Queueing Syst (2008) 59: 157–184
DOI 10.1007/s11134-008-9080-y

State space collapse for asymptotically critical multi-class fluid networks

Rosario Delgado

Received: 16 November 2007 / Revised: 7 July 2008 / Published online: 12 August 2008

Abstract We consider a class of fluid queueing networks with multiple fluid classes and feedback allowed, which are fed by N heavy tailed ON/OFF sources. We study the asymptotic behavior when $N \to \infty$ of these queueing systems in a heavy traffic regime (that is, when they are asymptotically critical). As performance processes we consider the workload W^N (the amount of time needed for each server to complete processing of all the fluid in queue), and the fluid queue Z^N (the quantity of each fluid class in the system). We show the convergence of $\sqrt{N} W^N$ and $\sqrt{N} Z^N$ (to $\hat{W}$ and $\hat{Z}$) in heavy traffic if *state space collapse* (SSC) holds. (SSC) is a condition that establishes a relationship between those components of $\hat{Z}$ that correspond to fluid classes processed by the same server, which implies that $\hat{Z} = \Delta \hat{W}$ for a deterministic *lifting* matrix Δ. Our main contribution is to prove that assuming that the other hypotheses are true, (SSC) is not only sufficient for this convergence, but necessary. Furthermore, we prove that processes $\hat{W}$ and $\hat{Z}$, conveniently scaled in time, converge to W (a reflected fractional Brownian motion) and Z $(= \Delta W)$. We illustrate the application of our results with some examples including a tandem queue.

Keywords State space collapse · Reflected fractional Brownian motion · Fluid model · Multi-class queueing network · Workload process · Fluid queue process · On-off sources · Heavy traffic · *Completely-S* matrix · Tandem queue

Mathematics Subject Classification (2000) Primary 60K25 · Secondary 60F05 · 68F17 · 60G15 · 90B22

Supported by project MEC-FEDER ref. MTM2006-06427.

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

1 Introduction

In this paper we investigate the asymptotic behavior of a family of multi-class fluid networks in a heavy traffic regime (that is, when they are asymptotically critical). The networks of this family consist of a different number of fluid classes, K, and stations, J, with $K \geq J \geq 1$, and they generalize the non-deterministic fluid model considered in [7], where $K = J$. We assume a non-idling FIFO (first-in-first-out) service discipline, and fluid classes processed at constant processing rates, $1/m_k$ for class k. Each fluid class can be served at only one station, and at each station there is a single server and an infinite buffer where fluid waiting to be processed accumulates. Moreover, feedback is allowed. The structure of these networks is that of the deterministic fluid analog of a multi-class queueing network introduced by Harrison [8].

Nevertheless, while in [8] the external arrival process is deterministic, ours is considered to be a non-deterministic aggregated cumulative process generated by a big number N of heavy tailed ON/OFF sources. It is known from [17] that the superposition of many ON/OFF sources with strictly alternating ON- and OFF-periods and whose ON- or OFF-period lengths are heavy tailed is a good model for modern high-speed network traffic, which exhibits long-range dependence and self-similar traffic patterns. The reason, as proved in Theorem 1 [17], is that the superposition of N of these ON/OFF sources generates an aggregate cumulative arrival process that conveniently scaled in time by a factor r, and in state space, converges in some sense, as N goes to infinity and after that, as r goes to infinity (note that order here is important), to a fractional Brownian motion (fBm). And it is known too that fBm is self-similar and when its Hurst parameter H is bigger than $1/2$, it has positively correlated increments. See [12] and references therein for more insight into the interest in dealing with self-similar and long-range dependent models when modeling modern high-speed network traffic.

As main performance processes, the (J-dimensional) immediate *workload process* W^N and the (K-dimensional) *fluid queue process* Z^N are introduced. For a server j, the corresponding component of the workload process at time t, $W_j^N(t)$, is defined to be the total quantity of time this server needs to complete the processing of all the fluid in queue (or being served) at station j at time t. For a fluid class k, the corresponding component of the fluid queue process at time t, $Z_k^N(t)$, is defined as the quantity of that class fluid in queue (or being served) at time t. There is an obvious relationship between these two processes: for any station j,

$$W_j^N = \sum_{k \in \{\text{classes served at station } j\}} m_k Z_k^N. \tag{1.1}$$

Debicki and Mandjes have showed in [6] that the convergence of the aggregated cumulative arrival process to the fBm given in [17] carries over to the stationary buffer content process in a heavy-traffic environment (that is, in the asymptotically critical situation): the (scaled) workload process converges to the (uni-dimensional) fBm, reflected appropriately to be non-negative, for single-class fluid models with only one station and without reentering (that is, in the case $K = J = 1$ and without feedback). In [7] the result of [6] was generalized to a multidimensional setting for a multi-server fluid model with feedback in which every server can only process one fluid class, that is, with $K = J \geq 1$: the (scaled) workload process was proved to converge to a J-dimensional reflected fractional Brownian motion (rfBm) on the first orthant. In the present work we generalize still more this result to the multi-class situation, in which each server can process one or more fluid classes (but, as usual, each fluid class can be processed only by one of the servers, and then $K \geq J \geq 1$). Multidimensional rfBm in the first orthant had been introduced, among others, by Konstantopoulos and Lin [12], although throughout this paper we use notations from [7], where the question of what condition on the reflection matrix ensures the existence of such a process is also considered (see the Appendix at the end here).

Semimartingale reflecting Brownian motions in the first orthant (SRBM) are diffusion processes that have been used as approximate models of multi-class open queueing networks, in a light-tailed setting, under the heavy traffic assumption and different service disciplines (including FIFO) by many authors, starting with Iglehart and Whitt [10, 11] who considered single-class networks without feedback and with FIFO service discipline. This was generalized to single-class networks with feedback by Reiman [15], and to feedforward multi-class networks by Peterson [14], where limit theorems to justify the diffusion approximations have been proved. See [19] for a summary of the heavy traffic limit theorems of this type for different single-class and multi-class networks up to date. Heavy traffic limits for multi-class queueing networks is a topic of much interest although the class of networks for which these limits have been proved is still small and it is known that not all multi-class networks with feedback can be approximated under heavy traffic by such reflecting processes (see [5], for instance). General approximation schemes can be seen in [9], but they are not always valid. A rigorous theory for multi-class networks in the light-tailed setting under heavy traffic and with feedback remained to be carried out when Williams' paper [21] appeared. In [21] Williams proves a generalization of the previous heavy traffic limit results to multi-class networks with feedback, by using an invariance principle proved by the same author in [20]. It was a first step in the direction of developing a theory by giving general sufficient conditions for a heavy traffic limit theorem to hold for open multi-class queueing networks with some disciplines (including FIFO), in the light-tailed setting, a SRBM in the first orthant being the limit process. These sufficient conditions are: the reflection matrix has to be well defined and *completely-S*, and a form of *state space collapse* must hold.

By combining the previous ideas and methodology developed by Williams among others, with the convergence results for heavy-tailed ON/OFF traffic to fractional Brownian motion of Taqqu et al. [17] already used in the previous work [7], in this paper we try to contribute to the development of a similar theory for multi-class fluid networks with feedback in the heavy-tailed setting, by obtaining sufficient conditions to ensure a heavy traffic limit result, where the limit process is a multidimensional rfBm in the first orthant (see Corollary 1).

More specifically, in Theorem 1 we prove that when the number of ON/OFF sources N converges to ∞, the limit of $\sqrt{N}W^N$, say $\hat{W}$, and the limit of $\sqrt{N}Z^N$, $\hat{Z}$, exist (both in the sense of the convergence of the finite-dimensional distributions). The result can be proved under four hypotheses, two of which are related to matrices defined from the model parameters: condition (HΔ), which refers to a $K \times J$ matrix Δ that is a *lifting* operator from $\mathbb{R}^J$ to $\mathbb{R}^K$ which relates $\hat{Z}$ and $\hat{W}$ as follows: $\hat{Z} = \Delta\hat{W}$, and condition (HR) on a $J \times J$ matrix R, related to Δ, which ensures the existence of the reflected fractional Brownian motion process with R as reflection matrix (and also that R is a *completely-S* matrix). See the Appendix for more details on assumption (HR) and for the definition of the multidimensional rfBm process. The other two hypotheses are: a *heavy traffic* condition, denoted by (HT) (see (3.10)), which states that the network is asymptotically critical, that is, its traffic intensity tends to be one (for any station) as N goes to infinity, and a kind of *state space collapse* condition denoted by (SSC) (see (3.12)).

The phenomenon of *state space collapse* was first established by Whitt in [18] for the single multi-class station but the *term* was first introduced by Reiman [16]. *State space collapse* condition has proved to be a key ingredient in the proof of heavy traffic limits for multi-class queueing networks in the light-tailed environment. See for instance [14] for feedforward multi-class queueing networks. In [4] Bramson proves that a form of *multiplicative state space collapse* holds for two families of multi-class networks (FIFO networks of Kelly type and head-of-the-line proportional processor sharing queueing networks), and by using that fact, Williams proves in [21] that under the heavy traffic condition and the *completely-S* assumption for the reflecting matrix, the *multiplicative state space collapse* condition implies (SSC), and a heavy traffic limit theorem holds in a light-tailed environment, a multidimensional SRBM in the first orthant being the workload limit process.

In this work we show that condition (SSC) also plays a key role in demonstrating a similar heavy traffic limit in our heavy-tailed environment for multi-class fluid networks, with a multidimensional rfBm as workload limit. (SSC) establishes a restriction in process $\hat{Z}$ in the sense that some relationships between those components corresponding to fluid classes processed at the same station must be satisfied; these relationships are established by means of some parameters of the model: the service rates m_k and the long run fluid rates into and out of stations for each class, λ_k. Roughly speaking we can say that from (SSC), with the knowledge of the workload process we do not need any additional information about the fluid queue process, because both processes are linked by means of a deterministic lifting operator Δ in this way:

$$\hat{Z} = \Delta\hat{W}, \quad \text{or} \quad \hat{Z}_k = \lambda_k \hat{W}_j \quad \text{for any fluid class } k \tag{1.2}$$

if j is the station that processes that class (see (6.4)). It is interesting to compare (1.2) to (1.1). If $K = J$ *state space collapse* condition vanishes, but for $K > J$ it does not. Indeed, by assuming that the other three hypotheses are true, this condition is sufficient and, what is more important, also necessary for the conclusion of Theorem 1, that is, it cannot be weakened nor dropped.

In Corollary 1 we present our heavy traffic limit result, which is a generalization to the multi-class setting of Theorem 1 in [7], by considering the processes obtained from the limit processes $\hat{W}$ and $\hat{Z}$ scaling in this way:

$$\frac{\hat{W}(r\cdot)}{r^H f(r)} \quad \text{and} \quad \frac{\hat{Z}(r\cdot)}{r^H f(r)}, \quad \text{where } r \text{ is the scaling factor },$$

f is a slowly varying at infinity function and $H \in (1/2, 1)$ is an adequate constant. We prove that these two processes converge respectively, as r goes to ∞, in the distributional sense, to respective processes W (which is a rfBm process of Hurst parameter H) and Z $(= \Delta W)$.

