First Frontier Monotonicity for Fluid Models of Multiclass EDF Queueing Networks

Łukasz Kruk(✉)

Institute of Mathematics, Maria Curie-Skłodowska University, Lublin, Poland
lkruk@hektor.umcs.lublin.pl

Abstract. We investigate fluid models of subcritical Earliest Deadline First (EDF) multiclass queueing networks with soft deadlines. For any such model, we show that after a time proportional to the size of the initial condition, the left endpoint of the cumulative fluid mass lead time distribution, called the first frontier, is nondecreasing. Moreover, in the strictly subcritical case, the first frontier actually increases as long as there is fluid mass in the system. Stability of strictly subcritical EDF fluid models and weak stability of subcritical EDF fluid models follow from the above findings as corollaries.

Keywords: Earliest Deadline First · Fluid model · Stability

1 Introduction

In the theory of real-time queueing systems, jobs with individual timing requirements (deadlines) are considered. Several kinds of such systems, reacting differently to customer deadline misses, are known. If deadlines are hard, even a single miss results in a system failure. A firm deadline can be missed, but a late task is considered worthless and hence it is removed from the system, either by the corresponding customer (reneging), or by the system controller. Finally, systems with soft deadlines permit lateness, although they may attempt to minimize it, and use the late jobs. Applications of various real-time queueing models include telecommunication networks, voice and video transmission services, manufacturing systems, real-time vehicle control or scheduling of medical services.

A natural service discipline in the above setting is Earliest Deadline First (EDF), under which the job with the shortest lead time, defined as the difference between its deadline and the current time, is selected for service. For single server, single customer class queueing systems, EDF is known to be optimal with respect to handling real-time service requests, according to a number of performance criteria, see Liu and Layland [25], Panwar and Towsley [27,28], Moyal [26] or Kruk et al. [20].

Q. Zhao and L. Xia (Eds.): VALUETOOLS 2021, LNICST 404, pp. 126–143, 2021.
https://doi.org/10.1007/978-3-030-92511-6_8

To our knowledge, the first informal analysis of stochastic EDF queueing system asymptotics was due to Lehoczky [22–24] and the first mathematically rigorous paper on this topic was the heavy traffic analysis of a G/G/1 EDF queue with soft deadlines by Doytchinov, Lehoczky and Shreve [13]. The latter result was generalized to multiclass feedforward networks by Yeung and Lehoczky [32], while Kruk et al. [20] provided its counterpart for a G/G/1 EDF queue with reneging. Kruk et al. [21] investigated a possibility of a further generalization of the results of [13,32] to multiclass acyclic networks.

Recall that in classical open Jackson networks, the customer interarrival and service times are independent, exponentially distributed and each station serves one customer class (in other words, the jobs at every server are homogeneous). In generalized open Jackson networks, there are renewal arrival processes and independent, identically distributed service times which need not follow exponential distributions, but the customer population at each server is homogeneous, as in the previous case. In this sense, traditional and generalized Jackson networks are single-class models. See, e.g., Chen and Yao [9] for more details. Open multiclass queueing networks, as considered e.g., in Harrison [14], further generalize both these notions by introducing many job classes, differing in their arrival processes, service time distributions and routing through the network, with a many-to-one relation between customer classes and servers. Queueing systems of this type arise in many application areas, e.g., manufacturing and communication networks, service operations or multiprocessor computer systems. Asymptotic theory of multiclass networks is notably more difficult than its counterpart for single-class systems. In particular, it is well known that a strictly subcritical stochastic multiclass queueing network may be unstable, see Rybko and Stolyar [29] or Seidman [30].

Fluid limits, arising from the temporal and spatial scaling of the system's performance processes by the same sequence of constants, and the corresponding fluid models, i.e., formal functional law of large numbers approximations, of EDF queueing systems were investigated by several authors. Decreusefond and Moyal [12] and Atar, Biswas and Kaspi [1] derived fluid limits for nonpreemptive EDF queues with firm deadlines. Atar et al. [3] proposed a unified framework, based on the measure-valued Skorokhod map, for establishing convergence to fluid approximations for queues under a number of service disciplines, including EDF. The scope of their method was extended to a many server EDF queue by Atar, Biswas and Kaspi [2], and recently to generalized EDF Jackson networks, with soft of firm deadlines, by Atar and Shadmi [4].

Results explaining asymptotic behavior of multiclass EDF queueing networks are still in short supply. It is known that strictly subcritical networks with firm deadlines are stable under a wide range of service disciplines, including EDF [15]. In the case of soft deadlines, the following facts were established. Bramson [8] demonstrated that subsequences of suitably rescaled sample paths of the performance processes for an EDF network without preemption converge to fluid limits satisfying the First In System, First Out (FISFO) fluid model equations. Note that the FISFO protocol is a special case of EDF with all the

customer initial lead times being equal (e.g., zero). Bramson [8] also showed that a class of strictly subcritical FISFO fluid models satisfying some additional technical assumptions is stable. Findings of this kind are important, because in many cases stability of fluid models implies stability of the underlying stochastic networks, see Dai [10]. In particular, the results from [8] imply stability of nonpreemptive, strictly subcritical EDF networks. The corresponding results for preemptive multiclass EDF queueing networks were provided by Kruk [15,18], who also extended Bramson's fluid model stability result to arbitrary strictly subcritical FISFO fluid models. Little is known about asymptotics of subcritical (in particular, critical) multiclass EDF, or even FISFO, fluid models. Their behavior seems to be well understood only in the feedforward case, in which such a model converges to its invariant manifold (Kruk [16,17]).

In this paper, we investigate subcritical EDF fluid models of multiclass queueing networks with soft deadlines. Our main object of study is the left endpoint of the cumulative lead time distribution corresponding to the current fluid model state, which we call the first frontier. Our main result states that after a time period proportional to the size of the initial condition, the first frontier in a subcritical EDF fluid model is nondecreasing. Moreover, in the strictly subcritical case, it actually increases at a rate bounded below by some positive constant, depending on the model primitives, as long as there is fluid mass in the system. These findings readily imply stability of strictly subcritical EDF fluid models, with the first frontier acting as the Lyapunov function. (In contrast, in previous work on convergence to equilibria for fluid models of First-In, First-Out (FIFO) networks of Kelly type [5] and head-of-the-line proportional processor sharing networks [6], suitably defined entropies were used as the corresponding Lyapunov functions.) In particular, we provide a new, simpler and perhaps more intuitive proof of the FISFO fluid model stability results from [8,18]. Let us mention, however, that although the original fluid model stability proof from [8] was notably different from ours, it was actually based on a similar observation of "the ability of the FISFO discipline to avoid idling near its oldest customer" ([8, p. 88]). Another simple consequence of our main theorem is weak stability of arbitrary subcritical EDF fluid models. The latter result, related to so-called rate stability of the underlying stochastic network, appears to be new even in the FISFO case.

