Sufficient Stability Conditions for Multi-class Constant Retrial Rate Systems

Konstantin Avrachenkov, E. Morozov, B. Steyaert

Sufficient Stability Conditions for Multi-class Constant Retrial Rate Systems

K. Avrachenkov[*], E. Morozov[†], and B. Steyaert[‡]

October 17, 2015

Abstract

We study multi-class retrial queueing systems with Poisson inputs, general service times, and an arbitrary numbers of servers and waiting places. A class-i blocked customer joins orbit i and waits in the orbit for retrial. Orbit i works like a single-server $\cdot/M/1$ queueing system with exponential retrial time regardless of the orbit size. Such retrial systems are referred to as retrial systems with constant retrial rate. Our model is motivated by several telecommunication applications, such as wireless multi-access systems, optical networks and transmission control protocols, but represents independent theoretical interest as well. Using a regenerative approach, we provide sufficient stability conditions which have a clear probabilistic interpretation. We show that the provided sufficient conditions are in fact also necessary, in the case of a single-server system without waiting space and in the case of symmetric classes. We also discuss a very interesting case, when one orbit is unstable, whereas the rest of the system is stable.

1 Introduction

We study a retrial system with N classes of customers, c servers and $K - c < \infty$ waiting spaces in the buffer. Customers from class i originally arrive according to a Poisson process with rate λ_i. If a customer from class i finds all c servers busy and the buffer full, he joins the i-th infinity-capacity orbit. If the i-th orbit is not empty, only one orbit customer attempts to rejoin the primary queue after exponentially distributed time intervals with rate $\mu_0^{(i)}$. Such a retrial model is referred to as a retrial model with *constant retrial rate.* Class i customers have i.i.d. service times $\{S_n^{(i)}, n \geq 1\}$ with a generic element $S^{(i)}$. We note however that the specific assignment rule of the service times depends on coupling procedures used in our analysis below. We also require that the service times belong to the so-called class D [30]. In this class, the random variables are absolutely continuous and have well-defined failure rates.

Because of the feedback nature of retrials, the analysis of retrial systems is very difficult [3, 13]. Single-class retrial systems with constant retrial rate have been investigated in a large number of contributions. Not pretending to be exhaustive, let us mention some relevant articles. In [15] a retrial system with

[*] Corresponding author. Inria Sophia Antipolis, 2004 Route des Lucioles, Sophia Antipolis, 06902, France, e-mail: K.Avrachenkov@sophia.inria.fr

[†] Institute of Applied mathematical research, Karelian Research Centre RAS, and Petrozavodsk State University, Russia, e-mail: emorozov@karelia.ru

[‡] SMACS Research Group, Ghent University, St-Pietersnieuwstraat 41, B-9000 Gent, Belgium, e-mail: bs@telin.ugent.be

constant retrial rate was introduced for the first time, and stability conditions were derived for the case of an $M/G/1/1$ primary queue. In [1] the author obtained stability conditions for the Markovian $M/M/2/2$ case. In [27] the authors deduced stability conditions for the $M/M/1/2$ case. For the general Markovian $M/M/c/K$ case, the authors of [27] have obtained decomposition results assuming ergodicity (stability). The ergodicity conditions for the multiserver Markovian $M/M/c/c$ case with recovery probability were derived in [2]. Finally, a sufficient stability condition of the general single-class retrial system with constant retrial rate described above was obtained in [8], which turned out to be also a necessary one for Markovian systems.

As is typically the case, the analysis, and in particular the stability issue of a multi-class system, is much more challenging than that of the single-class variant. There are only few works dedicated to the analysis of multi-class retrial systems. In [17] the author considered a two-class retrial system with batch arrivals and classical retrial discipline. In [14] the results of [17] were generalized to the case of more than two classes. In [28] the authors considered a multi-class retrial queue with classical retrial discipline and $M/M/c/c$ primary queue. In particular, they showed that as appropriately scaled retrial rates go to infinity, the system becomes close to the random order service priority system. In [26] the authors considered a multi-class retrial queue with one server, a classical retrial policy and with a non-preemptive priority mechanism between different classes. Then, in [18] the model of [26] was extended to associate different phases of service with different retrial classes. In [7] necessary stability conditions were obtained for the case of a single-server, multi-class retrial queue with constant retrial rates. Recently in [9] the authors studied the two-class Markovian retrial queueing system with one server, no buffer and with constant retrial rates. In particular, the authors obtained necessary and sufficient stability conditions for such a Markovian system.

Retrial systems with constant retrial rate can be adopted to model a range of telecommunication systems, such as a telephone exchange system [15], multiple access systems [10, 11], short TCP transfers [4, 5], optical packet switching networks [31, 32] as well as logistic systems [19]. Multi-class multi-orbit retrial systems are natural extensions of these applications for customers with different Quality of Service requirements, and TCP or optical transfers with several source-destination pairs. Let us briefly discuss optical networks as a specific example of application for our model. The primary queue models a wavelength channel and each class corresponds to a connection for reliable data transfer. Thus, several connections using the same wavelength are in competition and have to resend packets if the sent packets find the channel busy. The packets needed to be resent are then queued at the MAC layer in the source. The orbit represents this queue in our terminology. Therefore, with the help of our model we can study the constraints of the wavelength channel with several competing connections.

In the present work, using a dominating auxiliary system – which helps to break the feedback effect of retrials – and adopting a regenerative approach, we first provide sufficient stability conditions for a quite general multi-class retrial queue with constant retrial rates. The obtained sufficient conditions are actually not that far from the necessary conditions. In fact, in the particularly important case of a single server, no buffer, and either symmetric classes or a single customer class, the sufficient conditions coincide with the necessary conditions given in [7]. Then, in the second part of the paper we study a very interesting, and conceptually new, situation of "partial instability", when the orbit length of one class goes to infinity while the orbit lengths of the other classes remain bounded (tight). The analysis of such a system is especially difficult since the basic regenerative process describing its behavior is not positive recurrent. However, some performance metrics of such a system are well-behaving and can be calculated in explicit form thanks in fact to the “partial instability” of the system. Such “partial instability” can be viewed as a good property of the system that ensures a significant level of fairness if one of the sources ’misbehaves’. If one class of customers violates its service contract, this class will be severely punished while the other classes will suffer only slightly.

Let us also mention some limitations of our analysis. The analysis is based on the regenerative method, which has been previously used to establish sufficient stability conditions of various queueing models, including non-Markov systems, see for instance, [20, 21, 25] and the references therein. Being strong, this method requires the existence of the (infinite) sequence of the regeneration instances, and it can be a problem to construct such instances for a queueing system fed by a superposition of the independent renewal processes. In the context of this paper, such a problem arises if we assume non-Poisson inputs and/or non-exponential retial times. In this case, in the auxiliary dominating system $\hat{\Sigma}$ (see below), the summary input would be a superposition of the renewal inputs, and the regenerative method, in a continuous-time setting, faces difficulties to obtain synchronized (common) renewal points of the input, while this is a crucial step to construct regenerations of the basic queueing processes. (We would like to point out that there is no such problem in a discrete-time setting, see [23].) Thus, in this aspect the well-known fluid stability analysis is less restrictive, however it has other limitations. (see [12]; a discussion of the limitations and comparison of both methods can be found in the concluding section of paper [22].)

The paper is organized as follows. In Section 2, we present and prove the main results (Theorems 1, 2 and 3). A detailed proof of the basic Theorem 1 is rather long, and we therefore split it into two parts. The content of the first part, presented in Lemma 1, is important for a good understanding of the ensuing consideration and therefore is given in Section 2, while the second part of the proof of Theorem 1 is postponed to the Appendix. In Section 3, for the first time, we present the stability analysis of a “partially unstable” queueing system. Finally, Section 4 provides conclusions and directions for future research.

2 Sufficient stability conditions

Consider a c-server retrial system Σ with a finite waiting space of size K (including servers), and N classes of customers arriving according to Poisson processes with rates λ_i, exponential retrial times with rates $\mu_0^{(i)}$ independent of the orbit length, and class-dependent general service time distributions with rates μ_i, $i = 1, \dots, N$.

It is assumed that the service time distributions belong to class D [8, 30]. Namely, they are described by non-negative absolutely continuous random variables ξ_i, with cumulative distributions $F_{\xi_i}(t)$, densities $f_{\xi_i}(t)$, and with failure rates

$$r_{\xi_i}(t) := \frac{f_{\xi_i}(t)}{1 - F_{\xi_i}(t)},$$

defined for such t that $F_{\xi_i}(t) < 1$, $i = 1, \dots, N$.

