STABILITY OF CASCADE NETWORKS VIA FLUID MODELS

ROSARIO DELGADO[1] AND EVSEY MOROZOV[2]

Abstract. The fluid approach is applied to find stability conditions of a two-station cascade network (customers that are awaiting service at the the first queue can move to the second station, whenever it is free, to be served there immediately, but the opposite is not allowed). Each station is fed by a renewal input with general i.i.d. inter-arrival times and general i.i.d. service times (possibly different in the two stations). Then the stability analysis is extended to two classes of cascade networks with an arbitrary number $N \geq 2$ of stations. In *type-I* network, an awaiting customer from the queue i jumps to (free) station $i+1$ only, while in *type-II* network, station $j = 2, \ldots, N$ can help any previous station $1, \ldots, j-1$.

1. Introduction

In this paper, we first consider a basic queueing network consisting of two single-server infinite-capacity stations, where stations have independent general renewal inputs and i.i.d. general service times. Both inter-arrival time and service time distributions are (possibly) different in the two stations. Whenever the 2nd station becomes empty while customers are awaiting service at the 1st *queue* (that is, the server at station 1 is busy and there are customers waiting to be served at the queue of the 1st station), one customer jumps from the 1st queue to the 2nd station to be served there. However, the 1st station is unable to support the 2nd in the same manner. For this reason the system is called *cascade* ([20]). Then we extend analysis to cascade networks with $N \geq 3$ stations and two different types of interactions between stations. The recent paper [20] contains a detailed motivation of importance of such cascade models, and also a comprehensive survey of the existing literature on this topic including various applications. For this reason, in the following discussion we will only deal with major issues addressed in [20].

Cascade queueing networks are related with systems with *flexible* servers, where a server may transfer some service capacity to accommodate workload accumulated in another server. Such networks with flexible servers have been studied,

2010 *Mathematics Subject Classification.* 60K25, 60K20, 90B15, 90B22.

Key words and phrases. cascade queueing network, stability, fluid limit approach, Lyapunov function, Skorokhod Problem.

[1] Supported by the project MEC-FEDER ref. MTM2009-08869.

[2] Supported by the Program of strategic development of Petrozavodsk State University for 2012-2016 and by RFBR project 10-07-00017.

for instance, in [3, 13, 17, 22, 26]. Related to this is the concept of *cross-trained* servers, whereby some servers can serve a reduced set of customers types, whereas others accept all types [1, 2, 23, 24]. These schemes can be used to model a variety of real-life systems, including service centers, production systems, computer networks with rescheduling of jobs, and parallel computing systems where processors have overlapping capabilities [5]. There are also manufacturing applications in which machines may have differing primary functions although some overlapping secondary ones.

It is worth noting that the focus of the most related papers has been to establish an optimal server allocation to minimize a cost function. However, much less attention has been devoted to obtaining stability conditions. Some works study (sub)optimal scheduling disciplines in multi-class systems under heavy traffic regime (see [4, 6, 15, 18, 22] and references therein). In these papers the boundary of the stability region is defined based on the resource pooling (RP) or complete resource pooling (CRP) assumption. Such an assumption is formulated as a requirement that the input rate in a pool of servers must be less than (RP) or equal to (CRP) the maximal service rate of the pool, but concrete service rates are not specified there. At the same time, as papers [14] and [20] show, it is not a trivial problem in general, to find a condition for service rates which delimit the stability region (in terms of predefined parameters).

We also mention closely related work [25], where stability analysis of a two-server system operating as our cascade model is studied. The main difference is that in [25] customers arrived to station 1 have (non-preemptive) priority to be served by server 2, provided the queue in server 1 exceeds some threshold $C \geq 0$. To solve stability problem, the author adopts an *augmented* fluid flow approach because he states that traditional fluid analysis of [9] cannot capture the whole behavior of the system. (However, as it is shown in [20], the stability condition found in [25] is not tight.) Another closely related system is studied in the mentioned above paper [6]. It turns out to be that the RP *assumption* used in [6] to delimit the stability region is exactly the stability *condition* obtained in [20] by the regenerative method for the model with Poisson input to 1st station. In the present paper, the same result is proved by using the fluid limit approach but for general (unbounded and spread-out) i.i.d. inter-arrival and general i.i.d service times.

As an important source of motivation for our research we mention [14], where stability analysis of a particular case of our basic model is carried out, by using standard Foster's type arguments through a rather complicated proof. In general, stability analysis of cascade networks is not a trivial matter, and so far this question has been solved only for two-station systems, by using different techniques including the regenerative method, the augmented fluid approach and (in the Markovian setting) Foster's type arguments. We note that the developed proofs are not simple in general, and it is not obvious how to extend them beyond the two-station cascade networks.

Thus, the main motivation of this research has been to develop stability analysis that could be applied to general N-station cascade networks, $N \geq 2$. Foremost we develop stability analysis of a two-station cascade network. More exactly, we find conditions under which the underlying Markov process describing the network dynamics is *positive Harris recurrent*, and see that match those known ([6, 20]). For our stability analysis we use conventional fluid model approach. Although it is known that a state-dependent service discipline can be a source of problems when applying fluid stability analysis (see [7]), a deeper insight allows us to apply this approach to the general $N-$station situation, in spite of this difficulty, which in fact is present in our model.

Following [9], we first establish the stability of the fluid limit model associated to the queueing network, which allows to transform the initial stochastic problem into a (related) deterministic one. The stability of the fluid limit model means that, regardless of the initial state, the fluid limit of the queue-size process reaches zero in a finite time interval, and stays there from that time forward. In the paper [9], functional laws of large numbers for the renewal processes or, in other words, hydrodynamic scaling by the increasing value of the initial state, are used to obtain the stability of the fluid limit model via the solution of a Skorokhod problem. At that, the choice of an appropriate Lyapunov function is the key point of analysis. This paper gives an example of application of this methodology to an unconventional model.

We note that the fluid approach and the regenerative method are not equivalent stability analysis methods because, in general, they use different assumptions and lead to slightly different conclusions. For instance, regenerative approach establishes that the stability condition in Theorem 1 below is the stability criterion, while the fluid analysis only can show that it is a sufficient condition. (A detailed discussion on this topic is can be found in [20].)

The crucial fact in the proof of main Theorem 1 is that the fluid limit of the *workload process* turns out to be part of a solution of a linear Skorokhod Problem, while in [9] a similar fact for the fluid limit of the *queue-size process*, is used instead. As mentioned in [12], the workload process seems to be better adapted to the use of the methodology of the Skorokhod Problem than the queue-size process. On the other hand, as said before, the adequate choice of the Lyapunov function, which always has a simple form in our models, also plays an important role in the proofs.

The organization of the paper is as follows. In Section 2 we introduce main notations and definition of the Skorokhod Problem. Section 3 introduces the basic two-station cascade queueing network we deal with, and the queueing network equations that govern the processes associated to the network. Section 4 is devoted to the study of the stability of such a network. In particular, Section 4.1 introduces the *fluid limit model* associated to the network as well as the definitions of the *fluid limit* and the *stability of the fluid limit model*, and in Section 4.2 we state and prove main Theorem 1. In Sections 5, 6 we extend the stability

analysis to N-station cascade networks. Actually, we handle with two different generalizations of the basic two-station cascade system to more than two stations. In Section 5, we consider a three-station model, in which station i can only help strictly preceding station $i-1$, $i=2,3$. Then we generalize this model for N stations and call it *type I-cascade network*. In section 6 we study the following N-station *type-II cascade network*: if station j becomes idle, $j=2,\dots,N$, it can handle a customer that is waiting at any of the non-empty queues $1,\dots,j-1$. This study is also based on a detailed analysis of the three-station case.

2. Notations and basic definitions

In this section, we introduce main notations and definitions. Vector (in)equalities are interpreted component-wise. For any integer $d\geq 1$, let $\mathbb{R}^d_+=\{\,v\in\mathbb{R}^d\,:\,v\geq 0\,\}$ and $\mathbb{Z}^d_+=\{\,v=(v_1,\dots,v_d)\in\mathbb{R}^d\,:\,v_i\in\mathbb{Z}_+\,\}$. We denote by I the identity matrix (of the corresponding dimension). For a vector $v=(v_1,\dots,v_d)$, let $|v|=\sum_i |v_i|$. We say that a sequence of vectors $\{v^n\}$ converges to a vector v if $|v^n-v|\to 0$, $n\to\infty$, and denote it as $\lim_{n\to\infty} v^n=v$. (This convergence is equivalent to the convergence in the component-wise sense.) For $n\geq 1$, let $\phi_n\colon[0,\infty)\to\mathbb{R}^d$ be right continuous functions having limits on the left on $(0,+\infty)$, and let function $\phi\colon[0,+\infty)\to\mathbb{R}^d$ be continuous. We say that ϕ^n converges to ϕ as $n\to\infty$ *uniformly on compacts* (u.o.c.) if for any $T\geq 0$,

$$||\phi^n-\phi||_T\stackrel{\text{def}}{=}\sup_{t\in[0,T]}|\phi_n(t)-\phi(t)|\longrightarrow 0\,,$$

and write it as $\lim_{n\to\infty}\phi^n=\phi$. If function ϕ is differentiable at a point $s\in(0,\infty)$ then s is a *regular* point of ϕ, and we denote the derivative by $\dot\phi(s)$. We denote a process $X=\{X(t),\,t\geq 0\}$ as X^x if $X(0)=x$.