Our approach was inspired by a recent article [19], in which locally edge minimal fluid models for real-time resource sharing networks were investigated. Such fluid models may be regarded, in some sense, as fluid counterparts of EDF networks with shared resources. A crucial difference between resource sharing networks and multiclass queueing networks considered in this paper is that tasks in a resource sharing network require access to all the resources along their routes at the same time, while customers of a multiclass network are being served by different stations along their routes in succession. The main result of [19] is convergence of subcritical locally edge minimal resource sharing fluid models to the corresponding invariant manifold. The main idea of its proof was to show that the frontiers of such a fluid model increase, until they stabilize at fixed

levels. We hope that under suitable assumptions, this approach can be adapted to fluid models of multiclass EDF queueing networks, opening the way to a more satisfactory theory of diffusion approximations for such networks, in the spirit of Bramson [5–7] and Williams [31]. Our present paper may be considered as a step in this direction.

This paper is organized as follows. In Sect. 2, we briefly describe multiclass EDF queueing networks and we state a definition of the corresponding EDF fluid models. In Sect. 3, preliminary analysis of EDF fluid models is provided. Section 4 contains our main results and their proofs. Section 5 concludes.

2 EDF Queueing Networks and their Fluid Models

2.1 Basic Notation

The following notation will be used throughout the paper. Let $\mathbb{N} = \{1, 2, \ldots\}$ and let $\mathbb{R}$ denote the set of real numbers. The Borel σ-field on $\mathbb{R}$ will be denoted by $\mathcal{B}(\mathbb{R})$. For $a, b \in \mathbb{R}$, we write $a \vee b$ for the maximum of a and b, $a \wedge b$ for the minimum of a and b and a^+ for $a \vee 0$, respectively. The infimum taken over the empty set should be interpreted as ∞. For $n \in \mathbb{N}$, we write $\mathbb{R}^n$ to denote the n-dimensional Euclidean space. All vectors in the paper are to be interpreted as column vectors. For a vector $a = (a_1, ..., a_n) \in \mathbb{R}^n$, let $|a| \triangleq \sum_{i=1}^n |a_i|$. For $a, b \in \mathbb{R}^n$, $a = (a_1, ..., a_n)$, $b = (b_1, ..., b_n)$, the vector $(a_1 b_1, ..., a_n b_n)$ will be denoted by $a \circ b$. Vector inequalities should be interpreted componentwise, e.g., for $a \in \mathbb{R}^n$, $a = (a_1, ..., a_n)$, we write $a \geq 0$ if $a_1 \geq 0$,...,$a_n \geq 0$. For a matrix A, A' denotes the transpose of A. In matrix calculations, I denotes the identity matrix. If $A = [a_{ij}]$ is a square $n \times n$ matrix, then $||A|| = \max_{a \in \mathbb{R}^n : |a|=1} |Aa|$ is the matrix norm of A. It is easy to check that, due to our choice of the l^1 norm in $\mathbb{R}^n$, we have

$$||A|| = \sum_{i=1}^{n} \max_{j=1,...,n} |a_{ij}|. \tag{1}$$

2.2 Stochastic EDF Networks

This paper investigates asymptotic properties of a family of fluid models corresponding to open multiclass queueing networks with the EDF service discipline. To motivate the introduction of these fluid models, we first provide a brief description of the corresponding queueing networks.

We consider a network consisting of J single server stations, indexed by $j = 1, ..., J$. The network is populated by K customer classes, indexed by $k = 1, ..., K$. There is a stationary external arrival process with rate α_k associated with each class k. In particular, if $\alpha_k = 0$, there are no external arrivals to class k. We put $\alpha = (\alpha_1, ..., \alpha_K)$. A customer of class k receives service at a unique station j, written $k \in \mathcal{C}(j)$ or $j = s(k)$. Let m_k be the mean service time for the class k and let $m = (m_1, ..., m_K)$. Upon being served at j, a customer of class k immediately becomes a customer of class l with probability p_{kl}, independently

of its past history. Thus, the probability that a customer of class k leaves the network after completion of service equals $1 - \sum_{l=1}^{K} p_{kl}$. The routing matrix $P = (p_{kl})$ is assumed to be transient, i.e., such that the matrix

$$\Theta = (\theta_{kl}) \triangleq (I - P')^{-1} = I + P' + (P')^2 + ... \tag{2}$$

exists. We define the *total arrival rate* vector

$$\lambda = (\lambda_1, ..., \lambda_K) = \Theta\alpha. \tag{3}$$

Without loss of generality we assume that $\lambda_k > 0$ for each k. Next, we define the *traffic intensity* at station j as

$$\rho_j = \sum_{k \in \mathcal{C}(j)} m_k \lambda_k. \tag{4}$$

When $\rho_j \leq 1$ ($\rho_j < 1$, $\rho_j = 1$) for each j, the network is called *subcritical* (*strictly subcritical*, *critical*).

Class k customers entering the network have nonnegative *initial lead times* with cumulative distribution function (c.d.f.) G_k. Note that

$$G_k(x) = 0, \qquad x < 0. \tag{5}$$

For notational convenience, we define G_k for every $k = 1, ..., K$, including classes with no external arrival streams. For k such that $\alpha_k = 0$, G_k may be chosen in an arbitrary way, subject to (5), and this choice does not affect any further considerations. We put $G = (G_1, ..., G_K)$. To simplify the presentation, we assume that

$$y_k^* \triangleq \sup\{y \in \mathbb{R} : G_k(y) < 1\} < \infty, \qquad k = 1, ..., K. \tag{6}$$

To determine whether customers meet their timing requirements, one must keep track of each customer's lead time, where

$$\text{lead time} = \text{initial lead time} - \text{time elapsed since arrival}$$

for customers coming to the system after time zero and

$$\text{lead time} = \text{initial lead time} - \text{current time}$$

for *initial customers*, i.e., those who are present in the network at time zero.

Customers are served at each station according to the EDF discipline. That is, the customer with the shortest remaining lead time, regardless of class, is selected for service at each station. Late customers (customers with negative lead times) stay in the system until served to completion. Two types of EDF network protocols may be considered. In the *preemptive* case preemption occurs when a customer more urgent than the customer in service arrives (we assume preempt-resume and no set up, switch-over or other type of overhead). In EDF networks *without preemption* customer service continues until he is served to completion, even if a more urgent customer enters the station.

2.3 EDF Fluid Models

Fluid models are deterministic, continuous analogs of queueing networks, in which individual customers are replaced by a divisible commodity (fluid) of K classes, indexed by $k = 1, ..., K$, which change as the fluid moves between stations $j = 1, ..., J$ until it leaves the system. In analogy with customers of the corresponding queueing networks, class k fluid arrives exogeneously to a unique station $j = s(k)$ with rate α_k and initial lead time distribution G_k, it is processed at $s(k)$ with mean service time m_k and changes class to l with transition probability p_{kl} after service completion. As in the case of queueing networks, we say that a fluid model is *subcritical* (*strictly subcritical*, *critical*) if $\rho_j \leq 1$ ($\rho_j < 1$, $\rho_j = 1$) for each j, where ρ_j are given by (4). Fluid models are defined rigorously in terms of the appropriate *fluid model equations*.