For further use, we introduce the following stochastic r-ordering process [30]: $\xi \leq_r \eta$ iff

$$\inf_{t \geq 0} r_\xi(t) \geq \sup_{t \geq 0} r_\eta(t). \tag{1}$$

2.1 Auxiliary dominating system

Our derivation of the sufficient stability condition is based on the auxiliary system $\hat{\Sigma}$. The auxiliary loss system $\hat{\Sigma}$ only has the following difference with the original system Σ: the retrial Poisson inputs with rates $\mu_0^{(i)}$, $i = 1, \ldots, N$ do not have "gaps", unlike the streams from the orbits in the original system Σ which have gaps when the corresponding orbit is empty. It is convenient to imagine that the $\mu_0^{(i)}$-input is fed by an infinitely loaded virtual orbit, and that the rejected class-i customers are coloured and join the front of this virtual orbit i. (Of course, we could assume that a rejected customer joins the virtual orbit in a random order, since the stability analysis does not depend on the position of the customer in the virtual orbit.) Thus, if the class-i coloured customers are exhausted, then the uncoloured customers continue to arrive in the system, making retrial attempts according to a Poisson input with rate $\mu_0^{(i)}$. Denote by t_n, $n \ge 1$, the arrival instants of the merged (Poisson) input in system $\hat{\Sigma}$. Let the indicator $\hat{I}(t) = 1$ if the *primary system* (that is, all servers and the buffer) is completely full at instant t, and $\hat{I}(t) = 0$, otherwise. In particular, the primary system accepts an arrival at instant t, if $\hat{I}(t) = 0$. Then the process

$$\hat{B}(t) := \int_0^t \hat{I}(u)du,$$

is the length of the time period, within the interval $[0, t]$, when the primary system is completely full (and cannot accept new customers for service). Let $\hat{Q}(t)$ denote the number of customers in the primary system (of $\hat{\Sigma}$) at instant t. The process $\hat{Q} := \{\hat{Q}(t), t \ge 0\}$ is regenerative with regeneration instants which are defined as follows:

$$Z_0 := 0,\ Z_{n+1} = \min_k(t_k > Z_n :\ \hat{Q}(t_k^-) = 0),\ n \ge 0. \tag{2}$$

In other words, a regeneration of the process $\hat{Q}$ occurs when a new customer sees an empty primary system. (Notice that the process $\hat{Q}$ is independent on the number of coloured customers in the orbits.) Clearly, a new regeneration period is initiated by a class-i customer with probability

$$p_i := \frac{\lambda_i + \mu_0^{(i)}}{\sum_{j=1}^N (\lambda_j + \mu_0^{(j)})},\ i = 1, \ldots, N. \tag{3}$$

In this regard we note that one can define *i-type regenerations* when an i-class customer meets an empty primary system, but we do not consider them in this work. The process $\{\hat{B}(t), t \ge 0\}$, describing the time when the primary system in $\hat{\Sigma}$ is fully busy, is a cumulative process with the associated process of regenerations $\{Z_n\}$ because for this process, again, we do not distinguish between coloured and uncoloured customers. (For more details on cumulative processes, see for instance, [29].)

It is important to point out that in the multi-class system $\hat{\Sigma}$, the arrival process in the primary queue is a superposition of Poisson processes (either external type-i arrivals or stemming from the i-th orbit, $1 \le i \le N$). Hence, in order to analyse the process $\{\hat{B}(t), t \ge 0\}$, $\hat{\Sigma}$ can be treated as *a single-class system* with (generic) service time S, in which the total input rate is

$$\lambda := \sum_{j=1}^N (\lambda_j + \mu_0^{(j)}), \tag{4}$$

and each new arrival has service rate μ_i with probability p_i. Thus, the average service time of a customer is given by $\mathsf{E}S := \sum_i p_i/\mu_i$. This results in the following traffic intensity

$$\hat{\rho} := \lambda \mathsf{E}S = \sum_{i=1}^{N} \frac{\lambda_i + \mu_0^{(i)}}{\mu_i}. \tag{5}$$

Since the primary system has a finite buffer, the stationary probability $\hat{\mathsf{P}}_f$ that the primary system is full exists, and is defined as the (w.p.1) limit

$$\lim_{t\to\infty} \frac{\hat{B}(t)}{t} = \hat{\mathsf{P}}_f. \tag{6}$$

(Note that the limit in average in (6) then also exists, [29].) To prove the existence of the limit (6) we use the following regenerative arguments. Since the system has a finite buffer, it is easy to prove that the total workload $\hat{W}(t)$, at instant t, in the primary system is tight, and in particular,

$$\inf_t \mathsf{P}(\hat{W}(t) \le X_0) \ge c_0 > 0,$$

for some c_0 and a constant X_0 depending on c_0. (We could take c_0 arbitrary close to 1, this choice has no impact on the final results.) Then the probability that all primary arrivals and the orbit customers do not attack the system during time X_0 is not less than

$$\exp\{-\sum_i (\lambda_i + \mu_0^{(i)}) X_0\} > 0.$$

It means that a regeneration of the primary system occurs in a finite interval with a positive probability. (In this part of the analysis we again do not distinguish between coloured and uncoloured customers.) This immediately implies the positive recurrence of any process associated with the primary system in $\hat{\Sigma}$ and, in particular, the process of regenerations $\{Z_n\}$.

Let $\hat{J}(t) = 0$ if the primary system (in $\hat{\Sigma}$) is *empty* at instant t (and $\hat{J}(t) = 1$, otherwise). Denote by $C_i(t)$ the number of the coloured customers in orbit i at instant t, and introduce the $(N+1)$-dimensional process $\mathcal{C}(t) := \{\hat{J}(t), C_1(t), \dots, C_N(t)\}$, $t \ge 0$. Note that we do not need to assume a Markovian property for this process. The regenerations of the process $\{\mathcal{C}(t),\, t \ge 0\}$ are defined recursively as follows: $(T_0 := 0)$

$$T_{n+1} = \min(t_k > T_n : \mathcal{C}(t_k^-) = \mathbf{0}),\ n \ge 0, \tag{7}$$

when an arriving customer sees *an empty primary system with no coloured customers in all orbits.* Denote by T a generic regeneration period, that is the stochastic equality $T =_{st} T_2 - T_1$ holds (on the event $\{T_2 < \infty\}$). The regenerative process $\{\mathcal{C}(t),\, t \ge 0\}$ is called positive recurrent if

$$T_1 < \infty \ \text{ w.p.1 and } \ \mathsf{E}T < \infty. \tag{8}$$

Below we first consider the zero-delayed case when the first regeneration period is distributed as the generic period, that is $T_1 =_{st} T$. In particular, this means that the primary system and all coloured orbits are empty at instant $t = 0$, which is the arrival instant of the first customer. The general case is treated in Subsection 2.2.

Our first goal is to find sufficient conditions implying positive recurrence of the process $\{\mathcal{C}(t),\, t \ge 0\}$. Then we use a dominance property of the auxiliary system $\hat{\Sigma}$ to establish that the total orbit size process (that is the sum of the orbit sizes) in the original system Σ is also positive recurrent under the same conditions.

Theorem 1. *Assume that $\mathcal{C}(0)=0$ and that the following conditions hold:*

$$(\lambda_i + \mu_0^{(i)})\hat{\mathsf{P}}_f < \mu_0^{(i)},\ i = 1, \dots, N, \tag{9}$$

where $\hat{\mathsf{P}}_f$ is as defined in (6). *Then the regenerative process $\{\mathcal{C}(t),\, t \ge 0\}$ is positive recurrent.*

Because the proof of Theorem 1 is rather long, we split it into two steps, formulated as Lemma 1 and Lemma 2, respectively. Since the proof of Lemma 1 is very important for a good understanding of the further analysis, it is given in this section, while the proof of Lemma 2 is postponed to the Appendix.

To facilitate reading, let us first summarize the main steps of the proof. Our key observation is that, *for any fixed i*, the orbit size process $\{C_i(t),\, t \ge 0\}$ regenerates at the instants when a class-i arrival sees both the primary system empty and the coloured customers in orbit i absent. The distribution of the process $\{C_i(t),\, t \ge 0\}$ is insensitive to the type (coloured or uncoloured) of class-j customers arriving in the system, $j \ne i$. Then we show that, if condition (9) holds for some i, the coloured orbit size i is a positive recurrent regenerative process, and hence is tight, irrespective of whether the condition (9) is satisfied or not for other orbits $j \ne i$. This result is based on the monotonicity property of the corresponding loss system [30] and the insensitivity of the busy-time process to the customer type (coloured/uncoloured). Thus, each of the coloured orbit sizes is a tight process, and the total orbit size (of coloured customers) is a tight process. Finally, we show that the tightness implies positive recurrence, where a key role is played by the exponential distributions (corresponding to both the poisson input processes, and the processes of the attempts of the orbit customers to enter the primary system).