Two main ingredients in the proofs of these results are the *oscillation inequality* given by Bernard and el Kharroubi [1] (Lemma 1) for Theorem 1, and for Corollary 1, the *invariance principle* given by Williams [20] (Theorem 4.1) for the reflected Brownian motion process, which can also be applied to the reflected fractional Brownian motion process.

In a paper based on [12], Majewski [13] considers multi-class feedforward queueing networks with priority service discipline and FIFO within each priority class, driven by long-range dependent arrival and service time processes where feedback is not allowed: stations are numbered in such a way that a customer leaving a queue is routed to the next one, so the departures of the ith queue are the arrivals of the $(i+1)$th. Majewski proves a heavy traffic limit: given that cumulative arrival and service time processes approach heavy traffic in such a way that the corresponding normalized processes converge to a fractional Brownian motion, the scaled workload and queue length processes converge to multidimensional rfBm. Instead, in this paper we assume that the external arrival process for the multi-class fluid network is a non-deterministic aggregated cumulative process generated by a big number of heavy tailed ON/OFF sources, and that each fluid class is processed at a constant rate by using a FIFO service discipline. Moreover, the structure of the multi-class network allowing feedback is more complex and rich than that of [13], and can be adapted to several interesting examples, as Sect. 6 illustrates.

The importance of Theorem 1 and Corollary 1 in the present paper lies in the fact that they show the role played by condition (SSC) (which is redundant if $K = J$, so it did not appear in [7]) for proving the convergence, under heavy traffic, of the workload and fluid queue processes for a multi-class queueing fluid network under a FIFO service discipline with feedback, in the heavy-tailed environment, and that the workload limit process is a multidimensional rfBm. These results allow us to analyze a wide variety of real situations, as the examples considered in Sect. 6 show, where the more explicit form of condition (SSC) given by (6.2) is used. These examples are: a two-stage queueing system or tandem queue with feedback, considered to be one of the canonical "building blocks" in modern high speed communication networks (Sect. 6.1), and a network with a traffic stream and a $\bigvee$-system (a multi-class network with a single server), both with feedback allowed, in Sect. 6.3.

The organization of the rest of the paper is as follows. In Sect. 2 we set up notation and terminology. Definition of the reflected fractional Brownian motion process and assumption on the reflection matrix R (HR) are presented at the end, in the Appendix. The multi-class fluid network we consider is introduced in Sect. 3, where performance processes, model equations and the rest of assumptions are given. Section 4 presents scaled processes and the main results (Theorem 1 and Corollary 1). Section 5 deals with a kind of *multiplicative state space collapse*, which is a condition apparently weaker than that of *state space collapse*; by using that R is a *completely-S* matrix, which is a consequence of condition (HR), we show in Proposition 1 that they are, in fact, equivalent conditions (see Williams [21] and Bramson [3, 4] for the introduction of this kind of multiplicative condition in relation with state space collapse and heavy traffic limits).

2 Some definitions and terminology

For the convenience of the reader in this section we introduce some definitions and notations from [7], making our exposition self-explanatory.

For each integer $d \geq 1$, we will denote by I_d the d-dimensional identity matrix. Vectors will be column vectors and v^T means the transpose of a vector (or a matrix) v. Given $v = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, hereafter we will denote by $\mathrm{diag}(v)$ (or, equivalently, by $\mathrm{diag}(v_1, \ldots, v_d)$) the $d \times d$ diagonal matrix with diagonal elements $v_1, \ldots, v_d$. Let S denote the d-dimensional first orthant

$$S = \mathbb{R}^d_+ = \{v = (v_1, \ldots, v_d)^T \in \mathbb{R}^d : v_i \geq 0 \quad \forall i = 1, \ldots, d\}.$$

For a $d \times d'$ matrix $A = (a_{ij})_{i=1,\ldots,d, j=1,\ldots,d'}$, let $|A| \overset{\text{def}}{=} \max_{1 \leq j \leq d'}(\sum_{1 \leq i \leq d} |a_{ij}|)$. We will say that a sequence of $d \times d'$ matrices $\{A^n\}_n$ converges to a $d \times d'$ matrix A if $|A^n - A| \to 0$ as n tends to ∞ (this convergence is equivalent to the convergence in the componentwise sense), and we will denote it simply $\lim_{n\to\infty} A^n = A$. The same applies for the particular case $d' = 1$, which corresponds to d-dimensional vectors, with $|v| \overset{\text{def}}{=} \sum_{1 \leq i \leq d} |v_i|$.

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

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

We will say that $\omega^n \to \omega$ as $n \to \infty$ in $\mathcal{C}^d$ (*uniformly on compacts*) if for any $T \geq 0$, $\|\omega^n(\cdot) - \omega(\cdot)\|_T \to 0$, and we will denote it $\lim_{n\to\infty} \omega^n = \omega$. To measure the *oscillation* of ω we make the following definition: for any $T \geq 0$,

$$\mathrm{Osc}\left(\omega(\cdot), [0, T]\right) \overset{\text{def}}{=} \sup_{0 \leq s < t \leq T} |\omega(t) - \omega(s)| = \sup_{0 \leq s < t \leq T} \left(\sum_{1 \leq \ell \leq d} |\omega_\ell(t) - \omega_\ell(s)| \right).$$

Note that, in general, $\mathrm{Osc}(\omega(\cdot), [0, T]) \leq 2\|\omega(\cdot)\|_T$ and that $\mathrm{Osc}(\omega(\cdot), [0, T]) = \|\omega(\cdot)\|_T$ if $\omega(0) = 0$ and $\omega_\ell(t) \geq 0$ for any $t \in [0, T]$ and any $1 \leq \ell \leq d$.

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

$\mathcal{D}$-lim for the *convergence in distribution* on $\mathcal{C}^d$ (or *weak convergence*), P-lim for the *convergence in probability (uniformly on compacts)*, and $\widetilde{\lim}$ for the *convergence of the finite-dimensional distributions*.

3 The multi-class fluid network

3.1 Introducing the model

In this section we present the family of multi-class fluid networks we deal with: those whose input traffic alternates between heavy-tailed ON and OFF-periods, by following the notation of Williams [21].

We consider a network composed of J stations with a single server that processes continuous fluid, and an infinite buffer, at each one, and we distinguish among fluid of classes $1, \dots, K$, with $K \geq J$. Each fluid class can be processed by only one station but each station can process more than one class, and the many-to-one mapping from fluid classes to stations is denoted by s, $s : \{1, \dots, K\} \longrightarrow \{1, \dots, J\}$, $s(k)$ being the station where class k fluid is processed. We can also introduce the $J \times K$ (deterministic) *constituency matrix* $C = (C_{jk})$ by

$$C_{jk} \overset{\text{def}}{=} \begin{cases} 1 & \text{if } j = s(k), \\ 0 & \text{otherwise.} \end{cases}$$

For any j, $s^{-1}(j)$ is the *constituency of station* j, that is, the set of fluid classes that are served at that station.

This model is a generalization of that introduced in [7], where $K = J$ was assumed, and therefore matrix C was the identity (that is, each server could only process one fluid class).

We follow [7] in assuming that for each fluid class, say k, there are N i.i.d. ON/OFF sources, each one with its own 0/1-valued jump process $\{U_k^{(n)}(t), t \geq 0\}, n = 1, \dots, N$, on a common probability space, and that they are all independent. $U_k^{(n)}(t) = 1$ means that at time t source n of fluid class k is ON (and it is sending fluid to the network, at a deterministic traffic rate $\alpha_k^N \geq 0$), and $U_k^{(n)}(t) = 0$ means that it is OFF. We suppose that, independently of k, the lengths of the ON-periods are i.i.d., those of the OFF-periods are i.i.d., and the lengths of the ON- and OFF-periods are mutually independent. The ON- and OFF-period lengths may have different distributions but at least one of them must be heavy-tailed. Their respective expected values and variances are $\nu_{\text{on}}, \sigma_1^2$ for the ON- and $\nu_{\text{off}}, \sigma_2^2$ for the OFF-periods, and their respective density functions f_1 and f_2 are supposed to verify that as $x \to \infty$,

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

with L_1 and L_2 being positive slowly varying functions at infinity, and $1 < \beta_1, \beta_2 \leq 2$. We assume that both ν_{on} and ν_{off} are finite numbers but that σ_1^2 or σ_2^2 are infinite. If $\sigma_i^2 < +\infty$, then $\beta_i = 2$ and if $\sigma_i^2 = +\infty$, $\beta_i < 2$.

If all sources where ON, class k fluid would arrive at deterministic rate $\alpha_k^N \geq 0$, and the cumulative *external fluid traffic* up to time t would be deterministic and equal to $\alpha_k^N t$ (this was the case for the fluid model introduced by Harrison in [8]). Instead, as in [7], we define the cumulative external class k fluid generated up to time t (by the N sources) in this way:

$$E_k^N(t) \overset{\text{def}}{=} \alpha_k^N \int_0^t \frac{1}{N} \left(\sum_{n=1}^N U_k^{(n)}(u) \right) du.$$

The K-dimensional (non-deterministic) *aggregated cumulative external fluid traffic* process is $E^N = \{E^N(t) = (E_1^N(t), \ \dots, \ E_K^N(t))^T \mid t \geq 0\}$, whose component processes are all independent. We suppose, for the sake of simplicity, that at time

$t=0$ there is no accumulated fluid at the network (that is, $E^N(0)=0$). Let $\alpha^N = (\alpha_1^N,\dots,\alpha_K^N)^T$.

From now on we make the assumption that fluid at each server is processed in a first-in-first-out (FIFO) discipline, there is no capacity restriction in buffers, and service discipline is under a *non-idling* (or *work-conserving*) policy, that means that a server is never idle when there is fluid waiting to be processed at its station.

Suppose that class k fluid is processed at a constant rate $\mu_k > 0$ (independent of N) if station $s(k)$ were never idle and the server devoted all of its attention to class k. Let $m_k = 1/\mu_k$ be the *mean service rate* for class k fluid, $m=(m_1,\dots,m_K)^T$ and $\mu=(\mu_1,\dots,\mu_K)^T$.

Let $P_{k\ell}$ be the proportion of class k fluid that leaving station $s(k)$ goes next to station $s(\ell)$ as class ℓ fluid. We assume that for each k, $\sum_{\ell=1}^K P_{k\ell} \le 1$ and $1-\sum_{\ell=1}^K P_{k\ell} \ge 0$ is the proportion of class k fluid that leaving station $s(k)$ goes outside the network. Thus, $P=(P_{k\ell})_{k,\ell=1}^K$ is a sub-stochastic matrix. It is called the *"flow"* or *"routing" matrix of the network*, and it is assumed to have spectral radius less than one. Hence, $Q \overset{\text{def}}{=} (I_K - P^T)^{-1}$ is well defined.

3.2 Performance processes and model equations

The following descriptive processes Z^N, W^N and Y^N will be used to measure the performance of the queueing fluid network:

The *fluid queue process* Z^N is a K-dimensional process defined by: $Z_k^N(t)$ is the class k fluid that is in queue or being processed (at station $s(k)$) at time t. We assume that $Z^N(0)=0$.

