doi: 10.17951/a.2019.73.2.105-134

ANNALES
UNIVERSITATIS MARIAE CURIE-SKŁODOWSKA
LUBLIN – POLONIA

VOL. LXXIII, NO. 2, 2019 SECTIO A 105–134

ŁUKASZ KRUK

Stability of preemptive EDF queueing networks

Dedicated to Professor Yuri Kozitsky on the occasion of his 70th birthday

Abstract. We show stability of preemptive, strictly subcritical EDF networks with Markovian routing. To this end, we prove that the associated fluid limits satisfy the first-in-system, first-out (FISFO) fluid model equations and thus, by an extension of a result of Bramson (2001), the corresponding fluid models are stable. We also demonstrate that in a preemptive multiclass EDF network, after a time large enough to process all the initial customers to completion, the maximal number of partially served customers in the system over a finite time horizon converges to zero in L^1 under fluid scaling.

1. Introduction

A fundamental question in the theory of multiclass queueing networks is whether a given system 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. Apparently, there is no general criterion for this behavior; in particular the usual necessary traffic condition that $\rho_j < 1$ at each station, called strict subcriticality of the underlying queueing system, is not sufficient, see, e.g., [15]. On the positive side, the condition $\rho_j < 1$ for all j is sufficient

2010 *Mathematics Subject Classification.* Primary: 60K25, 90B15; Secondary: 68M20.

Key words and phrases. Multiclass queueing networks, deadlines, preemption, stability, fluid models, fluid limits.

for generalized Jackson networks [14] and multiclass networks with some disciplines, including first-in-first-out (FIFO) in networks of Kelly type [2], head-of-the-line proportional processor sharing [3], first-buffer-first-served and last-buffer-first-served [6, 7].

Dai [6], generalizing and systematizing earlier work of Rybko and Stolyar [15], 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 with soft (i.e., permitting lateness) customer deadlines and no 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 [5] showed that the fluid limits of the performance processes for a non-preemptive, strictly subcritical EDF network 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 [6], implies stability of the network under consideration.

It is natural to ask whether this stability result remains valid for preemptive EDF networks with soft deadlines. As it is observed in Bramson [5], 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. Kruk [11] showed how to overcome this difficulty under the assumption that the customer routes in the network are fixed. The main idea of the stability proof from [11] is that since 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 [5]. This is because 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 [9, 13, 17]. Once convergence to a FISFO fluid model is established, stability of the latter models proved in Bramson [5] and an argument similar to the proof of Theorem 4.2 in Dai [6] imply stability of preemptive EDF systems.

To our knowledge, the theorems presented in [11] were the first stability results for multiclass queueing networks with unbounded numbers of partially served customers and constituted the first application of the methodology of Dai [6] to such systems. Unfortunately, the arguments presented in [11] rely heavily on the assumption of fixed customer routes. Moreover, this assumption is not satisfied by a number of systems considered in the literature, for example by generalized Jackson networks.

The aim of this paper is to prove stability of general open, strictly subcritical preemptive multiclass EDF networks with soft deadlines and Markovian routing. The main idea of our argument is to divide customers into various types, according to the paths followed by them in the system. These types (paths) are counterparts of customer classes with fixed customer routes, considered in [11]. However, doing this does not immediately reduce the problem to the framework of [11], because, in general, the set of possible customer paths in the network is infinite. We deal with this problem by dividing this set of paths into two groups: one finite, but traversed with high probability, the other one infinite (cofinite), but very unlikely. Then, loosely speaking, we apply the methods developed in [11] (recalled above) to the first group and we show that the other one does not significantly alter the overall system performance. We hope that this proof strategy will turn out to be useful also for other queueing systems with infinite numbers of job types, in particular those, for which an initial assumption of fixed customer routes significantly simplifies the underlying analysis.

Along the way, we generalize Bramson's stability result from [5] for strictly subcritical, initially aging EDF fluid models satisfying an additional technical condition (see (5.1), to follow) to general (not necessarily initially aging) strictly subcritical EDF fluid models. It is noteworthy that Bramson conjectured the validity of such a generalization, see [5], pp. 88–89.

Our third contribution is to show that in a general preemptive multiclass EDF network, after a time large enough to process all the initial customers to completion, the maximal number of partially served customers in the system over a finite time horizon converges to zero in L^1 under fluid scaling. Although this fact is related to convergence of the fluid-scaled sample paths of the network performance processes to the FISFO fluid model solutions, it seems that none of these facts can be readily deduced from the other.

Together with [5] and [11], the results of this paper characterize asymptotic behavior of multiclass EDF networks with soft deadlines in the strictly subctricital case. It would be desirable to extend the analysis to the corresponding subcritical (in particular, critical) systems. Bramson ([5], p. 81) and Williams (private communication) posed a question whether the modular approach introduced by Bramson [4] and Williams [16] can be applied, at least in some situations, to subcritical EDF networks. The first step in this direction was made by Kruk [12], where the invariant manifold for the corresponding fluid models was characterized. Our paper also contributes to this project, because the issue of convergence of the fluid-scaled sample paths of the network performance processes to the FISFO fluid model solutions in the general preemptive EDF case is addressed here. (As in [5] and [11], this part of the analysis does not require the strict subcriticality assumption).

In spite of theoretical and practical importance of stochastic multiclass EDF queueing networks, there are still few mathematically rigorous results for such systems. Apart from the work recalled above, Yeung and Lehoczky [17] provided a diffusion approximation for measure-valued state descriptors of preemptive EDF feedforward networks. Their result has been generalized to the case of acyclic networks, with or without preemption, by Kruk, Lehoczky, Shreve, and Yeung [13]. 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. This amplifies the need for further research in this area.

The paper is organized as follows. Section 2 describes the model, provides background information on positive Harris recurrence of Markov processes and adjusts it to our setting. It also contains a formulation of Theorem 2.1, our main stability result. In Section 3, we present preemptive EDF queueing network equations, the corresponding FISFO fluid model equations and Theorem 3.1, assuring stability of an arbitrary strictly subcritical EDF fluid model. In Section 4 we formulate two important facts: Theorem 4.4, stating that the fluid limits of (properly shifted) performance processes describing a preemptive EDF network satisfy the FISFO fluid model equations, and Theorem 4.6, according to which the maximal number of partially served customers in the system over a finite time horizon converges to zero in L^1 under fluid scaling. We also derive Theorem 2.1 from Theorem 3.1 and the results presented in this section. Section 5 contains the proof of Theorem 3.1. In Section 6 we provide an auxiliary lemma, necessary for the last two sections. The proof of Theorem 4.4 is contained in Section 7. Finally, Section 8 contains proofs of Theorem 4.6 and an auxiliary state space collapse result from Section 4.

2. Terminology, background and the main result

2.1. Notation. The following notation will be used throughout the paper. Let $\mathbb{N} = \{0, 1, 2, \ldots\}$, let $\mathbb{Q}$, $\mathbb{R}$ denote the set of rational and real numbers, respectively. Let $\mathbb{R}_+ = [0, \infty)$, and let $\mathbb{R}^2_+ = (\mathbb{R}_+)^2$ be the nonnegative orthant. 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. For a vector

$a = (a_1, \ldots, a_n) \in \mathbb{R}^n$, let $|a| \triangleq \sum_{i=1}^n |a_i|$. All vectors in the paper are to be interpreted as column vectors. For a matrix A, A' denotes the transpose of A. For a finite set B, $|B|$ denotes the cardinality of B. The Borel σ-field on a topological space Y will be denoted by $\mathcal{B}(Y)$.

2.2. The model.

2.2.1. EDF networks. We consider a network consisting of J single server stations, indexed by $j = 1, \ldots, J$. The network is populated by K customer classes (or buffers), indexed by $k = 1, \ldots, K$. 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 $\alpha = (\alpha_1, \ldots, \alpha_K)$ and $\mathcal{E} = \{k \in \{1, \ldots, K\} : \alpha_k > 0\}$. 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 $m = (m_1, \ldots, m_K)$. Upon being served at j, a customer of class k immediately becomes a customer of class l with probability p_{kl}, independently of the network's 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 $\Theta = (q_{kl}) \triangleq (I - P')^{-1} = I + P' + (P')^2 + \ldots$ exists. We define the *total arrival rate* vector $\lambda = (\lambda_1, \ldots, \lambda_K) = \Theta\alpha$. Without loss of generality we assume that $\lambda_k > 0$ for each k. Next, we define the *traffic intensity* at station j as

$$\rho_j = \sum_{k \in \mathcal{C}(j)} m_k \lambda_k. \tag{2.1}$$

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

2.2.2. Stochastic primitives. We will now define the stochastic primitives for the model described in Section 2.2.1. 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 $k \in \mathcal{E}$,

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

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

and for some $n_k > 0$ and some nonnegative Borel function f_k with $\int_0^\infty f_k(x)dx > 0$, we have

$$\mathbb{P}(u_k(1) + \cdots + u_k(n_k) \in dx) \geq f_k(x)dx. \tag{2.4}$$

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 class k customers are a sequence of strictly positive, independent, and identically distributed random variables $v_k(i)$, $i = 1, 2, \dots$, where the index i denotes the order of arrival of customers to the buffer. We assume that for all k,

$$m_k \triangleq \mathbb{E}\, v_k(1) < \infty. \tag{2.5}$$

The *arrival rates* α_k, $k \in \{1, \dots, K\}$, are defined by $\alpha_k \triangleq 1/\mathbb{E} u_k(1)$ if $k \in \mathcal{E}$ and $\alpha_k \triangleq 0$ otherwise.

Customers entering the network through the buffer $k \in \mathcal{E}$ at times $U_k(i)$ have *initial lead times* $\ell_k(i)$, $i = 1, 2, \dots$, 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 $k \in \mathcal{E}$,

$$\mathbb{E}\, \ell_k(1) < \infty. \tag{2.6}$$

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

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

2.2.3. Lead times, service discipline. To determine whether customers meet their timing requirements, one must keep track of each customer's lead time, where

lead time = deadline - current time.

Customers are served at each station according to the preemptive EDF 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. 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. The (natural) assumption that $\ell_k(i) \geq 0$ was added only to simplify the exposition of the proofs. All our results are valid without this condition as long as $\ell_k(i)$ are integrable.

2.3. Markov process background. In EDF queueing systems, the individual customer lead times or some equivalent information must be kept to determine customer priorities. 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 [9, 13, 17].

Let $d = |\mathcal{E}|$ and let $S = (\mathbb{R}_+ \times \mathbb{R})^\infty$. Let

$$\Omega = \Big\{(q_k, k = 1, \dots, K, h_k, k = 1, \dots, K, r_k, k \in \mathcal{E}) \in \mathbb{N}^K \times S^K \times \mathbb{R}_+^d : \\ (h_k)_j = (0,0) \;\; \forall k = 1, \dots, K, \; j > q_k\}$$

be the *state space.* Under the product topology, Ω is a locally compact Polish space. The *state* of the process at any time is given by a point

$$x = (q_k, k = 1, \dots, K, h_k, k = 1, \dots, K, \; r_k, k \in \mathcal{E}) \in \Omega,$$

where for $k = 1, \dots, K$, q_k is the queue length at buffer k, h_k describes all customers present at buffer k 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_k are listed in the order of their arrivals to the buffer and ties are broken in an arbitrary manner. Let $\mathbf{0}$ denote the element of Ω describing the empty system, i.e., with $q_k = 0$, $h_k = ((0,0), (0,0), \dots)$ and $r_k = 0$ for all k. Let $q = (q_k)_{k=1,\dots,K}$ and $w = (w_k)_{k=1,\dots,K}$, where w_k is the sum of the residual service times of the customers listed in h_k. Let $r = (r_k)_{k\in\mathcal{E}}$ and let ℓ be the greatest lead time. For $x \in \Omega$, let $|x| = |q| + |w| + |r| + \ell^+$ be the "norm" of x.