Fluid models for EDF queueing networks consist of the six-tuples of vectors

$$\mathfrak{X}(t,s) = (Z(t,s), W(t,s), A(t,s), D(t,s), T(t,s), Y(t,s)), \quad t \geq 0,\ s \in \mathbb{R}, \quad (7)$$

where the vectors $Z(t,s), W(t,s), A(t,s), D(t,s), T(t,s)$ are indexed by $k = 1, ..., K$ and the vector $Y(t,s)$ is indexed by $j = 1, ..., J$. Here $Z_k(t,s)$ denotes the amount of class k fluid with lead times less than or equal to s at time t and $W_k(t,s)$ represents the workload for station $s(k)$ associated with this fluid, i.e., the amount of time necessary for the server $s(k)$ to process it to completion (provided that the station devotes all its capacity to it, without processing any other fluids at the same time). The quantity $A_k(t,s)$ ($D_k(t,s)$) denotes the amount of fluid with lead times at time t less than or equal to s which has arrived at (departed from) class k by time t and $T_k(t,s)$ represents the amount of work devoted to this fluid by server $s(k)$ by time t. Finally, $Y_j(t,s)$ denotes the cumulative idleness by time t at station j with regard to service of fluids with lead times at time t less than or equal to s. The vectors defining $\mathfrak{X}$ are the continuous analogs of the corresponding quantities in the EDF queueing network. We assume that all the components of $\mathfrak{X}$ are continuous and nonnegative, with $A(\cdot, s-\cdot)$, $D(\cdot, s-\cdot)$, $T(\cdot, s-\cdot)$, $Y(\cdot, s-\cdot)$ nondecreasing in each coordinate, $A(0,s) = D(0,s) = T(0,s) = 0$ and $Y(0,s) = 0$ for $s \in \mathbb{R}$. Similarly, we assume that every coordinate of $A(t,\cdot)$, $D(t,\cdot)$, $T(t,\cdot)$, $-Y(t,\cdot)$, $Z(t,\cdot)$ and $W(t,\cdot)$ is nondecreasing for all $t \geq 0$.

The *EDF fluid model equations*, defining the model, are:

$$A(t,s) = \alpha \circ \int_0^t G(s+\eta)\, d\eta + P' D(t,s), \quad (8)$$

$$Z(t,s) = Z(0, t+s) + A(t,s) - D(t,s), \quad (9)$$

$$T(t,s) = m \circ D(t,s), \quad (10)$$

$$\sum_{k \in \mathcal{C}(j)} T_k(t,s) + Y_j(t,s) = t, \qquad j = 1, ..., J, \quad (11)$$

$$Y_j(t, s-t) \text{ can only increase in } t \text{ if } \sum_{k \in \mathcal{C}(j)} Z_k(t, s-t) = 0, \quad j = 1, ..., J, \quad (12)$$

$$W(t,s) = m \circ Z(t,s), \quad (13)$$

where $t \geq 0$, $s \in \mathbb{R}$. A system (7) satisfying the Eqs. (8)–(13) will be called an *EDF fluid model*. The terms α, m, P, G and $\mathcal{C}(j)$, $j = 1, ..., J$, are the model data, given in advance. The above notions were introduced in [16], where more information, including a heuristic explanation of the form $\alpha \circ \int_0^t G(s+\eta)\, d\eta$ of the external arrival process in (8), may be found. Note that while the Eqs. (8)–(9), (11) hold regardless of the underlying scheduling protocol and the Eqs. (10), (13) are satisfied for fluid models of networks under various service disciplines, the Eq. (12) actually characterizes the EDF policy.

An important special case of the EDF fluid model equations may be obtained by putting $G_k(y) = \mathbb{I}_{[0,\infty)}(y)$ for each k, so that $y_k^* = 0$, $k = 1, ..., K$ (see (6)), and (8) takes the form

$$A(t,s) = \alpha\,(t + (s \wedge 0))^+ + P'D(t,s). \tag{14}$$

The equations (9)–(14) will be referred to as the *FISFO fluid model equations*. If we change the coordinates (t,s) to $(t,\tilde{s})$, where $\tilde{s} = s - t$, in (9)–(12), (14), we obtain the FISFO fluid model equations introduced by Bramson [8]:

$$\overline{A}(t,\tilde{s}) = \alpha(t \wedge \tilde{s}) + P'\overline{D}(t,\tilde{s}), \tag{15}$$

$$\overline{Z}(t,\tilde{s}) = \overline{Z}(0,\tilde{s}) + \overline{A}(t,\tilde{s}) - \overline{D}(t,\tilde{s}), \tag{16}$$

$$\overline{D}_k(t,\tilde{s}) = \overline{T}_k(t,\tilde{s})/m_k, \qquad k = 1, ..., K, \tag{17}$$

$$\sum_{k \in \mathcal{C}(j)} \overline{T}_k(t,\tilde{s}) + \overline{Y}_j(t,\tilde{s}) = t, \quad j = 1, ..., J, \tag{18}$$

$$\overline{Y}_j(t,\tilde{s}) \text{ can only increase in } t \text{ when } \sum_{k \in \mathcal{C}(j)} \overline{Z}_k(t,\tilde{s}) = 0, \tag{19}$$

for $t, \tilde{s} \geq 0$. In (15)–(19), the coordinate $\tilde{s}$ represents the arrival times of customers (fluids) to the network, rather than their lead times.

If we take *fluid limits*, i.e., the limits of sample paths along subsequences under scaling which is linear in both time and space (called *fluid* or *hydrodynamic scaling*) obtained from a single EDF network, then the initial lead time distributions disappear in the limit, giving rise to the FISFO fluid models. This was shown in [8] and [15,18] for nonpreemptive and preemptive EDF networks, respectively. Here, following [16,17], we study somewhat more general EDF fluid models, satisfying (8)–(13) with nontrivial lead time distributions G_k. The latter setup may turn out to be useful in asymptotic analysis of a *sequence* of EDF networks in which the initial lead time distributions dilate with the same rate as the space scaling parameter. Because of such lead time scaling, used e.g., in [13,20,21,32], the customer lead times are more "realistic", i.e., they are of the same order as the queue lengths and the sojourn times.

Let

$$\mathbf{K} = \left\{(k_1, ..., k_n) : n \in \mathbb{N}, k_1, ..., k_n \in \{1, ..., K\}, \alpha_{k_1} p_{k_1 k_2} ... p_{k_{n-1} k_n} > 0\right\},$$

where $p_{k_1 k_2} ... p_{k_{n-1} k_n}$ should be interpreted as 1 for $n = 1$. The elements of $\mathbf{K}$ will be called *multi-indices*. They represent paths of finite length which are being

followed with positive probability by customers (fluids) since their arrival to the network. For $\mathbf{k} = (k_1, ..., k_n) \in \mathbf{K}$, let $p_{\mathbf{k}} = p_{k_1 k_2} ... p_{k_{n-1} k_n}$ and $\alpha_{\mathbf{k}} = \alpha_{k_1} p_{\mathbf{k}}$. Also, for $\mathbf{k}$ as above, let $b(\mathbf{k}) = k_1$ and $e(\mathbf{k}) = k_n$ be the beginning and the end of the path $\mathbf{k}$, respectively. For $\mathbf{k} \in \mathbf{K}$, $k \in \{1, ..., K\}$ and $j \in \{1, ..., J\}$, we write $\mathbf{k} \in \tilde{\mathcal{C}}(k)$ if $e(\mathbf{k}) = k$ and $\mathbf{k} \in \overline{\mathcal{C}}(j)$ if $e(\mathbf{k}) \in \mathcal{C}(j)$.

