PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 28, Fasc. 2 (2008), pp. 179–202

# STABILITY OF TWO FAMILIES OF REAL-TIME QUEUEING NETWORKS

BY

ŁUKASZ KRUK (WARSZAWA)

*Abstract.* We study open multiclass queueing networks with renewal arrival streams and general service time distributions. Upon arrival to the network, customers from each class are assigned a random deadline drawn from a distribution associated with this class. We show that preemptive subcritical EDF networks with fixed customer routes are stable. We also prove that a broad class of (not necessarily subcritical) networks with reneging and Markovian routing, including EDF, FIFO, LIFO, SRPT, fixed priorities and processor sharing, is stable.

**2000 AMS Mathematics Subject Classification:** Primary: 60K25, 90B15; Secondary: 68M20.

**Key words and phrases:** Multiclass queueing networks, deadlines, reneging, stability, fluid models, fluid limits.

## 1. INTRODUCTION

A principal question in the theory of multiclass queueing networks is whether a given network is stable, i.e., the corresponding Markov process is positive Harris recurrent. The intuitive meaning of network stability is that the system performs well under reasonable workload: the queue lengths do not grow linearly with time and do not oscillate “wildly”, there is no mutual blocking and forced idleness of the servers when work is present in the system. Thus, stability of a network is a basic indicator of its proper design. It appears that there is no general criterion for this behavior; in particular, it is known that the usual necessary traffic condition that $\rho_j < 1$ at each station is not sufficient; see, e.g., [3], [4], [19]. On the positive side, the condition $\rho_j < 1$ for all $j$ is sufficient for generalized Jackson networks [16] and multiclass networks with some disciplines, including first-in-first-out (FIFO) in networks of Kelly type [5], head-of-the-line proportional processor sharing [6], first-buffer-first-served and last-buffer-first-served [8], [9].

Dai [8], generalizing and systematizing the earlier work of Rybko and Stolyar [19], provided a general framework for proving such stability results. Its main idea is to reduce the problem to showing stability of the corresponding fluid model, a deterministic analog of the network under consideration. This approach has been applied to various queueing systems. The result most relevant to this paper is stability of multiclass earliest-deadline-first (EDF) networks without preemption. The EDF discipline, also called earliest-due-date-first-served (EDDFS), is the rule where each customer has a deadline, assigned upon arrival at the network and maintained until departure, and a customer with the earliest deadline is selected for service at each station of the network. Bramson [7] showed that the fluid limits of the performance processes for a non-preemptive EDF network with $\rho_j < 1$ for all $j$ satisfy the first-in-system-first-out (FISFO) fluid model equations. He then proved that a sufficiently rich class of FISFO fluid models is stable. This, by a variation of Theorem 4.2 of Dai [8], implies stability of the network under consideration.

It is natural to ask whether this stability result remains valid for EDF networks with preemption. As observed in Bramson [7], this problem is more difficult and the analysis for the non-preemptive case does not generalize immediately to the preemptive setting. The main reason for this is that the number of partially served customers in a preemptive EDF system is unbounded, so it is not clear that the number of departed customers from a given class is asymptotically proportional to the service time devoted by the server to this class. This difficulty increases if we consider initial conditions with unbounded numbers of residual service times, which are natural for preemptive service protocols.

In this paper we show how to overcome this difficulty under the assumption that the customer routes in the network are fixed. We consider this to be a mild assumption since it is satisfied by many EDF networks of practical interest, arising, e.g., in manufacturing (see [13], [21]). Moreover, known examples of unstable systems with fixed customer routes (see, e.g., [3], [4], [19]) indicate that stability theory for such (in fact, even acyclic) systems is already interesting. The main idea of the argument is based on the observation that because the initial lead time distributions disappear in the limit, the asymptotic behavior of a preemptive EDF system does not differ from the behavior of the corresponding FISFO system. More precisely, after a time large enough to process all the initial customers to completion at every station, the fluid limits for a preemptive EDF system satisfy the FISFO fluid model equations introduced in Bramson [7]. Once the convergence to an FISFO fluid model is established, stability of the latter models proved in Bramson [7] and an argument similar to the proof of Theorem 4.2 in Dai [8] imply stability of preemptive EDF systems.

In spite of the theoretical and practical importance of stochastic EDF queueing networks, there are still few mathematically rigorous results for such systems. Apart from Bramson's work [7] recalled above, Doytchinov et al. [11] provided a diffusion approximation for measure-valued state descriptors of preemptive EDF GI/G/1 queues. Their result has been generalized by Yeung and Lehoczky [23] to preemptive EDF feedforward networks. A further generalization to the case of acyclic networks, with or without preemption, was given by Kruk et al. [14]. However, the latter result rests on a strong assumption implying the existence of a heavy traffic limit for the corresponding real-valued workload process. Currently, we are able to verify this assumption only in a number of special cases. Our stability result is a step forward in a process of filling this gap.

It turns out that some of the techniques developed for the EDF stability proofs can be used to show stability of general queueing networks with reneging. In fact, the analysis of the latter networks is much easier, since in this case a more direct approach is possible. In particular, there is no need to use fluid models, and consequently to show convergence to such models and their stability. Existing stability results for models with impatient customers treat the case of a single server queue (see [1], [15], [20], [22] and the references given therein) and feedforward networks [22]. Our analysis allows for Markovian routing of customers and for a broad range of service protocols, including EDF (preemptive or not), FIFO, LIFO (last-in-first-out), SRPT (shortest-remaining-processing-time-first), fixed priorities and processor sharing. Moreover, we do not require any condition on the traffic intensities $\rho_j$. Thus, as it should be expected, customer impatience is a universal stabilizing mechanism.

To our knowledge, the theorems presented in this paper are the first stability results for queueing systems with unbounded numbers of partially served customers. In particular, our stability theorem for preemptive EDF networks is the first application of the methodology of Dai [8] to such systems.

The paper is organized as follows. Section 2 describes the models, provides background information on positive Harris recurrence of Markov processes and adjusts it to our setting. In Section 3, we provide the formulation of the main results. Section 4 contains an auxiliary lead time estimate. In Section 5, we present the preemptive EDF queueing network equations and show that the fluid limits of the corresponding (properly shifted) performance processes satisfy the FISFO fluid model equations. In Section 6, we provide the proof of the stability theorem for preemptive EDF networks. The proof of the stability theorem for networks with reneging is contained in Section 7. Proofs of two technical auxiliary results are relegated to the Appendix.

## 2. TERMINOLOGY AND BACKGROUND

**2.1. Notation.** The following notation will be used throughout the paper. Let $\mathbb{R}$ denote the set of real numbers and let $\mathbb{R}_+ = [0, \infty)$. 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 $a \vee 0$. We also write $\lfloor a \rfloor$ for the largest integer less than or equal to $a$. For a vector $a = (a_1, \dots, a_n) \in \mathbb{R}^n$, let $|a| \triangleq \sum_{i=1}^n |a_i|$. All vectors in the paper are to be interpreted as column vectors unless indicated otherwise. For a finite set $B$, $|B|$ denotes the cardinality of $B$. The Borel $\sigma$-field on a topological space $Y$ will be denoted by $\mathcal{B}(Y)$. Finally, the space of right-continuous real-valued functions with left limits on $[0, \infty)$ will be denoted by $D[0, \infty)$.

### 2.2. The models.

**2.2.1.** *Open queueing networks with Markovian routing.* We consider general open queueing networks with Markovian routing. Such a network consists of $J$ single server stations, indexed by $j = 1, \dots, J$. Customers are members of *classes*, or *buffers*, $\mathbf{k} \in \mathbf{K}$, where $\mathbf{K}$ is a finite set of indices, with a customer of buffer $\mathbf{k}$ being served at a unique server $j$, written $\mathbf{k} \in \bar{\mathcal{C}}(j)$. The set of buffers with external arrival processes will be denoted by $\mathcal{E}$. We assume that $K \triangleq |\mathcal{E}| > 0$ and

$$\mathcal{E} = \{1, \dots, K\}. \tag{2.1}$$

Upon being served at $j$, a customer of buffer $\mathbf{k}$ immediately becomes a customer of buffer $\mathbf{k}'$ with probability $p_{\mathbf{k},\mathbf{k}'}$, independently of its past history. The *routing matrix* $P = (p_{\mathbf{k},\mathbf{k}'})$ is assumed to be transient, i.e., such that the matrix

$$\Theta \triangleq (I - P')^{-1} = I + P' + (P')^2 + \dots$$

exists, where $'$ denotes the transpose.

**2.2.2.** *Open networks with fixed customer routes.* An important special case of the structure described above arises when $p_{\mathbf{k},\mathbf{k}'} \in \{0, 1\}$ for each $\mathbf{k}, \mathbf{k}' \in \mathbf{K}$, i.e., the network routing is deterministic. We now introduce additional notation and terminology for this case.

The network is populated by $K$ *customer types*, indexed by $k = 1, \dots, K$. Customers of type $k$ arrive to the network and move through it according to a fixed route, until they eventually exit the system. Different customer types may visit stations in different orders and some servers may be visited by type $k$ customers more than once, i.e., the system is not necessarily feedforward or acyclic. We define the *path of type $k$ customers* as the sequence of servers they encounter along their way through the network and denote it by $\mathcal{P}(k) \triangleq \big(j_{k,1}, j_{k,2}, \dots, j_{k,m(k)}\big)$. In particular, type $k$ customers enter the system at station $j_{k,1}$ and leave it through station $j_{k,m(k)}$. If $j$ is a member of the list of station indices in $\mathcal{P}(k)$, we shall write $j \in \mathcal{P}(k)$. For $k = 1, \dots, K$, $j = 1, \dots, J$ and $i = 1, \dots, m(k)$, let

$$b(k, j, i) \triangleq \#[i' \in \{1, \dots, i\} : j_{k,i'} = j].$$

In other words, $b(k, j, i)$ is the number of occurrences of station $j$ among the first $i$ steps along the route of type $k$ customers. In particular, $\bar{b}(k, j) \triangleq b\big(k, j, m(k)\big)$ is the number of times the station index $j$ appears in $\mathcal{P}(k)$. For $j = 1, \dots, J$, we define

$$\mathcal{C}(j) \triangleq \{\text{indices of customer types which visit station } j\}\,.$$

We introduce *multi-indices* of the form $\mathbf{k} = (k, j, b)$, indexing the $b$-th visit of type $k$ customers at station $j$, where $k = 1, \dots, K$, $j \in \mathcal{P}(k)$, $b = 1, \dots, \bar{b}(k, j)$. The set of all such multi-indices (which will be identified with the customer classes or buffers introduced in Section 2.2.1) will be denoted by $\mathbf{K}$. To make the indexing of classes with external arrival streams uniform throughout the paper, we denote a multi-index $(k, j_{k,1}, 1) \in \mathcal{E}$, $k = 1, \ldots, K$, simply by $k$, so that (2.1) holds. For $j = 1, \ldots, J$, let

$$\bar{\mathcal{C}}(j) \triangleq \{(k, j, b) \in \mathbf{K} : k \in \mathcal{C}(j), b = 1, \ldots, \bar{b}(k, j)\}.$$

The routing matrix $P = (p_{\mathbf{k},\mathbf{k}'})$ corresponding to this network topology is given by $p_{\mathbf{k},\mathbf{k}'} \triangleq 1$ if $\mathbf{k} = \big(k, j_{k,i}, b(k, j_{k,i}, i)\big)$, $\mathbf{k}' = \big(k, j_{k,i+1}, b(k, j_{k,i+1}, i+1)\big)$ for some $k \in \{1, \ldots, K\}$, $i \in \{1, \ldots, m(k) - 1\}$, and $p_{\mathbf{k},\mathbf{k}'} \triangleq 0$ otherwise. Since the network is open and the customer routes are fixed, it is clear that the matrix $P$ is transient.

