Heavy-traffic analysis of a polling model with fluid queues and heavy-tailed On/Off sources

Rosario Delgado

Abstract We consider a network composed of d single-server workstations with an infinite buffer at each one, that processes continuous fluid whose arrival processes are generated by a big number of heavy-tailed On/Off sources. Class-j fluid is primarily assigned to queue j, $j = 1,\ldots,d$. Servers are disposed in cascade and each station can provide help to the previous ones when it becomes free of its own work. We prove a heavy-traffic limit theorem for an adequate "workload process" associated to this fluid network model. Our limit process is a d-dimensional reflected fractional Browian motion (rfBm) living in a convex polyhedron.

Keywords polling model · reflected fractional Brownian motion · convex polyhedron · On/Off sources · workload process · heavy-traffic limit · Skorokhod problem

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

1 Introduction

By following the model introduced in [7] for three stations, in this paper we study a generalization of the two-station cascade fluid model considered in [5]. Indeed, we investigate the asymptotic behavior under a heavy-traffic regime of a fluid queueing network consisting of d single-server workstations with arbitrary $d \geq 3$, that process or let pass through them d classes of some kind of fluid, such that for any $j = 1,\ldots,d$,

The author is supported by Ministerio de Economía y Competitividad, Gobierno de España, project ref. MTM2015 67802-P

Rosario Delgado
Departament de Matemàtiques
Universitat Autònoma de Barcelona
Edifici C- Campus de la UAB. Av. dels Til.lers s/n.
08193 Cerdanyola del Vallès, SPAIN
E-mail: delgado@mat.uab.cat

class-j fluid is primarily assigned to server j. Servers are disposed in cascade and each station is allowed to give assistance to any of the previous ones (one at a time) when it becomes free of its own work. For the heavy-traffic analysis purpose, we do not need to specify priorities among the previous stations. Indeed, whenever a server, say server j, become idle while there is fluid awaiting for processing at buffer of any previous station, a floodgate opens from one of them, say i, $1 \leq i < j$, and class-i fluid starts to be transferred to station j so that while the situation persists, it is simultaneously processed by servers i and j (possibly at different speeds). We assume that there is no travel delay (setup time). This situation persists until the buffer at station i empties (fluid runs out) or the arrival of fluid (class-j fluid from outside, or other class of fluid from other previous station) to station j starts, whichever happens first. In the latter case, transfer from station i immediately ceases (the floodgates close) to station j, while server i goes on with the processing of class-i fluid. In this sense, each station supports the previous ones, although the converse is not allowed. Class-i fluid can be processed simultaneously by station i and any subset of stations $\{i+1,\ldots,d\}$. For station j, priority is class-j fluid, and the service is exhaustive within each class. That is, if there is class-j fluid in the system, it is processed by server j (and perhaps other servers in $\{j+1,\ldots,d\}$) until the queue empties, and only then server j accepts class-k fluid with $k < j$ (by strictly following the priority ordering, that we do not specify), coming back to class-j fluid as soon as new class-j fluid arrives to the system. We assume that there is an infinite-capacity buffer at each station and that no server can be idle if there is fluid awaiting for processing at its own buffer or buffers of the previous stations (nonidling or work-conserving policy).

This system is a polling model portrayed schematically in Figure 1 for $d = 3$. This kind of queueing network with *flexible* servers, in which servers may transfer some service capacity to accommodate workload accumulated in another ones, can be found in the literature as models for a variety of 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 (skill-based systems). See for instance [6], [7] and references therein. The most related papers focus on an optimal server allocation to minimize a cost function of these network systems, while others are devoted to stability analysis, as [6], [7]. Up to our knowledge, there are very few previous work considering the question of the heavy-traffic behavior of this type of systems, and even the concept of *"heavy-traffic"* itself has not been deeply treated up to now in this context. An exception is the paper of Harrison [10], that considers dynamic scheduling in a two-station cascade system with two independent Poisson input streams, deterministic service and linear holding costs, under a heavy-traffic regime where the combined capacity of the two servers is approximately equal to the total input rate.

We assume that the process of external arrivals to the network system of each fluid class is a non-deterministic aggregated cumulative process generated by a large enough number of heavy tailed On/Off sources, N. The aggregated network traffic generated by the superposition of many On/Off sources with strictly alternating On- and Off-periods and whose On- and/or Off-periods lengths have high variability, has the properties of be self-similar and long-range dependent. This justifies our assump-

tion since actually it has been observed the presence of long-range dependence in broadband network traffic as well as that of self-similar traffic patterns in modern high-speed network traffic ([13], [1]). It is that these are the same properties that characterize the fractional Brownian motion (fBm) process, which is a self-similar process that has long-range dependent increments, which are positively correlated if its Hurst parameter H belongs to the interval $(1/2, 1)$.

Adequate definition of "workload process" is introduced in [5] for $d = 2$ (see justification and references therein). As mentioned in [5], we follow [10] and define the workload process for station j, as the total time of service that would be required to complete processing of all fluid in stations $1, \ldots, j$ at time t, if server j were required to complete the processing of all of them without help form other servers.

We consider a double sequence of systems indexed by r (a parameter of change of scale in time) 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 heavy-traffic limit theorem for the $d-$dimensional workload process with a convenient change of scale. Indeed, in Theorem 1 we prove that after adequate scaling, the workload processes converges to a d-dimensional *reflected fractional Brownian motion* (rfBm) process living in a convex polyhedron (which is not the positive orthant).

RfBm is a stochastic process that has been introduced previously in the context of heavy-traffic limit theorems. See for instance [1]-[3], in which the process lives in the positive orthant, and [4], where the rfBm process lives in a convex polyhedron with constant directions of reflection along each face. As in [5], in this paper we explore a different type of fluid network than that considered in [4] for which the rfBm process lives in a convex polyhedron with constant directions of reflection along each face also appears as the limit process in a heavy-traffic limit theorem.

A key ingredient in the proof of Theorem1 is an *Invariant Principle* that we introduce in Section 5 (Appendix), which is a version of the Invariance Principle for Semimartingale reflecting Brownian motions living in the closure of a domain with piecewise smooth boundaries presented in Theorem 4.3 [11]. This principle does not depend, in fact, on the specific law of the processes, as can be seen in [4], and we apply it to the rfBm process and to a sequence of convex polyhedra that approximates the convex polyhedron in which the limit rfBm process lives.

The organization of the paper is as follows. In Section 2 we give main notations and introduce the definitions of convex polyhedron and rfBm process on a convex polyhedron. In Section 3 we explore the d-station tree-cascade fluid network, first introducing the model, and then the processes used to measure its performance, the sequence of convex polyhedra, the normalization factors to scale these processes and the heavy-traffic condition. The heavy-traffic limit theorem is stated and proved in Section 4. Section 5 is an Appendix in which the Invariance Principle used in the proof of the heavy-traffic limit theorem, can be found.

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). As usual, we use the notation $\mathbb{R}^-$ for the half-line $(-\infty, 0]$ and for any $x \in \mathbb{R}$, $[x]$ denotes the greatest integer less than or equal to x. 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 $A = (a_{ij})_{i=1,\dots,d,\, j=1,\dots,m}$, 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 $x \in \mathbb{R}$). 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 same applies for the particular case $m = 1$, which corresponds to $d-$dimensional vectors, with $|v| \stackrel{\text{def}}{=} \sum_{1\leq i\leq d} |v_i|$ the ℓ_1-norm. The Euclidean norm on $\mathbb{R}^d$ is $||v|| = \big(\sum_{1\leq i\leq d} v_i^2\big)^{1/2} \leq |v|$. 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 $d(x, A)$ denote the distance between $x \in \mathbb{R}^d$ and $A \subset \mathbb{R}^d$, $d(x, A) = \inf\{\, ||x - y|| : y \in A\}$, with the convention $d(x, \emptyset) = +\infty$. For any $r > 0$, let $U_r(A)$ denote the closed set $\{x \in \mathbb{R}^d : d(x, A) \leq r\}$. For a set $A \subset \mathbb{R}^d$, we denote by A^c the complement of A in $\mathbb{R}^d$, that is, $A^c = \{x \in \mathbb{R}^d : x \notin A\}$.

