Alcom-FT Technical Report Series
ALCOMFT-TR-01-10

ANALYSIS OF A MULTI-CLASS QUEUEING SYSTEM

JEAN-FRANÇOIS DANTZER AND PHILIPPE ROBERT

ABSTRACT. The stability properties of the bandwidth allocation algorithm First Fit are analyzed for the distributions concentrated on three sizes for the requests. We give the explicit expression of the ergodicity condition of this model; it involves a quadratic functional of the input parameters. The stochastic processes describing these systems are string valued Markov process. To analyze these processes we introduce the notion of a smooth initial state. Starting from a smooth initial state the fluid limits of this system can be investigated. The fluid limits of interest are random dynamical systems in $\mathbb{R}^2$ which are products of random 2×2 matrices.

CONTENTS

1. INTRODUCTION

The model we consider here is a simplified description of a bandwidth allocation scheme, i.e. the allocation of different streams of messages in a communication network. The arriving messages are of different nature; to be transmitted they require different throughputs, i.e. variable portions of the offered bandwidth C of the network. The sum of throughputs required by the messages being transmitted at a given time must be less than C. If they are not being transmitted, the messages are stored in an infinite buffer in their order of arrival. When a message has finished its transmission, messages in the queue can be transmitted if there is enough room in the network, i.e. if the quantity C minus the sum of the throughputs of the messages being transmitted is large enough. The allocation algorithm considered here is the *First Fit Algorithm*: A message in the queue is allocated if its throughput

Date: December 6, 2000.

1991 *Mathematics Subject Classification.* Primary 60K25 90B12; Secondary 60J75 90B22.

Key words and phrases. Bin Packing Algorithms. Dynamic Bin-Packing. Ergodicity. Fluid Limits. Multi-class Queueing Systems. Bandwidth Allocation.

Partially supported by the IST Programme of the EU under contract number IST-1999-14186 (ALCOM-FT).

is less than the available bandwidth at that time and none of the other messages arrived before it in the queue can be transmitted.

For convenience we shall use the bin packing terminology: the network is a bin of size C, messages are items and the bandwidth required by a message is the size of the item. Items have the same distribution as some random variable S_1. A stream of such items arrives at rate λ at the bin and each item requires a service of mean 1. In this setting the First Fit algorithm can be described as follows: The sum of the items in the bin is less than C, the size of the bin. Following every event (arrival or departure), the queue is scanned from the beginning in search of an item whose size is smaller than the empty space left in the bin. This procedure is repeated until the end of the queue is reached. An item in the bin is served at speed 1. As we shall see, the probabilistic description of this model is not easy to handle; it involves an infinite dimensional vector space (a space of strings).

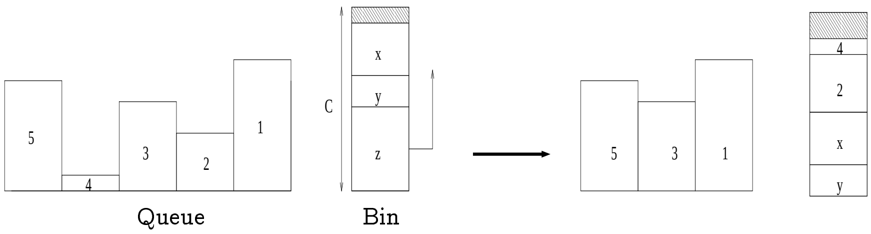

FIGURE 1. A departure for the First Fit algorithm

Related models. A similar problem has been analyzed by Kipnis and Robert [19] with the FIFO algorithm: An item enter in the bin only if all the items arrived before him have left the queue. In particular an item in the queue cannot access the bin if it is not the first item in the queue. The stability problem is simpler in this case: the vector of the sizes of the items in the bin and the size of the first item in the queue is a Markov process. The lengths of the items in the queue, the first one excepted, are i.i.d. random variables with distribution μ. To study the maximal throughput of this model, it is sufficient to calculate the output of the bin when the queue is saturated, i.e. when it contains an infinite number of items. For the First Fit algorithm the situation is quite different. Since the queue is scanned to accommodate items in the bin, the sizes of the items in the queue are unlikely to remain independent and with the same initial distribution μ. For example there should be less small items at the beginning of the queue than at the end. Furthermore, if we saturate the queue, the output will not give the maximal output of the queue: if the size of the items are uniformly distributed on $[0, 1]$, an infinite number of small items will be in the bin generating an infinite output.