**2.2.3.** *Stochastic primitives.* We will now define the stochastic primitives for the models described in Sections 2.2.1 and 2.2.2. The *customer interarrival times* are a sequence of strictly positive, i.i.d. random variables $u_k(i)$, $i = 1, 2, \ldots$, where the subscript $k \in \mathcal{E}$ indicates the customer class. We assume that for $k \in \mathcal{E}$

$$\mathbb{E}\, u_k(1) < \infty, \tag{2.2}$$

$$\mathbb{P}\big(u_k(1) \geqslant x\big) > 0 \quad \text{for all } x > 0, \tag{2.3}$$

and for some $n_k > 0$ and some nonnegative Borel function $f_k$ with

$$\int\limits_0^\infty f_k(x)dx > 0,$$

we have

$$\mathbb{P}\big(u_k(1) + \ldots + u_k(n_k) \in dx\big) \geqslant f_k(x)dx. \tag{2.4}$$

In other words, the interarrival times are integrable, unbounded and spread out. The residual interarrival times $u_k(0)$, $k = 1, \ldots, K$, are assigned fixed nonnegative values. The *arrival time* of the $n$-th customer of class $k$ to the system is given by $U_k(n) = \sum_{i=0}^{n-1} u_k(i)$, $n = 1, 2, \ldots$ The *service times* of buffer $\mathbf{k}$ customers are a sequence of strictly positive, i.i.d. random variables $v_{\mathbf{k}}(i)$, $i = 1, 2, \ldots$, where the index $i$ denotes the order of arrival of customers to the buffer. We assume that for all $\mathbf{k} \in \mathbf{K}$

$$m_{\mathbf{k}} \triangleq \mathbb{E}\, v_{\mathbf{k}}(1) < \infty. \tag{2.5}$$

The *arrival rates* $\alpha_{\mathbf{k}}$, $\mathbf{k} \in \mathbf{K}$, are defined by

$$\alpha_{\mathbf{k}} \triangleq \begin{cases} 1/\mathbb{E}u_{\mathbf{k}}(1) & \text{if } \mathbf{k} \in \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

We put $\alpha = (\alpha_{\mathbf{k}})_{\mathbf{k} \in \mathbf{K}}$. We define the *total arrival rate* vector $\lambda = \Theta\alpha$. Next, we define the *traffic intensity* at station $j$ as

$$\rho_j = \sum_{\mathbf{k} \in \bar{\mathcal{C}}(j)} m_{\mathbf{k}} \lambda_{\mathbf{k}}. \tag{2.7}$$

When $\rho_j < 1$ for each $j$, the network is called *strictly subcritical.*

Customers entering the network through the buffer $k \in \mathcal{E}$ at times $U_k(i)$ have *initial lead times* $\ell_k(i)$, $i = 1, 2, \ldots$, which are mutually independent nonnegative i.i.d. random variables. The *deadline* of such a customer is given by $\Delta_k(i) = U_k(i) + \ell_k(i)$. We assume that for $k \in \mathcal{E}$

$$\mathbb{E}\,\ell_k(1) < \infty. \tag{2.8}$$

We assume also that the sequences $\{u_k(i)\}_{i=1}^{\infty}$, $k \in \mathcal{E}$, and $\{v_{\mathbf{k}}(i)\}_{i=1}^{\infty}$, $\mathbf{k} \in \mathbf{K}$, are mutually independent. Moreover, we assume that the sequences $\{\ell_k(i)\}_{i=1}^{\infty}$, $k \in \mathcal{E}$, and $\{v_{\mathbf{k}}(i)\}_{i=1}^{\infty}$, $\mathbf{k} \in \mathbf{K}$, are mutually independent.

For each $\mathbf{k} \in \mathbf{K}$, the *initial condition* specifies $Q_{\mathbf{k}}(0)$, the number of *initial customers* present at the buffer $\mathbf{k}$ at time $0$, as well as their residual service times and initial lead times, which are denoted by $\tilde{v}_{\mathbf{k}}(i)$ and $\tilde{\ell}_{\mathbf{k}}(i)$, $i = 1, \ldots, Q_{\mathbf{k}}(0)$, respectively. We assume that $Q_{\mathbf{k}}(0)$ are fixed nonnegative integers, $\tilde{v}_{\mathbf{k}}(i)$ are fixed positive numbers, and $\tilde{\ell}_{\mathbf{k}}(i)$ are fixed real numbers. The deadlines of the initial customers are given by $\tilde{\Delta}_{\mathbf{k}}(i) = \tilde{\ell}_{\mathbf{k}}(i)$.

**2.2.4.** *Lead times, service disciplines.* 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.

In this paper, two types of queueing networks will be considered. The first one is a system with general Markovian routing (see Section 2.2.1) and stochastic primitives defined in Section 2.2.3, in which the customers are *impatient*: they *renege* (i.e., leave the network) immediately after their deadlines elapse. Assumption 2.1 on the service protocol considered in this case will be given in Section 2.3. Now we only mention that this assumption is very mild, allowing for a broad class of service disciplines. Here we assume that $\tilde{\ell}_{\mathbf{k}}(i) \geqslant 0$, since customers with negative lead times are never present at the network.

The other network type analyzed in this paper can be characterized by fixed customer routes (see Section 2.2.2), stochastic primitives defined in Section 2.2.3 and the EDF service discipline. That is, the customer with the shortest remaining lead time, regardless of class, is selected for service at each station. Preemption occurs when a customer more urgent than the customer in service arrives (we assume preempt-resume). There is no set up, switch-over or other type of overhead. Here we assume that the customers are *patient*: they stay in the system until served to completion, even if they get *late*, i.e., their lead times become negative. In this case, the (natural) assumption that $\ell_k(i) \geqslant 0$ was added only to simplify the exposition of the proofs. All our results concerning EDF networks without reneging are valid without this condition as long as $\ell_k(i)$ are integrable.

**2.3. Markov process background.** In the real-time queueing systems under consideration, the individual customer lead times or some equivalent information must be kept to determine customer priorities in the case of the EDF service discipline and to identify late customers leaving a reneging system. Similarly, to model protocols with preemption or simultaneous service of multiple customers, it is necessary to store the residual service time of every task, i.e., the current remaining amount of processing time required to fulfill its service time requirement. Since the number of customers present in the system at a given time is unbounded, it is necessary to model its evolution in an infinitely dimensional state space. In what follows, we use lists of infinite length to construct the state descriptor. An alternative approach utilizing finite Borel measures can be found, e.g., in [11], [14] and [23].

Let $d = |\mathbf{K}|$ and let $S = (\mathbb{R}_+ \times \mathbb{R})^\infty$. Let $\Omega = S^d \times \mathbb{R}_+^K$ be the *state space*. Under the product topology, $S^d$ and $\Omega$ are Polish spaces. The *state* of the process at any time is given by a point

$$x = (h_\mathbf{k}, \mathbf{k} \in \mathbf{K},\ r_k, k \in \mathcal{E}) \in \Omega, \tag{2.9}$$

where for $\mathbf{k} \in \mathbf{K}$, $h_\mathbf{k}$ describes all customers present at buffer $\mathbf{k}$ at this time so that each of them is listed in terms of his residual service time and lead time, and $r_k$ is the residual interarrival time for class $k \in \mathcal{E}$. We assume that the customers in $h_\mathbf{k}$ are listed in the order of their arrival to the buffer, ties are broken in an arbitrary manner and the empty spaces on the list $h_\mathbf{k}$ (i.e., not corresponding to any customer present in the buffer) are positioned after all the listed customers and they are filled with zeros. Let **0** denote the element of $S^d$ describing the empty system, i.e., with all coordinates equal to the sequence $\big((0,0),(0,0),\dots\big)$. Let $q = (q_\mathbf{k})_{\mathbf{k}\in\mathbf{K}}$ and $w = (w_\mathbf{k})_{\mathbf{k}\in\mathbf{K}}$, where $q_\mathbf{k}$ is the number of customers listed in $h_\mathbf{k}$ and $w_\mathbf{k}$ is the sum of their residual service times. Let $r = (r_k)_{k\in\mathcal{E}}$ and let $\ell$ be the greatest lead time. For $x \in \Omega$, let $|x| = |q| + |w| + |r| + \ell^+$ be the "norm" of $x$.

Fix one of the queueing systems described in Section 2.2. The process describing the evolution of this system is denoted by $X = \big(X(t), t \geqslant 0\big)$, where

$$X(t) = \big(H(t), R(t)\big) = \big(H_\mathbf{k}(t), \mathbf{k} \in \mathbf{K},\ R_k(t), k \in \mathcal{E}\big) \tag{2.10}$$

is the state of the system at time $t$. By definition, the process $X$ has right-continuous sample paths. In the case of a preemptive EDF system without reneging, it is easy to see that $X$ is a Markov process. In the case of networks with reneging and Markovian routing, we make the following assumption:

ASSUMPTION 2.1. The process $X$ defined by (2.10) has the Markov property.

Assumption 2.1 holds for a broad class of service disciplines with reneging, e.g., EDF (preemptive or not), FIFO, LIFO, SRPT, fixed priorities and processor sharing.

The evolution of the process $X$ between arrivals and departures is deterministic. Thus, $X$ is a piecewise-deterministic Markov (PDM) process, so it is actually strong Markov (see [10]).