Let $\mathscr{C}^d$ 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 $\mathscr{D}^d$ 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 $d \geq 1$. For each $T \geq 0$ and $\omega \in \mathscr{C}^d$, we define

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

We will say that $\omega^n \to \omega$ as $n \to +\infty$ in $\mathscr{C}^d$ (*uniformly on compacts*, u.o.c.) if for any $T \geq 0$, $||\omega^n(\cdot) - \omega(\cdot)||_T \to 0$, and we will denote it $\lim_{n\to+\infty} \omega^n = \omega$.

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 $\mathscr{D} - \lim_{n\to+\infty} X^n = X$ if the sequence of probability measures induced in $\mathscr{D}^d$ by $\{X^n\}_n$, say $\{P^n\}_n$, converges weakly to that induced by X, P. We denote the weak convergence of probability measures by $P^n \Rightarrow P$.

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 $\mathscr{C}^d$.

The multi-dimensional reflected fractional Brownian motion (rfBm) process on the positive orthant has been introduced, for instance, in Delgado [1,2] and Konstantopoulos and Lin [12]. In this paper we use the extension of this process to a convex

polyhedron with constant directions of reflection along each face, introduced in Delgado [4], reproduced here for the convenience of the reader.

Definition 1 (convex polyhedron) For any $d \geq 1$, a convex polyhedron S on $\mathbb{R}^d$ can be defined algebraically as the set of solutions to a systems of linear inequalities:

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

where $v^1, \ldots, v^d$ in $\mathbb{R}^d$, being Υ 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 $F_\ell = \{x \in S : \langle v^\ell, x\rangle = 0\}$ are the boundary faces of S. Note that $0 \in \partial S \subset S$. For each $x \in S$, let $\ell(x) = \{j = 1, \ldots, d : x \in F_j\}$.

The fact that the convex polyhedron is determined by the matrix Υ is made explicit, when convenient, by means of the notation $S(\Upsilon)$. It is assumed that the interior of $S(\Upsilon)$ is no empty and that the set $\{v^1, \ldots, v^d\}$ is minimal in the sense that no proper subset defines $S(\Upsilon)$, that is, for any strict subset $L \subset \{1, \ldots, d\}$, the set $\{x \in \mathbb{R}^d : \langle v^\ell, x\rangle \geq 0 \quad \text{for all } \ell \in L\}$ is strictly larger than $S(\Upsilon)$. This is equivalent to the assumption that each of the boundary faces F_ℓ has dimension $d-1$. 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* $u(y)$ for any y on its boundary, which are constant along each face, by using a $d \times d$ matrix R whose column vectors are denoted by $u^1, \ldots, u^d$ and are normalized in such a way that $\langle u^\ell, v^\ell\rangle = 1$ for any $\ell = 1, \ldots, d$, in the following way: if $\ell(y) = \{\ell\}$ for some ℓ, $u(y)$ is defined as u^ℓ. That is, u^ℓ is the reflection direction interior to F_ℓ. Otherwise,

$$u(y) \stackrel{\text{def}}{=} \{\xi \in \mathbb{R}^d : \xi = \sum_{i \in \ell(y)} \delta_i u^i, \text{ with some } \delta_i \geq 0 \text{ and such that } |\xi| = 1\}, \quad (1)$$

that is, the possible directions of reflection at points on the intersections of some faces are in the convex hull of the directions on the adjoint faces.

Remark 1 (See Remark 1 [5]) It is well known that if the matrix product $Q = \Upsilon R$ is a *completely-S* matrix, then fixed any point on the intersection of some faces, there is a nonnegative linear combination of the reflection directions given by the column vectors of R that points strictly inward $S(\Upsilon)$. Moreover, if matrix $Q = (q_{ij})_{i,j=1,\ldots,d}$ verifies that $q_{ij} \leq 0$ for all $i \neq j$, and the following condition, named *the generalized Harrison-Reiman condition*,

$$(\mathbf{HR}) : \text{The matrix } \Theta \text{ obtained from } Q - I_d \text{ by replacing its entries by their absolute values, has spectral radius strictly less than } 1,$$

then Q is a *completely-S* matrix.

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, θ, Γ), that is, it is a continuous Gaussian process starting from point x, with mean function $E\big(X(t)\big) = x + \theta t$ for any $t \geq 0$ (θ is the *drift vector*), 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 *(multidimensional) Skorokhod Problem* associated to X on the convex polyhedron $S(\Upsilon)$ with associated matrix of directions of reflection R.

Remark 2 Strong existence and uniqueness of the solution of a Skorokhod problem can be ensured if the column vectors of R, $\{u^1, \ldots, u^d\}$, are linearly independent, and matrix $Q = \Upsilon R$ verifies *the generalized Harrison-Reiman condition* **(HR)** introduced in Remark 1. See Remark 1 in Delgado [4] for a detailed justification of this assertion, based in the couple of papers of Dupuis and Ramanan [8,9].

Rather informally, we can say that the rfBm process starts in the interior of the convex polyhedron S and behaves like a fractional Brownian motion (fBm) being constrained to remain within S in the following way: when the fBm process touches the boundary of S, it is instantaneously "reflected" preventing its exit from it. For each ℓ, the $\ell-$th column vector of matrix R, u^ℓ, gives the direction of the reflection on F_ℓ, and V_ℓ gives its intensity. On the intersection of two or more faces, the direction of reflection is given by a linear combination of the corresponding vectors u^ℓ of the form given by (1).

3 The d-station polling fluid network

In this section we go deeper into the details of the fluid network with $d \geq 3$ stations in cascade yet introduced in Section 1. The particular case of such a network with

$d = 3$ is pictured in Figure 3. Stability of a similar system with the same structure but where each station is fed by a renewal input with general i.i.d. inter-arrival times and general i.i.d. service times for customers, in the particular case of $d = 3$, has been studied in [7], where sufficient conditions have been found.

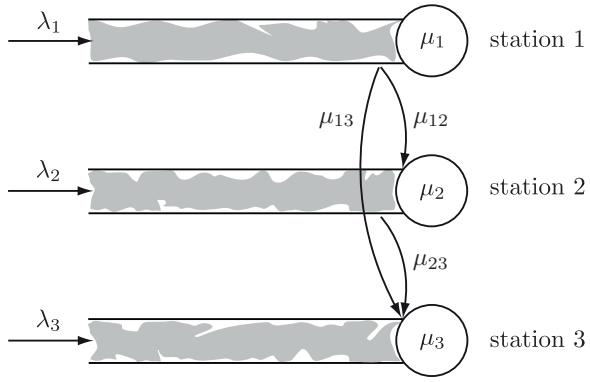

Fig. 1 A three-station tree-cascade fluid network.

