Verification of the Stability of a Two-Server Queueing System With Static Priority

Evsey Morozov
Institute of Applied Mathematical Research,
Karelian Research Centre RAS
Petrozavodsk State University
Petrozavodsk, Russia
emorozov@karelia.ru

Maria Maltseva
Petrozavodsk State University
Petrozavodsk, Russia
masha.mariam.maltseva@mail.ru

Bart Steyaert
SMACS Research Group, Department TELIN
Ghent University
Ghent, Belgium
bart.steyaert@UGent.be

***Abstract*—In this work, we use simulation to verify the stability conditions of the so-called N-model, which consists of two servers and two classes of external customers, both generated by Poisson inputs. Service times are server-dependent and, in each server, are i.i.d. When server 1 is occupied, and there are waiting customers in queue of server 1, then a class-1 customer jumps to server 2, thereby becoming a class-(1,2) customer. We consider a non-preemptive service priority: a class-1 customer starts service in server 2, when a class-2 customer, if any, finishes his service. Thus, server 2 assists server 1, while the reverse interaction is impossible. The purpose of this research is to verify the tightness of the stability condition found in [8] by fluid a approach, and to deduce a simpler sufficient stability condition, which is obtained in an explicit form by a regenerative approach. Moreover, our analysis includes verification of the conditions when the 1st server is stable, while the 2nd server is unstable. In addition, we verify by simulation a monotonicity property of this model: the idle stationary probability of server 1 attains a minimum when the 2nd server is permanently occupied by class-2 customers.**

I. Introduction

This model is a variation of the single-server $N-$model from [6], in which server 2 accepts class-1 customers when free, but gives preemptive priority to class-2 customers. For this reason, the stability conditions of the present model and the model from [6] are different. Following [8], we call the present model the N-*model with static priority* . In this N-model, which is very well motivated, see again [8] and also [9], [10], server 1 can be treated as *beneficiary*, while server 2 is the *donor*.

This queueing system configuration has a lot in common with the notion of flexible servers, meaning that some service capacity may be transferred from one server to another to accommodate varying demands. Also closely related are cross-trained servers (see [1], [2], [12], [13] and the references therein) where one server (or pool of servers) is fit to handle a limited set of customers, whereas a second server(pool) can handle more types of customers that enter. The described queueing networks can model a broad class of computer networks with rescheduling of jobs, multiserver systems with heterogeneous job types, etc. As stated in [3], this system can model a computing system where processors have overlapping capabilities.

The models with flexible servers mainly appear in analysis of multiclass multiserver systems to find an optimal allocation by minimizing a cost function. As a rule, the identification of stability region is based on the so-called *resource pooling (or complete resource pooling) assumption* implying that a full (aggregated) capacity of the system is used in when the system is heavy loaded. For instance, such an approach is applied in [14] for a multiclass system, to solve an optimal allocation problem by construction of a linear program solution. We stress again that cooperating servers join efforts to serve the corresponding customer, and it makes the stability analysis extremely hard. In this regard, we mention the paper [15] which shows that a correct stability analysis is not straightforward even under exponential assumptions.

It is worth mentioning that in order to develop the stability analysis of the non-preemptive scheduling discipline, the author of [8] has been forced to modify the well-known and recognized fluid stability analysis, because its classic form (see, for example, [4]) does not allow to analyse this complicated model.

However, there are some problems concerning the application of the stability analysis developed in [8]. The first problem is that this analysis holds only provided an "Assumption 1" holds, which in turns states that there exists a specific solution of a linear program related to this model. Thus before being able to find the stability conditions, we must first formulate and solve this linear program. Another important problem is that the basic stability condition in [8] (see also the analysis below) in general is not formulated in an explicit form. More precisely, this condition contains, besides the given first moments of the governing distributions (that determine the service times and inputs), also an unknown parameter. Actually this parameter is the stationary busy probability of the 1st server provided the 2nd server is overloaded by class-2 customers. Thus, to apply this condition in practice, this parameter must be estimated in advance.

In practice, the presence of this unknown parameter makes the stability analysis of this model highly difficult, because simulation can take a lot of effort, requires development of the corresponding algorithms, and yields numerical result only for the simulated system parameters with little additional insight. Hence, it is useful to search a simpler stability condition which can be practically implemented, but which may be less tight.

The contribution of this work is as follows. First of all, we give a simple proof of the necessary stability condition of the

ISSN 2305-7254

two-server model, using an alternative regenerative approach. Then we estimate by simulation the mentioned parameter, and verify both our stability conditions of each server and the stability criterion of the two-server system from [8]. Moreover, we verify by simulation an important monotonicity property of the model: the stationary busy probability of the 1st server attains minimum when the 2nd server is overloaded by class-2 customers. It opens the way for a full regenerative stability analysis of the system, which we leave open for further research at this point.

