Noname manuscript No.
(will be inserted by the editor)

State space collapse and stability of queueing networks

Rosario Delgado

the date of receipt and acceptance should be inserted later

Abstract We study the stability of subcritical multi-class queueing networks with feedback allowed and a work-conserving head-of-the-line service discipline. Assuming that the fluid limit model associated to the queueing network satisfies a *state space collapse* condition, we show that the queueing network is stable provided that any solution of an associated linear Skorokhod problem is attracted to the origin in finite time. We also give sufficient conditions ensuring this attraction in terms of the reflection matrix of the Skorokhod problem, by using an adequate Lyapunov function. *State space collapse* establishes that the fluid limit of the queue process can be expressed in terms of the fluid limit of the workload process by means of a *lifting* matrix.

Keywords fluid limit model, Lyapunov function, queueing network, Skorokhod problem, stability, state space collapse

Mathematics Subject Classification (2000) 60K25
60F05, 68G15, 60K20, 90B22

1 Introduction

We consider a multi-class queueing network endowed with J stations that serve K different customer classes, with $K \geq J \geq 1$. Each customer class can be served at only one station, and each station has a single server and an infinite-capacity buffer where customers to be served wait. We allow feedback and assume a work-conserving (or non-idling) HL (*head-of-the-line*) service discipline, that is, servers cannot be idle while there are customers waiting to be served, customers at each station are ordered taking into account their

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
Supported by project MEC-FEDER ref. MTM2009-08869

arrival times, and service is restricted to the oldest customer in each nonempty class. External inter-arrival times and service times are assumed to be i.i.d. and mutually independent.

For such a network, we study the question of its *stability*, that is, the *positive Harris recurrence* of the underlying Markov process describing the network dynamics. It is known that sub-criticality (or *light-traffic*), that is, traffic intensity strictly less than one at each station, is not sufficient in the general setting, although necessary, for the stability of a multi-class queueing network. There has been several attempts to find sufficient conditions in different particular situations, but the problem in general is not yet completely solved. For an interesting account of literature on the stability of queueing networks we refer the reader to [8] and [9] and references therein, and to the monograph [6].

Our approach is quite standard and consists in establishing the stability of the fluid limit model associated to the queueing network, which allows to reduce the initial stochastic problem to a deterministic one, based on the result due to Rybko and Stolyar [17] generalized by Dai (Theorem 4.2 [10]): *"a queueing network is stable if its corresponding fluid limit model is stable"*. For stability of the fluid limit model it is understood that the fluid limit of the queue process reaches zero in a finite amount of time and stays there, regardless of the initial inventory levels. Dai [10] uses a Lyapunov function to obtain the stability of the fluid limit model via the solution of a Skorokhod problem, for the generalized Jackson network ($K = J$) and for the single multi-class station ($J = 1$).

Inspired by [10], the present work describes the relationship between the stability of the queueing network and that of an associated linear Skorokhod problem (meaning that any solution of the problem is attracted to the origin in finite time). This relationship has also been studied by Chen and Zhang [9], which link the stability of a queueing network under a priority service discipline to the feasibility of a set of linear inequalities defined by network parameters (that is, to the stability of a linear Skorokhod problem). In our paper this connection has been made explicit for the first time in the general context, and it has been established through a kind of *state space collapse* condition, stated in terms of the fluid limit model associated to the network. In his works on FIFO and HLPPS (head-of-the-line proportional processor sharing) queueing networks [2] and [3], Bramson relates the stability of a network with the convergence to equilibria (which is the terminology he uses there for *state space collapse* in fluid limits) but using a different approach based on an entropy function.

The phenomenon of *state space collapse* had been first established by Whitt [19] for the single multi-class station but the *term* was first introduced by Reiman [16]. This kind of condition has proved to be a key ingredient in the proof of *heavy-traffic limits* (when the traffic intensity goes to one) for multi-

class queueing networks in the *light-tailed* environment. For references, see for instance [7] or the introduction of [12].

In a rather informal way, we can introduce *state space collapse* as follows: let $W_j(t)$ be the total quantity of time server j needs to complete the service of all customers in queue (or being served) at station j at time t, and $Z_k(t)$ be the quantity of class k customers in queue (or being served) at time t. If we scale the queue process Z and the workload process W by a factor r, we prove in Proposition 1 that both (fluid) limits (u.o.c.)

$$\bar{Z}(t) = \lim_{r\to\infty} \frac{Z(r\,t)}{r} \quad \text{and} \quad \bar{W}(t) = \lim_{r\to\infty} \frac{W(r\,t)}{r} \quad \text{exist almost surely,} \tag{1}$$

and since Z and W satisfy the queueing network equations, then $\bar{Z}$ and $\bar{W}$ are also solution of the deterministic analog of these equations, named the *fluid model equations*. *State space collapse* (in fluid limits) condition is a restriction on process $\bar{Z}(t)$ in the sense that some relationship between those components corresponding to customer classes served at the same station must be satisfied, which is equivalent to say that a deterministic "lifting" operator Δ from $\mathbb{R}^J$ to $\mathbb{R}^K$ exists such that

$$\bar{Z} = \Delta\,\bar{W}\,. \tag{2}$$

Roughly speaking, this assumption means that with the knowledge of the workload process we do not need any additional information about the queue process, that is, for each customer class, we can recover the corresponding queue from the workload at the station at which this class is served, even if more than one customer class is served at the same station.

State space collapse condition appearing in previous works in relation with heavy-traffic limit theorems can be stated as

$$Z^* = \Delta\,W^*$$

where Z^* and W^* are the typically the *diffusion limits*

$$Z^*(t) = \lim_{r\to\infty} \frac{Z(r^2\,t)}{r} \quad \text{and} \quad W^*(t) = \lim_{r\to\infty} \frac{W(r^2\,t)}{r}\,.$$

The heavy-traffic limit establishes that in the light-tailed environment, W^* is a semi-martingale reflecting Brownian motion (SRBM), that is, it is a solution of a Skorokhod problem on the positive orthant associated to a Brownian motion process (see [21]). This result is proved under heavy-traffic assumption (the convergence to 1 of the traffic intensity at each station), state-space collapse $Z^* = \Delta\,W^*$, and the assumption of the *completely-$\mathcal{S}$* condition on the reflection matrix of the Skorokhod problem. The relationship between heavy-traffic limit theorems for a queueing network and its associated fluid limit model has been already considered in several works. See for instance [7] and references therein. In that paper a heavy-traffic limit theorem is proved whose proof uses the *uniform convergence* of the critically loaded fluid model associated to

the queueing network; more specifically, the authors prove that under heavy-traffic, uniform convergence of the fluid model implies *state space collapse* $Z^* = \Delta W^*$. *State space collapse* has shown to be a key ingredient in proving heavy-traffic limit theorems in other scenarios too, as is the case of [12], where it is proved that assuming the other hypotheses hold, *state space collapse* is not only sufficient but necessary to obtain a heavy-traffic limit result.

In this paper we discuss in a general setting the connection between stability of a network, which must be always considered in the *subcritical case*, and stability of the associated linear Skorokhod problem, through the phenomenon of *state space collapse*. More specifically, Theorem 1 shows that under any work-conserving HL service discipline, a subcritical multi-class queueing network satisfying *state space collapse* (in fluid limits) is stable provided that its associated linear Skorokhod problem is stable. Theorem 2 gives general conditions on the reflecting matrix of a linear Skorokhod problem under which it is stable, and as an immediate consequence of both results, Corollary 1 establishes that

"The subcritical queueing network satisfying state space collapse (in fluid limits) is stable provided that the reflection matrix of the associated linear Skorokhod problem is a completely-$\mathcal{S}$ matrix."

The crucial fact in the proof of Theorem 1 is that $\bar{W}$ turns out to be part of a solution of a linear Skorokhod problem, while in [10] a similar fact for $\bar{Z}$ is used instead. Workload seems to be better adapted to the use of the methodology of the Skorokhod problems than the queue process, and this gave us the opportunity to make explicit the relationship between the stability of a multi-class queueing network and that of the linear Skorokhod problem, through *state space collapse* (in fluid limits). On the other hand, a key point in the proof of Theorem 2 is the adequate choice of what is named a *Lyapunov function*. Lyapunov functions are commonly used to prove the stability of queueing systems, and the one introduced in Theorem 2 has proved its usefulness for this objective.

Proposition 2 shows that stability of the fluid limit model (which implies sub-criticality) yields that $\frac{d}{dt}\bar{D}_k(t) = \lambda_k$ for any fluid class k, $\bar{D}$ being the fluid limit of the departure process associated to the network and λ_k being the long run customer class k rate into and out of the corresponding station. As can be seen in Proposition 3, this condition is not only necessary but sufficient for the stability of the fluid limit model associated to any subcritical queueing network under FIFO service discipline.