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

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

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

Fluid queue and immediate *workload* processes measure the congestion and delay in the network, while *idle-time* process measures utilization of resources.

We further define some other additional processes associated to the fluid model, namely A^N, D^N, F^N and L^N, that will be useful in proofs.

A^N and D^N are K-dimensional processes defined by: $A_k^N(t)$ is the total class k fluid arriving to server $s(k)$ up to time t, including both feedback flow and external input, and $D_k^N(t)$ is the total amount of class k fluid departing station $s(k)$ (both being routed to other stations or departing the network), up to time t. We assume that $A^N(0)=D^N(0)=0$. By definition, we have that if t and h are non-negative, for any k,

$$D_k^N(t+h) - D_k^N(t) \le h\mu_k. \tag{3.1}$$

F^N is another K-dimensional process defined by $F^N \overset{\text{def}}{=} P^T D^N$, that is, $F_\ell^N(t) = \sum_{k=1}^K P_{k\ell} D_k^N(t)$ is the total amount of class ℓ fluid that is fed back to station $s(\ell)$ (due to the fraction of the amount $D_k^N(t)$ of class k fluid that leaving station $s(k)$ is next routed to station $s(\ell)$ as class ℓ fluid, summed over all fluid classes).

Finally, L^N is a J-dimensional process defined by $L^N \overset{\text{def}}{=} CMA^N$ where $M \overset{\text{def}}{=} \text{diag}(m_1, \dots, m_K)$. Since server $s(k)$ requires an amount of time $m_k a$ to process a quantity a of fluid of class k, the amount of time that a server, say j, would need in order to process the total quantity of fluid arriving in its station by time t equals $L_j^N(t) = \sum_{k \in s^{-1}(j)} m_k A_k^N(t)$.

Processes E^N, A^N, D^N, W^N, Z^N, Y^N, F^N and L^N are related by the following *model equations*: for any $t \geq 0$,

$$A^N(t) = E^N(t) + F^N(t), \tag{3.2}$$

$$W^N(t) = L^N(t) - et + Y^N(t), \tag{3.3}$$

$$Z^N(t) = A^N(t) - D^N(t), \tag{3.4}$$

$$D^N(t + C^T W^N(t)) = A^N(t), \tag{3.5}$$

$$W^N(t) = CMZ^N(t), \tag{3.6}$$

where $e = 1 \in \mathbb{R}^J$. The interpretation of these equations is clear: in (3.3), $W_j^N(t)$ is the amount of time required for server j to complete processing of all fluid buffered or being served at station j at time t, which equals $L_j^N(t)$ (the cumulative total amount of time required for server j to complete processing of fluid arrived at station j up to time t) minus the amount of time, $t - Y_j^N(t)$, that the server at station j has been busy (working) up to time t. Equation (3.5) is

$$D_k^N(t + W_{s(k)}^N(t)) = A_k^N(t), \quad \text{for all } k = 1, \dots, K,$$

and reflects the fact that we are assuming a FIFO service discipline.

Remark 1 Therefore, by (3.4) and (3.5) we have that

$$Z_k^N(t) = D_k^N(t + W_{s(k)}^N(t)) - D_k^N(t),$$

and by (3.1), for any k,

$$Z_k^N \leq \mu_k W_{s(k)}^N \quad \bigl(\text{in matrix form, } Z^N \leq M^{-1} C^T W^N\bigr). \tag{3.7}$$

That is, by (3.6) we can express the workload in terms of the fluid queue process, the natural thing, but for the reverse we only have the inequality given by (3.7). Roughly speaking, we can say that the *state space collapse* condition stated below establishes the existence of a deterministic operator which in the limit, when the number of sources N tends to ∞, expresses the fluid queue in terms of the workload.

We define λ^N to be the unique K-dimensional vector solution to the *traffic equation*, that in this case takes the form:

$$\lambda^N = \alpha^N \frac{\nu_{\text{on}}}{\nu_{\text{on}} + \nu_{\text{off}}} + P^T \lambda^N \tag{3.8}$$

(recall that ν_{on} and ν_{off} are the expected values of the ON- and OFF-period lengths, respectively), that is,

$$\lambda^N = Q\alpha^N \frac{\nu_{\text{on}}}{\nu_{\text{on}} + \nu_{\text{off}}}$$

and it can be interpreted as the long run class k fluid rate into and out of station $s(k)$ (see Theorem 2 in [7], which is a Functional Weak Law of Large Numbers for processes A^N and D^N there that justifies this interpretation).

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

$$\rho_j^N \overset{\text{def}}{=} \sum_{k \in s^{-1}(j)} \lambda_k^N m_k \quad \big(\text{in matrix form, } \rho^N = CM\lambda^N\big), \tag{3.9}$$

and introduce, by following [21], a $K \times J$ matrix $\Delta^N = \big(\Delta_{kj}^N\big)$ by

$$\Delta_{kj}^N \overset{\text{def}}{=} \begin{cases} \frac{\lambda_k^N}{\rho_j^N} & \text{if } k \in s^{-1}(j), \\ 0 & \text{otherwise.} \end{cases}$$

3.3 Model assumptions

The following assumptions will be needed throughout the paper. In order to get an asymptotic result we assume that the network is asymptotically critical, that is, $\lim_{N\to\infty} \rho_j^N = 1$ for any j. This kind of assumption is typically known in the literature as *heavy traffic*, and expresses the requirement that the limit traffic intensity is one by specifying the velocity of convergence of ρ^N to $\rho = e$, which will be faster than $N^{-1/2}$ in our case:

(HT) *Heavy-traffic assumption*

$$\lim_{N\to\infty} \sqrt{N}\big(\rho^N - e\big) = 0. \tag{3.10}$$

As a consequence of this assumption, we deduce the existence of

$$\alpha \overset{\text{def}}{=} \lim_{N\to\infty} \alpha^N,$$

$$\lambda \overset{\text{def}}{=} \lim_{N\to\infty} \lambda^N \left(= Q\alpha \frac{\nu_{\text{on}}}{\nu_{\text{on}} + \nu_{\text{off}}} \right), \quad \text{and}$$

$$\Delta \overset{\text{def}}{=} \lim_{N\to\infty} \Delta^N, \quad \text{whose elements are} \quad \Delta_{kj} = \begin{cases} \lambda_k & \text{if } k \in s^{-1}(j), \\ 0 & \text{otherwise ,} \end{cases}$$

that is, $\Delta = \operatorname{diag}(\lambda)C^T$. It is required that matrix Δ satisfies the following technical restriction.

(HΔ) *Hypothesis on matrix* Δ

$$CMQ\Delta \quad \text{is invertible.}$$

Remark 2 Note that in the particular case $K = J$, this condition is trivially accomplished. Also note that under assumptions (HT) and (HΔ) we also have that $CMQ\Delta^N$ is invertible for N big enough. Therefore, if we define $R^N \overset{\text{def}}{=} (I_J + CMQP^T\Delta^N)^{-1}$ (for N big enough), and $R \overset{\text{def}}{=} (I_J + CMQP^T\Delta)^{-1}$, we have that these matrices are well defined by assumption (HΔ), because

$$I_J + CMQP^T\Delta^N = CMQ\Delta^N \quad \text{and} \quad I_J + CMQP^T\Delta = CMQ\Delta.$$

Moreover, (HT) implies that

$$\lim_{N\to\infty} R^N = R. \tag{3.11}$$

Our final assumption is a form of *state space collapse* that expresses a relationship between the scaled immediate workload and the fluid queue length processes.

(SSC) *Assumption of state space collapse*

$$\text{P-}\lim_{N\to\infty} \sqrt{N}\big(Z^N - \Delta^N W^N\big) = 0. \tag{3.12}$$

With regard to this condition see Remark 1, and also note that by (3.6), if we define $\varepsilon^N \overset{\text{def}}{=} Z^N - \Delta^N W^N$ we have that $\varepsilon^N = (I_K - \Delta^N CM)Z^N$ and, in the particular case $K = J$, taking into account that $\Delta^N CM = I_K$, $\varepsilon^N = 0$. So, if $K = J$ condition (SSC) is trivially accomplished. In general this is not the case, and it must be imposed. On the other hand, by applying (3.7) we can always ensure that

$$\varepsilon^N \le \big(M^{-1}C^T - \Delta^N\big)W^N \quad \text{(which is non-negative).} \tag{3.13}$$

4 State space collapse and the heavy traffic limit

First we introduce the *scaled (in space) processes* associated to the fluid model (we use a hat to denote them):

$$\hat{E}^N(t) \overset{\text{def}}{=} \sqrt{N}\left(E^N(t) - \alpha^N t \frac{\nu_{\text{on}}}{\nu_{\text{on}} + \nu_{\text{off}}}\right),$$

$$\hat{A}^N(t) \overset{\text{def}}{=} \sqrt{N}\big(A^N(t) - \lambda^N t\big),$$

$$\hat{D}^N(t) \overset{\text{def}}{=} \sqrt{N}\big(D^N(t) - \lambda^N t\big),$$

$$\hat{F}^N(t) \overset{\text{def}}{=} \sqrt{N}\big(F^N(t) - P^T\lambda^N t\big),$$

$$\begin{aligned}
&\hat{W}^N(t) \overset{\text{def}}{=} \sqrt{N}W^N(t),\\
&\hat{Z}^N(t) \overset{\text{def}}{=} \sqrt{N}Z^N(t),\\
&\hat{Y}^N(t) \overset{\text{def}}{=} \sqrt{N}Y^N(t),\\
&\hat{L}^N(t) \overset{\text{def}}{=} \sqrt{N}\big(L^N(t) - \rho^N t\big),\\
&\hat{\varepsilon}^N(t) \overset{\text{def}}{=} \sqrt{N}\varepsilon^N(t).
\end{aligned}$$

Note that with this notation, (SSC) condition can be rewritten as

$$\text{P-}\lim_{N\to\infty} \hat{\varepsilon}^N = 0.$$

The following scaled equations are obtained by substituting scaled processes into model equations (3.2)–(3.6), and will be used to determine the behavior of the scaled immediate workload and the fluid queue processes, $\hat{W}^N$ and $\hat{Z}^N$, as N goes to infinity:

$$\hat{A}^N = \hat{E}^N + \hat{F}^N \quad \text{(by using (3.2) and (3.8))} \tag{4.1}$$

$$\hat{L}^N = CM\hat{A}^N \quad \text{(by using that } \rho^N = CM\lambda^N \text{ by (3.9))}, \tag{4.2}$$

$$\hat{W}^N = \hat{L}^N + \hat{Y}^N + \hat{\gamma}^N \quad \text{by (3.3), if we introduce,} \tag{4.3}$$

$$\hat{\gamma}^N(t) \overset{\text{def}}{=} \sqrt{N}\big(\rho^N - e\big)t,$$

$$\hat{W}^N = CM\hat{Z}^N \quad \text{by using (3.6)},$$

$$\hat{Z}^N = \hat{A}^N - \hat{D}^N \quad \text{(by (3.4))}, \tag{4.4}$$

$$\hat{F}^N = P^T\hat{D}^N \quad \text{by definition of } F^N \text{, and by (3.5)}, \tag{4.5}$$

$$\hat{A}^N(t) = \hat{D}^N\left(t + C^T\frac{\hat{W}^N(t)}{\sqrt{N}}\right) + \operatorname{diag}(\lambda^N)C^T\hat{W}^N(t), \tag{4.6}$$

$$\hat{\varepsilon}^N = \hat{Z}^N - \Delta^N\hat{W}^N = \big(I_K - \Delta^N CM\big)\hat{Z}^N \tag{4.7}$$

$$\text{(and therefore, } \hat{Z}^N = \hat{\varepsilon}^N + \Delta^N\hat{W}^N). \tag{4.8}$$

