JAMI
Journal of Applied Mathematics and Informatics

J. Appl. Math. & Informatics Vol. **42***(2024), No. 2, pp. 291 - 303*
https://doi.org/10.14317/jami.2024.291

FLUID MODEL SOLUTION OF FEEDFORWARD NETWORK OF OVERLOADED MULTICLASS PROCESSOR SHARING QUEUES

AMAL EZZIDANI*, ABDELGHANI BEN TAHAR AND MOHAMED HANINI

Abstract. In this paper, we consider a feedforward network of overloaded multiclass processor sharing queues and we give a fluid model solution under the condition that the system is initially empty. The main theorem of the paper provides sufficient conditions for a fluid model solution to be linear with time. The results are illustrated through examples.

AMS Mathematics Subject Classification : 60K20, 60K25, 60K30.
Key words and phrases : Fluid model, overloaded queues, multiclass processor sharing.

1. Introduction

Open multiclass queueing networks are an important tool to model several systems in computer systems. Thus, the Processor Sharing (PS) discipline has been widely used in the analysis of such models [14] and [8]. It might be seen as an idealization of time-sharing protocols used in computer systems. The benefit is that, unlike in other disciplines, the system will not be blocked by a big job because jobs are processed simultaneously. Moreover, serving many jobs might result in substantial overhead, which therefore lowers overall performance. This is referred to as an overloaded multiclass processor sharing queue. Furthermore, building on the work [1] for a single queue, there is ample motivation to generalize this model to a feedforward network of such queues. Here, feedforward includes both feedback into the same queue (self-feedback) or feedforward to a higher numbered queue. We call this system a feedforward network of overloaded multiclass processor sharing queues abbreviated as a feedforward network of overloaded MPS queues. As mentioned before, our inspiration for this problem is drawn from a specific case of multiclass single queue system [1]. A natural next

Received December 18, 2022. Revised January 19, 2024. Accepted January 20, 2024.
*Corresponding author.

step would therefore be to incorporate the finite number of multiclass queues in addition to Processor sharing discipline. Furthermore, with the high load in networks, it is very interesting to take into count the case where the queues are overloaded. Moreover, the importance to consider this network lies in its ability to model and optimize complex systems. This type of network enables the representation of environments where multiple processing steps are involved, with a fair allocation of resources among processes. The significance lies in its capacity to capture real-time system dynamics, facilitating the design and optimization of queue management policies to enhance overall system performance. Specifically, it has meaningful applications in areas such as computer networks, service operations, and other contexts where equitable resource allocation is crucial for ensuring efficiency and user satisfaction (e.g. [10, 15, 17]).

To study the evolution of this stochastic system, we use a fluid model solution which produces deterministic and continuous equations. The significance of this proposed technique lies in its ability to bridge the gap between the stochastic and deterministic representations, facilitating a deeper understanding of the system's behavior and enabling the application of well-established mathematical techniques for analysis and prediction.

In the literature, several authors studied the fluid limits of the PS-queue. The authors in [7, 2, 16, 3] began by taking either Poisson arrivals or exponential service times. Work in [7] established that the queue length grows asymptotically linearly with time for overloaded $M/M/1$ processor sharing queue. Authors of [2] considered a fluid approximation for the queue length process. In [16] the authors studied the multiclass $M/M/1$ discriminatory processor sharing queue. In [3] Frolkova and Zwart considered a variation of the processor sharing queue, inspired by freelance job websites where multiple freelancers compete for a single job. They developed fluid limit approximations for the overloaded PS-model with multiple (possibly infinitely many) service stages. Moreover, there are important studies for the $GI/GI/1$ processor sharing queue, in the literature [6, 4, 5, 11, 12, 19], in which the authors established properties of fluid model solution such as existence, uniqueness and asymptotic behaviors for critically loaded and overloaded queues, for zero or nonzero initial customers as well as for limited processor sharing queue. These studies were extended to the multiclass case in [1], where customers may re-enter the queue as a different class. In [9], a critical fluid model was studied. The authors developed a strategy for analyzing the long run behavior of critical fluid model solutions using a notion of relative entropy. In [18] the diffusion approximation of the queue length process was developed under a state space collapse. Furthermore , in [13] the authors described a fluid model with time-varying input that approximates a multiclass many-server queue with time-varying arrivals (specifically, the multiclass $G/GI/N+GI$ queue). They demonstrated the use of the restricted fluid model with a constant input rate to approximately solve scheduling control problems for a queue with a constant arrival rate. In [20] Fluid deterministic models offer precise and manageable optimization formulations to facilitate the development

of practical, implementable policies particularly in the context of time-varying systems. The emphasis lies lie in the practical application of fluid models to address diverse challenges in service and healthcare operations management.

In this paper we consider a feedforward network of overloaded MPS queues, composed of J queues and K customer classes. Each queue has a single server, an infinite storage capacity and can receive K_j customer classes. All customers present in a given queue are served simultaneously according to the egalitarian processor sharing discipline: at any time, customers can be served at a rate that is inverse of the total number of customers in that queue.
Here, our goal is to produce a fluid model solution of feedforward network of overloaded MPS queues under the condition that all queues have no initial mass. As one might expect, to describe the state of the fluid model in the literature [6, 5, 11, 19], the authors used the known load factor. However, the novelty here is the appearance of another reel factor which will be influenced on the system state instead of the known load factor. The main idea is induction on the number of queues. We suppose that the first queue has the load factor strictly greater than one, then by [1] the fluid model solution at that queue grows linearly with time. Thus, to go to the downstream queues $j = 2, \ldots, J$, we must construct and use this new factor. We shown that existence of fluid model solution to downstream queues is based on this factor and that the fluid model solution at each queue grows linearly with time. From this point we note that the adjective "overloaded" means that this factor in each queue is greater than one.