The process describing the evolution of the EDF system is denoted by $X = (X(t), t \geq 0)$, where

$$\begin{aligned} X(t) &= (Q(t), H(t), R(t)) \\ &= (Q_k(t), k = 1, \dots, K, \; H_k(t), k = 1, \dots, K, \; R_k(t), k \in \mathcal{E}) \end{aligned}$$

is the state of the system at time t. By definition, the process X has right-continuous sample paths. It is easy to see that X is a Markov process. 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 [8]).

A Markov process X on the state space Ω is *Harris recurrent* if there exists a σ-finite measure ν 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 $\tau_A = \inf\{t \geq 0 : X(t) \in A\}$. It is known that Harris recurrence implies the existence of a unique (up to a multiplicative constant) invariant measure, see e.g., [10]. If this measure is finite, X is called *positive Harris recurrent.*

2.4. Main result. Recall that a queueing network is *stable* when the underlying Markov process is positive Harris recurrent. The following theorem is the main result of this paper.

Theorem 2.1. *All strictly subcritical EDF queueing networks with preemption which satisfy* (2.2)–(2.6) *are stable.*

3. Preemptive EDF network equations and fluid models

Let $E(t,s) = (E_k(t,s))_{k=1,\dots,K}$, $t \geq 0$, $s \in \mathbb{R}$, denote the *external arrival process* defined as follows. If $k \in \mathcal{E}$, then $E_k(t,s)$ is equal to the number of external arrivals by time t of type k customers with deadlines at time t less than or equal to s, otherwise $E_k(t,s) \equiv 0$. For $k = 1,\dots,K$, $t \geq 0$ and $s \in \mathbb{R}$, let $Z_k(t,s)$ denote the number of class k customers who are visiting station $j = s(k)$ at time t with deadlines at time t less than or equal to s. Let $Z(t,s) = (Z_k(t,s))_{k=1,\dots,K}$. Similarly, the vectors $A(t,s) = (A_k(t,s))_{k=1,\dots,K}$, $D(t,s) = (D_k(t,s))_{k=1,\dots,K}$, $T(t,s) = (T_k(t,s))_{k=1,\dots,K}$ denote the number of arrivals and departures, and the cumulative service time by time t corresponding to each class k of customers with deadlines at time t less than or equal to s. Let $Y_j(t,s)$, $j = 1,\dots,J$, denote the cumulative idleness by time t at station j with regard to service of customers with deadlines at time t less than or equal to s and let $Y(t,s) = (Y_j(t,s))_{j=1,\dots,J}$. For $k = 1,\dots,K$, $t,t' \geq 0$ and $s \in \mathbb{R}$, let $S_k(t',t,s)$ denote the number of service completions of class k customers having deadlines at time t less than or equal to s, by the time the station $j = s(k)$ has spent t' units of time serving these customers. Finally, for $k = 1,\dots,K$, $n \in \mathbb{N}$, $t \geq 0$ and $s \in \mathbb{R}$ let the *routing vector* $\Phi_k(n,t,s) = (\Phi_{k,1}(n,t,s),\dots,\Phi_{k,K}(n,t,s))$ be the number of the first n departures from class (buffer) k with deadlines at time t less than or equal to s that are routed to each class.

For $t \geq 0$ and $s \in \mathbb{R}$, let $\mathfrak{X}(t,s) = (A(t,s), D(t,s), T(t,s), Y(t,s), Z(t,s))$. Note that $Q(t) = (Q_k(t))_{k=1,\dots,K} = \lim_{s\to\infty} Z(t,s)$ is the queue length vector. Let $W(t) = (W_k(t))_{k=1,\dots,K}$ denote the unfinished work in the system, i.e., $W_k(t)$ is the sum of the residual service times of customers in buffer k at time t. We will sometimes use 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* (compare [5]):

$$A(t,s) = E(t,s) + \sum_{k=1}^{K} \Phi_k(D_k(t,s),t,s), \tag{3.1}$$

$$Z(t,s) = Z(0,s) + A(t,s) - D(t,s), \tag{3.2}$$

$$D_k(t,s) = S_k(T_k(t,s),t,s), \quad k = 1,\dots,K, \tag{3.3}$$

$$\sum_{k\in\mathcal{C}(j)} T_k(t,s) + Y_j(t,s) = t, \quad j = 1,\dots,J, \tag{3.4}$$

$$Y_j(t,s) \text{ can only increase in } t \text{ when } \sum_{k\in\mathcal{C}(j)} Z_k(t,s) = 0,\ j = 1,\dots,J, \tag{3.5}$$

valid for every for $t \geq 0$ and $s \in \mathbb{R}$. The equation (3.5) means that $Y_j(t_1,s) < Y_j(t_2,s)$ implies that $\sum_{k\in\mathcal{C}(j)} Z_k(t,s) = 0$ for some $t \in [t_1,t_2]$. The equations (3.1)–(3.4) are general properties of queueing networks and they do not depend on the service discipline under consideration. The equation (3.5) is specific to preemptive EDF networks. Indeed, for any s, the server idleness with regard to customers with deadlines not greater than s 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 (3.1)–(3.5) are the *FISFO fluid model equations* (see [5]):

$$\overline{A}(t,s) = \alpha(t\wedge s) + P'\overline{D}(t,s), \tag{3.6}$$

$$\overline{Z}(t,s) = \overline{Z}(0,s) + \overline{A}(t,s) - \overline{D}(t,s), \tag{3.7}$$

$$\overline{D}_k(t,s) = \overline{T}_k(t,s)/m_k, \quad k=1,\dots,K, \tag{3.8}$$

$$\sum_{k\in\mathcal{C}(j)} \overline{T}_k(t,s) + \overline{Y}_j(t,s) = t, \quad j=1,\dots,J, \tag{3.9}$$

$$\overline{Y}_j(t,s) \text{ can only increase in } t \text{ when } \sum_{k\in\mathcal{C}(j)} \overline{Z}_k(t,s) = 0,\ j=1,\dots,J, \tag{3.10}$$

where $t,s\geq 0$. In 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 $s\geq 0$. 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 $t\geq 0$ and that $\overline{Z}_k(t,s)\geq 0$, $k=1,\dots,K$. Let

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

Following [5], we additionally assume that

$$\overline{\mathfrak{X}}(t,s) = \overline{\mathfrak{X}}(t,t), \qquad 0\leq t\leq s. \tag{3.11}$$

We also define $\overline{Q}(t) = \lim_{s\to\infty}\overline{Z}(t,s) = \overline{Z}(t,t)$, where the last equation follows from (3.11).

As in the case of queueing networks, we say that a fluid model is *strictly subcritical* if $\rho_j<1$ for each j, where ρ_j is defined by (2.1). We also say that a FISFO fluid model is *stable* if there exists $c>0$ such that for all solutions of the equations (3.6)–(3.10), $\overline{Q}(t)=0$ for $t\geq c|\overline{Q}(0)|$.

The following result extends Theorem 2 of Bramson [5] to arbitrary strictly subcritical FISFO fluid models.

Theorem 3.1. *Any strictly subcritical FISFO fluid model is stable.*

4. Fluid limits and network stability

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

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

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} v_k(i)(\omega)=m_k, \qquad k=1,\dots,K, \tag{4.2}$$

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

By (2.2), (2.5), (2.6) and the strong law of large numbers, $\mathbb{P}(G)=1$.

We consider sequences of points $x_n=(q_n,h_n,r_n)$, $q_n\in\mathbb{N}^K$, $h_n\in S^K$, $r_n\in\mathbb{R}^d_+$, 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 $\overline{r}=(\overline{r}_1,\dots,\overline{r}_k)\in[0,1]^d$, $\overline{\ell}\in[0,1]$. By (4.1) and (4.4), on G

$$\frac{1}{|x_n|}N_k^{x_n}(|x_n|t)\to\alpha_k(t-\overline{r}_k)^+ \tag{4.5}$$

uniformly on compacts (u.o.c.) in t (see Lemma 4.2 in [6]).

Lemma 4.1 (Lemma 4.1 [11]). *Let $T_0>0$. Let a sequence x_n satisfy* (4.4) *and let*

$$\mathcal{L}_n=\max_{k\in\mathcal{E}}\ \max_{1\le i\le N_k^{x_n}(|x_n|T_0)}\ell_k(i). \tag{4.6}$$

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

Lemma 4.2. *Let x_n satisfy* (4.4) *and let $k\in\{1,\dots,K\}$. On the set G,*

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

u.o.c. in $t,s\ge 0$.

The proof of this lemma is the same as the proof of Lemma 5.1 in [11].

Let γ_k be the expected number of visits to all buffers in the network by a customer entering the network at the class $k\in\mathcal{E}$ and let $\gamma=\max_{k\in\mathcal{E}}\gamma_k$.

Lemma 4.3. *Let*

$$C=2\gamma|m|(1+|\alpha|)+4. \tag{4.8}$$

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_η such that for $\omega\in G_1$ and η sufficiently large (depending on ω),*

$$\bar{V}^{x_\eta}(\omega)\le C\,|x_\eta|, \tag{4.9}$$

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

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 ℓ_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. Using (4.5), we see that on the set G the number of these priority customers is bounded above by

$$|q_n| + |N^{x_n}(\ell_n^+)| \le |x_n| \left(1 + |\alpha| l_n^+/|x_n|\right) + o(|x_n|) \le |x_n| \left(1 + |\alpha|\right) + o(|x_n|).$$

Let $I_k^{x_n}$ be the set of indices $n \in \mathbb{N}$ corresponding to the class k service times $v_k(i)$ of the priority customers in the network with initial state x_n. Proceeding as in the proof of (A.7) of [5], we can show that there exist an integer-valued random variable N and a set $\tilde{G} \subseteq G$ with $\mathbb{P}(\tilde{G}) = 1$ such that for $\omega \in \tilde{G}$ and $n \ge N(\omega)$ we have

$$|I_k^{x_n}(\omega)| \le 2\gamma(1 + |\alpha|)|x_n| + |x_n|/|m|. \tag{4.10}$$

Under the EDF service discipline, the index i of the arrival of a customer of class k is independent of $v_k(i)$. Thus, by (2.5), (4.4) and the weak law of large numbers, for $k = 1, \dots, K$ we have

$$\frac{1}{|x_n|} \left| \sum_{i \in I_k^{x_n}} v_k(i) - m_k |I_k^{x_n}| \right| \xrightarrow{P} 0, \quad n \to \infty.$$

By Theorem 20.5 in [1], there exist a set $G_1 \subseteq \tilde{G}$ with $\mathbb{P}(G_1) = 1$ and a subsequence η such that for every $\omega \in G_1$ and $k = 1, \dots, K$, we have

$$\frac{1}{|x_\eta|} \left| \sum_{i \in I_k^{x_\eta}(\omega)} v_k(i)(\omega) - m_k |I_k^{x_\eta}(\omega)| \right| \to 0, \quad \eta \to \infty.$$

Therefore, the sum of the service times of the priority customers in the network with initial state x_η is bounded above on the set G_1 by

$$\begin{aligned} V^{x_\eta} = |w_\eta| + \sum_{k=1}^{K} \sum_{i \in I_k^{x_\eta}} v_k(i) &\le |x_\eta| + \sum_{k=1}^{K} m_k |I_k^{x_\eta}| + o(|x_\eta|) \\ &\le |x_\eta|(2 + 2\gamma|m|(1 + |\alpha|)) + o(|x_\eta|), \end{aligned}$$

where the second inequality follows from (4.10). This, together with (4.8), implies that for every $\omega \in G_1$ there exists $\eta_0 = \eta_0(\omega)$ such that

$$V^{x_\eta}(\omega) \le (C - 1)|x_\eta|, \qquad \eta \ge \eta_0(\omega). \tag{4.11}$$

Note that because all the priority customers arrive at the preemptive EDF system with initial state x_η by time ℓ_η^+, $V^{x_\eta} + \ell_\eta^+$ 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 (4.11), for $\omega \in G_1$ and $\eta \geq \eta_0(\omega)$, (4.9) holds. □

For $t_0 \geq 0$, we introduce the *time shift operator* Δ_{t_0} acting on the coordinates of the process $\mathfrak{X}$ as follows: for $t \geq 0$, $s \in \mathbb{R}$, we have

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

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 $t \geq 0$. 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 theorem plays a crucial role in the proof of Theorem 2.1. 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.