Besides this, in Theorem 3 we prove that condition (5.3) in [6]: "$\varepsilon > 0$ exists such that $\frac{d}{dt}\bar{D}_k(t) \geq \lambda_k + \varepsilon$ for any k if $\bar{Z}(t) > 0$", is sufficient for the stability of any subcritical multi-class HL queueing network. This result is similar to Theorem 5.2 [6] and we provide a brief proof of it using a different Lyapunov function, by following the ideas of Theorem 2.

The organization of the paper is as follows: in Section 2 we introduce general notations and the definitions of the linear Skorokhod problem and of

its stability. Section 3 introduces the multi-class queueing network we deal with and the queueing network equations that govern the processes associated to the network. Section 4, where our main results are stated and proved, is devoted to the study of the stability of the network: Subsection 4.1 introduces the *fluid limit model* associated to the network and in Subsection 4.2 we introduce *state space collapse* (in fluid limits) and our main results, which give sufficient conditions for the stability of the queueing network. Finally, Section 5 considers three particular cases: the FIFO service discipline, the generalized Jackson network ($K = J$) and the $\bigvee$−system ($J = 1$), and two illustrating examples: a tandem queue and a two parallel-server system.

2 Notations and basic definitions

Vectors will be column vectors, v' means the transpose of a vector (or a matrix) v, and vector inequalities must to be interpreted componentwise. For any $d \geq 1$, given $v = (v_1, \ldots, v_d)' \in \mathbb{R}^d$, hereafter we let $diag(v)$ (or, equivalently, by $diag(v_1, \ldots, v_d)$) stand for the $d \times d$ diagonal matrix with diagonal elements $v_1, \ldots, v_d$. Let $\mathbb{R}^d_+ = \{ v \in \mathbb{R}^d : v \geq 0 \}$ be the d−dimensional positive orthant, and write I for the d−dimensional identity matrix, whatever $d \geq 1$ be. Let $\mathbb{Z}^d_+ = \{ v = (v_1, \ldots, v_d)' \in \mathbb{R}^d : v_i \in \mathbb{Z}_+ \}$.

For a d−dimensional vector, $v = (v_1, \ldots, v_d)'$, let $|v| = \sum_{1 \leq i \leq d} |v_i|$. We will say that a sequence of d−dimensional vectors $\{v^n\}_n$ converges to a d−dimensional vector v if $|v^n - v| \to 0$ as n tends to ∞ (this convergence is equivalent to the convergence in the componentwise sense), and we will denote it simply by $\lim_{n\to\infty} v^n = v$.

For $n \geq 1$, let $\omega_n \colon [0, \infty) \to \mathbb{R}^d$ be right continuous functions having limits on the left on $(0, \infty)$, and let $\omega \colon [0, \infty) \to \mathbb{R}^d$ be continuous. We will say that ω^n converges to ω as $n \to \infty$ *uniformly on compacts* (u.o.c.) if for any $T \geq 0$,

$$||\omega^n(\cdot) - \omega(\cdot)||_T = \sup_{t \in [0,T]} |\omega_n(t) - \omega(t)| \to 0 \,,$$

and it is customary to write $\lim_{n\to\infty} \omega^n = \omega$.

For any process $X = \{X(t), t \geq 0\}$ and state x, X^x denotes the process X starting from x (that is, conditioned to $X(0) = x$).

A square matrix R is called $\mathcal{S}$ matrix if there exists a vector $u > 0$ such that $Ru > 0$, and it is called *completely-$\mathcal{S}$* matrix if all of its principal submatrices are $\mathcal{S}$ matrices.

A non-singular square matrix is called $\mathcal{M}$ matrix if all its off-diagonal entries are non-positive, all its diagonal entries are positive and has non-negative row sums with at least one of them positive. It is known that any $\mathcal{M}$ matrix has inverse and its inverse has all entries non-negative.

Definition 1 (Linear Skorokhod problem)
For any $T \geq 0$, R a $J \times J$ matrix, $\theta \in \mathbb{R}^J$ and $x \in \mathbb{R}^J_+$, we say that the pair of $J-$dimensional stochastic processes defined on the same probability space (W, Y) with continuous paths, is a *solution of the linear Skorokhod problem* $\mathrm{LSP}(\theta, x, R)$ in the positive orthant $\mathbb{R}^J_+$ restricted to $[T, +\infty)$ if:

(i) $W(t) \in \mathbb{R}^J_+ \quad$ for all $t \geq T$,
(ii) $W(t) = x + \theta(t - T) + R\,(Y(t) - Y(T))$ a.s., $t \geq T$,
(iii) Y has non $-$ decreasing paths on $[T, +\infty)$ and for any $j = 1, \ldots, J$, Y_j increases only when W is on face $\{w \in \mathbb{R}^J_+ : w_j = 0\}$, that is, $\int_T^\infty W_j(t)\, d\, Y_j(t) = 0$.

(Then, $x = W(T) \geq 0$.) R is called the *reflection matrix* of the Skorokhod problem.

Remark 1 In the one-dimensional case, the existence of a solution of the Skorokhod problem with reflection matrix R is assured if $R > 0$ for any θ and x (see Theorem I.1.2 [14]). For the $J-$dimensional case, Theorem 2 [1] shows that the *completely-$\mathcal{S}$* property is a sufficient assumption on matrix R (also necessary in some cases) for the existence of the Skorokhod problem. Proposition 4.2 [20] shows that under a stronger assumption on R, there exists strong path-wise uniqueness of the solution (see Remark 6 below).

Definition 2 (Stability of a linear Skorokhod problem)
We say that a linear Skorokhod problem $\mathrm{LSP}(\theta, x, R)$ in the positive orthant $\mathbb{R}^J_+$ restricted to $[T, +\infty)$ and with R being a *completely-$\mathcal{S}$* matrix, is *stable* if $t_R \geq 0$ exists such that for any solution pair (W, Y) of the problem,

$$W(t) = 0 \quad \forall t \geq T + t_R\,|x|\,.$$

Informally speaking, we say then that W is attracted to the origin in a finite time.

3 The multi-class queueing network

In this section we introduce the multi-class queueing network we deal with. Although it can be found in the literature (see [5], [6], [10] and [21], among others), we state here definitions and notations for the convenience of the reader.

We consider an open network composed of $J \geq 1$ single-server stations labeled $j = 1, \ldots, J$, which give some service to the customers. Each station has an infinite-capacity buffer where customers wait for service. We distinguish among customers of classes $k = 1, \ldots, K$, with $K \geq J$. Each customer class can be served at only one station but at each station more than one customer class can be served, and s denotes the many-to-one map from customer classes to stations, $s : \{1, \ldots, K\} \longrightarrow \{1, \ldots, J\}$, $s(k)$ being the station where class k

customers are served. We also introduce the $J \times K$ (deterministic) *constituency matrix* $C = \big(C_{jk}\big)_{j,k}$ by

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

For any j, $s^{-1}(j)$ is the *constituency of station* j, that is, the set of customer classes served at this station. Let $\alpha_k \geq 0$ be the arrival rate (from outside) for class k customers and $\alpha = (\alpha_1, \ldots, \alpha_K)'$. Let $m_k > 0$ be the mean service time for class k customers, $m = (m_1, \ldots, m_K)'$ and $M = diag(m)$.

Although it is not considered any specific distribution for the inter-arrival times nor the service times, we assume the standard mild regularity conditions (which are automatically accomplished by the exponential distribution). See for instance $(1.2) - (1.5)$ in [10].

Customers at each station are served in a *head-of-the-line* (HL) and *work-conserving* (or *non-idling*) discipline. Upon being served at station $s(k)$, with probability $P_{k\ell}$ a class k customer leaving station $s(k)$ goes next to station $s(\ell)$ to be served there as a class ℓ customer. We assume $\sum_{\ell=1}^{K} P_{k\ell} \leq 1$ for any k. Then, $1 - \sum_{\ell=1}^{K} P_{k\ell} \geq 0$ is the probability that upon service at station $s(k)$, a class k customer 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 strictly less than one. Hence, $Q = \big(I - P'\big)^{-1} = I + P' + (P')^2 + \cdots$ is well defined (that is, the network is *open* since all customers eventually leave the network after receiving service at a finite number of stations).

We define λ to be the unique $K-$dimensional vector solution to the *traffic equation* $\lambda = \alpha + P'\,\lambda$ $(= Q\,\alpha)$. Then, λ_k can be interpreted as the long run class k customers rate into and out of station $s(k)$. We also define the *fluid traffic intensity* for station j as

$$\rho_j = \sum_{k \in s^{-1}(j)} \lambda_k\, m_k \quad \big(\text{or } \rho = C\,M\,\lambda\big)\,. \tag{3}$$

Sub-criticality (light-traffic) condition is: $\rho_j < 1$ for any $j = 1, \ldots, J$ or in vector form, $\rho < e$, with $e = (1, \ldots, 1)'$.