The paper is organized as follows: Section 2 contains the model description, including the definition of fluid model solution. Section 3 presents the main result, which is the construction of the solution (Theorem 3.4). A particular case of processor sharing tandem queue stated in section 4. Section 5 is devoted to some examples. We conclude this work in Section 6.

To state the features of this system, we need to introduce some notations.

Notations: In this paper, we will let J denote the number of queues and K denote the number of customer classes. Then $J, K \in \mathbb{N}$ and $J \leq K$. Further we let $\mathcal{K}_j$ to denote the set of classes belonging to queue j of dimension is K_j. Let $\mathcal{J}$ be the set of all queues and $\mathcal{K} = \cup_{j\in\mathcal{J}}\mathcal{K}_j$ be the set of all classes belonging to queues $1, \ldots, j$. Vectors will be normally arranged as a column. As an exception, the vector e stands for a row vector of ones. The transpose of a vector or matrix A is denoted by A'. For a vector $u \in \mathbb{R}_+^K$, we denote by $u^{\mathbf{j}}$ the sub vector of u whose component indices are in $\mathcal{K}_j$. The $K \times K$ diagonal matrix whose entries are given by the components of K-dimensional vector x will be denoted by $diag(x)$. The space of finite, nonnegative Borel measures on $\mathbb{R}_+$ endowed with the topology of weak convergence is denoted by $\mathcal{M}$. We write $\langle g, \mu\rangle = \int g d\mu$ for $\mu \in \mathcal{M}$ and a Borel measurable function g which is integrable with respect to μ.
For a differentiable function g we write $\dot{g}(x) = \frac{d}{dx}g(x)$.
We will use $\mathbb{E}$ to denote the expectation operator with whatever space the relevant random element is defined on.

2. Fluid Model

In this section, we give a description of a fluid model for a feedforward network of overloaded MPS queues.
The model has three parameters, $\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_+^K$, $\nu = (\nu_1, \ldots, \nu_K)$ the vector of Borel probability measures ν_k on $\mathbb{R}_+^K$ that does not charge the origin ($\nu_k(\{0\}) = 0$) and has a finite first moment denoted β_k, and the nonnegative matrix $P = (p_{kl}, k, l \in \mathcal{K})$. These parameters correspond to parameters in the queueing system. Specifically, α_k corresponds to the long-run average rate at which customers of class k arrive from exterior to the system, the probability measure ν_k corresponds to the distribution of i.i.d. service times for those customers with the distribution function B_k with $\hat{B}_k$ is its Laplace-Stieltjes transform, and P is the routing matrix where p_{kl} is routing probability from class k to class l. Here, we only consider the routing classes from a given class to downstream queues, then, matrix P is upper triangular block matrix, where for each j the sub-matrices $P^{\mathbf{j}}$ and $P^{\mathbf{jj'}}$ are given by: $P^{\mathbf{j}} = (p_{kl},\ k, l \in \mathcal{K}_j)$ and $P^{\mathbf{jj'}} = (p_{kl},\ k \in \mathcal{K}_j, l \in \mathcal{K}_{j'})$.
The triple (α, ν, P) is referred to as the data for the fluid model. The network is assumed to be open, that is, $Q = I + P' + P'^2 + \ldots$ is finite matrix which is equivalent to requirement that $(I - P')$ is invertible with $Q = (I - P')^{-1}$.
Define the vector $\lambda = Q\alpha$. The global arrival rate to the class k is then λ_k and the load factor of each queue j is

$$\rho_j = \sum_{k \in \mathcal{K}_j} \beta_k \lambda_k. \tag{1}$$

A fluid model solution is two real functions A, $D : \mathbb{R}_+ \to \mathbb{R}_+^K$, and one measure-valued vectors of continuous mappings $\mu = (\mu_1, ..., \mu_K) : \mathbb{R}_+ \to \mathcal{M}^K$ such that A and D are continuous and nondecreasing componentwise, and for every $j \in \mathcal{J},\ k \in \mathcal{K}_j$ the following equations hold

$$A_k(t) = \alpha_k t + \sum_{i=1}^{j} \sum_{l \in \mathcal{K}_i} p_{lk} D_l(t), \tag{2}$$

$$\langle 1, \mu_k(t) \rangle = A_k(t) - D_k(t), \tag{3}$$

$$\mu_k(t)([x, \infty)) = \int_0^t \nu_k([x + S_j(s,t), \infty)) dA_k(s) \tag{4}$$

for all $t, x \in \mathbb{R}_+$, where for each $0 \leq s \leq t$

$$S_j(s,t) = \int_s^t \frac{1}{\langle 1, e^{\mathbf{j}} \mu^{\mathbf{j}}(u) \rangle} du \tag{5}$$

where $\inf_{s \leq u \leq t} \langle 1, e^{\mathbf{j}} \mu^{\mathbf{j}}(u) \rangle > 0$ and $S_j(s,t) = +\infty$ for all $0 < s < t$ such that $\langle 1, e^{\mathbf{j}} \mu^{\mathbf{j}}(t) \rangle = 0$ for all $0 < s < t$.
Here, $S_j(s,t)$ is the fluid version of the cumulative service process, and represents the cumulative amount of service time at queue j on a time interval $[s,t]$.

Equation (2) captures the flow conservation equations, taking external arrivals and internal routing into account. Equation (3) relates queue lengths at class k with input/outputs. Equation (4) gives a dynamic equation that describes how the fluid model solution evolves through time.