The paper is organized as follows. In Section 2, the model is described in detail, and the main notations are given. Section 3 contains in brief the regenerative proof of the necessary stability condition. Also, we deduce in an explicit form the stationary solution for a special case of the model. In this section, a simpler set of stability conditions are discussed as well. Finally, in Section 4, simulation results are given, which demonstrate the difference between stability regions given by different stability conditions, in addition to the mentioned monotonicity property of the busy probability.

II. Description of the model

We consider a two-server queueing model, both with infinite-capacity buffers. The 1st server is fed by a Poisson input with rate λ_1, and the 2nd server is fed by Poisson input with rate λ_2. Class-1 customers can be served both servers 1 and 2, while class-2 customers can be accommodated by server 2 only. Moreover, an arbitrary class-1 customer waiting in the 1st queue, jumps to server 2 and has non-preemptive priority over class-2 customers: it starts service immediately after the class-2 customer being served (if any) leaves server 2. We call such a customer a $(1, 2)$-customer (or (12)-customer). It is unimportant for stability analysis which waiting class-1 customer makes this jump.

We assume that the service times of class-i customers $\{S_k^{(i)},\ k \geq 1\}$ are i.i.d. with rate $\mu_i = 1/\mathsf{E}S^{(i)} \in (0,\ \infty)$, $i = 1, 2, (1, 2)$. (In what follows, we omit the serial index to denote a generic element of an i.i.d sequence.) All sequences are assumed to be independent.

We denote $Q_i(t)$ the *queue size* (the number of customers in the server and waiting for service) in server i at instant t^- , $i = 1, 2$. We assume an arbitrary *work-conserving service discipline* in each pool, in particular, an *arbitrary waiting class-1 customer* may jump to server 2 (provided $Q_1(t) > 1$) because stability/instability does not depend on the order of customers, which (in each class) are stochastically undistinguishable.

The regenerations of server 1 are generated by class-1 customers arriving to an empty system. Denote by $V_1(t)$, $B_1(t)$ the arrived and departed work, respectively, in the interval $[0,\ t]$. Also, denote by $L_1(t)$ the work *lost by server 1* in the interval $[0,\ t]$. In other words, it is the summary (not realised) unfinished work in server 1 of class-(1,2) customers. Denote by $A_1(t)$ the number of class-1 and $A_{12}(t)$ the number of class-(1,2) customer arrivals in $(0,\ t]$. Denote by $W_1(t)$ the remaining work (workload) at instant t^+, and let $\rho_i = \lambda_i/\mu_i$, $i = 1, 2$.

The stability analysis in [8] has been developed to obtain stability of the entire 2-server system, and in particular, the stability of the isolated 1st server is not considered. Nevertheless, it may be sometimes useful to know the stability condition of the 1st server regardless of the state of the 2nd server. Introduce a basic process

$$X(t) = Q_1(t) + Q_2(t),\ t \geq 0,$$

describing the total number of customers in the system.

Let T_n be the nth regeneration instant of the process $\{X(t)\}$, which is generated by a customer arriving in an empty system. More exactly, if t_n is the nth arrival instant in the superposed input (Poisson) process, then regenerations are defined recursively as follows: set $T_0 = 0$ and

$$T_{n+1} = \inf(t_k > T_n : X(t_k^-) = 0),\ n \geq 0. \tag{1}$$

We recall that the regeneration period lengths $T_{n+1} - T_n$ are i.i.d., and the values of the regenerative process $\{X(t)\}$ belonging to different regeneration periods are i.i.d. as well. Note that, for the non-preemptive priority, we can not consider regeneration of the 1st server solely because of a dependence between the states of both servers. (However, it is possible for the system with preemptive priority, because in this model class-1 customers "do not see" the class-2 customers.) Let T be the generic regeneration period length of the 1st server system generated by a class-1 customer arriving in an empty system. In other words, T is distributed as any difference $T_{n+1} - T_n$.

III. Stability analysis

In this section, we apply a regenerative approach to obtain a simple proof of the necessary stability condition of the 1st server solely.

To find the necessary stability condition of the 1st server, we assume that it is stable (*positive recurrent*), that is $\mathsf{E}T_1 < \infty$ [7]. This condition implies the existence of stationary distribution the process $\{X(t)\}$, that is $X(t) \Rightarrow X$, $t \to \infty$, and any associated processes we consider below. (Here $\Rightarrow$ means convergence in distribution, and X is the stationary number of customers in the system.)