A Markov process $X$ on the state space $\Omega$ is *Harris recurrent* if there exists a $\sigma$-finite measure $\nu$ on $\mathcal{B}(\Omega)$ such that whenever $A \in \mathcal{B}(\Omega)$, $\nu(A) > 0$, we have $\mathbb{P}_x(\tau_A < \infty) = 1$ for all $x \in \Omega$, where $\tau_A = \inf\{t \geqslant 0 : X(t) \in A\}$. It is known that Harris recurrence implies the existence of a unique (up to a multiplicative constant) invariant measure; see, e.g., [12]. If this measure is finite, $X$ is called *positive Harris recurrent.*

Let $P^t(x, A)$, $x \in \Omega$, $A \in \mathcal{B}(\Omega)$, $t \geqslant 0$, be the transition probability of $X$, i.e., $P^t(x, A) = \mathbb{P}_x\big(X(t) \in A\big)$. A nonempty set $A \in \mathcal{B}(\Omega)$ is called *petite* if for some nontrivial measure $\nu$ on $\mathcal{B}(\Omega)$ and some probability distribution $p$ on $(0, \infty)$

$$\nu(B) \leqslant \int_0^\infty P^t(x, B)\, p(dt) \qquad \text{for all } x \in A,\ B \in \mathcal{B}(\Omega).$$

PROPOSITION 2.1 ([17], Theorem 4.1). *Let $A$ be a closed petite set, suppose that $P_x(\tau_A < \infty) = 1$ for each $x \in \Omega$ and that for some $\delta > 0$*

$$\sup_{x \in A} \mathbb{E}_x[\tau_A(\delta)] < \infty, \tag{2.11}$$

*where $\tau_A(\delta) = \inf\{t \geqslant \delta : X(t) \in A\}$. Then $X$ is positive Harris recurrent.*

LEMMA 2.1. *Under the assumptions* (2.3) *and* (2.4), *for any $\zeta > 0$ it follows that $A = \{x \in \Omega : |x| \leqslant \zeta\}$ is a closed petite set.*

This lemma is analogous to Lemma 3.2 in [8] and it can be proved in a similar way.

The following proposition, which is very useful in stability theory for queueing networks, reduces the problem of proving the positive Harris recurrence of a Markov process to checking the condition (2.12) on the asymptotic behavior of this process as the initial condition gets large. The latter condition can be verified either directly or with the use of suitable fluid models.

PROPOSITION 2.2. *If there exists $\delta > 0$ such that*

$$\lim_{|x| \to \infty} \frac{1}{|x|} \mathbb{E}_x\big|X(\delta|x|)\big| = 0, \tag{2.12}$$

*then* (2.11) *holds for $A = \{x \in \Omega : |x| \leqslant \zeta\}$ with some $\zeta > 0$. Consequently, $X$ is positive Harris recurrent.*

The proof of this proposition is the same as the proof of Theorem 3.1 in [8] (see also the proof of Theorem 2.1 (ii) in [18]).

## 3. MAIN RESULTS

Recall that a queueing network is *stable* when the underlying Markov process is positive Harris recurrent. The following theorems give the main results of this paper.

THEOREM 3.1. *All strictly subcritical EDF queueing networks with preemption and fixed customer routes which satisfy* (2.2)–(2.5) *and* (2.8) *are stable.*

THEOREM 3.2. *All queueing networks with reneging satisfying* (2.2)–(2.5), (2.8) *and Assumption* 2.1 *are stable.*

Let us stress that in Theorem 3.2 we do not require that the network be strictly subcritical.

## 4. BASIC LEAD TIME ESTIMATE

Let $k \in \mathcal{E}$, $t \geqslant 0$, and let $x \in \Omega$ be the initial state of the network. Let $N_k^x(t) = \max\{n \geqslant 0 : U_k(n) \leqslant t\}$. Let $G$ be the set of elementary events $\omega$ for which

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} u_k(i)(\omega) = \mathbb{E}u_k(1), \qquad k \in \mathcal{E}, \tag{4.1}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} v_{\mathbf{k}}(i)(\omega) = m_{\mathbf{k}}, \qquad \mathbf{k} \in \mathbf{K}, \tag{4.2}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \ell_k(i)(\omega) = \mathbb{E}\,\ell_k(1), \qquad k \in \mathcal{E}. \tag{4.3}$$

By (2.2), (2.5), (2.8) and the strong law of large numbers, $\mathbb{P}(G) = 1$.

We consider sequences of points $x_n = (q_n, r_n)$, $q_n \in S^d$, $r_n \in \mathbb{R}_+^K$, such that

$$\lim_{n\to\infty} |x_n| = \infty, \qquad \lim_{n\to\infty} \frac{r_n}{|x_n|} = \overline{r}, \qquad \lim_{n\to\infty} \frac{\ell_n^+}{|x_n|} = \overline{\ell} \tag{4.4}$$

for some $\overline{r} = (\overline{r}_1, \ldots, \overline{r}_k) \in \mathbb{R}_+^K$, $\overline{\ell} \in [0, 1]$.

LEMMA 4.1. *Let* $T_0 > 0$. *Assume that a sequence* $x_n$ *satisfies* (4.4) *and let*

$$\mathcal{L}_n = \max_{k\in\mathcal{E}} \max_{1\leqslant i\leqslant N_k^{x_n}(|x_n|T_0)} \ell_k(i). \tag{4.5}$$

*Then* $\lim_{n\to\infty} \mathcal{L}_n(\omega)/|x_n| = 0$ *for every* $\omega \in G$.

Proof. Fix $\omega \in G$. Our aim is to show that for $k \in \mathcal{E}$

$$\frac{1}{|x_n|} \max_{1\leqslant i\leqslant N_k^{x_n}(|x_n|T_0)(\omega)} \ell_k(i)(\omega) \to 0. \tag{4.6}$$

By (4.1) and (4.4), it follows that on the set $G$

$$\frac{1}{|x_n|} N_k^{x_n}(|x_n|t) \to \alpha_k(t - \overline{r}_k)^+ \tag{4.7}$$

uniformly on compacts (u.o.c.) in $t$ (see Lemma 4.2 in [8]). Therefore, to prove (4.6), it suffices to verify that

$$\frac{1}{|x_n|} \max_{1 \leqslant i \leqslant |x_n| T'} \ell_k(i)(\omega) \to 0, \tag{4.8}$$

where $T' = \alpha_k T_0 + 1$. By (4.3) and the functional strong law of large numbers, the process

$$M_k^n(t) = \frac{1}{|x_n|} \sum_{i=1}^{\lfloor |x_n| t \rfloor} \big(\ell_k(i) - \mathbb{E}\ell_k(i)\big)$$

converges to zero u.o.c. in $t \geqslant 0$ on the set $G$. Hence,

$$\frac{1}{|x_n|} \max_{1 \leqslant i \leqslant |x_n| T'} \ell_k(i)(\omega) \leqslant \frac{1}{|x_n|} \mathbb{E}\ell_k(1) + j_{T'}\big(M_k^n(\omega)\big) \to 0,$$

where for $f \in D[0, \infty)$, $j_{T'}(f) = \sup_{0 \leqslant t \leqslant T'} |f(t) - f(t-)|$, and (4.8) follows. ■

## 5. PREEMPTIVE EDF NETWORK EQUATIONS AND FLUID MODELS

In this section we analyze fluid limits for preemptive EDF queueing networks with fixed customer routes described in Sections 2.2.2, 2.2.3 and 2.2.4.

Let $E(t,s) = \big(E_{\mathbf{k}}(t,s)\big)_{\mathbf{k} \in \mathbf{K}}$, $t \geqslant 0$, $s \in \mathbb{R}$, denote the *external arrival process* defined as follows. If $\mathbf{k} = (k, j_{k,1}, 1)$ for some $k$, then $E_{\mathbf{k}}(t,s)$ is equal to the number of external arrivals to the system (or, equivalently, to station $j_{k,1}$) by time $t$ of type $k$ customers with lead times at time $t$ less than or equal to $s - t$; otherwise $E_{\mathbf{k}}(t,s) \equiv 0$. Let $\mathbf{k} = (k, j, b) \in \mathbf{K}$, $t \geqslant 0$ and $s \in \mathbb{R}$, let $Z_{\mathbf{k}}(t,s)$ denote the number of type $k$ customers who are visiting station $j$ for the $b$-th time along their route at time $t$ with lead times at time $t$ less than or equal to $s - t$. Let $Z(t,s) = \big(Z_{\mathbf{k}}(t,s)\big)_{\mathbf{k} \in \mathbf{K}}$. Similarly, the vectors $A(t,s) = \big(A_{\mathbf{k}}(t,s)\big)_{\mathbf{k} \in \mathbf{K}}$, $D(t,s) = \big(D_{\mathbf{k}}(t,s)\big)_{\mathbf{k} \in \mathbf{K}}$, $T(t,s) = \big(T_{\mathbf{k}}(t,s)\big)_{\mathbf{k} \in \mathbf{K}}$ denote the number of arrivals and departures, and the cumulative service time by time $t$ corresponding to each class $\mathbf{k}$ of customers with lead times at time $t$ less than or equal to $s - t$. Let $Y_j(t,s)$, $j = 1, \ldots, J$, denote the cumulative idleness by time $t$ at station $j$ with regard to service of customers with lead times at time $t$ less than or equal to $s - t$ and let $Y(t,s) = \big(Y_j(t,s)\big)_{j=1,\ldots,J}$. For $\mathbf{k} = (k, j, b) \in \mathbf{K}$, $t, t' \geqslant 0$ and $s \in \mathbb{R}$, let $S_{\mathbf{k}}(t', t, s)$ denote the number of service completions at station $j$ of type $k$ customers visiting this station for the $b$-th time along their route and having lead times at time $t$ less than or equal to $s - t$, by the time the station $j$ has spent $t'$ units of time serving these customers. For $t \geqslant 0$ and $s \in \mathbb{R}$, let

$$\mathfrak{X}(t,s) = \big(A(t,s), D(t,s), T(t,s), Y(t,s), Z(t,s)\big).$$

Let $Q(t) = \big(Q_{\mathbf{k}}(t)\big)_{\mathbf{k}\in\mathbf{K}} = \lim_{s\to\infty} Z(t,s)$ be the queue length vector and let $W(t) = \big(W_{\mathbf{k}}(t)\big)_{\mathbf{k}\in\mathbf{K}}$ denote the unfinished work in the system, i.e., $W_{\mathbf{k}}(t)$ is the sum of the residual service times of customers in buffer $\mathbf{k}$ at time $t$. We will sometimes use a superscript $x \in \Omega$ such as in $\mathfrak{X}^x(t,s)$ to indicate that the process starts at state $x$. For $c > 0$, $c\mathfrak{X}(t,s)$ denotes componentwise multiplication.

The process $\mathfrak{X}(t,s)$ satisfies the following *network equations*:

(5.1) $A(t,s) = E(t,s) + P'D(t,s)$;

(5.2) $Z(t,s) = Z(0,s) + A(t,s) - D(t,s)$;

(5.3) $D_{\mathbf{k}}(t,s) = S_{\mathbf{k}}\big(T_{\mathbf{k}}(t,s), t, s\big),\ \mathbf{k} \in \mathbf{K}$;

(5.4) $\sum_{\mathbf{k}\in\bar{\mathcal{C}}(j)} T_{\mathbf{k}}(t,s) + Y_j(t,s) = t,\ j = 1,\dots,J$;

(5.5) $Y_j(t,s)$ can only increase in $t$ when $\sum_{\mathbf{k}\in\bar{\mathcal{C}}(j)} Z_{\mathbf{k}}(t,s) = 0,\ j = 1,\dots,J$,

valid for every $t \geqslant 0$ and $s \in \mathbb{R}$.

