma$, in $\mathbb{R}^2$, such that

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

This type of condition has been introduced previously in [4], where it is justified.

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

$$\widehat{W}_j^{r,N}(t) \stackrel{\rm def}{=} \sqrt{N}\, \frac{W_j^{r,N}(rt)}{r^H L^{1/2}(r)}, \tag{22}$$

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

$$\widehat{E}_j^{r,N}(t) \stackrel{\rm def}{=} \sqrt{N}\, \frac{E_j^N(rt) - \tilde{\alpha}_j^N rt}{r^H L^{1/2}(r)} \quad (j=1,2). \tag{23}$$

From (11) and (12) we obtain

$$\widehat{V}_1^{r,N} \stackrel{\text{def}}{=} \big(1+\frac{\mu_2^{r,N}}{\mu_{21}^{r,N}}\big)\widehat{Y}_1^{r,N} + \frac{\mu_{12}^{r,N}}{\mu_1^{r,N}}(r_1^{r,N}r_2^{r,N}-1)\widehat{T}_{12}^{r,N},$$
$$\widehat{V}_2^{r,N} \stackrel{\text{def}}{=} \big(1+\frac{\mu_1^{r,N}}{\mu_{12}^{r,N}}\big)\widehat{Y}_2^{r,N} + \frac{\mu_{21}^{r,N}}{\mu_2^{r,N}}(r_1^{r,N}r_2^{r,N}-1)\widehat{T}_{21}^{r,N}.$$

From (4) it follows that

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

and from (10),

$$\widehat{T}_{12}^{r,N}(t) = \frac{\sqrt{N}r^{1-H}}{L^{1/2}(r)}\int_0^t 1_{\{\widehat{\widehat{W}}_1^{r,N}(s)>0,\widehat{\widehat{W}}_2^{r,N}(s)=0\}}ds,$$
$$\widehat{T}_{21}^{r,N}(t) = \frac{\sqrt{N}r^{1-H}}{L^{1/2}(r)}\int_0^t 1_{\{\widehat{\widehat{W}}_1^{r,N}(s)=0,\widehat{\widehat{W}}_2^{r,N}(s)>0\}}ds.$$

The following lemma provides a Skorokhod decomposition that will prove extremely useful in the proof of Theorem 1 in the next section.

**Lemma 3.** (*Skorokhod decomposition*) The scaled processes $\widehat{W}^{r,N}$, $\widehat{E}^{r,N}$ and $\widehat{V}^{r,N}$ are related by means of:

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

with

$$\widehat{X}^{r,N}(t) = M^{r,N}\widehat{E}^{r,N}(t) + \frac{\sqrt{N}r^{1-H}}{L^{1/2}(r)}R^{r,N}(\rho^{r,N}-e)t. \tag{24}$$

*Proof:* From (22), (18) and (23) we obtain

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

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

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

which is our claim, due to the fact that

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

**Lemma 4.** The column vectors of matrix $R$ given by (20) are linearly independent. Moreover, as $r_1 r_2 > 1$, matrix $\Psi = \Upsilon R = \begin{pmatrix} 1 & -\frac{\mu_{12}}{\mu_1} \\ -\frac{\mu_{21}}{\mu_2} & 1 \end{pmatrix}$ verifies that the entries outside the main diagonal are nonpositive and also condition **(HR)** (see Remark 1). (Analogously for the $(r,N)-$model, for any $r$ and $N$ big enough.)

The proof of this lemma is straightforward and omitted.

# 4 The heavy-traffic limit

Our goal 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 $S$, when $N$ first and then $r$, tend to infinity in this order, under heavy-traffic. For similar heavy-traffic limits, see Theorem 1 [2] and Theorem 1 [4].