Lemma 1. *Under the assumptions of Theorem 1, there exist a constant $\varepsilon > 0$ and a deterministic sequence $z_j \to \infty$ such that*

$$\inf \mathsf{P}(C_i(z_j) = 0) \ge \varepsilon. \tag{10}$$

Proof. For each i, denoting by $\{t_n^{(i)}\}$ the arrival instants of the class-i arrivals, we define the regeneration instants $\{T_n^{(i)}\}$ of the process $\{C_i(t),\, t \ge 0\}$ as follows: $T_0^{(i)} := 0$ and

$$T_{n+1}^{(i)} = \min_k (t_k^{(i)} > T_n^{(i)} : \hat{J}(t_k^{(i)-}) = C_i(t_k^{(i)^-}) = 0),\ n \ge 0. \tag{11}$$

In other words, regeneration occurs when both the primary system and the i-th coloured orbit are empty. (Note that by the property of exponentials, we could define other types of regenerations, in particular, when the i-orbit has fixed size k.) For future use, we also define the indicator $\hat{J}_i(t) = 0$, if the i-th (coloured) orbit is empty at instant t, and $\hat{J}_i(t) = 1$, otherwise.

Now we fix an arbitrary i, and denote by $\hat{D}_i(t)$ the total number of $\mu_0^{(i)}$-arrivals in system $\hat{\Sigma}$ in the interval $[0,\, t]$. (It is useful to note that the process $\{\hat{D}_i(t),\, t \ge 0\}$, is a cumulative process, like the process $\{\hat{B}(t),\, t \ge 0\}$.) Relying on the renewal theory, as $t \to \infty$, we have

$$\frac{\hat{D}_i(t)}{t} \to \mu_0^{(i)},\ \text{w.p.1 and } \frac{\mathsf{E}\hat{D}_i(t)}{t} \to \mu_0^{(i)}. \tag{12}$$

It can be easily shown by coupling that the total number of the rejected $\mu_0^{(i)}$- and λ_i-customers arriving in the system $\hat{\Sigma}$ in interval $[0,\, t]$ is stochastically equivalent to the number of the arrivals generated by a Poisson process $\Lambda(\cdot)$ (with rate $\lambda_i + \mu_0^{(i)}$) in interval $[0,\, \hat{B}(t)]$. To obtain this equivalence, we "couple together" all time periods, in the interval $[0,\, t]$, where the primary system is completely full and then shift the "coupled" interval to the origin. Then we consider the Poisson input $\Lambda_i(\cdot)$ over interval $[0,\,\hat{B}(t)]$, and it follows that the next relation holds for the number of λ_i- and $\mu_0^{(i)}$-customers arriving in the system $\hat{\Sigma}$ in the interval $[0,t]$: $\hat{A}_i(t) =_{st} \Lambda_i(\hat{B}(t))$. Evidently, $\hat{B}(t) \to \infty$ w.p.1, and we obtain, as $t \to \infty$,

$$\frac{\Lambda_i(\hat{B}(t))}{t} = \frac{\Lambda_i(\hat{B}(t))}{\hat{B}(t)} \cdot \frac{\hat{B}(t)}{t} \to (\lambda_i + \mu_0^{(i)})\hat{\mathsf{P}}_f, \tag{13}$$

implying

$$\frac{1}{t}[\hat{D}_i(t) - \Lambda_i(\hat{B}(t)] \to \mu_0^{(i)} - (\lambda_i + \mu_0^{(i)})\hat{\mathsf{P}}_f := \delta_i > 0. \tag{14}$$

Notice that $\mathsf{E}\Lambda_i(x) = (\lambda_i + \mu_0^{(i)})x$, and

$$\mathsf{E}\Lambda_i(\hat{B}(t)) = \int_0^t \mathsf{E}\Lambda_i(x)\mathsf{P}(\hat{B}(t) \in dx) = (\lambda_i + \mu_0^{(i)})\mathsf{E}\hat{B}(t),$$

giving, as $t \to \infty$,

$$\frac{\mathsf{E}\Lambda_i(\hat{B}(t))}{t} \to (\lambda_i + \mu_0^{(i)})\hat{\mathsf{P}}_f. \tag{15}$$

Now it follows that

$$\frac{1}{t}\mathsf{E}[\hat{D}_i(t) - \Lambda_i(\hat{B}(t))] \to \delta_i. \tag{16}$$

Since (due to the zero initial state) the number of class-i coloured customers arriving from the orbit to the primary system during interval $[0,\,t]$, $\Gamma_i(t)$, cannot exceed $\Lambda_i(\hat{B}(t))$, we obtain

$$\liminf_{t\to\infty} \frac{1}{t}\mathsf{E}[\hat{D}_i(t) - \Gamma_i(t)] \geq \delta_i. \tag{17}$$

The inequality (17) also holds for an arbitrary initial state $C_i(0) = x_i$, because then $\Gamma_i(t) \leq \Lambda_i(\hat{B}(t)) + x_i$. (We will use this fact below when considering the stability analysis under an arbitrary initial state in Subsection 2.2.) On the other hand, a key observation is that the difference $\Delta_i(t) := \hat{D}_i(t) - \Gamma_i(t)$ is stochastically upper bounded by the number of uncoloured $\mu_0^{(i)}$-arrivals during interval $[0,\,t]$. Denote the total time of absence of i-class coloured customers in orbit i, within interval $[0,\,t]$, by $U_i(t)$. To be more precise, to obtain $U_i(t)$, we couple together the different subperiods (if any) when class-i coloured customers are absent within time interval $[0,\,t]$. Thus, the following stochastic inequality holds

$$\Delta_i(t) \leq_{st} \hat{D}_i(U_i(t)) + 1. \tag{18}$$

By (16), (17) we have

$$\liminf_{t\to\infty} \frac{1}{t}\mathsf{E}\hat{D}_i(U_i(t)) \geq \delta_i. \tag{19}$$

Since $\mathsf{E}\hat{D}_i(t) = \mu_0^{(i)}t$, then by (18),

$$\mathsf{E}\Delta_i(t) \leq \int_0^t \mathsf{E}(\hat{D}_i(x))\mathsf{P}(U_i(t) \in dx) + 1 = \mu_0^{(i)}\mathsf{E}U_i(t) + 1, \tag{20}$$

and it follows that

$$\liminf_{t\to\infty} \frac{1}{t}\mathsf{E}U_i(t) \geq \frac{\delta_i}{\mu_0^{(i)}} > 0. \tag{21}$$

Since

$$\mathsf{E}U_i(t) = \int_0^t \mathsf{P}(C_i(u) = 0)du, \tag{22}$$

we obtain from (21) that

$$\limsup_{t\to\infty} \mathsf{P}(C_i(t) = 0) > 0, \tag{23}$$

and inequality (10) follows. ■

Now we show that (10) indeed implies the statement of Theorem 1. Define the remaining time for regeneration of the process $\{\mathcal{C}(t)\}$ at instant t as

$$T(t) = \min\{T_n - t : T_n - t > 0\}.$$

Lemma 2. *If* (10) *holds then*

$$T(t) \not\Rightarrow \infty, \;\; t \to \infty, \tag{24}$$

where $\Rightarrow$ stands for convergence in probability. The proof of Lemma 2 can be found in the Appendix.

It now follows from (24) and from renewal theory, see [16], that the basic regenerative orbit size process $\{\mathcal{C}(t)\}$ is positive recurrent, that is $\mathsf{E}T < \infty$. The proof of Theorem 1 is hereby completed. ■

Remark 1. It is important to note that the tightness (and even positive recurrence) of the i-th coloured orbit-size process $\{C_i(t),\, t \geq 0\}$ only holds under condition $(\lambda_i + \mu_0^{(i)})\hat{\mathsf{P}}_f < \mu_0^{(i)}$, regardless of whether the other stability conditions (9) hold or not. This is because, in the analysis of orbit i, we do not distinguish between coloured and uncoloured customers of other types, and hence relations (14)-(23) remain unchanged.

Remark 2. The importance of Theorem 1 is defined by a possibility to find the value of the probability $\hat{\mathsf{P}}_f$. In several important cases, the probability $\hat{\mathsf{P}}_f$ can be found in explicit form. For instance, in the important particular case $c = K$ (e.g., in optical networks there is often no buffering) one can use the celebrated Erlang loss formula to obtain

$$\hat{\mathsf{P}}_f = \frac{\hat{\rho}^c/c!}{\sum_{k=0}^{c} \hat{\rho}^k/k!},$$

with $\hat{\rho}$ defined in (5). In [8] an interested reader can find more explicit formulae for $\hat{\mathsf{P}}_f$ in various scenarii.