Definition. Let $d\geq 1$ and R be a $d\times d$ matrix, $x\in\mathbb{R}^d_+$ and X be a $d-$dimensional stochastic process defined on some probability space with continuous paths and $X(0)=0$. We say that the pair of $d-$dimensional stochastic processes (W,Y) with continuous paths and defined on the same probability space is a *solution of the Skorokhod problem* associated to the process X with initial state x and *reflection matrix* R in the orthant $\mathbb{R}^d_+$ if

(i) $W(t)\in\mathbb{R}^d_+\quad$ for all $t\geq 0\,$,
(ii) $W(t)=x+X(t)+RY(t)$ with probabilty 1 (w.p.1)$\,,\ t\geq 0\,$,
(iii) Y has non $-$ decreasing paths on $[0,+\infty)$ and
$Y_j(0)=0$ and Y_j increases only when $W_j=0$, that is :
$\int_0^\infty W_j(t)\,dY_j(t)=0$, $j=1,\dots,d\,$.

(The deterministic version of the Skorokhod problem is also known as *dynamic complementarity problem* [9].)

Remark 1. In the one-dimensional case, for any given initial state x, the solution of the Skorokhod problem with reflection matrix R exists if $R > 0$ (Theorem I.1.2 [16]). In the $d-$dimensional case, an assumption on matrix R is known implying strong path-wise uniqueness of the solution (see condition (II) in Proposition 4.2 [27] or condition (HR) in [12], for technical details). In our models $R = I$, and this assumption is trivially satisfied.

3. The two-station cascade queueing network

In this section, we describe the two-station cascade queueing network in detail. Each station has an infinite-capacity buffer for awaiting customers who arrive from outside. As soon as station 2 becomes free, an awaiting customer at buffer of station 1 (if any) switches to station 2 and starts service immediately. We call class-i exogenous customers who arrive at station $i = 1, 2$, and class-$(1, 2)$ customers jumping from station 1 to station 2. (For stability analysis, it does not matter which class-1 customer makes the jump becoming a class-$(1, 2)$ customer.) In what follows, we use index i (respectively, double index $1, 2$) to denote the quantities related to class-i (respectively, class-$(1, 2)$) customers. Let $\{\xi_i(j),\, j \geq 2\}$ be the i.i.d. inter-arrival times of class-i customers ($i = 1, 2$) and let $\{\eta_k(j),\, j \geq 2\}$ be the i.i.d. service times of class-k customers, $k = 1, 2, (1, 2)$. All sequences are assumed to be mutually independent. We will omit corresponding index to denote a generic element of an i.i.d sequence. The residual arrival time $\xi_i(1)$ of the first class-i customer, entering the network after instant 0, is independent of $\{\xi_i(j),\, j \geq 2\}$, $i = 1, 2$. Also the residual service time $\eta_k(1)$ of a class-k customer initially being served is independent of $\{\eta_k(j),\, j \geq 2\}$, and $\eta_k(1) =_{st} \eta_k$ if class k is initially empty, $k = 1, 2, (1, 2)$ ($=_{st}$ means stochastic equality.)

As depicted in Fig. 1, we introduce the arrival rate $\lambda_i = 1/\mathsf{E}\xi_i$ of class-i customers ($i = 1, 2$) and the mean service time $m_k = \mathsf{E}\eta_k$ of class-k customers with the rate $\mu_k = m_k^{-1},\ k = 1,\ 2,\ (1, 2)$.

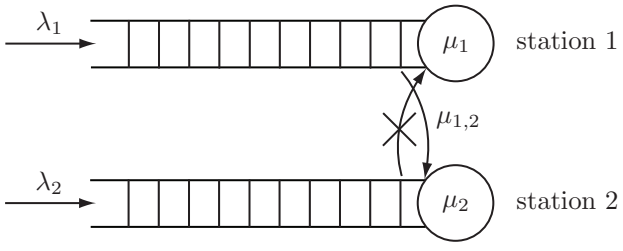

Figure 1. A cascade network with two stations

Define the 3-dimensional diagonal matrix M with entries $m_1,\ m_2,\ m_{1,2}$, respectively. We impose the following standard conditions (see [9]):

$$\mathsf{E}\,\eta_k < \infty, \quad k = 1, 2, (1, 2); \tag{1}$$

$$\mathsf{E}\,\xi_i < \infty, \quad i = 1,\, 2; \tag{2}$$

$$\mathsf{P}(\xi_i \geq x) > 0,\ i = 1, 2;\ \text{for any } x \in [0,\, \infty). \tag{3}$$

Also we assume that inter-arrival times are *spread out*, that is some integers $r_i > 1$ and functions $f_i \geq 0$ with $\int_0^\infty f_i(y)\,dy > 0$ exist such that

$$\mathsf{P}\left(a \leq \sum_{\ell=2}^{r_i} \xi_i(\ell) \leq b\right) \geq \int_a^b f_i(y)\,dy, \quad \text{for any } 0 \leq a < b,\ \ i = 1, 2. \tag{4}$$

We assume that customers waiting in queue at each station are processed in a *head-of-the-line* (HL) and *work-conserving* (or *non-idling*) discipline: a server is never idle when there are some customers at the station. In our case this means that server at station 1 is never idle if there are waiting class-1 customers, and server at station 2 is never idle if there are customers waiting at *any buffer*. We also assume that the queueing discipline is *non-preemptive*.

Now we introduce the primitives E, S, Υ describing the queueing network:

The exogenous arrival process $E = \{E(t) := (E_1(t),\, E_2(t)),\, t \geq 0\}$, where

$$E_i(t) = \max\left\{\, n \geq 1 \,:\, \sum_{\ell=1}^{n} \xi_i(\ell) \leq t\right\}$$

is the total number of class-i arrivals in interval $[0,\, t]$, $i = 1,\, 2$.

The process of served customers $S = \{S(t) := (S_1(t),\, S_2(t),\, S_{1,2}(t)),\, t \geq 0\}$, where the renewal process

$$S_k(t) = \max\left\{\, n \geq 1 \,:\, \sum_{\ell=1}^{n} \eta_k(\ell) \leq t\right\},$$

is the total number of class-k customers served in $[0,\, t]$ if the server *devotes all time to class* k. (By definition, $\max \emptyset = 0$).

Finally, define the cumulative service time process

$$\Upsilon = \{\Upsilon(n) := (\Upsilon_1(n_1),\, \Upsilon_2(n_2),\, \Upsilon_{1,2}(n_{1,2})),\, n := (n_1,\, n_2,\, n_{1,2}) \in \mathbb{N}^3\}\,,$$

where

$$\Upsilon_k(n_k) = \sum_{\ell=1}^{n_k} \eta_k(\ell)$$

is the cumulative service time of the first n_k class-k customers. Note that $S_k(t) = \max\{n \geq 1 : \Upsilon_k(n) \leq t\}$ and that $E(0) = S(0) = \Upsilon(0) = 0$ by definition.

Now we introduce the descriptive processes which will be used to measure the performance of the networks. Let $A_k(t)$ be the number of class-k effective arrivals in interval $[0,\, t]$, $k = 1, 2, (1, 2)$. More exactly, $A_1(t)$ equals the number of class-1 arrivals which *did not jump to station 2 in interval* $[0,\, t]$. Thus, the total number of (external) arrivals in $[0,\, t]$ to the 1st station equals $A_1(t) + A_{1,2}(t)$. Let $D_k(t)$ be the number of class-k departures (from the system) in interval $[0,\, t]$, and let $Z_k(t)$ be the number of class-k customers in the network (in queue or being

served) at time t, $k = 1, 2, (1,2)$. We assume that initial number of customers $Z(0) = (Z_1(0), Z_2(0), Z_{1,2}(0))$ is arbitrary and independent of all above given quantities. Note that $Z_{1,2}(t) \in \{0, 1\}$, $t \geq 0$. Also let $T_k(t)$ be the total service time devoted to class-k customers in interval $[0, t]$, that is

$$T_1(t) = \int_0^t 1_{\{Z_1(u)\geq 1\}}\, du, \quad T_2(t) = \int_0^t 1_{\{Z_2(u)\geq 1\,,\, Z_{1,2}(u)=0\}}\, du,$$
$$T_{1,2}(t) = \int_0^t 1_{\{Z_{1,2}(u)=1\}}\, du.$$

where $1_{\{A\}}$ denotes indicator function. For $j = 1, 2$, let $W_j(t)$ be the (remaining) workload at station j at instant t, and $Y_j(t)$ be the idle time of server j in interval $[0, t]$. Now we consider the following processes

$$\begin{aligned}
A &= \{A(t) := (A_1(t), A_2(t), A_{1,2}(t)),\, t \geq 0\},\\
D &= \{D(t) := (D_1(t), D_2(t), D_{1,2}(t)),\, t \geq 0\},\\
Z &= \{Z(t) := (Z_1(t), Z_2(t), Z_{1,2}(t)),\, t \geq 0\},\\
T &= \{T(t) := (T_1(t), T_2(t), T_{1,2}(t)),\, t \geq 0\},\\
W &= \{W(t) := (W_1(t), W_2(t)),\, t \geq 0\},\\
Y &= \{Y(t) := (Y_1(t), Y_2(t)),\, t \geq 0\}.
\end{aligned}$$