Suppose that for each fluid class 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, where $U_j^{(n)}(t) = 1$ means that at time t source n is On (and it is sending fluid to station j, at a constant rate), and $U_j^{(n)}(t) = 0$ 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^{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

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

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

$$\int_x^{+\infty} f^{\text{on}}(u)\, du \sim x^{-\beta^{\text{on}}} L^{\text{on}}(x) \quad \text{and} \quad \int_x^{+\infty} f^{\text{off}}(u)\, du \sim x^{-\beta^{\text{off}}} L^{\text{off}}(x)\,, \qquad (2)$$

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

We define the cumulative external class-j fluid arrived up to time t (by the N sources) at station j by:

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

where $\alpha_j^N > 0$ is the (possibly dependent on N) deterministic rate at which class-j 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 $E^N = \{E^N(t) = \left(E_1^N(t), \ldots, E_d^N(t)\right)^T, t \geq 0\}$, are assumed to be independent. We also assume $E^N(0) = 0$. We assume that fluid at each server is processed in a first-in-first-out (FIFO) basis.

Let $\lambda^N = (\lambda_1^N, \ldots, \lambda_d^N)^T$, where $\lambda_j^N \stackrel{\text{def}}{=} \alpha_j^N \frac{\tilde{\mu}^{\text{on}}}{\tilde{\mu}^{\text{on}} + \tilde{\mu}^{\text{off}}}$ can be thought as the long run fluid rate for fluid class j. Assume that $\lambda = \lim_{N \to +\infty} \lambda^N$ exists, $\lambda = (\lambda_1, \ldots, \lambda_d)^T$. This implies that $\alpha = (\alpha_1, \ldots, \alpha_d)^T = \lim_{N \to +\infty} (\alpha_1^N, \ldots, \alpha_d^N)^T$ 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 for any $j = 1, \ldots, d$, server j processes or lets pass through it class-j fluid at a constant rate $\mu_j^{r,N} > 0$ if station j were devoted all time to it (that is, $1/\mu_j^{r,N}$ is the *mean service time* for class-j fluid at station j). By the other hand, class-j fluid is processes at a constant rate $\mu_{j\ell}^{r,N} > 0$, not necessarily equal to $\mu_j^{r,N}$ nor to $\mu_\ell^{r,N}$, if station $\ell \in \{j+1, \ldots, d\}$ devoted all time to this fluid class ($1/\mu_{j\ell}^{r,N}$ is the *mean service time* for class-j fluid processed by server ℓ). Let $\mu^{r,N} = (\mu_j^{r,N}, \mu_{j\ell}^{r,N})_{j,\ell \in \{1,\ldots,d\}, j<\ell}$ as column vector. We assume that $\lim_{N \to +\infty} \mu^{r,N}$ exists and does not depend on r; we denote it simply by μ. Finally, we also introduce the *fluid traffic intensity* for the system, $\rho^{r,N} = (\rho_1^{r,N}, \ldots, \rho_d^{r,N})^T$, by

$$\rho_1^{r,N} \stackrel{\text{def}}{=} \frac{\lambda_1^N}{\mu_1^{r,N}}, \qquad \rho_j^{r,N} \stackrel{\text{def}}{=} \sum_{\ell=1}^{j-1} \frac{\lambda_\ell^N - \mu_\ell^{r,N}}{\mu_{\ell j}^{r,N}} + \frac{\lambda_j^N}{\mu_j^{r,N}} \quad j = 2, \ldots, d. \tag{3}$$

The *heavy-traffic condition* establishes that the *fluid traffic intensity* $\rho^{r,N}$ tends to $e = (1, \ldots, 1)^T \in \mathbb{R}^d$ in some sense that will be specified later.

Let introduce the following notation: for all $1 \leq i < j \leq d$, $s_{ij} = \frac{\mu_i}{\mu_{ij}}$, and for all pair (r,N), $s_{ij}^{r,N} = \frac{\mu_i^{r,N}}{\mu_{ij}^{r,N}}$. Throughout this work we suppose the following assumption holds:

Assumption (s): If $d \geq 3$, $1 \leq i < j < k \leq d$ and r and N are big enough, then

$$s_{ij}^{r,N}\, s_{jk}^{r,N} = s_{ik}^{r,N}.$$

Remark 3 Assumption (s) is accomplished, for instance, if for all $1 \leq i < j \leq d$, $\mu_{ij}^{r,N} = f(j-i)\, \mu_j^{r,N}$ with $f(x) = e^{ax}$ for some $a \in \mathbb{R}$, that is, in case that $\frac{\mu_{ij}^{r,N}}{\mu_j^{r,N}}$, which is the rate ratio or relative difference measure to compare the rate at which server j processes class-i and class-j fluids, grows (if $a > 0$) or decays (if $a < 0$) at a rate proportional to the difference $j - i$. Case $a = 0$ corresponds to $\mu_{ij}^{r,N} = \mu_j^{r,N}$.

3.1 Performance processes

Some nonnegative processes will be used to measure the performance of the two-station cascade fluid network:

The *workload process* $W^{r,N} = (W_1^{r,N}, \ldots, W_d^{r,N})^T$ is introduced analogously to [5]: for any station $j = 1, \ldots, d$, $W_j^{r,N}(t)$ represents the total time of service that would be required to complete processing of all class-k fluid in the system at time t for $k = 1, \ldots, j$, if server j were required to complete the processing of all of them without help from other servers. We assume that $W^{r,N}(0) = 0$. We denote by $\widetilde{W}_j^{r,N}(t)$ the portion of the workload $W_j^{r,N}(t)$ that is exclusively due to fluid arriving from outside the system, that is, due to all class-j fluid. That is,

$$\widetilde{W}_j^{r,N}(t) \stackrel{\text{def}}{=} \frac{E_j^N(t)}{\mu_j^{r,N}} - \Big(T_j^{r,N}(t) + \sum_{\ell=j+1}^{d} \frac{1}{s_{j\ell}^{r,N}} T_{j\ell}^{r,N}(t)\Big) \geq 0 \quad \text{if } j = 1, \ldots, d-1\,, \tag{4}$$

$$\widetilde{W}_d^{r,N}(t) \stackrel{\text{def}}{=} \frac{E_d^N(t)}{\mu_d^{r,N}} - T_d^{r,N}(t) \geq 0\,,$$

where $T_j^{r,N}(t)$ is the total service time devoted to class-j fluid (by server j) in the interval $[0,t]$, and $T_{j\ell}^{r,N}(t)$ is the total service time devoted to class-j by server ℓ in the same time interval. Indeed, $E_j^N(t)/\mu_j^{r,N}$ would be the amount of time required by server j to process all the class-j fluid arrived up to time t to the system, while $T_j^{r,N}(t) + \sum_{\ell=j+1}^{d} \frac{1}{s_{j\ell}^{r,N}} T_{j\ell}^{r,N}(t)$ represents the part of this time yet consumed at instant t, by server j, which is $T_j^{r,N}(t)$, and if $j < d$ also by other servers to which part of this fluid has been transferred, which is $T_{j\ell}^{r,N}(t)$ conveniently rescaled since the service time for class-j fluid is different when processed by server j and when processed by server ℓ, for $\ell = j+1, \ldots, d$. If $j = d$, it only appears the term $T_d^{r,N}(t)$.

Then, by definition of workload $W_j^{r,N}$ and recurrence, we have that

$$W_1^{r,N}(t) = \widetilde{W}_1^{r,N}(t), \tag{5}$$

$$W_j^{r,N}(t) = \widetilde{W}_j^{r,N}(t) + \check{W}_j^{r,N}(t), \text{ with } \check{W}_j^{r,N}(t) = s_{j-1\,j}^{r,N} W_{j-1}^{r,N}(t),\ j = 2, \ldots, d, \tag{6}$$

that is, to obtain $W_j^{r,N}(t)$ from $\widetilde{W}_j^{r,N}(t)$ we add $W_{j-1}^{r,N}(t)$ conveniently rescaled representing the amount of time required by server j to process all the class-ℓ fluid for $\ell = 1, \ldots, j-1$ stored at the system at time t.