Theorem 4.4. *Let C be as in* (4.8). *For every sequence x_n in* (4.4), *there exist a set $G' \subseteq G$ with $\mathbb{P}(G') = 1$ and a subsequence x_ξ such that for each $\omega \in G'$ and each subsequence x_ϑ of x_ξ (possibly depending on ω), there exists a further subsequence x_ζ of x_ϑ (depending on ω) 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*

$$\lim_{n\to\infty} \Delta_{C|x_\zeta|}\mathfrak{X}^{x_\zeta}(t|x_\zeta|, s|x_\zeta|)(\omega)/|x_\zeta| \tag{4.12}$$

satisfies the FISFO fluid model equations (3.6)–(3.10), *together with the condition* (3.11).

The proof of Theorem 4.4 will be given in Section 7.

To show Theorem 2.1, we need the following proposition, which will be proved in Section 8.

Proposition 4.5 (State space collapse). *Let x_n be a sequence satisfying* (4.4). *Let C be given by* (4.8) *and let G' be as in the proof of Theorem 4.4. Let $\omega \in G'$, and let x_ζ be a subsequence (depending on ω) constructed in the proof of Theorem 4.4. Then for each $k = 1, \ldots, K$ and $t \geq 0$,*

$$\lim_{\zeta\to\infty} \frac{1}{|x_\zeta|}\left|W_k^{x_\zeta}((t+C)|x_\zeta|) - m_k Q_k^{x_\zeta}((t+C)|x_\zeta|)\right| = 0. \tag{4.13}$$

Using the above results, we can prove Theorem 2.1 by repeating, with minor changes, the proof of Theorem 3.1 in [11]. In particular, our Theorem 4.4 and Proposition 4.5 should be quoted instead of Propositions 5.2, 6.1 of

[11], respectively, and the process $N_k^{\mathbf{0}}$ from the proof of Theorem 3.1 in [11] should be replaced by the process $N_{0,k}^{\mathbf{0}}$ defined by (6.1), to follow.

For $t \geq 0$, let $\mathcal{P}(t)$ denote the number of *partially served customers* in the system at time t, i.e., those who have received some service in the time interval $[0, t]$, but they have not been fully served by time t.

Theorem 4.6. *Let C be given by* (4.8) *and let x_n be a sequence satisfying* (4.4)*. Then for every $T_0 > C$, we have*

$$\lim_{n\to\infty} \frac{1}{|x_n|}\mathbb{E}\left[\max_{C\leq t\leq T_0} \mathcal{P}^{x_n}(t|x_n|)\right] = 0. \tag{4.14}$$

Moreover, if $\overline{\ell} = 0$ in (4.4)*, then the constant C in* (4.14) *can be replaced by* 0*.*

The proof of this result will be given in Section 8.

Theorem 4.6 is closely related to Proposition 4.5 and to the assertion in Theorem 4.4 that the fluid limits (4.12) satisfy (3.8). This relation stands behind similarity of the proofs of these facts. However, it seems that none of them can be immediately deduced from another. For example, a longer service time increases the probability of the customer being preempted, so if $\mathcal{P}(t)$ is relatively small, it does not directly imply that the time devoted by the server to class k customers is roughly proportional to their mean service time multiplied by the number of departed customers from this class. Conversely, the latter relation does not rule out the possibility of $\mathcal{P}(t)$ being nonnegligible, since there may be a lot of customers preempted just after entering into service, before the server spends a lot of time working on them.

5. Proof of Theorem 3.1

In this section we prove Theorem 3.1. We will first introduce some additional notation and terminology. We introduce the set of *multi-indices*

$$\mathbf{K} = \big\{(k_1, \ldots, k_n) : n \geq 1,\ k_1, \ldots, k_n \in \{1, \ldots, K\},\ \alpha_{k_1} p_{k_1 k_2} \cdots p_{k_{n-1} k_n} > 0\big\},$$

where $p_{k_1 k_2} \cdots p_{k_{n-1} k_n}$ should be interpreted as 1 for $n = 1$. The elements of $\mathbf{K}$ represent paths of finite length which are being followed with positive probability by customers since their arrival to the network. For $\mathbf{k} = (k_1, \ldots, k_n) \in \mathbf{K}$, let $|\mathbf{k}| = n$ be the length of the path $\mathbf{k}$, let $p_{\mathbf{k}} = p_{k_1 k_2} \cdots p_{k_{n-1} k_n}$, $\alpha_{\mathbf{k}} = \alpha_{k_1} p_{\mathbf{k}}$, $m_{\mathbf{k}} = m_{k_n}$ and 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 $\mathbf{k} \in \mathbf{K}$ and $k \in \{1, \ldots, K\}$, we write $\mathbf{k} \in \tilde{\mathcal{C}}(k)$ if $e(\mathbf{k}) = k$. We will refer to customers following the path $\mathbf{k} \in \mathbf{K}$ as *type* $\mathbf{k}$ *customers.*

Definition 5.1 ([5], p. 88). *A queueing network (resp. its fluid model) is initially aging if the set $\{1, \ldots, K\}$ of its customer classes can be divided into two disjoint subsets $\mathcal{K}_1$ and $\mathcal{K}_2$ such that*

(a) the classes from $\mathcal{K}_1$ are not accessible from classes in $\mathcal{K}_2$,

(b) for any given $k \in \mathcal{K}_1$, all $\mathbf{k} \in \mathbf{K}$ with $e(\mathbf{k}) = k$ have the same length and this common length is not greater than $|\mathcal{K}_1|$.

The following result will be our starting point.

Theorem 5.2 (Theorem 2 [5]). *Assume that an initially aging FISFO fluid model is strictly subcritical and satisfies*

$$\sum_{k \in \mathcal{K}_2} m_k \lambda_k \le \frac{1}{4}. \tag{5.1}$$

Then, it is stable.

We will show below that Theorem 5.2 actually implies stability of *any* strictly subcritical FISFO fluid model.

Proof of Theorem 3.1. Let

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

be an arbitrary strictly subcritical FISFO fluid model. We will now construct an initially aging strictly subcritical FISFO fluid model

$$\tilde{\mathfrak{X}}(t,s) = (\tilde{A}(t,s), \tilde{D}(t,s), \tilde{T}(t,s), \tilde{Y}(t,s), \tilde{Z}(t,s)), \qquad t,s \ge 0,$$

satisfying (5.1) and such that $\overline{\mathfrak{X}}$ is a projection (in a suitable sense) of $\tilde{\mathfrak{X}}$. Let $H \in \mathbb{N}$. The customer classes in $\tilde{\mathfrak{X}}$ will be labeled by ordered pairs $\tilde{k} = (k,h)$ where $k = 1,\dots,K$ and $h = 0,\dots,H$. The corresponding arrival and service rates are defined by $\alpha_{(k,0)} = \alpha_k$, $\alpha_{(k,h)} = 0$, $h > 0$, $m_{(k,h)} = m_k$, and the nonzero entries in the corresponding routing matrix $\tilde{P}$ are given by $p_{(k,h),(l,h+1)} = p_{kl}$, $h < H$, $p_{(k,H),(l,H)} = p_{kl}$, $k,l = 1,\dots,K$. The set of stations $1,\dots,J$ in $\tilde{\mathfrak{X}}$ is the same as in the original fluid model $\overline{\mathfrak{X}}$ and $(k,h) \in \mathcal{C}(j)$ in $\tilde{\mathfrak{X}}$ iff $k \in \mathcal{C}(j)$ in $\overline{\mathfrak{X}}$. It is easy to see that this defines the routing structure of an initially aging fluid model with $\mathcal{K}_1 = \{(k,h) : k = 1,\dots,K, h < H\}$ and $\mathcal{K}_2 = \{(k,H) : k = 1,\dots,K\}$. Also, for H large enough and $(\lambda_{\tilde{k}}) = (I - \tilde{P}')^{-1}(\alpha_{\tilde{k}})$, we have

$$\sum_{\tilde{k} \in \mathcal{K}_2} m_{\tilde{k}} \lambda_{\tilde{k}} = \sum_{k=1}^{K} m_k \lambda_{(k,H)} = \sum_{k=1}^{K} m_k \left((P')^H \Theta \alpha \right)_k \le \frac{1}{4},$$

since $(P')^H \to 0$ as $H \to \infty$.

Let w_h, $h = 0,\dots,H$, be fixed nonnegative numbers (weights) such that $\sum_{h=0}^{H} w_h = 1$. For $t,s \ge 0$, $k = 1,\dots,K$ and $h = 0,\dots,H$, let $\tilde{D}_{(k,h)}(t,s) = w_h \overline{D}_k(t,s)$, $\tilde{Z}_{(k,h)}(0,s) = w_h \overline{Z}_k(0,s)$, $\tilde{A}_{(k,0)}(t,s) = \alpha_k(t \wedge s)$ and let

$$\tilde{A}_{(k,h)}(t,s) = \sum_{l=1}^{K} p_{(l,h-1),(k,h)} \tilde{D}_{(l,h-1)}(t,s) = \sum_{l=1}^{K} p_{lk} \tilde{D}_{(l,h-1)}(t,s),\ 0<h<H,$$

$$
\begin{aligned}
\tilde{A}_{(k,H)}(t,s) &= \sum_{l=1}^{K} p_{(l,H-1),(k,H)} \tilde{D}_{(l,H-1)}(t,s) + \sum_{l=1}^{K} p_{(l,H),(k,H)} \tilde{D}_{(l,H)}(t,s) \\
&= \sum_{l=1}^{K} p_{lk} \left(\tilde{D}_{(l,H-1)}(t,s) + \tilde{D}_{(l,H)}(t,s) \right).
\end{aligned}
$$

It is easy to check that the pair $\tilde{A}$, $\tilde{D}$ satisfies (3.6) (with $\tilde{P}$ instead of P). Moreover, for any k,

$$
\begin{aligned}
\sum_{h=0}^{H} \tilde{A}_{(k,h)}(t,s) &= \alpha_k(t \wedge s) + \sum_{l=1}^{K} p_{lk} \sum_{h=0}^{H} \tilde{D}_{(l,h)}(t,s) \\
&= \alpha_k(t \wedge s) + \sum_{l=1}^{K} p_{lk} \overline{D}_l(t,s) = \overline{A}_k(t,s),
\end{aligned}
\tag{5.2}
$$

where the third equality follows from (3.6). Define $\tilde{Z}$ by the right-hand side of (3.7) (with $\overline{A}$, $\overline{D}$, $\overline{Z}(0,\cdot)$ replaced by $\tilde{A}$, $\tilde{D}$ and $\tilde{Z}(0,\cdot)$, respectively). By (5.2) and (3.7),

$$
\begin{aligned}
\sum_{h=0}^{H} \tilde{Z}_{(k,h)}(t,s) &= \sum_{h=0}^{H} \tilde{Z}_{(k,h)}(0,s) + \sum_{h=0}^{H} \tilde{A}_{(k,h)}(t,s) - \sum_{h=0}^{H} \tilde{D}_{(k,h)}(t,s) \\
&= \overline{Z}_k(0,s) + \overline{A}_k(t,s) - \overline{D}_k(t,s) = \overline{Z}_k(t,s).
\end{aligned}
\tag{5.3}
$$