(In general, these processes depend on the accepted service discipline.) Also processes A, D, T and Y are nondecreasing and satisfy initial conditions $A(0) = D(0) = T(0) = Y(0) = 0$. Moreover, the following *queueing network equations* hold for $t \geq 0$:

$$A_1(t) = E_1(t) - A_{1,2}(t)\,, \quad A_2(t) = E_2(t), \tag{5}$$
$$Z(t) = Z(0) + A(t) - D(t)\,, \tag{6}$$
$$D_k(t) = S_k\,(T_k(t)), \quad k = 1, 2, (1,2)\,, \tag{7}$$
$$T_1(t) + Y_1(t) = t\,, \quad (T_2(t) + T_{1,2}(t)) + Y_2(t) = t\,, \tag{8}$$
$$\int_0^\infty W_j(t)\, d\, Y_j(t) = 0 \quad j = 1, 2\,, \tag{9}$$
$$W_1(t) = \Upsilon_1(Z_1(0) + A_1(t)) - T_1(t)\,, \tag{10}$$
$$\begin{aligned} W_2(t) &= \Upsilon_2(Z_2(0) + A_2(t))\\ &\quad + \Upsilon_{1,2}(Z_{1,2}(0) + A_{1,2}(t)) - (T_2(t) + T_{1,2}(t))\,, \end{aligned} \tag{11}$$
$$D_{1,2}(t) \leq A_{1,2}(t) \leq D_{1,2}(t) + 1. \tag{12}$$

Equations (5)–(11) are self-explained, while the inequalities (12) hold because $Z_{1,2}(\cdot) \in \{0, 1\}$. By HL service discipline, we have additional relations

$$\Upsilon_k(D_k(t)) \leq T_k(t) < \Upsilon_k(D_k(t) + 1)\,, \quad k = 1, 2, (1,2), \quad t \geq 0. \tag{13}$$

Below we consider the following (basic) Markov process $\Psi := \big(A,\, D,\, T,\, Z,\, W,\, Y\big)$.

Remark 2. The previous equations do not specify the service discipline. For each particular HL discipline there exists at least one more queueing network equation. For instance, since server at station 2 cannot be idle if there are customers waiting in buffer 1 the following additional equation (not use below, however) holds

$$\int_0^\infty \nu_1(t)\, d\, Y_2(t) = 0,$$

where $\nu_1(t)$ is the number of customers in buffer 1 at instant t.

4. Stability analysis of the two-station cascade queueing network

By definition, a queueing network is *stable* if its associated underlying Markov process is *positive Harris recurrent* (see [7], [9] for details). The main criterion of the positive Harris recurrence is the limit Theorem 3.1 in [9], which borrows from Theorem 2.1(ii) [19]. To make it practical, we will apply the standard *fluid approximation*, which is based on the *fluid limit model* associated with the queueing network. It is known that the queueing network is stable (that is, the limit in Theorem 3.1 [9] holds) whenever the corresponding fluid limit model is stable (see Theorem 4.2 in [9]). (The definition of stability is given in Definition 3 below.) Thus, to prove stability of the network, it is enough to establish stability of the associated fluid limit model.

Without loss of generality, we may assume that the fluid limits of the initial remaining inter-arrival times and the initial remaining service times (for each class of customers) both are zero (see [8] or Lemma 5.3 in [9]).

4.1. The fluid limit model.

Next proposition is similar to Theorem 4.1 [9] (see also Lemma 3.1 in [11]).

Lemma 1. *For almost all sample paths and any sequence of initial states* $\{z_n\} \subset \mathbb{Z}_+^3$ *with* $\lim_{n\to\infty} |z_n| = \infty$, *there are a subsequence* $\{z_{n_j}\} \subseteq \{z_n\}$ *with* $\lim_{j\to\infty} |z_{n_j}| = \infty$, *a vector* $\bar z \in \mathbb{R}_+^3$, *and a stochastic process* $\bar\Psi = (\bar A,\, \bar D,\, \bar T,\, \bar Z,\, \bar W,\, \bar Y)$ *such that the following u.o.c. fluid limit exists*

$$\lim_{j\to\infty} \frac{1}{|z_{n_j}|}\, \Psi^{z_{n_j}}(|z_{n_j}|t) := \bar\Psi(t), \quad t \ge 0, \tag{14}$$

where $\bar Z(0) = \bar z$. *Furthermore, the components of* $\bar\Psi$ *satisfy the following (deterministic) equations:*

$$\bar A_1(t) = \lambda_1\, t - \bar A_{1,2}(t)\,, \quad \bar A_2(t) = \lambda_2\, t, \tag{15}$$

$$\bar Z(t) = \bar z + \bar A(t) - \bar D(t), \tag{16}$$

$$\bar D(t) = M^{-1}\, \bar T(t)\,, \tag{17}$$

$$\bar{T}_1(t) + \bar{Y}_1(t) = t\,, \quad (\bar{T}_2(t) + \bar{T}_{1,2}(t)) + \bar{Y}_2(t) = t\,, \tag{18}$$

$$\int_0^\infty \bar{W}_j(t)\, d\,\bar{Y}_j(t) = 0 \quad j = 1,\, 2\,, \tag{19}$$

$$\bar{W}_1(t) = m_1\,(\bar{z}_1 + \bar{A}_1(t)) - \bar{T}_1(t)\,, \tag{20}$$

$$\bar{W}_2(t) = m_2\,(\bar{z}_2 + \bar{A}_2(t)) + m_{1,2}\,(\bar{z}_{1,2} + \bar{A}_{1,2}(t)) - (\bar{T}_2(t) + \bar{T}_{1,2}(t))\,, \tag{21}$$

$$\bar{A}_{1,2}(t) = \bar{D}_{1,2}(t)\,. \tag{22}$$

In addition, for any $0 \le s \le t$,

$$0 \le \bar{T}_k(t+s) - \bar{T}_k(s) \le t, \;\; k = 1,\, 2,\, (1,2)\,. \tag{23}$$

Proof. As in [9], we apply the (functional) Strong Law of Large Numbers to renewal processes to obtain (u.o.c) limits (for $i = 1, 2$ and $k = 1, 2, (1, 2)$):

$$E_i^{z_{n_j}}(|z_{n_j}|t)/|z_{n_j}| \longrightarrow \lambda_i t,\; S_k^{z_{n_j}}(|z_{n_j}|t)/|z_{n_j}| \longrightarrow \\ \mu_k\, t, \Upsilon_k([|z_{n_j}|\, t])/|z_{n_j}|| \longrightarrow m_k\, t,\; j \longrightarrow \infty,$$

w. p. 1. Then equations (15), (16), (18)–(21) follow directly from (5), (6), (8)–(11), respectively. Moreover, (17) and (22) follow from (7), (12) and the convergence

$$\lim_{j\to\infty} \frac{D_k^{z_{n_j}}(|z_{n_j}|\, t)}{|z_{n_j}|} = \lim_{j\to\infty} \frac{S_k^{z_{n_j}}\left(T_k^{z_{n_j}}(|z_{n_j}|\, t)\right)}{|z_{n_j}|} = \mu_k\, \bar{T}_k(t) = \bar{D}_k(t)\,. \tag{24}$$

Note that the existence of the process $\bar{T} = \{\bar{T}(t)\}$ with

$$\bar{T}(t) := \lim_{j\to\infty} \frac{T^{z_{n_j}}(|z_{n_j}|\, t)}{|z_{n_j}|} \text{ (u.o.c.)}$$

is based on the inequalities $T^{z_{n_j}}(|z_{n_j}|\, t) \le |z_{n_j}|\, t$ and

$$\frac{T^{z_{n_j}}(|z_{n_j}|\, t)}{|z_{n_j}|} - \frac{T^{z_{n_j}}(|z_{n_j}|\, s)}{|z_{n_j}|} \le t - s, \quad 0 \le s \le t$$

(see also [9]). We also note that equation (17) comes from (7) (by (24)), or from (13). □

Remark 3. *Fluid model equations* (15)-(22) may not have in general a unique solution and can be treated as the *"limit"* of the corresponding queueing network equations (see Section 4.3 in [7]). Any limit $\left(\bar{z},\, \bar{\Psi}\right)$ in (14) is called a *fluid limit* associated to the queueing network (see [9]). Thus, Proposition 1 states that any *fluid limit* satisfies the *fluid model equations* (15)-(22), and also (23).