The *cumulative idle-time process* $Y^{r,N} = (Y_1^{r,N}, \ldots, Y_d^{r,N})^T$ 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,

$$Y_j^{r,N}(t) = \int_0^t 1_{\{W_1^{r,N}(s) = \cdots = W_j^{r,N}(s) = 0\}}\, ds\,, \tag{7}$$

and the *total service time process* and the *cumulative idle-time process* are related by means of the following equalities:

$$Y_1^{r,N}(t) = t - T_1^{r,N}(t)\,,\; Y_j^{r,N}(t) = t - \big(T_j^{r,N}(t) + \sum_{i=1}^{j-1} T_{ij}^{r,N}(t)\big),\; j = 2,\dots,d\,. \tag{8}$$

Note that with this notations, for any $j = 1,\dots,d$ and for any pair (i,j) with $1 \le i < j \le d$, respectively, we can write

$$\begin{aligned} T_j^{r,N}(t) &= \int_0^t 1_{\{\widetilde{W}_j^{r,N}(s)>0\}} ds\,, \\ T_{ij}^{r,N}(t) &= \int_0^t 1_{\{\widetilde{W}_i^{r,N}(s)>0,\, \widetilde{W}_j^{r,N}(s)=0\} \cap \tilde{\tau}_{ij}^{r,N}(s)}\, ds\,, \end{aligned} \tag{9}$$

where $\tilde{\tau}_{ij}^{r,N}(s)$ depends on the priorities among stations $1,\dots,j-1$ when server j becomes idle. Only to mention two examples, if priority is given to the nearest station, then

$$\tilde{\tau}_{ij}^{r,N}(s) = \{\widetilde{W}_{i+1}^{r,N}(s) = \cdots = \widetilde{W}_{j-1}^{r,N}(s) = 0\}\,,$$

but

$$\tilde{\tau}_{ij}^{r,N}(s) = \{i = argmax_{\ell=1,\dots,j-1} \widetilde{W}_\ell^{r,N}(s)\}$$

if priority is given to the previous station with a greater workload due exclusively to fluid arriving from outside. What is important is that we do not need to specify $\tilde{\tau}_{ij}^{r,N}(\cdot)$ for performing the heavy-traffic analysis of our system.

Finally, we introduce the process $V^{r,N} = (V_1^{r,N},\dots,V_d^{r,N})^T$ by:

$$V_1^{r,N}(t) \stackrel{\text{def}}{=} Y_1^{r,N}(t) + \sum_{\ell=2}^{d} \frac{1}{s_{1\ell}^{r,N}}\, Y_\ell^{r,N}(t), \tag{10}$$

$$\begin{aligned} V_j^{r,N}(t) \stackrel{\text{def}}{=}\; & \big(Y_j^{r,N}(t) + \sum_{h=1}^{j-1} T_{hj}^{r,N}(t)\big) \\ & + \sum_{\ell=j+1}^{d} \frac{1}{s_{j\ell}^{r,N}} \big(Y_\ell^{r,N}(t) + \sum_{h=1}^{j-1} T_{h\ell}^{r,N}(t)\big),\; 1 < j < d, \end{aligned} \tag{11}$$

$$V_d^{r,N}(t) \stackrel{\text{def}}{=} Y_d^{r,N}(t) + \sum_{h=1}^{d-1} T_{hd}^{r,N}(t)\,. \tag{12}$$

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

Let we define $G_1 = \{(x_1,\dots,x_d) \in \mathbb{R}^d : x_1 \ge 0\}$,

$$G_j = \{(x_1,\dots,x_d) \in \mathbb{R}^d : x_j \ge s_{j-1\,j}\, x_{j-1}\},\ \text{for } j = 2,\dots,d,$$

and $S = \cap_{j=1}^{d} G_j$. Then, S is the convex polyhedron in $\mathbb{R}^d$ determined by matrix

$$\Upsilon = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ -s_{12} & 1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & -s_{23} & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -s_{d-1\,d} & 1 \end{pmatrix}, \tag{13}$$

and the set of row vectors $\{v^1, \ldots, v^d\}$ is minimal, n^j are the inward unit normal to the closed half spaces G_j, with $n^j = v^j/||v^j||$, and the boundary faces are

$$F_1 = \{(x_1, \ldots, v_d) \in S : x_1 = 0\} \text{ and}$$
$$F_j = \{(x_1, \ldots, x_d) \in S : x_j = s_{j-1\,j}\, x_{j-1}\}, \text{ for } j = 2, \ldots, d.$$

The boundary of S is $\partial S = \cup_{j=1}^{d} F_j$. The sequence of convex polyhedron $S^{r,N} = S^{r,N}(\Upsilon^{r,N})$ is introduced analogously to S. The matrices of reflection $R^{r,N}$ and R are introduced in Lemmas 1 and 2, respectively, in the next subsection. Matrices $\Upsilon^{r,N}$ converge to Υ as $N \to +\infty$ (and the limit does not depend on r).

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) fluid model depend on r and N.

3.3 Scaled processes

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 [13] (see also Delgado [1], [2]). Set $a^{\text{on}} = \frac{\Gamma(2-\beta^{\text{on}})}{(\beta^{\text{on}}-1)}$ and $a^{\text{off}} = \frac{\Gamma(2-\beta^{\text{off}})}{(\beta^{\text{off}}-1)}$, where β^{on} and β^{off} are defined by (2). The normalization factors used below depend on b, defined by $b \stackrel{\text{def}}{=} \lim_{t\to+\infty} \frac{L^{\text{on}}(t)}{L^{\text{off}}(t)}\, t^{\beta^{\text{off}}-\beta^{\text{on}}}$, which exists although it could be infinite. If $0 < b < +\infty$ (implying $\beta^{\text{on}} = \beta^{\text{off}}$ and $b = \lim\limits_{t\to+\infty} \dfrac{L^{\text{on}}(t)}{L^{\text{off}}(t)}$), set $\beta = \beta^{\text{on}} = \beta^{\text{off}}$, $L = L^{\text{off}}$ and

$$\sigma^{2,\text{lim}} \stackrel{\text{def}}{=} \frac{2\left((\tilde{\mu}^{\text{off}})^2\, a^{\text{on}}\, b + (\tilde{\mu}^{\text{on}})^2\, a^{\text{off}}\right)}{\left(\tilde{\mu}^{\text{on}} + \tilde{\mu}^{\text{off}}\right)^3 \Gamma(4-\beta)}. \tag{14}$$

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

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

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

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

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

$$H \stackrel{\text{def}}{=} \frac{3-\beta}{2}. \tag{15}$$

Therefore, $H \in (\frac{1}{2}, 1)$.

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

$$\textbf{(HTd)} \quad \begin{cases} \lim\limits_{N\to+\infty} \sqrt{N}(\rho^{r,N} - e) = \widehat{\gamma}^r & \text{for some } \widehat{\gamma}^r \in \mathbb{R}^{d-1}\times\mathbb{R}^- \quad \text{and} \\ \lim\limits_{r\to+\infty} \frac{r^{1-H}}{L^{1/2}(r)}\,\widehat{\gamma}^r = \gamma & \text{for some } \gamma \in \mathbb{R}^{d-1}\times\mathbb{R}^-. \end{cases}$$

Remark 4 Heavy-traffic condition **(HT)** generalizes that introduced in Delgado [1] since $\widehat{\gamma}^r$ was taken there to be identically zero, by following Delgado [3], where motivation for this kind of generalization in terms of what is known as *"thin control"* is given. Moreover, it also generalizes that of [5] for $d > 2$.