Finally, for $t, s \geq 0$, define $\tilde{T}_{(k,h)}(t,s) = m_{(k,h)} \tilde{D}_{(k,h)}(t,s)$, $k = 1, \ldots, K$, $h = 0, \ldots, H$, and $\tilde{Y}_j(t,s) = \overline{Y}_j(t,s)$, $j = 1, \ldots, J$. By definition, the pair $\tilde{D}$, $\tilde{T}$ satisfies a suitable counterpart of (3.8). Moreover, since (3.8) implies

$$
\begin{aligned}
\sum_{(k,h) \in \mathcal{C}(j)} \tilde{T}_{(k,h)}(t,s) &= \sum_{k \in \mathcal{C}(j)} m_k \sum_{h=0}^{H} \tilde{D}_{(k,h)}(t,s) = \sum_{k \in \mathcal{C}(j)} m_k \overline{D}_k(t,s) \\
&= \sum_{k \in \mathcal{C}(j)} \overline{T}_k(t,s),
\end{aligned}
$$

it is easy to see that $\tilde{\mathfrak{X}}$ satisfies a suitable counterpart of (3.9). Similarly, (5.3) implies that $\tilde{\mathfrak{X}}$ satisfies a suitable counterpart of (3.10). To summarize, $\tilde{\mathfrak{X}}$ satisfies all the FISFO fluid model equations. The remaining conditions for a fluid model (positivity, monotonicity, etc.) for $\tilde{\mathfrak{X}}$ follow readily from those for $\overline{\mathfrak{X}}$. Hence, $\tilde{\mathfrak{X}}$ is an initially aging strictly subcritical FISFO fluid model. By Theorem 5.2, there exists a finite constant $c > 0$ such that $\tilde{Q}(t) = 0$ for $t \geq c|\tilde{Q}(0)|$, where $\tilde{Q}(t) = \lim_{s \to \infty} \tilde{Z}(t,s)$. However, (5.3), implies that $|\tilde{Q}(t)| = |Q(t)|$ for all $t \geq 0$, so $Q(t) = 0$ for $t \geq c|Q(0)|$ and the fluid model $\overline{\mathfrak{X}}$ is also stable. □

Remark. The set of customer classes and the routing structure for the fluid model $\tilde{\mathfrak{X}}$ defined above are the same as for a network $\mathcal{N}'$ used in [5], pp. 87-88, to reduce the question of stability of arbitrary strictly subcritical EDF *queueing networks* without preemption to stability of those of them which are initially aging and satisfy (5.1). Our approach, although clearly related to the one in [5], is different, since we use a similar method to obtain stability of *fluid models* rather than queueing networks. There are, of course, other possible extensions of the set of customer classes of $\mathfrak{X}$ and its routing structure which can be used for a construction of $\tilde{\mathfrak{X}}$. One such choice is $\mathcal{K}_1 = \{\mathbf{k} \in \mathbf{K} : |\mathbf{k}| < H\}$ and $\mathcal{K}_2 = \{\mathbf{k} \in \mathbf{K} : |\mathbf{k}| = H\}$ with nonzero transition probabilities $p_{(k_1,\ldots,k_h),(k_1,\ldots,k_{h+1})} = p_{k_h,k_{h+1}}$ for $h < H$ and $p_{(k_1,\ldots,k_H),(k_1,\ldots,k_{H-1},l)} = p_{k_H,l}$.

6. An auxiliary lemma

This section contains a technical Lemma 6.1, necessary for the proofs of Theorem 4.4 and 4.6. The following additional notation and terminology will be used in the sequel.

Recall the set of multi-indices $\mathbf{K}$ defined in Section 5. For $m \in \mathbb{N}$ and $k, l = 1, \ldots, K$, let $\mathbf{K}_m = \{\mathbf{k} \in \mathbf{K} : |\mathbf{k}| > m\}$, $\mathbf{K}_{m,l} = \mathbf{K}_m \cap \tilde{\mathcal{C}}(l)$ and let $\mathbf{K}_{m,k,l} = \{\mathbf{k} \in \mathbf{K}_{m,l} : b(\mathbf{k}) = k\}$. For $x \in \Omega$, $t \geq 0$ and $\mathbf{k} \in \mathbf{K}$, let $N^x_{\mathbf{k}}(t)$ be the number of type $\mathbf{k}$ customers who have arrived at the network with initial state x by time t. Also, for $m \in \mathbb{N}$ and $k, l = 1, \ldots, K$, let

$$N^x_{m,k,l}(t) = \sum_{\mathbf{k} \in \mathbf{K}_{m,k,l}} N^x_{\mathbf{k}}(t),$$

$$N^x_{m,l}(t) = \sum_{\mathbf{k} \in \mathbf{K}_{m,l}} N^x_{\mathbf{k}}(t) = \sum_{k=1}^{K} N^x_{m,k,l}(t),$$

$$\begin{aligned} \alpha_{m,k,l} &= \sum_{\mathbf{k} \in \mathbf{K}_{m,k,l}} \alpha_{\mathbf{k}} = \alpha_k \left\{P^{m+1} + P^{m+2} + \ldots\right\}_{k,l} \\ &= \left\{(P')^{m+1} + (P')^{m+2} + \ldots\right\}_{l,k} \alpha_k = \left\{(P')^{m+1}\Theta\right\}_{l,k} \alpha_k, \\ \alpha_{m,l} &= \sum_{\mathbf{k} \in \mathbf{K}_{m,l}} \alpha_{\mathbf{k}} = \sum_{k=1}^{K} \alpha_{m,k,l} = \left\{(P')^{m+1}\Theta\alpha\right\}_l = \left\{(P')^{m+1}\lambda\right\}_l . \end{aligned} \tag{6.1}$$

Note that $N^x_{m,l}(t)$ is, in general, *not* equal to the number of customers arriving at the network up to time t which will eventually visit class l by a path longer than m. Indeed, every such arriving customer increases the count in $N^x_{m,l}(t)$ by the number of his visits to class l in more than m steps before his departure.

In analogy with the performance processes D_k, T_k, defined for $k = 1, \ldots, K$, for $\mathbf{k} \in \mathbf{K}$, $t \geq 0$ and $s \in \mathbb{R}$, let $D_{\mathbf{k}}(t, s)$ denote the number of departures from class $k = e(\mathbf{k})$ of customers with deadlines at time t less than or equal to s which have arrived at class k following the path $\mathbf{k}$, and let $T_{\mathbf{k}}(t,s)$ denote the cumulative service time at station $s(e(\mathbf{k}))$ by time t corresponding to such customers. Similarly, in analogy with the performance processes Q_k, W_k, defined for $k=1,\dots,K$, for $\mathbf{k}=(k_1,\dots,k_n)\in\mathbf{K}$ and $t\geq 0$, let $Q_{\mathbf{k}}(t)$ denote the number of customers present in class k_n at time t who have arrived at this class following the path $\mathbf{k}$ and let $W_{\mathbf{k}}(t)$ denote the workload (i.e., the sum of the residual service times) for station $s(k_n)$ corresponding to these customers. By service times of type $\mathbf{k}$ customers we shall mean the service times of class $k=e(\mathbf{k})$ customers who have arrived at class k following the path $\mathbf{k}$.

For $k=1,\dots,K$ and $t_1,t_2\in\mathbb{R}$, let $B^n_k(t_1,t_2)$ denote the set of $j=1,2,\dots,$ for which the customer corresponding to the service time $v_k(j)$ has entered the network with initial state x_n in the time interval $(t_1|x_n|,t_2|x_n|]\cap(0,\infty)$. In particular, $B^n_k(t_1,t_2)=\emptyset$ if $t_1\geq t_2$. Similarly, for $\mathbf{k}=(k_1,\dots,k_l)\in\mathbf{K}$ and $t_1,t_2\geq 0$, let $B^n_{\mathbf{k}}(t_1,t_2)$ denote the set of $j=1,2,\dots,$ corresponding to the class k_l service times $v_{k_l}(j)$ of those type $\mathbf{k}$ customers on the l-th step of their routes who have entered the network with initial state x_n in the time interval $(t_1|x_n|,t_2|x_n|]$. Finally, for $k=1,\dots,K$, $m\in\mathbb{N}$ and $t\geq 0$, let $B^n_{m,k}(t)=\bigcup_{\mathbf{k}\in\mathbf{K}_{m,k}}B^n_{\mathbf{k}}(0,t)$.

Lemma 6.1. *Let a sequence x_n satisfy* (4.4) *and let the set G_1 and the subsequence x_η be as in Lemma 4.3. There exists a set $G'\subseteq G_1$ with $\mathbb{P}(G')=1$ and a subsequence x_ξ of x_η such that on G' for every $\mathbf{k}\in\mathbf{K}$, $m\in\mathbb{N}$ and $l=1,\dots,K$, we have*

$$\frac{1}{|x_n|}N^{x_n}_{\mathbf{k}}(|x_n|t)\to\alpha_{\mathbf{k}}(t-\overline{r}_{b(\mathbf{k})})^+,\tag{6.2}$$

$$\frac{1}{|x_n|}N^{x_n}_{m,l}(|x_n|t)\to\tilde{\alpha}_{m,l}(t)\triangleq\sum_{k=1}^{K}\alpha_{m,k,l}(t-\overline{r}_k)^+,\tag{6.3}$$

u.o.c. in t as $n\to\infty$ and for every $r_0>0$, $k=1,\dots,K$ and $\mathbf{k}\in\mathbf{K}$, as $\xi\to\infty$,

$$\sup_{0\leq t_1<t_2\leq r_0}\left|\frac{1}{|x_\xi|}\sum_{i\in B^\xi_k(t_1,t_2)}v_k(i)-m_k(\tilde{\alpha}_{0,k}(t_2)-\tilde{\alpha}_{0,k}(t_1))\right|\to 0,\tag{6.4}$$

$$\begin{aligned}\sup_{0\leq t_1<t_2\leq r_0}\Bigg|\frac{1}{|x_\xi|}&\sum_{i\in B^\xi_{\mathbf{k}}(t_1,t_2)}v_{e(\mathbf{k})}(i)\\&-\alpha_{\mathbf{k}}m_{\mathbf{k}}((t_2-\overline{r}_{b(\mathbf{k})})^+-(t_1-\overline{r}_{b(\mathbf{k})})^+)\Bigg|\to 0,\end{aligned}\tag{6.5}$$

$$\sup_{0\le t\le r_0}\left|\frac{1}{|x_\xi|}\sum_{i\in B^\xi_{m,k}(t)} v_k(i)-m_k|B^\xi_{m,k}(t)|\right|\to 0. \tag{6.6}$$

Moreover, on G' for every $k,k'=1,\ldots,K$ and every t,s of the form $t=C+t'$, $s=C+s'$, where C was defined by (4.8), $t'\ge 0$, $t',s'\in\mathbb{Q}$, we have, as $\xi\to\infty$,

$$\frac{1}{|x_\xi|}\left|\Phi^{x_\xi}_{k,k'}(D^{x_\xi}_k(t|x_\xi|,s|x_\xi|),t|x_\xi|,s|x_\xi|)-p_{kk'}D^{x_\xi}_k(t|x_\xi|,s|x_\xi|)\right|\to 0. \tag{6.7}$$

Proof. Proceeding as in the proof of (4.5) we can show that for every $\mathbf{k}\in\mathbf{K}$ there exists a set $G_{\mathbf{k}}$ with $\mathbb{P}(G_{\mathbf{k}})=1$ on which (6.2) holds u.o.c. in t. A similar reasoning, using strong laws of large numbers for partial sums and arrival processes, shows that for the sequence x_n, $m\in\mathbb{N}$ and $k,l=1,\ldots,K$, there exists a set $G_{m,k,l}$ with $\mathbb{P}(G_{m,k,l})=1$ on which u.o.c. in t,

$$\frac{1}{|x_n|}N^{x_n}_{m,k,l}(|x_n|t)\to\alpha_{m,k,l}(t-\overline{r}_k)^+.$$

Put $G_{m,l}=\bigcap_{k=1}^K G_{m,k,l}$. Then $\mathbb{P}(G_{m,l})=1$ and (6.3) holds on $G_{m,l}$ u.o.c. in t.