The equation (5.5) means that $Y_j(t_1,s) < Y_j(t_2,s)$ implies that

$$\sum_{\mathbf{k}\in\bar{\mathcal{C}}(j)} Z_{\mathbf{k}}(t,s) = 0 \quad \text{for some } t \in [t_1,t_2].$$

The equations (5.1)–(5.4) are general properties of queueing networks without reneging and they do not depend on the service discipline under consideration. The equation (5.5) is specific to preemptive EDF networks. Indeed, for any $s$, the server idleness with regard to customers with lead times not greater than $s - t$ cannot increase at time $t$ in the presence of such customers if and only if the server is working under the preemptive EDF protocol.

It turns out that the deterministic analogs of the equations (5.1)–(5.5) are the *FISFO fluid model equations* (see [7]):

(5.6) $\overline{A}(t,s) = \alpha(t\wedge s) + P'\overline{D}(t,s)$;

(5.7) $\overline{Z}(t,s) = \overline{Z}(0,s) + \overline{A}(t,s) - \overline{D}(t,s)$;

(5.8) $\overline{D}_{\mathbf{k}}(t,s) = \overline{T}_{\mathbf{k}}(t,s)/m_{\mathbf{k}},\ \mathbf{k} \in \mathbf{K}$;

(5.9) $\sum_{\mathbf{k}\in\bar{\mathcal{C}}(j)} \overline{T}_{\mathbf{k}}(t,s) + \overline{Y}_j(t,s) = t,\ j = 1,\dots,J$;

(5.10) $\overline{Y}_j(t,s)$ can only increase in $t$ when $\sum_{\mathbf{k}\in\bar{\mathcal{C}}(j)}\overline{Z}_{\mathbf{k}}(t,s)=0,\ j=1,\dots,J$,

where $t,s \geqslant 0$.

By analogy with the processes $A, D, T, Y, Z$, we assume that $\overline{A}(\cdot, s)$, $\overline{D}(\cdot, s)$, $\overline{T}(\cdot, s)$, $\overline{Y}(\cdot, s)$ are nondecreasing in each coordinate, $\overline{A}(0, s) = \overline{D}(0, s) = \overline{T}(0, s) = 0$ and $\overline{Y}(0, s) = 0$ for $s \geqslant 0$. Similarly, we assume that every coordinate of $\overline{A}(t, \cdot)$, $\overline{D}(t, \cdot)$, $\overline{T}(t, \cdot)$, $-\overline{Y}(t, \cdot)$, $\overline{Z}(t, \cdot)$ is nondecreasing for all $t \geqslant 0$ and that $\overline{Z}_{\mathbf{k}}(t, s) \geqslant 0$, $\mathbf{k} \in \mathbf{K}$. Let $\overline{Q}(t) = \lim_{s\to\infty} \overline{Z}(t, s)$ and let

$$\overline{\mathfrak{X}}(t, s) = \big(\overline{A}(t, s), \overline{D}(t, s), \overline{T}(t, s), \overline{Y}(t, s), \overline{Z}(t, s)\big).$$

As in the case of queueing networks, we say that a fluid model is *strictly subcritical* if $\rho_j < 1$ for each $j$, where $\rho_j$ is defined by (2.7). We also say that a FISFO fluid model is *stable* if there exists $c > 0$ such that, for all solutions of the equations (5.6)–(5.10), $\overline{Q}(t) = 0$ for $t \geqslant c|\overline{Q}(0)|$.

PROPOSITION 5.1. *A strictly subcritical FISFO fluid model of a network with fixed customer routes is stable.*

This follows immediately from Theorem 2 of Bramson [7], because the sets $\mathcal{K}_1 = \mathbf{K}$ and $\mathcal{K}_2 = \emptyset$ have all the properties required by this theorem. Alternatively, we may take $\mathcal{K}_2 = \{\mathbf{k}_0\}$ and $\mathcal{K}_1 = \mathbf{K} - \mathcal{K}_2$, where $\mathbf{k}_0 = \big(1, j_{1,m(1)}, \bar{b}(1, j_{1,m(1)})\big)$, because $\sum_{k\in\mathcal{K}_2} m_{\mathbf{k}}\lambda_{\mathbf{k}} = m_{\mathbf{k}_0}\lambda_{\mathbf{k}_0} \leqslant \rho_{j_{1,m(1)}} < 1$, which, by the remark before Theorem 2 in Bramson [7], is sufficient for the theorem to hold.

LEMMA 5.1. *Let $x_n$ satisfy* (4.4) *and let $\mathbf{k} = (k, j, b) \in \mathbf{K}$. On the set $G$,*

$$\frac{1}{|x_n|} E^{x_n}_{\mathbf{k}}(|x_n|t, |x_n|s) \to \alpha_{\mathbf{k}}\big((t \wedge s) - \overline{r}_k\big)^+ \tag{5.11}$$

*u.o.c. in $t, s \geqslant 0$.*

Proof. Let $k \in \mathcal{E}$. Fix $T_0 > 0$. We claim that, for $s \leqslant t \leqslant T_0$,

$$N^{x_n}_k\big((|x_n|s - \mathcal{L}_n)^+\big) \leqslant E^{x_n}_k(|x_n|t, |x_n|s) \leqslant N^{x_n}_k(|x_n|s). \tag{5.12}$$

Indeed, if $|x_n|s < \mathcal{L}_n$, the first inequality in (5.12) is obvious. Assume that $|x_n|s \geqslant \mathcal{L}_n$. At time $|x_n|t$, the time since the arrival of a customer who has entered the network by time $(|x_n|s - \mathcal{L}_n)^+ = |x_n|s - \mathcal{L}_n$ is at least $|x_n|(t - s) + \mathcal{L}_n$. The initial lead time of this customer is bounded above by $\mathcal{L}_n$, so his lead time at time $|x_n|t$ is not greater than $|x_n|(s - t)$ and the first inequality in (5.12) holds. Since $\ell_k(i) \geqslant 0$ for all $i$, a customer with lead time at time $|x_n|t$ not greater than $|x_n|(s - t)$ must have entered the network by time $|x_n|s$. This explains the second inequality in (5.12). Dividing (5.12) by $|x_n|$, using (2.6), (4.4), (4.7), Lemma 4.1 and the fact that $s \leqslant t$, we have

$$\frac{1}{|x_n|} E^{x_n}_k(|x_n|t, |x_n|s) \to \alpha_k\big((t \wedge s) - \overline{r}_k\big)^+. \tag{5.13}$$

Now, let $t < s \leqslant T_0$. By (4.4) and Lemma 4.1, on the set $G$ for $n$ large enough, $\mathcal{L}_n \leqslant |x_n|(s-t)$. For such $n$,

$$E_k^{x_n}(|x_n|t, |x_n|s) = N_k^{x_n}(|x_n|t) = N_k^{x_n}\big(|x_n|(t \wedge s)\big).$$

Dividing by $|x_n|$ and using (2.6), (4.4), (4.7), we again have (5.13). If $\mathbf{k} \neq k$ for all $k \in \mathcal{E}$, then $E_{\mathbf{k}}^{x_n} \equiv 0$ and $\alpha_{\mathbf{k}} = 0$. Thus, in any case, the convergence (5.11) for any $\mathbf{k} \in \mathbf{K}$ and fixed $t, s$ holds true. Finally, since $E_{\mathbf{k}}^{x_n}(t,s)$ is nondecreasing in both variables and the limit $\alpha_{\mathbf{k}}\big((t \wedge s) - \overline{r}_k\big)^+$ is continuous, it is not hard to see that (5.11) is actually u.o.c. in $t$ and $s$ (see the proofs of Lemma 4.1 in [8] and Proposition 3.4 in [11] for similar arguments). ■

LEMMA 5.2. *Let*

$$C = (1+|\alpha|) \sum_{\mathbf{k} \in \mathbf{K}} m_{\mathbf{k}} + 3. \tag{5.14}$$

*For every sequence $x_n$ in* (4.4), *there exist a set $G_1 \subseteq G$ with $\mathbb{P}(G_1) = 1$ and a subsequence $x_\eta$ such that for $\omega \in G_1$ and $\eta$ sufficiently large*,

$$\bar{V}^{x_\eta}(\omega) \leqslant C\,|x_\eta|, \tag{5.15}$$

*where $\bar{V}^{x_\eta}$ is the departure time of the last initial customer from the network with initial state $x_\eta$.*

Proof. Let a sequence $x_n$ satisfy (4.4). In a preemptive EDF network with the initial state $x_n$, the initial customers, together with customers arriving at the network after time zero with deadlines not greater than $\ell_n^+$, form a *priority class*, i.e., as long as these customers are present at any station of the network, all the service capacity of this station is devoted to them. Since the initial lead times of the arriving customers are nonnegative, this priority class has at most $|q_n| + |N^{x_n}(\ell_n^+)|$ members. Let $\mathbf{k} \in \mathbf{K}$ and let $i_1^n, \ldots, i_{p_{\mathbf{k}}^n}^n$ be the indices of the service times in the sequence $v_{\mathbf{k}}(i)$, $i = 1, 2, \ldots$, corresponding to the priority customers in the network with the initial state $x_n$. We have $p_{\mathbf{k}}^n \leqslant |q_n| + |N^{x_n}(\ell_n^+)|$, $\mathbf{k} \in \mathbf{K}$. Under the EDF service discipline, the index $i$ of the $b$-th arrival at station $j$ of a customer of class $k$ is independent of $v_{\mathbf{k}}(i)$, where $\mathbf{k} = (k, j, b)$. Thus, $\big(v_{\mathbf{k}}(i_1^n), \ldots, v_{\mathbf{k}}(i_{p_{\mathbf{k}}^n}^n)\big)$ have the same distribution as $\big(v_{\mathbf{k}}(1), \ldots, v_{\mathbf{k}}(p_{\mathbf{k}}^n)\big)$. In particular, the sum of the service times of the priority customers in the network with the initial state $x_n$, which will be denoted by $V^{x_n}$, is bounded by the random variable with the same distribution as

$$\tilde{V}^{x_n} = |w_n| + \sum_{\mathbf{k} \in \mathbf{K}} \sum_{i=1}^{|q_n|+|N^{x_n}(\ell_n^+)|} v_{\mathbf{k}}(i) \leqslant |x_n| + \sum_{\mathbf{k} \in \mathbf{K}} \sum_{i=1}^{|x_n|+|N^{\mathbf{0}}(|x_n|)|} v_{\mathbf{k}}(i).$$

By the assumptions (2.2), (2.5), (4.4), (5.14) and the weak law of large numbers, we have $\big(\tilde{V}^{x_n} - (C-2)|x_n|\big)^+ \xrightarrow{P} 0$. Hence, $\big(V^{x_n} - (C-2)|x_n|\big)^+ \xrightarrow{P} 0$. By Theorem 20.5 in [2], there exist a set $G_1$ with $\mathbb{P}(G_1) = 1$ and a subsequence $\eta$ such that, for every $\omega \in G_1$, we have $\big(V^{x_\eta}(\omega) - (C-2)|x_\eta|\big)^+ \to 0$. Thus, for $\omega \in G_1$ and $\eta$ large enough,

$$V^{x_\eta}(\omega) \leqslant (C-1)|x_\eta|. \tag{5.16}$$

Note that since all the priority customers arrive at the preemptive EDF system with initial state $x_\eta$ by time $\ell^+_\eta$, $V^{x_\eta} + \ell^+_\eta$ is the upper bound for the time by which all the priority customers leave this system. Indeed, as long as the priority customers are present at the network, at least one server works on these customers. Consequently, by (5.16), for $\omega \in G_1$ and $\eta$ sufficiently large, (5.15) holds. ∎