Note that $B_1(t) = t - I_1(t)$, where $I_1(t)$ is an empty time of server 1 in the interval $[0,\ t]$. We have the following balance equation for any $t \geq 0$:

$$V_1(t) = W_1(t) + L_1(t) + t - I_1(t). \tag{2}$$

By the Strong Law of Large Numbers (SLLN), with probability (w.p.) 1,

$$\frac{V_1(t)}{t} = \frac{\sum_{k=1}^{A_1(t)} S_k^{(1)}}{A_1(t)} \frac{A_1(t)}{t} \to \rho_1,\ \ t \to \infty. \tag{3}$$

Denote by $\mathcal{L}(t)$ the set of numbers of class-1 customers that have jumped to server 2 in the interval $[0,\ t)$, and $|\mathcal{L}(t)| = A_{12}(t)$ as its capacity. Then

$$L_1(t) = \sum_{k \in \mathcal{L}(t)} S_k^{(1)}. \tag{4}$$

Now, by the SLLN, w.p.1 as $t \to \infty$,

$$\begin{aligned} \frac{1}{t} L_1(t) \quad &= \quad \frac{1}{t} \sum_{k \in \mathcal{L}(t)} S_k^{(1)} \\ &=_{st} \quad \frac{\sum_{k \in \mathcal{L}(t)} S_k^{(1)}}{A_{12}(t)} \frac{A_{12}(t)}{A_1(t)} \frac{A_1(t)}{t} \to \rho_1 \mathsf{P}_\ell, \end{aligned} \tag{5}$$

where $=_{st}$ means stochastic equality, and the limit

$$\mathsf{P}_\ell := \lim_{t\to\infty} \frac{A_{12}(t)}{A_1(t)}, \tag{6}$$

exists and is the stationary probability that a class-1 customer jumps from server 1 to server 2. Denote indicator $\mathbf{1}(t) = 1$, if server 1 is busy at instant t, and $\mathbf{1}(t) = 0$, otherwise, and let P_b be the stationary busy probability of server 1. Then, by positive recurrence of the *cumulative process* $\{B_1(t),\, t \geq 0\}$, w.p.1

$$\lim_{t\to\infty} \frac{B_1(t)}{t} = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbf{1}(u)du = \mathsf{P}_b, \tag{7}$$

see [11]. By positive recurrence, $W(t) = o(t), t \to \infty$ w.p.1 [11], and (2)-(7) imply the following important relation connecting the stationary busy probability and stationary loss probability of the 1st server:

$$\mathsf{P}_\ell = 1 - \frac{\mathsf{P}_b}{\rho_1}. \tag{8}$$

Because a (generic) interarrival time τ is exponential, then

$$\mathsf{P}(\tau > S^{(i)}) > 0,\, i = 1, 2. \tag{9}$$

Then it is easy to show that the positive limit exists [11]

$$\lim_{t\to\infty} \frac{I_1(t)}{t} > 0. \tag{10}$$

Because $t = B_1(t) + I_1(t)$, then it follows from (2), (7) and (10) that the necessary stability condition of the 1st server is

$$\rho_1(1 - \mathsf{P}_\ell) < 1. \tag{11}$$

Next, we present the necessary stability conditions of the original two-server system. Denote $W_2(t)$ as the current workload in the 2nd server at instant t, $B_2(t) = t - I_2(t)$ the accumulated busy time of the 2nd server, $I_2(t)$ – the summary idle time of server 2, and $V_{12}(t)$ – the work that arrived in server 2 from server 1, in the interval $[0,\, t]$, $i = 1, 2$. We assume positive recurrence, $\mathsf{E}T < \infty$, and write the balance equation for the work $V_2(t)$ that arrived in server 2 in interval $[0,\, t]$:

$$V_2(t) + V_{12}(t) = W_2(t) + t - I_2(t). \tag{12}$$

Note that

$$V_{12}(t) = \sum_{k\in\mathcal{L}(t)} S_k^{(12)} =_{st} \sum_{k=1}^{A_{12}(t)} S_k^{(12)}.$$

Again, by the positive recurrence, $W_2(t) = o(t)$, $t \to \infty$ [11], and as above, we obtain that the following limits w.p.1 exist

$$\lim_{t\to\infty} \frac{V_{12}(t)}{t} = \frac{\lambda_1}{\mu_{12}}\mathsf{P}_\ell,\ \lim_{t\to\infty} \frac{V_2(t)}{t} = \rho_2,\ \lim_{t\to\infty} \frac{I_2(t)}{t} > 0. \tag{13}$$

Using (8) we obtain from (12), (13) the *necessary stability condition of the entire system*:

$$\rho_2 + \frac{\lambda_1}{\mu_{12}}\mathsf{P}_\ell = \rho_2 + \frac{\lambda_1 - \mu_1\mathsf{P}_b}{\mu_{12}} < 1. \tag{14}$$

Finally, we note that the regenerative approach also allows to establish stability of the 1st server and instability of the 2nd server, if the following conditions hold:

$$\lambda_1 < \mu_{12} + \mu_1; \tag{15}$$
$$\lambda_1 > \mu_1\mathsf{P}_b + \mu_{12} - \rho_2\mu_{12}. \tag{16}$$

Remark. We note that the previous analysis can be extended to the model where each server is replaced by a pool with an arbitrary number of servers working in parallel.