In what follows, we fix an EDF queueing network under consideration, which, for the sake of construction of the corresponding fluid models, is completely determined by α, m, P, G, together with the sets $\mathcal{C}(j)$, $j = 1, ..., J$. We also assume a condition compatible to (6) on the initial states under consideration: there exists a constant $C_0 < \infty$ such that

$$Z(0, C_0) = \lim_{s \to \infty} Z(0, s). \tag{20}$$

In particular, the support of the measure with the distribution function $Z_k(0, \cdot)$ is bounded above by C_0 for each k. This is the case, for example, if all the mass present in the system at time zero has arrived at some prior times with initial lead time distributions G_l, $l = 1, ..., K$, upon arrival and $C_0 = \max_{k:\alpha_k > 0} y_k^*$. In most cases, without loss of generality we may additionally assume that

$$C_0 \geq \max_{k:\alpha_k > 0} y_k^*. \tag{21}$$

3 Preliminary Analysis

Consider an EDF fluid model and let $t \geq 0$, $s \in \mathbb{R}$. From (9) we have

$$D(t, s) = Z(0, t + s) - Z(t, s) + A(t, s). \tag{22}$$

Plugging this to (8), we obtain

$$(I - P')A(t, s) = \alpha \circ \int_0^t G(s + \eta)\, d\eta + P' \left(Z(0, t + s) - Z(t, s) \right),$$

which, after multiplication by Θ from the left and using (2), yields

$$A(t, s) = \Theta \left(\alpha \circ \int_0^t G(s + \eta)\, d\eta + P' \left(Z(0, t + s) - Z(t, s) \right) \right). \tag{23}$$

This, in turn, together with (22) and the identity

$$I + \Theta P' = \Theta \tag{24}$$

resulting from (2), implies

$$D(t, s) = \Theta \left(\alpha \circ \int_0^t G(s + \eta)\, d\eta + Z(0, t + s) - Z(t, s) \right). \tag{25}$$

It is not hard to see that for any EDF fluid model $\mathfrak{X}$, the function $\mathfrak{X}(t, s - t)$ is Lipschitz continuous in t for all $t \geq 0$, $s \in \mathbb{R}$. The relations (20)–(21), (25)

and monotonicity of $D(t,\cdot)$, $Z(t,\cdot)$ imply that $D(t,s)$ is Lipschitz continuous in s as long as $t+s \geq C_0$. This can be proved as in [17, p. 543]. Consequently, $\mathfrak{X}(t,s)$ is Lipschitz continuous in both t and s for $t+s \geq C_0$. Under the additional assumption that $Z(0,\cdot)$ is Lipschitz continuous, $\mathfrak{X}(t,s)$ is Lipschitz in both variables for all $t \geq 0$, $s \in \mathbb{R}$.

We will now make an intuitively clear observation that the conditions (6) and (20)–(21) imply the absence of mass with lead times greater than C_0 in the system.

Lemma 1. *For $k = 1, \ldots K$, $t \geq 0$ and $s \geq C_0$, we have*

$$D_k(t,s) = D_k(t,C_0), \tag{26}$$
$$Z_k(t,s) = Z_k(t,C_0). \tag{27}$$

Proof: Fix $k = 1, \ldots K$, $t \geq 0$, $s \geq C_0$. Let $l \in \{1, \ldots, K\}$ be such that $\theta_{kl}\alpha_l > 0$ (or, equivalently, there exists $\mathbf{k} \in \mathbf{K}$ with $b(\mathbf{k}) = l$ and $e(\mathbf{k}) = k$). By (6) and (21), we have

$$\int_0^t G_l(s+\eta)\, d\eta = \int_0^t G_l(C_0+\eta)\, d\eta. \tag{28}$$

Now let $l \in \{1, \ldots, K\}$ be arbitrary. By (20), we have

$$Z_l(0, t+s) = Z_l(0, t+C_0). \tag{29}$$

The Eqs. (28)–(29), together with (25) and the fact that $D_k(t,\cdot)$ is nondecreasing, implies that $\{\Theta Z(t,s)\}_k \leq \{\Theta Z(t,C_0)\}_k$. However, $Z(t,\cdot)$ is also nondecreasing and all the entries of Θ are nonnegative by (2), so

$$Z(t,s) \geq Z(t,C_0) \tag{30}$$

and $\Theta Z(t,s) \geq \Theta Z(t,C_0)$. We have obtained

$$\{\Theta Z(t,s)\}_k = \{\Theta Z(t,C_0)\}_k. \tag{31}$$

The Eqs. (25), (28)–(29) and (31), imply (26). Finally, (30), together with (2) and (31), proves (27). □

For $k = 1, \ldots K$ and $t \geq 0$, define the fluid queue lengths

$$Q_k(t) = \lim_{s\to\infty} Z_k(t,s) = Z_k(t,C_0), \tag{32}$$

where the last equality follows from Lemma 1. Let $Q(t) = (Q_1(t), \ldots, Q_K(t))$.

For $k = 1, \ldots, K$ and $y \in \mathbb{R}$, let us define $H_k(y) = \int_y^\infty (1 - G_k(\eta))\, d\eta$. Each function H_k is finite by (6). We put $H = (H_1, \ldots, H_K)$. The following lemma generalizes the second part of Proposition 4.1 from [16] to EDF fluid models which are not necessarily invariant.

Lemma 2. *For $k = 1, ...K$, $s_1 \le s_2$ and $t \ge C_0 - s_1$, the condition*

$$D_k(t, s_2) = D_k(t, s_1) \tag{33}$$

implies

$$\{\Theta\left(Z(t, s_2) - Z(t, s_1)\right)\}_k = \{\Theta\left(\alpha \circ [H(s_1) - H(s_2)]\right)\}_k . \tag{34}$$

Proof: Fix $s_1 \le s_2$ and let $t \ge 0$. From (25), we get

$$\begin{aligned}\Theta\left(Z(t, s_2) - Z(t, s_1)\right) = \Theta\left(Z(0, t + s_2) - Z(0, t + s_1)\right) \\ + \Theta\left(\alpha \circ \left(\int_{t+s_1}^{t+s_2} G(\eta)\, d\eta - \int_{s_1}^{s_2} G(\eta)\, d\eta\right)\right) \\ - \left(D(t, s_2) - D(t, s_1)\right).\end{aligned}$$

This, together with (6) and (20)–(21), implies that for $t \ge C_0 - s_1$,

$$\Theta\left(Z(t, s_2) - Z(t, s_1)\right) = \Theta\left(\alpha \circ [H(s_1) - H(s_2)]\right) - \left(D(t, s_2) - D(t, s_1)\right). \tag{35}$$

The Eq. (34) is an immediate consequence of (33) and (35). □

The following proposition is the main result of this section. It implies that after a time period proportional to the size of the initial condition, the initial fluid mass leaves the system.