**Theorem 1.** (*heavy-traffic limit*)

Under the heavy-traffic condition **(HT)** (and $r_1 r_2 > 1$), the following limits exist in $\mathscr{C}^2$:

$$\widehat{\widehat{W}}^{r} = \mathscr{D}-lim_{N\to+\infty}\widehat{W}^{r,N}, \quad W = \mathscr{D}-lim_{r\to+\infty}\widehat{\widehat{W}}^{r},$$

and $W$ is a two-dimensional rfBm process on the convex polyhedron $S$, and associated data ($x =$

$0, H, \theta = R\gamma, \Gamma, R)$, where $H \in (\frac{1}{2}, 1)$ is defined by (21), $\gamma \in \mathbb{R}^2$ is given by condition **(HT)**,

$$\Gamma = \sigma^{2,\lim} \begin{pmatrix} \frac{\alpha_1^2}{\mu_1^2} + \frac{\alpha_2^2}{\mu_{21}^2} & \frac{\alpha_1^2}{\mu_1 \mu_{12}} + \frac{\alpha_2^2}{\mu_2 \mu_{21}} \\ \frac{\alpha_1^2}{\mu_1 \mu_{12}} + \frac{\alpha_2^2}{\mu_2 \mu_{21}} & \frac{\alpha_1^2}{\mu_{12}^2} + \frac{\alpha_2^2}{\mu_2^2} \end{pmatrix} \tag{25}$$

with $\sigma^{2,\lim}$ given by Section 3.3, and $R$ given by (20).

Note that is $p_{11} = p_{22} = 0$, then $R = I$ and Theorem 1 provides a heavy-traffic limit for the $X-$model without feedback.

*Proof:*

Fix $r > 0$. Let us first show that Proposition 1 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 the pair $(S^{r,N}, R^{r,N})$ verifies conditions (A1)-(A5) [8] for any $N$, which is clear from Lemma 4 (see [4] for details). 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 for $N$ big enough. Indeed,

(i) By (16) and (17) and applying scaling, we can check that for all $t \geq 0$, $\widehat{W}^{r,N}(t) \in S^{r,N}$. Indeed, for all $r$ and for $N$ big enough,

$$\widehat{W}_1^{r,N} - \frac{\mu_{12}^{r,N}}{\mu_1^{r,N}} \widehat{W}_2^{r,N} = \frac{\mu_{12}^{r,N}}{\mu_1^{r,N}} (r_1^{r,N} r_2^{r,N} - 1) \widehat{\widehat{W}}_2^{r,N} \geq 0,$$

$$\widehat{W}_2^{r,N} - \frac{\mu_{21}^{r,N}}{\mu_2^{r,N}} \widehat{W}_1^{r,N} = \frac{\mu_{21}^{r,N}}{\mu_2^{r,N}} (r_1^{r,N} r_2^{r,N} - 1) \widehat{\widehat{W}}_1^{r,N} \geq 0.$$

(ii) Lemma 3 gives the Skorokhod decomposition.

(iii) By (11), $\widehat{V}_1^{r,N}(t)$ can only increase if $\widehat{\widehat{W}}_2^{r,N} = 0$, that is, if $-\frac{\mu_1^{r,N}}{\mu_{12}^{r,N}} \widehat{W}_1^{r,N} + \widehat{W}_2^{r,N} = 0$ by (16) and (17), which means $\widehat{W}^{r,N} \in F_1^{r,N}$. Symmetrically, $\widehat{V}_2^{r,N}(t)$ can only increase if $\widehat{\widehat{W}}_1^{r,N} = 0$, which means $\widehat{W}^{r,N} \in F_2^{r,N}$.

(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 [11] and Theorem 7.2.5 [12]. Indeed, for any $j = 1, 2$, from (23) and (2) we can write

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

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

$$\mathscr{D} - \lim_{r \to +\infty} \widehat{\widehat{E}}^r = B^H, \tag{26}$$

$B^H$ being a two-dimensional fBm process with data $(x = 0, H, \theta = 0, diag(\alpha)^2 \sigma^{2,\lim})$, which is condition (a) in Proposition 1. Combining (24), **(HT)** and the *continuous mapping theorem*, according to the above limit $\widehat{\widehat{E}}^r$, we deduce the existence of $\widehat{\widehat{X}}^r = \mathscr{D} - \lim_{N \to +\infty} \widehat{X}^{r,N}$, which verifies that

$$\widehat{\widehat{X}}^r(t) = M \widehat{\widehat{E}}^r(t) + \frac{r^{1-H}}{L^{1/2}(r)} R\widehat{\gamma} t, \tag{27}$$