Now we will find an alternative expression to (4.3) for $\hat{W}^N$ in the following way: first we substitute $\hat{D}^N$ from (4.4) into (4.5), obtaining $\hat{F}^N = P^T(\hat{A}^N - \hat{Z}^N)$, and in turn by substituting this expression into (4.1) we obtain

$$\hat{A}^N = Q\big(\hat{E}^N - P^T\hat{Z}^N\big). \tag{4.9}$$

Substituting (4.9) into (4.2) and the resulting into (4.3) yields

$$\begin{aligned}
\hat{W}^N &= CMQ\hat{E}^N + \hat{\gamma}^N - CMQP^T\hat{Z}^N + \hat{Y}^N\\
&= R^N\big(CMQ\hat{E}^N + \hat{\gamma}^N - CMQP^T\hat{\varepsilon}^N + \hat{Y}^N\big)
\end{aligned}$$

by using (4.8), which can be rewritten as

$$\hat{W}^N = \hat{X}^N + R^N \hat{Y}^N \tag{4.10}$$

with

$$\hat{X}^N \overset{\text{def}}{=} R^N CMQ\big(\hat{E}^N - P^T \hat{\varepsilon}^N\big) + R^N \hat{\gamma}^N. \tag{4.11}$$

Remark 3 Note that processes appearing in expression (4.10) verify: $\hat{W}^N$ has continuous paths; for any $t \geq 0$, a.s. and $\hat{W}^N(t) \in S = \mathbb{R}_+^J$; $\hat{Y}^N$ has continuous and non-decreasing paths, and for each j, a.s., $\hat{Y}_j^N(0) = 0$ and

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

In Theorem 1 below we study the relationship between *state space collapse* (SSC) and the asymptotic behavior, when $N \to \infty$, of the fluid model introduced in Sect. 3 whose model equations are (3.2)–(3.6), under the preceding hypotheses. More specifically, we prove that (SSC) is a necessary and sufficient condition for the existence of the scaled workload limit and fluid queue limit processes $\hat{W}^N$ and $\hat{Z}^N$.

Theorem 1 *Assume that conditions* ($H\Delta$) *and* (HT) *of Sect.* 3, *and* (HR) *introduced in the Appendix*, *for matrix* R, *hold.*

Then, *condition* (SSC) *is necessary and sufficient for the existence of*

$$\hat{W} = \tilde{\lim}_{N\to\infty} \hat{W}^N \quad \textit{and} \quad \hat{Z} = \tilde{\lim_{N\to\infty}} \hat{Z}^N \quad (\textit{and if they exist, then } \hat{Z} = \Delta\hat{W}).$$

Proof Step 1: *Sufficiency.* We first mention that a slight modification in the proof of Theorem 1 [7] actually shows that there exists

$$\tilde{\lim_{N\to\infty}} (\hat{E}^N, \hat{\varepsilon}^N, \hat{\gamma}^N) = (\hat{E}, 0, 0),$$

where $\hat{E}$ is a process with continuous paths. For a fuller treatment we refer the reader to [7], where it can be seen that the convergence of $\hat{E}^N$ to $\hat{E}$ is in the sense of the convergence of the finite-dimensional distributions, because

$$\hat{E}_k^N(t) = \alpha_k^N \int_0^t \frac{1}{\sqrt{N}} \sum_{n=1}^N \left(U_k^{(n)}(u) - \frac{\nu_{\text{on}}}{\nu_{\text{on}} + \nu_{\text{off}}}\right) du,$$

and then it can be shown by the usual Central Limit Theorem, as in [17], that

$$\tilde{\lim_{N\to\infty}} \hat{E}^N (= \hat{E}) = \alpha^T \int_0^\cdot G(u)du,$$

being $\{G(t), t \geq 0\}$ a K-dimensional driftless Gaussian and stationary process. Convergence in the weak sense is not proved because unlike the limit considered in

Corollary 1 below (where we deal with a Gaussian process with stationary increments and an adequate variance function), here the tightness criterion given by Theorem 12.3 [2], cannot be applied successfully.

Following the above and by using (4.11) and the *continuous mapping theorem* (see Corollary 1 of Theorem 5.1 in [2]), we obtain that $\hat{X}=\tilde{\lim}_{N\to\infty}\hat{X}^N$ exists and can be expressed as

$$\hat{X}=RCMQ\hat{E}. \tag{4.12}$$

We proceed now to show the corresponding convergence for processes $\hat{W}^N$ and $\hat{Y}^N$: first of all note that we can rewrite (4.10) as

$$\hat{W}^N=\Big(\hat{X}^N+\big(R^N-R\big)\hat{Y}^N\Big)+R\hat{Y}^N. \tag{4.13}$$

On account of that and since R is *completely-$\mathcal{S}$* by assumption (HR), we can apply the *oscillation inequality* given in Lemma 1 [1] to obtain that a constant $C_R>0$ exists, which only depends on R, such that for any $T\geq 0$,

$$\operatorname{Osc}\Big(\hat{Y}^N(\cdot),[0,T]\Big)\leq C_R\operatorname{Osc}\Big(\hat{X}^N(\cdot)+(R^N-R)\hat{Y}^N(\cdot),[0,T]\Big), \tag{4.14}$$

$$\operatorname{Osc}\Big(\hat{W}^N(\cdot),[0,T]\Big)\leq C_R\operatorname{Osc}\Big(\hat{X}^N(\cdot)+(R^N-R)\hat{Y}^N(\cdot),[0,T]\Big). \tag{4.15}$$

By (3.11), R^N converges to R as $N\to\infty$ and hence, N_0 exists such that for any $N\geq N_0$, $C_R|R^N-R|<1/2$. Thus, (4.14) implies that if $N\geq N_0$,

$$\begin{aligned}\|\hat{Y}^N(\cdot)\|_T&=\operatorname{Osc}\Big(\hat{Y}^N(\cdot),[0,T]\Big)\\&\leq 2C_R\operatorname{Osc}\Big(\hat{X}^N(\cdot),[0,T]\Big)\leq 4C_R\|\hat{X}^N(\cdot)\|_T.\end{aligned} \tag{4.16}$$

By the continuity of $\hat{E}$ (which yields that $\hat{X}$ is continuous by (4.12)), we have that for any $T\geq 0$ and for any $\varepsilon>0$, a constant $K_\varepsilon>0$ and N_1 exist such that if $N\geq N_1$, $P(\|\hat{X}^N(\cdot)\|_T\leq\frac{K_\varepsilon}{4C_R})\geq 1-\varepsilon$. Therefore, from (4.16) we conclude that if $N\geq N_1\vee N_0$, $P(\|\hat{Y}^N(\cdot)\|_T\leq K_\varepsilon)\geq 1-\varepsilon$.

Furthermore, fixed ε and K_ε, N_2 exists such that for any $N\geq N_2$, $|R^N-R|<\frac{\varepsilon}{K_\varepsilon}$, and consequently, for any $N\geq\max\{N_0,N_1,N_2\}$,

$$P\big(|(R^N-R)|\|\hat{Y}^N(\cdot)\|_T\geq\varepsilon\big)\leq\varepsilon,$$

that is, P-$\lim_{N\to\infty}(R^N-R)\hat{Y}^N=0$.

According to $\hat{W}^N-(R^N-R)\hat{Y}^N=\hat{X}^N+R\hat{Y}^N$, which is a consequence of (4.13), it may be concluded that, if they exist, $\tilde{\lim}_{N\to\infty}(\hat{X}^N+R\hat{Y}^N)=\tilde{\lim}_{N\to\infty}\hat{W}^N$.

Since R verifies (HR), and $\hat{X}=\tilde{\lim}_{N\to\infty}\hat{X}^N$ has continuous paths, from Remark 6 we deduce the existence of a unique strong path-wise solution of the R-regularization problem of $\hat{X}$, which coincides with $(\tilde{\lim}_{N\to\infty}\hat{W}^N,\tilde{\lim}_{N\to\infty}\hat{Y}^N)$. Therefore, if we denote $\tilde{\lim}_{N\to\infty}\hat{Y}^N$ by $\hat{Y}$ and $\tilde{\lim}_{N\to\infty}\hat{W}^N$ by $\hat{W}$, we have that the unique solution of the R-regularization problem of $\hat{X}$ is $(\hat{W}, \hat{Y})$, and hence

$$\hat{W} = \hat{X} + R\hat{Y}. \tag{4.17}$$

From (4.8) and the *continuous mapping theorem*, the existence of

$$\hat{Z} = \tilde{\lim_{N\to\infty}} \hat{Z}^N, \quad \text{with } \hat{Z} = \Delta\hat{W}, \tag{4.18}$$

follows.

Step 2: *Necessity.* The proof of necessity is based in writing $\hat{\varepsilon}^N$, by using (4.7), (4.4) and (4.6), in this way:

$$\hat{\varepsilon}^N = \hat{Z}^N - \Delta^N \hat{W}^N = B_1^N + B_2^N, \quad \text{with}$$

$$B_1^N(t) = \hat{D}^N\left(t + C^T \frac{\hat{W}^N(t)}{\sqrt{N}}\right) - \hat{D}^N(t) \text{ and}$$

$$B_2^N = \left(\operatorname{diag}(\lambda^N)C^T - \Delta^N\right)\hat{W}^N.$$

We assume condition (HT) and, in consequence,

$$\lim_{N\to\infty}\left(\operatorname{diag}(\lambda^N)C^T - \Delta^N\right) = \operatorname{diag}(\lambda)C^T - \Delta = 0.$$

By assumption of the existence of $\hat{W} = \tilde{\lim}_{N\to\infty}\hat{W}^N$, it follows that $\tilde{\lim}_{N\to\infty} B_2^N = 0$. Let us now examine B_1^N. Taking into account that $\tilde{\lim}_{N\to\infty} C^T \frac{\hat{W}^N(t)}{\sqrt{N}} = 0$ and the random time change theorem (see (17.9) in [2]), from the existence of

$$\hat{D} = \tilde{\lim_{N\to\infty}} \hat{D}^N \left(= Q\left(\hat{E} - P^T\hat{Z}\right) - \hat{Z}\right) \tag{4.19}$$

(we will prove this fact below), we deduce that $\tilde{\lim}_{N\to\infty} B_1^N = 0$, which finishes the proof of the convergence $\tilde{\lim}_{N\to\infty}\hat{\varepsilon}^N = 0$. Moreover, assumption (SSC) follows if we prove tightness, and tightness can be easily checked by using (3.13), (4.15) and (4.16), from which it may be concluded that a positive constant κ exists such that for any $T \ge 0$, $\|\hat{\varepsilon}^N(\cdot)\|_T \le \kappa\|\hat{X}^N(\cdot)\|_T$. Hence, from the continuity of $\hat{X} = \tilde{\lim}_{N\to\infty}\hat{X}^N$ we have that for any $T \ge 0$ and for any $\varepsilon > 0$, a positive constant $K'_\varepsilon > 0$ exists such that

$$P\left(\|\hat{\varepsilon}^N(\cdot)\|_T \le K'_\varepsilon\right) \ge 1 - \varepsilon \quad \text{(for } N \text{ big enough)},$$

that is, sequence $\{\hat{\varepsilon}^N\}_N$ is tight.