Let $B(x) = diag\,(B_k(x), k \in \mathcal{K})$, $\hat{B}(x) = diag\left(\hat{B}(x), k \in \mathcal{K}\right)$ and $\beta = diag(\beta_k, k \in \mathcal{K})$. From equations (2)-(5), a fluid model of feedforward networks of MPS-queues is presented by fluid model at each station j which is a triple $(A^{\mathbf{j}}(t), D^{\mathbf{j}}(t), \mu^{\mathbf{j}}(t))$ where it expression, in a vectorial form, is given by:

$$A^{\mathbf{j}}(t) = \alpha^{\mathbf{j}} t + \sum_{i=1}^{j-1} P'^{\mathbf{ji}} D^{\mathbf{i}}(t) + P'^{\mathbf{j}} D^{\mathbf{j}}(t), \tag{6}$$

$$Z^{\mathbf{j}}(t) = A^{\mathbf{j}}(t) - D^{\mathbf{j}}(t), \tag{7}$$

$$\mu^{\mathbf{j}}(t)([x, \infty)) = \int_0^t (I - B^{\mathbf{j}})(x + S_j(s,t)) dA^{\mathbf{j}}(s). \tag{8}$$

particularly, by taking $x = 0$ in the above equation, one have the total mass at each queue j $Z^{\mathbf{j}}(t) = \langle 1, e^{\mathbf{j}} \mu^{\mathbf{j}}(t) \rangle$ is given by

$$Z^{\mathbf{j}}(t) = \int_0^t (I - B^{\mathbf{j}})(S_j(s,t)) dA^{\mathbf{j}}(s). \tag{9}$$

3. Construction of the Fluid Model Solution

In this section, we give a fluid model solution when all queues are initially empty.

Define the set

$$\Lambda_d = \left\{ u \in (\mathbb{R}_+^*)^d \ : \ \sum_{k=1}^{d} u_k \beta_k > 1 \right\}. \tag{10}$$

Recall that $\beta_k := \langle \chi, \nu_k \rangle$, the finite first moment of the Borel probability measure ν_k.

Proposition 3.1. *For all nonnegative $d \times d$-matrix A such that $\sum_{n=0}^{\infty} A'^n = (I - A')^{-1}$, for all vector of the Borel probability measure $\nu = (\nu_1, \ldots, \nu_d)$ and for all $\lambda \in \Lambda_d$, there exists a unique, positive real number θ such that*

$$\theta = e(I - A')(I - \hat{B}(\theta)A')^{-1}(I - \hat{B}(\theta))\lambda, \tag{11}$$

where $\hat{B}_k(\theta) := \mathbb{E}\left[e^{-\theta \nu_k}\right]$ for all $k = 1, \ldots, d$.

Proof. Consider the function $g : \mathbb{R}_+^* \to \mathbb{R}_+^*$ as follows:

$$g(x) = e(I - A')(I - \hat{B}(x)A')^{-1}(I - \hat{B}(x))\lambda.$$

We have $g(0) = 0$ and $\lim_{x \to +\infty} g(x) = e(I - A')\lambda > 0$. Then it suffices to prove that $\lim_{x \to 0^+} \dot{g}(x) > 1$.

According to [1], we have $\left(I - \hat{B}(x)A'\right)^{-1} = \sum_{n=0}^{\infty} \hat{B}^n(x) A'^n$ then

$\left(I-\hat{B}(x)A'\right)^{-1}\left(I-\hat{B}(x)\right)=\sum_{n=0}^{\infty}\hat{B}^{n}(x)A'^{n}-\sum_{n=0}^{\infty}\hat{B}^{n}(x)A'^{n}\hat{B}(x)$.
Using the following formulas

$$\begin{aligned}
\frac{d}{dx}\left(\hat{B}^{n}(x)A'^{n}\right)_{kl} &= n\dot{\hat{B}}_{k}(x)\hat{B}_{k}^{n-1}(x){A'_{kl}}^{n},\\
\frac{d}{dx}\left(\hat{B}^{n}(x)A'^{n}\hat{B}(x)\right)_{kl} &= n\dot{\hat{B}}_{k}(x)\hat{B}_{k}^{n-1}(x){A'_{kl}}^{n}\hat{B}_{l}(x)+\hat{B}_{k}^{n}(x){A'_{kl}}^{n}\dot{\hat{B}}_{l}(x),\\
\lim_{x\to 0^{+}}\dot{\hat{B}}(x) &= -\beta,\\
\lim_{x\to 0^{+}}\frac{d}{dx}\left(\hat{B}^{n}(x)A'^{n}\right) &= -n\beta A'^{n},\\
\lim_{x\to 0^{+}}\frac{d}{dx}\left(\hat{B}^{n}(x)A'^{n}\hat{B}(x)\right) &= -(n\beta A'^{n}+A'^{n}\beta).
\end{aligned}$$

Straightforward computations lead to

$$\lim_{x\to 0^{+}}\dot{g}(x) = e\beta\lambda. \tag{12}$$

Since $\lambda\in\Lambda_d$ then $\lim_{x\to 0^{+}}\dot{g}(x)>1$. This implies that that the function g admits a unique fixed point θ. This proves the proposition. □