Coffman and Stolyar [6] analyzed the stability of the algorithms First Fit and Best Fit when the services are constant equal to 1. Because of the hypothesis on the services, the problem is related to static bin packing problems. They prove that the natural condition $\lambda E(S_1) < C$ is sufficient for the stability in the case of a symmetrical distribution of the sizes; in Coffman *et al.* [5] the sufficiency for stability of the condition $\lambda E(S_1) < C$ is considered in a more complex communication network.

The First Fit algorithm with items having two possible sizes has been analyzed in Dantzer *et al.* [9]. In that paper the stability condition has been established and, more interesting, a curious transient behavior analyzed. The present paper is a continuation of this work. The case we analyze here requires a much more detailed analysis of the evolution of the string structure than that was necessary in [9].

Markov processes on strings occur also naturally in the multi-class queueing networks. Rybko and Stolyar [28], Bramson [2, 3] analyzed such Markov processes. In their work, due to clever arguments, the string structure is not really taken into account in their analysis. Dumas [11] presented an analysis of the fluid equations for the string structure of Bramson's networks.

In the same vein, Malyshev *et al.* [16, 23] and Serfozo [30] (and the references therein) investigated quite general models with strings but with a dynamic depending only on a finite number of components at the end (or the beginning) of the string. For the string valued processes we consider here, the dynamic depends, a priori, of all the components of the string since the queue is scanned as long as there is an empty space in the bin.

Assumptions. The items arrive according to a Poisson process $\mathcal{N}_\lambda$ with parameter λ; for $t \geq 0$, the quantity $\mathcal{N}_\lambda(]0,t])$ denotes the number of arrivals between 0 and t. The capacity of the bin C is equal to 4.

The sizes (S_i) of the items form an i.i.d. sequence with a common distribution $F(dx)$ given by

$$F(dx) = p\delta_1 + q\delta_2 + r\delta_3$$

where δ_x is the Dirac measure in x and p, q, r are non negative numbers such that $p+q+r=1$. An item of size s will also be called an item s.

The set of the possible sizes is denoted by $\mathcal{T} = \{1,2,3\}$ and $\mathcal{T}^{(\mathbb{N})}$ is the set of finite vectors with coordinates in $\mathcal{T}$, if $x \in \mathcal{T}^{(\mathbb{N})}$, $\|x\|$ denotes the number of coordinates of x and $\emptyset$ is the empty vector.

The sojourn times of the items in the bin is an i.i.d. sequence with an exponential distribution with parameter 1.

An overview. In this paper we give a necessary and sufficient condition under which the size of the queue converges in distribution. If this condition has some interesting features (it is expressed with a quadratic functional of the input parameters), this is not the main point of the paper. The string valued Markov processes describing these models are complex to analyze. In this respect the paper presents an approach to deal with such processes. To keep the presentation simple we have chosen the simplest of these complicated models. A companion paper Robert[27] explores other aspects of these string processes.

To study the ergodicity properties of a finite dimensional Markov process, a standard approach is the following: The behavior of the process is analyzed when the initial state is such that a subset S of the coordinates is "large". When all the possible subsets S have been considered, the ergodicity condition generally follows easily.

In the case considered here, the processes can travel in infinitely many directions. To study the stability properties of these processes, one cannot recursively exhaust all the possibilities by inspection as it is the case in a finite dimensional setting. One of the conclusions of the paper is that it is better to consider the evolution of the *distribution* of the process at some random times rather than looking at the evolution of the *states* that the process visits at some random times as it is usually the case. This is related, in some sense, to the case of continuous state space Markov chains: the recurrence of the chain is defined not in term of the number of visits to some specified states, but by the fact that, at some random times, the Markov chain has a specified distribution. Notice that despite our framework is discrete (the state space is countable), these ideas are useful.

The framework of these Markov processes complicates technically the proofs of the results, even in some "simple" cases. See for example the section 3 where the ergodicity condition is quite intuitive, but its proof requires some discussion on the possible bifurcations of the system. This situation seems unavoidable, especially when the ergodicity condition is not natural at all (see Section 5).

The paper is organized as follows. We first prove that under some hypothesis, the Markov process describing the First Fit algorithm is ergodic if the "natural" condition is satisfied, i.e. if the load of the system is less than 1 (see Dantzer [9] for a discussion on this condition). In the other cases, the analysis is more intricate. The notion of smooth distribution on the state space is introduced. It is shown that at some random time the distribution of the process is smooth. The section 5 studies the fluid limits of the *distributions* of the process. The associated dynamical system turns out to be a product of random 2×2 matrices in $\mathbb{R}_+^2$; its stability properties are analyzed. All these results are then used to derive the ergodicity and transience conditions for the Markov processes.