To finish Step 2 we have to prove (4.19), as explained above. Indeed, (4.19) follows from (4.4) and the existence of

$$\tilde{\lim_{N\to\infty}} \hat{A}^N = \tilde{\lim_{N\to\infty}} Q\left(\hat{E}^N - P^T\hat{Z}^N\right) = Q\left(\hat{E} - P^T\hat{Z}\right),$$

which is a consequence of (4.9). Note that the existence of $\hat{E}$, which was shown in Step 1, does not need assumption (SSC). □

We now introduce the *scaled (in time) processes* associated to the fluid model, indexed by r, where r (the factor of scaling in time) tends to infinity through a strictly increasing sequence of strictly positive real numbers. For this we introduce some notation previously used in [17] and [7]. Set $\beta_{\min} = \min(\beta_1, \beta_2)(\in (1,2))$. For any $j = 1, 2$, set $a_j = \frac{\Gamma(2-\beta_j)}{(\beta_j-1)}$ if $\sigma_j^2 = +\infty$ and $a_j = \frac{\sigma_j^2}{2}$ if $\sigma_j^2 < +\infty$. The normalization factors used below depend on whether b, defined by $b \overset{\text{def}}{=} \lim_{t\to\infty} t^{\beta_2-\beta_1}\frac{L_1(t)}{L_2(t)}$, is finite or not. If $0 \le b < \infty$ we have that $\beta_{\min} = \beta_2$. Set $L = L_2$ and

$$\sigma^2 = \frac{2(\nu_{\text{off}}^2 a_1 b + \nu_{\text{on}}^2 a_2)}{(\nu_{\text{on}} + \nu_{\text{off}})^3 \Gamma(4-\beta_{\min})}.$$

If, on the other hand, $b = \infty$, $\beta_{\min} = \beta_1$. Then set $L = L_1$ and

$$\sigma^2 = \frac{2\nu_{\text{off}}^2 a_1}{(\nu_{\text{on}} + \nu_{\text{off}})^3 \Gamma(4-\beta_{\min})}.$$

Let us define $H \overset{\text{def}}{=} \frac{3-\beta_{\min}}{2}$.

Remark 4 In Corollary 1 below quantity H plays the role of the Hurst parameter of the *reflected fractional Brownian motion process* (rfBm), to which the scaled in time workload process converges. Definition of this process can be found in the Appendix. Note that $\beta_{\min} \in (1,2)$ implies $H \in (\frac{1}{2}, 1)$. In particular, $H > \frac{1}{2}$ (the condition on the Hurst parameter corresponding to the long-range dependence behavior of the rfBm process) is due to the fact that $\beta_{\min} < 2$, that is, that the ON- or OFF-period lengths (at least one of them) have infinite variance (heavy tails). As is mentioned in [17], if both period lengths were light-tailed (with finite variances), then $\beta_1 = \beta_2 = 2$ and $H = \frac{1}{2}$, which would correspond to the ordinary Brownian motion process, whose increments are independent.

Corollary 1 *Under the assumptions of Theorem 1, suppose that condition* (SSC) *also holds. Then,* $\hat{X} = \tilde{\lim}_{N\to\infty} \hat{X}^N$ *and* $\hat{Y} = \tilde{\lim}_{N\to\infty} \hat{Y}^N$ *exist, and if we define the scaled in time limit processes by*

$$\hat{\hat{W}}^r(t) \overset{\text{def}}{=} \frac{\hat{W}(rt)}{r^H L^{1/2}(r)}, \qquad \hat{\hat{X}}^r(t) \overset{\text{def}}{=} \frac{\hat{X}(rt)}{r^H L^{1/2}(r)},$$

$$\hat{\hat{Y}}^r(t) \overset{\text{def}}{=} \frac{\hat{Y}(rt)}{r^H L^{1/2}(r)} \quad \textit{and} \quad \hat{\hat{Z}}^r(t) \overset{\text{def}}{=} \frac{\hat{Z}(rt)}{r^H L^{1/2}(r)},$$

we also have the following:

(i) $W = \mathcal{D}\text{-}\lim_{r\to\infty} \hat{\hat{W}}^r$, $X = \mathcal{D}\text{-}\lim_{r\to\infty} \hat{\hat{X}}^r$ *and* $Y = \mathcal{D}\text{-}\lim_{r\to\infty} \hat{\hat{Y}}^r$ *exist*,

(ii) $W = X + RY$ *and it is a rfBm on* $S = \mathbb{R}_+^J$ *with associated data*

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

where $\Gamma = \sigma^2 RCMQ\,\text{diag}(\alpha)^2 Q^T MC^T R^T$, *and*

(iii) $Z = \mathcal{D}\text{-}\lim_{r\to\infty} \hat{\hat{Z}}^r$ *also exists*, *and* $Z = \Delta W$.

Remark 5 Assumption (HR) holds trivially for matrix R if $K = J$ because $(\Delta)^{-1}M^{-1} = I_J$ and $C = I_J$ in that case, and then we obtain that

$$\begin{aligned} R &= (I_J + G)^{-1} = (CMQ\Delta)^{-1} = \Delta^{-1}Q^{-1}M^{-1} \\ &= \Delta^{-1}M^{-1} - \Delta^{-1}P^TM^{-1} = I_J + \big(-\Delta^{-1}P^TM^{-1}\big), \end{aligned}$$

and $\Theta = \Delta^{-1}P^TM^{-1}$ has the same spectral radius as P, which is supposed to be strictly less than one. Therefore, for $K = J$ assumption (HR) is accomplished, and also conditions (SSC) and (HΔ), as we have seen above. Therefore, for $K = J$ heavy traffic (HT) is the only hypothesis we need to prove the rfBm limit for the scaled (both in space and time) workload process, as was established in Theorem 1 of [7].

Proof of Corollary 1 The existence of $\hat{X} = \tilde{\lim}_{N\to\infty}\hat{X}^N$ and $\hat{Y} = \tilde{\lim}_{N\to\infty}\hat{Y}^N$ is proved in the Step 1 of the proof of Theorem 1, and we also have by (4.12) and (4.17) that

$$\hat{X} = RCMQ\hat{E} \quad \text{and} \quad \hat{W} = \hat{X} + R\hat{Y}. \tag{4.20}$$

Furthermore, with a similar proof to that of Theorem 1 [7], we can obtain that

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

where B^H is a K-dimensional drift-less fractional Brownian motion with associated data $(x = 0, H = \frac{3-\beta_{\min}}{2}, \theta = 0, \Gamma = \sigma^2\,\mathrm{diag}(\alpha)^2)$. Here, as we have pointed out in the proof of the previous theorem, the convergence in the weak sense can be proved, as in [17], by using the tightness criterion.

As a consequence, we have that $\mathcal{D}\text{-}\lim_{r\to\infty}\hat{\hat{X}}^r = X$ exists, with $X = RCMQB^H$ by (4.20), which is a J-dimensional fBm with associated data $(x = 0, H = \frac{3-\beta_{\min}}{2}, \theta = 0, \Gamma)$, being Γ the matrix

$$\Gamma = \sigma^2 RCMQ\,\mathrm{diag}(\alpha)^2 Q^T MC^T R^T.$$

Moreover, from (4.20) we conclude that $\hat{W}$, $\hat{X}$ and $\hat{Y}$ verify the hypotheses of the *invariance principle* of Theorem 4.1 [20] with matrix R, taking into account that $\mathcal{D}\text{-}\lim_{r\to\infty} \hat{\hat{X}}^r = X$, and R is a *Completely-$\mathcal{S}$* matrix by assumption (HR).

Therefore, $\{(\hat{\hat{W}}^r, \hat{\hat{X}}^r, \hat{\hat{Y}}^r)\}_r$ inherits tightness from sequence $\{\hat{\hat{X}}^r\}_r$ and consequently, by assumption (HR) (see Corollary 4.3 [20]),

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

where $W = X + RY$, and conditions of Definition 1 (see Appendix) are satisfied. Hence W is a J-dimensional rfBm on $S = \mathbb{R}_+^J$ with associated data $(x = 0, H = \frac{3-\beta_{\min}}{2}, \theta = 0, \Gamma, R)$, and (i) and (ii) are proved.

Finally, we deduce (iii) from (4.18).

We mention that Theorem 4.1 [20] gives the convergence in the distributional sense on $\mathcal{D}^J$, the space of functions from $[0,\infty)$ to $\mathbb{R}^J$ which are right continuous and have finite left hand limits, with the Skorokhod topology. Our convergence is taken in the distributional sense on $\mathcal{C}^J$, and is implied by the convergence on $\mathcal{D}^J$ because the Skorokhod topology relativized to $\mathcal{C}^J$ coincides with the uniform topology over compacts. □

5 Multiplicative state space collapse

It is possible, by following [4], to introduce an assumption related to (SSC) which is a kind of *multiplicative state space collapse*, in our setting:

(MSSC) *Multiplicative state space collapse*

$$\text{P-}\lim_{N\to\infty} \frac{\hat{\varepsilon}^N}{\|W^N(\cdot)\|_T \vee 1} = 0.$$

It is obvious that (SSC) implies (MSSC), because for any N and t, if we introduce the notation

$$\hat{\zeta}^N(t) \overset{\text{def}}{=} \frac{\hat{\varepsilon}^N(t)}{\|W^N(\cdot)\|_T \vee 1},$$

we have that $|\hat{\zeta}^N(t)| \le |\hat{\varepsilon}^N(t)|$. We will see in Proposition 1 below that in fact they are equivalent (that is, (MSSC) also implies (SSC)) if matrix R is *Completely-$\mathcal{S}$*. Previously, in Lemma 1 we will establish a technical result that is needed in the proof of Proposition 1.