We can introduce the *scaled processes* associated with the (r,N) fluid model and use a hat to denote them: $\widehat{W}^{r,N}$, $\widehat{E}^{r,N}$, $\widehat{Y}^{r,N}$, $\widehat{V}^{r,N}$ and $\widehat{T}^{r,N}$ by

$$\widehat{E}_j^{r,N}(t) \stackrel{\text{def}}{=} \sqrt{N}\,\frac{E_j^N(rt) - \lambda_j^N rt}{r^H L^{1/2}(r)}, \tag{16}$$

$$\widehat{W}_j^{r,N}(t) \stackrel{\text{def}}{=} \sqrt{N}\,\frac{W_j^{r,N}(rt)}{r^H L^{1/2}(r)} \quad (\text{for } j = 1,\dots,d), \tag{17}$$

and analogously to $\widehat{W}$ for the rest of the processes.

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

Lemma 1 *Under Assumption (s) the scaled processes are related by means of*

$$\widehat{W}^{r,N}(t) = \widehat{X}^{r,N}(t) + R^{r,N}\,\widehat{V}^{r,N}(t), \tag{18}$$

with $\widehat{X}^{r,N} = (\widehat{X}_1^{r,N},\dots,\widehat{X}_d^{r,N})^T$ *defined by*

$$\widehat{X}_1^{r,N}(t) \stackrel{\text{def}}{=} \frac{\widehat{E}_1^{r,N}(t)}{\mu_1^{r,N}} + \frac{\sqrt{N}\,r^{1-H}}{L^{1/2}(r)}\,(\rho_1^{r,N} - 1)\,t, \tag{19}$$

$$\widehat{X}_j^{r,N}(t) \stackrel{\text{def}}{=} \sum_{i=1}^{j-1} \frac{\widehat{E}_i^{r,N}(t)}{\mu_{ij}^{r,N}} + \frac{\widehat{E}_j^{r,N}(t)}{\mu_j^{r,N}} + \frac{\sqrt{N}\,r^{1-H}}{L^{1/2}(r)}\,(\rho_j^{r,N} - 1)\,t, \quad j = 2,\dots,d, \tag{20}$$

where $\rho^{r,N}$ *is defined by* (3)*, and* $R^{r,N}$ *is the matrix*

$$R^{r,N} \stackrel{\text{def}}{=} \begin{pmatrix} 1 & -\frac{1}{s_{12}^{r,N}} & 0 & 0 & \cdots & \cdots & 0 \\ s_{12}^{r,N} & 0 & -\frac{1}{s_{23}^{r,N}} & 0 & \cdots & \cdots & 0 \\ s_{13}^{r,N} & 0 & 0 & -\frac{1}{s_{34}^{r,N}} & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ s_{1d-1}^{r,N} & 0 & 0 & 0 & \cdots & \cdots & -\frac{1}{s_{d-1d}^{r,N}} \\ s_{1d}^{r,N} & 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}. \tag{21}$$

Proof:

We prove (18) for each component of process $\widehat{W}^{r,N}$, $\widehat{W}_j^{r,N}$, by induction on $j = 1,\ldots,d$.

The base case: $j = 1$.

By (5) first, and then by (4) and (8) with $j = 1$, we have that

$$\begin{aligned} \widehat{W}_1^{r,N}(t) &= \sqrt{N}\,\frac{\widetilde{W}_1^{r,N}(rt)}{r^H L^{1/2}(r)} \\ &= \frac{\sqrt{N}}{r^H L^{1/2}(r)}\Big(\frac{E_1^N(rt)-\lambda_1^N rt}{\mu_1^{r,N}} + \frac{\lambda_1^N}{\mu_1^{r,N}}\,rt - rt + Y_1^{r,N}(rt) - \sum_{\ell=2}^{d}\frac{1}{s_{1\ell}^{r,N}}\,T_{1\ell}^{r,N}(t)\Big) \\ &= \frac{\widehat{E}_1^{r,N}(t)}{\mu_1^{r,N}} + \frac{\sqrt{N}}{r^H L^{1/2}(r)}\,(\rho_1^{r,N}-1)\,rt + \widehat{Y}_1^{r,N}(t) - \sum_{\ell=2}^{d}\frac{1}{s_{1\ell}^{r,N}}\,\widehat{T}_{1\ell}^{r,N}(t) \\ &= \widehat{X}_1^{r,N}(t) + \big(\widehat{V}_1^{r,N}(t) - \frac{1}{s_{12}^{r,N}}\,\widehat{V}_2^{r,N}(t)\big), \end{aligned} \tag{22}$$

where in the last equality we used (19), (10) and (11) with $j = 2$. We can check that (22) corresponds exactly to the first component of process $\widehat{W}^{r,N}$ in (18).

Inductive step: we assume that the result is proved up to j, with $j \in \{1,\ldots,d-1\}$, and then we will prove that it also holds for component $j+1$.

Indeed, by (6) and (4) for $j+1$ we have that

$$\begin{aligned} \widehat{W}_{j+1}^{r,N}(t) &= \sqrt{N}\,\frac{W_{j+1}^{r,N}(rt)}{r^H L^{1/2}(r)} = \sqrt{N}\,\frac{\widetilde{W}_{j+1}^{r,N}(rt)+\breve{W}_{j+1}^{r,N}(rt)}{r^H L^{1/2}(r)} \\ &= \frac{\sqrt{N}}{r^H L^{1/2}(r)}\Big(\frac{E_{j+1}^N(rt)}{\mu_{j+1}^{r,N}} - \big(T_{j+1}^{r,N}(rt) + \sum_{\ell=j+2}^{d}\frac{1}{s_{j+1\ell}^{r,N}}T_{j+1\ell}^{r,N}(rt)\big) + s_{jj+1}^{r,N}W_j^{r,N}(rt)\Big) \\ &= \frac{\widehat{E}_{j+1}^N(t)}{\mu_{j+1}^{r,N}} + \frac{\sqrt{N}\,r^{1-H}}{L^{1/2}(r)}\,\frac{\lambda_{j+1}^N}{\mu_{j+1}^{r,N}}\,t \\ &- \big(\widehat{T}_{j+1}^{r,N}(t) + \sum_{\ell=j+2}^{d}\frac{1}{s_{j+1\ell}^{r,N}}\widehat{T}_{j+1\ell}^{r,N}(t)\big) + s_{jj+1}^{r,N}\,\widehat{W}_j^{r,N}(t) \end{aligned} \tag{23}$$

(where the summatory does not appear in the case $j=d-1$). By the induction hypothesis,

$$\begin{aligned}\widehat{W}_j^{r,N}(t) &= \widehat{X}_j^{r,N}(t) + \big(s_{1j}^{r,N}\widehat{V}_1^{r,N}(t) - \frac{1}{s_{j\,j+1}^{r,N}}\widehat{V}_{j+1}^{r,N}(t)\big)\\ &= \sum_{i=1}^{j-1}\frac{\widehat{E}_i^{r,N}(t)}{\mu_{ij}^{r,N}} + \frac{\widehat{E}_j^{r,N}(t)}{\mu_j^{r,N}} + \frac{\sqrt{N}\,r^{1-H}}{L^{1/2}(r)}(\rho_j^{r,N}-1)t\\ &+ \big(s_{1j}^{r,N}\widehat{V}_1^{r,N}(t) - \frac{1}{s_{j\,j+1}^{r,N}}\widehat{V}_{j+1}^{r,N}(t)\big)\end{aligned}$$

by (20). By replacing this expression into (25) and taking into account (3) we obtain

$$\begin{aligned}\widehat{W}_{j+1}^{r,N}(t) &= \sum_{i=1}^{j}\frac{\widehat{E}_i^{r,N}(t)}{\mu_{i\,j+1}^{r,N}} + \frac{\widehat{E}_{j+1}^{r,N}(t)}{\mu_{j+1}^{r,N}} + \frac{\sqrt{N}\,r^{1-H}}{L^{1/2}(r)}\rho_{j+1}^{r,N}t - \widehat{T}_{j+1}^{r,N}(t)\\ &- \sum_{\ell=j+2}^{d}\frac{1}{s_{j+1\,\ell}^{r,N}}\widehat{T}_{j+1\,\ell}^{r,N}(t) + \big(s_{1\,j+1}^{r,N}\widehat{V}_1^{r,N}(t) - \widehat{V}_{j+1}^{r,N}(t)\big)\,. \qquad (24)\end{aligned}$$