For $t_0 \geqslant 0$, we introduce the *time shift operator* $\Delta_{t_0}$ acting on the coordinates of the process $\mathfrak{X}$ as follows for $t, s \geqslant 0$:

$$\begin{aligned}
\Delta_{t_0}A(t,s) &= A(t+t_0, s+t_0) - A(t_0,t_0),\\
\Delta_{t_0}D(t,s) &= D(t+t_0, s+t_0) - D(t_0,t_0),\\
\Delta_{t_0}T(t,s) &= T(t+t_0, s+t_0) - T(t_0,t_0),\\
\Delta_{t_0}Y(t,s) &= Y(t+t_0, s+t_0) - Y(t_0,t_0),\\
\Delta_{t_0}Z(t,s) &= Z(t+t_0, s+t_0).
\end{aligned}$$

Let $\Delta_{t_0}\mathfrak{X} = (\Delta_{t_0}A, \Delta_{t_0}D, \Delta_{t_0}T, \Delta_{t_0}Y, \Delta_{t_0}Z)$ and let $\Delta_{t_0}Q(t) = Q(t+t_0)$ for $t \geqslant 0$. Intuitively, the processes $\Delta_{t_0}\mathfrak{X}$, $\Delta_{t_0}Q$ describe the dynamics of the queueing system under consideration "restarted" at time $t_0$.

The following proposition plays a crucial role in the proof of Theorem 3.1. Its intuitive meaning is that, after a time large enough to process all the initial customers to completion at every station, the fluid limits for a preemptive EDF system satisfy the FISFO fluid model equations.

PROPOSITION 5.2. *Let $C$ be as in* (5.14). *For every sequence $x_n$ in* (4.4), *there exist a set $G' \subseteq G$ with $\mathbb{P}(G') = 1$ and a subsequence $x_\eta$ such that, for each $\omega \in G'$ and each subsequence $x_\vartheta$ of $x_\eta$* (*possibly depending on* $\omega$), *there exists a further subsequence $x_\zeta$ of $x_\vartheta$* (*depending on* $\omega$) *on which $\Delta_{C|x_\zeta|}\mathfrak{X}^{x_\zeta}(t|x_\zeta|, s|x_\zeta|)(\omega)/|x_\zeta|$ converges u.o.c. in t and s and*

$$\lim_{n\to\infty} \Delta_{C|x_\zeta|}\mathfrak{X}^{x_\zeta}(t|x_\zeta|, s|x_\zeta|)(\omega)/|x_\zeta| \tag{5.17}$$

*satisfies the FISFO fluid model equations* (5.6)–(5.10).

The main idea of the proof of Proposition 5.2 is based on the observation that because the initial lead time distributions disappear in the limit, the asymptotic behavior of a preemptive EDF system does not differ from the behavior of the corresponding FISFO system. In particular, under fluid scaling the number of customers coming to the system in a small time interval is small, so the corresponding fluid limits are continuous. Also, since the order of service does not differ significantly from FISFO, the number of partially served customers at each station and the work associated with them are negligible in the limit. The latter finding is analogous to "crushing lemmas" from the papers on diffusion limits for EDF systems (see [11], [14], [23]). The formal, technical proof of Proposition 5.2 has been relegated to the Appendix.

### 6. PROOF OF THEOREM 3.1

To show Theorem 3.1 we will need the following proposition, which will be proved in the Appendix.

PROPOSITION 6.1 (state space collapse). *Let $x_n$ be a sequence satisfying* (4.4). *Let $C$ be given by* (5.14) *and let $G'$ be as in the proof of Proposition* 5.2. *Let $\omega \in G'$, and let $x_\zeta$ be a subsequence* (*depending on* $\omega$) *constructed in the proof of Proposition* 5.2. *Then for each* $\mathbf{k} \in \mathbf{K}$ *and* $t \geqslant 0$

$$\lim_{\zeta\to\infty} \frac{1}{|x_\zeta|} \left|W_{\mathbf{k}}^{x_\zeta}\big((t+C)|x_\zeta|\big) - m_{\mathbf{k}} Q_{\mathbf{k}}^{x_\zeta}\big((t+C)|x_\zeta|\big)\right| = 0. \tag{6.1}$$

REMARK 6.1. Most of the arguments presented in this paper can be generalized to open preemptive EDF networks with Markovian routing. However, we have been unable to show that in this general case the limit (5.17) satisfies the fluid model equation (5.8). It is also unclear how to prove Proposition 6.1 in this generality. This is why our analysis is limited to the case of deterministic customer routes.

We shall now prove Theorem 3.1. The main idea of the proof, which is due to Dai [8], is to approximate the sample paths of the (suitably scaled) performance processes under consideration by the corresponding fluid models and to use stability of the latter models to show (2.12).

Proof of Theorem 3.1. Let

$$\delta = C + c(1 + |\alpha| C), \tag{6.2}$$

where $c$ is the constant appearing in the definition of a stable FISFO fluid model, and $C$ is given by (5.14). By Proposition 2.2, it suffices to show (2.12). If (2.12) is false, there exist $\epsilon > 0$ and a sequence $x_n \in \Omega$ such that $|x_n| \to \infty$ and

$$\mathbb{E}_{x_n}\big|X(\delta|x_n|)\big| \geqslant \epsilon |x_n| \tag{6.3}$$

for every $n$. Without loss of generality we can assume that the sequence $x_n$ satisfies (4.4). Let the set $G'$ and the subsequence $x_\eta$ be as in Proposition 5.2. We will first show that on $G'$

$$\lim_{\eta\to\infty} \bigl|X^{x_\eta}(\delta|x_\eta|)\bigr|/|x_\eta| = 0. \tag{6.4}$$

If this is not the case, there exist $\omega \in G'$, $\epsilon_1 > 0$ and a subsequence $x_\vartheta$ of the sequence $x_\eta$ such that for every $\vartheta$

$$\bigl|X^{x_\vartheta}(\delta|x_\vartheta|)(\omega)\bigr| \geqslant \epsilon_1|x_\vartheta|. \tag{6.5}$$

By Proposition 5.2, the sequence $x_\vartheta$ contains a subsequence $x_\zeta$ such that the limit (5.17) satisfies the FISFO fluid model equations (5.6)–(5.10). By (4.7) and the fact that $G' \subseteq G$, we have

$$\begin{aligned}|\Delta_{C|x_\zeta|}Q^{x_\zeta}(0)(\omega)| = \bigl|Q^{x_\zeta}(C|x_\zeta|)(\omega)\bigr| &\leqslant |x_\zeta| + \bigl|N^{x_\zeta}(C|x_\zeta|)(\omega)\bigr| \\ &\leqslant (1+|\alpha|C)|x_\zeta| + o(|x_\zeta|).\end{aligned}$$

Consequently, $\lim_{\zeta\to\infty} |\Delta_{C|x_\zeta|}Q^{x_\zeta}(0)(\omega)|/|x_\zeta| \leqslant 1 + |\alpha|C$. This, together with Proposition 5.1, yields

$$\begin{aligned}\lim_{\zeta\to\infty} &\bigl|Q^{x_\zeta}(\delta|x_\zeta|)(\omega)\bigr|/|x_\zeta| \\ &= \lim_{\zeta\to\infty} \bigl|\Delta_{C|x_\zeta|}Q^{x_\zeta}\bigl(c(1+|\alpha|C)|x_\zeta|\bigr)(\omega)\bigr|/|x_\zeta| = 0.\end{aligned} \tag{6.6}$$

This, in turn, together with Proposition 6.1, implies that

$$\lim_{\zeta\to\infty} \bigl|W^{x_\zeta}(\delta|x_\zeta|)(\omega)\bigr|/|x_\zeta| = 0. \tag{6.7}$$

Using the fact that $G' \subseteq G$ and arguing as in the proof of Lemma 4.3a in [8], we can show that

$$\lim_{\zeta\to\infty} \bigl|R^{x_\zeta}(\delta|x_\zeta|)(\omega)\bigr|/|x_\zeta| = 0 \tag{6.8}$$

(recall from (2.10) that $R(t)$ is the vector of the residual interarrival times at $t$). Denote the positive part of the greatest lead time in the system at time $t$ by $L_+(t)$. Lemma 4.1, together with the fact that, by Lemma 5.2 and (6.2), for large $\zeta$ there are no initial customers at time $\delta|x_\zeta|$ in the system with initial state $x_\zeta$, implies

$$\lim_{\zeta\to\infty} L_+^{x_\zeta}(\delta|x_\zeta|)(\omega)/|x_\zeta| = 0. \tag{6.9}$$

By (6.6)–(6.9), $\lim_{\zeta\to\infty} \bigl|X^{x_\zeta}(\delta|x_\zeta|)(\omega)\bigr|/|x_\zeta| = 0$, which contradicts (6.5). We have proved (6.4).

Arguing as in the proof of Lemma 5.2 we can show that for every $\mathbf{k} = (k, j, b)$ $W_{\mathbf{k}}^{x_\eta}(\delta|x_\eta|)$ is bounded by a random variable with the same distribution as

$$\tilde{W}_{\mathbf{k}}^{x_\eta}(\delta|x_\eta|) = |x_\eta| + \sum_{i=1}^{N_k^{\mathbf{0}}(\delta|x_\eta|)} v_{\mathbf{k}}(i).$$

By Wald's identity,

$$\mathbb{E}_{x_\eta} W_{\mathbf{k}}(\delta|x_\eta|) \leqslant \mathbb{E}\tilde{W}_{\mathbf{k}}^{x_\eta}(\delta|x_\eta|) = |x_\eta| + m_{\mathbf{k}}\mathbb{E}N_k^{\mathbf{0}}(\delta|x_\eta|). \tag{6.10}$$

Using (6.4), (6.10) and arguing as in the proofs of Lemmas 4.3b and 4.5 in [8], we get

$$\lim_{\eta\to\infty} \mathbb{E}_{x_\eta}\big|X(\delta|x_\eta|)\big|/|x_\eta| = 0,$$

which contradicts (6.3). ■

## 7. PROOF OF THEOREM 3.2

In this section we prove Theorem 3.2. The argument is similar to the proof of Theorem 3.1, but simpler. In particular, we shall now verify (2.12) directly, without using a fluid model. The main idea of the proof is that, because of reneging, after a time large enough to process all the initial customers to completion, the number of customers present in the system cannot be very large.