If in some cases an explicit expression for $\hat{\mathsf{P}}_f$ is not available, one can rely on an estimate of $\hat{\mathsf{P}}_f$, for instance obtained by simulations. Since $\hat{\mathsf{P}}_f$ is a loss probability of a stable finite buffer queueing system, the estimate of $\hat{\mathsf{P}}_f$ can be obtained quickly and with high accuracy. Moreover, as one is mostly interested in the values of $\hat{\mathsf{P}}_f$ for the boundary of the stability region, the value of $\hat{\mathsf{P}}_f$ near the stability boundary is quite high and its estimation is easy (for more details see [6, 7]).

2.2 Stability analysis of the auxiliary system for an arbitrary initial state

Now we show that the statement of Theorem 1 also holds for an arbitrary initial state $\mathcal{C}(0) = \mathbf{x}$ of the system, where $\mathbf{x} = (x_0, \ldots, x_N)$.

Theorem 2. *If the conditions of Theorem 1 hold, then the regenerative process* $\{\mathcal{C}(t),\, t \geq 0\}$ *is positive recurrent under an arbitrary initial state* $\mathcal{C}(0) = \mathbf{x}$.
Proof. It follows from (14) that

$$\liminf_{t\to\infty} \frac{\Delta_i(t)}{t} \geq \delta_i. \tag{25}$$

Moreover, it is easy to see that (25) holds for any fixed state $\mathcal{C}(0) = \mathbf{x}$, see the comment after expression (17). Now we obtain from (18) that

$$\liminf_{t\to\infty} \frac{\hat{D}_i(U_i(t))}{U_i(t)} \frac{U_i(t)}{t} \geq \delta_i, \tag{26}$$

implying

$$\liminf_{t\to\infty} \frac{U_i(t)}{t} > 0. \tag{27}$$

Hence, the total time the coloured orbit i spends in the state $\{0\}$ is infinite w.p.1 under an arbitrary initial state.

Because the total workload process $\{\hat{W}(t),\, t \geq 0\}$ in the primary system (servers and buffer), is stochastically dominated by the tight process $\{\sum_{r=1}^{K-c} \phi_r + \sum_{k=1}^{c} S_k(t),\, t \geq 0\}$, then $\hat{W}$ is also tight and moreover positive recurrent regenerative. (It is easy to show, similar to the above reasoning, by standard arguments using a contradiction. Note that in this proof, as mentioned, we do not distinguish between coloured and uncoloured customers.) Moreover, because the input process is Poisson, then the weak limit $\hat{W}(t) \Rightarrow \hat{W}(\infty)$ exists as well. Denote the generic regeneration period of the process $\hat{W}$ by $\hat{T}$. Then for an arbitrary initial state $\hat{W}(0) = y$ and an arbitrary (fixed) $\hat{\varepsilon} > 0$, there exists a bounded set $\hat{\mathcal{D}}$ such that $\{0, y\} \in \hat{\mathcal{D}}$ and

$$\inf_{t\geq 0} \mathsf{P}(\hat{W}(t) \in \hat{\mathcal{D}}) \geq 1 - \hat{\varepsilon}. \tag{28}$$

Note that w.p.1,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t I(\hat{W}(u) \in \hat{\mathcal{D}})du = \frac{\mathsf{E}\int_0^{\hat{T}} I(\hat{W}(u) \in \hat{\mathcal{D}})du}{\mathsf{E}\hat{T}} = \mathsf{P}(\hat{W}(\infty) \in \hat{\mathcal{D}})\ . \tag{29}$$

Because

$$\frac{1}{t} \int_0^t I(\hat{W}(u) \in \hat{\mathcal{D}})du \leq 1,$$

then by dominance convergence, we can also write that

$$\frac{1}{t}\mathsf{E} \int_0^t I(\hat{W}(u) \in \hat{\mathcal{D}})du = \frac{1}{t} \int_0^t \mathsf{P}(\hat{W}(u) \in \hat{\mathcal{D}})du \to \mathsf{P}(\hat{W}(\infty) \in \hat{\mathcal{D}}) \geq 1 - \hat{\varepsilon}. \tag{30}$$

Thus both processes $\{\hat{W}(t),\, t \geq 0\}$, and $\{C_i(t),\, t \geq 0\}$, are positive recurrent regenerative (and in particular, tight) under arbitrary initial states. Moreover, the coupled process $\{(\hat{W}(t),\, C_i(t)),\, t \geq 0\}$ is also positive recurrent regenerative under an arbitrary initial state $(y,\, x)$. Then, for any $\varepsilon_0 > 0$ there exists a bounded set $\mathcal{D}_i$ such that (in an evident notation)

$$\frac{1}{t}\int_0^t I\Big((\hat{W}(u),\, C_i(u)) \in \mathcal{D}_i\Big)du \to \mathsf{P}\Big((\hat{W}(\infty),\, C_i(\infty)) \in \mathcal{D}_i\Big) \geq 1 - \varepsilon_0 > 0.$$

Hence, the total time the process $\{(\hat{W}(t),\, C_i(t)),\, t \geq 0\}$ spends in the bounded set $\mathcal{D}_i$ is infinite w.p.1. This analysis is immediately extended to the entire process $\{\mathcal{C}(t),\, t \geq 0\}$, and it follows that the time this process spends in a bounded set $\mathcal{B} \subseteq \{0,\, 1\} \times R_+^N$, $\mathbf{0} \in \mathcal{B}$, is infinite w.p.1. On the other hand, it can be shown exactly as in [8] that the time which this process spends in this set during the first regeneration period is finite w.p.1. Then it follows that the first regeneration period is finite w.p.1 under an arbitrary initial state $\mathcal{C}(0)$. It remains to recall that, as it follows from above, the average standard regeneration period (under a zero initial state) is finite as well. It then follows that the process $\mathcal{C}$ is positive recurrent under an arbitrary initial state: $T_1 < \infty \ w.p.1$, and $\mathsf{E}T < \infty$. ■

2.3 Stability analysis of the original retrial system

Let us now formulate and prove our main stability result. Define the indicator $I(t) = 1$ when the primary system in Σ (that is all servers and buffer) is completely full at instant t, and $I(t) = 0$, otherwise. Then the total time, within interval $[0,\, t]$, when the primary system in Σ is completely full (and cannot accept other customers) is defined as

$$B(t) := \int_0^t I(u)du.$$

Let the indicator $J(t) = 0$ if the primary system in Σ is empty at instant t, and $J(t) = 1$, otherwise. Also let $Y_i(t)$ be the i-th orbit size in the system Σ at instant t, $i = 1, \ldots, N$. Define the basic $(N+1)$-dimensional process $\mathcal{Y}(t)$ describing the original system Σ as follows:

$$\mathcal{Y}(t) := \Big(J(t),\, Y_1(t),\, \ldots, Y_N(t)\Big),\ \ t \geq 0.$$

Note that we do not need a Markovian property for the basic process $\{\mathcal{Y}(t),\, t \geq 0,\}$ and, in particular, there are other candidates that could be considered, instead of the component $J(t)$. (This particular choice for $J(t)$ is motivated mainly by the simplicity of the definition of the regenerations of the basic process.) Denote by $\{u_n\}$ the input instants of the merged (Poisson) input of the λ-arrivals (with rate $\sum_i \lambda_i$) in the primary system of Σ. Then, the regeneration instants of the process $\{\mathcal{Y}(t),\, t \geq 0,\}$ are defined as follows:

$$R_0 := 0,\ \ R_{n+1} := \min_k (u_k > R_n : \mathcal{Y}(u_k^-) = \mathbf{0}),\ \ n \geq 0. \tag{31}$$