Definition ([7], [9]). The *fluid limit model* associated to a queueing network is *stable* if the component $\bar{Z}$ of any fluid limit $\left(\bar{z},\, \bar{\Psi}\right)$ satisfies condition

$$\bar{Z}(t) = 0 \text{ for all } t \geq t_1\,|\bar{z}| \tag{25}$$

for some constant $t_1 \geq 0$ (depending on the input and service rates only).

Because $z_n := (z_n^{(1)},\, z_n^{(2)},\, z_n^{(1,2)})$ with $z_{n_j}^{(1,2)} \in \{0,\, 1\}$, then w. p. 1

$$\bar{Z}_{1,2}(0) = \bar{z}_{1,2} = \lim_{j\to\infty} \frac{1}{|z_{n_j}|}\, Z_{1,2}^{z_{n_j}}(0) = \lim_{j\to\infty} \frac{1}{|z_{n_j}|}\, z_{n_j}^{(1,2)} = 0,$$

while $\lim_{j\to\infty} |z_{n_j}| = \infty$. By the same reason,

$$\bar{Z}_{1,2}(t) = 0,\ t \geq 0\,. \tag{26}$$

(The latter can also be obtained if we substitute $\bar{D}_{1,2}(t)$ from (16) to (15) and use equalities (22) and $\bar{z}_{1,2} = 0$.) Finally, (20), (21), (16), (17) and (26) imply

$$\bar{W}_1(t) = m_1\, \bar{Z}_1(t), \tag{27}$$

$$\bar{W}_2(t) = m_2\, \bar{Z}_2(t) + m_{1,2}\, \bar{Z}_{1,2}(t) = m_2\, \bar{Z}_2(t),\ \ t \geq 0. \tag{28}$$

Thus $\bar{Z}$ can be expressed in terms of $\bar{W}$ as

$$\bar{Z}_1(t) = \mu_1\, \bar{W}_1(t), \quad \bar{Z}_2(t) = \mu_2\, \bar{W}_2(t),\ t \geq 0\,. \tag{29}$$

This property has been introduced and studied in the context of the fluid limits in [12] and is known as *state space collapse* (in fluid limits). It is shown in [12] that expression (29), which is a kind of condition traditionally appearing in relation with heavy-traffic limit theorems, is a sufficient condition for the stability of the so-called subcritical multi-class queuing network with feedback and a work-conserving HL service discipline, provided that its associated linear Skorokhod problem is stable.

Remark 4. Relation (25) corresponds to a *strong* stability. Weaker notions may also be of interest (see, for instance, [8]). In particular, the *fluid limit model* associated with the queueing network is *weakly stable* if $\bar{Z}(t) \equiv 0$ when $\bar{z} = 0$. The *path-wise stability* means that w. p. 1, $\bar{D}(t) = \bar{A}(t) + \bar{z}$. It is obvious that (strong) stability implies both weak and path-wise stability.

4.2. The stability analysis. In this section we present a stability result for the original two-station cascade network. It has been mentioned in the Introduction, that stability condition (30) (in Theorem 1 below) was proved in [20] for the model with Poisson input to 1st station and a general renewal input to the 2nd station. At that the unboundedness and spread-outness of the inter-arrival times (to the 2nd station) are not required in [20]. The discrepancy is caused by the different approaches used. In particular, regenerative methodology used in [20] requires to construct synchronized regenerations of the merged input stream to the network. Note that the regenerative analysis in [20] covers arbitrary initial

conditions and allows to show that condition (30) is in fact a stability criterion (that is, not only a sufficient but a necessary condition for the stability of the corresponding model).

The main motivation to present here a new proof for the two-station model is that it then allows to extend analysis to $N-$station cascade networks in Sections 5, 6. (This proof, in our opinion, is also shorter and simpler than the corresponding proofs in [20] and in the earlier work [14], and does not use Poisson arrivals)

Theorem 1. *The two-station cascade queueing network is stable under conditions* (1)–(4) *and*

$$\frac{(\lambda_1 - \mu_1)^+}{\mu_{1,2}} + \frac{\lambda_2}{\mu_2} < 1. \tag{30}$$

The proof of Theorem 1 relies on the following statement (which is adaptation of Lemma 5.1 [9] to our setting).

Lemma 2. *Let (W, Y) be the solution of the Skorokhod problem associated to process X with reflection matrix I and an arbitrary (fixed) initial state $X(0) = x$ on the orthant $\mathbb{R}^2_+$, and let $s > 0$ be fixed. Assume that*

$$W(s) + X(t+s) - X(s) \geq \theta\, t \quad \textit{for all } t \geq 0\,, \tag{31}$$

where $\theta := (\theta_1, \theta_2) < 0$. Then for any regular point t,

$$\dot{Y}(t) \leq -\theta. \tag{32}$$

Proof. The proof uses Propositions 1 and 2 from [21]. We summarize these results for our setting as follows: if $(\hat{w}, \hat{y})$ is the solution of the Skorokhod problem associated to the process $\hat{x}$ with initial state $\hat{x}(0) = x_0$ and reflection matrix I, then $\hat{y}$ is the unique least element of the set

$$U(x) = \{\text{nondecreasing } y \geq 0 : x(t) + y(t) \geq 0,\ t \geq 0\}\,,$$

with $x(t) = \hat{x}(t) + x_0$. Now we take

$$X^0(t) := \theta\, t,\ Y^0(t) := -X^0(t) = -\theta\, t,\ t \geq 0.$$

Thus $Y^0(0) = 0$ and the process $Y^0 = \{Y^0(t)\}$ is nondecreasing. Since (W, Y) is the solution of the Skorokhod problem associated to process X with reflection matrix I and initial state x, then $W(t) = x + X(t) + Y(t)$. For a fixed $s > 0$, we can write for any $t \geq 0$

$$0 \leq W(t+s) = W(s) + X(t+s) - X(s) + Y(t+s) - Y(s) = \tilde{X}(t) + \tilde{Y}(t)\,,$$

where

$$\tilde{X}(t) = W(s) + X(t+s) - X(s),\ \ \tilde{Y}(t) = Y(t+s) - Y(s).$$

Therefore, $(W(\cdot+s), \tilde{Y}(\cdot))$ is the solution of the Skorokhod problem associated to process $\tilde{X}(\cdot) - \tilde{X}(0)$ with reflection matrix I and initial state $\tilde{X}(0) = W(s) \in \mathbb{R}^2_+$. As a consequence, $\tilde{Y}$ is the least element of $U(\tilde{X})$. On the other hand, $\tilde{X}(t) \geq \theta\, t = X^0(t),\ t \geq 0$, by (31). Then $\tilde{X}(t) + Y^0(t) \geq X^0(t) + Y^0(t) = 0$, and this implies $Y^0 \in U(\tilde{X})$. As a consequence, $\tilde{Y}(t) \leq Y^0(t)$, $t \geq 0$, or

$$Y(t+s) - Y(s) \leq -\theta\, t, \quad t \geq 0,$$

for each fixed $s > 0$. Thus (32) follows. □

Proof of Theorem 1. We split the proof into three main steps.
First step. Fix a fluid limit $(\bar{z},\, \bar{\Psi})$.

It is easy to obtain from (16) - (18) and (26), that

$$\bar{Z}_1(t) = \bar{z}_1 + \bar{A}_1(t) - \mu_1\, t + \mu_1\, \bar{Y}_1(t), \tag{33}$$

$$\bar{Z}_2(t) = \bar{z}_2 + \bar{A}_2(t) - \mu_2\, t + \mu_2\, \bar{Y}_2(t) + \mu_2\, \bar{T}_{1,2}(t), \tag{34}$$

$$\bar{Z}_{1,2}(t) = \bar{z}_{1,2} + \bar{A}_{1,2}(t) - \mu_{1,2}\, \bar{T}_{1,2}(t) = \bar{A}_{1,2}(t) - \mu_{1,2}\, \bar{T}_{1,2}(t)\,. \tag{35}$$

Since $\bar{Z}_{1,2}(t) = 0$ then (34),(35) give

$$\bar{Z}_2(t) = \bar{z}_2 + \bar{A}_2(t) + r_2\, \bar{A}_{1,2}(t) - \mu_2\, t + \mu_2\, \bar{Y}_2(t)\,, \tag{36}$$

where $r_2 := \mu_2/\mu_{1,2}$. Using (27), (28), (33) and (36) we obtain

$$\bar{W}_1(t) = m_1\, \bar{z}_1 + m_1\, \bar{A}_1(t) - t + \bar{Y}_1(t), \tag{37}$$

$$\bar{W}_2(t) = m_2\, \bar{z}_2 + m_2\, \bar{A}_2(t) + m_{1,2}\, \bar{A}_{1,2}(t) - t + \bar{Y}_2(t). \tag{38}$$

Define $\bar{w} := (\bar{w}_1,\, \bar{w}_2) = (m_1\, \bar{z}_1,\, m_2\, \bar{z}_2)$ and the following process $X = (X_1,\, X_2)$:

$$X_1(t) = m_1\, \bar{A}_1(t) - t, \tag{39}$$

$$\begin{aligned} X_2(t) &= m_2\, \bar{A}_2(t) + m_{1,2}\, \bar{A}_{1,2}(t) - t \\ &= (m_2\, \lambda_2 - 1)\, t + m_{1,2}\, \lambda_1\, t - m_{1,2}\, \bar{A}_1(t). \end{aligned} \tag{40}$$

(In the 2nd equality in (40) equation (15) is used.) Then (37) and (38) imply

$$\bar{W}(t) = \bar{w} + X(t) + \bar{Y}(t). \tag{41}$$

It is easy to check that $(\bar{W},\, \bar{Y})$ is the solution of the Skorokhod problem associated to X with reflection matrix I and initial state $\bar{w} \in \mathbb{R}^2_+$.