Proof of Theorem 3.2. Let $\delta = 3/2$. Again, by Proposition 2.2, it is sufficient to show (2.12). If (2.12) is false, there exist $\epsilon > 0$ and a sequence $x_n \in \Omega$ such that $|x_n| \to \infty$ and (6.3) holds for every $n$. Without loss of generality we can assume that $|x_n|$ increases with $n$ and the sequence $x_n$ satisfies (4.4). For $\mathbf{k} \in \mathbf{K}$ and $0 \leqslant t_1 < t_2$, let $B_{\mathbf{k}}^n(t_1, t_2)$ denote the set of $j = 1, 2, \ldots$ for which the customer corresponding to the service time $v_{\mathbf{k}}(j)$ has entered the network with initial state $x_n$ in the time interval $(t_1|x_n|, t_2|x_n|]$. Let $\gamma$ denote the maximal expected number of visits to all buffers in the network by a customer entering the network at any $\mathbf{k} \in \mathcal{E}$. Proceeding similarly to the proof of Proposition 3.2 in [7], we can show that there exists a set $G' \subseteq G$ with $\mathbb{P}(G') = 1$ such that on $G'$ for each $r > 0$ and $\mathbf{k} \in \mathbf{K}$

$$\frac{1}{|x_n|}\Big|\sum_{i\in B_{\mathbf{k}}^n(0,r)} v_{\mathbf{k}}(i) - m_{\mathbf{k}}|B_{\mathbf{k}}^n(0,r)|\Big| \to 0 \tag{7.1}$$

and for every $\mathbf{k} \in \mathbf{K}$, $r_1 < r_2$ we have

$$|B_{\mathbf{k}}^n(0,r_2)| - |B_{\mathbf{k}}^n(0,r_1)| \leqslant 4|\alpha|\gamma(r_2 - r_1)|x_n| \tag{7.2}$$

for $n$ large enough. We want to show that on $G'$

$$\lim_{n\to\infty} \big|X^{x_n}(\delta|x_n|)\big|/|x_n| = 0. \tag{7.3}$$

First note that $\ell_n \leqslant |x_n|$, so because of reneging, none of the initial customers of the network with initial state $x_n$ is present at this network at time $\delta|x_n|$. Put $T_0 = \delta$ in Lemma 4.1. Because of reneging, the customers present at the network with initial state $x_n$ at time $\delta|x_n|$ must have entered it after time $\delta|x_n| - \mathcal{L}_n$. Thus, on $G'$, by (4.7) and Lemma 4.1,

$$\begin{aligned} \big|Q^{x_n}(\delta|x_n|)\big| &\leqslant \big|N^{x_n}(\delta|x_n|) - N^{x_n}(\delta|x_n| - \mathcal{L}_n)\big| \\ &\leqslant |\alpha|\mathcal{L}_n + o(|x_n|) = o(|x_n|), \end{aligned} \tag{7.4}$$

$$\big|W^{x_n}(\delta|x_n|)\big| \leqslant \sum_{\mathbf{k}\in\mathbf{K}} \sum_{i\in B^n_{\mathbf{k}}(\delta-\mathcal{L}_n/|x_n|,\delta)} v_{\mathbf{k}}(i). \tag{7.5}$$

By (7.4), on the set $G'$ we have

$$\lim_{n\to\infty} \big|Q^{x_n}(\delta|x_n|)\big|/|x_n| = 0. \tag{7.6}$$

Fix $\epsilon \in (0, 1/2)$. By (7.5), (7.1) and (7.2), on the set $G' \cap [\mathcal{L}_n \leqslant \epsilon|x_n|]$ we have

$$\begin{aligned} \big|W^{x_n}(\delta|x_n|)\big| &\leqslant \sum_{\mathbf{k}\in\mathbf{K}} \sum_{i\in B^n_{\mathbf{k}}(\delta-\epsilon,\delta)} v_{\mathbf{k}}(i) \\ &= \sum_{\mathbf{k}\in\mathbf{K}} m_{\mathbf{k}}\big(|B^n_{\mathbf{k}}(0,\delta)| - |B^n_{\mathbf{k}}(0,\delta-\epsilon)|\big)|x_n| + o(|x_n|) \\ &\leqslant 4|\alpha|\gamma\epsilon \sum_{\mathbf{k}\in\mathbf{K}} m_{\mathbf{k}}\, |x_n| \;+\; o(|x_n|) \end{aligned} \tag{7.7}$$

for $n$ large enough. By (7.7), Lemma 4.1 and the fact that $\epsilon \in (0, 1/2)$ is arbitrary,

$$\lim_{n\to\infty} \big|W^{x_n}(\delta|x_n|)\big|/|x_n| = 0 \tag{7.8}$$

on $G'$. The proof of the fact that

$$\lim_{n\to\infty} \big|R^{x_n}(\delta|x_n|)\big|/|x_n| = \lim_{n\to\infty} L^{x_n}_{+}(\delta|x_n|)/|x_n| = 0$$

on $G'$, where $L_+(t)$ is the greatest lead time in the system at time $t$, is similar to the corresponding argument in the proof of Theorem 3.1. This shows (7.3). Using (7.3) and arguing as in the proofs of Lemmas 4.3b and 4.5 in [8], we get

$$\lim_{n\to\infty} \mathbb{E}_{x_n}\big|X(\delta|x_n|)\big|/|x_n| = 0,$$

which contradicts (6.3). ■

## 8. APPENDIX

This appendix contains the proofs of Propositions 5.2 and 6.1.

Proof of Proposition 5.2. Let a sequence $x_n$ satisfy (4.4) and let the set $G_1$ and the subsequence $x_\eta$ be as in Lemma 5.2.

For $\mathbf{k} = (k, j, b) \in \mathbf{K}$ and $t_1, t_2 \geqslant 0$, let $B^n_{\mathbf{k}}(t_1, t_2)$ denote the set of $j = 1, 2, \ldots$ for which the customer corresponding to the service time $v_{\mathbf{k}}(j)$ has entered the network with initial state $x_n$ in the time interval $(t_1|x_n|, t_2|x_n|]$. In particular, $B^n_{\mathbf{k}}(t_1, t_2) = \emptyset$ if $t_1 \geqslant t_2$. By the independence of the interarrival times and the service times, together with the weak law of large numbers, we have

$$\frac{1}{|x_\eta|}\Big|\sum_{i\in B^\eta_{\mathbf{k}}(t_1,t_2)} v_{\mathbf{k}}(i) - m_{\mathbf{k}}|B^\eta_{\mathbf{k}}(t_1,t_2)|\Big| \xrightarrow{P} 0. \tag{8.1}$$

However, (4.7) implies that on the set $G$, for $0 \leqslant t_1 < t_2$ we get

$$\begin{aligned}|B^\eta_{\mathbf{k}}(t_1,t_2)| &= N^{x_\eta}_k(t_2|x_\eta|) - N^{x_\eta}_k(t_1|x_\eta|)\\ &= \alpha_k|x_\eta|\big((t_2-\overline{r}_k)^+ - (t_1-\overline{r}_k)^+\big) + o(|x_\eta|),\end{aligned}$$

and hence (8.1) yields

$$\left|\frac{1}{|x_\eta|}\sum_{i\in B^\eta_{\mathbf{k}}(t_1,t_2)} v_{\mathbf{k}}(i) - \alpha_k m_{\mathbf{k}}\big((t_2-\overline{r}_k)^+ - (t_1-\overline{r}_k)^+\big)\right| \xrightarrow{P} 0. \tag{8.2}$$

Using (8.2) and arguing as in the proof of (A.1) in [7] or in the proof of Proposition 3.4 in [11], we get, for every $r_0 > 0$,

$$\sup_{0\leqslant t_1<t_2\leqslant r_0}\left|\frac{1}{|x_\eta|}\sum_{i\in B^\eta_{\mathbf{k}}(t_1,t_2)} v_{\mathbf{k}}(i) - \alpha_k m_{\mathbf{k}}\big((t_2-\overline{r}_k)^+ - (t_1-\overline{r}_k)^+\big)\right| \xrightarrow{P} 0.$$

By Theorem 20.5 in [2], there exist a set $G_2$ with $\mathbb{P}(G_2) = 1$ and a subsequence (still denoted by $\eta$) such that on $G_2$ we have the pointwise convergence

$$\sup_{0\leqslant t_1<t_2\leqslant r_0}\left|\frac{1}{|x_\eta|}\sum_{i\in B^\eta_{\mathbf{k}}(t_1,t_2)} v_{\mathbf{k}}(i) - \alpha_k m_{\mathbf{k}}\big((t_2-\overline{r}_k)^+ - (t_1-\overline{r}_k)^+\big)\right| \to 0. \tag{8.3}$$

Let $G' = G \cap G_1 \cap G_2$. We have $\mathbb{P}(G') = 1$. Fix $\omega \in G'$. Consider an arbitrary subsequence $\vartheta$ of the sequence $\eta$. For $t, s \geqslant 0$, let

$$\begin{aligned}\overline{\mathfrak{X}}^{(\vartheta)}(t,s) &= \big(\overline{A}^{(\vartheta)}(t,s), \overline{D}^{(\vartheta)}(t,s), \overline{T}^{(\vartheta)}(t,s), \overline{Y}^{(\vartheta)}(t,s), \overline{Z}^{(\vartheta)}(t,s)\big)\\ &= \Delta_{C|x_\vartheta|}\mathfrak{X}^{x_\vartheta}(t|x_\vartheta|, s|x_\vartheta|)/|x_\vartheta|.\end{aligned}$$

The coordinate mappings of $\overline{\mathfrak{X}}^{(\vartheta)}(\omega)$ inherit the monotonicity properties from the corresponding coordinate mappings of $\mathfrak{X}^{x_\vartheta}(\omega)$. Thus, by Helley's choice theorem (see, e.g., [2], Theorem 25.9 and the remark in the proof of Theorem 29.3), there exists a subsequence $\zeta$ and a right-continuous function

$$\overline{\mathfrak{X}}(t,s) = \big(\overline{A}(t,s), \overline{D}(t,s), \overline{T}(t,s), \overline{Y}(t,s), \overline{Z}(t,s)\big), \qquad t,s \geqslant 0$$

(both depending on $\omega$) such that each coordinate map of $\overline{\mathfrak{X}}^{(\zeta)}(\omega)$ converges to the corresponding coordinate map of $\overline{\mathfrak{X}}$ at every point of continuity of the latter function. Since a monotone function has at most countably many discontinuities, (5.1)–(5.2), (5.4) and Lemma 5.1 imply that $\overline{\mathfrak{X}}$ satisfies (5.6)–(5.7), (5.9). In particular, because $\overline{T}_{\mathbf{k}}(\cdot,s)$ and $\overline{Y}_j(\cdot,s)$ are nondecreasing, (5.9) implies that the functions $\overline{T}(t,s)$ and $\overline{Y}(t,s)$ are Lipschitz in $t$. We will show that they are also continuous in $s$. Let $T_0 > C+1$. Suppose that for some $\mathbf{k} = (k,j,b) \in \mathbf{K}$, $0 \leqslant t < T_0 - C - 1$ and $s > 0$,

$$2\epsilon \triangleq \overline{T}_{\mathbf{k}}(t,s) - \overline{T}_{\mathbf{k}}(t,s-) > 0. \tag{8.4}$$