By the other hand, by (8) for $j+1$ we can write

$$\widehat{T}_{j+1}^{r,N}(t) = \frac{\sqrt{N}\,r^{1-H}}{L^{1/2}(r)}t - \widehat{Y}_{j+1}^{r,N}(t) - \sum_{i=1}^{j}\widehat{T}_{i\,j+1}^{r,N}(t)\,,$$

which can be replaced into (24) implying that

$$\begin{aligned}\widehat{W}_{j+1}^{r,N}(t) &= \widehat{X}_{j+1}^{r,N}(t) + s_{1\,j+1}^{r,N}\widehat{V}_1^{r,N}(t)\\ &- \Big(\sum_{\ell=j+2}^{d}\frac{1}{s_{j+1\,\ell}^{r,N}}\widehat{T}_{j+1\,\ell}^{r,N}(t) + \sum_{\ell=j+2}^{d}\frac{1}{s_{j+1\,\ell}^{r,N}}\big(\widehat{Y}_\ell^{r,N}(t) + \sum_{h=1}^{j}\widehat{T}_{h\ell}^{r,N}(t)\big)\Big), \qquad (25)\end{aligned}$$

by applying (20) and (11) with $j+1$ (if $j+1<d$) or (12) (if $j+1=d$), where the two summatories indexed by ℓ do not appear in the latter case. Therefore, if $j+1=d$ we obtain directly

$$\widehat{W}_d^{r,N}(t) = \widehat{X}_d^{r,N}(t) + s_{1\,d}^{r,N}\widehat{V}_1^{r,N}(t)$$

as desired. Otherwise, $j+1<d$ and we can easily check that

$$\sum_{\ell=j+2}^{d}\frac{1}{s_{j+1\,\ell}^{r,N}}\widehat{T}_{j+1\,\ell}^{r,N}(t) + \sum_{\ell=j+2}^{d}\frac{1}{s_{j+1\,\ell}^{r,N}}\big(\widehat{Y}_\ell^{r,N}(t) + \sum_{h=1}^{j}\widehat{T}_{h\ell}^{r,N}(t)\big) = \frac{1}{s_{j+1\,j+2}^{r,N}}\widehat{V}_{j+2}^{r,N}(t)$$

by (11) if $j+2<d$, and by (12) if $j+2=d$, implying that (25) can be rewritten as

$$\widehat{W}_{j+1}^{r,N}(t) = \widehat{X}_{j+1}^{r,N}(t) + s_{1\,j+1}^{r,N}\widehat{V}_1^{r,N}(t) - \frac{1}{s_{j+1\,j+2}^{r,N}}\widehat{V}_{j+2}^{r,N}(t)\,,$$

which corresponds to the component $j+1$ of $\widehat{W}^{r,N}(t)$ in expression (18), what ends up the proof. □

Lemma 2 *For the (r,N) fluid model and under Assumption (s), the column vectors of matrix $R^{r,N}$ given by* (21) *are linearly independent and the product of matrices $Q^{r,N} = \Upsilon^{r,N} R^{r,N}$ verify that the entries outside the main diagonal are nonpositive and also condition* **(HR)**, *where matrix $\Upsilon^{r,N}$ is given by* (13) *by adding superscript r,N in all the entries (except zeros and ones). Moreover, by taking the limit as $N \to +\infty$ in* (21)*, which does not depend on r, we introduce*

$$R \stackrel{\text{def}}{=} \lim_{N\to+\infty} R^{r,N} = \begin{pmatrix} 1 & -\frac{1}{s_{12}} & 0 & 0 & \cdots & \cdots & 0 \\ s_{12} & 0 & -\frac{1}{s_{23}} & 0 & \cdots & \cdots & 0 \\ s_{13} & 0 & 0 & -\frac{1}{s_{34}} & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ s_{1d-1} & 0 & 0 & 0 & \cdots & \cdots & -\frac{1}{s_{d-1d}} \\ s_{1d} & 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

and the column vectors of matrix R are linearly independent, and $Q = \Upsilon R$ verifies that the entries outside the main diagonal are nonpositive and also condition **(HR)** *too, where matrix Υ is given by* (13)*.*

Proof:

Indeed, we only show the justification for matrices R and Υ, since for their (r,N) counterparts, the proof is analogous. First of all we can see that the column vectors of R are linearly independent since its determinant is $\neq 0$. Then,

$$Q = \Upsilon R = \begin{pmatrix} 1 & -\frac{1}{s_{12}} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & -\frac{1}{s_{23}} & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & -\frac{1}{s_{34}} & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -\frac{1}{s_{d-1d}} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

which verifies that the entries of the main diagonal are all equal to 1 while the others are nonpositive numbers, as well as condition **(HR)** in Remark 1. Indeed, if Θ denotes the matrix obtained from $Q - I_d$ by replacing all the entries by their absolute value, we have that the spectral radius of matrix

$$\Theta = \begin{pmatrix} 0 & \frac{1}{s_{12}} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \frac{1}{s_{23}} & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & \frac{1}{s_{34}} & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \frac{1}{s_{d-1d}} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \text{ is } 0 < 1. \quad \square$$

Remark 5 In the main result of the paper (Theorem 1 below), we will prove that matrix R introduced in Lemma 2 plays the role of the matrix of directions of reflection on the boundary faces of a rfBm process on the convex polyhedron $S(\Upsilon)$, where Υ is introduced in (13). Then, the column vectors of matrix R are the reflection vectors on the boundary faces.

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 d-dimensional rfBm process in the convex polyhedron $S(\Upsilon)$, when N first and then r, tend to infinity in this order, under heavy-traffic, where Υ is given by (13) and $S = S(\Upsilon)$ is drawn in Figure 2 for $d = 3$, and with matrix of directions of reflection R given by Lemma 2.

Theorem 1 (heavy-traffic limit for the d-station tree-cascade fluid network)
Under the heavy-traffic condition **(HTd)** *and Assumption (s), there exist:*

$$\widehat{\widehat{W}}^{r} = \mathscr{D} - lim_{N\to+\infty} \widehat{W}^{r,N} \ (in\ \mathscr{D}^d) \quad and \quad W = \mathscr{D} - lim_{r\to+\infty} \widehat{\widehat{W}}^{r} \ (in\ \mathscr{C}^d),$$

and W is a d-dim. rfBm process on the convex polyhedron $S(\Upsilon)$ with associated data $(x = 0, \theta = \gamma, H, \Gamma, R)$, where $H \in (\frac{1}{2}, 1)$ is defined by (15)*, $\gamma \in \mathbb{R}^{d-1} \times \mathbb{R}^-$ is given by condition* **(HTd)***, and Γ is the symmetric matrix whose entries are:*

$$\Gamma_{jk} = \sigma^{2,\lim} \sum_{i=1}^{j} \frac{\alpha_i^2}{\mu_{ij}\,\mu_{ik}}, \quad 1 \le j \le k \le d \tag{26}$$

with the convention $\mu_{\ell\ell} = \mu_\ell$ for all $\ell = 1, \ldots, d$, and $\sigma^{2,\lim}$ given by (14)*.*

Proof:

The proof of this result is similar to that of Theorem 1 [5]. of which it is a generalization, and therefore we only highlight the main particularities.