By the weak law of large numbers, together with the independence of the service times on the interarrival times, the initial lead times and the routing,

$$\frac{1}{|x_\eta|}\left|\sum_{i\in B^\eta_k(t_1,t_2)} v_k(i)-m_k|B^\eta_k(t_1,t_2)|\right|\xrightarrow{P}0. \tag{6.8}$$

However, (6.3) implies that on the set $G_{0,k}$, for $0\le t_1<t_2$

$$|B^\eta_k(t_1,t_2)|=N^{x_\eta}_{0,k}(t_2|x_\eta|)-N^{x_\eta}_{0,k}(t_1|x_\eta|)=|x_\eta|(\tilde\alpha_{0,k}(t_2)-\tilde\alpha_{0,k}(t_1))+o(|x_\eta|),$$

and hence (6.8) yields

$$\left|\frac{1}{|x_\eta|}\sum_{i\in B^\eta_k(t_1,t_2)} v_k(i)-m_k(\tilde\alpha_{0,k}(t_2)-\tilde\alpha_{0,k}(t_1))\right|\xrightarrow{P}0. \tag{6.9}$$

Using (6.9) and arguing as in the proof of (A.1) in [5] or in the proof of Proposition 3.4 in [9], we get, for every $r_0>0$,

$$\sup_{0\le t_1<t_2\le r_0}\left|\frac{1}{|x_\eta|}\sum_{i\in B^\eta_k(t_1,t_2)} v_k(i)-m_k(\tilde\alpha_{0,k}(t_2)-\tilde\alpha_{0,k}(t_1))\right|\xrightarrow{P}0. \tag{6.10}$$

By Theorem 20.5 in [1], there exist a set G_2 with $\mathbb{P}(G_2)=1$ and a subsequence ξ of the sequence η such that on G_2, we have pointwise convergence (6.4) for every $r_0>0$ and $k=1,\ldots,K$.

Arguing as in the proof of (6.10), but using (6.2) instead of (6.3), we can check that for every $r_0 > 0$,

$$\begin{aligned}\sup_{0\le t_1<t_2\le r_0}\Bigg|\frac{1}{|x_\xi|}\sum_{i\in B^\xi_{\mathbf{k}}(t_1,t_2)} v_{e(\mathbf{k})}(i)\qquad\qquad\\ - \alpha_{\mathbf{k}} m_{\mathbf{k}}((t_2-\overline{r}_{b(\mathbf{k})})^+ - (t_1-\overline{r}_{b(\mathbf{k})})^+)\Bigg| \xrightarrow{P} 0.\end{aligned} \tag{6.11}$$

As in (6.4), we want to refine (6.11) to almost sure convergence for each $\mathbf{k}\in\mathbf{K}$. To this end, we use the fact that $\mathbf{K}$ is countable and enumerate its elements, getting $\mathbf{K}=\{\mathbf{k}_1,\mathbf{k}_2,\dots\}$. By Theorem 20.5 in [1], there exist a set $\tilde G_1$ with $\mathbb{P}(\tilde G_1)=1$ and a subsequence ξ^1 of ξ such that on $\tilde G_1$, we have pointwise convergence

$$\begin{aligned}\sup_{0\le t_1<t_2\le r_0}\Bigg|\frac{1}{|x_{\xi^j}|}\sum_{i\in B^{\xi^j}_{\mathbf{k}}(t_1,t_2)} v_{e(\mathbf{k})}(i)\qquad\qquad\\ - \alpha_{\mathbf{k}} m_{\mathbf{k}}((t_2-\overline{r}_{b(\mathbf{k})})^+ - (t_1-\overline{r}_{b(\mathbf{k})})^+)\Bigg| \to 0\end{aligned} \tag{6.12}$$

for $j=1$, $\mathbf{k}=\mathbf{k}_1$ and every $r_0>0$. Using Theorem 20.5 in [1] again, we get a set $\tilde G_2\subset\tilde G_1$ with $\mathbb{P}(\tilde G_2)=1$ and a subsequence ξ^2 of ξ^1 such that on $\tilde G_2$, we have pointwise convergence (6.12) for all $r_0>0$ and $j=2$, $\mathbf{k}=\mathbf{k}_1,\mathbf{k}_2$. Proceeding in this way, for every $n\ge 2$ we construct a subsequence ξ^n of ξ^{n-1} and a set $\tilde G_n$ with $\mathbb{P}(\tilde G_n)=1$ such that on $\tilde G_n$ (6.12) holds for all $r_0>0$, $j=n$ and $\mathbf{k}=\mathbf{k}_1,\dots,\mathbf{k}_n$. Using the Cantor diagonal procedure, we extract a subsequence (still denoted by ξ for convenience) of each sequence ξ^n along which (6.5) holds on the set $G_3=\bigcap_{n=1}^\infty \tilde G_n$ for all $r_0>0$ and *every* $\mathbf{k}\in\mathbf{K}$.

An argument similar to the one presented above shows that there exist a further subsequence (still denoted by ξ) and a set G_4 with $\mathbb{P}(G_4)=1$ on which for every $k=1,\dots,K$, $m\in\mathbb{N}$ and $r_0>0$ we have (6.6).

Proceeding as in the proof of (A7) [5], we can show that for each $t\ge 0$,

$$\mathbb{P}\left[|A^{x_\xi}(t|x_\xi|,\infty)|\le 4\gamma|\alpha|t|x_\xi| \quad \text{for } \xi \text{ large enough}\right]=1, \tag{6.13}$$

where $A^{x_\xi}(t|x_\xi|,\infty)=\lim_{s\to\infty}A^{x_\xi}(t|x_\xi|,s)$. The equation (3.2) implies that for every $x\in\Omega$,

$$\begin{aligned}|D^x(t,s)| &= |A^x(t,s)+Z^x(0,s)-Z^x(t,s)| \le |A^x(t,s)+Z^x(0,s)|\\ &\le |A^x(t,\infty)|+|x|.\end{aligned}$$

This, together with (6.13) implies that for each fixed $t \geq 0$ and $s \in \mathbb{R}$,

$$\mathbb{P}\left[|D^{x_\xi}(t|x_\xi|, s|x_\xi|)| \leq (4\gamma|\alpha|t+1)|x_\xi| \quad \text{for } \xi \text{ large enough}\right] = 1.$$

This, in turn, together with the independence of the i.i.d. routing vectors on other network primitives and the weak law of large numbers, yields

$$\frac{1}{|x_\xi|}\left|\Phi^{x_\xi}_{k,k'}(D^{x_\xi}_k(t|x_\xi|, s|x_\xi|), t|x_\xi|, s|x_\xi|) - p_{kk'}D^{x_\xi}_k(t|x_\xi|, s|x_\xi|)\right| \xrightarrow{P} 0$$

for each $k, k' \in \{1, \ldots, K\}$ and fixed $t \geq 0$, $s \in \mathbb{R}$. Using this fact and proceeding as in the next to last paragraph, we can extract from the sequence ξ a further subsequence (still denoted by ξ) and a set G_5 with $\mathbb{P}(G_5) = 1$ on which for every $k, k' = 1, \ldots, K$ and every t, s of the form $t = C + t'$, $s = C + s'$, where C was defined by (4.8), $t' \geq 0$, $t', s' \in \mathbb{Q}$, (6.7) holds.

Let $G' = G_1 \cap G_2 \cap G_3 \cap G_4 \cap G_5 \cap \bigcap_{\mathbf{k}\in\mathbf{K}} G_{\mathbf{k}} \cap \bigcap_{m=1}^{\infty}\bigcap_{k=1}^{K} G_{m,k}$. We have $\mathbb{P}(G') = 1$ and the set G', together with the sequence x_ξ, satisfy (6.2)–(6.7). □

7. Proof of Theorem 4.4

Let a sequence x_n satisfy (4.4) and let the set G_1 and the subsequence x_η be as in Lemma 4.3. Let the set G' and the subsequence x_ξ be as in Lemma 6.1. Fix $\omega \in G'$. Consider an arbitrary subsequence ϑ of the sequence ξ. For $t \geq 0$ and $s \in \mathbb{R}$, let

$$\begin{aligned}
\overline{\mathfrak{X}}^{(\vartheta)}(t,s) &= (\overline{A}^{(\vartheta)}(t,s), \overline{D}^{(\vartheta)}(t,s), \overline{T}^{(\vartheta)}(t,s), \overline{Y}^{(\vartheta)}(t,s), \overline{Z}^{(\vartheta)}(t,s)) \\
&= \Delta_{C|x_\vartheta|}\mathfrak{X}^{x_\vartheta}(t|x_\vartheta|, s|x_\vartheta|)/|x_\vartheta|, \\
\overline{\mathfrak{X}}^{(\vartheta)}_0(t,s) &= (\overline{A}^{(\vartheta)}(t,s), \overline{D}^{(\vartheta)}(t,s), \overline{T}^{(\vartheta)}(t,s), \overline{Y}^{(\vartheta)}(t,s), \overline{Z}^{(\vartheta)}(0,s)).
\end{aligned}$$

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 selection theorem (see, e.g., [1], Theorem 25.9 and the remark in the proof of Theorem 29.3), there exists a subsequence ζ and a right-continuous function

$$\overline{\mathfrak{X}}_0(t,s) = (\overline{A}(t,s), \overline{D}(t,s), \overline{T}(t,s), \overline{Y}(t,s), \overline{Z}(0,s)), \qquad t \geq 0,\ s \in \mathbb{R},$$

(both depending on ω) such that each coordinate map of $\overline{\mathfrak{X}}^{(\zeta)}_0(\omega)$ converges to the corresponding coordinate map of $\overline{\mathfrak{X}}_0$ at every point of continuity of the latter function. Define $\overline{Z}(t,s)$ for $t > 0$, $s \in \mathbb{R}$, by (3.7) and let

$$\overline{\mathfrak{X}}(t,s) = (\overline{A}(t,s), \overline{D}(t,s), \overline{T}(t,s), \overline{Y}(t,s), \overline{Z}(t,s)), \qquad t \geq 0,\ s \in \mathbb{R}.$$

(The introduction of the auxiliary processes $\overline{\mathfrak{X}}^{(\vartheta)}_0$, $\overline{\mathfrak{X}}_0$ in the above argument is necessary, because $Z(t,s)$ is not necessarily monotone in t.) Since the coordinates of $\overline{\mathfrak{X}}^{(\zeta)}_0$ (with the exception of $\overline{Y}^{(\zeta)}_j(t,s) = t - \overline{T}^{(\zeta)}_j(t,s)$) are distribution functions of nonnegative measures, the mapping $\overline{\mathfrak{X}}_0$ inherits this property, and hence its coordinate functions have at most countably many discontinuities. Consequently, the set $I(\overline{\mathfrak{X}})$ of discontinuities of $\overline{\mathfrak{X}}_0$ (and hence of $\overline{\mathfrak{X}}$) in $\mathbb{R}_+ \times \mathbb{R}$ is contained in a countable union of vertical and horizontal lines. Since $\overline{\mathfrak{X}}^{(\zeta)}(\omega)(t,s) \to \overline{\mathfrak{X}}(t,s)$ for every $(t,s) \in (\mathbb{R}_+ \times \mathbb{R}) \setminus I(\overline{\mathfrak{X}})$ and $\overline{\mathfrak{X}}$ is right-continuous, the equation (3.4) implies that $\overline{\mathfrak{X}}$ satisfies (3.9). In particular, because $\overline{T}_k(\cdot, s)$, $\overline{Y}_j(\cdot, s)$ are nondecreasing, (3.9) 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 $k \in \{1,\ldots,K\}$, $0 \le t < T_0 - C - 1$ and $s \in \mathbb{R}$,

$$2\epsilon \triangleq \overline{T}_k(t,s) - \overline{T}_k(t,s-) > 0. \tag{7.1}$$