Let $s_1$, $s_2$ be such that $0 < s_1 < s < s_2$,

$$s_2 - s_1 < \epsilon/(\alpha_k m_{\mathbf{k}}), \tag{8.5}$$

and the function $\overline{T}_{\mathbf{k}}$ is continuous at the points $(t,s_1)$, $(t,s_2)$. By (8.4) and the monotonicity of $T(t,s)$ in $s$, for $\zeta$ large enough we have

$$\epsilon|x_\zeta| \leqslant T_{\mathbf{k}}^{x_\zeta}\big((t+C)|x_\zeta|, (s_2+C)|x_\zeta|\big) - T_{\mathbf{k}}^{x_\zeta}\big((t+C)|x_\zeta|, (s_1+C)|x_\zeta|\big). \tag{8.6}$$

In other words, the cumulative work done by time $(t+C)|x_\zeta|$ by server $j$ on type $k$ customers with lead times at time $(t+C)|x_\zeta|$ belonging to the interval $\big((s_1-t)|x_\zeta|, (s_2-t)|x_\zeta|\big]$ during their $b$-th visit at station $j$ is at least $\epsilon|x_\zeta|$. It is easy to check that these customers arrived at the network in the time interval $\big((s_1+C)|x_\zeta| - \mathcal{L}_\zeta, (s_2+C)|x_\zeta|\big]$. By (8.3), we have

$$\begin{aligned} \epsilon|x_\zeta| &\leqslant \sum_{i \in B_{\mathbf{k}}^{\zeta}(s_1+C-\mathcal{L}_\zeta/|x_\zeta|, s_2+C)} v_{\mathbf{k}}(i) \\ &\leqslant \alpha_k m_{\mathbf{k}}\big((s_2-s_1)|x_\zeta| + \mathcal{L}_\zeta\big) + o(|x_\zeta|). \end{aligned} \tag{8.7}$$

This, by (8.5) and Lemma 4.1, yields a contradiction for sufficiently large $\zeta$. We have proved continuity of $\overline{T}(t,s)$ in $s$ (the argument actually shows that $\overline{T}_{\mathbf{k}}(t,s)$ is Lipschitz in $s$ with the Lipschitz constant $\alpha_k m_{\mathbf{k}}$). By (5.9), $\overline{Y}$ is Lipschitz in both variables, so $(\overline{T}^{(\zeta)}, \overline{Y}^{(\zeta)})(t,s) \to (\overline{T}, \overline{Y})(t,s)$ for any $t,s \geqslant 0$. As in the proof of Lemma 5.1, it is easy to see that this convergence is u.o.c. in $t$ and $s$.

We will now show that $\overline{\mathfrak{X}}$ satisfies (5.8). Let $\mathbf{k} = (k, j, b) \in \mathbf{K}$ and $T_0 > 0$. For every $0 \leqslant t, s < T_0 - C - 1$, we have

$$T_{\mathbf{k}}^{(\zeta)}\big((t+C)|x_\zeta|, (s+C)|x_\zeta|\big) = T_{\mathbf{k},1}^{(\zeta)} + T_{\mathbf{k},2}^{(\zeta)}(t,s) + T_{\mathbf{k},3}^{(\zeta)}(t,s), \tag{8.8}$$

$$D_{\mathbf{k}}^{(\zeta)}\big((t+C)|x_\zeta|, (s+C)|x_\zeta|\big) = D_{\mathbf{k},1}^{(\zeta)} + D_{\mathbf{k},2}^{(\zeta)}(t,s), \tag{8.9}$$

where the quantities on the right-hand side of (8.8) and (8.9) are defined as follows.

First, $T_{\mathbf{k},1}^{(\zeta)}$ is the sum of the service times of type $k$ customers present in the system at time $0$ corresponding to their $b$-th visit at server $j$. In particular, this quantity does not depend on $t$ and $s$. Let us note that for every $t \geqslant 0$ the lead times at time $(t+C)|x_\zeta|$ of the initial customers are bounded by $\ell_\zeta^+ - (t+C)|x_\zeta|$, which in turn is dominated by $|x_\zeta|(s-t)$ for all $s \geqslant 0$ since $\ell_\zeta^+ \leqslant |x_\zeta|$ and $C > 1$. Also, by Lemma 5.2, all the initial customers have been served to completion at every station by time $C|x_\zeta|$. Hence, $T_{\mathbf{k},1}^{(\zeta)}$ is the portion of $T_{\mathbf{k}}^{(\zeta)}\big((t+C)|x_\zeta|, (s+C)|x_\zeta|\big)$ devoted to the initial customers.

The quantity $T_{\mathbf{k},2}^{(\zeta)}(t,s)$ is the sum of the service times of type $k$ customers visiting $j$ for the $b$-th time who arrived at the system after time $0$, have been fully served by time $(t+C)|x_\zeta|$ during this visit at $j$, and have lead times at $(t+C)|x_\zeta|$ not greater than $|x_\zeta|(s-t)$.

Next, $T_{\mathbf{k},3}^{(\zeta)}(t,s)$ is the time devoted by server $j$ to those type $k$ customers visiting $j$ for the $b$-th time who arrived at the system after time $0$, have lead times at time $(t+C)|x_\zeta|$ not greater than $|x_\zeta|(s-t)$, and have been only partially served by time $(t+C)|x_\zeta|$ during this visit at $j$.

Finally, $D_{\mathbf{k},1}^{(\zeta)}$ ($D_{\mathbf{k},2}^{(\zeta)}(t,s)$) is the number of customers whose service times are counted in $T_{\mathbf{k},1}^{(\zeta)}$ ($T_{\mathbf{k},2}^{(\zeta)}(t,s)$).

It is easy to see that to show (5.8) it suffices to verify the relations

$$T_{\mathbf{k},2}^{(\zeta)}(t,s) = m_{\mathbf{k}} D_{\mathbf{k},2}^{(\zeta)}(t,s) + o(|x_\zeta|), \tag{8.10}$$

$$T_{\mathbf{k},3}^{(\zeta)}(t,s) = o(|x_\zeta|). \tag{8.11}$$

For $t \geqslant 0$, let $a_{\mathbf{k}}^{(\zeta)}(t)$ be the arrival time at the network with initial state $x_\zeta$ of the type $k$ customer who was the last one to receive service at station $j$ by time $(t+C)|x_\zeta|$ during his $b$-th visit at $j$. Every type $k$ customer who arrived at the network before $a_{\mathbf{k}}^{(\zeta)}(t) - \mathcal{L}_\zeta$ (in particular, by $a_{\mathbf{k}}^{(\zeta)}(t) - \mathcal{L}_\zeta - 1$) has already finished his $b$-th visit at $j$ by time $(t+C)|x_\zeta|$. Similarly, type $k$ customers who arrived at the network after $a_{\mathbf{k}}^{(\zeta)}(t) + \mathcal{L}_\zeta$ cannot preempt the type $k$ customer who arrived at time $a_{\mathbf{k}}^{(\zeta)}(t)$, and hence such customers have not received any service by time $(t+C)|x_\zeta|$ during their $b$-th visit at $j$. Every customer who has entered the network by time $(s+C)|x_\zeta| - \mathcal{L}_\zeta$ has lead time by time $(t+C)|x_\zeta|$ not greater than $|x_\zeta|(s-t)$. On the other hand, each customer who has entered the network after time $(s+C)|x_\zeta|$ has lead time by time $(t+C)|x_\zeta|$ greater than $|x_\zeta|(s-t)$. These facts, together with (4.7) and (8.3), imply

$$\begin{aligned}(8.12)\quad &\alpha_k m_{\mathbf{k}}\Big(\Big(\big(a_{\mathbf{k}}^{(\zeta)}(t)-\mathcal{L}_\zeta-1\big)/|x_\zeta|\Big)\wedge(s+C-\mathcal{L}_\zeta/|x_\zeta|)-\overline{r}_k\Big)^{+}+o(1)\\ &=\frac{1}{|x_\zeta|}\sum_{i\in B_{\mathbf{k}}^{\zeta}(0,((a_{\mathbf{k}}^{(\zeta)}(t)-\mathcal{L}_\zeta-1)/|x_\zeta|)\wedge(s+C-\mathcal{L}_\zeta/|x_\zeta|))} v_{\mathbf{k}}(i)\\ &\leqslant\frac{1}{|x_\zeta|}\,T_{\mathbf{k},2}^{(\zeta)}(t,s)\leqslant\frac{1}{|x_\zeta|}\sum_{i\in B_{\mathbf{k}}^{\zeta}(0,((a_{\mathbf{k}}^{(\zeta)}(t)+\mathcal{L}_\zeta)/|x_\zeta|)\wedge(s+C))} v_{\mathbf{k}}(i)\\ &=\alpha_k m_{\mathbf{k}}\Big(\Big(\big(a_{\mathbf{k}}^{(\zeta)}(t)+\mathcal{L}_\zeta\big)/|x_\zeta|\Big)\wedge(s+C)-\overline{r}_k\Big)^{+}+o(1),\end{aligned}$$

$$\begin{aligned}(8.13)\quad &\alpha_k\Big(\Big(\big(a_{\mathbf{k}}^{(\zeta)}(t)-\mathcal{L}_\zeta-1\big)/|x_\zeta|\Big)\wedge(s+C-\mathcal{L}_\zeta/|x_\zeta|)-\overline{r}_k\Big)^{+}+o(1)\\ &=\frac{1}{|x_\zeta|}N_k^{x_\zeta}\Big((a_{\mathbf{k}}^{(\zeta)}(t)-\mathcal{L}_\zeta-1)\wedge\big((s+C)|x_\zeta|-\mathcal{L}_\zeta\big)\Big)\\ &\leqslant\frac{1}{|x_\zeta|}\,D_{\mathbf{k},2}^{(\zeta)}(t,s)\leqslant\frac{1}{|x_\zeta|}N_k^{x_\zeta}\Big((a_{\mathbf{k}}^{(\zeta)}(t)+\mathcal{L}_\zeta)\wedge\big((s+C)|x_\zeta|\big)\Big)\\ &=\alpha_k\Big(\Big(\big(a_{\mathbf{k}}^{(\zeta)}(t)+\mathcal{L}_\zeta\big)/|x_\zeta|\Big)\wedge(s+C)-\overline{r}_k\Big)^{+}+o(1).\end{aligned}$$

Lemma 4.1, (4.4) and (8.12)–(8.13) imply (8.10). Similarly, we get

$$\begin{aligned}0\leqslant\frac{1}{|x_\zeta|}T_{\mathbf{k},3}^{(\zeta)}(t,s)&\leqslant\frac{1}{|x_\zeta|}\sum_{i\in B_{\mathbf{k}}^{\zeta}(((a_{\mathbf{k}}^{(\zeta)}(t)-\mathcal{L}_\zeta-1)/|x_\zeta|),(a_{\mathbf{k}}^{(\zeta)}(t)+\mathcal{L}_\zeta)/|x_\zeta|)} v_{\mathbf{k}}(i)\\ &\leqslant\alpha_k m_{\mathbf{k}}(2\mathcal{L}_\zeta+1)/|x_\zeta|+o(1)=o(1),\end{aligned}$$

and (8.11) holds true. We have proved that $\overline{\mathfrak{X}}$ satisfies (5.8).