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

Secondly, since hypothesis (b) is accomplished by Lemma 4 and Remark 1, we can apply Proposition 1 to obtain that there exists the following limit:

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

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 $S$ with associated matrix of directions of reflection $R$. The repeated application of Proposition 1 to the sequence $\{(\widehat{\widehat{W}}^r, \widehat{\widehat{X}}^r, \widehat{\widehat{V}}^r)\}_r$, enables us to complete the proof. Indeed, from (27), (26), **(HT)** and the *continuous mapping theorem*, we can ensure the existence of $\mathscr{D} - \lim_{r \to +\infty} \widehat{\widehat{X}}^r = X$, with $X(t) = MB^H(t) + R\gamma t$, which is a two-dimensional fBm process with associated data $\big(x = 0, H, \theta = R\gamma, \Gamma\big)$, where $\Gamma = \sigma^{2,\lim} M diag(\alpha)^2 M^T$ is given by (25). Moreover, by Lemma 4 we can assert that

(b) in Proposition 1 holds, by Remark 1, and by Proposition 1 it follows the existence of

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

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

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

**Acknowledgements:** The author is supported by Ministerio de Economía y Competitividad, Gobierno de España, project ref. MTM2015 67802-P (MINECO/FEDER, UE).

## References

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

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

[3] R. Delgado, State space collapse and heavy-traffic for a packet-switched network with On/Off sources and a fair bandwidth sharing policy, *Telecommunication Systems* , Vol. 62, No. 2, 2016, pp. 461-479.

[4] R. Delgado, A two-queue polling model with priority on one queue and heavy-tailed On/Off sources: a heavy-traffic limit, to appear in *Queueing Systems*, Vol. 83, No. 1, 2016, pp. 57-85.

[5] R. Delgado, A Heavy-Traffic Limit of a Two-Station Fluid Model with Heavy-Tailed On/Off Sources, Feedback and Flexible Servers, *International Journal of Mathematical and Computational Methods*, Vol. 1, No. 1, 2016, pp. 149-158.

[6] R. Delgado, E. Morozov, Stability analysis of some networks with interacting servers. *Analytical and Stochastic Modeling Techniques and Applications (ASMTA) 2014, Lecture Notes in Computer Science*, Vol. 8499, 2014, pp. 1-15.

[7] M. Harrison, Heavy-traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review pollcies, *Ann. Appl. Probab.*, Vol. 8, No. 3, 1998, pp. 822-848.

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

[9] O. Perry, W. Whitt, An ODE for an overloaded X model involving a stochastic averaging principle, *Stochastic Systems*, Vol. 1, No. 1, 2011, pp. 59-108.

[10] O. Perry, W. Whitt, A fluid limit for an overloaded X model via a stochastic averaging principle, *Mathematics of Operations Research*, Vol. 38, No. 2, 2013, pp. 294-349.

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

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

## Appendix: The invariance principle

Kang and Williams prove in Theorem 4.3 [8] 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 [4] 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 [4] 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 [8]:

**Assumption (h)** *For each positive integer n, there are processes $W^n, X^n$ having paths in $\mathscr{D}^d$ and $V^n$ having paths in $\mathscr{C}^d$ defined on some probability space $(\Omega^n, \mathscr{F}^n, P^n)$ 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 $\mathscr{C}-$tight.*

**Proposition 1.** (*Invariance Principle*) Suppose Assumption (h) and assumptions (A1)-(A5) [8] hold, and also that $\lim_{n\to+\infty} R^n = R$. Then, the sequence $\{(W^n, X^n, V^n)\}_n$ is $\mathscr{C}-$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 $(\Omega, \mathscr{F}, P)$, 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 $(x, H, \theta, \Gamma)$, 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 $(x, H, \theta, \Gamma, R)$.