Lemma 1 *Assume that R is a Completely-$\mathcal{S}$ matrix and that* (MSSC) *holds. Therefore, for any $T \ge 0$ and for any $\varepsilon > 0$, a constant $C_{R,\varepsilon} > 1$ which only depends on R and ε, and $\tilde{N}$ exist such that for any $N \ge \tilde{N}$, we have that*

$$P\big(\|\hat{W}^N(\cdot)\|_T \vee 1 \le C_{R,\varepsilon}\big) \ge 1-\varepsilon$$

(*equivalently, $\|\hat{W}^N(\cdot)\|_T$ is bounded in probability, that is, a constant $\tilde{C}_{R,\varepsilon} > 0$ and $\tilde{N}$ exist such that for any $N \ge \tilde{N}$, $P(\|\hat{W}^N(\cdot)\|_T \le \tilde{C}_{R,\varepsilon}) \ge 1-\varepsilon$, by taking $\tilde{C}_{R,\varepsilon} = C_{R,\varepsilon} - 1$*).

Proof By (4.15) and (4.16) we have that a constant $C_R > 0$ which only depends on R exists such that for any $N \ge N_0$ (N_0 is the fixed value that appears in the proof of expression (4.16)),

$$\begin{aligned}
\|\hat{W}^N(\cdot)\|_T &= \operatorname{Osc}\left(\hat{W}^N(\cdot), [0,T]\right) \\
&\le C_R \operatorname{Osc}\left(\hat{X}^N(\cdot), [0,T]\right) + \frac{1}{2}\operatorname{Osc}\left(\hat{Y}^N(\cdot), [0,T]\right) \\
&\le 2C_R \operatorname{Osc}\left(\hat{X}^N(\cdot), [0,T]\right).
\end{aligned} \tag{5.1}$$

By defining $\hat{\xi}^N \overset{\text{def}}{=} CMQ\hat{E}^N + \hat{\gamma}^N$, (4.11) gives that

$$\begin{aligned}
&\operatorname{Osc}\Big(\hat{X}^N(\cdot), [0,T]\Big)\\
&\quad = \operatorname{Osc}\Big(R^N\big(\hat{\xi}^N(\cdot) - CMQP^T\hat{\zeta}^N(\cdot)\big(\|\hat{W}^N(\cdot)\|_T \vee 1\big)\big), [0,T]\Big)\\
&\quad \le |R^N| \operatorname{Osc}\Big(\hat{\xi}^N(\cdot), [0,T]\Big)\\
&\qquad + |R^N CMQP^T|\big(\|\hat{W}^N(\cdot)\|_T \vee 1\big) \operatorname{Osc}\Big(\hat{\zeta}^N(\cdot), [0,T]\Big)\\
&\quad \le 2|R^N|\|\hat{\xi}^N(\cdot)\|_T + 2|R^N CMQP^T|\big(\|\hat{W}^N(\cdot)\|_T \vee 1\big)\|\hat{\zeta}^N(\cdot)\|_T. \qquad (5.2)
\end{aligned}$$

By assumption (MSSC), it follows that $\|\hat{\zeta}^N(\cdot)\|_T \to 0$ in probability as N goes to infinity. Consequently, N_1 exists such that for any $N \ge N_0 \vee N_1$,

$$P\bigg(4C_R|R^N CMQP^T|\|\hat{\zeta}^N(\cdot)\|_T > \frac{1}{2}\bigg) \le \frac{\varepsilon}{2}.$$

From (5.1) and (5.2) we deduce the following chain of inclusions:

$$\begin{aligned}
&\bigg\{4C_R|R^N CMQP^T|\|\hat{\zeta}^N(\cdot)\|_T \le \frac{1}{2}\bigg\}\\
&\quad \subseteq \bigg\{\|\hat{W}^N(\cdot)\|_T \le 4C_R|R^N|\|\hat{\xi}^N(\cdot)\|_T + \frac{1}{2}\big(\|\hat{W}^N(\cdot)\|_T \vee 1\big)\bigg\}\\
&\quad \subseteq \Big\{\big(\|\hat{W}^N(\cdot)\|_T \vee 1\big) \le 8C_R|R^N|\|\hat{\xi}^N(\cdot)\|_T + 1\Big\},
\end{aligned}$$

and therefore

$$P\Big(\big(\|\hat{W}^N(\cdot)\|_T \vee 1\big) \le 8C_R|R^N|\|\hat{\xi}^N(\cdot)\|_T + 1\Big) \ge 1 - \frac{\varepsilon}{2}.$$

On the other hand, the continuity of $\hat{E}$ implies that of

$$\hat{\xi} = \widetilde{\lim_{N\to\infty}} \hat{\xi}^N = CMQ\hat{E},$$

and then for any $T \ge 0$ and for any $\varepsilon > 0$, a constant $\kappa_\varepsilon > 0$ and N_2 exist such that if $N \ge N_2$,

$$P\Big(\|\hat{\xi}^N(\cdot)\|_T \le \kappa_\varepsilon\Big) \ge 1 - \frac{\varepsilon}{2}.$$

By using now that $|R^N| \le 2|R|$ if N is big enough, say $N \ge N_3$, since $R^N \to R$, we have that for any $N \ge \tilde{N}$, with $\tilde{N} = \max\{N_0, N_1, N_2, N_3\}$,

$$P\Big(\big(\|\hat{W}^N(\cdot)\|_T \vee 1\big) \le C_{R,\varepsilon}\Big) \ge 1 - \varepsilon, \quad \text{where } C_{R,\varepsilon} = 16C_R|R|\kappa_\varepsilon + 1(> 1),$$

which finishes the proof. □

Proposition 1 *Assume that R is a Completely-$\mathcal{S}$ matrix. Then*,

$$\text{(MSSC)} \Longrightarrow \text{(SSC)}.$$

Proof We have that $\hat{\varepsilon}^N(t) = \hat{\zeta}^N(t)(\|\hat{W}^N(\cdot)\|_T \vee 1)$ for any t. Since we suppose that assumption (MSSC) holds, we obtain that for any $\varepsilon > 0$ and for any $T \geq 0$, N' exists such that for any $N \geq N'$,

$$P\left(\|\hat{\zeta}^N(\cdot)\|_T \geq \frac{\varepsilon}{C_{R,\varepsilon/2}}\right) \leq \frac{\varepsilon}{2},$$

by taking $C_{R,\varepsilon/2}$ to be the constant that appears in Lemma 1 by substituting ε for $\varepsilon/2$ there. Lemma 1 implies that $\tilde{N}$ exists such that for any $N \geq \tilde{N}$, $P(\|\hat{W}^N(\cdot)\|_T \vee 1 > C_{R,\varepsilon/2}) \leq \frac{\varepsilon}{2}$, and therefore, for any $N \geq N' \vee \tilde{N}$,

$$\begin{aligned} P\left(\|\hat{\varepsilon}^N(\cdot)\|_T \geq \varepsilon\right) &= P\left(\|\hat{\zeta}^N(\cdot)\|_T \left(\|\hat{W}^N(\cdot)\|_T \vee 1\right) \geq \varepsilon\right) \\ &\leq P\left(\|\hat{\zeta}^N(\cdot)\|_T \geq \frac{\varepsilon}{C_{R,\varepsilon/2}}\right) + P\left(\|\hat{W}^N(\cdot)\|_T \vee 1 > C_{R,\varepsilon/2}\right) \leq \varepsilon, \end{aligned}$$

and the proof now is finished because we have proved that for any $T \geq 0$, $\|\hat{\varepsilon}^N(\cdot)\|_T$ converges to 0 in probability as $N \to \infty$, that is, $\text{P-}\lim_{N\to\infty} \hat{\varepsilon}^N = 0$. □

6 Examples

We start this section by considering the example of a tandem queue with feedback. We will examine this example in some detail.

6.1 The tandem queue with feedback

Consider a fluid tandem queue, which is a network with two stations ($J = 2$) and three fluid classes ($K = 3$). Class 1 fluid enters the system from outside (at rate $\alpha_1^N > 0$) and it is processed by server 1. After being processed (at constant processing rate $1/m_1$) by the first server, this fluid goes into station 2 as class 3 fluid, where it is processed at constant processing rate $1/m_3$. After that, a proportion $q \in (0, 1]$ of fluid goes outside the network but the proportion $p = 1 - q$ goes back to station 1 to be reprocessed as class 2 fluid, at constant processing rate $1/m_2$, and then goes again to station 2 as class 3 fluid, and so on. This model, which is a two-stage queueing system, seems adequate, for example, for situations in which there is recycling, that is, quality control inspection is performed after first stage at the second one, and fluid that does not meet quality standards is sent back to station 1 for reprocessing.

In that scenario, $\alpha_1^N > 0$ but $\alpha_2^N = \alpha_3^N = 0$ (the system only allows external arrivals of class 1 fluid). Constituency and flow matrices are, respectively,

$$C = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & p & 0 \end{pmatrix}$$

(note that P is a sub-stochastic matrix with spectral radius $\sqrt{p}$, which is strictly less than 1). Therefore, by (3.6) we can express the workload process in terms of the fluid queue process as $W^N = CMZ^N$, that is,

$$\begin{cases} W_1^N = m_1 Z_1^N + m_2 Z_2^N, & \text{and} \\ W_2^N = m_3 Z_3^N. & \end{cases}$$

We can isolate from the second equation and obtain $Z_3^N = \frac{1}{m_3} W_2^N$, but we cannot do the same with Z_1^N nor Z_2^N. On the other hand, we also know that $Z^N \leq M^{-1} C^T W^N$ (see (3.7)), that is,

$$\begin{cases} Z_1^N \leq \frac{1}{m_1} W_1^N, & \\ Z_2^N \leq \frac{1}{m_2} W_1^N, & \text{and} \\ Z_3^N \leq \frac{1}{m_3} W_2^N & \text{(actually, this is an equality).} \end{cases}$$

Fluid traffic intensity is $\rho^N = (\rho_1^N, \rho_2^N)^T$, with $\rho_1^N = \lambda_1^N m_1 + \lambda_2^N m_2$ and $\rho_2^N = \lambda_3^N m_3$, being $\lambda^N = Q\alpha^N \nu$ (by ν we denote throughout this section constant $\frac{\nu_{\text{on}}}{\nu_{\text{on}}+\nu_{\text{off}}}$). Taking into account that $Q = (I_K - P^T)^{-1}$, we have that

$$Q = \frac{1}{q} \begin{pmatrix} q & 0 & 0 \\ p & 1 & p \\ 1 & 1 & 1 \end{pmatrix}$$

and thus

$$\lambda_1^N = \nu\alpha_1^N, \qquad \lambda_2^N = \nu\frac{p}{q}\alpha_1^N \quad \text{and} \quad \lambda_3^N = \nu\frac{1}{q}\alpha_1^N. \tag{6.1}$$

Heavy traffic condition (HT) can be stated as:

$$\boxed{\alpha_1 \Bigl(= \lim_{N\to\infty} \alpha_1^N\Bigr) = \frac{q}{m_3 \nu} > 0 \quad \text{and} \quad qm_1 + pm_2 = m_3}$$

by using that $\lambda_1^N m_1 + \lambda_2^N m_2 \longrightarrow 1$, $\lambda_3^N m_3 \longrightarrow 1$, and that by (6.1),

$$\lambda_1^N \to \lambda_1 = \nu\alpha_1, \qquad \lambda_2^N \to \lambda_2 = \nu\frac{p}{q}\alpha_1 \quad \text{and} \quad \lambda_3^N \to \lambda_3 = \nu\frac{1}{q}\alpha_1.$$

Under (HT) we have

$$\Delta = \begin{pmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ 0 & \lambda_3 \end{pmatrix} \quad \text{and} \quad CMQ\Delta = \begin{pmatrix} 1 + \frac{p}{q}\frac{m_2}{m_3} & \frac{p}{q}\frac{m_2}{m_3} \\ \frac{1}{q} & \frac{1}{q} \end{pmatrix},$$

which turns out to be an invertible matrix. Thus, assumption (HΔ) is accomplished.