Notations. An element X of the state space $\mathcal{S}$ of the Markov process describing the storage process can be written as $X = (B, L)$, where L and B are elements of $\mathcal{T}^{(\mathbb{N})}$, the set of finite vectors with coordinates in $\mathcal{T}$. The vector $B = (b_j; j = 1, \dots, \|B\|)$ describes the sizes of the items in the bin, since these items fit in the bin,

$$\sum_{j=1}^{\|B\|} b_j \leq C,$$

and the vector $L = (l_i; l = 1, \dots, \|L\|)$ represents the state of the queue. Since the First Fit algorithm scans the queue from the beginning in search of an item that may fit in the bin. Any item in the queue cannot fit in the bin, i.e. for any $i = 1, \dots, \|L\|$ the following inequality holds

$$l_i + \sum_{j=1}^{\|B\|} b_j > C.$$

Since $C = 4$, the possible values of B are the following

$$\emptyset,\ (1),\ (1,1),\ (1,1,1),\ (1,1,1,1),\ (1,1,2),\ (1,2),\ (1,3),\ (2),\ (2,2),\ (3).$$

Notice that the order of the components in B has no importance for the dynamic of the system, for this reason we shall consider B as a set. The order is important for the vector L since the First Fit discipline checks if the first coordinate l_1 fits in the bin, then the coordinates l_2, l_3, and so on. The vector L is a string of 1, 2, 3.

If $(X(t)) = ((B(t), L(t))$ is the state of the system at time t, $(X(t))$ is a Markov process with the following transitions :

(1) *Arrival.* At rate λ an item of size s arrives at the bin. If it does not fit in the bin, the element s is concatenated at the end of the vector $L(t)$.
(2) *Departure.* At rate 1, each item in the bin leaves the bin. In the case of a departure, the first element of the queue that fits, if any, is moved in the bin, and then the second, and so on.

It is not difficult to show that $(X(t))$ is an irreducible Markov process on $\mathcal{S}$. We shall say that the model is stable when $(X(t))$ is an ergodic Markov process on $\mathcal{S}$. In Dantzer *et al.* [9] it has been proved that the condition $\lambda\mathbb{E}(S_1) \leq C$ is necessary for the stability of the system, i.e. the Markov process $(X(t))$ is transient if $\lambda\mathbb{E}(S_1) > C$.

Definition 1. *The norm $\|X\|$ of the state $X = (B, L) \in \mathcal{S}$ is the sum of the $\|B\|$ and $\|L\|$. The* load $W(X(t))$ *of* $(X(t)) = (B(t), L(t)) = ((b_i(t)), (l_j(t)))$ *is defined as*

$$W(X(t)) = \sum_{i=1}^{\|B(t)\|} b_i\sigma_i^0(t) + \sum_{j=1}^{\|L(t)\|} l_j\sigma_j(t),$$

where, for $i \in \{1, \dots, \|B\|\}$ and $j \in \{1, \dots, \|L\|\}$, $\sigma_i^0(t)$ is the residual service time of the item $b_i(t)$ and $(\sigma_j(t))$ the service time of the item $l_j(t)$.

Notice that the load of the system increases at rate $\lambda\mathbb{E}(S_1)$ in average and decreases at most at rate 4.

Our results concerning ergodicity use the formalism of fluid limits. The next section recalls some of the results in this domain.

2. Fluid limits

In this section $(X(t))$ is an irreducible Markov process on some countable space $\mathcal{S}$ embedded in a normed space. We assume that the bounded subsets of $\mathcal{S}$ are finite. The rescaled process is defined by

$$X_x(t) = \frac{\|X(t\|x\|)\|}{\|x\|}, \tag{1}$$

since $X(0) = x$, $\|X_x\|(0) = 1$. The time variable and the space variable are scaled by a factor $\|x\|$. The following theorem is the combination of two results, one due to Filonov [15] and the other due to Rybko and Stolyar [28], it gives an ergodicity criterion.

Theorem 2. *If there exist an integrable stopping time U, constants K and $\varepsilon > 0$ such that*

$$\limsup_{\|x\|\to+\infty} \frac{\mathbb{E}_x\left(\|X(U)\|\right)}{\|x\|} \leq 1 - \varepsilon, \tag{2}$$

$$\limsup_{\|x\|\to+\infty} \frac{\mathbb{E}_x\left(U\right)}{\|x\|} \leq K, \tag{3}$$

the Markov process $(X(t))$ is ergodic.