The *exogenous arrival process* and the *process of served customers* are respectively defined by: $E_k(t)$ is the cumulative amount of class k customers arrived to the system from outside up to time t, and $S_k(t)$ is the cumulative number of class k customers served at station $s(k)$ up to time t if server $s(k)$ devotes all attention to class k. The *cumulative service time* process $\Upsilon = \{\Upsilon(n)\}_{n \in \mathbb{N}^K}$ is defined by: $\Upsilon_k(n_k)$ is the total amount of service required for the first n_k class k customers in being served (including the remaining service time at time 0 for the first one). Then, $E(0) = S(0) = \Upsilon(0) = 0$. Finally, the *routing* process $\Phi = \{\Phi(n)\}_{n \in \mathbb{N}}$ is defined by: $\Phi_k(n) = \sum_{i=1}^{n} \phi^k(i)$, where $\phi^k(i) = \{\phi_\ell^k(i),\, \ell = 1, \ldots, K\}$ varying $i \in \mathbb{N}$, are i.i.d. random vectors

(independent of the inter-arrival and service time processes), at most with one component equal to 1, the others being 0; the nonzero component corresponds to the class to which the ith class k customer is converted next, with $\phi^k(i) = 0$ indicating the departure of the customer from the network.

The following descriptive processes, A, D, T, Z, W and Y, will be used to measure the performance of the queueing network: $A_k(t)$ is the cumulative number of arrivals (from outside and by feedback) by time t to customer class k and $D_k(t)$ is the cumulative number of departures from class k (to other classes or to outside). $T_k(t)$ is the cumulative amount of service time spent on customer class k by time t. $Z_k(t)$ is the amount of customers of class k in the network (in queue or being served) at time t. $W_j(t)$ denotes the workload or amount of time required by server at station j to complete service of all customers in queue or being served at time t, and $Y_j(t)$ is the cumulative amount of time that the server at station j has been idle in the interval $[0, t]$. By definition, T and Y are nondecreasing processes which depend on the service discipline, and $A(0) = D(0) = T(0) = Y(0) = 0$. These processes are related by means of the following *queueing network equations*:

$$A(t) = E(t) + \sum_{k=1}^{K} \Phi_k \left(D_k(t) \right), \tag{4}$$

$$Z(t) = Z(0) + A(t) - D(t), \tag{5}$$

$$D_k(t) = S_k \left(T_k(t) \right) \quad \text{for any } k = 1, \ldots, K, \tag{6}$$

$$C\,T(t) + Y(t) = e\,t, \tag{7}$$

$$\int_0^\infty W_j(t)\, d\, Y_j(t) = 0 \quad \text{for any } j = 1, \ldots, J, \tag{8}$$

$$W(t) = C\,\Upsilon(Z(0) + A(t)) - C\,T(t). \tag{9}$$

Equation (7) is nothing but the statement that "total time can be split into working time plus idle time". Equation (8) expresses that idle time Y_j can only increase when workload W_j is zero; this is exactly the definition of the assumed *non-idling* or *work-conserving* discipline. Equations (4), (5), (6) and (9) are self-explained.

As we assume throughout this work that the service discipline is *head-of-the-line* (HL), that is, only the oldest customer of each class can receive service, we have the additional queueing network equation:

$$\Upsilon(D(t)) \leq T(t) < \Upsilon(D(t) + e). \tag{10}$$

The previous equations do not specify the service discipline. For each particular HL discipline we would have one more queueing network equation. For example, in the special case of a FIFO (first-in-first-out) discipline, the additional equation is:

$$D_k(t + W_{s(k)}(t)) = Z_k(0) + A_k(t) \quad \text{for any } k = 1, \ldots, K. \tag{11}$$

Corresponding equations for the GHLPPS (generalized-head-of-the-line proportional processor sharing) and SBP (static buffer priority) disciplines can be found in (4.12) and (4.14) [7], respectively."

Let us denote by Ψ the process $\Psi(\cdot) = \big(A(\cdot),\, D(\cdot),\, T(\cdot),\, Z(\cdot),\, W(\cdot),\, Y(\cdot)\big)$.

4 Stability of a network

By definition, a queueing network is **stable** if the associated underlying (strongly) Markov process is **positive Harris recurrent** (see [6], among others, for details). The main criterion for the *positive Harris recurrence* of a Markov process is the limit in Theorem 3.1 [10], which borrow from Theorem 2.1(ii) [15]. Unfortunately this limit is not easy to check directly, so in order to make it practical we will use the standard *fluid approximation*, which is based in the introduction of the *fluid limit model* associated to the queueing network. It is known that the queueing network is stable (that is, the limit in Theorem 3.1 [10] holds) whenever the corresponding fluid limit model is stable (see Theorem 4.2 [10]). Then, in order to obtain conditions ensuring the stability of the network we can reduce ourselves to the research of sufficient conditions for the stability of the fluid limit model.

4.1 The fluid limit model

Next proposition is analogue to Theorem 4.1 [10] (see also Lemma 3.1 in [11]), and it makes it legitimate the definition of the *fluid limit* associated to a queueing network. A initial state for process $\Psi(\cdot)$ will be determined by the initial number of customers of each class in the system, say $z \in \mathbb{Z}_+^K$, since for the other components of $\Psi(\cdot)$, the initial state can be assumed to be 0, except for the workload, whose initial state is a function of z.

Proposition 1 *If $\rho < e$, for almost all sample paths and any sequence of initial states $\{z_n\}_n \subset \mathbb{Z}_+^K$ with $\lim_{n\to\infty} |z_n| = +\infty$, a subsequence $\{z_{n_j}\}_j \vdash \{z_n\}_n$ with $\lim_{j\to\infty} |z_{n_j}| = +\infty$, a vector $\bar{z} \in \mathbb{R}_+^K$, and stochastic processes $\bar{A}$, $\bar{D}$, $\bar{T}$, $\bar{Z}$, $\bar{W}$ and $\bar{Y}$ exist, such that*

$$\lim_{j\to+\infty} \frac{1}{|z_{n_j}|}\, \Psi^{z_{n_j}}(|z_{n_j}|t) = \bar{\Psi}(t) \quad \text{(u.o.c.)} \quad \text{and} \quad \bar{Z}(0) = \bar{z} \tag{12}$$

with $\bar{\Psi}(\cdot) = \big(\bar{A}(\cdot),\, \bar{D}(\cdot),\, \bar{T}(\cdot),\, \bar{Z}(\cdot),\, \bar{W}(\cdot),\, \bar{Y}(\cdot)\big)$. Furthermore, these processes satisfy the following equations, which are the deterministic analogs of the queueing network equations (4)-(10) *obtained by replacing the random vectors*

governing the system (E, S, Υ and Φ) by their respective means:

$$\bar{A}(t) = \alpha\, t + P'\, \bar{D}(t)\,, \tag{13}$$

$$\bar{Z}(t) = \bar{z} + \bar{A}(t) - \bar{D}(t) \tag{14}$$

$$= \bar{z} + \alpha\, t - (I - P')\, M^{-1}\, \bar{T}(t)\,, \tag{15}$$

$$\bar{D}(t) = M^{-1}\, T(t)\,, \tag{16}$$

$$C\, \bar{T}(t) + \bar{Y}(t) = e\, t\,, \tag{17}$$

$$\int_0^\infty \bar{W}_j(t)\, d\, \bar{Y}_j(t) = 0 \quad \text{for all } j = 1, \ldots,\, J\,, \tag{18}$$

$$\bar{W}(t) = C\, M\, (\bar{z} + \bar{A}(t)) - C\, \bar{T}(t)\,. \tag{19}$$

If the service discipline is FIFO, we have from (11) *the additional equation:*

$$\bar{D}_k(t + \bar{W}_{s(k)}(t)) = \bar{z}_k + \bar{A}_k(t) \quad \text{for any } k = 1,\, \ldots,\, K\,. \tag{20}$$

Fluid model equations (13)-(19) (and (20) for the FIFO case) may not have in general a unique solution and can be thought as the "limit" of the corresponding queueing network equations.

Definition 3 (The fluid limit)
Any limit $\big(\, \bar{z},\, \bar{\Psi}(\cdot)\, \big)$ in (12) is called a *fluid limit* associated to the queueing network.

With this definition, Proposition 1 establishes that any *fluid limit* satisfies the *fluid model equations* (13)-(19) (and additionally (20) if the service discipline is FIFO).