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

$$s_2 - s_1 < \epsilon/(\alpha_{0,k} m_k), \tag{7.2}$$

and the function $\overline{T}_k$ is continuous at the points (t,s_1), (t,s_2). By (7.1) and the monotonicity of $T(t,s)$ in s, for ζ large enough we have

$$\epsilon|x_\zeta| \le T_k^{x_\zeta}((t+C)|x_\zeta|, (s_2+C)|x_\zeta|) - T_k^{x_\zeta}((t+C)|x_\zeta|, (s_1+C)|x_\zeta|). \tag{7.3}$$

In other words, the cumulative work done by time $(t+C)|x_\zeta|$ by server $j = s(k)$ on class k customers with deadlines at time $(t+C)|x_\zeta|$ belonging to the interval $((s_1+C)|x_\zeta|, (s_2+C|x_\zeta|]$ is at least $\epsilon|x_\zeta|$. It is easy to check that these customers arrived at the network in the time interval $((s_1 + C)|x_\zeta| - \mathcal{L}_\zeta, (s_2+C)|x_\zeta|]$. By (6.4), we have

$$\begin{aligned}
\epsilon|x_\zeta| &\le \sum_{i \in B_k^\zeta(s_1+C-\mathcal{L}_\zeta/|x_\zeta|, s_2+C)} v_k(i) \\
&\le m_k\left(\tilde{\alpha}_{0,k}(s_2+C) - \tilde{\alpha}_{0,k}(s_1+C-\mathcal{L}_\zeta/|x_\zeta|)\right)|x_\zeta| + o(|x_\zeta|) \\
&\le m_k\alpha_{0,k}\left((s_2-s_1)|x_\zeta| + \mathcal{L}_\zeta\right) + o(|x_\zeta|).
\end{aligned} \tag{7.4}$$

This, by (7.2) and Lemma 4.1, yields a contradiction for sufficiently large ζ. We have proved continuity of $\overline{T}(t,s)$ in s (the argument actually shows that $\overline{T}_k(t,s)$ is Lipschitz in s with the Lipschitz constant $m_k\alpha_{0,k}$). By (3.9), $\overline{Y}$ is Lipschitz in both variables, so $(\overline{T}^{(\zeta)}, \overline{Y}^{(\zeta)})(\omega)(t,s) \to (\overline{T}, \overline{Y})(t,s)$ for any $t, s \ge 0$. As in the proof of Lemma 4.2, it is easy to see that this convergence is u.o.c. in t, s.

We will now show that $\overline{\mathfrak{X}}$ satisfies (3.8) Let $\mathbf{k} = (k_1, \ldots, k_n) \in \mathbf{K}$, $T_0 > 0$ and let $0 \le t, s < T_0 - C - 1$. The first step in the justification of (3.8) is to show

$$\begin{aligned}
&T_{\mathbf{k}}^{(\zeta)}((t+C)|x_\zeta|, (s+C)|x_\zeta|) - T_{\mathbf{k}}^{(\zeta)}(C|x_\zeta|, C|x_\zeta|) \\
&\quad = m_{\mathbf{k}}\Big(D_{\mathbf{k}}^{(\zeta)}((t+C)|x_\zeta|, (s+C)|x_\zeta|) - D_{\mathbf{k}}^{(\zeta)}(C|x_\zeta|, C|x_\zeta|)\Big) + o(|x_\zeta|),
\end{aligned} \tag{7.5}$$

which is a pre-limit analog of (3.8), but for a fixed path $\mathbf{k}$ rather than a customer class k. This follows by an argument similar to the proof of (5.8) in [11], with (4.7) and (8.3) there replaced by our (6.2) and (6.5), respectively.

Fix $\epsilon > 0$ and choose m so large that

$$\alpha_{m,k} T_0 \max(m_k, 1) < \epsilon, \qquad k = 1, \dots, K. \tag{7.6}$$

For $0 \le t, s < T_0 - C - 1$ and $k = 1, \dots, K$, we have

$$\begin{aligned} \overline{T}_k^{(\zeta)}(t,s) &= \sum_{\mathbf{k} \in \tilde{\mathcal{C}}(k)} \overline{T}_{\mathbf{k}}^{(\zeta)}(t,s) \\ &= \sum_{\mathbf{k} \in \tilde{\mathcal{C}}(k) \setminus \mathbf{K}_{m,k}} \overline{T}_{\mathbf{k}}^{(\zeta)}(t,s) + \sum_{\mathbf{k} \in \mathbf{K}_{m,k}} \overline{T}_{\mathbf{k}}^{(\zeta)}(t,s), \end{aligned} \tag{7.7}$$

$$\begin{aligned} \overline{D}_k^{(\zeta)}(t,s) &= \sum_{\mathbf{k} \in \tilde{\mathcal{C}}(k)} \overline{D}_{\mathbf{k}}^{(\zeta)}(t,s) \\ &= \sum_{\mathbf{k} \in \tilde{\mathcal{C}}(k) \setminus \mathbf{K}_{m,k}} \overline{D}_{\mathbf{k}}^{(\zeta)}(t,s) + \sum_{\mathbf{k} \in \mathbf{K}_{m,k}} \overline{D}_{\mathbf{k}}^{(\zeta)}(t,s). \end{aligned} \tag{7.8}$$

By (7.5),

$$\sum_{\mathbf{k} \in \tilde{\mathcal{C}}(k) \setminus \mathbf{K}_{m,k}} \overline{T}_{\mathbf{k}}^{(\zeta)}(t,s) \ = \ m_k \sum_{\mathbf{k} \in \tilde{\mathcal{C}}(k) \setminus \mathbf{K}_{m,k}} \overline{D}_{\mathbf{k}}^{(\zeta)}(t,s) + o(1). \tag{7.9}$$

On the other hand, by (6.3), (6.6) and (7.6), for ζ large enough we have

$$\begin{aligned} 0 \le \sum_{\mathbf{k} \in \mathbf{K}_{m,k}} \overline{D}_{\mathbf{k}}^{(\zeta)}(t,s) \le \frac{1}{|x_\zeta|} N_{m,k}^{x_\zeta}(|x_\zeta|(t+C)) = \tilde{\alpha}_{m,k}(t+C) + o(1) \\ \le \alpha_{m,k} T_0 + o(1) < \epsilon, \end{aligned} \tag{7.10}$$

$$\begin{aligned} 0 &\le \sum_{\mathbf{k} \in \mathbf{K}_{m,k}} \overline{T}_{\mathbf{k}}^{(\zeta)}(t,s) \le \frac{1}{|x_\eta|} \sum_{i \in B_{m,k}^{\eta}(t+C)} v_k(i) \\ &= \frac{m_k}{|x_\eta|} |B_{m,k}^{\eta}(t+C)| + o(1) = \frac{m_k}{|x_\zeta|} N_{m,k}^{x_\zeta}(|x_\zeta|(t+C)) + o(1) \\ &\le \alpha_{m,k} m_k T_0 + o(1) < \epsilon. \end{aligned} \tag{7.11}$$

The relations (7.7)–(7.11) imply that at every point $(t,s) \in \mathbb{R}_+^2$ of continuity of the function $\overline{D}$, $|\overline{T}_k(t,s) - m_k \overline{D}_k(t,s)| \le \epsilon$. Since $\epsilon > 0$ is arbitrary, we actually have (3.8) at any such point and hence, by right-continuity of both $\overline{T}$ and $\overline{D}$, at any $t, s \ge 0$. We have proved that $\overline{\mathfrak{X}}$ satisfies (3.8). In particular, by Lipschitz continuity of $\overline{T}$, the function $\overline{D}$ is Lipschitz in both variables. Thus, $(\overline{D}^{(\zeta)}, \overline{T}^{(\zeta)}, \overline{Y}^{(\zeta)})(\omega)(t,s) \to (\overline{D}, \overline{T}, \overline{Y})(t,s)$ u.o.c. in $t, s \ge 0$.

We will now show that $\overline{\mathfrak{X}}$ satisfies (3.6). By (6.7), for every $t, s \ge 0$, $t, s \in \mathbb{Q}$ and $l = 1, \dots, K$, we have

$$
\begin{aligned}
&\overline{A}_l^{(\zeta)}(t,s) = \Delta_{C|x_\zeta|} A_l^{x_\zeta}(t|x_\zeta|, s|x_\zeta|)/|x_\zeta| \\
&= \frac{1}{|x_\zeta|}\left(A_l^{x_\zeta}((t+C)|x_\zeta|, (s+C)|x_\zeta|) - A_l^{x_\zeta}(C|x_\zeta|, C|x_\zeta|)\right) \\
&= \frac{1}{|x_\zeta|}\left(E_l^{x_\zeta}((t+C)|x_\zeta|, (s+C)|x_\zeta|) - E_l^{x_\zeta}(C|x_\zeta|, C|x_\zeta|)\right) \\
&+ \frac{1}{|x_\zeta|}\sum_{k=1}^{K}\Big(\Phi_{k,l}^{x_\zeta}(D_k^{x_\zeta}((t+C)|x_\zeta|, (s+C)|x_\zeta|), (t+C)|x_\zeta|, (s+C)|x_\zeta|) \\
&\qquad\qquad - \Phi_{k,l}^{x_\zeta}(D_k^{x_\zeta}(C|x_\zeta|, C|x_\zeta|), C|x_\zeta|, C|x_\zeta|)\Big) \qquad (7.12) \\
&= \alpha_l\left((t+C)\wedge(s+C) - C)\right) + o(1) \\
&+ \frac{1}{|x_\zeta|}\sum_{k=1}^{K} p_{kl}\left(D_k^{x_\zeta}((t+C)|x_\zeta|, (s+C)|x_\zeta|) - (D_k^{x_\zeta}(C|x_\zeta|, C|x_\zeta|)\right) \\
&= \alpha_l(t\wedge s) + \sum_{k=1}^{K} p_{kl}\Delta_{C|x_\zeta|} D_k^{x_\zeta}(t|x_\zeta|, s|x_\zeta|)/|x_\zeta| + o(1) \\
&= \alpha_l(t\wedge s) + \sum_{k=1}^{K} p_{kl}\overline{D}_k^{(\zeta)}(t,s) + o(1),
\end{aligned}
$$

where the third equation follows from (3.1) and the fourth one is a consequence of Lemma 4.2, together with (6.7). Letting $\zeta \to \infty$ in (7.12), we get

$$
\lim_{\zeta\to\infty} \overline{A}^{(\zeta)}(t,s) = \tilde{A}(t,s) \triangleq \alpha(t\wedge s) + P'\overline{D}(t,s) \qquad (7.13)
$$

for every $t, s \geq 0$, $t, s \in \mathbb{Q}$. However, the functions $\overline{A}_l^{(\zeta)}$, $\tilde{A}_l$, $l = 1, \ldots, K$, are nondecreasing in both variables and $\tilde{A}$ is continuous, so it is not hard to check that (7.13) actually holds for every $t, s \geq 0$. In particular, $\tilde{A} = \overline{A}$ and the proof of (3.6) is complete. Consequently, the function $\overline{A}$ is Lipschitz in both variables and thus $(\overline{A}^{(\zeta)}, \overline{D}^{(\zeta)}, \overline{T}^{(\zeta)}, \overline{Y}^{(\zeta)})(\omega)(t,s) \to (\overline{A}, \overline{D}, \overline{T}, \overline{Y})(t,s)$ u.o.c. in $t, s \geq 0$.