The limit as $N \to \infty$.
In order to see that $(\widehat{W}^{r,N}, \widehat{X}^{r,N}, \widehat{V}^{r,N})_N$ verifies (i)-(iv) in Assumption (h) in the Appendix, we take into account that:

(i) First, $\widehat{W}_1^{r,N} \ge 0$ since $W_1^{r,N} = \widetilde{W}_1^{r,N} \ge 0$ by (5). Second, by (6), for any $j = 2, \ldots, d$, $\widehat{W}_j^{r,N}(t) \ge s_{j-1\,j}^{r,N} \widehat{W}_{j-1}^{r,N}(t)$, that is, $\widehat{W}^{r,N}(t) \in S^{r,N}$ for all $t \ge 0$.

(ii) follows from Lemma 1.

(iii) by (10) and (7) we have that

$$\begin{aligned}\widehat{V}_1^{r,N}(t) &= \widehat{Y}_1^{r,N}(t) + \sum_{\ell=2}^{d} \frac{1}{s_{1\ell}^{r,N}} \widehat{Y}_\ell^{r,N}(t) \\ &= \frac{\sqrt{N}\, r^{1-H}}{L^{1/2}(r)} \Big(\int_0^t 1_{\{\widehat{W}_1^{r,N}(s)=0\}}\, ds + \sum_{\ell=2}^{d} \frac{1}{s_{1\ell}^{r,N}} \int_0^t 1_{\{\widehat{W}_1^{r,N}(s)=\cdots=\widehat{W}_\ell^{r,N}(s)=0\}}\, ds \Big),\end{aligned}$$

and therefore, $\widehat{V}_1^{r,N}$ only can increase if $\widehat{W}_1^{r,N}=0$, that is,

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

using that $F_1^{r,N}=\{(x_1,\dots,v^d))\in S^{r,N}: x_1=0\}$. Analogously, for $j=2,\dots,d-1$ we have by (11), (7) and (9) that

$$\widehat{V}_j^{r,N}(t)=\big(\widehat{Y}_j^{r,N}(t)+\sum_{h=1}^{j-1}\widehat{T}_{hj}^{r,N}(t)\big)+\sum_{\ell=j+1}^{d}\frac{1}{s_{j\ell}^{r,N}}\big(\widehat{Y}_\ell^{r,N}(t)+\sum_{h=1}^{j-1}\widehat{T}_{h\ell}^{r,N}(t)\big)=$$
$$\frac{\sqrt{N}r^{1-H}}{L^{1/2}(r)}\Big(\big(\int_0^t 1_{\{\widehat{W}_1^{r,N}(s)=\cdots=\widehat{W}_j^{r,N}(s)=0\}}ds+\sum_{h=1}^{j-1}\int_0^t 1_{\{\widehat{\widetilde{W}}_h^{r,N}(s)>0\widehat{\widetilde{W}}_j^{r,N}(s)=0\}\cap\widehat{\widetilde{\tau}}_{hj}^{r,N}(s)}ds\big)$$
$$+\sum_{\ell=j+1}^{d}\frac{1}{s_{1\ell}^{r,N}}\big(\int_0^t 1_{\{\widehat{W}_1^{r,N}(s)=\cdots=\widehat{W}_\ell^{r,N}(s)=0\}}ds+\sum_{h=1}^{j-1}\int_0^t 1_{\{\widehat{\widetilde{W}}_h^{r,N}(s)>0,\widehat{\widetilde{W}}_\ell^{r,N}(s)=0\}\cap\widehat{\widetilde{\tau}}_{h\ell}^{r,N}(s)}ds\big)\Big),$$

where $\widehat{\widetilde{\tau}}_{ij}^{r,N}(\cdot)$ is defined analogously to $\widetilde{\tau}_{ij}^{r,N}(\cdot)$ but substituting $\widetilde{W}$ by $\widehat{\widetilde{W}}$.

Then, for $j=2,\dots,d-1$, $\widehat{V}_j^{r,N}$ is only allowed to increase if $\widehat{\widetilde{W}}_j^{r,N}=0$, that by (6) implies that

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

since $F_j^{r,N}=\{(x_1,\dots,x_d)\in S^{r,N}: x_j=s_{j-1\,j}^{r,N}x_{j-1}\}$. Finally, if $j=d$, by (12), (7) and (9),

$$\widehat{V}_d^{r,N}(t)=\widehat{Y}_d^{r,N}(t)+\sum_{h=1}^{d-1}\widehat{T}_{hd}^{r,N}(t)=$$
$$\frac{\sqrt{N}r^{1-H}}{L^{1/2}(r)}\Big(\big(\int_0^t 1_{\{\widehat{W}_1^{r,N}(s)=\cdots=\widehat{W}_d^{r,N}(s)=0\}}ds+\sum_{h=1}^{d-1}\int_0^t 1_{\{\widehat{\widetilde{W}}_h^{r,N}(s)>0,\widehat{\widetilde{W}}_d^{r,N}(s)=0\}\cap\widehat{\widetilde{\tau}}_{hd}^{r,N}(s)}ds\big),$$

which implies by similar arguments that

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

(iv) It is a consequence of the fact that by the heavy-traffic condition **(HTd)** we can ensure the existence of $\widehat{\widehat{X}}^r=\mathscr{D}-\lim_{N\to+\infty}\widehat{X}^{r,N}$, with $\widehat{\widehat{X}}^r=(\widehat{\widehat{X}}_1^r,\dots,\widehat{\widehat{X}}_d^r)^T$ given by

$$\widehat{\widehat{X}}_1^r(t)=\frac{\widehat{\widehat{E}}_1^r(t)}{\mu_1}+\frac{r^{1-H}}{L^{1/2}(r)}\,\widehat{\gamma}_1^r t\,,$$
$$\widehat{\widehat{X}}_j^r(t)=\sum_{i=1}^{j-1}\frac{\widehat{\widehat{E}}_i^r(t)}{\mu_{ij}}+\frac{\widehat{\widehat{E}}_j^r(t)}{\mu_j}+\frac{r^{1-H}}{L^{1/2}(r)}\,\widehat{\gamma}_j^r t,\quad j=2,\dots,d,$$

that implies the continuity of the paths of process $\widehat{\widehat{X}}^r$.

Moreover, hypothesis (b) in Proposition 1 is accomplished by Lemma 2, and as a consequence, in $\mathscr{D}^d$ there exists

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

and $(\widehat{\widehat{W}}^{r},\widehat{\widehat{V}}^{r})$ 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 limit as $r\to\infty$.

By **(HTd)**, there exists the limit $\mathscr{D}-\lim_{r\to+\infty}\widehat{\widehat{X}}^{r}=X=(X_1,\ldots,X_d)^T$, with

$$X_1(t)=\frac{B_1^H(t)}{\mu_1}+\gamma_1 t,\quad X_j(t)=\sum_{i=1}^{j-1}\frac{B_i^H(t)}{\mu_{ij}}+\frac{B_j^H(t)}{\mu_j}+\gamma_j t,\quad j=2,\ldots,d,$$

which is a d-dimensional fBm process with associated data $\left(x=0,\theta=\gamma,H,\Gamma\right)$, where Γ is the $d\times d$ positive definite matrix given by (26).