Second step. Define vector $\theta = (\theta_1,\, \theta_2)$ as

$$\begin{cases} \theta_1 = m_1 \big(\lambda_1 - (\mu_1 + \mu_{1,2})\big), \\ \theta_2 = m_2 \big(\lambda_2 - \mu_2\big). \end{cases}$$

To show that $\theta < 0$, we first note that $\theta_2 < 0$ immediately by (30). Besides, if $\lambda_1 \leq \mu_1$, then $\lambda_1 < \mu_1 + \mu_{1,2}$ implying $\theta_1 < 0$, while if $\lambda_1 > \mu_1$, then (30) can be written as $\lambda_1\, \mu_2 + \lambda_2\, \mu_{1,2} < \mu_2\, (\mu_1 + \mu_{1,2})$, and it again implies $\theta_1 < 0$. We

now prove that $\bar{W}(s) + X(t+s) - X(s) \geq \theta\, t$ for any $t \geq 0$ and any fixed $s > 0$. Indeed, using sequentially (39), (15), (22), (17) and (23) we obtain

$$\begin{aligned}
&\bar{W}_1(s) + X_1(t+s) - X_1(s) \\
&\qquad = \bar{W}_1(s) + m_1\, \bar{A}_1(t+s) - (t+s) - m_1\, \bar{A}_1(s) + s \\
&\qquad = \bar{W}_1(s) + m_1\, (\bar{A}_1(t+s) - \bar{A}_1(s)) - t \\
&\qquad = \bar{W}_1(s) + (m_1\, \lambda_1 - 1)\, t - m_1\, (\bar{A}_{1,2}(t+s) - \bar{A}_{1,2}(s)) \\
&\qquad = \bar{W}_1(s) + (m_1\, \lambda_1 - 1)\, t - m_1\, (\bar{D}_{1,2}(t+s) - \bar{D}_{1,2}(s)) \\
&\qquad = \bar{W}_1(s) + (m_1\, \lambda_1 - 1)\, t - m_1\, \mu_{1,2}\, (\bar{T}_{1,2}(t+s) - \bar{T}_{1,2}(s)) \\
&\qquad \geq (m_1\, \lambda_1 - 1)\, t - m_1\, \mu_{1,2}\, t = \theta_1\, t\,,
\end{aligned}$$

since $\bar{W}_1(s) \geq 0$. For the second component we similarly obtain

$$\begin{aligned}
&\bar{W}_2(s) + X_2(t+s) - X_2(s) \\
&\quad = \bar{W}_2(s) + (m_2\, \lambda_2 - 1)\, (t+s) + m_{1,2}\, \bar{A}_{1,2}(t+s) - (m_2\, \lambda_2 - 1)\, s - m_{1,2}\, \bar{A}_{1,2}(s) \\
&\quad \geq (m_2\, \lambda_2 - 1)\, t + m_{1,2}\, (\bar{A}_{1,2}(t+s) - \bar{A}_{1,2}(s)) \geq \theta_2\, t\,,
\end{aligned}$$

since $\bar{W}_2(s) \geq 0$ and $\bar{A}_{1,2}(t+s) - \bar{A}_{1,2}(s) \geq 0$. Therefore, Lemma 1 can be applied to $(\bar{W},\, \bar{Y})$, and for any regular point s of $\bar{Y}$ we obtain

$$\dot{\bar{Y}}(s) + \theta \leq 0\,. \tag{42}$$

Third step. We take the Lyapunov function in the following from

$$f(t) = \frac{m_{1,2}}{m_1}\, \bar{W}_1(t) + \bar{W}_2(t), \quad t \geq 0\,, \tag{43}$$

which can be written as (see (37), (38)):

$$\begin{aligned}
f(t) &= \frac{m_{1,2}}{m_1}\, \bar{w}_1 + m_{1,2}\, \bar{A}_1(t) - \frac{m_{1,2}}{m_1}\, t + \frac{m_{1,2}}{m_1}\, \bar{Y}_1(t) \\
&\quad + \bar{w}_2 + m_2\, \lambda_2\, t + m_{1,2}\, (\lambda_1\, t - \bar{A}_1(t)) - t + \bar{Y}_2(t) \\
&= f(0) + (m_2\, \lambda_2 - 1)\, t + \frac{m_{1,2}}{m_1}\, (m_1\, \lambda_1 - 1)\, t + \frac{m_{1,2}}{m_1}\, \bar{Y}_1(t) + \bar{Y}_2(t) \\
&= f(0) + \frac{m_{1,2}}{m_1}\, (\theta_1\, t + \bar{Y}_1(t)) + (\theta_2\, t + \bar{Y}_2(t)) + t\,,
\end{aligned}$$

where equality $(m_1\, \lambda_1 - 1) m_{1,2}/m_1 = \theta_1\, m_{1,2}/m_1 + 1$ is used. Thus, $\bar{Y}$ and f have the same points of differentiability, which verify

$$\dot{f}(t) = \frac{m_{1,2}}{m_1}\, (\theta_1 + \dot{\bar{Y}}_1(t)) + (\theta_2 + \dot{\bar{Y}}_2(t)) + 1\,. \tag{44}$$

Let $t > 0$ be such a point of differentiability and such that $f(t) > 0$ (if any). Therefore $\max(\bar{W}_1(t),\, \bar{W}_2(t)) > 0$, and (19) implies that $\min(\dot{\bar{Y}}_1(t),\, \dot{\bar{Y}}_2(t)) = 0$.

If $\dot{\bar{Y}}_1(t) = 0$ then it follows from the inequality $\theta_2 + \dot{\bar{Y}}_2(t) \le 0$ (see (42)) and from (44) that

$$\dot{f}(t) \le \frac{m_{1,2}}{m_1}\,\theta_1 + 1\,.$$

Analogously, if $\dot{\bar{Y}}_2(t) = 0$, then

$$\dot{f}(t) \le \theta_2 + 1\,.$$

Thus, at any regular point $t > 0$ with $f(t) > 0$,

$$\begin{aligned}\dot{f}(t) \le \frac{m_{1,2}}{m_1}\,\theta_1 + \theta_2 + 1 &= (m_2\,\lambda_2 - 1) + \frac{m_{1,2}}{m_1}\,(m_1\,\lambda_1 - 1)\\ &= m_{1,2}\left(\frac{\lambda_2 - \mu_2}{r_2} + \lambda_1 - \mu_1\right) := -C < 0.\end{aligned}$$

By assumption (30), $\dot{f}(t) < 0$. Then, by Lemma 5.2 [9] f is non-increasing, and $f(t) = 0$ for $t \ge \frac{f(0)}{C}$. Moreover,

$$f(0) = \frac{m_{1,2}}{m_1}\,\bar{w}_1 + \bar{w}_2 = m_{1,2}\,\bar{z}_1 + m_2\,\bar{z}_2 \le (m_2 + m_{1,2})\,|\bar{z}|\,. \tag{45}$$

Eventually, it follows from (29), (26), (45) that

$$\bar{Z}(t) = 0 \quad \text{for } t \ge t_1\,|\bar{z}|\,, \quad \text{where } t_1 = \frac{m_2 + m_{1,2}}{C}.$$

Thus (25) holds. □

5. The type-I cascade network

We first consider a three-station type-I cascade network, where station 3 helps station 2, which, in turn, helps station 1, as showed in Fig. 2. The analysis of this network is a useful intermediate step which allows then extend easily the above developed stability analysis to N-station cascade networks with arbitrary N.