Stability condition (14) is identical to the condition obtained by a modified fluid approach in [8]. However, it is proved in [8] that (14) is also the stability criterion, if we replace P_b by the stationary busy probability $\mathsf{P}_b^{(o)}$ of the 1st server *provided there are always class-2 customers in the queue of the 2nd server*. (In other words, the 2nd server would be permanently busy accommodating class-2 customers if class-1 customers do not jump to server 2.) Denote the corresponding idle probability $\pi_o(0) = 1 - \mathsf{P}_b^{(o)}$. Then the stability criterion becomes

$$\rho_2 + \frac{\lambda_1 - \mu_1(1 - \pi_o(0))}{\mu_{12}} < 1. \tag{17}$$

Thus, to apply in practice stability condition (14) or (17), we must know probability P_b (or $\pi_o(0)$), which in general can not be expressed in terms of the first moments of the interarrival time and service times only, but rather depends on their distributions. Thus, this quantity needs to be estimated by simulation.

An exception, when explicit solution is easy available, is exponential service times in both servers and moreover, $\mu_2 = \mu_{12}$. Recall that $\rho_1 = \lambda_1/\mu_1$ and denote

$$\sigma = \frac{\lambda_1}{\mu_1 + \mu_2}.$$

Then the corresponding Kolmogorov equations for the stationary probabilities of the state of the 1st server (provided the 2nd one is permanently full) are as follows:

$$\begin{aligned} \lambda_1\pi_o(0) &= \mu_1\pi_o(1), \\ \lambda_1\pi_o(k) &= (\mu_1 + \mu_2)\pi_o(k+1),\ k \geq 1, \end{aligned} \tag{18}$$

where $\pi_o(k) = \mathsf{P}(Q_1 = k)$ is the stationary probability that there are k customers in server 1. (That is, Q_1 is a weak limit: $Q_1(t) \Rightarrow Q_1$.) Fig. 1 (reproduced from [8]) illustrates the transition rates of the Markov chain $\{Q_1(t)\}$.

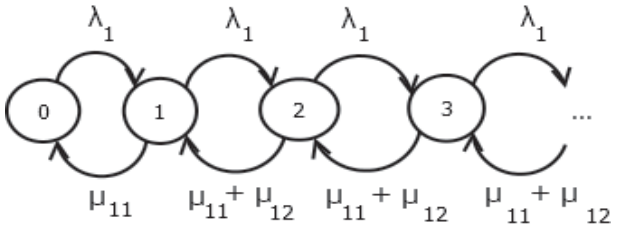

Fig. 1. Transition rates for the system with saturated server 2

To explain (18), we note that when $\mu_2 = \mu_{12}$, the process $\{Q_1(t)\}$ is a Markov chain because, by the memoryless property of exponential distribution, the state of the

2nd server is insensitive to type of customer that occupies server 2. It is remarkable that relations (18) are valid both for the non-preemptive priority and for preemptive-resume priority. Solution of (18) together with normalization condition $\sum_0^\infty \pi_o(k) = 1$ gives the following explicit expressions:

$$\begin{aligned} \pi_o(k+1) &= \sigma^k \rho_1 \pi_o(0),\ k \geq 0, \\ \pi_o(0) &= \frac{1-\sigma}{1-(1-\rho_1)\sigma}, \end{aligned} \tag{19}$$

which coincide with the ones that are given in [8]. Taking into account that, when the 1st queue is positive, both servers join efforts to serve class-1 customers, the inequality

$$\sigma = \frac{\lambda_1}{\mu_1 + \mu_{1\,2}} < 1 \tag{20}$$

guarantees stability of the 1st queue in the worst-case scenario when the capacity of the 1st server is insufficient to process the arriving load. Thus if condition (20) is valid, then the process $\{Q_1(t)\}$ will be stable. It holds both for non-preemptive and preemptive-resume priority, because, under overloaded conditions for server 1, class-1 customers occupy server 2 permanently, and these two different types of priorities become indistinguishable. Provided the 2nd server is overloaded, class-1 customers are always waiting in the 1st queue before jumping to server 2. This waiting time can be interpreted as a "lost part" of the joint capacity $\mu_1 + \mu_{12}$. On the other hand, the condition $\mathsf{P}_b < \rho_1$ holds since not all arriving work is processed by server 1, due to the collaboration with server 2. At the same time, stability of the 1st server is possible even if $\rho_1 > 1$. In summary, we expect that indeed the following inequality

$$\sigma \leq \mathsf{P}_b \leq \min(\rho_1,\, 1), \tag{21}$$

holds in all scenarios where the 1st server is stable, and in particular for $\mathsf{P}_b = \mathsf{P}_b^{(0)}$. Taking into account (21), we can conclude that if the following condition

$$\rho_2 + \sigma < 1, \tag{22}$$

holds, then stability criterion (17) holds as well, and thus (22) is a sufficient stability condition of the original system.