If the variable $X(t)$ has a second moment for all $t \geq 0$, for a fixed $K \geq 0$ sufficiently large, the hitting time

$$H = \inf\{t \geq 0 / \|X(t)\| \leq K\},$$

has a second moment of order $\|x\|^2$*, i.e.*

$$\limsup_{\|x\|\to+\infty} \frac{\mathbb{E}_x\left(H^2\right)}{\|x\|^2} < +\infty. \tag{4}$$

The condition (2) requires that at some random time, $U/\|x\|$, the norm of the rescaled process $(X_x(t))$ is, in average, below its initial value. This suggests the analysis of the sequences of processes $(X_x(t))$, when $\|x\|$ tends to infinity. The limit of one of its converging subsequences is called a *fluid limit.* If one can prove that every fluid limit converges almost surely to 0 after some time T, then up to an integrability argument, Theorem 2 can be applied.

These scaling ideas are difficult to trace back. The origin of this criterion is the Lyapounov stability test of ordinary differential equations (see Hirsch and Smale [18] for the classical results). Has'minskii [17] seems to have been the first to use this test in a stochastic context, to prove the stability of stochastic differential equations.

The discovery of some unexpected phenomena for the stability of queueing systems, Bramson [2, 3], Dumas [10], Lu and Kumar [22], Kumar and Seidman [20] and Rybko and Stolyar [28] among others, gave an impulse to the studies in this domain recently. Chen and Mandelbaum [4] used fluid limits to study Jackson networks. Dai [7] set a framework to apply these methods to prove Harris ergodicity for some queueing networks. Concerning transience criteria, Dai [8], Meyn [25], Puhalskii and Rybko [26] obtained partial counterparts to the ergodicity results. In the context of diffusions, related ideas are used to prove the ergodicity of diffusions living in a domain with a boundary (see Dupuis and Williams [12] and the references therein).

The relations (2) and (3) imply that one "controls"the process in space and time when it starts very far away from some fixed state. In a finite dimensional context, one has consider the process when some of the coordinates of the initial state are large. In general, Theorem 2 can then be applied when all the possibilities for the large coordinates have been considered. Applying Theorem 2 for our process turns out to be more difficult since the process can go to infinity in infinitely many ways. This is not strictly true as we shall see. We prove that the process may diverge only along some "smooth patterns". We establish that, starting from any large arbitrary state, the process will eventually travel along some smooth states. (This notion of smooth state will be precized later).

3. When the natural condition is sufficient for ergodicity

We study a case where it is not necessary to know much about the structure of the L-component of the initial state. The following lemma gives an estimation of the wasted space when there are only two possible sizes: 1 and 2.

Lemma 3. *Under the conditions* $\lambda\mathbb{E}(S_1) < 4$*, if* $r = 0$ *(only items* 1 *and* 2 *arrive) and*

$$\tau = \inf\{t \geq 0/\|L(t)\| = 0\} \quad \textit{and} \quad D = \int_0^\tau 1_{\{b_1(t)+\cdots+b_{\|B(t)\|}<4\}}\, dt,$$

then there exist some constants K_1 *and* K_2 *such that*

$$\mathbb{E}_x(D) \leq K_1 \log(1 + \|x\|) + K_2.$$

for any $x = (l, b) \in \mathcal{S}$.

Proof. The variable D is the duration of time during which the bin is not full during a busy period. Notice first that there is no waste of space as long as there are items 1 in the L-component of $(X(t))$. We can therefore assume that l is a string of items 2. In this context the only possibility to waste space with a non empty queue is when the state $(B(t))$ of the bin is $(1,1,1)$ or $(2,1)$. We set $A_0 = l$ and $T_0 = 0$ and by induction we define

$$T_{n+1} = \inf(t > T_n / \|B(t-)\| = 4, \|B(t)\| < 4,\\ \text{and all the items 2 present at time } T_n \text{ are served at time } t)$$

and $A_n = \|L(T_n)\|$ for $n \geq 1$. (Notice that $L(T_n)$ is necessarily a (possibly empty) string of items 2). The sequence $(B(T_n), A_n)$ is clearly a Markov chain.

If b the initial state of the bin is $(1,1,1)$. As long as there is at least an item 1 in the queue, because of the First Fit discipline, the items 2 are ignored. Since $\lambda p \leq \lambda \mathbb{E}(S_1) < 4$, after an integrable amount of time not depending on $\|l\|$, at least two places will be vacant in the bin and consequently an item 2 will enter the bin. (In this situation the number of items 1 is the number of customers of an $M/M/4$ queue with parameter λp for the input rate and 1 for the service rate).