Proposition 1. *Let $m_0 = \min_{k=1,...,K} m_k$. There exists a time Υ such that*

$$\Upsilon \le C_0 + m_0(|\Theta Q(0)| + |\lambda| C_0), \tag{36}$$

$$Z(t, C_0 - t) = 0, \qquad t \ge \Upsilon. \tag{37}$$

Proof: Let

$$\Upsilon = \inf\{t \ge C_0 : Z(t, C_0 - t) = 0\}. \tag{38}$$

We will first prove the bound (36). If $\Upsilon = C_0$, there is nothing to prove. Assume that $\Upsilon > C_0$. For every $t \in [C_0, \Upsilon)$, there exists $k \in \{1, ..., K\}$ such that $Z_k(t, C_0 - t) > 0$. By continuity of the fluid model, for some $\epsilon > 0$ we have $Z_k(\tilde{t}, C_0 - \tilde{t}) > 0$ for every $\tilde{t} \in [C_0, \Upsilon) \cap (t - \epsilon, t + \epsilon)$ and thus, by (12), for $j = s(k)$, the function $Y_j(\cdot, C_0 - \cdot)$ does not increase on $[C_0, \Upsilon) \cap (t - \epsilon, t + \epsilon)$. Consequently, for any $t \in [C_0, \Upsilon)$,

$$\sum_{j=1}^{J} (Y_j(t, C_0 - t) - Y_j(C_0, 0)) \le (J - 1)(t - C_0),$$

and thus, by (11),

$$\sum_{k=1}^{K} (T_k(t, C_0 - t) - T_k(C_0, 0)) \ge t - C_0.$$

This, in turn, together with (10), implies that for $t \in [C_0, \Upsilon)$,

$$\sum_{k=1}^{K} D_k(t, C_0 - t) \geq \sum_{k=1}^{K} (D_k(t, C_0 - t) - D_k(C_0, 0)) \geq (t - C_0)/m_0. \tag{39}$$

For $t \geq 0$, plugging $s = C_0 - t$ into (25), we get

$$\Theta Z(t, C_0 - t) = \Theta Z(0, C_0) + \Theta\Big(\alpha \circ \int_0^t G(C_0 - t + \eta) d\eta\Big) - D(t, C_0 - t). \tag{40}$$

By (5), for each k and $t \in [C_0, \Upsilon)$,

$$\int_0^t G_k(C_0 - t + \eta) d\eta = \int_0^t G_k(C_0 - \eta) d\eta = \int_0^{C_0} G_k(C_0 - \eta) d\eta \leq C_0. \tag{41}$$

Moreover, by (2), all the entries of Θ are nonnegative, so $\Theta Z(t, C_0 - t) \geq 0$ and the relations (3), (32), (40)–(41) imply

$$D(t, C_0 - t) \leq \Theta Q(0) + \lambda C_0. \tag{42}$$

The relations (39) and (42) imply that $t \leq C_0 + m_0(|\Theta Q(0)| + |\lambda| C_0)$ for any $t \in [C_0, \Upsilon)$, so (36) follows.

We will now justify (37). By (38) and continuity of the fluid model, we have $Z(\Upsilon, C_0 - \Upsilon) = 0$. Let $t > \Upsilon$. Using (40) twice and observing that the equalities in (41) hold for any $t \geq C_0$, we get

$$\begin{aligned}\Theta Z(t, C_0 - t) = \Theta Z(t, C_0 - t) - \Theta Z(\Upsilon, C_0 - \Upsilon) \;&=\; \Theta\Big(\alpha \circ \int_0^t G(C_0 - t + \eta) d\eta\Big) \\ &\quad - \Theta\Big(\alpha \circ \int_0^{\Upsilon} G(C_0 - \Upsilon + \eta) d\eta\Big) - (D(t, C_0 - t) - D(\Upsilon, C_0 - \Upsilon)) \\ &= -(D(t, C_0 - t) - D(\Upsilon, C_0 - \Upsilon)) \;\leq\; 0.\end{aligned}$$

However, $\Theta Z(t, C_0 - t) \geq 0$, so actually $\Theta Z(t, C_0 - t) = 0$, yielding (37) by invertibility of Θ (see (2)). □

Corollary 1. *Let Υ be as in Proposition 1. For $t \geq \Upsilon$ and $s_1 \leq s_2$, we have*

$$\Theta(Z(t, s_2) - Z(t, s_1)) \leq \Theta(\alpha \circ [H(s_1) - H(s_2)]). \tag{43}$$

Proof: If $s_1 \geq C_0 - t$, then (43) follows from (35) and monotonicity of $D(t, \cdot)$. On the other hand, for $s \leq C_0 - t$, we have $Z(t, s) = 0$ by (37) and the function H is nonincreasing, so as long as $t \geq \Upsilon$, validity of (43) for $C_0 - t \leq s_1 \leq s_2$ implies its validity for any $s_1 \leq s_2$. □

Let $t \geq \Upsilon$, $s_1 \leq s_2$. By Corollary 1, (2) and monotonicity of increments of $Z(t, \cdot)$,

$$Z(t, s_2) - Z(t, s_1) \leq \Theta(Z(t, s_2) - Z(t, s_1)) \leq \Theta(\alpha \circ [H(s_1) - H(s_2)]), \tag{44}$$

i.e., for each class k,

$$Z_k(t,s_2) - Z_k(t,s_1) \le \sum_{\mathbf{k}\in\tilde{\mathcal{C}}(k)} \alpha_{b(\mathbf{k})} p_{\mathbf{k}} [H_{b(\mathbf{k})}(s_1) - H_{b(\mathbf{k})}(s_2)]$$
$$= \sum_{\mathbf{k}\in\tilde{\mathcal{C}}(k)} \alpha_{\mathbf{k}} [H_{b(\mathbf{k})}(s_1) - H_{b(\mathbf{k})}(s_2)]. \tag{45}$$

In particular, since $H_k \equiv 0$ on $[y_k^*, \infty)$ by (6), we have

$$Z_k(t, y_k^{**}) = \lim_{s\to\infty} Z_k(t,s), \tag{46}$$

where

$$y_k^{**} = \max_{\mathbf{k}\in\tilde{\mathcal{C}}(k)} y^*_{b(\mathbf{k})}. \tag{47}$$

We will use a notion of a time-shifted EDF fluid model, introduced in [17]. For $\theta \ge 0$, we define the time-shift operator Δ_θ acting on the coordinates of a fluid model $\mathfrak{X}$ of the network under consideration as follows: for $t \ge 0$, $s \in \mathbb{R}$, $\Delta_\theta Z(t,s) = Z(t+\theta, s)$, $\Delta_\theta W(t,s) = W(t+\theta, s)$, $\Delta_\theta A(t,s) = A(t+\theta, s) - A(\theta, t+s)$, $\Delta_\theta D(t,s) = D(t+\theta, s) - D(\theta, t+s)$, $\Delta_\theta T(t,s) = T(t+\theta, s) - T(\theta, t+s)$ and $\Delta_\theta Y(t,s) = Y(t+\theta, s) - Y(\theta, t+s)$. Let

$$\Delta_\theta \mathfrak{X} = (\Delta_\theta Z, \Delta_\theta W, \Delta_\theta A, \Delta_\theta D, \Delta_\theta T, \Delta_\theta Y).$$

It is not hard to see that for any $\theta \ge 0$, $\Delta_\theta \mathfrak{X}$ satisfies the EDF fluid model equations (8)–(13).