In addition, $\lim_{N\to+\infty}R^{r,N}=R$, independent of r, and by Lemma 2, $Q=\Upsilon R$ satisfies assumption **(HR)**. Then, by Proposition 1 again, there exists

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

where $W=X+RV$ is a d-dimensional rfBm process on the convex polyhedron $S(\Upsilon)$ with associated data $(x=0,\theta=\gamma,H,\Gamma,R)$. □

5 Appendix: An Invariance Principle for rfBm processes living in convex polyhedra

Kang and Williams prove in Theorem 4.3 [11] 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, and a crucial ingredient in its proof is an oscillation inequality for solutions of a perturbed Skorokhod problem (Theorem 4.1 [11]). This invariance principle is used in [11], in particular, to give sufficient conditions for validating approximations involving SRBMs in convex polyhedra with a constant reflection vector field on each face. As showed in Lemma 4 [4], this principle does not depend on the specific law of the processes and can be applied to the rfBm process instead of SRBM.

In this section we recall a different version of this principle stated in [5], also applied to the rfBm process, as in [4], but considering a sequence of convex polyhedra. This result gives sufficient conditions for validating approximations involving rfBm processes in convex polyhedra with a constant reflection vector field on each face, in such a way the approximating processes live in a sequence of convex polyhedra. This sequence of convex polyhedra approximates the convex polyhedron in which the limit rfBm process lives.

Remark 6 Results reported in this section are based on the following assumptions on a convex polyhedron $S = S(\Upsilon)$ on $\mathbb{R}^d$ with matrix of directions of reflection R (see Kang and Williams [11]):

(A1) $S = \cap_{j=1}^{d} G_j$ with $\emptyset \neq G_j \neq \mathbb{R}^d$ and $\partial G_j = F_j$ is C^1 for each $j = 1, \ldots, d$. (For the definition of what the feature C^1 of a boundary means, we refer the reader to Section 1.1 [11].)

(A2) For each $\varepsilon \in (0,1)$ there exists $R(\varepsilon) > 0$ such that for each $j = 1, \ldots, d$, $x \in F_j$ and $y \in S$ satisfying $||x-y|| < R(\varepsilon)$, we have that $\langle n^j, y-x \rangle \geq -\varepsilon ||x-y||$.

(A3) The function $D : [0,\infty) \to [0,\infty]$ defined such that $D(0) = 0$ and for $r > 0$, $D(r) = \sup_{\emptyset \neq \mathscr{J} \subset \{1,\ldots,d\}} \sup\{d\big(x, \cap_{j \in \mathscr{J}} F_j\big) : x \in \cap_{j \in \mathscr{J}} U_r(F_j)\}$ satisfies $\lim_{r \to 0} D(r) = 0$.

(A4) If $\{u^\ell(\cdot)\}_{\ell=1,\ldots,d}$ are the reflection vector fields, there is a constant $L > 0$ such that for each $\ell = 1, \ldots, d$, $u^\ell(\cdot)$ is uniformly Lipschitz continuous function from $\mathbb{R}^d$ into $\mathbb{R}^d$ with Lipschitz constant L and $||u^\ell(x)|| = 1$ for each $x \in \mathbb{R}^d$.

(A5) There is a constant $a \in (0,1)$ and there are vector valued functions $b(\cdot) = (b_1(\cdot), \ldots, b_d(\cdot))^T$ and $c(\cdot) = (c_1(\cdot), \ldots, c_d(\cdot))^T$ from ∂S into $\mathbb{R}^d_+$ such that for each $x \in \partial S$,

(i) $\sum_{i \in \ell(x)} b_i(x) = 1$, $\min_{j \in \ell(x)} \langle \sum_{i \in \ell(x)} b_i(x)\, n^i, v^j \rangle \geq a$,
(ii) $\sum_{i \in \ell(x)} c_i(x) = 1$, $\min_{j \in \ell(x)} \langle \sum_{i \in \ell(x)} c_i(x)\, v^i, n^j \rangle \geq a$,

When the reflection vector fields are constant at each face, as in our case, assumption (A4) holds trivially, while assumption (A5) is equivalent to the most easily verifiable Assumption 5.1 [11], reproduced here for convenience of the reader:

Assumption 5.1 [11]: For each *maximal* $\mathscr{K} \subset \{1, \ldots, d\}$ (that is, $\mathscr{K} \neq \emptyset$ with $F_\mathscr{K} \neq \emptyset$ and $F_\mathscr{K} \neq F_\mathscr{L}$ for any $\mathscr{L} \supset \mathscr{K}$ such that $\mathscr{L} \neq \mathscr{K}$, where $F_\mathscr{K}$ denotes $\cap_{\ell \in \mathscr{K}} F_\ell$),

(S.a) there is a positive linear combination $u = \sum_{i \in \mathscr{K}} b_i u^i$ with $b_i > 0$, such that $\langle u, v^i \rangle > 0$ for all $i \in \mathscr{K}$,
(S.b) there is a positive linear combination $v = \sum_{i \in \mathscr{K}} c_i v^i$ with $c_i > 0$, such that $\langle u^i, v \rangle > 0$ for all $i \in \mathscr{K}$.

The invariance principle (Proposition 1 below) requires the following additional assumption, which is a version of the Assumption 4.1 in Kang and Williams [11] (see [5]):

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.*

We can state the following invariance principle, which is a version of Theorem 4.3 [11], and is proved in [5]:

Proposition 1 (invariance principle) *Suppose that* Assumption (h) *and assumptions* (A1)-(A5) *hold. Then, the sequence of processes* $\{(W^n, X^n, V^n)\}_n$ *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* $S(\Upsilon)$ *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* $S(\Upsilon)$ *with associated matrix of directions of reflection* R *has a unique strong solution,*

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

References

1. R. Delgado, A reflected fBm limit for fluid models with ON/OFF sources under heavy-traffic, *Stochastic Processes and Their Applications* **117**, 188-201 (2007).
2. R. Delgado, State space collapse for asymptotically critical multi-class fluid networks, *Queueing Systems* **59**, 157-184 (2008).
3. R. Delgado, Heavy-traffic limit for a feed-forward fluid model with heterogeneous heavy-tailed On/Off sources, *Queueing Systems* **74(1)**, 41-63 (2013).
4. R. Delgado, A packet-switched network with On/Off sources and a fair bandwidth sharing policy: state space collapse and heavy traffic, *Telecommunication Systems* **62(2)**, 461-479 (2016).
5. R. Delgado, A two-queue polling model with priority on one queue and heavy-tailed On/Off sources: a heavy-traffic limit, *Queueing Systems* **83(1)**, 57-85 (2016).
6. R. Delgado, E. Morozov, Stability analysis of cascade networks via fluid models. Performance Evaluation **82**, 39-54 (2014).
7. R. Delgado, E. Morozov, Stability analysis of some networks with interacting servers. ASMTA 2014, LNCS 8499, 1-15 (2014).
8. P. Dupuis, K. Ramanan, Convex duality and the Skorokhod Problem I, *Probability Theory and Related Fields* **115**, 153-195 (1999).
9. P. Dupuis, K. Ramanan, Convex duality and the Skorokhod Problem II, *Probability Theory and Related Fields* **115**, 197-236 (1999).
10. M. Harrison, Heavy-traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review pollcies, *Ann. Appl. Probab.* **8(3)**, 822-848 (1998).
11. W. N. Kang, R. J. Williams, An invariance principle for semimartingale reflecting Brownian motions in domains with piecewise smooth boundaries, *Ann. Appl. Probab.* **17**, 741-779 (2007).
12. T. Konstantopoulos, S. J. Lin, Fractional Brownian approximations of stochastic networks, in: Stochastic Networks, Stability and Rare Events, Lecture Notes in Statistics **117**, 257-274 (1996).
13. M. S. Taqqu, W. Willinger, R. Sherman, Proof of a fundamental result in self-similar traffic modeling, *Comput. Commun. Rev.* **27**, 5-23 (1997).