IV. Simulation results

In this section, we present some simulation results. It is worth mentioning that the source work [8] does not contain simulation result illustrating corresponding theoretical statements. Oppositely, we think that simulations are very useful and illustrative to demonstrate the usefulness of the developed theory.

First, we estimate the idle probability of the 1st server $\pi_o(0)$, with a saturated 2nd server, using the explicit formula (19):

$$\pi_o(0) = 1 - \frac{\lambda_1\mu_{12} + \lambda_1\mu_1}{\mu_1^2 + \mu_1\mu_{12} + \lambda_1\mu_{12}}. \tag{23}$$

Fig. 2 illustrates the evident convergence of the sample mean estimate $\hat{\pi}_o(0)$ to the theoretical value $\pi_o(0) = 0.63$ in (23) as the number of customer arrivals increases.

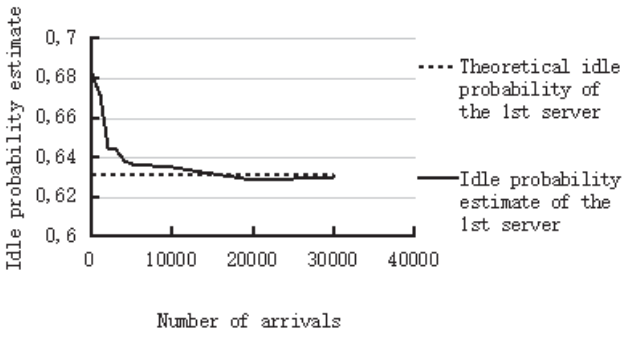

Fig. 2. Theoretical value $\pi_o(0)$ vs. estimated value $\hat{\pi}_o(0)$ with λ_1 = 1, λ_2 = 7, μ_1 = 2, μ_{12} = 5, μ_2 = 5

Now we demonstrate the following monotonicity property of the probability $\pi(0)$. Let $\hat{\pi}_o(0)$ be the estimate of the probability $\pi(0)$ when the 2nd server is saturated by class-2 customers, and $\hat{\pi}(0)$ be the estimate of the probability $\pi(0)$ in the original system. Then we show that the estimate $\hat{\pi}(0)$ attains a minimum when the 2nd server is permanently busy, that is $\hat{\pi}(0) \geq \hat{\pi}_o(0)$. Intuitively, this property is clear because, with the saturated 2nd server, class-1 customers in general will jump to the 2nd server less intensively than in the original model. Hence, the 1st server will be idle during less time, implying $\pi_o(0) \leq \pi(0)$. This property is confirmed for 1) exponential service times (Fig. 3) with rates $\lambda_1 = 1$, $\lambda_2 = 8$, $\mu_1 = 2$, $\mu_{12} = \mu_2 = 5$, and for 2) Pareto service time distribution (Fig. 4),

$$F(x) = 1 - (1/x)^{k_i},\ x \geq 1,\ i = 1, 2, (12),$$

with parameters λ_1 = 0.2, λ_2 = 0.5, k_1 = 3, $k_{12} = k_2$ = 4. Recall that the mean Pareto service time $S^{(i)}$ equals

$$\mathsf{E}S^{(i)} =: \frac{1}{\mu_i} = \frac{k_i}{k_i - 1},\ i = 1, 2, (12).$$

For this example, the service rates are equal to $\mu_1 = 0.68, \mu_{12} = \mu_2 = 0.75$,

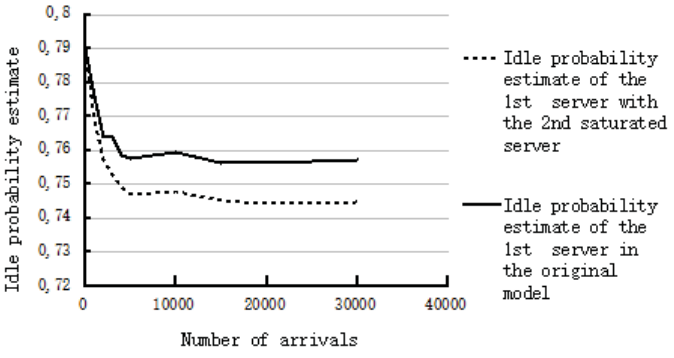

Fig. 3. The monotonicity of $\hat{\pi}(0)$ for exponential service times

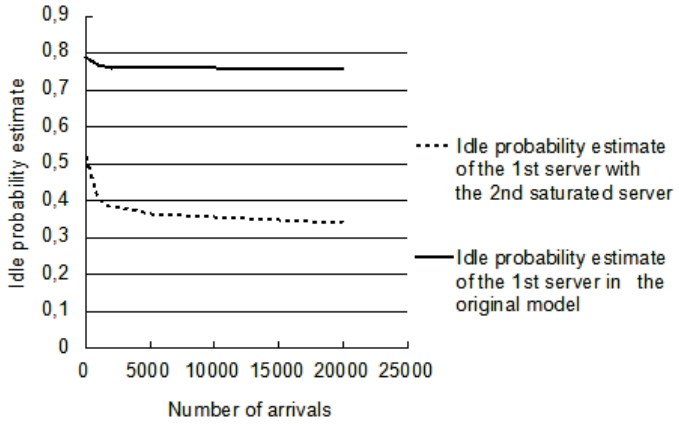

Fig. 4. The monotonicity of $\hat{\pi}(0)$ for Pareto service times

Now we verify the stability/instability of the system and of each server, for different service time distributions, by means of the conditions obtained above. We denote n_i the number of class-i customer arrivals, and use $n_1 = n_2 = 20000$ in all experiments below.