In what follows, to simplify the presentation, we will additionally assume that (44) (and consequently (45)–(47)) holds for *every* $t \ge 0$, (rather than $t \ge \Upsilon$) and $s_1 \le s_2$. Formally, this can be achieved by considering the fluid model $\Delta_\Upsilon \mathfrak{X}$ instead of $\mathfrak{X}$. The bound (36) assures that the length of the time interval "neglected" in this way is comparable to (and controlled by) the "size" of the initial condition. The first consequence of this additional assumption is the following refinement of Lemma 1.

Lemma 3. *For $k = 1, \ldots K$, $t \ge 0$ and $s \ge y_k^{**}$, we have*

$$D_k(t,s) = D_k(t, y_k^{**}), \tag{48}$$
$$Z_k(t,s) = Z_k(t, y_k^{**}). \tag{49}$$

Note that (49) follows immediately from (46) and monotonicity of $Z_k(t,\cdot)$. The equation (48) follows from (47) and (49) by an argument similar to the justification of (26).

4 First Frontier Monotonicity and Fluid Model Stability

For $j = 1, \ldots, J$, let

$$\bar{y}_j = \max_{k\in\mathcal{C}(j)} y_k^{**} = \max_{\mathbf{k}\in\bar{\mathcal{C}}(j)} y^*_{b(\mathbf{k})}, \tag{50}$$

and let $\bar{y}_{max} = \max_{j=1,...,J} \bar{y}_j = \max_{k:\alpha_k>0} y_k^*$. For $t \geq 0$, define

$$F_j(t) = \inf\Big\{u \in \mathbb{R} : \sum_{k\in\mathcal{C}(j)} Z_k(t,u) > 0\Big\} \wedge \bar{y}_j, \qquad j = 1,...,J,$$

$$F^{(1)}(t) = \inf\Big\{u \in \mathbb{R} : \sum_{k=1}^{K} Z_k(t,u) > 0\Big\} \wedge \bar{y}_{max},$$

$$\mathcal{J}^{(1)}(t) = \big\{j \in \{1,...,J\} : F_j(t) = F^{(1)}(t)\big\}.$$

The quantity $F_j(t)$ will be called the *frontier* at station j at time t. Accordingly, $F^{(1)}(t)$ will be called the *first frontier* at time t and $\mathcal{J}^{(1)}(t)$ is the set of servers with the first frontier at this time. By (49)–(50), we have $F_j(t) = \bar{y}_j$ if and only if $\sum_{k\in\mathcal{C}(j)} Q_k(t) = 0$, while $F^{(1)}(t) = \bar{y}_{max}$ if and only if $Q(t) = 0$.

For $j = 1,...,J$, the function F_j is upper semi-continuous, i.e., for every $t_0 \geq 0$, we have $\limsup_{t\to t_0} F_j(t) \leq F_j(t_0)$. To see this, choose any sequence $t_n \to t_0$ such that $F_j(t_n) \to a$ for some a. Then

$$0 = \sum_{k\in\mathcal{C}(j)} Z_k(t_n, F_j(t_n)) \to \sum_{k\in\mathcal{C}(j)} Z_k(t_0, a)$$

by continuity of Z, and hence $a \leq F_j(t_0)$. Similarly, the function $F^{(1)}$ is upper semi-continuous.

We will now investigate the dynamics of $F^{(1)}$. By upper semicontinuity, the function $F^{(1)}$ does not have downward jumps. Let $t \geq 0$ be such that $Q(t) \neq 0$, equivalently, $F^{(1)}(t) < \bar{y}_{max}$. Let df, dt be small positive numbers such that

$$dt \leq df \tag{51}$$

(it is convenient to think about them as infinitesimal quantities). By (23)–(24), for any s, we have

$$\begin{aligned} A(t+dt,s) - A(t,s+dt) &= \Theta\left(\alpha \circ \left(\int_0^{t+dt} G(s+\eta)d\eta - \int_0^t G(s+dt+\eta)d\eta\right)\right) \\ &\quad - \Theta P'\left(Z(t+dt,s) - Z(t,s+dt)\right) \\ &= \Theta\left(\alpha \circ \int_0^{dt} G(s+\eta)\, d\eta\right) \\ &\quad - \Theta P'\left(Z(t+dt,s) - Z(t,s+dt)\right) \\ &\leq \Theta\left(\alpha \circ \int_0^{dt} G(s+\eta)\, d\eta\right) + (\Theta - I)Z(t,s+dt). \end{aligned}$$

Consequently, the vector of fluid masses with lead times not greater than s at time $t+dt$ which can be served in the time interval $[t, t+dt]$ is bounded above as follows:

$$Z(t,s+dt) + A(t+dt,s) - A(t,s+dt) \leq \Theta\left(\alpha \circ \int_0^{dt} G(s+\eta)\, d\eta\right) + \Theta Z(t,s+dt). \tag{52}$$

Putting $s = F^{(1)}(t) + df - dt$ and using (52), the equality $Z(t, F^{(1)}(t)) = 0$, (51), (44), (3) and the definition of the functions H_k, we get

$$\begin{aligned} &Z(t, F^{(1)}(t) + df) + A(t + dt, F^{(1)}(t) + df - dt) - A(t, F^{(1)}(t) + df) \\ &\qquad \le \Theta\left(\alpha \circ \int_0^{dt} G(F^{(1)}(t) + df - dt + \eta)\, d\eta\right) + \Theta Z(t, F^{(1)}(t) + df) \\ &\quad = \Theta\left(\alpha \circ \int_{F^{(1)}(t)+df-dt}^{F^{(1)}(t)+df} G(\eta)\, d\eta\right) + \Theta(Z(t, F^{(1)}(t) + df) - Z(t, F^{(1)}(t))) \\ &\quad \le \Theta\left(\alpha \circ \int_{F^{(1)}(t)}^{F^{(1)}(t)+df} G(\eta)\, d\eta\right) + \Theta(\alpha \circ [H(F^{(1)}(t)) - H(F^{(1)}(t) + df)]) \\ &\quad = \Theta(\alpha\, df) = \lambda\, df. \end{aligned} \tag{53}$$

Let $j \in \{1, ..., J\}$. By (4), (10) and (53), the service time necessary for station j to process the fluids with lead times not greater than $F^{(1)}(t) + df - dt$ at time $t + dt$ present at the station by this time is bounded above by

$$\sum_{k \in \mathcal{C}(j)} m_k \lambda_k\, df = \rho_j\, df.$$

On the other hand, because of (12), as long as there are fluids with lead times not greater than some threshold at station j, the server will devote its full capacity to them. This leads to the main result of this paper:

Theorem 1. *The first frontier function $F^{(1)}$ corresponding to a subcritical EDF fluid model satisfying (44) for every $t \ge 0$, $s_1 \le s_2$, is nondecreasing. Moreover, in such a model the condition $Q(t) \ne 0$ implies that*

$$F^{(1)}(t + dt) \ge F^{(1)}(t) + (1/\max_{j=1,...,J} \rho_j - 1)\, dt. \tag{54}$$

Proof: For $t \ge 0$ such that $Q(t) \ne 0$, letting $df = dt/\max_{j=1,...,J} \rho_j$ in the above calculations, we see that for small dt, the time necessary for each station j to process to completion all the fluids with lead times not greater than $F^{(1)}(t) + (1/\max_{j=1,...,J} \rho_j - 1)\, dt$ at time $t + dt$ is not greater than dt, so (54) follows. Note that in the subcritical case, even if $Q(t) = 0$, we can still repeat the above calculations with $df = dt$, getting $F^{(1)}(t + dt) \ge F^{(1)}(t) = \bar{y}_{max}$. This, by the definition of $F^{(1)}$, implies that in this case actually $F^{(1)}(t+dt) = F^{(1)}(t) = \bar{y}_{max}$, i.e., $Q(t + dt) = 0$. □

The following corollary establishes stability of strictly subcritical EDF fluid models (more precisely, a variant of the notion of fluid model stability, originally introduced by Dai [10], which is suitable in our context).