We will now prove (3.11). To this end, by (3.6)–(3.9), it suffices to show

$$
\overline{T}(t,s) = \overline{T}(t,t), \qquad (7.14)
$$

$$
\overline{Z}(0,s) = \overline{Z}(0,t), \qquad (7.15)
$$

for $0 \leq t < s$. The proof of (7.14) is similar to the argument showing continuity of $\overline{T}(t,s)$ in s. Suppose that (7.14) is false, i.e., for some $k \in \{1, \ldots, K\}$ and $0 \leq t < s < T_0 - C - 1$,

$$
2\epsilon \triangleq \overline{T}_k(t,s) - \overline{T}_k(t,t) > 0. \qquad (7.16)
$$

By (7.16) and the monotonicity of $T(t,s)$ in s, for ζ large enough, we have

$$(7.17)\quad \epsilon|x_\zeta| \le T_k^{x_\zeta}((t+C)|x_\zeta|,(s+C)|x_\zeta|) - T_k^{x_\zeta}((t+C)|x_\zeta|,(t+C)|x_\zeta|).$$

In other words, the cumulative work done by time $(t+C)|x_\zeta|$ by server $j=s(k)$ on class k customers with deadlines at time $(t+C)|x_\zeta|$ belonging to the interval $((t+C)|x_\zeta|,(s+C)|x_\zeta|]$ is at least $\epsilon|x_\zeta|$. It is easy to check that these customers arrived at the network in the time interval

$$((t+C)|x_\zeta|-\mathcal{L}_\zeta,(s+C)|x_\zeta|]\cap[0,(t+C)|x_\zeta|] = ((t+C)|x_\zeta|-\mathcal{L}_\zeta,(t+C)|x_\zeta|].$$

Arguing as in (7.4), with $s_1=s_2=t$, and using Lemma 4.1, we obtain a contradiction with (7.16) for sufficiently large ζ, which proves (7.14).

It remains to prove (7.15). By right-continuity of $\overline{Z}(0,\cdot)$, it actually suffices to show (7.15) under the additional assumption that both t and s are the points of continuity of $Z(0,\cdot)$. Suppose that for some $k\in\{1,\dots,K\}$ and $0\le t<s<T_0-C-1$, such that $\overline{Z}(0,\cdot)$ is continuous at both t and s,

$$(7.18)\qquad 2\epsilon \triangleq \overline{Z}_k(0,s)-\overline{Z}_k(0,t)>0.$$

For ζ large enough, by (7.18), (3.2), Lemma 4.3, (3.6), (3.8) and (7.14), we have

$$\begin{aligned}\epsilon|x_\zeta| &\le Z_k^{x_\zeta}(C|x_\zeta|,(s+C)|x_\zeta|) - Z_k^{x_\zeta}(C|x_\zeta|,(t+C)|x_\zeta|)\\ &\le A_k^{x_\zeta}(C|x_\zeta|,(s+C)|x_\zeta|) - A_k^{x_\zeta}(C|x_\zeta|,(t+C)|x_\zeta|)\\ &= \left(\overline{A}_k(0,s)-\overline{A}_k(0,t)\right)|x_\zeta| + o(|x_\zeta|) = o(|x_\zeta|).\end{aligned}$$

This contradiction proves (7.15), and hence (3.11).

Using (3.11), together with (3.6)–(3.9), and arguing as on p. 90 of [5], we get Lipschitz continuity of $\overline{Z}(t,s)$ in s. Thus, $\overline{\mathfrak{X}}(t,s)$ is Lipschitz in both variables and consequently, u.o.c. in t and s,

$$(7.19)\qquad \overline{\mathfrak{X}}^{(\zeta)}(\omega)(t,s)\to\overline{\mathfrak{X}}(t,s).$$

We now show that $\overline{\mathfrak{X}}$ satisfies (3.10). Let $T_0>0$, $s\ge 0$. By (3.5), we have

$$(7.20)\qquad \int_0^{T_0}\sum_{k\in\mathcal{C}(j)}\overline{Z}_k^{(\zeta)}(t,s)\,\overline{Y}_j^{(\zeta)}(dt,s)=0.$$

By Lemma 4.4 in [6], (7.19)-(7.20) imply

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

and (3.10) follows.

The coordinate mappings of $\overline{\mathfrak{X}}$ inherit the nonnegativity and monotonicity properties from the corresponding coordinates of $\overline{\mathfrak{X}}^{(\vartheta)}(\omega)$, and hence from those of $\mathfrak{X}^{x_\vartheta}(\omega)$. It remains to verify the initial conditions

$$\overline{A}(0,s)=\overline{D}(0,s)=\overline{T}(0,s)=0,\quad \overline{Y}(0,s)=0,\qquad s\geq 0. \tag{7.21}$$

First note that (7.21) holds for $s=0$. Indeed, by definition

$$\begin{aligned}\overline{A}(0,0)&=\lim_{\zeta\to\infty}\overline{A}^{(\zeta)}(0,0)=\lim_{\zeta\to\infty}\Delta_{C|x_\zeta|}A^{x_\zeta}(0,0)/|x_\zeta|\\&=\lim_{\zeta\to\infty}\left(A^{x_\zeta}(C|x_\zeta|,C|x_\zeta|)-A^{x_\zeta}(C|x_\zeta|,C|x_\zeta|)\right)/|x_\zeta|=0.\end{aligned}$$

Similarly $\overline{D}(0,0)=\overline{T}(0,0)=0$ and $\overline{Y}(0,0)=0$. Thus, by continuity of A,D,T,Y, in order to establish (7.21), it suffices to verify that for every $0<s_1<s_2$, we have $\overline{A}(0,s_1)=\overline{A}(0,s_2)$, $\overline{D}(0,s_1)=\overline{D}(0,s_2)$, $\overline{T}(0,s_1)=\overline{T}(0,s_2)$ and $\overline{Y}(0,s_1)=\overline{Y}(0,s_2)$. This, however, follows from the fact that, by Lemma 4.1, for ζ large enough, no customer arriving at the system after time 0 has lead time greater than $s_1|x_\zeta|$ at time $C|x_\zeta|$, while Lemma 4.3 implies that every initial customer has already left the system by that time. Consequently, for such large ζ,

$$\begin{aligned}&\overline{A}^{(\zeta)}(0,s_2)-\overline{A}^{(\zeta)}(0,s_1)\\&\quad=\left(A^{x_\zeta}(C|x_\zeta|,(s_2+C)|x_\zeta|)-A^{x_\zeta}(C|x_\zeta|,(s_1+C)|x_\zeta|)\right)/|x_\zeta|=0.\end{aligned}$$

Letting $\zeta\to\infty$, we get $\overline{A}(0,s_1)=\overline{A}(0,s_2)$. The proofs of the remaining inequalities are similar. □

8. Proofs of Proposition 4.5 and Theorem 4.6

Proof of Proposition 4.5. By Lemma 4.3, for ζ sufficiently large, all initial customers have left the system with initial state x_ζ by time $C|x_\zeta|$. Let $T_0>C+1$ and let $0\leq t\leq T_0-C-1$. Arguing as in the proof of Proposition 6.1 in [11], one may check that for every $\mathbf{k}\in\mathbf{K}$,

$$\lim_{\zeta\to\infty}\frac{1}{|x_\zeta|}\left|W_{\mathbf{k}}^{x_\zeta}((t+C)|x_\zeta|)-m_{\mathbf{k}}Q_{\mathbf{k}}^{x_\zeta}((t+C)|x_\zeta|)\right|=0. \tag{8.1}$$

Fix $\epsilon>0$ and choose m so large that $\alpha_{m,k}m_kT_0<\epsilon/2$, $k=1,\dots,K$. We have

$$\begin{aligned}W_k^{x_\zeta}((t+C)|x_\zeta|)&=\sum_{\mathbf{k}\in\tilde{\mathcal{C}}(k)}W_{\mathbf{k}}^{x_\zeta}((t+C)|x_\zeta|)\\&=\sum_{\mathbf{k}\in\tilde{\mathcal{C}}(k)\setminus\mathbf{K}_{m,k}}W_{\mathbf{k}}^{x_\zeta}((t+C)|x_\zeta|)+\sum_{\mathbf{k}\in\mathbf{K}_{m,k}}W_{\mathbf{k}}^{x_\zeta}((t+C)|x_\zeta|),\end{aligned} \tag{8.2}$$

$$\begin{aligned}Q_k^{x_\zeta}((t+C)|x_\zeta|)&=\sum_{\mathbf{k}\in\tilde{\mathcal{C}}(k)}Q_{\mathbf{k}}^{x_\zeta}((t+C)|x_\zeta|)\\&=\sum_{\mathbf{k}\in\tilde{\mathcal{C}}(k)\setminus\mathbf{K}_{m,k}}Q_{\mathbf{k}}^{x_\zeta}((t+C)|x_\zeta|)+\sum_{\mathbf{k}\in\mathbf{K}_{m,k}}Q_{\mathbf{k}}^{x_\zeta}((t+C)|x_\zeta|).\end{aligned} \tag{8.3}$$

As in (7.10)–(7.11), for ζ large enough,

$$\begin{aligned}0 &\le \frac{1}{|x_\zeta|}\sum_{\mathbf{k}\in\mathbf{K}_{m,k}} Q_k^{x_\zeta}((t+C)|x_\zeta|) \le \frac{1}{|x_\zeta|} N_{m,k}^{x_\zeta}(|x_\zeta|(t+C)) \\ &= \tilde{\alpha}_{m,k}(t+C) + o(1) \le \alpha_{m,k}T_0 + o(1) < \epsilon/(2m_k),\end{aligned} \tag{8.4}$$

$$\begin{aligned}0 &\le \frac{1}{|x_\zeta|}\sum_{\mathbf{k}\in\mathbf{K}_{m,k}} W_{\mathbf{k}}^{x_\zeta}((t+C)|x_\zeta|) \le \frac{1}{|x_\eta|}\sum_{i\in B_{m,k}^{\eta}(t+C)} v_k(i) \\ &= \frac{m_k}{|x_\eta|}|B_{m,k}^{\eta}(t+C)| + o(1) = \frac{m_k}{|x_\zeta|} N_{m,k}^{x_\zeta}(|x_\zeta|(t+C)) + o(1) < \epsilon/2.\end{aligned} \tag{8.5}$$