State space collapse (SSC) can be expressed under (HT) in the following way:

$$\text{P-}\lim_{N\to\infty} \bigl(\lambda_2^N Z_1^N - \lambda_1^N Z_2^N\bigr) \quad \text{exists and equals zero,}$$

because

$$I_K - \Delta^N CM = \begin{pmatrix} 1-\frac{m_1\lambda_1^N}{\rho_1^N} & -\frac{m_2\lambda_1^N}{\rho_1^N} & 0 \\ -\frac{m_1\lambda_2^N}{\rho_1^N} & 1-\frac{m_2\lambda_2^N}{\rho_1^N} & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{m_2\lambda_2^N}{\rho_1^N} & -\frac{m_2\lambda_1^N}{\rho_1^N} & 0 \\ -\frac{m_1\lambda_2^N}{\rho_1^N} & \frac{m_1\lambda_1^N}{\rho_1^N} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Finally, with respect to condition (HR) we have that

$$R = \left(CMQ\Delta\right)^{-1} = \begin{pmatrix} 1 & -p\frac{m_2}{m_3} \\ -1 & q+p\frac{m_2}{m_3} \end{pmatrix} = I_2 + \Theta, \quad \text{with } \Theta = \begin{pmatrix} 0 & -p\frac{m_2}{m_3} \\ -1 & p(\frac{m_2}{m_3}-1) \end{pmatrix}$$

and we can ensure that the spectral radius of the matrix obtained from Θ by replacing its elements by their absolute values, is strictly less than 1

$$\boxed{\text{if } p < \frac{m_3}{2m_2 - m_3} \quad \text{in the case } m_3 < m_2}$$

(otherwise, the spectral radius is always strictly less than 1).

As a consequence of all these facts, we can see that Theorem 1 and Corollary 1 for that tandem queue could be rewritten in this way:

Corollary 2 *For the tandem queue considered in this section and with notations of the previous sections, under assumptions*:

$$(h_1) \quad \begin{cases} \alpha_1^N \longrightarrow \frac{q}{m_3\nu}, \\ qm_1 + pm_2 = m_3, \\ p < \frac{m_3}{2m_2 - m_3} \quad \textit{if } m_3 < m_2 \end{cases}$$

we have that condition

$$(h_2) \quad \text{P-}\lim_{N\to\infty} \left(\lambda_2^N Z_1^N - \lambda_1^N Z_2^N\right) \quad \textit{exists and equals zero}$$

is necessary and sufficient for the existence of $\hat{W}$ and $\hat{Z}$ (which must verify $p\hat{Z}_1 = q\hat{Z}_2$), and $\hat{Z} = \Delta\hat{W}$. Moreover, under (h_1) and (h_2), we have that

(i) *W, X and Y exist,*
(ii) *$W = X + RY$ and it is a rfBm on $S = \mathbb{R}_+^2$ with associated data*

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

$$R = \begin{pmatrix} 1 & -p\frac{m_2}{m_3} \\ -1 & q+p\frac{m_2}{m_3} \end{pmatrix} \textit{ and } \Gamma = \sigma^2\alpha_1^2 \begin{pmatrix} m_1^2 & m_1(m_3-m_1) \\ m_1(m_3-m_1) & (m_3-m_1)^2 \end{pmatrix}$$

with $\alpha_1 = \frac{q}{m_3\nu}$, and

(iii) *Z also exists, and $Z = \Delta W$, that is,*

$$\begin{cases} Z_1 = \frac{q}{m_3} W_1, \\ Z_2 = \frac{p}{m_3} W_1, \\ Z_3 = \frac{1}{m_3} W_2. \end{cases}$$

6.2 More insight into condition (SSC)

In general, if we do not know the structure of the network, we cannot obtain expressions as explicit as in the previous example for condition (SSC) nor for condition (HT), which states that the limit values of parameters m and λ verify

$$\rho_j = \sum_{k \in s^{-1}(j)} \lambda_k m_k = 1 \quad \text{for any } j = 1, \dots, J.$$

Nevertheless, it is quite straightforward to show that condition (SSC) can be expressed in general in the following way: for any k such that $s^{-1}(s(k)) \diagdown \{k\} \neq \emptyset$,

$$\boxed{\text{P-}\lim_{N \to \infty} \Big((\rho_{s(k)}^N - \lambda_k^N m_k) Z_k^N - \sum_{k' \in s^{-1}(s(k)), k' \neq k} \lambda_k^N m_{k'} Z_{k'}^N \Big) \quad \text{exists and equals zero}} \tag{6.2}$$

Indeed, we have that $I_K - \Delta^N CM$ is a $K \times K$ matrix with diagonal square matrix boxes (and zeros outside), one for each station. Take one of these boxes, say that corresponding to station j, and assume that fluid classes processed at that station are $j_1, \dots, j_{r_1}$. Then, row ℓ of this box (for $\ell = 1, \dots, r_1$) has the form:

$$\left(-\frac{\lambda_{j_\ell}^N m_{j_1}}{\rho_j^N}, -\frac{\lambda_{j_\ell}^N m_{j_2}}{\rho_j^N}, \dots, -\frac{\lambda_{j_\ell}^N m_{j_{\ell-1}}}{\rho_j^N}, 1 - \frac{\lambda_{j_\ell}^N m_{j_\ell}}{\rho_j^N}, -\frac{\lambda_{j_\ell}^N m_{j_{\ell+1}}}{\rho_j^N}, \dots, -\frac{\lambda_{j_\ell}^N m_{j_{r_1}}}{\rho_j^N} \right).$$

Therefore, by Theorem 1, under (HT), (HΔ), (HR) and (SSC), we know that $\hat{Z}$ exists, and verifies that for any k such that $s^{-1}(s(k)) \diagdown \{k\} \neq \emptyset$,

$$\hat{Z}_k = \frac{\lambda_k}{1 - \lambda_k m_k} \sum_{k' \in s^{-1}(s(k)), k' \neq k} m_{k'} \hat{Z}_{k'}, \tag{6.3}$$

and moreover $\hat{W}$ exists and $\hat{Z} = \Delta \hat{W}$. Note that condition (SSC) as stated in (6.2), expresses some relationships between those components of $\hat{Z}$ that correspond to the fluid classes processed at the same station (see (6.3)). In particular, if for one station there is only one fluid class processed there, there is no restriction in the corresponding component of $\hat{Z}$. Corollary 1 (analogously to Corollary 2) says that under the same hypotheses, both the rfBm process W, which is the limit of the workload process, and the queue limit process Z, exist, and moreover $Z = \Delta W$, that is,

$$\boxed{Z_k = \lambda_k W_{s(k)}, \quad \text{for any } k = 1, \dots, K} \tag{6.4}$$

We can observe that formula (6.4) is compatible with (3.6), which implies that in the limit, for any station j, $W_j = \sum_{k\in s^{-1}(j)} m_k Z_k$, because

$$\sum_{k\in s^{-1}(j)} m_k Z_k = \sum_{k\in s^{-1}(j)} m_k \lambda_k W_{s(k)} = W_j \sum_{k\in s^{-1}(j)} m_k \lambda_k = W_j \rho_j = W_j,$$

and also with (3.7), because $\lambda_k \le \mu_k (= \frac{1}{m_k})$ due to the fact that $\lambda_k m_k \le \rho_{s(k)} = 1$. Finally, we point out that (6.4) can be thought of as another approach to the interpretation of λ_k as the long run class k fluid rate into and out of station $s(k)$.

We can apply comments in this subsection to other examples than the tandem queue. Only as an illustration, we consider two more examples in the next subsection.

6.3 Two more examples

A queueing network with a traffic stream Consider a queueing network similar to that introduced by Majewski [13] (Sect. 6). In that network, a traffic stream has to traverse several (say J) stations which are additionally loaded with long-range dependent background traffic. We have then J stations and $K = 2J$ fluid classes. $\alpha_1^N > 0$ is the arrival rate for class 1 fluid (the incoming stream) to station 1, which also serves class 2 fluid (that arrives from outside at rate $\alpha_2^N > 0$ and after being served at this station leaves the system). When class 1 fluid finishes service at station 1, it is sent to station 2 as class 3 fluid. Class 4 fluid also arrives at station 2 from outside, at a rate $\alpha_4^N > 0$, and after being served leaves the system. Class 3 fluid, when served, next goes to station 3 as class 5 fluid, and so on. At the last station, when service finishes, all the fluid leaves the system.

With this notation, class 1 fluid is the traffic stream that must traverse all the J stations by changing class (from class 1 at station 1 to class 3 at station $2, \dots,$ to class $2j-1$ at station j). Moreover, at any station j there is an extra input of class $2j$ fluid that arrives from outside, at rate $\alpha_{2j}^N > 0$. Then, station j processes two fluid classes: $2j-1$, which is the traffic stream, and $2j$, which comes from outside, at respective constant processing rates $1/m_{2j-1}$ and $1/m_{2j}$. Note that $\alpha_3^N = \cdots = \alpha_{2J-1}^N = 0$, and by using that $\lambda^N = Q\alpha^N \nu$ (recall that $\nu = \frac{\nu_{\mathrm{on}}}{\nu_{\mathrm{on}}+\nu_{\mathrm{off}}}$) we have:

$$\text{for any station } j, \quad \lambda_{2j}^N = \alpha_{2j}^N \nu, \qquad \lambda_{2j-1}^N = \alpha_1^N \nu.$$

As a consequence, the restriction on the parameters introduced by (HT) is:

$$\boxed{\exists \alpha_1 = \lim_{N\to\infty} \alpha_1^N > 0, \alpha_{2j} = \lim_{N\to\infty} \alpha_{2j}^N > 0 \quad \text{and} \quad \nu\big(\alpha_1 m_{2j-1} + \alpha_{2j} m_{2j}\big) = 1 \quad \forall j}$$

It can be easily seen that conditions (HΔ) and (HR) are satisfied, because $CMQ\Delta$ is a lower-triangular matrix with all its diagonal elements equal to 1, under (HT) (and, consequently, the same applies for $R = (CMQ\Delta)^{-1}$).