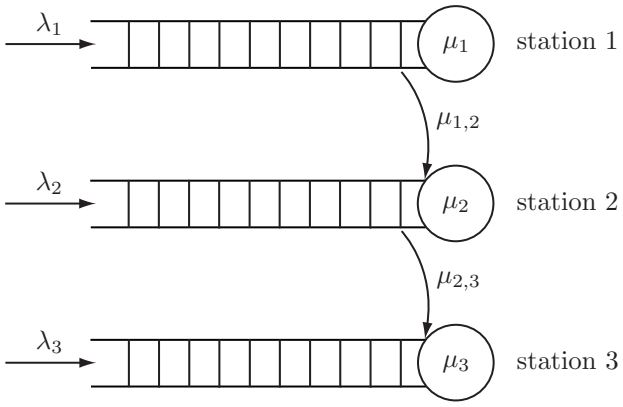

Figure 2. A type-I cascade network with three stations

We consider class-i customers, arriving to station i from outside $i = 1, 2, 3$, and class-$(i-1, i)$ customers jumping (one at a time) from queue $i-1$ to (empty) station $i = 2, 3$. Let λ_i be the arrival rate of the (exogenous) class-i customers with mean service time m_i and service rate $\mu_i = 1/m_i$, $i = 1, 2, 3$. Also let $m_{i-1,\,i}$ be the mean service time of class-$(i-1,\, i)$ customers with rate $\mu_{i-1,\,i} = 1/m_{i-1,\,i}$, $i = 2, 3$. In what follows, we use double index $i-1, i$ to denote the quantities related to class-$(i-1,\, i)$ customers. The total service time devoted to customers of each class by time t is defined, respectively, as

$$
\begin{aligned}
T_1(t) &= \int_0^t 1_{\{Z_1(u)\geq 1\}}\, du\,, \quad T_2(t) = \int_0^t 1_{\{Z_2(u)\geq 1\,,\; Z_{1,2}(u)=0\}}\, du\,,\\
T_3(t) &= \int_0^t 1_{\{Z_3(u)\geq 1\,,\; Z_{2,3}(u)=0\}}\, du\,, \quad T_{1,2}(t) = \int_0^t 1_{\{Z_{1,2}(u)=1\}}\, du\,,\\
T_{2,3}(t) &= \int_0^t 1_{\{Z_{2,3}(u)=1\}}\, du\,. \qquad (46)
\end{aligned}
$$

Then the network equations associated to this model are similar to (5)-(12) with (5) replaced by

$$
A_1(t) = E_1(t) - A_{1,2}(t)\,, \quad A_2(t) = E_2(t) - A_{2,3}(t)\,, \quad A_3(t) = E_3(t). \qquad (47)
$$

Moreover, (8), (11) and (12) should be replaced, respectively by

$$
\begin{aligned}
&T_1(t) + Y_1(t) = t\,, \quad (T_2(t) + T_{1,2}(t)) + Y_2(t) = t\,, \quad (T_3(t) + T_{2,3}(t)) + Y_3(t) = t\,,\\
&W_2(t) = \Upsilon_2(Z_2(0) + A_2(t)) + \Upsilon_{1,2}(Z_{1,2}(0) + A_{1,2}(t)) - (T_2(t) + T_{1,2}(t))\\
&W_3(t) = \Upsilon_3(Z_3(0) + A_3(t)) + \Upsilon_{2,3}(Z_{2,3}(0) + A_{2,3}(t)) - (T_3(t) + T_{2,3}(t))\,,\\
&D_{1,2}(t) \leq A_{1,2}(t) \leq D_{1,2}(t) + 1\,, \quad D_{2,3}(t) \leq A_{2,3}(t) \leq D_{2,3}(t) + 1\,. \qquad (48)
\end{aligned}
$$

(Note that $Z_{1,\,2}(0)$, $Z_{2,\,3}(0) \in \{0,\, 1\}$.) The associated fluid equations are analogous to (15)-(22) *mutatis mutandis*. In particular, (15) becomes now

$$
\bar{A}_1(t) = \lambda_1\, t - \bar{A}_{1,2}(t)\,, \quad \bar{A}_2(t) = \lambda_2\, t - \bar{A}_{2,3}(t)\,, \quad \bar{A}_3(t) = \lambda_3\, t\,. \qquad (49)
$$

Moreover, equations (21) and (22) are now, respectively,

$$
\begin{aligned}
\bar{W}_2(t) &= m_2\,(\bar{z}_2 + \bar{A}_2(t)) + m_{1,2}\,(\bar{z}_{1,2} + \bar{A}_{1,2}(t)) - (\bar{T}_2(t) + \bar{T}_{1,2}(t))\,,\\
\bar{W}_3(t) &= m_3\,(\bar{z}_3 + \bar{A}_3(t)) + m_{2,3}\,(\bar{z}_{2,3} + \bar{A}_{2,3}(t)) - (\bar{T}_3(t) + \bar{T}_{2,3}(t))\,,\\
\bar{A}_{1,2}(t) &= \bar{D}_{1,2}(t)\,, \quad \bar{A}_{2,3}(t) = \bar{D}_{2,3}(t)\,. \qquad (50)
\end{aligned}
$$

We observe now that for any $t \geq 0$,

$$
\begin{aligned}
&\bar{Z}_{1,2}(t) = \bar{Z}_{2,3}(t) = 0\,,\\
&\bar{A}_1(t) + \bar{A}_{1,2}(t) = \lambda_1\, t\,, \quad \bar{A}_2(t) + \bar{A}_{2,3}(t) = \lambda_2\, t\,, \quad \bar{A}_3(t) = \lambda_3\, t, \qquad (51)\\
&\bar{W}_1(t) = m_1\, \bar{Z}_1(t)\,, \quad \bar{W}_2(t) = m_2\, \bar{Z}_2(t)\,, \quad \bar{W}_3(t) = m_3\, \bar{Z}_3(t)\,.
\end{aligned}
$$

CRM Preprint Series number 1119

To see a consistency of the following statement with Theorem 1, we rewrite condition (30) as the following inequalities

$$\lambda_2 < \mu_2\,, \quad \lambda_1 + \frac{\lambda_2}{r_2} < \mu_1 + \frac{\mu_2}{r_2}\,, \tag{52}$$

where $r_2 := \mu_2/\mu_{1,2}$. These inequalities are the analogue of conditions (55), (56) below where $r_3 := \mu_3/\mu_{2,3}$. Note that the second condition in (52) implies $\lambda_1 < \mu_1 + \frac{\mu_2}{r_2}$, which corresponds to conditions (53), (54) below.

Theorem 2. *Assume that conditions* (1)-(4) *hold for all (exogenous) inputs and all service times in the three-station type-I cascade network. Then sufficient stability conditions for this network are:*

$$\lambda_1 < \mu_1 + \frac{\mu_2}{r_2}, \tag{53}$$

$$\lambda_2 < \mu_2 + \frac{\mu_3}{r_3}, \tag{54}$$

$$\lambda_3 < \mu_3, \tag{55}$$

$$\lambda_1 + \frac{\lambda_2}{r_2} + \frac{\lambda_3}{r_2\, r_3} < \mu_1 + \frac{\mu_2}{r_2} + \frac{\mu_3}{r_2\, r_3}. \tag{56}$$

Proof. The proof of Theorem 2 is close to the proof of Theorem 1, and we only outline such points where any significant difference exists. As in Step 1, we can write $W(t) = \bar{w} + X(t) + \bar{Y}(t)$, where $X(t) := (X_1(t),\, X_2(t),\, X_3(t))$ with

$$\begin{aligned}
X_1(t) &= m_1\, \bar{A}_1(t) - t\,,\\
X_2(t) &= m_2\, \bar{A}_2(t) + m_{1,2}\, \bar{A}_{1,2}(t) - t = m_2\, \bar{A}_2(t) - m_{1,2}\, \bar{A}_1(t) + (m_{1,2}\, \lambda_1 - 1)\, t\,,\\
X_3(t) &= m_3\, \bar{A}_3(t) + m_{2,3}\, \bar{A}_{2,3}(t) - t = (m_3\, \lambda_3 - 1)\, t + m_{2,3}\, \lambda_2\, t - m_{2,3}\, \bar{A}_2(t)\,,
\end{aligned}$$

and $(\bar{W},\, \bar{Y})$ is the solution of the Skorokhod problem associated to X with reflection matrix I and initial state $\bar{w} = (m_1\, \bar{z}_1,\, m_2\, \bar{z}_2,\, m_3\, \bar{z}_3) \in \mathbb{R}^3_+$. The second step is done similarly by introducing $\theta = (\theta_1,\, \theta_2,\, \theta_3)$ as

$$\begin{cases}
\theta_1 = m_1\, \big(\lambda_1 - (\mu_1 + \mu_{1,2})\big),\\
\theta_2 = m_2\, \big(\lambda_2 - (\mu_2 + \mu_{2,3})\big),\\
\theta_2 = m_3\, \big(\lambda_3 - \mu_3\big).
\end{cases}$$

Then it follows from assumptions (53)-(56), that $\theta < 0$. Note also that

$$\begin{aligned}
0 \le \bar{A}_{1,2}(t+s) - \bar{A}_{1,2}(s) &= \mu_{1,2}\, \big(\bar{T}_{1,2}(t+s) - \bar{T}_{1,2}(s)\big) \le \mu_{1,2}\, t,\\
0 \le \bar{A}_{2,3}(t+s) - \bar{A}_{2,3}(s) &= \mu_{2,3}\, \big(\bar{T}_{2,3}(t+s) - \bar{T}_{2,3}(s)\big) \le \mu_{2,3}\, t,
\end{aligned}$$

Finally, we define the Lyapunov function as

$$f(t) = \frac{m_{2,3}}{m_2}\, \frac{m_{1,2}}{m_1}\, \bar{W}_1(t) + \frac{m_{2,3}}{m_2}\, \bar{W}_2(t) + \bar{W}_3(t), \quad t \ge 0\,,$$

which can be expressed similarly to the third step in the proof of Theorem 1 as

$$f(t) = f(0) + \frac{m_{2,3}}{m_2}\,\frac{m_{1,2}}{m_1}\,(\theta_1\,t + \bar{Y}_1(t)) + \frac{m_{2,3}}{m_2}\,(\theta_2\,t + \bar{Y}_2(t)) + (\theta_3\,t + \bar{Y}_3(t)) + \frac{m_{2,3}}{m_2}t\,.$$

Then by Lemma 5.2 [9], function f is non-increasing and $f(t) = 0$ for $t \geq \frac{f(0)}{C}$ with

$$C := -r_2\,m_{2,3}\left(\lambda_1 - \mu_1 + \frac{\lambda_2 - \mu_2}{r_2} + \frac{\lambda_3 - \mu_3}{r_2\,r_3}\right). \tag{57}$$

It follows from assumption (56) that $C > 0$. □

Previous analysis can be directly extended to a general type-I N-station cascade system.