We now introduce the standard definition of stability of a fluid limit model. The idea is that the fluid limit model is stable if the component $\bar{Z}(\cdot)$ of any fluid limit reaches zero in a finite amount of time, for any initial state of the system.

Definition 4 (Stability of the fluid limit model)
We say that the *fluid limit model* associated to a queueing network is *stable* if $t_1 \geq 0$, which only depends on α, m and P exists, such that for any fluid limit $\big(\, \bar{z},\, \bar{\Psi}(\cdot)\, \big)$, its $\bar{Z}(\cdot)$ component satisfies

$$\bar{Z}(t) = 0 \;\; \forall t \geq t_1\, |\bar{z}|\,. \tag{21}$$

If restricted to $[T,\, +\infty)$ with $T \geq 0$,

$$\bar{Z}(t) = 0 \;\; \forall t \geq T + t_1\, |\bar{Z}(T)|\,.$$

Remark 2 Taking into account that from (19), (14) and (16) we have

$$\bar{W}(t) = C\, M\, \big(\, \bar{z} + \bar{A}(t) - M^{-1}\, \bar{T}(t)\, \big) = C\, M\, \bar{Z}(t) \tag{22}$$

that can be written for any j as $\bar{W}_j(t) = \sum_{k \in s^{-1}(j)} m_k \, \bar{Z}_k(t)$, and $m_k > 0$, we see that (21) is equivalent to

$$\bar{Z}(t) = 0 \;\; \forall t \geq t_2 \, |\bar{w}| \, ,$$

$\bar{w}$ being defined as $\bar{W}(0)$, with $t_2 \geq 0$ only dependent on α, m and P. Indeed, this is an immediate consequence that for any j, $\bar{w}_j = \sum_{k \in s^{-1}(j)} m_k \, \bar{z}_k$ and then,

$$\big(\min_{k=1,\ldots,K} m_k \big) \, |\bar{z}| \leq |\bar{w}| \leq \big(\max_{k=1,\ldots,K} m_k \big) \, |\bar{z}| \, .$$

By following the same reasoning, we see that in (21) it is equivalent to write $\bar{Z}(\cdot) = 0$ or $\bar{W}(\cdot) = 0$ since $\bar{Z}(t) = 0 \Leftrightarrow \bar{W}(t) = 0$.

Remark 3 Definition 4 corresponds to *strong* stability. Weaker notions, which can be found for instance in [8], may also be of interest, such as that of *weak* stability or *pathwise* stability: we say that the *fluid limit model* associated to the queueing network is *weakly stable* if $\bar{Z}(t) \equiv 0$ when $\bar{z} = 0$, while *pathwise stability* means that the fluid limit of the departures process, $\bar{D}(t)$, must be almost surely equal to the fluid limit of the arrivals process, $\bar{A}(t)$, plus the initial value $\bar{z}$. It is obvious that (strong) stability implies weak stability.

The next result shows that if the fluid limit model associated to a queueing network under any work-conserving HL service discipline is stable, then the fluid model equations have a unique solution starting from some time that depends on the initial state. In particular, we obtain that (strong) stability implies pathwise stability. It could be stated restricted to any parameter set $[T, +\infty)$ with $T \geq 0$, although we have chosen $T = 0$ for simplicity.

Proposition 2 *If the fluid limit model associated to the queueing network is stable, then there exist $t_1 \geq 0$ which only depends on α, m, and P, such that the fluid model equations* (13)-(19) *have a unique solution for any $t \geq t_1 \, |\bar{z}|$, which is given by:*

$$\begin{aligned}
&\bar{D}(t) = Q \, \bar{z} + \lambda \, t \, , \\
&\bar{A}(t) = P' \, Q \, \bar{z} + \lambda \, t \, , \\
&\bar{Z}(t) = \bar{W}(t) = 0 \, , \\
&\bar{T}(t) = M \, Q \, \bar{z} + M \, \lambda \, t \, , \\
&\bar{Y}(t) = -C \, M \, Q \, \bar{z} + (e - \rho) \, t \, .
\end{aligned}$$

As a consequence, $\forall t \geq t_1 \, |\bar{z}|$, $\bar{D}(t) = \bar{A}(t) + \bar{z}$ (pathwise stability) and

$$\bar{D}(t+s) - \bar{D}(t) = \lambda \, s \;\; \forall s \geq 0 \; \forall t \geq t_1 \, |\bar{z}| \, , \tag{23}$$

which is equivalent to $\frac{d}{dt} \, \bar{D}(t) = \lambda$.

Proof: Stability implies sub-criticality and by (15) and (16) and the definition of matrix Q, we have that

$$\bar{Z}(t) = \bar{z} + \alpha\, t - Q^{-1}\, \bar{D}(t)\,,$$

and $t_1 \geq 0$ only depending on α, m and P exists such that for any $t \geq t_1\, |\bar{z}|\,,$

$$0 = \bar{Z}(t) = \bar{z} + \alpha\, t - Q^{-1}\, \bar{D}(t) \Rightarrow \bar{D}(t) = Q\, \bar{z} + \lambda\, t\,, \tag{24}$$

which is due to the fact that $\lambda = Q\, \alpha\,.$ On the other hand, by (13) and (24),

$$\bar{A}(t) = \alpha\, t + P'\, \bar{D}(t) = \alpha\, t + P'\, Q\, \bar{z} + P'\, \lambda\, t = P'\, Q\, \bar{z} + \lambda\, t\,,$$

since $I + P'\, Q = Q\,.$ By (22), $\bar{W} = C\, M\, \bar{Z}(t) = 0$ for any $t \geq t_1\, |\bar{z}|\,,$ and by (16) and (24),

$$\bar{T}(t) = M\, \bar{D}(t) = M\, Q\, \bar{z} + M\, \lambda\, t\,.$$

Finally, this last expression and (17) give that $\bar{Y}(t) = e\, t - C\, \bar{T}(t) = (e - \rho)\, t - C\, M\, Q\, \bar{z}\,.$ $\square$

Remark 4 It is straightforward to deduce from Proposition 2 that also

$$\bar{A}(t+s) - \bar{A}(t) = \lambda\, s \quad \Big(\Leftrightarrow \ \frac{d}{d\, t}\, \bar{A}(t) = \lambda \Big)\,.$$

With analogous reasoning we have that if $\rho < e\,,$ the fluid limit model is weakly stable and $\bar{z} = 0,$ therefore the unique solution of the fluid model equations is given by:

$$\begin{aligned} &\bar{D}(t) = \bar{A}(t) = \lambda\, t\,, \quad \bar{Z}(t) = \bar{W}(t) = 0\,, \\ &\bar{T}(t) = M\, \lambda\, t\,, \quad \bar{Y}(t) = (e - \rho)\, t\,, \quad \forall t \geq 0\,. \end{aligned}$$

4.2 Sufficient conditions for stability

For any fluid limit $(\bar{z},\, \bar{\Psi}(\cdot))$ associated to the queueing network, by (22) we have seen how can be expressed the workload limit $\bar{W}$ in terms of the limit number of customers $\bar{Z}\,.$ In the next definition we introduce *state space collapse* (in fluid limits), which establishes that $\bar{Z}(t)$ in its turn can be expressed in terms of $\bar{W}(t)$ by means of a *"lifting"* deterministic operator.

Definition 5 (State space collapse)
We say that the fluid limit model associated to a queueing network (or that the network itself) satisfies *state space collapse* with "lifting" matrix $\Delta = \big(\Delta_{kj}\big)$ $(\Delta_{kj} \geq 0$ for any $k,\, j\,)$ if $t_\Delta \geq 0$ exists such that for any fluid limit $\big(\, \bar{z},\, \bar{\Psi}(\cdot)\,\big)$, its $\bar{W}(\cdot)$ and $\bar{Z}(\cdot)$ components satisfy that

$$\bar{Z}(t) = \Delta\, \bar{W}(t) \quad \text{for any } \ t \geq t_\Delta\,.$$

Remark 5 The matrix Δ typically depends on the structure of the network and/or on the service discipline as can be seen in [7], where the *uniform convergence with lifting matrix* Δ for a fluid model is introduced. This concept is similar to the *state space collapse* introduced here in the above definition. In the proof of Theorems 3.1-3.3 [7], where three different HL service disciplines are assumed, the authors show that for a single station and if $\rho = 1$, the fluid model converge uniformly with lifting matrices dependent on the service discipline each one. The disciplines considered are: FIFO (first-in-first-out), GHLPPS (generalized-head-of-the-line proportional processor sharing) and SBP (static buffer priority).