A system of a feedforward network of MPS queues is reduced to the standard MPS queue when $J=1$ and $K=K_1$. The above proposition corresponds to parameters in a MPS queue with $d=K_1$, $A=P^{\mathbf{1}}$ and $\lambda=\lambda^{\mathbf{1}}=Q^{\mathbf{1}}\alpha^{\mathbf{1}}$. Then, we define a vector $m^{\mathbf{1}}(\lambda^{\mathbf{1}})=\left(m_k(\lambda^{\mathbf{1}}),k\in\mathcal{K}_1\right)$ as follows:

$$m^{\mathbf{1}}(\lambda^{\mathbf{1}})=(I-P'^{\mathbf{1}})(I-\hat{B}^{\mathbf{1}}(\theta)P'^{\mathbf{1}})^{-1}(I-\hat{B}^{\mathbf{1}}(\theta))\lambda^{\mathbf{1}}, \tag{13}$$

thus $\theta=e\ m^{\mathbf{1}}(\lambda^{\mathbf{1}})$.
In the same way, we will generalise the result above to a feedforward network of MPS queues.
The result in Proposition 3.1 motivates the following definition. Set

$$\Lambda_K=\left\{u\in(\mathbb{R}_+^*)^K\ :\ \sum_{k\in\mathcal{K}_j}u_k\beta_k>1\ \textit{for all}\ j\in\mathcal{J}\right\}, \tag{14}$$

the K-vector $m(u)=\left(m^{\mathbf{1}}(u^{\mathbf{1}}),\ldots,m^{\mathbf{J}}(u^{\mathbf{J}})\right)$ where $m^{\mathbf{j}}(u^{\mathbf{j}})=\left(m_k(u^{\mathbf{j}}),k\in\mathcal{K}_j\right)$ for all $j\in\mathcal{J}$ and the K-vector

$$M(u)=R\ m(u)\ \textit{for all}\ u\in\Lambda_K, \tag{15}$$

where R is a triangular inferior $K\times K$-matrix, diagonal blocs $R^{\mathbf{j}}$ are null matrices and $R^{\mathbf{ji}}=P'^{\mathbf{ji}}Q^{\mathbf{i}}$ for $j>i$.

Proposition 3.2. *Suppose that (α,ν,P) is a data for feedforward network of MPS queues. There exists a unique vector $\bar{\lambda}\in\Lambda_K$ such that,*

$$\bar{\lambda}=\alpha-M(\bar{\lambda})+P'\bar{\lambda}, \tag{16}$$

where $M(\bar{\lambda})$ is defined by (15).

Proof. We apply induction on the numbers of queues. Starting with $j = 1$, Equation (16) is equivalent to $\bar{\lambda}^{\mathbf{1}} = \alpha^{\mathbf{1}} + P'^{\mathbf{1}}\bar{\lambda}^{\mathbf{1}}$.
By Proposition 3.1, since $\sum_{k\in\mathcal{K}_1}\lambda_k\beta_k > 1$, this is analogous to the parameters of a standard MPS queue. Then we have $\bar{\lambda}^{\mathbf{1}} = \lambda^{\mathbf{1}}$ and $m^{\mathbf{1}}(\bar{\lambda}^{\mathbf{1}})$ is given by (13). Next, we assume that the vectors $\bar{\lambda}^{\mathbf{i}} \in \Lambda_{K_i}$ and $m^{\mathbf{i}}(\bar{\lambda}^{\mathbf{i}})$ are well defined for all $i = 1, \ldots, j-1$. We go on to j. In a system of feedforward network of MPS queues, the rate of exogenous arrivals at queue j is given by:

$$a^{\mathbf{j}} = \alpha^{\mathbf{j}} + \sum_{i=1}^{j-1} P'^{\mathbf{ji}}\left(\bar{\lambda}^{\mathbf{i}} - Q^{\mathbf{i}}m^{\mathbf{i}}(\bar{\lambda}^{\mathbf{i}})\right). \tag{17}$$

We pose $\bar{\lambda}^{\mathbf{j}} = Q^{\mathbf{j}}a^{\mathbf{j}}$. This is equivalent to $\bar{\lambda}^{\mathbf{j}} = a^{\mathbf{j}} + P'^{\mathbf{j}}\bar{\lambda}^{\mathbf{j}}$. This implies, by using (17),

$$\bar{\lambda}^{\mathbf{j}} = \alpha^{\mathbf{j}} - \sum_{i=1}^{j-1} P'^{\mathbf{ji}}Q^{\mathbf{i}}m^{\mathbf{i}}(\bar{\lambda}^{\mathbf{i}}) + \sum_{i=1}^{j} P'^{\mathbf{ji}}\bar{\lambda}^{\mathbf{i}}. \tag{18}$$

One observe that $\sum_{i=1}^{j-1} P'^{\mathbf{ji}}Q^{\mathbf{i}}m^{\mathbf{i}}(\bar{\lambda}^{\mathbf{i}}) = \left(M(\bar{\lambda})\right)^{\mathbf{j}}$ where the matrix $M(\bar{\lambda})$ is a K-vector defined by (15).
Since (18) is verified for all $j \in \mathcal{J}$, then we obtain (16). Moreover the vector $m(\lambda)$ is given by:

$$m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}}) = (I - P'^{\mathbf{j}})(I - \hat{B}^{\mathbf{j}}P'^{\mathbf{j}})^{-1}(I - \hat{B}^{\mathbf{j}})\bar{\lambda}^{\mathbf{j}}, \tag{19}$$

for all $j \in \mathcal{J}$. This proves the proposition. □

Denoted by $\bar{\rho}_j := \sum_{k\in\mathcal{K}_j}\bar{\lambda}_k\beta_k$ for all $j \in \mathcal{J}$, then $\bar{\rho}_j > 1$ for all $j \in \mathcal{J}$. Here the system is called *feedforward network of overloaded MPS queues.*

Corollary 3.3. *The link between $\bar{\rho}_j$ and ρ_j is given as: $\bar{\rho}_1 = \rho_1$ and for each $j = 2, \ldots, J$*

$$\bar{\rho}_j \quad = \quad \rho_j - e^{\mathbf{j}}\beta^{\mathbf{j}}\left(M(\bar{\lambda})\right)^{\mathbf{j}}. \tag{20}$$

Proof. Equation (18) implies that for all $j \in \mathcal{J}$ we have $\bar{\lambda}^{\mathbf{j}} = \lambda^{\mathbf{j}} - \sum_{i=1}^{j-1} P'^{\mathbf{ji}}Q^{\mathbf{i}}m^{\mathbf{i}}(\bar{\lambda}^{\mathbf{i}})$.
Multiplying by $e^{\mathbf{j}}\beta^{\mathbf{j}}$, then we obtain (20). □

The following theorem is the main result of this paper.