Theorem 3. *The process $\{\mathcal{Y}(t),\, t \geq 0\}$ is positive recurrent regenerative for any initial state $\mathcal{Y}(0) = x$ under conditions* (9)*, where $\hat{\mathsf{P}}_f$ is the stationary probability of the primary queue being full in the auxiliary system $\hat{\Sigma}$ with exponential service times with parameters $\hat{\mu}_i$ for class-i customers. The parameters $\hat{\mu}_i$ are chosen in such a way that*

$$S^{(i)} \leq_r \exp(\hat{\mu}_i),\ \ i = 1, \ldots, N, \tag{32}$$

where $\exp(\mu)$ *denotes an exponential random variable with parameter* μ*.*

Proof. First we note that under condition (32) the service time distributions satisfy the monotonicity property (1) from [8]. In particular, the primary system in $\hat{\Sigma}$ is always full if the primary system in Σ is full. Now fix some i and observe only class-i customers in both systems. Then, the number of blocked primary class-i customers (or λ_i-*customers*) and the number of the attempts of the blocked class-i customers in the primary system of Σ in interval [0, t] is not larger than the number of *all class-i customers* (coloured) joining virtual orbit i in the auxiliary system $\hat{\Sigma}$, in the same interval. On the other hand, invoking the same monotonicity property, if the primary system in $\hat{\Sigma}$ can accept a new customer, so can the primary system in Σ. This can be expressed as

$$\{t : I(t) = 1\} \subseteq \{t : \hat{I}(t) = 1\},$$

implying $\hat{B}(t) \geq B(t)$ as well. We use coupling when needed to synchronize the arrival instants in both systems and to take the same (corresponding) service times for the customers which arrive to the orbits at the same time instant, and then (if needed) resample service times to have the same service times for the customers entering the primary systems in both Σ and $\hat{\Sigma}$. (Necessary details can be found in the proof of Lemma 1 in [8].) Further, if a (class-i) coloured customer makes a successful attempt to enter the primary system at some instant t, then, provided $Y_i(t) > 0$, a class-i orbit customer enters the primary system in Σ at the same instant. In other words, the number of arrivals to orbit i (departures from the orbit) in the system Σ is not larger (is not less, within the periods where the orbit is not empty) than that in system $\hat{\Sigma}$. Note that the coloured customers go to the front of the queue of the virtual orbit i and behave exactly as the blocked customers in Σ. In particular, when such a customer makes an unsuccessful attempt to enter the primary system, he immediately rejoins the orbit, exactly as in the original system Σ. Thus, the orbit sizes are ordered as $Y_i(t) \leq_{st} C_i(t)$, and this evidently holds for any value of i. It remains to recall that the processes $\{C_i(t), t \geq 0\}$ are positive recurrent regenerative. This completes the proof. ■

We would like to note that the derived sufficient conditions are not too restrictive. In fact, they coincide with the necessary conditions in the important particular case of a single server, no additional waiting space, and a symmetric parameter setting. Let us recall from [7] the necessary stability conditions for the case of a single server and no additional waiting space:

$$(\lambda_i + \mu_0^{(i)}) \sum_{j=1}^{N} \rho_j < \mu_0^{(i)}, \; i = 1, \ldots, N, \tag{33}$$

where $\rho_j := \lambda_j/\mu_j$. In this particular case, the probability of the primary queue being full in the auxiliary system, $\hat{\mathsf{P}}_f$, has a simple expression

$$\hat{\mathsf{P}}_f = \frac{\sum_j (\rho_j + \hat{\rho}_j)}{1 + \sum_j (\rho_j + \hat{\rho}_j)}, \tag{34}$$

where $\hat{\rho}_j := \mu_0^{(j)}/\mu_j$. If we consider the symmetric case $\rho_j = \rho$ and $\hat{\rho}_j = \hat{\rho}$, we can write the sufficient conditions as a single inequality

$$(\rho + \hat{\rho}) \frac{N(\rho + \hat{\rho})}{1 + N(\rho + \hat{\rho})} < \hat{\rho},$$

or, equivalently,

$$\rho N(\rho + \hat{\rho}) + \hat{\rho} N(\rho + \hat{\rho}) < \hat{\rho} + \hat{\rho} N(\rho + \hat{\rho}),$$

which in fact coincides with the necessary condition

$$(\rho + \hat{\rho}) N \rho < \hat{\rho}.$$

The above derivation implies that the obtained sufficient conditions are also necessary in the case of symmetric classes and in the case of a single class.

Since we have used the rough dominating system for establishing sufficient conditions, there is no reason to expect that in general our sufficient conditions are also necessary. And indeed, this is not the case. We can see how they are different on the example of a two-class, single-server and bufferless retrial queue. The necessary stability conditions [7] are given by

$$(\rho_i + \hat{\rho}_i)(\rho_1 + \rho_2) < \hat{\rho}_i, \quad i = 1, 2,$$

or, equivalently,

$$\hat{\rho}_i > \rho_i \frac{\rho_1 + \rho_2}{1 - \rho_1 - \rho_2} \quad i = 1, 2.$$

This corresponds to the quadrant in Figure 1. In fact, in [9] it was shown that these conditions are necessary and sufficient in the case of exponential service.

On the other hand, using the explicit expression (34) for $\hat{\mathsf{P}}_f$, we can transform the sufficient conditions to the following form

$$\rho_i[(\rho_1 + \hat{\rho}_1) + (\rho_2 + \hat{\rho}_2)] < \hat{\rho}_i, \quad i = 1, 2.$$

This is a set of linear inequalities with respect to $\hat{\rho}_i, i = 1, 2$, and the stability region is depicted as the shaded area in Figure 1. We can therefore deduce that in the general non-symmetric case, the sufficient conditions that we derived are not (always) necessary. However, in the case of a system with buffer (meaning additional waiting spaces), deriving necessary and sufficient conditions can be very challenging.

We note that if in the bufferless multi-orbit symmetric single-server system the sufficient conditions (9) hold, then it can be easily verified (by summing the inequalities in (9) and using (34)) that the necessary conditions (33) hold as well.

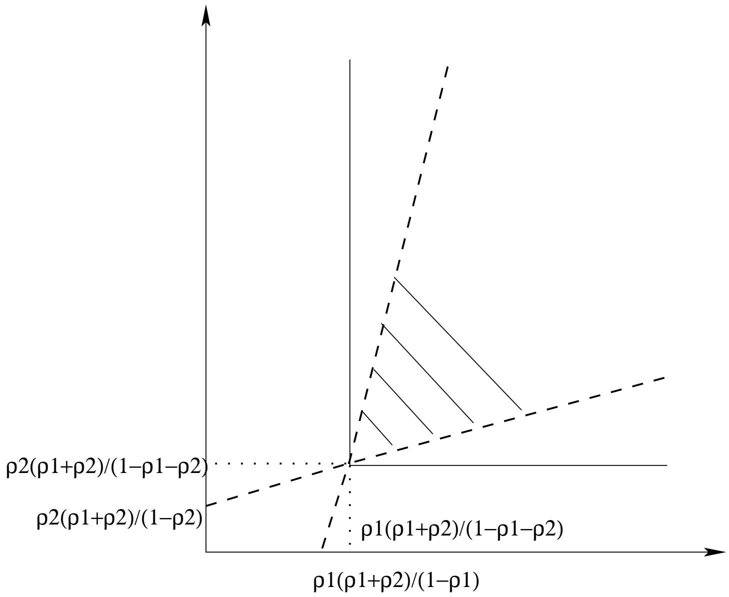

Figure 1: Stability regions for the 2-class single-server model.

3 Analysis of a partially unstable regime

In this section, we analyze the multi-class constant retrial rate *single-server bufferless system* Σ, with i.i.d. exponential service times $\{S_n^{(i)},\, n \geq 1\}$ (with rate μ_i) for class-i customers, $i = 1, \ldots, N$, in some specific cases where the stability conditions are violated. Denote by $\rho_i = \lambda_i/\mu_i$, $i = 1, \ldots, N$. It is proven in [7] that in the stationary original N-orbit constant retrial rate system the busy probability is equal to $P_b = \sum_{i=1}^N \rho_i$, and that the necessary stability (positive recurrence) conditions are

$$(\lambda_i + \mu_0^{(i)}) \sum_{i=1}^{N} \rho_i < \mu_0^{(i)},\ i = 1, \ldots, N, \tag{35}$$

where $\rho_i := \lambda_i/\mu_i$. Thus, if at least one inequality in the above conditions is violated, then the basic process $\{\mathcal{Y}(t),\, t \geq 0\}$ can not be positive recurrent. In other words, denoting by R a generic regeneration period of system Σ (see (31)), we obtain that $\mathsf{E}R = \infty$. Also recall that in the associated loss system $\hat{\Sigma}$, see (34),

$$\hat{\mathsf{P}}_f = \frac{\sum_{i=1}^N (\rho_i + \hat{\rho}_i)}{1 + \sum_{i=1}^N (\rho_i + \hat{\rho}_i)},$$

with $\hat{\rho}_i := \mu_0^{(i)}/\mu_i$. Finally, we note that the exponential service times satisfy stochastic ordering (1); see [30].