State space collapse is not generally an easy condition to check. In [7] the authors say literally that "checking uniform convergence of a fluid model may involve entropy arguments (for FIFO networks of Kelly type and HLPPS networks), comparisons with Markov chains (for single FIFO and GHLPPS networks) and piecewise linear Lyapunov functions (for SBP networks)". Moreover, we can not say much in general on the structure of the lifting matrix, but by (22) we have that *state space collapse* with "lifting" matrix Δ implies that $(I - C\,M\,\Delta)\,\bar{W}(t) = 0$ and therefore $I = C\,M\,\Delta$ if $\bar{W}(t) \neq 0$ for some t. As a consequence, $\Delta_{kj} = 0$ if $j \neq s(k)$, and if δ_k stands for $\Delta_{k\,s(k)}$, we obtain that

$$\sum_{k \in s^{-1}(j)} \delta_k\, m_k = 1 \quad \text{for all } j = 1, \ldots, J\,. \tag{25}$$

In particular, if $s^{-1}(j) = \{k\}$ for some station j, δ_k must be equal to $\frac{1}{m_k}$. The consistency of expression (25) is guaranteed since by (3), $\delta_k = \frac{\lambda_k}{\rho_{s(k)}}$ satisfy it.

From the previous remark, loosely speaking we can say that *state space collapse* assumption expresses that class-k customers contribute with a proportion δ_k to the workload fluid limit $\bar{W}$ at the station at which this class is served. That is, the customer classes served at any station are mixed in a fixed way in the station's queue (there is a fixed concentration of customer classes throughout each queue).

It can be easily seen that *state space collapse* is a necessary condition for stability of the fluid limit model: if $\bar{Z}(t) = 0$ (then, $\bar{W}(t) = 0$), obviously we have that $\bar{Z}(t) = \Delta\,\bar{W}(t)$ (with any "lifting" matrix Δ). Our objective now is to find conditions that jointly with *state space collapse* and sub-criticality be sufficient for the stability. In Theorem 1 we give such a condition: the stability of a linear Skorokhod problem associated to the fluid limit model. Theorem 2 gives sufficient conditions on the reflection matrix for the stability of the linear Skorokhod problem.

Theorem 1 *Assume that the fluid limit model associated to a subcritical queueing network under a HL work-conserving discipline, satisfies state space collapse for some "lifting" matrix* Δ*.*

Then, the queueing network is stable provided that the associated linear Skorokhod problem LSP(θ, x, R) *restricted to* $[t_\Delta, +\infty)$ *is stable, where* $R = (C\,M\,Q\,\Delta)^{-1}$, $\theta = R\,(\rho - e)$ *and* $x \in \mathbb{R}^J_+$ *is arbitrary.*

Proof of Theorem 1:
Fix a fluid limit $(\bar{z}, \bar{\Psi}(\cdot))$. By isolating $\bar{D}(t)$ from (14) we obtain

$$\bar{D}(t) = \bar{z} + \bar{A}(t) - \bar{Z}(t)\,,$$

which can be substituted into (13), and isolating $\bar{A}(t)$ from the resulting expression we have that

$$\bar{A}(t) = \lambda\,t + Q\,P'\,\bar{z} - Q\,P'\,\bar{Z}(t)\,. \tag{26}$$

By *state space collapse* assumption we can replace $\bar{Z}(t)$ by $\Delta\,\bar{W}(t)$ in (26) for any $t \geq t_\Delta \geq 0$, and by substituting into (19) we obtain that for any $t \geq t_\Delta$,

$$\begin{aligned}\bar{W}(t) - \bar{W}(t_\Delta) = C\,M\,\lambda\,(t - t_\Delta) - C\,M\,Q\,P'\,\Delta\,\big(\bar{W}(t) - \bar{W}(t_\Delta)\big)\\ - e\,(t - t_\Delta) + \bar{Y}(t) - \bar{Y}(t_\Delta)\,,\end{aligned}$$

by using (17). And by isolating $\bar{W}(t) - \bar{W}(t_\Delta)$ in its turn from this expression and taking into account that $I + C\,M\,Q\,P'\,\Delta = C\,M\,Q\,\Delta$ and $\rho = C\,M\,\lambda$, we have finally that

$$\begin{aligned}\bar{W}(t) - \bar{W}(t_\Delta) = R\,(\rho - e)\,(t - t_\Delta) + R\,\big(\bar{Y}(t) - \bar{Y}(t_\Delta)\big) \quad \text{or}\\ \bar{W}(t) = \bar{W}(t_\Delta) + R\,(\rho - e)\,(t - t_\Delta) + R\,\big(\bar{Y}(t) - \bar{Y}(t_\Delta)\big)\end{aligned} \tag{27}$$

with $R = (C\,M\,Q\,\Delta)^{-1}$.

Then, we have showed that the pair of processes $(\bar{W}(\cdot), \bar{Y}(\cdot))$ is a solution of the LSP(θ, x, R) restricted to $[t_\Delta, +\infty)$ on the positive orthant, where $\theta = R\,(\rho - e) \in \mathbb{R}^J$ and $x = \bar{W}(t_\Delta) \in \mathbb{R}^J_+$. By the hypothesis of stability of the linear Skorokhod problem, $t_R \geq 0$ exists such that $\bar{W}(t) = 0$ for all $t \geq t_\Delta + t_R\,|\bar{W}(t_\Delta)|$, and the same applies for $\bar{Z}$:

$$\bar{Z}(t) = 0 \quad \text{for all} \quad t \geq t_\Delta + t_1\,|\bar{Z}(t_\Delta)|$$

with $t_1 = t_R\,|C\,M\,e| \geq 0$, since $\bar{W} = C\,M\,\bar{Z}$, which implies that $|\bar{W}(t_\Delta)| \leq |C\,M\,e|\,|\bar{Z}(t_\Delta)|$.

Then, the fluid limit model is stable restricted to $[t_\Delta, +\infty)$ (t_1 is independent of the fixed fluid limit). And finally, as usual, application of Theorem 4.2 [10] ensures the stability of the queueing network. □

Theorem 2 *The linear Skorokhod problem* LSP(θ, x, R) *with parameter set* $[T, +\infty)$ *for any* $T \geq 0$, $\theta = R\,(\rho - e)$ *(with* $\rho < e$*) and* $x \in \mathbb{R}^J_+$, *is stable provided that* R *is an invertible completely-*$\mathcal{S}$ *matrix such that all entries of* R^{-1} *are non-negatives.*

The proof of this result relies on two previous lemmas.

Lemma 1 *Assume that $\rho < e$. Let (W, Y) any solution of* LSP(θ, x, R) *on the positive orthant restricted to some interval $[T, +\infty)$, $T \geq 0$, with R an invertible completely-$\mathcal{S}$ matrix such that all entries of R^{-1} are non-negative. Therefore,*

$$Y(t+s) - Y(s) \leq (e-\rho)\, t \quad \text{for all } s \geq T,\, t \geq 0\,,$$

and hence for any j, $\dfrac{d}{d\,s}\, Y_j(s) \leq 1 - \rho_j$ if $Y(\cdot)$ is differentiable at s.

Proof of Lemma 1: this result is analogous to Lemma 5.1 [10] restricted to the particular case of a linear Skorokhod problem, and can be proved restricted to any parameter set $[T, +\infty)$ in a similar way, by using the non-negativeness of the elements of matrix R^{-1} ($= C\,M\,Q\,\Delta$ here, $= M\,Q$ in [10]), and the fact that

$$-R^{-1}\left(R\,(\rho - e)\right) t = (e-\rho)\, t\,. \quad \square$$

Lemma 2 *[Lemma 5.2 [10]] Let $T \geq 0$ and $f : [T, +\infty) \longrightarrow [0, +\infty)$ be a nonnegative function that is absolutely continuous and let $\kappa > 0$ be a constant. Suppose that for almost surely all points t of differentiability of $f(\cdot)$, we have that $\frac{d}{d\,t}\, f(t) \leq -\kappa$ whenever $f(t) > 0$. Then f is non-increasing and $f(t) = 0$ for any $t \geq T + \frac{f(T)}{\kappa}$.*

(Actually, Lemma 5.2 [10] is enounced with $T = 0$, but the elementary proof of this result is analogous in our setting.)

Proof of Theorem 2:

Lemma 1 can be applied to any solution $(W(\cdot), Y(\cdot))$ of the LSP(θ, x, R) restricted to $[T, +\infty)$ and then for any j,

$$\frac{d}{d\,s}\, Y_j(s) \leq 1 - \rho_j \tag{28}$$

if Y is differentiable at point $s > T$.