Corollary 2. *In a strictly subcritical EDF fluid model $\mathfrak{X}$ we have*

$$Q(t) = 0, \qquad t \ge \Upsilon + \frac{\bar{y}_{max} + \Upsilon - C_0}{1/\max_{j=1,...,J} \rho_j - 1}, \tag{55}$$

where m_0 and Υ are as in Proposition 1.

PROOF: Let $\mathfrak{X}' = \triangle_{\Upsilon}\mathfrak{X}$ and let $Q'(t) = Q(t+\Upsilon)$, $F'^{(1)}(t) = F^{(1)}(t+\Upsilon)$ be the fluid queue length vector and the first frontier corresponding to $\mathfrak{X}'$, respectively. Then $\mathfrak{X}'$ satisfies (44) for every $t \geq 0$, $s_1 \leq s_2$ (see the discussion before Lemma 3) and (55) is equivalent to

$$Q'(t) = 0, \qquad t \geq \frac{\bar{y}_{max} + \Upsilon - C_0}{1/\max_{j=1,\ldots,J} \rho_j - 1}. \tag{56}$$

By (37), $Z'(0, C_0 - \Upsilon) = Z(\Upsilon, C_0 - \Upsilon) = 0$, so $F'^{(1)}(0) \geq C_0 - \Upsilon$. By Theorem 1, $F'^{(1)}(t+dt) \geq F'^{(1)}(t) + Cdt$ as long as $Q(t) \neq 0$, where $C = 1/max_{j=1,\ldots,J}\rho_j - 1 > 0$. In particular,

$$\tau = \inf\{t \geq 0 : Q'(t) = 0\} = \inf\{t \geq 0 : F'^{(1)}(t) = \bar{y}_{max}\} \leq (\bar{y}_{max} + \Upsilon - C_0)/C$$

and $F'^{(1)} \equiv \bar{y}_{max}$ on $[\tau, \infty)$ by monotonicity of $F'^{(1)}$, so (56) follows. □

As a special case of Corollary 2, we get stability of strictly subcritical FISFO fluid models:

Corollary 3. *In a strictly subcritical FISFO fluid model $\mathfrak{X}$ with $C_0 = 0$ in (20), we have $Q(t) = 0$ for $t \geq c|Q(0)|$, where*

$$c = \frac{m_0 ||\Theta||}{1 - \max_{j=1,\ldots,J} \rho_j},$$

m_0 is as in Proposition 1 and $||\Theta|| = \sum_{i=1}^{K} \max_{j=1,\ldots,K} \theta_{ij}$ (see (1)).

This result was first proved, under some additional technical assumptions, as Theorem 2 in Bramson [8], with a somewhat larger constant, namely

$$c = \frac{11(\sum_{j=1}^{J} \rho_j + 2)|e_k' M \Theta|}{1 - \max_{j=1,\ldots,J} \rho_j} = \frac{11(\sum_{j=1}^{J} \rho_j + 2) \sum_{i=1}^{K} \sum_{j=1}^{K} m_i \theta_{ij}}{1 - \max_{j=1,\ldots,J} \rho_j},$$

where $e_K = (1, ..., 1) \in \mathbb{R}^K$ and M is the $K \times K$ diagonal matrix with $m_1, ..., m_K$ on the main diagonal. Kruk [18, Theorem 3.1], extended the scope of Bramson's proof to arbitrary strictly subcritical FISFO fluid models, at a price of an additional technical argument. Our present approach, based on Theorem 1, seems to be simpler and somewhat stronger than the FISFO fluid model stability arguments from [8,18]. In particular, we can also establish some kind of EDF fluid model stability in the general subcritical case.

Definition 1 (Compare [11], Definition 6). *A fluid model is said to be weakly stable if for every fluid model solution with $Q(0) = 0$, we have $Q(t) = 0$, $t \geq 0$.*

Note that by (2) and (8)–(9), in a weakly stable EDF fluid model with $Q(0) = 0$, we have

$$A(t,s) = D(t,s) = \Theta\left(\alpha \circ \int_0^t G(s+\eta)\, d\eta\right), \qquad t \geq 0, s \in \mathbb{R}.$$

This notion is related to *rate stability of the corresponding stochastic network*, introduced by Dai and Prabhakar [11], which in our context means

$$\lim_{r\to\infty}\frac{1}{r}D(rt,rs)=\lim_{r\to\infty}\frac{1}{r}A(rt,rs)=\Theta\left(\alpha\circ\int_0^t G(s+\eta)\,d\eta\right)$$

for any $t\geq 0$, $s\in\mathbb{R}$, where A and D are the arrival and departure processes in the *stochastic network* approximated by the EDF fluid model solutions.

Corollary 4. *Any subcritical EDF fluid model is weakly stable.*

Proof: Let $\mathfrak{X}$ be a subcritical EDF fluid model such that $Q(0)=0$. Then (20) holds with $C_0=0$. With this choice, we have $\Upsilon=0$, where Υ is defined by (38). Moreover, (37) holds by the proof of Proposition 1. However, for $C_0=0$, the condition (21) may fail and consequently Lemma 2 may no longer hold. Nevertheless, for $t\geq 0$ and $s_1\leq s_2$, by (25), the equality $Q(0)=0$ and monotonicity of $D(t,\cdot)$, we get

$$\begin{aligned}\Theta\left(Z(t,s_2)-Z(t,s_1)\right)&=\Theta\left(Z(0,t+s_2)-Z(0,t+s_1)\right)\\&\quad+\Theta\left(\alpha\circ\left(\int_{t+s_1}^{t+s_2}G(\eta)\,d\eta-\int_{s_1}^{s_2}G(\eta)\,d\eta\right)\right)\\&\quad-\left(D(t,s_2)-D(t,s_1)\right)\\&\leq\Theta\left(\alpha\circ\left(\int_{t+s_1}^{t+s_2}d\eta-\int_{s_1}^{s_2}G(\eta)\,d\eta\right)\right)\\&\quad-\left(D(t,s_2)-D(t,s_1)\right)\\&\leq\Theta(\alpha\circ[H(s_1)-H(s_2)]),\end{aligned}$$

so Corollary 1 and its consequences (44)–(47) hold. Therefore, our justification of Theorem 1 is valid in the current setting. Hence, for any $t\geq 0$, $F^{(1)}(t)=F^{(1)}(0)=\bar{y}_{max}$ which, together with (47), implies that $Q(t)=0$. □

5 Conclusion

We studied fluid models of subcritical EDF multiclass queueing networks with soft job deadlines. We established monotonicity of the first frontier function $F^{(1)}$ corresponding to such a model. Stability of strictly subcritical EDF fluid models and weak stability of their subcritical counterparts follow easily from this finding.