Theorem 4. *Consider an initially empty system, $\mathcal{Y}(0) = 0$. Assume that*

$$(\lambda_j + \mu_0^{(j)}) \sum_i \rho_i \quad \geq \quad \mu_0^{(j)}, \tag{36}$$

for some j, and that

$$(\lambda_i + \mu_0^{(i)})\hat{\mathsf{P}}_f \quad < \quad \mu_0^{(i)},\ i = 1, \ldots, N,\ i \neq j. \tag{37}$$

Then the orbit size $Y_j(t) \Rightarrow \infty$, the orbit sizes $\{Y_i(t),\, t \geq 0\}$, $i \neq j$, are tight, and there exists the limiting busy probability $\mathsf{P}_b := \lim_{t\to\infty} \mathsf{E}B(t)/t$ such that

$$\mathsf{P}_b = \frac{\lambda_j + \mu_0^{(j)} + \mu_j \sum_{i\neq j} \rho_i}{\lambda_j + \mu_0^{(j)} + \mu_j}, \tag{38}$$

if $(\lambda_j + \mu_0^{(j)}) \sum_i \rho_i > \mu_0^{(j)}$, and

$$\mathsf{P}_b = \sum_{i=1}^{N} \rho_i, \tag{39}$$

if $(\lambda_j + \mu_0^{(j)}) \sum_i \rho_i = \mu_0^{(j)}$.

Proof. First, for simplicity of presentation, assume that $N = 2$, and that

$$(\lambda_1 + \mu_0^{(1)})\hat{\mathsf{P}}_f \quad < \quad \mu_0^{(1)}, \tag{40}$$

$$(\lambda_2 + \mu_0^{(2)})(\rho_1 + \rho_2) \quad \geq \quad \mu_0^{(2)}. \tag{41}$$

Let $V_i(t)$ be the class-i workload that arrived in interval $[0,\, t]$, $i = 1, 2$. We can write the balance equation

$$V(t) = V_1(t) + V_2(t) = \hat{W}(t) + W_1(t) + W_2(t) + B(t), \tag{42}$$

where $W_i(t)$ is the workload in orbit i at instant t, and $\hat{W}(t)$ is the workload in the primary queue at instant t. Since, under condition (40), the class-1 orbit size in system Σ is (stochastically) dominated by the positive recurrent class-1 coloured orbit size in system $\hat{\Sigma}$, then in particular, $W_1(t) = o(t)$, $t \to \infty$ w.p.1, and moreover the process $\{W_1(t),\, t \geq 0\}$ is tight. Thus, only orbit 2 maybe a "source of instability". Then it is easy to show by contradiction (see the proof of Theorem 1 after formula (10)) that the 2^{nd} orbit size $Y_2(t) \Rightarrow \infty$, and in particular, $\mathsf{P}(Y_2(t) = 0) \to 0,\ \ t \to \infty$. Define

$$I_2(t) = \int_0^t I(Y_2(u) = 0)du\ , \tag{43}$$

the total time, within interval $[0,\, t]$, when the $2nd$ orbit is empty. Then it follows that

$$\mathsf{E}I_2(t) = o(t),\ t \to \infty. \tag{44}$$

Denote by $\Delta_2(t) \geq 0$ the difference between the number of $\mu_0^{(2)}$-arrivals in system $\hat{\Sigma}$ and the number of class-2 retrial attempts in system Σ, in interval $[0,\, t]$. Because this difference is caused by the number of $\mu_0^{(2)}$-arrivals arriving (in system $\hat{\Sigma}$) during the idle time of orbit 2 (in system Σ), in interval $[0,\, t]$, then

$$\mathsf{E}\Delta_2(t) = \mu_0 \mathsf{E}I_2(t) = o(t),\ \ t \to \infty. \tag{45}$$

Finally, denote by $\Delta W_2(t)$ the difference between the workloads generated by $\mu_0^{(2)}$-arrivals (in $\hat{\Sigma}$), and class-2 retrial attempts (in Σ), caused by the difference $\Delta_2(t)$. Thus,

$$\Delta W_2(t) =_{st} \sum_{n=1}^{\Delta_2(t)} S_n^{(2)},$$

and (using an independence between $S_n^{(2)}$ and the event $\{\Delta(t) > n-1\}$) we obtain

$$\mathsf{E}\Delta W_2(t) = \mathsf{E}\Delta_2(t)\mathsf{E}S^{(2)} = o(t),\ \ t \to \infty. \tag{46}$$

Denote by $B_2(t)$ the total busy time of the server in system Σ_2, which differs from the original system Σ only in that system Σ_2 has a *permanently loaded orbit 2*. In other words, system Σ_2 is fed (besides by class-1 customers) by the *Poisson input of* $\mu_0^{(2)}$*-customers*. Moreover, in system Σ_2, we will not colour blocked $\mu_0^{(2)}$-customers and may assume that such customers either join an infinite virtual orbit (being the source of $\mu_0^{(2)}$-input) or, equivalently, are lost. Note that the class-1 orbit size process in system Σ_2 satisfies condition (40), and, in spite of a similarity between both systems, Σ_2 obeys a nice regeneration property (unlike system Σ). In particular, the process $\{B_2(t),\, t \geq 0\}$ describing the accumulated busy time of the servers in interval $[0,\, t]$ in system Σ_2 is positive recurrent regenerative. Then the limits

$$\lim_{t\to\infty} \frac{\mathsf{E}B_2(t)}{t} = \lim_{t\to\infty} \frac{\mathsf{E}B(t)}{t} = \mathsf{P}_b, \tag{47}$$

exist, where we take into account that, from (46),

$$\mathsf{E}B(t) = \mathsf{E}B_2(t) + o(t),\, t \to \infty.$$

It remains to find P_b in an explicit form. First we return to equation (42) and notice that, by regenerative arguments [29], w.p.1,

$$\frac{V(t)}{t} \to \rho_1 + \rho_2,\ \ \frac{\hat{W}(t)}{t} \to 0,\ \ \frac{W_1(t)}{t} \to 0, \tag{48}$$

and moreover, the convergence in average holds as well:

$$\frac{\mathsf{E}V(t)}{t} \to \rho_1 + \rho_2, \quad \frac{\mathsf{E}\hat{W}(t)}{t} \to 0, \ \frac{\mathsf{E}W_1(t)}{t} \to 0. \tag{49}$$

Thus, it follows from (42) that the following limit exists as well:

$$\frac{\mathsf{E}W_2(t)}{t} \to \rho_1 + \rho_2 - \mathsf{P}_b, \tag{50}$$

which shows, in particular, that in all cases $\mathsf{P}_b \le \rho_1 + \rho_2$, see [7]. Moreover, if $\mathsf{E}W_2(t)/t \to 0$ then $\mathsf{P}_b = \rho_1 + \rho_2$. On the other hand, $W_2(t)$ is the difference between the workload that arrived in orbit 2 and the workload that departed from orbit 2, in interval $[0,\, t]$. Denote by $A(t) = t - B(t)$ the time when the server is free in the interval $[0,\, t]$. Note that the workload that arrived in orbit 2 in $[0,\, t]$ is stochastically equivalent to the workload $V_2(B(t))$ that arrived in the system during the busy time $B(t)$, and the workload that departed from orbit 2 in $[0,\, t]$ is stochastically equivalent to the workload that customers, generated by the Poisson process with rate $\mu_0^{(2)}$ (and with service rate μ_2), bring in the server in interval $[0,\, A(t)]$, *provided orbit 2 is not empty.* Let now indicator $I(t) = 0$ if the server (in the original system) is empty at time instant t. Then the time when both orbit 2 and the server are empty, is defined as the length

$$\beta_2(t) := \#\{s \in [0,\, t] : W_2(s) = 0,\, I(s) = 0\} \le I_2(t).$$

Then it follows from (44) that $\mathsf{E}\beta_2(t) = o(t)$, $t \to \infty$. Since

$$W_2(t) =_{st} V_2(B(t)) - [\hat{D}_2(A(t)) - \hat{D}_2(\beta_2(t))], \tag{51}$$

where $\hat{D}$ denotes the Poisson process with rate $\mu_0^{(2)}$, we obtain

$$\mathsf{E}W_2(t) = \frac{1}{\mu_2}\Big[\int_0^t \lambda_2 x \mathsf{P}(B(t) \in dx) - \int_0^t \mu_0^{(2)} x \mathsf{P}(A(t) \in dx)\Big] + o(t). \tag{52}$$

This, in turn, implies

$$\frac{\mathsf{E}W_2(t)}{t} \to \frac{\lambda_2 \mathsf{P}_b - \mu_0^{(2)}(1 - \mathsf{P}_b)}{\mu_2}. \tag{53}$$