By (8.1)–(8.5), for ζ large enough,

$$\begin{aligned}&\frac{1}{|x_\zeta|}\left|W_k^{x_\zeta}((t+C)|x_\zeta|) - m_k Q_k^{x_\zeta}((t+C)|x_\zeta|)\right| \\ &\quad\le \sum_{\mathbf{k}\in\tilde{\mathcal{C}}(k)\setminus\mathbf{K}_{m,k}} \frac{1}{|x_\zeta|}\left|W_{\mathbf{k}}^{x_\zeta}((t+C)|x_\zeta|) - m_{\mathbf{k}} Q_{\mathbf{k}}^{x_\zeta}((t+C)|x_\zeta|)\right| \\ &\quad\quad + \frac{1}{|x_\zeta|}\sum_{\mathbf{k}\in\mathbf{K}_{m,k}} W_{\mathbf{k}}^{x_\zeta}((t+C)|x_\zeta|) + \frac{m_k}{|x_\zeta|}\sum_{\mathbf{k}\in\mathbf{K}_{m,k}} Q_k^{x_\zeta}((t+C)|x_\zeta|) < \epsilon,\end{aligned}$$

and (4.13) follows. □

For the proof of Theorem 4.6, let $\mathcal{P}_k(t)$, $k = 1,\ldots,K$, and $\mathcal{P}_{\mathbf{k}}(t)$, $\mathbf{k} \in \mathbf{K}$, denote the number of partially served class k and type $\mathbf{k}$ customers, respectively, present at the system at time t. We further decompose $\mathcal{P}_{\mathbf{k}}(t)$ into $\mathcal{P}_{\mathbf{k},0}(t)$, counting those type $\mathbf{k}$ customers partially served at time t who were already present in the system at time 0, and $\mathcal{P}_{\mathbf{k},1}(t) = \mathcal{P}_{\mathbf{k}}(t) - \mathcal{P}_{\mathbf{k},0}(t)$, counting type $\mathbf{k}$ customers partially served at time t who arrived at the system after time 0. Note that at any time t there may be at most one type $\mathbf{k}$ partially served job which was already present in the system at time 0, because the corresponding customers move along the path $\mathbf{k}$ one by one in the order determined by their deadlines. Hence

$$\mathcal{P}_{\mathbf{k}}(t) \le \mathcal{P}_{\mathbf{k},\mathbf{1}}(t) + 1. \tag{8.6}$$

Clearly,

$$\mathcal{P}(t) = \sum_{k=1}^{K} \mathcal{P}_k(t), \tag{8.7}$$

and for $k = 1,\ldots,K$ and $m \in \mathbb{N}$, we have

$$\mathcal{P}_k(t) = \sum_{\mathbf{k}\in\tilde{\mathcal{C}}(k)} \mathcal{P}_{\mathbf{k}}(t) = \sum_{\mathbf{k}\in\tilde{\mathcal{C}}(k)\setminus\mathbf{K}_{m,k}} \mathcal{P}_{\mathbf{k}}(t) + \sum_{\mathbf{k}\in\mathbf{K}_{m,k}} \mathcal{P}_{\mathbf{k}}(t). \tag{8.8}$$

Proof of Theorem 4.6. We will prove only the second claim. The proof of the first one is similar, but simpler, since in this case, by Lemma 4.3, we do not have to take the initial customers into account. If the second claim is false, there exist $T_0 > C$, $\epsilon > 0$ and a sequence $x_n \in \Omega$ satisfying (4.4) such that

$$\mathbb{E}\left[\max_{0\le t\le T_0} \mathcal{P}^{x_n}(t|x_n|)\right] \ge \epsilon|x_n| \tag{8.9}$$

for every n. Let the set G' and the subsequence x_ξ be as in Lemma 6.1. For $k = 1, \dots, K$ and $m \in \mathbb{N}$, let s_k^m be the probability that a customer entering the network at class k visits more than m (not necessarily distinct) classes before he exits the system and let $s^m = \sum_{k=1}^K s_k^m$. We have $\lim_{m\to\infty} s^m = 0$, because the network is open. Let $I_{m,k}^{x_\xi}$ be the number of initial class k customers visiting more than m (not necessarily distinct) classes before exiting the network with initial state x_ξ and let $I_m^{x_\xi} = \sum_{k=1}^K I_{m,k}^{x_\xi}$. We claim that as $\xi \to \infty$,

$$\left(I_{m,k}^{x_\xi} - 3s_k^m|x_\xi|/2\right)^+ \xrightarrow{P} 0. \tag{8.10}$$

Suppose that (8.10) fails. Without loss of generality, extracting a subsequence if necessary, we may assume either that $Q_k^{x_\xi}(0) \to \infty$ as $\xi \to \infty$, or that $Q_k^{x_\xi}(0)$ are bounded uniformly in ξ. In the first case, $I_{m,k}^{x_\xi}/Q_k^{x_\xi}(0) \xrightarrow{P} s_k^m$ as $\xi \to \infty$ by the weak law of large numbers and $Q_k^{x_\xi}(0) \le |x_\xi|$, so (8.10) holds. In the second case (8.10) follows from (4.4) and the fact that $I_{m,k}^{x_\xi} \le Q_k^{x_\xi}(0)$. We have proved (8.10).

Using Theorem 20.5 in [1], together with the Cantor diagonal procedure, and arguing as in the proof of (6.5), we can construct a set $G'' \subseteq G'$ with $\mathbb{P}(G'') = 1$ and a subsequence of x_ξ (still denoted by x_ξ) such that on G'' we have pointwise convergence $\left(I_{m,k}^{x_\xi} - 3s_k^m|x_\xi|/2\right)^+ \to 0$ as $\xi \to \infty$ for *every* $m \in \mathbb{N}$, $k = 1, \dots, K$. In particular, for $\omega \in G''$, $m \in \mathbb{N}$, ξ large enough and every $t \ge 0$,

$$\sum_{\mathbf{k}\in\mathbf{K}_m} \mathcal{P}_{\mathbf{k},0}^{x_\xi}(t|x_\xi|) \le I_m^{x_\xi} \le 2s^m|x_\xi|. \tag{8.11}$$

We will show that on G'',

$$\lim_{\xi\to\infty} \max_{0\le t\le T_0} \mathcal{P}^{x_\xi}(t|x_\xi|)/|x_\xi| = 0. \tag{8.12}$$

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

$$\max_{0\le t\le T_0} \mathcal{P}^{x_\vartheta}(t|x_\vartheta|)(\omega)| \ge \epsilon_1 |x_\vartheta|. \tag{8.13}$$

Choose m so large that $s^m \le \epsilon_1/6$ and $\alpha_{m,k}T_0 K < \epsilon_1/3$ for $k = 1,\dots,K$. Proceeding as in (7.10) or (8.4), on the set G'' for $k = 1,\dots,K$, ϑ large enough and all $t \in [0,T_0]$,

$$\frac{1}{|x_\vartheta|}\sum_{\mathbf{k}\in\mathbf{K}_{m,k}} \mathcal{P}^{x_\vartheta}_{\mathbf{k},1}(t|x_\vartheta|) \le \frac{1}{|x_\vartheta|} N^{x_\vartheta}_{m,k}(T_0|x_\vartheta|) \le \alpha_{m,k}T_0 + o(1) < \frac{\epsilon_1}{3K}. \tag{8.14}$$

Let $\epsilon_2 = \epsilon_1/(3K|\tilde{\mathcal{C}}(k) \setminus \mathbf{K}_{m,k}|)$. We will now show that for any $\mathbf{k} \in \tilde{\mathcal{C}}(k) \setminus \mathbf{K}_{m,k}$ and ϑ large enough, on the set G' we have

$$\max_{0\le t\le T_0} \mathcal{P}^{x_\vartheta}_{\mathbf{k}}(t|x_\vartheta|) \le \epsilon_2 |x_\vartheta|. \tag{8.15}$$

For $\mathbf{k} = (k_1,\dots,k_n) \in \tilde{\mathcal{C}}(k) \setminus \mathbf{K}_{m,k}$ and $t \ge 0$, let $b^{(\vartheta)}_{\mathbf{k}}(t)$ be the arrival time at the network with initial state x_ϑ of the type $\mathbf{k}$ customer who was the last one to receive service at station $j = s(e(\mathbf{k}))$ by time $t|x_\vartheta|$ among those routed to the class $k_n = e(\mathbf{k})$ along the path $\mathbf{k}$. Each type $\mathbf{k}$ customer who arrived at the network before $b^{(\vartheta)}_{\mathbf{k}}(t) - (\mathcal{L}_\vartheta \vee l^+_\vartheta)$ (in particular, by $b^{(\vartheta)}_{\mathbf{k}}(t) - (\mathcal{L}_\vartheta \vee l^+_\vartheta) - 1$) has already received full service at k_n, the n-th class on his route, by time $t|x_\vartheta|$. Type $\mathbf{k}$ customers who arrived at the network after $b^{(\vartheta)}_{\mathbf{k}}(t) + (\mathcal{L}_\vartheta \vee l^+_\vartheta)$ cannot receive service before the type $\mathbf{k}$ customer who arrived at time $b^{(\vartheta)}_{\mathbf{k}}(t)$. Consequently, none of these two groups of customers contributes to $\mathcal{P}^{x_\vartheta}_{\mathbf{k},1}(t|x_\vartheta|)$ and hence, uniformly in $0 \le t \le T_0$, we have the bounds

$$\begin{aligned}\mathcal{P}^{x_\vartheta}_{\mathbf{k},1}(t|x_\vartheta|) &\le N^{x_\vartheta}_{\mathbf{k}}\Big(b^{(\vartheta)}_{\mathbf{k}}(t) + (\mathcal{L}_\vartheta \vee l^+_\vartheta)\Big) - N^{x_\vartheta}_{\mathbf{k}}\Big(b^{(\vartheta)}_{\mathbf{k}}(t) - (\mathcal{L}_\vartheta \vee l^+_\vartheta) - 1\Big)\\ &= \alpha_{\mathbf{k}}(2(\mathcal{L}_\vartheta \vee l^+_\vartheta) + 1) + o(|x_\vartheta|) \ = \ o(|x_\vartheta|),\end{aligned}$$

where the equalities follow from (6.2), Lemma 4.1 and the fact that $\ell = 0$ in (4.4). This, together with (8.6), proves (8.15).

By (8.7)–(8.8), (8.11) and (8.14)–(8.15), for ϑ large enough, we get

$$\max_{0\le t\le T_0} \mathcal{P}^{x_\vartheta}(t|x_\vartheta|)(\omega)| < \epsilon_1 |x_\vartheta|,$$

which contradicts (8.13). We have proved (8.12).

Finally,

$$0 \leq \max_{0 \leq t \leq T_0} \mathcal{P}^{x_\xi}(t|x_\xi|)/|x_\xi| \leq 1 + \sum_{k=1}^{K} N_k^{x_\xi}(T_0|x_\xi|)/|x_\xi|$$
$$\leq 1 + \sum_{k=1}^{K} N_k^{\mathbf{0}}(T_0|x_\xi|)/|x_\xi|,$$

and it was shown in the proof of Lemma 4.5 in [6] that for $k = 1, \dots, K$, the sequences $\{N_k^{\mathbf{0}}(T_0|x_\xi|)/|x_\xi|\}$ are uniformly integrable. Hence the sequence $\{\max_{0\leq t\leq T_0} \mathcal{P}^{x_\xi}(t|x_\xi|)/|x_\xi|\}$ is also uniformly integrable, and thus, by (8.12), it converges to 0 in L^1 as $\xi \to \infty$. This, however, contradicts (8.9). □

Acknowledgement. The author thanks the anonymous referee for his helpful suggestions improving the paper.

References

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

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

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

[4] Bramson, M., *State space collapse with application to heavy traffic limits for multiclass queueing networks*, Queueing Syst. Theory Appl. **30** (1998), 89–148.

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

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

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

[8] Davis, M. H. A., *Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models*, Journal of the Royal Statistical Society. Series B **46** (1984), 353–388.

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

[10] Getoor, R. K., *Transience and recurrence of Markov processes*, Séminaire de Probabilités XIV **284**, 397–409, Springer, New York, 1979.

[11] Kruk, Ł, *Stability of two families of real-time queueing networks*, Probab. Math. Stat. **28** (2008), 179–202.

[12] Kruk, Ł, *Invariant states for fluid models of EDF networks: nonlinear lifting map*, Probab. Math. Stat. **30** (2010), 289–315.

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

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

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

[16] Williams, R. J., *Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse*, Queueing Syst. Theory Appl. **30** (1998), 27–88.

[17] 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.

Łukasz Kruk
Institute of Mathematics
Maria Curie-Skłodowska University
20-031 Lublin
Poland
e-mail: `lkruk@hektor.umcs.lublin.pl`

Received March 8, 2019