Theorem 3.4. *Suppose that (α, ν, P) is a data for feedforward network of MPS queues. Assume that there exist $\bar{\lambda} \in \Lambda_K$ and $m(\bar{\lambda})$ defined by (19) for all $j \in \mathcal{J}$. Then a fluid model solution $(A(t), D(t), \mu(t))$ for all $t \geq 0$ is given by:*

$$\left(A^{\mathbf{j}}(t), D^{\mathbf{j}}(t), \mu^{\mathbf{j}}(t)\right) = \left((I - \hat{B}^{\mathbf{j}})^{-1}m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ t, (I - \hat{B}^{\mathbf{j}})^{-1}\hat{B}^{\mathbf{j}}m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ t, \zeta^{\mathbf{j}}t\right) \tag{21}$$

for each $j \in \mathcal{J}$ and for all $t \geq 0$, where for $k \in \mathcal{K}_j$, ζ_k is a measure that is absolutely continuous with respect to Lebesgue measure and $\langle 1, \zeta_k\rangle = m_k(\bar{\lambda}^{\mathbf{j}})$.
As a consequence,

$$Z^{\mathbf{j}}(t) \quad = \quad m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ t \quad \textit{for all } t \geq 0. \tag{22}$$

Proof. Starting with $j = 1$, since $\rho_1 > 1$ then by Theorem 4.2 of [1], a fluid model solution of queue 1 is given by:

$$\left(A^{\mathbf{1}}(t), D^{\mathbf{1}}(t), \mu^{\mathbf{1}}(t)\right) = \left((I - \hat{B}^{\mathbf{1}})^{-1} m^{\mathbf{1}}(\bar{\lambda}^{\mathbf{1}})\ t, (I - \hat{B}^{\mathbf{1}})^{-1} \hat{B}^{\mathbf{1}} m^{\mathbf{1}}(\bar{\lambda}^{\mathbf{1}})\ t, \zeta^{\mathbf{1}} t\right), \tag{23}$$

where, for $k \in \mathcal{K}_1$, ζ_k is given by

$$\zeta_k(A) = \int_A p_k(x)dx \text{ for all measurable } A \subseteq \mathbb{R}_+. \tag{24}$$

ζ_k is Borel measure on $\mathbb{R}_+$ with density

$$p_k(x) = \frac{m_k(\bar{\lambda}^{\mathbf{1}})}{(I - \hat{B}_k)} \int_x^\infty e^{\mathbf{1}} m^{\mathbf{1}}(\bar{\lambda}^{\mathbf{1}}) \exp\left(-e^{\mathbf{1}} m^{\mathbf{1}}(\bar{\lambda}^{\mathbf{1}})\ (y - x)\right) dB_k(y). \tag{25}$$

Note that $\int_0^\infty p_k(x)dx = m_k(\bar{\lambda}^{\mathbf{1}})$ and $\langle 1, \zeta_k \rangle = m_k(\bar{\lambda}^{\mathbf{1}})$.
For $j \geq 2$, since $\bar{\lambda} \in \Lambda_K$, then the vectors $\bar{\lambda}^{\mathbf{j}}$ and $m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})$ are given by (18) and (19) respectively.
For each $k \in \mathcal{K}_j$, define ζ_k a Borel measure on $\mathbb{R}_+$:

$$\zeta_k(A) = \int_A p_k(x)dx \text{ for all measurable } A \subseteq \mathbb{R}_+,$$

with density $p_k(x) = \frac{m_k(\bar{\lambda}^{\mathbf{j}})}{(I - \hat{B}_k)} \int_x^\infty e^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}}) \exp\left(-e^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ (y - x)\right) dB_k(y)$ and one observe that $\int_0^\infty p_k(x)dx = m_k(\bar{\lambda}^{\mathbf{j}})$ and $\langle 1, \zeta_k \rangle = m_k(\bar{\lambda}^{\mathbf{j}})$.
Next, we verify, by inspection, that $A^{\mathbf{j}}$, $D^{\mathbf{j}}$ and $\mu^{\mathbf{j}}$ are solutions of the equations (6), (7) and (8). We have

$$\begin{aligned} a^{\mathbf{j}} t + P'^{\mathbf{j}} D^{\mathbf{j}}(t) &= \left(I - \hat{B}^{\mathbf{j}}\right)^{-1} \left(I - P'^{\mathbf{j}} \hat{B}^{\mathbf{j}}\right) m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}}) + P'^{\mathbf{j}} \left(I - \hat{B}^{\mathbf{j}}\right)^{-1} \hat{B}^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ t \\ &= \left(\left(I - P'^{\mathbf{j}} \hat{B}^{\mathbf{j}}\right) + P'^{\mathbf{j}} \hat{B}^{\mathbf{j}}\right) \left(I - \hat{B}^{\mathbf{j}}\right)^{-1} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ t \\ &= \left(I - \hat{B}^{\mathbf{j}}\right)^{-1} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ t \end{aligned}$$

which is the first component of (21). Moreover,

$$\begin{aligned} A^{\mathbf{j}}(t) - D^{\mathbf{j}}(t) &= (I - \hat{B}^{\mathbf{j}})^{-1} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ t - (I - \hat{B}^{\mathbf{j}})^{-1} \hat{B}^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ t \\ &= (I - \hat{B}^{\mathbf{j}})^{-1} (I - \hat{B}^{\mathbf{j}}) m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ t \\ &= m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ t. \end{aligned}$$