By (50) and (53) we finally obtain, if $(\lambda_2 + \mu_0^{(2)})(\rho_1 + \rho_2) > \mu_0^{(2)}$,

$$\mathsf{P}_b = \frac{\lambda_2 + \mu_0^{(2)} + \rho_1\mu_2}{\lambda_2 + \mu_0^{(2)} + \mu_2}, \tag{54}$$

where $\rho_1 < 1$ (by the positive recurrence of the workload process $Y_1(t)$, $t \ge 0$, in the system Σ_2). Otherwise, if $(\lambda_2 + \mu_0^{(2)})(\rho_1 + \rho_2) = \mu_0^{(2)}$, then $\mathsf{P}_b = \rho_1 + \rho_2$. This analysis is immediately extended to a general N-orbit system under the assumptions of Theorem 4, in which case the orbit sizes $Y_i(t)$, $i \ne j$, are tight, being dominated by the positive recurrent orbit size processes in the associated system where the input from orbit j is replaced by Poisson input with rate $\mu_0^{(j)}$ (see the definition of system Σ_2), while the orbit size $Y_j(t) \Rightarrow \infty$. In particular, it implies that $\sum_{i \ne j} \rho_i < 1$. Then, directly repeating previous arguments, we obtain (38) and (39). Finally, we recall that the tightness of the processes $\{Y_i(t),\, t \ge 0\}$, $i \ne j$ and the convergence $Y_j(t) \Rightarrow \infty$ have been established above as well. ■

Remark 3. The analysis of the unstable regime developed above is simplified for the single-orbit system (with Poisson input with rate λ, exponential service time with rate μ and orbit rate μ_0). Denote $\rho = \lambda/\mu$. Because in this case

$$\hat{\mathsf{P}}_b = \frac{\lambda + \mu_0}{\lambda + \mu_0 + \mu}, \tag{55}$$

then the necessary stability condition $\lambda\rho < \mu_0(1-\rho)$ coincides with the sufficient condition $(\lambda + \mu_0)\hat{\mathsf{P}}_b < \mu_0$. As a result, we can easily find that if in this system $\lambda\rho = \mu_0(1-\rho)$, then $\mathsf{P}_b = \rho$, and if $\lambda\rho > \mu_0(1-\rho)$, then $\mathsf{P}_b = \hat{\mathsf{P}}_b$.

4 Conclusion

We have analyzed multi-class retrial queueing systems with constant retrial rates, Poisson inputs, general service times, and an arbitrary numbers of servers and waiting places. Our model is motivated by several telecommunication applications such as multiple access protocols, TCP routing and optical switching networks, but also represents independent theoretical interest. Using a regenerative approach, we have provided sufficient stability conditions which have a clear probabilistic interpretation. We show that the provided sufficient conditions are in fact also necessary in the case of a single-server system without waiting space and in the case of symmetric classes. We have also studied a very interesting situation of "partial instability", when the orbit length of one class goes to infinity while the orbit lengths of the other classes remain tight. To the best of our knowledge, this situation has not been considered before in the literature. In fact, such "partial instability" can be viewed as a good property of the system that ensures a significant level of fairness. If one class of customers violates its service contract, this class will be severely punished while the other classes will suffer only slightly. This effect deserves special investigation for a broader class of systems. Of course, another obvious future research direction is refining the sufficient conditions or even establishing necessary and sufficient conditions. The analysis of stability in more complex settings can be done by simulations to explore the effect of various system parameters.

Acknowledgements

The authors thank the anonymous referees for a careful reading of the first version of the paper.

KA acknowledges the support of EU COST ACROSS IC 1304. The research of EM is supported by the Russian Foundation for Basic Research, projects 15-07-02341, 15-07-02354, 15-07-02360, the Program of strategic development of Petrozavodsk State University and Campus France. EM also thanks the organising committee of the ECQT2014 for the support of his visit to the conference. The research of BS has been funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

References

[1] Artalejo, J.R. (1996). Stationary analysis of the characteristics of the M/M/2 queue with constant repeated attempts. *Opsearch*, **33**, 83–95.

[2] Artalejo, J.R., Gómez-Corral, A., and Neuts, M.F. (2001). Analysis of multiserver queues with constant retrial rate. *European Journal of Operational Research*, **135**, 569–581.

[3] Artalejo, J.R. and Gómez-Corral, A. (2008). *Retrial Queueing Systems: A Computational Approach.* Springer.

[4] Avrachenkov, K., and Yechiali, U. (2008). Retrial networks with finite buffers and their application to Internet data traffic. *Probability in the Engineering and Informational Sciences*, **22**, 519–536.

[5] Avrachenkov, K., and Yechiali, U. (2010). On tandem blocking queues with a common retrial queue. *Computers and Operations Research*, **37(7)**, 1174–1180.

[6] Avrachenkov, K., Goricheva, R. S., Morozov, E. V. (2011). Verification of stability region of a retrial queuing system by regenerative method. *Proceedings of the International Conference "Modern Probabilistic Methods for Analysis and optimization of Information and Telecommunication Networks" Minsk*, 22–28.

[7] Avrachenkov, K., Morozov, E., Nekrasova, R., Steyaert, B. (2014). Stability analysis and simulation of N-class retrial system with constant retrial rates and Poisson inputs. *Asia-Pacific Journal of Operational Research*, **31(2)** 1440002 (18 pages). World Scientific Publishing Co. and Operational Research Society of Singapore. DOI: 10.1142/S0217595914400028

[8] Avrachenkov, K. and Morozov, E. (2014). Stability analysis of GI/G/c/K retrial queue with constant retrial rate. *Math. Meth. Oper. Res.*, **79**, 273–291. DOI 10.1007/s00186-014-0463-z.

[9] Avrachenkov, K., Nain, P., and Yechiali, U. (2014). A retrial system with two input streams and two orbit queues. *Queueing Systems*, **77(1)**, 1–31.

[10] Choi, B.D., Shin, Y.W., and Ahn, W.C. (1992). Retrial queues with collision arising from unslotted CSMA/CD protocol. *Queueing Systems*, **11**, 335–356.

[11] Choi, B.D., Rhee, K.H., and Park, K.K. (1993). The M/G/1 retrial queue with retrial rate control policy. *Probability in the Engineering and Informational Sciences*, **7**, 29–46.

[12] Dai, J. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *The Annals of Applied Probability*, **5**, 49–77.

[13] Falin, G.I. and Templeton, J.G. (1997). *Retrial queues.* Springer.

[14] Falin, G.I. (1988). On a Multiclass Batch Arrival Retrial Queue *Advances in Applied Probability*, **20**, 483–487.

[15] Fayolle, G. (1986). A simple telephone exchange with delayed feedback. In Boxma, O.J., Cohen J.W., and Tijms, H.C. (eds.), *Teletraffic Analysis and Computer Performance Evaluation*, **7**, Elsivier Science Publishers B. V., Amsterdam, 245–253.

[16] Feller, W. (1971). *An Introduction to Probability Theory and its Applications. I* (2*nd* ed.), New York: Wiley.

[17] Kulkarni, V.G., (1986) Expected waiting times in a multiclass batch arrival retrial queue. *Journal of Applied Probability*, **23**, 144–154.

[18] Langaris C. and Dimitriou I. (2010) A queueing system with n-phases of service and $(n-1)$-types of retrial customers. *European Journal of Operational Research*, **205**, 638–649.

[19] Lillo, R.E. (1996). A G/M/1-queue with exponential retrial. *TOP*, **4(1)**, 99–120.

[20] Morozov, E. (2004). Weak regeneration in modeling of queueing processes. *Queueing Systems*, **46**, 295–315.

[21] Morozov, E. (2007). A multiserver retrial queue: Regenerative stability analysis. *Queueing Systems*, **56**, 157–168.

[22] Morozov, E. and Steyaert, B. (2013). Stability analysis of a two-station cascade queueing network. *Annals of Operations Research*, **13**, 1–26.

[23] Morozov, E., Fiems, D., and Bruneel, H. (2011). Stability analysis of multiserver discrete-time queueing systems with renewal-type server interruptions, *Performance Evaluation*, **68**, 1261–1275.

[24] Morozov, E. (1997). The tightness in the ergodic analysis of regenerative queueing processes. *Queueing Systems*, **27**, 179–203.

[25] Morozov, E. and Delgado, R. (2009). Stability analysis of regenerative queues, *Automation and Remote control*, **70(12)**, 1977–1991.

[26] Moutzoukis E. and Langaris C. (1996). Non-preemptive priorities and vacations in a multiclass retrial queueing system. *Stochastic Models*, **12(3)**, 455–472.