Our main result implies the existence of a finite limit

$$F^{(1)}(\infty):=\lim_{t\to\infty}F^{(1)}(t)\tag{57}$$

in each subcritical EDF fluid model. In fact, Corollary 2 assures that in the strictly subcritical case, we have $F^{(1)}(\infty)=\bar{y}_{max}$ and this limit is actually attained in finite time. It would be interesting to characterize the rate of convergence in (57) for other subcritical EDF fluid models, in particular, to find

sufficient conditions for this limit to be attained in finite time. More generally, we would like to find sufficient conditions for convergence of the frontier $F_j(t)$ at each station j as $t \to \infty$ for a subcritical EDF fluid model. This would imply convergence of the fluid model states to the invariant manifold (so-called asymptotic stability of the fluid model). The latter result may be very helpful in the development of diffusion limits for the corresponding stochastic networks.

References

1. Atar, R., Biswas, A., Kaspi, H.: Fluid limits of G/G/1+G queues under the non-preemptive earliest-deadline-first discipline. Math. Oper. Res. **40**, 683–702 (2015)
2. Atar, R., Biswas, A., Kaspi, H.: Law of large numbers for the many-server earliest-deadline-first queue. Stochast. Process. Appl. **128**(7), 2270–2296 (2018)
3. Atar, R., Biswas, A., Kaspi, H., Ramanan, K.: A Skorokhod map on measure-valued paths with applications to priority queues. Ann. Appl. Probab. **28**, 418–481 (2018)
4. Atar, R., Shadmi, Y.: Fluid limits for earliest-deadline-first networks. arXiv arXiv:2009.07169v1 (2020)
5. Bramson, M.: Convergence to equilibria for fluid models of FIFO queueing networks. Queueing Syst. Theor. Appl. **22**, 5–45 (1996)
6. Bramson, M.: Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. Queueing Syst. Theor. Appl. **23**, 1–26 (1996)
7. Bramson, M.: State space collapse with application to heavy traffic limits for multiclass queueing networks. Queueing Syst. Theor. Appl. **30**, 89–148 (1998)
8. Bramson, M.: Stability of earliest-due-date, first-served queueing networks. Queueing Systems. Theor. Appl. **39**, 79–102 (2001)
9. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks. Springer Science+Business Media, LLC, New York (2001). https://doi.org/10.1007/978-1-4757-5301-1
10. Dai, J.G.: On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. Ann. Appl. Probab. **5**, 49–77 (1995)
11. Dai, J.G., Prabhakar, B.: The throughput of data switches with and without speedup. In: Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. 19th Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 2, pp. 556–564 (2000). https://doi.org/10.1109/INFCOM.2000.832229
12. Decreusefond, L., Moyal, P.: Fluid limit of a heavily loaded EDF queue with impatient customers. Markov Process. Relat. Fields **14**(1), 131–158 (2008)
13. Doytchinov, B., Lehoczky, J.P., Shreve, S.E.: Real-time queues in heavy traffic with earliest-deadline-first queue discipline. Ann. Appl. Probab. **11**, 332–379 (2001)
14. Harrison, J.M.: Brownian models of queueing networks with heterogeneous customer populations. In: Fleming, W., Lions, P.L. (eds.) Stochastic Differential Systems, Stochastic Control Theory and Applications. The IMA Volumes in Mathematics and Its Applications, vol. 10, pp. 147–186. Springer, New York (1988). https://doi.org/10.1007/978-1-4613-8762-6_11
15. Kruk, Ł: Stability of two families of real-time queueing networks. Probab. Math. Stat. **28**, 179–202 (2008)
16. Kruk, Ł: Invariant states for fluid models of EDF networks: nonlinear lifting map. Probab. Math. Stat. **30**, 289–315 (2010)

17. Kruk, Ł: An open queueing network with asymptotically stable fluid model and unconventional heavy traffic behavior. Math. Oper. Res. **36**, 538–551 (2011)
18. Kruk, Ł: Stability of preemptive EDF queueing networks. Ann. Univ. Mariae Curie-Skłodowska Math. A. **73**, 105–134 (2019)
19. Kruk, Ł.: Minimal and locally edge minimal fluid models for resource sharing networks. Math. Oper. Res. (2021). https://doi.org/10.1287/moor.2020.1110
20. Kruk, Ł, Lehoczky, J.P., Ramanan, K., Shreve, S.E.: Heavy traffic analysis for EDF queues with reneging. Ann. Appl. Probab. **21**, 484–545 (2011)
21. Kruk, Ł, Lehoczky, J.P., Shreve, S.E., Yeung, S.-N.: Earliest-deadline-first service in heavy traffic acyclic networks. Ann. Appl. Probab. **14**, 1306–1352 (2004)
22. Lehoczky, J.P.: Using real-time queueing theory to control lateness in real-time systems. Perform. Eval. Rev. **25**, 158–168 (1997)
23. Lehoczky, J.P.: Real-time queueing theory. In: Proceedings of the IEEE Real-Time Systems Symposium, pp. 186–195 (1998)
24. Lehoczky, J.P.: Scheduling communication networks carrying real-time traffic. In: Proceedings of the IEEE Real-Time Systems Symposium, pp. 470–479 (1998)
25. Liu, C.L., Layland, J.W.: Scheduling algorithms for multiprogramming in a hard real-time environment. J. Assoc. Comput. Mach. **20**(1), 40–61 (1973)
26. Moyal, P.: Convex comparison of service disciplines in real time queues. Oper. Res. Lett. **36**(4), 496–499 (2008)
27. Panwar, S.S., Towsley, D.: On the optimality of the STE rule for multiple server queues that serve customers with deadlines. Technical Report 88-81, Department of Computer and Information Science, University Massachusetts, Amherst (1988)
28. Panwar, S.S., Towsley, D.: Optimality of the stochastic earliest deadline policy for the G/M/c queue serving customers with deadlines. In: 2nd ORSA Telecommunications Conference. ORSA (Operations Research Society of America), Baltimore, MD (1992)
29. Rybko, A.N., Stolyar, A.L.: Ergodicity of stochastic processes describing the operations of open queueing networks. Probl. Inf. Transm. **28**, 199–220 (1992)
30. Seidman, T.I.: "First come, first served" can be unstable! IEEE Trans. Automat. Control **39**, 2166–2171 (1994)
31. Williams, R.J.: Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. Queueing Syst. Theor. Appl. **30**, 27–88 (1998)
32. Yeung, S.-N., Lehoczky, J.P.: Real-time queueing networks in heavy traffic with EDF and FIFO queue discipline. Working paper, Department of Statistics, Carnegie Mellon University (2001)