By (5.6)–(5.9) and the Lipschitz continuity of $(\overline{T},\overline{Y})$, $\overline{\mathfrak{X}}(t,s)$ is Lipschitz in both variables, and consequently

$$(8.14)\qquad \overline{\mathfrak{X}}^{(\zeta)}(t,s)\to\overline{\mathfrak{X}}(t,s)$$

u.o.c. in $t$ and $s$.

Finally, we show that $\overline{\mathfrak{X}}$ satisfies (5.10). Let $T_0>0$, $s\geqslant 0$. By (5.5), we have

$$(8.15)\qquad \int_0^{T_0}\sum_{\mathbf{k}\in\bar{\mathcal{C}}(j)}\overline{Z}_{\mathbf{k}}^{(\zeta)}(t,s)\,\overline{Y}_j^{(\zeta)}(dt,s)=0.$$

By Lemma 4.4 in [8], (8.14) and (8.15) imply

$$\int_0^{T_0}\sum_{\mathbf{k}\in\bar{\mathcal{C}}(j)}\overline{Z}_{\mathbf{k}}(t,s)\,\overline{Y}_j(dt,s)=0$$

and (5.10) is satisfied. ■

Proof of Proposition 6.1. By Lemma 5.2, for $\zeta$ sufficiently large, all initial customers have left the system with initial state $x_\zeta$ by time $C|x_\zeta|$. Let $T_0 > C+1$ and let $0 \leqslant t \leqslant T_0 - C - 1$. Let $a^{(\zeta)}_{\mathbf{k}}(t)$, $\mathbf{k} \in \mathbf{K}$, be as in the proof of Proposition 5.2. Recall that, for every $\mathbf{k} = (k, j_{k,i}, b) \in \mathbf{K}$, each type $k$ customer who arrived at the network before $a^{(\zeta)}_{\mathbf{k}}(t) - \mathcal{L}_\zeta$ (in particular, by $a^{(\zeta)}_{\mathbf{k}}(t) - \mathcal{L}_\zeta - 1$) has already been served to completion by time $(t+C)|x_\zeta|$ during his $b$-th visit at $j_{k,i}$. Also, a customer of type $k$ who arrived at the network after $a^{(\zeta)}_{\mathbf{k}}(t) + \mathcal{L}_\zeta$ has not received any service by time $(t+C)|x_\zeta|$ during his $b$-th visit at $j_{k,i}$, and hence, if he is already in the network at time $(t+C)|x_\zeta|$, he is either a class $\mathbf{k}$ customer who has not received any service or he is still "upstream", i.e., a member of some class $\mathbf{l} = (k, j_{k,i'}, b')$, $i' < i$. To show (6.1), we need to analyze two cases: $\mathbf{k} = (k, j_{k,1}, 1)$ for some $k$ and $\mathbf{k} = (k, j_{k,i}, b)$ for some $k$, $b$ and $i > 1$. We will consider only the second case; the proof in the first one is similar, but simpler, since if $\mathbf{k} = (k, j_{k,1}, 1)$, there is no need to take "upstream" customers into account. Let $\mathbf{k} = (k, j_{k,i}, b)$, $i > 1$, and let $\mathbf{k}' = (k, j_{k,i-1}, b')$. By (4.7), (8.3) and the facts recalled above, we have

$$\begin{aligned}
&\alpha_k m_{\mathbf{k}}\Big(\big(a^{(\zeta)}_{\mathbf{k}'}(t) - \mathcal{L}_\zeta - 1 - \bar{r}_k|x_\zeta|\big)^+ - \big(a^{(\zeta)}_{\mathbf{k}}(t) + \mathcal{L}_\zeta - \bar{r}_k|x_\zeta|\big)^+\Big)^+ + o(|x_\zeta|)\\
&= \sum_{i\in B^{\zeta}_{\mathbf{k}}((a^{(\zeta)}_{\mathbf{k}}(t)+\mathcal{L}_\zeta)/|x_\zeta|,(a^{(\zeta)}_{\mathbf{k}'}(t)-\mathcal{L}_\zeta-1)/|x_\zeta|)} v_{\mathbf{k}}(i) \leqslant W^{x_\zeta}_{\mathbf{k}}\big((t+C)|x_\zeta|\big)\\
&\leqslant \sum_{i\in B^{\zeta}_{\mathbf{k}}((a^{(\zeta)}_{\mathbf{k}}(t)-\mathcal{L}_\zeta-1)/|x_\zeta|,(a^{(\zeta)}_{\mathbf{k}'}(t)+\mathcal{L}_\zeta)/|x_\zeta|)} v_{\mathbf{k}}(i)\\
&= \alpha_k m_{\mathbf{k}}\Big(\big(a^{(\zeta)}_{\mathbf{k}'}(t) + \mathcal{L}_\zeta - \bar{r}_k|x_\zeta|\big)^+ - \big(a^{(\zeta)}_{\mathbf{k}}(t) - \mathcal{L}_\zeta - 1 - \bar{r}_k|x_\zeta|\big)^+\Big)^+ + o(|x_\zeta|),
\end{aligned}$$

$$\begin{aligned}
&\alpha_k\Big(\big(a^{(\zeta)}_{\mathbf{k}'}(t) - \mathcal{L}_\zeta - 1 - \bar{r}_k|x_\zeta|\big)^+ - \big(a^{(\zeta)}_{\mathbf{k}}(t) + \mathcal{L}_\zeta - \bar{r}_k|x_\zeta|\big)^+\Big)^+ + o(|x_\zeta|)\\
&= \Big(N^{x_\zeta}_k\big(a^{(\zeta)}_{\mathbf{k}'}(t) - \mathcal{L}_\zeta - 1\big) - N^{x_\zeta}_k\big(a^{(\zeta)}_{\mathbf{k}}(t) + \mathcal{L}_\zeta\big)\Big)^+ \leqslant Q^{(\zeta)}_{\mathbf{k}}(t)\\
&\leqslant \Big(N^{x_\zeta}_k\big(a^{(\zeta)}_{\mathbf{k}'}(t) + \mathcal{L}_\zeta\big) - N^{x_\zeta}_k\big(a^{(\zeta)}_{\mathbf{k}}(t) - \mathcal{L}_\zeta - 1\big)\Big)^+\\
&= \alpha_k\Big(\big(a^{(\zeta)}_{\mathbf{k}'}(t) + \mathcal{L}_\zeta - \bar{r}_k|x_\zeta|\big)^+ - \big(a^{(\zeta)}_{\mathbf{k}}(t) - \mathcal{L}_\zeta - 1 - \bar{r}_k|x_\zeta|\big)^+\Big)^+ + o(|x_\zeta|).
\end{aligned}$$

This, together with Lemma 4.1, shows (6.1). ■

## REFERENCES

[1] F. Bacelli, P. Boyer and G. Hebuterne, *Single server queues with impatient customers*, Adv. in Appl. Probab. 16 (1984), pp. 887–905.

[2] P. Billingsley, *Probability and Measure*, 2nd edition, Wiley, New York 1986.

[3] M. Bramson, *Instability of FIFO queueing networks*, Ann. Appl. Probab. 4 (1994), pp. 414–431.

[4] M. Bramson, *Instability of FIFO queueing networks with quick service times*, Ann. Appl. Probab. 4 (1994), pp. 693–718.

[5] M. Bramson, *Convergence to equilibria for fluid models of FIFO queueing networks*, Queueing Syst. 22 (1996), pp. 5–45.

[6] M. Bramson, *Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks*, Queueing Syst. 23 (1996), pp. 1–26.

[7] M. Bramson, *Stability of earliest-due-date, first-served queueing networks*, Queueing Syst. 39 (2001), pp. 79–102.

[8] J. G. Dai, *On positive Harris recurrence of multiclass queueing networks*: *a unified approach via fluid limit models*, Ann. Appl. Probab. 5 (1995), pp. 49–77.

[9] J. G. Dai and G. Weiss, *Stability and instability for fluid models of reentrant lines*, Math. Oper. Res. 21 (1996), pp. 115–134.

[10] M. H. A. Davis, *Piecewise-deterministic Markov processes*: *a general class of non-diffusion stochastic models*, J. Roy. Statist. Soc. Ser. B 46 (1984), pp. 353–388.

[11] B. Doytchinov, J. P. Lehoczky and S. E. Shreve, *Real-time queues in heavy traffic with earliest-deadline-first queue discipline*, Ann. Appl. Probab. 11 (2001), pp. 332–379.

[12] R. K. Getoor, *Transience and recurrence of Markov processes*, in: *Séminaire de Probabilités XIV*, Lecture Notes in Math. No. 784, Springer, New York 1980, pp. 397–409.

[13] W. Hopp and M. Spearman, *Factory Physics: Foundations of Manufacturing Management*, Irwin, Chicago 1996.

[14] Ł. Kruk, J. P. Lehoczky, S. E. Shreve and S.-N. Yeung, *Earliest-deadline-first service in heavy traffic acyclic networks*, Ann. Appl. Probab. 14 (2004), pp. 1306–1352.

[15] R. Lillo and M. Martin, *Stability in queues with impatient customers*, Stoch. Models 17 (2001), pp. 375–389.

[16] S. P. Meyn and D. Down, *Stability of generalized Jackson networks*, Ann. Appl. Probab. 4 (1994), pp. 124–148.

[17] S. P. Meyn and R. J. Tweedie, *Stability of Markovian processes III*: *Foster–Lyapunov criteria for continuous-time processes*, Adv. in Appl. Probab. 25 (1993), pp. 518–548.

[18] S. P. Meyn and R. J. Tweedie, *State-dependent criteria for convergence of Markov chains*, Ann. Appl. Probab. 4 (1994), pp. 149–168.

[19] A. N. Rybko and A. L. Stolyar, *Ergodicity of stochastic processes describing the operations of open queueing networks*, Probl. Inf. Transm. 28 (1992), pp. 199–220.

[20] R. E. Stanford, *Reneging phenomena in single channel queues*, Math. Oper. Res. 4 (1979), pp. 162–178.

[21] J. A. Stankovic, M. Spuri, K. Ramamritham and G. C. Buttazzo, *Deadline Scheduling for Real-Time Systems*, Springer, 1998.

[22] A. R. Ward and N. Bambos, *On stability of queueing networks with job deadlines*, J. Appl. Probab. 40 (2003), pp. 293–304.

[23] S.-N. Yeung and J. P. Lehoczky, *Real-time queueing networks in heavy traffic with EDF and FIFO queue discipline*, working paper, 2001, Department of Statistics, Carnegie Mellon University.

Institute of Mathematics
Maria Curie-Skłodowska University,
pl. Marii Curie-Skłodowskiej 1
20-031 Lublin, Poland
*E-mail*: lkruk@hektor.umcs.lublin.pl

*Received on 27.12.2006;*
*revised version on 17.9.2007*