Since $\langle 1, \zeta_k \rangle = m_k(\bar{\lambda}^{\mathbf{j}})\ t$ for $k \in \mathcal{K}_j$, then we have the second component of (21). Afterward that we us check the third component of (21). In the first hand, we have

$$\begin{aligned} \langle 1_{[x,+\infty)}, \zeta_k \rangle &= \int_x^\infty p_k(y)dy \\ &= \frac{e^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})}{1 - \hat{B}_k} m_k(\bar{\lambda}^{\mathbf{j}}) \int_x^\infty \int_y^\infty \exp(e^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ (y - z)) d\nu_k(z)dy \end{aligned}$$

$$
\begin{aligned}
&= \frac{e^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})}{1-\hat{B}_k} m_k(\bar{\lambda}^{\mathbf{j}}) \int_x^\infty \int_x^z \exp(e^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ (y-z)) dy\ d\nu_k(z) \\
&= \frac{m_k(\bar{\lambda}^{\mathbf{j}})}{1-\hat{B}_k} \int_x^\infty \left(1-\exp(e^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ (x-z))\right) d\nu_k(z). \qquad (26)
\end{aligned}
$$

On the other hand, since $S_j(s,t) = \log(t/s)/(e^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}}))$ for all $0 < s < t$,

$$
\begin{aligned}
&\int_0^t \nu_k([x + S_j(s,t), \infty))\ dA_k(s) \\
&= \frac{m_k(\bar{\lambda}^{\mathbf{j}})}{1-\hat{B}_k} \int_0^t \nu_k([x + \log(t/s)/(e^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})), \infty))\ ds \\
&= te^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}}) \frac{m_k(\bar{\lambda}^{\mathbf{j}})}{1-\hat{B}_k} \int_0^\infty \nu_k([x+u, \infty)) \exp(-e^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ u) du \\
&= te^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}}) \frac{m_k(\bar{\lambda}^{\mathbf{j}})}{1-\hat{B}_k} \int_x^\infty \int_v^\infty \exp(-e^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})(v-x)) d\nu_k(y) dv \\
&= t \frac{m_k(\bar{\lambda}^{\mathbf{j}})}{1-\hat{B}_k} \int_x^\infty (1-\exp(-e^{\mathbf{j}} m^{\mathbf{j}}(\bar{\lambda}^{\mathbf{j}})\ (y-x))) d\nu_k(y). \qquad (27)
\end{aligned}
$$

We identify the formulas (26) and (27), we find the result. □

4. Particular Case: Processor Sharing Tandem Queues

A system of processor sharing tandem queues is a particular case of feedforward network of multiclass processor sharing queues, that is, a sequence of J tandem queues where each one has a single class, external customers arrive at the first queue according to a renewal input process with rate α and having a general service times distribution. Upon completing service, customers leave the current queue and enter to the next. We have the following parameters:

$$
p_{ji} = \begin{cases} 1 & \text{if } i = j+1 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad Q_{ji} = \begin{cases} 1 & \text{if } j \geq i \\ 0 & \text{otherwise} \end{cases} \qquad (28)
$$

Particularly, in the following corollary the service in each queue j is distributed as $Exp(\mu_j)$. Then we have $\rho_j = \alpha/\mu_j$ for $j \geq 1$ and $\bar{\rho}_j = \mu_{j-1}/\mu_j$ for $j \geq 2$.

Corollary 4.1. *Suppose that (α, ν, P) is a data for processor sharing tandem queues. Let $j = 1, \ldots, J$, we assume that the service at each queue j follows the exponential distribution with the mean rate of μ_j. If $\alpha > \mu_{j-1} > \mu_j$ for each $j = 2, \ldots, J$ then we have, for all $t \geq 0$,*

$$
\begin{aligned}
(A_1(t)\ ,\ D_1(t)\ ,\ Z_1(t)) &= (\alpha t\ ,\ \mu_1 t\ ,\ (\alpha - \mu_1)t)\,, \\
(A_j(t)\ ,\ D_j(t)\ ,\ Z_j(t)) &= (\mu_{j-1} t\ ,\ \mu_j t\ ,\ (\mu_{j-1} - \mu_j)t) \text{ for } j = 2, \ldots, J.
\end{aligned}
$$

5. Examples

We give in this section some examples.

***Example 1*: Processor Sharing Tandem Queues**
As illustrated in Figure 1, we consider a system of processor sharing tandem queues where $J = 2$. Customers arrive to queue 1 from exterior with rate α. Their service is distributed as $\text{Exp}(\mu_1)$ and when they complete service, they go to queue 2: $P_{12} = 1$. In queue 2, customers have no external arrivals. Their service is distributed as $\text{Exp}(\mu_2)$ and when they complete service, they exit the system: $P_{21} = P_{22} = 0$. We assume that there is zero unit of initial customers in each queue: $Z_1(0) = Z_2(0) = 0$.