[27] Ramalhoto, M.F., and Gómez-Corral, A. (1998). Some decomposition formulae for $M/M/r/r+d$ queues with constant retrial rate. *Stochastic Models*, **14**, 123–145.

[28] Shin, Y.W., and Moon, D.H. (2013). M/M/c Retrial Queue with Multiclass of Customers. *Methodology and Computing in Applied Probability*, 1–19.

[29] Smith, W.L. (1955). Regenerative stochastic processes. *Proceedings of Royal Society, Ser. A*, **232**, 6–31.

[30] Whitt, W. (1981). Comparing counting processes and queues. *Advances in Applied Probability*, **13**, 207–220.

[31] Wong, E.W.M., Andrew, L.L.H., Cui, T., Moran, B., Zalesky, A., Tucker, R.S., and Zukerman, M. (2009). Towards a bufferless optical internet. *Journal of Lightwave Technology*, **27**, 2817–2833.

[32] Yao, S., Xue, F., Mukherjee, B., Yoo, S.J.B., and Dixit, S. (2002). Electrical ingress buffering and traffic aggregation for optical packet switching and their effect on TCP-level performance in optical mesh networks. *IEEE Communications Magazine*, **40(9)**, 66–72.

5 Appendix

The proof of Lemma 2. Recall definition (11) and define the remaining time for regeneration at instant t as

$$T_i(t) = \min_n (T_n^{(i)} - t : T_n^{(i)} - t > 0). \tag{56}$$

Thus, at each instant z_j (satisfying (10)), the coloured orbit i is empty, and we first show that, with a positive probability, the next arrival can start a new i-type regeneration period. That is, its arrival instant will be the first regeneration point, $T_k^{(i)}$, for some value of k, after instant z_j. Denote the remaining service time at server k as $S_k(t)$ ($S_k(t) = 0$, if the server is empty), $k = 1, \dots, c$. Then it is known that the processes $\{S_k(t),\, t \ge 0\}$ are tight for any value of k [24]. Because the sequence $\{z_j\}$ is deterministic, the sequences $\{S_k(z_j),\, j \ge 1\}$ are tight for any value k as well. Also note that the number of customers waiting in the buffer is limited by $K - c$. Denote by $S^{(i)}$ a generic service time of a class-i customer. Then the service time of *any customer* in the system is (stochastically) upper bounded by the random variable $\phi := \max_i S^{(i)}$ with finite mean. It then follows that the workload accumulated in the buffer is (stochastically) upper bounded by $\phi_1 + \cdots + \phi_{K-c}$, where $\{\phi_i\}$ are i.i.d. random variables, distributed as ϕ. It now follows from the tightness and from (10) that there exists a constant $C < \infty$ such that

$$\min_j \mathsf{P}\Big(C_i(z_j) = 0,\, \sum_{r=1}^{K-c} \phi_r + \sum_{k=1}^{c} S_k(z_j) \le C\Big) \ge \frac{\varepsilon}{2}. \tag{57}$$

We note that the event within parentheses in (57) means that the i-th coloured orbit and the primary system become empty at instant $t + C$, provided no new customer arrives in interval $[z_j,\, z_j + C]$. If, in addition, a new class-i customer arrives in the interval $[z_j + C,\, z_j + C + x]$, then a regeneration is initiated in this interval. Consequently, it easily follows that for any $x \ge 0$

$$\inf_j \mathsf{P}(T_i(z_j) \le x + C) \ge \frac{\varepsilon}{2}(e^{-(\lambda_i+\mu_0^{(i)})C} - e^{-(\lambda_i+\mu_0^{(i)})(C+x)})e^{-\sum_{l\neq i}(\lambda_l+\mu_0^{(l)})(C+x)} > 0, \tag{58}$$

where the product of the two exponential terms in the right-hand side represents the probability that class-i and only class-i customers arrive in the interval $[z_j + C,\, z_j + C + x]$. Since the lower bound is uniform in j, we have that $T_i(t) \not\Rightarrow \infty$, and it follows that $\mathsf{E}T_i < \infty$, where T_i is the generic regeneration period of the process $\{C_i(t),\, t \ge 0\}$ [16]. In particular, the process $\{C_i(t)\}$ is tight. This conclusion holds for any value of i, and thus the summary orbit size, $\{\sum_i C_i(t) := \mathbf{C}(t),\, t \ge 0\}$, is a tight process as well. In particular, for any value of $\varepsilon_1 > 0$, there exists a constant $\mathcal{C}_0$ such that

$$\inf_t \mathsf{P}(\mathbf{C}(t) \le \mathcal{C}_0) \ge 1 - \varepsilon_1. \tag{59}$$

As a next step we show that, at each instant t, it is possible to unload both the primary system and all coloured orbits, to obtain a new regeneration point T_n in an interval $[t,\, t + D]$ with a positive probability, where D is a finite constant. The main idea is to unload coloured orbit 1, then orbit 2, etc., within interval $[t,\, t + D]$, provided that *no new arrivals occur during this interval.* Recall that the remaining service times $S_k(t)$ are tight, so, by (59), we can take a finite constant $\mathcal{D}_0$ such that

$$\inf_t \mathsf{P}(\mathbf{C}(t) \le \mathcal{C}_0,\, \sum_{r=1}^{K-c} \phi_r + \sum_{k=1}^{c} S_k(t) \le \mathcal{D}_0) \ge 1 - \frac{\varepsilon_1}{2}. \tag{60}$$

Fix an arbitrary instant t and a constant $\Delta > 0$ and denote by

$$\gamma = \min_i (1 - e^{-\Delta \mu_0^{(i)}}).$$

It is easy to see that the probability that a coloured customer (if any) makes an attempt to enter the primary system in interval $[t,\, t+\Delta]$ is not less than γ. Note that $\mathsf{P}(C_i(t) \le \mathcal{C}_0) \ge 1 - \varepsilon_1$ for any value of i. Now, for a given constant $\zeta > 0$, take constant a such that $\mathsf{P}(S^{(i)} \le a) \ge \zeta,\ i = 1, \dots, N$. Assume that an orbit customer enters an empty primary system at an instant z. Then, provided no new arrivals and new retrial attempts from other non-empty orbits (if any) occur, the next orbit customer can start service in interval $[z,\, z + a + \Delta]$ with a probability which is lower bounded by $\zeta\Delta$. Denote by $\sigma = (\zeta\Delta)^{\mathcal{C}_0}$ and note that σ is a lower bound of the probability that at least $\mathcal{C}_0$ orbit customers of any class can be served one by one, provided the primary system is empty and no new arrivals enter the system. Consider for a moment the case of two orbits, that is $N = 2$. Recall that, for convenience only, we unload the orbits in increasing order, that is first orbit 1, followed by orbit 2. Thus, *provided no new arrivals of both classes and no retrial attempts of class-*2 *customers happened since instant* t, orbit 1 becomes empty during interval

$$[t,\, t + \mathcal{D}_0 + \mathcal{C}_0(\Delta + a)],$$

with a probability which is not less than

$$(1 - \frac{\varepsilon_1}{2})\sigma > 0.$$

In a similar way we can completely unload the second orbit during the interval

$$[t,\, t + \mathcal{D}_0 + 2\mathcal{C}_0(\Delta + a)],$$

with a probability which is not less than $(1 - \varepsilon_1/2)\sigma^2$, provided no new arrivals/attempts happened since instant t. If we take now into account the probability of the event $\{$*no new arrivals/attempts happened since instant* $t\}$, then we obtain that the probability to unload the primary system and both coloured orbits in this interval is lower bounded by the quantity

$$(1 - \varepsilon_1/2)\sigma^2 \exp\Big\{-\Big(\mathcal{D}_0 + 2\mathcal{C}_0(\Delta + a)\Big)\sum_{i=1}^{2}(\lambda_i + \mu_0^{(i)})\Big\} > 0, \tag{61}$$

which is independent of t. Note that this lower bound is definitely not tight but simple and suitable for our purpose. Consider now the general case of N orbits. Continuing in a similar way, we find that both the primary system and all coloured orbits are completely unloaded in the interval

$$[t,\, t + \mathcal{D}_0 + N\mathcal{C}_0(\Delta + a)], \tag{62}$$

with a probability which is lower bounded by

$$(1 - \varepsilon_1/2)\sigma^N \exp\Big\{-\Big(\mathcal{D}_0 + N\mathcal{C}_0(\Delta + a)\Big)\sum_{i=1}^{N}(\lambda_i + \mu_0^{(i)})\Big\} > 0. \tag{63}$$

It is then easy to see that in a finite interval, with a positive probability, a new customer arrives observing an empty primary system and all coloured orbits empty. Consequently, (24) follows from this property. ■