We introduce the Lyapunov function

$$g(t) = e'\, R^{-1}\, W(t)\, (\geq 0) \quad \text{for } t \geq T\,, \tag{29}$$

to which we apply Lemma 2. By using that $W(t) = W(T) + \theta\,(t-T) + R\left(Y(t) - Y(T)\right)$ and substituting it into (29),

$$\begin{aligned} g(t) &= e'\, R^{-1}\, W(T) + e'\,(\rho - e)\,(t-T) + e'\left(Y(t) - Y(T)\right) \\ &= g(T) + \sum_{j=1}^{J} \Big((\rho_j - 1)\,(t-T) + \big(Y_j(t) - Y_j(T)\big) \Big)\,, \end{aligned}$$

and thus the points of differentiability of $Y_j(\cdot)$ are the points of differentiability of $g(\cdot)$, and if $t > T$ is one of these points,

$$\frac{d}{d\,t}\,g(t) = \sum_{j=1}^{J} \left((\rho_j - 1) + \frac{d}{d\,t}\,Y_j(t) \right), \tag{30}$$

being $\frac{d}{d\,t}\,g(t)$ non positive by (28).

Let $t > T$ be a point such that $g(t) > 0$ (if any). By definition of g given by (29) and using that all elements of R^{-1} are nonnegative, some $i \in \{1, \ldots, J\}$ exists such that $W_i(t) > 0$. Then, $\displaystyle\int_T^\infty W_i(t)\, dY_i(t) = 0$ implies that $\frac{d}{d\,t}\,Y_i(t) = 0$ since W has continuous paths, and by (30) we have that

$$\frac{d}{d\,t}\,g(t) \le \rho_i - 1 \le \max_{j=1,\ldots,J} \rho_j - 1 \quad (\,< 0 \quad \text{because } \rho < e\,). \tag{31}$$

We have then proved that with $\kappa = 1 - \max\limits_{j=1,\ldots,J} \rho_j > 0$, $\frac{d}{d\,t}\,g(t) \le -\kappa$ at any point $t > T$ of differentiability of $g(\cdot)$ such that $g(t) > 0$. Lemma 2 ensures in this situation that $g(\cdot)$ is non-increasing and $g(t) = 0$ for any $t \ge T + \frac{g(T)}{\kappa}$.

Finally it only remains to take into account that

$$g(T) = e'\,R^{-1}\,W(T) \le e'\,R^{-1}\,e\,|W(T)|\,,$$

and therefore,

$$g(t) = 0 \ \text{ for any } t \ge T + t_R\,|W(T)|\,, \quad \text{with } t_R = \frac{e'\,R^{-1}\,e}{1 - \max_{j=1,\ldots,J} \rho_j} > 0$$

independent of the chosen solution of the LSP(θ, x, R), $(W(\cdot), Y(\cdot))$. As a consequence of (29) and the non-negativeness of the elements of R^{-1}, we deduce that also $W(t) = 0$ for any $t \ge T + t_R\,|W(T)|$ and the linear Skorokhod problem in $[T, \infty)$ is proved to be stable. $\square$

From Theorems 1 and 2, we derive the following immediate corollary since $R = (C\,M\,Q\,\Delta)^{-1}$ (assuming that $C\,M\,Q\,\Delta$ is invertible), and all entries of $R^{-1} = C\,M\,Q\,\Delta$ are non-negative, which implies that the linear Skorokhod problem LSP(θ, x, R) with $\theta = R\,(\rho - e)$, $\rho < e$, is stable if R is a *completely-$\mathcal{S}$* matrix:

Corollary 1 *The subcritical queueing network satisfying state space collapse (in fluid limits) with "lifting" matrix Δ, is stable provided that $R = (C\,M\,Q\,\Delta)^{-1}$ is a completely-$\mathcal{S}$ matrix.*

Remark 6 We can find in the literature different sufficient conditions for the *completely-$\mathcal{S}$* assumption on matrix R:

a) **(HR) condition:**

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

In Proposition 4.2 of [20] it is showed that **(HR)** (condition (II) there), which is stronger than *completely-$\mathcal{S}$*, ensures the strong pathwise uniqueness of the solution of the Skorokhod problem with reflection matrix R.

b) A non-singular $\mathcal{M}$ matrix is *completely-$\mathcal{S}$* (see Lemma 3.1 [8]), and moreover it is invertible and its inverse has all of its entries non-negative.

c) A lower triangular matrix whose inverse exists and is non-negative is *completely-$\mathcal{S}$* . (See Lemma 3.2 [8].)

Theorem 3 below asserts that condition (5.3) in Bramson's book [6] (see also condition (4.9) in [4]) is a sufficient condition for the stability. This result is similar to Theorem 5.2 [6] but we provide a shortener proof which uses a different Lyapunov function and follows the ideas of the proof of Theorem 2 here, although *state space collapse* assumption is not needed at all. Condition (5.3) is not easy to check as shows Theorem 5 [4], where this is done for one of the two families of fluid models considered in the paper.

Theorem 3 *Under each HL work-conserving discipline, a subcritical multiclass queueing network is stable provided that for any fluid limit* $(\bar{z}, \bar{\Psi}(\cdot))$ *associated to the network, its* $\bar{D}(\cdot)$ *and* $\bar{Z}(\cdot)$ *components satisfy that some* $\varepsilon > 0$ *exists such that if* t *is a point of differentiability of* $\bar{D}(\cdot)$ *for which* $\bar{Z}(t) > 0$*, then*

$$\frac{d}{d\,t}\,\bar{D}_k(t) \geq \lambda_k + \varepsilon \quad \textit{for any } \; k = 1, \ldots, K\,.$$

Proof: Fix a fluid limit $(\bar{z}, \bar{\Psi}(\cdot))$ associated to the queueing network, and introduce the Lyapunov function

$$h(t) = e'\,C\,M\,Q\,\bar{Z}(t)\,(\geq 0)\;\forall t \geq 0\,,$$

to which we will apply Lemma 2. By (15) and (16), we can write the Lyapunov function in the following way:

$$h(t) = e'\,C\,M\,Q\,\bar{z} + e'\,C\,M\,\lambda\,t - e'\,C\,M\,\bar{D}(t) = h(0) + e'\,\rho\,t - e'\,C\,M\,\bar{D}(t)\,.$$

Thus, the points of differentiability of $\bar{D}(\cdot)$ are the same as that of $h(\cdot)$, and if t is one of these points,

$$\frac{d}{d\,t}\,h(t) = e'\,\rho - e'\,C\,M\,\frac{d}{d\,t}\,\bar{D}(t)\,.$$

If in addition $h(t) > 0$ (which implies $\bar{Z}(t) > 0$ by definition of function $h(\cdot)$ and the non-negativeness of the elements of $C\,M\,Q$), some $\varepsilon > 0$ exists by hypothesis such that

$$\frac{d}{d\,t}\,h(t) \leq e'\,\rho - e'\,C\,M\,\lambda - e'\,C\,M\,e\,\varepsilon = -e'\,C\,M\,e\,\varepsilon = -\kappa\,,$$

with $\kappa = e' C M e \varepsilon > 0$. By Lemma 2 we obtain that $h(\cdot)$ is a non-increasing function and that $h(t) = 0$ for all $t \geq \frac{h(0)}{\kappa}$. Taking into account that $h(0) = e' C M Q \bar{z}$, this implies that

$$h(t) = 0 \text{ for any } t \geq t_1 |\bar{z}|, \quad \text{with } t_1 = \frac{e' C M Q e}{e' C M e \varepsilon} > 0.$$

Then, the same applies for $\bar{Z}(t)$, and this concludes the proof. □

5 Particular cases and examples

5.1 FIFO service discipline

Theorem 1 shows the stability of the fluid limit model associated to any subcritical multi-class HL queueing network satisfying *state space collapse*, provided that stability of a particular linear Skorokhod problem holds. Stability of the fluid limit implies in turn (23) by Proposition 2. We will show now that reciprocally, condition (23) yields stability of the fluid limit model associated to any subcritical multi-class queueing network under FIFO discipline (and as a consequence, also stability of the queueing network).

Proposition 3 *Under a FIFO service discipline, any subcritical multi-class queueing network is stable if for any fluid limit* $(\bar{z}, \bar{\Psi}(\cdot))$ *associated to the network, its* $\bar{D}(\cdot)$ *component satisfies* (23).