By (6.2) condition (SSC) can be expressed as:

$$\boxed{\text{P-}\lim_{N\to\infty} \big(\alpha_{2j}^N Z_{2j-1}^N - \alpha_1^N Z_{2j}^N\big) \quad \text{exists and equals zero, for any} j = 1, \dots, J}$$

As a consequence, by (6.3) we will have that $\hat{Z}_{2j-1} = \frac{\alpha_1}{\alpha_{2j}} \hat{Z}_{2j}$ for any j, and finally (6.4) becomes:

$$\boxed{Z_{2j-1} = \alpha_1 \nu W_j \quad \text{and} \quad Z_{2j} = \alpha_{2j} \nu W_j, \quad \text{for any } j = 1, \ldots, J}$$

A $\bigvee$-system with feedback allowed As a final example consider now a $\bigvee$-system consisting of multiple fluid classes ($K > 1$ classes) served by a single server ($J = 1$). We assume that fluid classes from 1 to $K - 1$ arrive at the server from outside at respective rates $\alpha_1^N, \ldots, \alpha_{K-1}^N > 0$. After being processed with constant processing rates $1/m_1, \ldots, 1/m_{K-1} > 0$ respectively, respective proportions $q_1, \ldots, q_{K-1}$ of fluid leave the system, and the rest, $p_1, \ldots, p_{K-1}$ with $p_\ell = 1 - q_\ell$, come back to the server as class K fluid to be served (with a FIFO service discipline) with a constant processing rate $1/m_K > 0$. After being served, a proportion of $q_K (> 0)$ class K fluid leaves the system and the rest ($p_K = 1 - q_K$) comes back to the server as class K fluid again. Note that $p_\ell = 0$ for all ℓ corresponds to the non-feedback model, and that it is allowed to have feedback only for some classes. It could be possible as well that some $p_\ell, \ell = 1, \ldots, K - 1$, be equal to 1, but p_K must be strictly less than 1 .

In this example, matrix P is identically zero except for its last column, whose elements are $p_1, \ldots, p_K$, and has spectral radius $\sqrt{p_K} < 1$. Taking into account that $\alpha_K^N = 0$, we obtain that

$$\lambda_\ell^N = \nu \alpha_\ell^N \quad \text{for } \ell = 1, \ldots, K - 1, \quad \text{and} \quad \lambda_K^N = \nu \frac{1}{q_K} \sum_{s=1}^{K-1} p_s \alpha_s^N .$$

Condition (HT) can be written as:

$$\boxed{\exists \alpha_\ell = \lim_{N \to \infty} \alpha_\ell^N > 0 \text{ for } \ell = 1, \ldots, K - 1, \quad \text{and} \quad \nu \sum_{s=1}^{K-1} \alpha_s \big(m_s + \frac{p_s}{q_K} m_K \big) = 1}$$

Under (HT) conditions (HΔ) and (HR) are trivially satisfied in this example, because

$$CMQ\Delta = 1 + \nu \frac{1}{q_K^2} \left(\sum_{s=1}^{K-1} p_s \alpha_s \right) m_K (\geq 1).$$

Condition (SSC) is expressed by means of:

$$\text{P-}\lim_{N \to \infty} \left((\rho^N - \nu \alpha_\ell^N m_\ell) Z_\ell^N - \nu \alpha_\ell^N \sum_{k \neq \ell} m_k Z_k^N \right) = 0 \quad \text{for } \ell = 1, \ldots, K - 1, \text{and}$$

$$\text{P-}\lim_{N \to \infty} \left(q_K \left(\sum_{s=1}^{K-1} \alpha_s^N m_s \right) Z_K^N - \left(\sum_{s=1}^{K-1} p_s \alpha_s^N \right) \sum_{k=1}^{K-1} m_k Z_k^N \right) = 0$$

and as a consequence, (6.3) and (6.4) become, respectively,

$$\boxed{\hat{Z}_\ell = \begin{cases} \frac{\nu\alpha_\ell}{1-\nu\alpha_\ell m_\ell} \sum\limits_{k\neq\ell} m_k \hat{Z}_k & \text{if } \ell = 1,\dots,K-1 \\ \frac{1}{q_K}\left(\frac{\sum\limits_{s=1}^{K-1} p_s\alpha_s}{\sum\limits_{s=1}^{K-1}\alpha_s m_s}\right)\sum\limits_{k=1}^{K-1} m_k\hat{Z}_k & \text{if } \ell = K \end{cases}}$$

$$\boxed{Z_\ell = \begin{cases} \nu\alpha_\ell W_1 & \text{if } \ell = 1,\dots,K-1 \\ \frac{1}{q_K}\nu\Big(\sum\limits_{s=1}^{K-1} p_s\alpha_s\Big)W_1 & \text{if } \ell = K \end{cases}}$$

Acknowledgements The author wishes to thank the anonymous referees for careful reading and very helpful comments that resulted in an overall improvement of the paper.

Appendix

Besides [7], from where we draw the definition and notation, the multidimensional reflected fractional Brownian motion (rfBm) process has also been introduced in other papers. For instance, in [12] a single-class queueing network with long-range dependent arrival and service processes is considered, and it is shown that the normalized queue length converges to a d-dimensional rfBm process, being d the number of nodes or servers. The definition of this process, as picked up from [7], is as follows:

Definition 1 (rfBm) A *reflected fractional Brownian motion* on $S = \mathbb{R}^J_+$ associated with data $(x, H, \theta, \Gamma, R)$, where $x, \theta \in S$, $H \in (0, 1)$ and Γ and R are $J \times J$ matrices, being Γ a positive definite one, is a J-dimensional process $W = \{W(t) = (W_1(t), \dots, W_J(t))^T, t \geq 0\}$ such that

(i) W has continuous paths and $W(t) \in S$ for all $t \geq 0$ a.s.,

(ii) $W = X + RY$ a.s., with X and Y two J-dimensional processes defined on the same probability space verifying:

(iii) X is a fBm with associated data (x, H, θ, Γ), that is, it is a continuous Gaussian process starting from x, with mean value $E(X(t)) = x + \theta t$ for any $t \geq 0$ (θ is the *drift vector*), and with covariance function given by: if $t, s \geq 0$,

$$\mathrm{Cov}\big(X(t), X(s)\big) = E\Big(\big(X(t) - (x + \theta t)\big)\big(X(s) - (x + \theta s)\big)^T\Big) = \Gamma_H(s, t)\Gamma,$$

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

(iv) Y has continuous and non-decreasing paths, and for each $j = 1, \dots, J$, a.s., $Y_j(0) = 0$ and $\int_0^\infty 1_{\{W_j(s)>0\}} dY_j(s) = 0$ (that means, Y_j can only increase when W is on face $F_j = \{y \in S = \mathbb{R}^J_+ : y_j = 0\}$).

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

To get an idea, rfBm starts in the interior of S and behaves like a fBm until it touches the boundary of S, formed by faces F_j. Therefore, it is instantaneously "reflected", by preventing the exit of S. For each j, the jth column of the *reflection matrix R* gives the direction of the reflection on face F_j, and component Y_j of process Y gives its intensity. Two fundamental properties of fBm justify the general interest in it from the modelling point of view: fBm is a self-similar process and has long-range dependent increments, which are positively correlated if $1/2 < H < 1$ (the most frequently encountered in modeling).

Remark 6 Proposition 4.2 [20] shows that condition (HR) stated below (denoted as condition (II) there), which is stronger than the *completely-$\mathcal{S}$* assumption, is sufficient to have strong path-wise uniqueness of the solution of the *R-regularization* problem of X.

(HR) *Hypothesis on the Reflection Matrix R*

R can be expressed as $I_J + \Theta$, with Θ a $J \times J$ matrix such that the matrix obtained from Θ by replacing its elements by their absolute values, has spectral radius less than 1.

References

1. Bernard, A., el Kharroubi, A.: Régulations déterministes et stochastiques dans le premier orthant de $\mathbb{R}^n$. Stoch. Stoch. Rep. **34**, 149–167 (1991)
2. Billingsley, P.: Convergence of Probability Measures. Wiley, New York (1968)
3. Bramson, M.: Convergence to equilibria for fluid models of FIFO queueing networks. Queueing Syst. **22**, 5–45 (1996)
4. Bramson, M.: State space collapse with application to heavy traffic limits for multi-class queueing networks. Queueing Syst. **30**, 89–148 (1998)
5. Dai, J.G., Wang, Y.: Nonexistence of Brownian models for certain multi-class queueing networks. Queueing Syst. **13**, 41–46 (1993)
6. Debicki, K., Mandjes, M.: Traffic with an fBm limit: convergence of the stationary workload process. Queueing Syst. **46**, 113–127 (2004)
7. Delgado, R.: A reflected fBm limit for fluid models with ON/OFF sources under heavy traffic. Stoch. Process. Appl. **117**, 188–201 (2007)
8. Harrison, J.M.: Balanced fluid models of multi-class queueing networks: a heavy traffic conjecture. In: Kelly, F.P., Williams, R.J. (eds.) Stochastic Networks. IMA Volumes in Mathematics and its Applications, vol. 71, pp. 1–20. Springer, New York (1995)
9. Harrison, J.M., Nguyen, V.: Brownian models of multi-class queueing networks: current status and open problems. Queueing Syst. **13**, 5–40 (1993)
10. Iglehart, D.L., Whitt, W.: Multiple channel queues in heavy traffic I. Adv. Appl. Probab. **2**, 150–177 (1970)
11. Iglehart, D.L., Whitt, W.: Multiple channel queues in heavy traffic II. Adv. Appl. Probab. **2**, 355–364 (1970)
12. Konstantopoulos, T., Lin, S.J.: Fractional Brownian approximations of stochastic networks. In: Stochastic Networks, Stability and Rare Events. Lecture Notes in Statistics, vol. 117, pp. 257–274 (1996)
13. Majewski, K.: Fractional Brownian heavy traffic approximations of multi-class feedforward queueing networks. Queueing Syst. **50**, 199–230 (2005)

14. Peterson, W.P.: Diffusion approximations for networks of queues with multiple customers types. Math. Oper. Res. **16**, 90–118 (1991)
15. Reiman, M.I.: Open queueing networks in heavy traffic. Math. Oper. Res. **9**(3), 441–458 (1984)
16. Reiman, M.I.: Some diffusion approximations with state space collapse. In: Baccelli, F., Fayolle, G. (eds.) Proc. of the Internat. Seminar on Modeling and Performance Evaluation Methodology. Lecture Notes in Control and Information Sciences, pp. 209–240. Springer, New York (1984)
17. Taqqu, M.S., Willinger, W., Sherman, R.: Proof of a fundamental result in self-similar traffic modeling. Comput. Commun. Rev. **27**, 5–23 (1997)
18. Whitt, W.: Weak convergence theorems for priority queues: preemptive resume discipline. J. Appl. Probab. **8**, 74–94 (1971)
19. Williams, R.J.: On the approximation of queueing networks in heavy traffic. In: Kelly, F.P., Zachary, S., Ziedins, I. (eds.) Stochastic Networks: Theory and Applications, pp. 35–56. Oxford Univ. Press, Oxford (1996)
20. Williams, R.J.: An invariance principle for semimartingale reflecting Brownian motions in an orthant. Queueing Syst. **30**, 5–25 (1998)
21. Williams, R.J.: Diffusion approximations for open multi-class queueing networks: sufficient conditions involving state space collapse. Queueing Syst. **30**, 27–88 (1998)