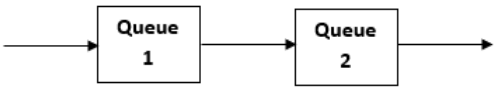

FIGURE 1. Network of processor sharing tandem queues with $J = 2$

We summarize the following parameters:

$$P = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad Q = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \quad \lambda = Q(\alpha, 0)' = (\alpha, \alpha)'.$$

Let $\alpha = 5/4$, $\mu_1 = 1$ and $\mu_2 = 1/4$. Then $\alpha > \mu_1 > \mu_2$ and we have $\rho_1 = 5/4 > 1$ and $\bar{\rho}_2 = 4 > 1$. By Corollary 4.1, for all $t \geq 0$ we have

$$\begin{aligned} (A_1(t)\ ,\ D_1(t)\ ,\ Z_1(t)) &= \left(\frac{5}{4}t\ ,\ t\ ,\ \frac{1}{4}t\right) \\ (A_2(t)\ ,\ D_2(t)\ ,\ Z_2(t)) &= \left(t\ ,\ \frac{1}{4}t\ ,\ \frac{3}{4}t\right). \end{aligned}$$

Example 2
Consider a system of feedforward network of MPS queues with $J = 2$ and $K = 3$. As shown in Figure 2, The fist queue has two classes. Customers of class 1 arrive from the exterior with rate α_1, their service is distributed as $\text{Exp}(\mu_1)$ and when they complete service, they turn into customers of class 2: $P_{12} = 1$. Customers of class 2 have no external arrivals ($\alpha_2 = 0$). Their service is distributed as $\text{Exp}(\mu_2)$ and when they complete service, they move to queue 2: $P_{23} = 1$. The initial situation is that there no unit of fluid of classes 1 and 2: $Z_1(0) = Z_2(0) = 0$. The second queue with single class. Customers have no external arrivals ($\alpha_3 = 0$), their service is distributed as $\text{Exp}(\mu_3)$, and when they complete service, they exit the system: $P_{31} = P_{32} = P_{33} = 0$. There is zero unit of initial customers: $Z_3(0) = 0$.

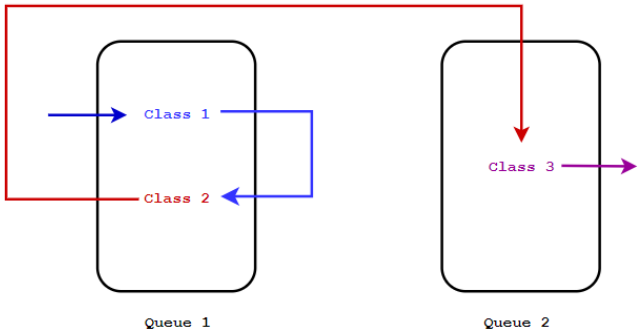

FIGURE 2. Feedforward network of MPS queues with $J = 2$ and $K = 3$

We have the following parameters:

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \quad Q = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \quad \lambda = Q(\alpha_1, 0, 0)' = (\alpha_1, \alpha_1, \alpha_1)'.$$

The load factor of the first queue is $\rho_1 = \alpha_1 (1/\mu_1 + 1/\mu_2)$. For the same values of μ_1 and μ_2 (e.g. $\mu_1 = \mu_2 = 1$), when the arrival rate is $(\alpha_1, \alpha_2) = (1, 0)'$, the load factor is $\rho_1 = 2 > 1$. Calculating

$$\hat{B}^{\mathbf{1}} = \begin{pmatrix} \frac{\mu_1}{m_1+m_2+\mu_1} & 0 \\ 0 & \frac{\mu_2}{m_1+m_2+\mu_2} \end{pmatrix} \quad \textit{and} \quad \left(I - P'^{\mathbf{1}} \hat{B}^{\mathbf{1}}\right)^{-1} = \begin{pmatrix} 1 & 0 \\ \frac{\mu_2}{m_1+m_2+\mu_2} & 1 \end{pmatrix}.$$

Then the vector $(m_1, m_2)' = \left((3-\sqrt{5})/2, \sqrt{5}-2\right)'$. This gives

$$\begin{aligned} (A_1(t)\ ,\ D_1(t)\ ,\ Z_1(t)) &= \left(t\ ,\ \frac{\sqrt{5}-1}{2}t\ ,\ \frac{3-\sqrt{5}}{2}t\right), \\ (A_2(t)\ ,\ D_2(t)\ ,\ Z_2(t)) &= \left(\frac{\sqrt{5}-1}{2}t\ ,\ \frac{3-\sqrt{5}}{2}t\ ,\ (\sqrt{5}-2)t\right). \end{aligned}$$

In queue 2, $\bar{\rho}_2$ is given by (20) as $(\alpha_1 - m_1 - m_2)/\mu_3$. Let $\mu_3 = 1/2$ then $\bar{\rho}_2 = 3 - \sqrt{5} > 1$. Thus $m_3 = (5 - 2\sqrt{5})/2$ and

$$(A_3(t)\ ,\ D_3(t)\ ,\ Z_3(t)) = \left((3-\sqrt{5})t\ ,\ \frac{1}{2}t\ ,\ \frac{5-2\sqrt{5}}{2}t\right).$$

6. Conclusion

In this paper, we introduced a system of a feedforward network of overloaded MPS queues. Our main contribution is obtaining a fluid model solution to this system under the assumption that all queues are initially empty and showing

that this solution is linear with time.
This paper constitutes the first of the two major steps of our study of feedforward network of overloaded MPS queues. It opens the way to the study of the fluid approximation for nonzero initial conditions. Formal proofs of this case will be developed in later works. Moreover, the limit theorem for feed-forward networks of critical MPS-queues not only fortifies the current scholarly inquiry but also serves as a conduit for delving into the investigation of diffusion limits. Furthermore, the upcoming stages of our research will be centered on the expansion and application of this network in practical domains such as computer systems, communication networks, and web servers. The potential implications of our approach pave the way for substantial advancements in the management and optimization of these intricate systems.

Conflicts of interest : The authors declare no conflict of interest.