Note that by (21) (22), the difference

$$\sigma - \frac{\lambda_1 - \mu_1(1 - \pi_o(0))}{\mu_{12}} := \Delta \tag{24}$$

satisfies the inequalities

$$0 < \Delta \leq \rho_1(1 + \frac{\mu_{12}}{\mu_1}),$$

and this indicates a possible difference between stability regions, which are delimited by condition (22) and *minimal condition* (17), respectively.

The following scenarios are possible.

1. Stability of both servers.

2. The 1st server is stable, the 2nd one is unstable.

3. Instability of both servers.

For each of these cases, we verify inequalities (21).

Case 1. Fig. 5 illustrates the stability of the second server under condition (17), with $\pi_o(0)$ replaced by the estimate $\hat{\pi}(0)$ in the *original model*, that is

$$\lambda_1 < \mu_1(1 - \hat{\pi}(0)) + \mu_{12} - \rho_2\mu_{12} \tag{25}$$

for $\lambda_1 = 0.2, \lambda_2 = 0.5$ and Pareto service times with parameters $k_1 = 3$, $k_{12} = k_2 = 4$, implying $\mu_1 = 0,68$, $\mu_{12} = \mu_2 = 0.75$. (The behaviour of the 1st queue is similar, and the corresponding picture is omitted.) It is more convenient to rewrite inequality (21) as

$$1 - \min(\rho_1, 1) \leq \pi(0) \leq 1 - \sigma. \tag{26}$$

Then we obtain $\hat{\pi}(0) \approx 0.75 \in (0.701, 0.858)$, where the bounds of interval (here and below) correspond to the upper and the lower bounds in inequalities (26).

Note that behaviour of the 1st queue is similar.

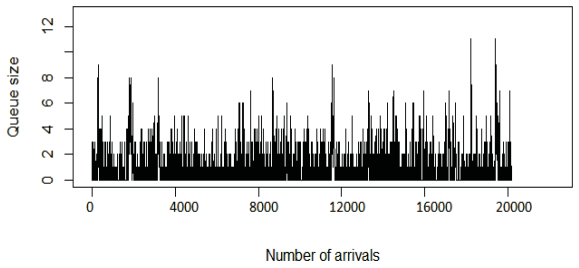

Fig. 5. Stability of the 2nd server with Pareto service times

Case 2. Note that conditions $\lambda_1 < \mu_1$ + μ_{12} and

$$\lambda_1 > \mu_1(1 - \hat{\pi}_o(0)) + \mu_{12} - \rho_2\mu_{12} \tag{27}$$

imply stability of the 1st server, and instability of the 2nd server. However, we replace $\hat{\pi}_o(0)$ by the observed estimate $\hat{\pi}(0)$, in which case in general (27) does not imply instability of the 2nd server. However, this allows us to simplify the analysis (because we do not care about the behaviour of the 2nd server in advance). On the other hand, if the 2nd server becomes unstable, then $\hat{\pi}(0) = \hat{\pi}_o(0)$.

Figs. 6, 7 illustrate the stability of the 1st server and instability of the 2nd server with input rates λ_1 = 0.2, λ_2 = 100 and Pareto service times with parameters k_1 = 3.1, k_{12} = k_2 = 3.2, implying service rates μ_1 = 0.68, μ_{12} =μ_2 = 0.69. Also $\hat{\pi}(0) = \hat{\pi}_o(0) \approx 0.75 \in (0.705, 0.854)$, and (26) holds.