Denote $r_i = \mu_i/\mu_{i-1,\,i}$ and $R_i = \prod_{\ell=2}^{i} r_\ell,\ i = 2,\,\ldots,\,N$, and let $R_1 := 1$. The following statement is an immediate extension of Theorem 2, and by this reason is given with no proof.

Theorem 3. *Assume that assumptions* (1)-(4) *hold for all exogenous inputs and all service times in the type-I N-station cascade queueing network. Then sufficient stability conditions for this network are:*

$$\begin{aligned} \lambda_i &< \mu_i + \frac{\mu_{i+1}}{r_{i+1}},\ i = 1, \ldots, N-1, \\ \lambda_N &< \mu_N, \\ \sum_{i=1}^{N} \frac{\lambda_i}{R_i} &< \sum_{i=1}^{N} \frac{\mu_i}{R_i}. \end{aligned} \tag{58}$$

6. The type-II cascade network

In this section, we extend the stability analysis to a type-II cascade network with N stations. In this model, if a station j becomes idle, $j = 2, \ldots, N$, it can help any of the non-empty upstream queues $i = 1, \ldots, j-1$ (as above, allowing at most one such a class-$(i,\,j)$ customer to be at station j at any instant). First, we again study the stability of three-station network in detail, and then formulate stability result for any number of stations $N \geq 2$ with no proof. As depicted in Fig. 3, we denote by λ_i the arrival rate of class-i (exogenous) customers with mean service time m_i and service rate $\mu_i = m_i^{-1}$. Also we keep notation $m_{i,j}$ and $\mu_{i,j}$ for mean service time and service rate, respectively, of $(i,\,j)$-class customers.

The cumulative service time devoted to each customer class $1, 2, (1,\,2)$ and $(2, 3)$ in interval $[0,\,t]$ are defined as above (see (46)), while for classes 3 and $(1, 3)$ they are defined by

$$T_3(t) = \int_0^t 1_{\{Z_3(u)\geq 1\,,\ Z_{2,3}(u)=0,\,Z_{1,3}(u)=0\}}\,du,$$

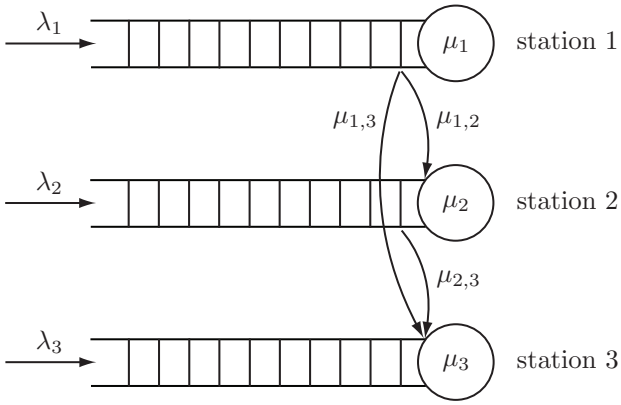

FIGURE 3. A type-II cascade network with three stations

$$T_{1,3}(t) = \int_0^t 1_{\{Z_{1,3}(u)=1\}}\, du\,. \tag{59}$$

In the used above notation, the network equations associated to this model are similar to the modifications of equations (5)-(12) considered in section 5. For instance, instead of (47) we have now

$$A_1(t) = E_1(t) - (A_{1,2}(t) + A_{1,3}(t))\,, \quad A_2(t) = E_2(t) - A_{2,3}(t)\,, \quad A_3(t) = E_3(t). \tag{60}$$

Moreover, corresponding expressions in (48) related to station 3 should be, respectively, replaced by

$$\Big(T_3(t) + T_{2,3}(t) + T_{1,3}(t)\Big) + Y_3(t) = t,$$

$$W_3(t) = \Upsilon_3\Big(Z_3(0) + A_3(t)\Big) + \Upsilon_{2,3}\Big(Z_{2,3}(0) + A_{2,3}(t)\Big) + \Upsilon_{1,3}\Big(Z_{1,3}(0) + A_{1,3}(t)\Big) - \Big(T_3(t) + T_{2,3}(t) + T_{1,3}(t)\Big)\,,$$

$$D_{1,3}(t) \le A_{1,3}(t) \le D_{1,3}(t) + 1\,. \tag{61}$$

The associated fluid equations are analogous to that of Section 5, *mutatis mutandis*, taking into account that

$$\bar{A}_1(t) = \lambda_1\, t - (\bar{A}_{1,2}(t) + \bar{A}_{1,3}(t))\,, \quad \bar{A}_2(t) = \lambda_2\, t - \bar{A}_{2,3}(t)\,, \quad \bar{A}_3(t) = \lambda_3\, t\,,$$

$$\bar{W}_3(t) = m_3\,(\bar{z}_3 + \bar{A}_3(t)) + m_{2,3}\,(\bar{z}_{2,3} + \bar{A}_{2,3}(t)) + m_{1,3}\,(\bar{z}_{1,3} + \bar{A}_{1,3}(t)) - (\bar{T}_3(t) + \bar{T}_{2,3}(t) + \bar{T}_{1,3}(t)),$$

$$\bar{A}_{1,3}(t) = \bar{D}_{1,3}(t)\,.$$

We observe that for any $t \ge 0$,

$$\bar{Z}_{1,2}(t) = \bar{Z}_{2,3}(t) = \bar{Z}_{1,3}(t) = 0\,,$$

$$\bar{A}_1(t) + \bar{A}_{1,2}(t) + \bar{A}_{1,3}(t) = \lambda_1\, t,\ \bar{A}_2(t) + \bar{A}_{2,3}(t) = \lambda_2\, t,\ \bar{A}_3(t) = \lambda_3\, t,$$

$$\bar{W}_1(t) = m_1\,\bar{Z}_1(t)\,, \quad \bar{W}_2(t) = m_2\,\bar{Z}_2(t)\,, \quad \bar{W}_3(t) = m_3\,\bar{Z}_3(t)\,.$$

As before, we denote $r_i = \mu_i/\mu_{i-1,i}$, $i = 2, 3$. Only to keep analysis simple, in the following stability statement we impose an extra condition (62) in comparison with type-I network.

Theorem 4. *Assume that in the three-station type-II cascade network, conditions* (1)-(4) *hold for all exogenous inputs and all service times, and moreover,*

$$r_2 = \frac{\mu_{2,3}}{\mu_{1,3}}. \tag{62}$$

Then the network is stable if conditions (54)-(56) *hold.*

Proof. The proof of Theorem 4 is similar to that of Theorem 2, and thus we will only discuss the main differences. First, we obtain $\bar{W}(t) = \bar{w} + X(t) + \bar{Y}(t)$, where

$$\begin{aligned}
&X_1(t) = m_1\,\bar{A}_1(t) - t\,, \quad X_2(t) = m_2\,\bar{A}_2(t) + m_{1,2}\,\bar{A}_{1,2}(t) - t,\\
&X_3(t) = m_3\,\bar{A}_3(t) + m_{2,3}\,\bar{A}_{2,3}(t) + m_{1,3}\,\bar{A}_{1,3}(t) - t\\
&\qquad = (m_3\,\lambda_3 - 1)\,t + (m_{2,3}\,\lambda_2 + m_{1,3}\,\lambda_1)\,t - m_{2,3}\,\bar{A}_2(t) - m_{1,3}\,(\bar{A}_1(t) + \bar{A}_{1,2}(t))\,.
\end{aligned}$$

At the second step we take

$$\begin{cases}
\theta_1 = m_1\,\big(\lambda_1 - (\mu_1 + \mu_{1,2} + \mu_{1,3})\big),\\
\theta_2 = m_2\,\big(\lambda_2 - (\mu_2 + \mu_{2,3})\big),\\
\theta_2 = m_3\,\big(\lambda_3 - \mu_3\big).
\end{cases}$$