Proof: By (14) and (20), under FIFO service discipline we have that for any fluid limit $(\bar{z}, \bar{\Psi}(\cdot))$, its components satisfy that for any $k = 1, \ldots, K$,

$$\bar{Z}_k(t) = \bar{z}_k + \bar{A}_k(t) - \bar{D}_k(t) = \bar{D}_k(t + \bar{W}_{s(k)}(t)) - \bar{D}_k(t), \tag{32}$$

and assuming (23), $t_1 \geq 0$ exists such that for any $t \geq t_1 |\bar{z}|$ and for any $k = 1, \ldots, K$,

$$\bar{D}_k(t + \bar{W}_{s(k)}(t)) - \bar{D}_k(t) = \lambda_k \bar{W}_{s(k)}(t).$$

Substituting this expression into (32) yields

$$\bar{Z}_k(t) = \lambda_k \bar{W}_{s(k)}(t)$$

and we conclude from Remark 5 and (25) that

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

if $\bar{W}(t) \neq 0$ for some $t \geq t_1 |\bar{z}|$, which contradicts sub-criticality. Consequently, $\bar{W}(t)$ (and $\bar{Z}(t)$) must be equal to zero for any $t \geq t_1 |\bar{z}|$, which finishes the proof. □

5.2 The generalized Jackson network (K=J)

Since when $K = J$ the queueing network considered in this paper has the same structure as the one introduced by Jackson [13], by only differing in the fact that we allow general inter-arrival and service time distributions, which do not have to be necessarily exponentials, it is usually referred as "generalized Jackson network".

For such a network we can easily see that *state space collapse* assumption is accomplished. Indeed, if we assume without loss of generality that $s(j) = j$ for any $j = 1, \ldots, J$, then $C = I$, (22) is $\bar{W} = M\,\bar{Z}$ ($\bar{W}_k = m_k\,\bar{Z}_k$ for any k), and we can isolate $\bar{Z} = M^{-1}\,\bar{W}$ and trivially obtain *state space collapse* with "lifting" matrix $\Delta = M^{-1}$ (customer class k contributes with a proportion $\frac{1}{m_k}$ to the workload fluid limit $\bar{W}_{s(k)}$). Then,

$$R = (C\,M\,Q\,\Delta)^{-1} = I - M\,P'\,M^{-1}$$

satisfies condition **(HR)** in Remark 6, since $M\,P'\,M^{-1}$ has the same spectral radius than matrix P, assumed to be < 1 .Then, R is a *completely-$\mathcal{S}$* matrix and by Corollary 1, under each HL work-conserving discipline,

$$\rho < e \implies \text{stability of the network}\,,$$

which is a well known result (see Theorem 5.1 [10]).

5.3 The $\bigvee$ −system (J=1)

Proposition 4 *In the particular case $J = 1$ (a multi-class station or $\bigvee$ −system), sub-criticality assumption $\rho < 1$ is sufficient for the stability of the queueing network.*

Proof: We can also derive this known result (see Theorem 6.1 [10]) from our approach, by following the reasoning of Theorem 1 but without using *state space collapse* assumption, with the obvious modifications. So, instead of (27) we have in this particular case that the components of any fluid limit $(\bar{z}, \bar{\Psi}(\cdot))$ satisfy that

$$\bar{W}(t) = C\,M\,\bar{Z}(t) = C\,M\,\bar{z} + (\rho - 1)\,t + \bar{Y}(t) + C\,M\,Q\,P'\,\bar{z} - C\,M\,Q\,P'\,\bar{Z}(t)\,,$$

and using that $I + Q\,P' = Q$, we have that

$$C\,M\,Q\,\bar{Z}(t) = C\,M\,Q\,\bar{z} + (\rho - 1)\,t + \bar{Y}(t)\,.$$

This implies that $(C\,M\,Q\,\bar{Z}(\cdot), \bar{Y}(\cdot))$ is a solution to the LSP(θ, x, R) on $[0, +\infty)$ with $\theta = \rho - 1$, $R = 1$ and $x = C\,M\,Q\,\bar{z} \in \mathbb{R}_+$. Theorem 2 ensures the stability of the Skorokhod problem. Therefore, $t_R \geq 0$ exists such that $C\,M\,Q\,\bar{Z}(t) = 0$ for all $t \geq t_R\,C\,M\,Q\,\bar{z}$. Taking into account that $C\,M\,Q$ is a $K-$dimensional vector with all its elements positive (because matrix Q cannot have a column with all of its entries equal to zero), we then have that

$$\bar{Z}(t) = 0 \quad \text{for all} \quad t \geq t_1\,|\bar{z}|\,, \quad \text{with} \quad t_1 = t_R\,|C\,M\,Q| \geq 0\,. \qquad \square$$

5.4 A tandem queue of Kelly type with feedback

Consider a tandem queue in a network with two stations ($J = 2$) and three customer classes ($K = 3$). Class 1 customers enter the system from outside (at rate $\alpha_1 > 0$) and are served at station 1. After being served (at rate $1/m_1$) these customers go into station 2 as class 3 customers, where they are served at rate $1/m_3$. After that, with probability $q \in (0, 1]$, they go outside the network and with probability $p = 1 - q$ they go back to station 1 to be served again, now as class 2 customers, at service rate $1/m_2$, and then go again to station 2 as class 3 customers, and so on. This model, which is a two-stage queueing system, is adequate for situations in which there is recycling, that is, quality control inspection is performed after first stage at the second one, and customers (or items) that do not meet quality standards are sent back to station 1 to be served (or reprocessed) again, at a possible different service rate. The flow of customers through the system is depicted in Fig. 1.

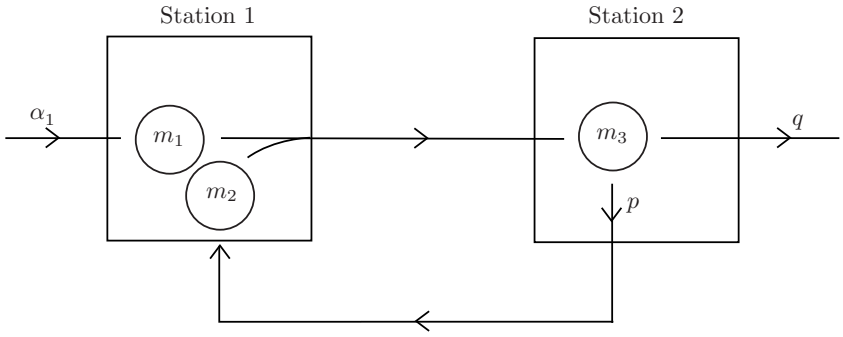

Fig. 1 a tandem queue with feedback

To simplify, we consider the particular case $m_1 = m_2$ (Kelly type). In that scenario, $\alpha_1 > 0$ but $\alpha_2 = \alpha_3 = 0$ (the system only allows external arrivals of class 1 customers). 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). Fluid traffic intensity is $\rho = (\rho_1, \rho_2)'$, with $\rho_1 = \lambda_1 m_1 + \lambda_2 m_2$ and $\rho_2 = \lambda_3 m_3$, being $\lambda = Q\alpha$. Taking into account that $Q = (I - P')^{-1}$, we have that

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

and thus

$$\lambda_1 = \alpha_1\,, \quad \lambda_2 = \frac{p}{q}\,\alpha_1 \quad \text{and} \quad \lambda_3 = \frac{1}{q}\,\alpha_1\,.$$

Then, sub-criticality condition $\rho < e$ is reduced to condition

$$q > \alpha_1\,(m_2 \vee m_3)\,.$$

State space collapse assumption can be expressed as:

$$\bar{Z}_1 = \delta_1\,\bar{W}_1\,,\ \bar{Z}_2 = \delta_2\,\bar{W}_1 \text{ with } \delta_1 + \delta_2 = \frac{1}{m_2} \quad \Big(\Leftrightarrow \delta_2\,\bar{Z}_1 = \delta_1\,\bar{Z}_2\Big) \tag{33}$$

On the other hand,

$$\Delta = \begin{pmatrix} \delta_1 & 0 \\ \delta_2 & 0 \\ 0 & \delta_3 \end{pmatrix} \quad \text{and} \quad C\,M\,Q\,\Delta = \frac{1}{q}\begin{pmatrix} 1 & p\,\frac{m_2}{m_3} \\ \frac{m_3}{m_2} & 1 \end{pmatrix},$$

which turns out to be an invertible matrix since $q > 0$. Condition **(HR)** in Remark 6 holds because

$$R = \big(CMQ\Delta\big)^{-1} = I + \Theta, \text{ with } \Theta = \begin{pmatrix} 0 & -p\,\frac{m_2}{m_3} \\ -\frac{m_3}{m_2} & 0 \end{pmatrix}$$

Therefore, as a consequence of Corollary 1 we have:

Corollary 2 *Assume an HL work-conserving discipline that satisfies state space collapse. Then, the subcritical tandem queue of Kelly type is stable.*

(Compare this result with Lemma 1 [18] when the HL service discipline is assumed to be FIFO and the inter-arrival and service times are exponentially distributed.)

5.5 A two parallel-server system of Kelly type with feedback

Consider now the system with two parallel servers ($J = 2$) whose structure is showed in Fig. 2 below. Class 1 customers, which arrive from outside at rate $\alpha_1 > 0$, are served by server 1, while server 2 serves class 2 customers, arrived from outside at rate $\alpha_2 > 0$. After finishing service at station 1, any customer goes with probability p_1 next to station 2 as a class 3 customer while with probability $q_1 = 1 - p_1$ exits the system.