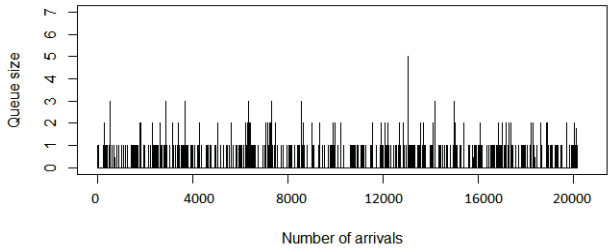

Fig. 6. Stability of the 1st server with Pareto service times

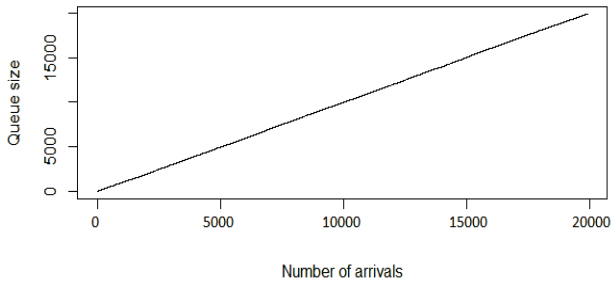

Fig. 7. Instability of the 2nd server with Pareto service times

Case 3. Now we consider instability of both servers, with the input rates $\lambda_1 = 10, \lambda_2 = 7$ and with Pareto service times with parameters $k_1 = k_{12} = k_2 = 3$, under condition $\lambda_1 > \mu_1 + \mu_{12}$, implying $\sigma > 1$. Note that instability of the 1st server implies instability of the 2nd server as well. This yields service rates $\mu_1 = \mu_{12} = \mu_2 = 0.67$ and, as we expect, the estimate $\hat{\pi}(0) = \hat{\pi}_o(0) \to 0$. Fig. 8 illustrates the instability of the 1st server, in view of the unlimited (linear) growth of the queue size in the 1st server. Note that behaviour of the 2nd queue is similar, and the illustration is omitted.

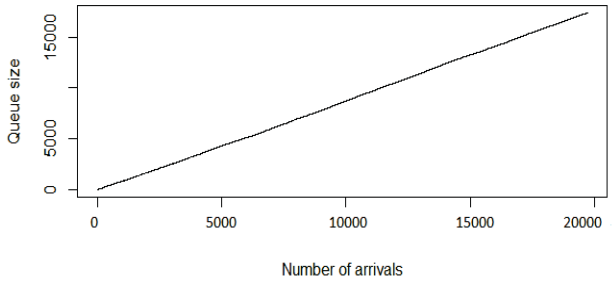

Fig. 8. Instability of the 1st server with Pareto service times

We also consider a Weibull service time distribution

$$F(x) = 1 - e^{-x^{k_i}},\ x \geq 0,\ i = 1, 2, (12).$$

Recall that the mean of Weibull service time $S^{(i)}$ is expressed by the formula

$$\mathsf{E}S^{(i)} = \Gamma(1 + \frac{1}{k_i}) =: \frac{1}{\mu_i},\ i = 1, 2, (12),$$

where Γ is the Gamma function.

Fig. 9 shows the stability of the 2nd server for parameters $\lambda_1 = 0.2, \lambda_2 = 0.5, k_1 = k_{12} = k_2 = 2$, implying service rates $\mu_1 = \mu_{12} = \mu_2 = 1.13$. In this case inequality (26) holds as well, since $\hat{\pi}(0) \approx 0.85 \in (0.82,\ 0.91)$. (The behaviour of the 1st queue is similar, and not shown.)

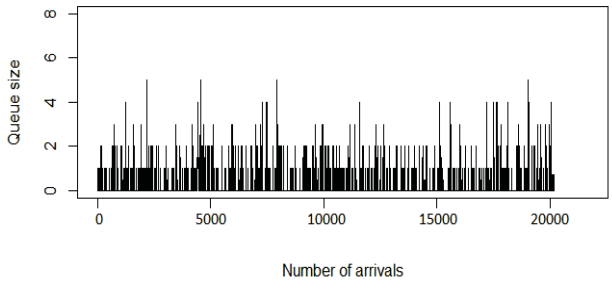

Fig. 9. Stability of the 2nd server with light-tailed Weibull service times

Fig. 10, 11 illustrate stability of the 1st server and instability of the 2nd server for Weibull service times, with parameters λ_1 = 1, λ_2 = 7, k_1 = 2, $k_{12} = k_2$ = 5, implying service rates μ_1 = 1.13, $\mu_{12} = \mu_2$ = 1.09. In this case $\hat{\pi}(0) \approx 0.14 \in (0.11,\ 0.55)$, and (26) holds again.