Data availability : Not applicable

References

1. A. Ben Tahar and A. Jean-Marie, *The fluid limit of the multiclass processor sharing queue*, Queueing Systems Theory Appl. **71** (2012), 347-404.
2. H. Chen, O. Kella and G. Weiss, *Fluid approximations for a processor sharing queue*, Queueing Systems Theory Appl. **27** (1997), 99-125.
3. M. Frolkova, B. Zwart, *Fluid limit of a PS-queue with multistage service*, Probability in the Engineering and Informational Sciences **33** (2019), 1-27.
4. H.C. Gromoll, *Diffusion Approximation for a Processor Sharing Queue in Heavy Traffic*, Ann. Appl. Prob. **14** (2004), 555-611.
5. H.C. Gromoll and L. Kruk, *Heavy traffic limit for a processor sharing queue with soft deadlines*, Ann. Appl. Prob. **17** (2007), 1049-1101.
6. H.C. Gromoll, A.L. Puha and R.J. Williams, *The fluid limit of a heavily loaded processor sharing queue*, Ann. Appl. Probab. **12** (2002), 797-859.
7. A. Jean-Marie and P. Robert, *On the transient behavior of the processor sharing queue*, Queueing Systems Theory Appl. **17** (1994), 129-136.
8. Y. Lu, *A fluid limit for processor-sharing queues weighted by functions of remaining amount of services*, Probability in the Engineering and Informational Sciences (2019), 1-8.
9. A.J. Mulvany, A.L. Puha and R.J. Williams, *Asymptotic behavior of a critical fluid model for a multiclass processor sharing queue via relative entropy*, Queueing Systems Theory Appl. **93** (2019), 351-397.
10. M. Nabe, M. Murata, H. Miyahara, *Analysis and Modelling of World Wide Web Traffic for Capacity Dimensioning of Internet Access Lines*, Performance Evaluation **34** (1998), 249–271.
11. A.L. Puha, A.L. Stolyar and R.J. Williams, *The fluid limit of an overloaded processor sharing queue*, Math. Oper. Res. **31(2)** (2006) 316-350.
12. A.L. Puha and R.J. Williams, *Asymptotic behavior of a critical fluid model for a processor sharing queue via relative entropy*, Stochastic Systems **6** (2016), 251-300.
13. A.L. Puha and Amy R. Ward, *Fluid limits for multiclass many-server queues with general reneging distributions and head-of-the-line scheduling*, Mathematics of Operations Research **47** (2022), 1192-1228.

14. J. Ruuskanen, T. Berner, K. Arzen and A. Cervin, *Improving the mean-field fluid model of processor sharing queueing networks for dynamic performance models in cloud computing*, Performance Evaluation **151** (2021), 69-70.
15. I.I. Trub, *Answers to Queries to Internet: An Optimal Generation Strategy*, Autom. Remote Control **64** (2003), 935–942.
16. I.M. Verloop, U. Ayesta and R. Núñez-Queija, *Heavy-traffic analysis of a multiple-phase network with discriminatory processor sharing*, Operations Research **59** (2011), 648-660.
17. V.M. Vishnevsky, Teoreticheskie osnovy proektirovaniya kompyuternykh setei, *The Theoretical Fundamentals for Design of Computer Networks*, Tekhnosphera, Moscow, 2003.
18. R. Williams, *Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse*, Queueing Systems Theory Appl. **30** (1998), 27-88.
19. J. Zhang, J. Dai and B. Zwart, *Law of large number limits of limited processor-sharing queues*, Math. Oper. Res. **34** (2009), 937-970.
20. N. Zychlinski, *Applications of fluid models in service operations management*, Queueing Syst. Theory Appl. **103** (2023), 161–185.

Amal Ezzidani is a Ph.D. student in Applied Mathematics at Computer, Networks, Mobility and Modeling laboratory, Faculty of Sciences and Technologies, Hassan 1st University, Settat, Morocco. He has completed his Bachelors degree in Mathematics and Applications at the Faculty of Sciences, Semlalia University, Marrakech, Morocco, in 2014, and he received the Masters degree in Mathematics and Applications from Faculty of Sciences and technologies, Hassan 1st University, Settat, Morocco, in 2016. His current research interests Probability, Discrete stochastic processes, queueing network models.

Hassan First University of Settat, Faculty of Sciences and Techniques, Computer, Networks, Mobility and Modeling laboratory: IR2M, 26000, Settat, Morocco.
e-mail: a.ezzidani@uhp.ac.ma

Abdelghani Ben Tahar received his PhD degree in applied mathematics from Hassan II University, Casablanca, Morocco in 2001. He was in an INRIA post-doctoral at Rocquencourt and a CNRS post-doctoral fellows at LIRMM, Montpellier, France, and a lecturer at LMRS (UMR 6085 CNRS-Univ. Rouen). Since 2009 he is a Professor at Hassan 1st University, Settat, Morocco. His current elds of research interests are focusing on networks performance evaluation and queueing network models.

Hassan First University of Settat, Faculty of Sciences and Techniques, Computer, Networks, Mobility and Modeling laboratory: IR2M, 26000, Settat, Morocco.
e-mail: abdelghani.bentahar@uhp.ac.ma

Mohamed Hanini is currently a professor at the department of Mathematic sand computer science in the Faculty of Sciences and techniques, Hassan 1st University Settat, Morocco. He obtained his PhD degree in mathematics and computer in 2013. He is the author and co-author of several papers related to the elds of modeling and performance evaluation of communication networks, cloud computing and security. He participated as TPC member and as an organizing committee member in international conferences and workshops, and he worked as reviewer for several international journals.

Hassan First University of Settat, Faculty of Sciences and Techniques, Computer, Networks, Mobility and Modeling laboratory: IR2M, 26000, Settat, Morocco.
e-mail: mohamed.hanini@uhp.ac.ma