After finishing service at station 2, independently of its class, any customer goes with probability p_2 next to station 1 as a class 4 customer, while with probability $q_2 = 1 - p_2$ exits the system.

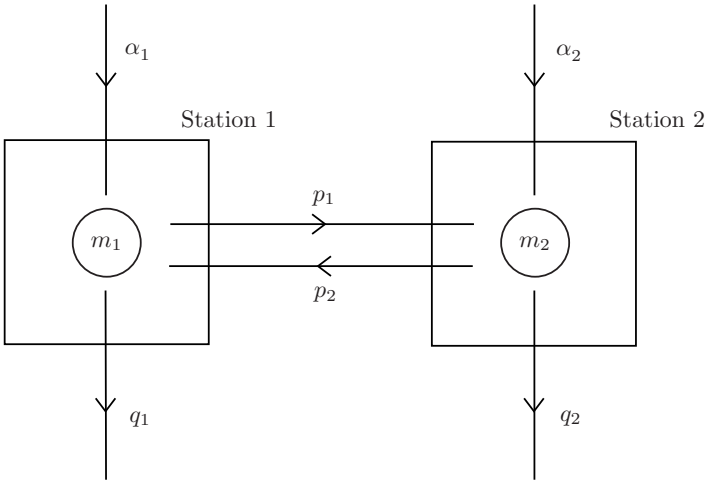

Fig. 2 a two parallel-server system with feedback

We have then $K = 4$ customer classes with $\alpha_3 = \alpha_4 = 0$, and the constituency and the fluid matrices are, respectively,

$$C = \begin{pmatrix} 1\,0\,0\,1 \\ 0\,1\,1\,0 \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} 0 & 0 & p_1 & 0 \\ 0 & 0 & 0 & p_2 \\ 0 & 0 & 0 & p_2 \\ 0 & 0 & p_1 & 0 \end{pmatrix} .$$

Assume that $0 \leq p_1, p_2 \leq 1$ but $p_1\, p_2 < 1$ (to ensure that the spectral radius of P is strictly less than 1). Then, matrix $Q = (I - P')^{-1}$ is

$$Q = \frac{1}{1 - p_1\, p_2} \begin{pmatrix} 1 - p_1\, p_2 & 0 & 0 & 0 \\ 0 & 1 - p_1\, p_2 & 0 & 0 \\ p_1 & p_1\, p_2 & 1 & p_1 \\ p_1\, p_2 & p_2 & p_2 & 1 \end{pmatrix}$$

and thus

$$\lambda = Q\, \alpha = (\alpha_1,\, \alpha_2,\, \frac{p_1\, \alpha_1 + p_1\, p_2\, \alpha_2}{1 - p_1\, p_2},\, \frac{p_2\, \alpha_2 + p_1\, p_2\, \alpha_1}{1 - p_1\, p_2})'$$

The mean service rates, assumed to depend only on the server (and not on the class), are $m_1 > 0$ for server 1 and $m_2 > 0$ for server 2.

Sub-criticality condition $\rho < e$ can be written as two symmetric conditions, by using that $\rho_1 = (\lambda_1 + \lambda_4)\, m_1$ and $\rho_2 = (\lambda_2 + \lambda_3)\, m_2$:

$$\begin{cases} \alpha_1 + p_2\, \alpha_2 < \frac{1 - p_1\, p_2}{m_1} \\ \alpha_2 + p_1\, \alpha_1 < \frac{1 - p_1\, p_2}{m_2} \end{cases} .$$

State space collapse assumption can be expressed as

$$\delta_4\,\bar{Z}_1 = \delta_1\,\bar{Z}_4\,,\ \delta_3\,\bar{Z}_2 = \delta_2\,\bar{Z}_3 \quad \text{with} \quad \delta_1+\delta_4 = \frac{1}{m_1}\,,\ \delta_2+\delta_3 = \frac{1}{m_2} \tag{34}$$

Then,

$$\Delta = \begin{pmatrix} \delta_1 & 0 \\ 0 & \delta_2 \\ 0 & \delta_3 \\ \delta_4 & 0 \end{pmatrix} \quad \text{and} \quad C\,M\,Q\,\Delta = \frac{1}{1-p_1\,p_2}\begin{pmatrix} 1 & \frac{m_1}{m_2}\,p_2 \\ \frac{m_2}{m_1}\,p_1 & 1 \end{pmatrix}$$

which has determinant $\frac{1}{1-p_1\,p_2} \neq 0$. Moreover,

$$R = \big(CMQ\Delta\big)^{-1} = I + \Theta, \text{ with } \Theta = \begin{pmatrix} 0 & -\frac{m_1}{m_2}\,p_2 \\ -\frac{m_2}{m_1}\,p_1 & 0 \end{pmatrix}$$

and the spectral radius of the matrix obtained from Θ by replacing its elements by their absolute values is $\sqrt{p_1\,p_2} < 1$, so condition **(HR)** in Remark 6 is accomplished without restriction on the parameters of the network. Corollary 1 then gives:

Corollary 3 *Assume an HL work-conserving discipline that satisfies state space collapse. Then, the subcritical two parallel-server queue of Kelly type is stable.*

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.

References

1. Bernard, A., el Kharroubi, A. (1991) Régulations déterministes et stochastiques dans le premier orthant de $\mathbb{R}^n$. Stoch. Stoch. Rep. **34**, 149-167.
2. Bramson, M. (1996) Convergence to equilibria for fluid models of FIFO queueing networks. Queueing Syst. **22**, 5-45.
3. Bramson, M. (1996) Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. Queueing Syst. **23**, 1-26.
4. Bramson, M.(1998) Stability of two families of queueing networks and a discussion of fluid limits. Queueing Syst. **28**, 7-31.
5. Bramson, M. (1998) State space collapse with application to heavy traffic limits for multi-class queueing networks. Queueing Syst. **30**, 89-148.
6. Bramson, M. (2008) Stability of queueing networks. Lecture Notes in Mathematics **1950**, École d'Été de Probabilités de Saint-Flour XXXVI-2006, Springer.
7. Bramson, M., Dai, J. G. (2001) Heavy traffic limits for some queueing networks. Ann. Appl. Prob. **11(1)**, 49-90.
8. Chen, H. (1995) Fluid approximations and stability of multiclass queueing networks: work-conserving disciplines. Ann. Appl. Prob. **5(3)**, 637-665.
9. Chen, H., Zhang, H. (2000) Stability of multiclass queueing networks under priority service disciplines. Oper. Res. **48(1)**, 026-037.
10. Dai, J. G. (1995) On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. Ann. Appl. Prob. **5(1)**, 49-77.

11. Dai, J. G., Meyn, S. P. (1995) Stability and convergence of moments for multiclass queueing networks via fluid limit models. IEEE Trans. Aut. Cont. **40(11)**, 1889-1904.
12. Delgado, R. (2008) State space collapse for asymptotically critical multi-class fluid networks. Queueing Syst. **59**, 157-184.
13. Jackson, J. R. (1957) Networks of waiting lines. Oper. Res. **5**, 518-521.
14. N. El Karoui, M. Chaleyat-Maurel, Un problème de réflexion et ses aplications au temps local et aux équations différentielles stochastiques sur $\mathbb{R}$. Cas continu, *Société Mathématique de France, Astérisque* **52-53** (1978) 117-144.
15. Meyn, S. P., Tweedie, R. L. (1994) State-dependent criteria for convergence of Markov chains. Ann. Appl. Probab. **4**, 149-168.
16. Reiman, M. I. (1984) Some diffusion approximations with state space collapse, in: *Proc. of the Internat. Seminar on Modeling and Performance Evaluation Methodology*, Lecture Notes in Control and Information Sciences, eds. F. Baccelli and G. Fayolle, 209-240. Springer, New York.
17. Rybko, A. N., Stolyar, A. L. (1992) Ergodicity of stochastic processes describing the operations of open queueing networks. Problemy Peredachi Informatsii **28**, 2-26.
18. Tang, J., Zhao, Y. Q. (2008) Stationary tail asymptotics of a tandem queue with feedback. Ann Oper Res **160**, 173-189.
19. Whitt, W. (1971) Weak convergence theorems for priority queues: preemptive resume discipline. Jour. Appl. Prob. **8**, 74-94.
20. Williams, R. J. (1998) An invariance principle for semimartingale reflecting Brownian motions in an orthant. Queueing Syst. **30**, 5-25.
21. Williams, R. J. (1998) Diffusion approximations for open multi-class queueing networks: sufficient conditions involving state space collapse. Queueing Syst. **30**, 27-88.