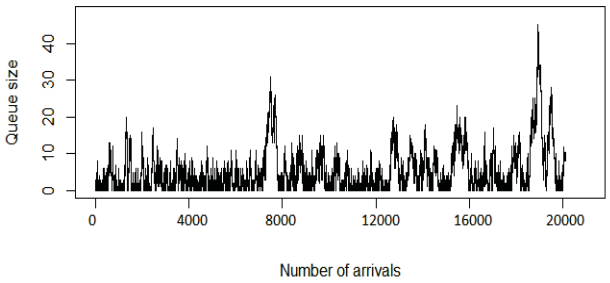

Fig. 10. Stability of the 1st server with light-tailed Weibull service times

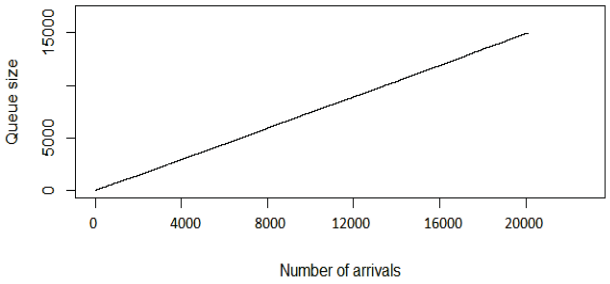

Fig. 11. Instability of the 2nd server with light-tailed Weibull service times

Also we obtain instability of both servers (under condition $\lambda_1 > \mu_1 + \mu_{12}$) with light-tailed Weibull service times with parameters λ_1 = 10, λ_2 = 20, $k_1 = k_{12} = k_2$ = 2, implying $\mu_1 = \mu_{12} = \mu_2$ = 1.13. Also in this case $\hat{\pi}(0) \to 0$. The queue size in both servers increases linearly, as the number of arrivals increases, as for Pareto service times, see Fig. 8.

It seems tempting to use a simpler but less tight stability condition, for instance,

$$\lambda_1 < \mu_1 + \mu_{12} - \rho_2\mu_{12}, \tag{28}$$

where the unknown parameter $\pi_0(0)$ is replaced by 0. Of course, this may lead to an error, as the following example shows. We take parameters λ_1 = 5, λ_2 = 0.5, k_1 = 3, $k_{12} = k_2$ = 4, implying μ_1 = 0.667, $\mu_{12} = \mu_2$ = 0.75 and $\hat{\pi}(0) \approx 0.74$. In this case condition (28) holds, but we indeed observe instability of the 2nd server (as on Fig. 11), because in this case the tighter condition (25) is violated.

V. Conclusion

In this work, we verify the stability of a special case of the so-called N-model with two interacting servers, where server 2 helps to serve class-1 customers arriving in the 1st server. It is well-known that stability analysis of computer systems with interacting servers is a challenging problem. We outline a new regenerative proof of the necessary stability conditions of this model (obtained earlier by the fluid approach in ([8]),

and this approach for this problem is new to the best of our knowledge. Another contribution is that we present stability conditions which allow to distinguish stability and instability regions of each server individually. Moreover, we demonstrate the monotonicity property of the estimate of the 1st server idle probability. Theoretical results are illustrated by numerical examples, obtained by simulation, which confirm our findings.

ACKNOWLEDGMENT

The research of EM and MM is in part supported by Russian Foundation for Basic Research, projects 18-07-00147, 18-07-00156. Part of the research of the third author has been funded by the Interuniversity Attraction Poles Program initiated by the Belgian Science Policy Office. This research is done under support of Institute of Applied Mathematical Research, Karelian Research Centre RAS.

REFERENCES

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

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

[3] S. L. Bell, R. J. Williams, "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, pp. 2255-2260.

[4] J. G. Dai, "On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models", *Ann. Appl. Prob.*, vol. 5, no. 1, 1995, pp. 49-77.

[5] E. Morozov, "The tightness in the ergodic analysis of regenerative queueing processes", *Queueing Systems*, no. 27, 1997, pp. 179-203.

[6] R. Delgado, E. Morozov, "Stability analysis of cascade networks via fluid models", *Performance Evalutation*, vol. 82, 2014, pp. 39-54.

[7] E. Morozov, R. Delgado, "Stability analysis of regenerative queues", *Automation and Remote control*, vol. 70, no. 12, 2009, pp. 1977-1991.

[8] T. Tezcan, "Stability analysis of N-model systems under a static priority rule", *Queueing Systems*, vol. 73, 2013, pp. 235-259.

[9] W. Whitt, "Blocking when service is required from several facilities simultaneously", *AT&T Technical Journal*, vol. 64, issue 8, 1985, pp. 1807-1856.

[10] D. Wong, N. Paciorek, T. Walsh, J. DiCelie, M. Young, B. Peet, "Concordia: An infrastructure for collaborating mobile agents. International Workshop on Mobile Agents MA 1997: Mobile Agents", *LNCS*, Springer, vol. 1219, 1997, pp. 86-97.

[11] W. L. Smith, "Regenerative stochastic processes", *Proc. Roy. Soc.*, ser. A 232, 1955, p. 6-31.

[12] E. Tekin, W.J. Hopp, M.P. Van Oyen, "Pooling strategies for call center agent cross-training", *IIE Transactions*, vol. 41, no. 6, 2009, pp. 546-561.

[13] D. Terekhov, J.C. Beck, "An extended queueing control model for facilities with front room and back room operations and mixed-skilled workers", *European Journal of Operational Research*, vol. 198, issue 1, 2009, pp. 223-231.

[14] S. Andradottir, H. Ayhan, G. D. Down, "Dynamic server allocation for queueing networks with flexible servers", *Operations Research*, vol. 51, no. 6, 2003, pp. 952 - 968.

[15] R. D. Foley, D. R. McDonald, "Large deviations of a modified Jackson network: stability and roughasymptotics", *The Annals of Applied Probability*, 15(1B), 2005, pp. 519 - 541.