One can check that $\theta = (\theta_1,\, \theta_2,\, \theta_3) < 0$ by the assumptions of Theorem. Finally, the Lyapunov function is taken as

$$f(t) = \frac{m_{1,3}}{m_1}\,\bar{W}_1(t) + \frac{m_{2,3}}{m_2}\,\bar{W}_2(t) + \bar{W}_3(t), \quad t \geq 0\,,$$

which, by assumption (62), can be written exactly as in Theorem 2:

$$f(t) = \frac{m_{2,3}}{m_2}\,\frac{m_{1,2}}{m_1}\,\bar{W}_1(t) + \frac{m_{2,3}}{m_2}\,\bar{W}_2(t) + \bar{W}_3(t), \quad t \geq 0\,.$$

After a simple algebra we can rewrite function f as

$$\begin{aligned}
f(t) =& f(0) + \frac{m_{2,3}}{m_2}\,\frac{m_{1,2}}{m_1}\,(\theta_1\,t + \bar{Y}_1(t)) + \frac{m_{2,3}}{m_2}\,(\theta_2\,t + \bar{Y}_2(t))\\
&+ (\theta_3\,t + \bar{Y}_3(t)) + \big(2 + \frac{m_{2,3}}{m_2}\big)\,t,\ t \geq 0.
\end{aligned}$$

Then, by Lemma 5.2 [9], we obtain that function f is non-increasing and $f(t) = 0$ for $t \geq \frac{f(0)}{C}$, where constant C is given by (57). □

Remark 5. If $\mu_{1,3} = \mu_{2,3}$ and $\mu_{1,2} = \mu_2$, then equality (62) holds with $r_2 = 1$. In this case, the assumptions of Theorem 4 are weaker than the assumptions of Theorem 2. In this sense the three-station type-II cascade network is more efficient than the corresponding type-I network.

Above obtained result allows us to extend (with no proof) previous analysis to type-II N-station cascade system. For simplicity, we assume that the system is *homogeneous*, that is, regardless of class, all customers served at the same station have the same service rate. In other words, we assume that

$$\mu_j = \mu_{i,j},\ i = 1, \dots, j-1;\ j = 2, \dots, N. \tag{63}$$

Thus, in the homogeneous network $r_j \equiv 1$.

Theorem 5. *Assume that conditions* (1)-(4) *hold for all exogenous inputs and all service times in the homogenous type-II N-station cascade network. Then sufficient stability conditions for this network are:*

$$\lambda_i < \sum_{j=i}^{N} \mu_j\,, \quad i = 2, \dots, N,$$

$$\sum_{i=1}^{N} \lambda_i < \sum_{i=1}^{N} \mu_i. \tag{64}$$

Finally, we note (as above) that since in the *homogeneous* system $r_i = 1$, $i = 2, \dots, N$, then the assumptions of Theorem 5 are strictly weaker than the assumptions of Theorem 3. In other words, the homogeneous type-II N-station cascade system is more efficient than the corresponding homogeneous type-I system.

7. Conclusion

In this paper, we present a stability analysis of a wide class of cascade networks with two different types of interactions between servers: type-I cascade networks, where each free server i can assist to serve customers waiting in the preceding queue $i-1$ only, while in type-II networks, each server i can help any preceding queue $1, \dots, i-1$. We first study two- and three-station networks in detail. This analysis then allows to formulate with no proof the stability results for general N-station networks.

To analyze type-II cascade networks, we consider homogeneous network where service rate at a given server does not depend on customer class. Stability conditions obtained in the work are consistent with that have been found for two-station network by other methods in [14] and [20]. The key element of analysis is the construction of the relevant Lyapunov function, which then is used to establish positive recurrence of the basic Markov process describing stochastic

dynamics of the network. As the main tool, a solution of a Skorokhod problem is applied to fluid limits obtained, in turn, by application of functional strong law of large numbers to initial equations describing stochastic behavior of the network.

8. Acknowledgements

EM thanks CRM for nice conditions during his visit in 2011, when the main part of this research has been done.

References

[1] Agnihothri S.R., Mishra A.K., Simmons D.E. (2003). Workforce cross-training decisions in field service systems with two job types. *Journal of the Operational Research Society.* 54(4), 410–418.

[2] Ahghari M., Balcioglu B. (2009). Benefits of cross-training in a skill-based routing contact center with priority queues and impatient customers. *IIE Transactions* 41, 524–536.

[3] Ahn, H.-S., Duenyas, I., Zhang, Q.R. (2004). Optimal control of a flexible server. *Adv. Appl. Probab.,* 36, 139–170.

[4] Andradottir S., Ayhan H., Down G. D. (2003). Dynamic server allocation for queueing networks with flexible servers. *Operations Reserach*, 51(6), 952–968.

[5] Bell, S. L., R. J. Williams. (1999). Dynamic scheduling of a server system with two parallel servers: asymptotic optimality of a continuous review threshold policy in heavy traffic. *Proceedings of the 38 Conference on Decision and Control*, Phoenix, Arizona, Dec.1999, 2255–2260.

[6] Bell, S. L., R. J. Williams. (2001). Dynamic scheduling of a system with two parallel servers in heavy traffic with complete resource pooling: Asymptotic optimality of a continuous review threshold policy. *Ann. Probab.*, 11, 608–649.

[7] Bramson, M. (2008) Stability of queueing networks. Lecture Notes in Mathematics, 1950, École d'Été de Probabilités de Saint-Flour XXXVI-2006, Springer.

[8] Chen, H. (1995) Fluid approximations and stability of multiclass queueing networks: work-conserving disciplines. Ann. Appl. Prob., 5(3), 637–665.

[9] Dai, J. G. (1995) On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. Ann. Appl. Prob., 5(1), 49–77.

[10] Dai, J. G. (1995). Stability of open multiclass queueing networks via fluid models. In *Stoch. Networks*, F.Kelly and R.Williams editors, v.71, 71–90, NY, Springer.

[11] Dai, J. G., Meyn, S. P. (1995) Stability and convergence of moments for multiclass queueing networks via fluid limit models. IEEE Trans. Aut. Cont., 40(11), 1889–1904.

[12] Delgado, R. (2010) State space collapse and stability of queueing networks. Math Meth Oper Res., 72, 477–499.

[13] Down D.G., Lewis M.E. (2006). Dynamic Load Balancing in Parallel Queueing Systems: Stabilityand Optimal Control. *European Journal of Operational Research.* 168, 509–519.

[14] Foley R.D., McDonald D.R. (2005). Large deviations of a modified Jackson network: Stability and rough asymptotics. *Annals of Applied Probability.* 15(1B), 519–541.

[15] Harrison M., Lopez, M. J. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, 33, 339–368.

[16] N. El Karoui, M. Chaleyat-Maurel, Un problème de réflexion et ses aplications au temps local et aux équations différentielles stochastiques sur $\mathbb{R}$. Cas continu, *Société Mathématique de France, Astérisque*, 52-53 (1978) 117–144.

[17] Kirkizlar E., Andradottir S., Ayhan H. (2010). Robustness of efficient server assignment policies to service time distributions in finite-buffered lines. *Naval Research Logistics*. 57(6), 563–582.
[18] Mandelbaum A., Stolyar A.L. (2004). Scheduling flexible servers with convex delay costs: heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research*, 52(6), 836–855.
[19] Meyn, S. P., Tweedie, R. L. (1994) State-dependent criteria for convergence of Markov chains. Ann. Appl. Probab., 4, 149–168.
[20] Morozov, E., Steyaert, B. (2011) Stability analysis of a two-staion cascade queueing network. Ann. Oper. Res., 13, 1–26.
[21] Reiman, M. I. (1984) Open queueing networks in heavy traffic. Mathematics of Operations Research, 9(3), 441–458.
[22] Stolyar A. L., Tezcan T. (2010). Control of systems with flexible multi-server pools: a shadow routing approach. *Queueing Systems*. 66 1–51.
[23] Tekin E., Hopp W.J., Van Oyen M.P. (2009). Pooling strategies for call center agent cross-training. *IIE Transactions*. 41(6), 546–561.
[24] Terekhov D., Beck J.C. (2009). An extended queueing control model for facilities with front room and back room operations and mixed-skilled workers. *European Journal of Operational Research*. 198(1), 223–231.
[25] Tezcan, T. (2009). Augmented Fluid Models and Stability of Queuing Systems. https://netfiles.uiuc.edu/ttezcan/www/afm070809.pdf.
[26] Tsai Y.C., Argon N.T. (2008). Dynamic server assignment policies for assembly-type queues with flexible servers. *Naval Research Logistics*. 55(3), 234–251.
[27] Williams, R. J. (1998) An invariance principle for semimartingale reflecting Brownian motions in an orthant. Queueing Syst., 30, 5–25.

Rosario Delgado
Departament de Matemàtiques
Universitat Autònoma de Barcelona

Evsey Morozov
Institute of Applied Mathematical Research
Russian Academy of Sciences and Petrozavodsk State University

CRM Preprint Series number 1119