If $b = (2,1)$. We have two cases to discuss.

(1) $\lambda p < 2$. This condition clearly implies that, with probability 1, at some time there will no item 1 in the system and, consequently, a second item 2 will enter the bin. The expected value of this duration of time is easily seen to be bounded with respect to $\|l\|$. Starting from that time, only items 2 are served as long as the initial items 2 are present (since these items are located at the beginning of the queue, the First Fit algorithm selects them). When the initial items 2 have been served the queue is an i.i.d. string of items 1 and 2. With probability 1 at least two items 1 enter in the bin. Later, when the number of items 1 in the system is 1 the system will waste some space, this is precisely the definition of time T_1, A_1 is the number of items at that time.

(2) $\lambda p > 2$. This condition implies that, if the state of the bin does not change, the arriving items 1 will saturate two places in the bin. In this case, the number of items 1 is the number of customers of a transient $M/M/2$ queue starting with one customer (in the bin at time 0). A change in the state of the bin may occur only if this transient queue is empty.
   - a) The $M/M/2$ queue never reaches the empty state. After some small amount of time (i.e. its expected value is bounded with respect to $\|l\|$), the bin will be full with an item 2 and two items 1. The condition $\lambda \mathbb{E}(S_1) < 4$ implies that $\lambda q < 1$, therefore, with probability 1 after some period of time the system will not contain any item 2. At that moment the state of the bin will be $(1,1,1,1)$. (Recall that $\lambda p \leq \lambda \mathbb{E}(S_1) < 4$). It is easily seen that, with probability 1, the total number of items 1 will be less than 2. An item 2 will be in the bin at that time, this is the starting situation.
   - b) The queue reaches the empty state. Two items 2 occupy the bin. The initial items 2 are served. In this situation, T_1 is the next time there is some wasted space.

Notice that the case a) occurs only a geometrically distributed number of times. Hence, the duration of time between time 0 and T_1 when the bin is not full has a bounded expected value (with respect to $\|l\|$).

Using the proposition 16 of the appendix, it is easy to check that there exists some constant $c > 0$ such that the following convergence holds in L_1 and almost surely

$$\lim_{\|x\|\to+\infty} \frac{\mathbb{E}_x(T_1)}{\|x\|} = c.$$

(Its calculation is possible but not interesting for our purpose). If I is the duration of time between 0 and T_1 when the bin is not full, the expected value of the load at time T_1 satisfies the following inequality (all the service times are i.i.d. exponentially distributed random variables with parameter 1),

$$\mathbb{E}_x(W(X(T_1))) \leq \mathbb{E}_x(W(x)) + \mathbb{E}\left(\sum_{i=1}^{\mathcal{N}_\lambda([0,T_1])} S_i\right) - 4\mathbb{E}_x(T_1 - I).$$

using Wald's formula (T_1 is a stopping time), we get

$$\mathbb{E}_x(W(X(T_1))) \leq \mathbb{E}_x(W(x)) + (\lambda\mathbb{E}(S_1) - 4)\mathbb{E}_x(T_1) + 4\mathbb{E}_x(I),$$

since there are no items 1 in the queue at 0 and T_1, we have $\mathbb{E}_x(W(X(0))) = 2\|x\|$ and $\mathbb{E}_x(W(X(T_1))) = 3+2\mathbb{E}_x(A_1)$. The quantity $\mathbb{E}_x(I)$ being bounded with respect to $\|x\|$, it follows that

$$2\limsup_{\|x\|\to+\infty} \frac{\mathbb{E}_x(A_1)}{\|x\|} = \limsup_{\|x\|\to+\infty} \frac{\mathbb{E}_x(W(X(T_1)))}{\|x\|} \leq 2 + c(\lambda\mathbb{E}(S_1) - 4),$$

where $W(\cdot)$ is the load (see Definition 1 page 5). Consequently, there exist a_0 and $\alpha < 1$ such that for $\|x\| > a_0$,

$$\mathbb{E}_x(A_1) \leq \alpha\|x\|, \tag{5}$$

$$\gamma = -\log\left(\frac{1+\alpha\|x\|}{1+\|x\|}\right) > 0. \tag{6}$$

If we set

$$\nu = \inf\{n \geq 1/A_n \leq a_0\},$$

the sequence

$$(Z_n) = (\log(1 + A_{n\wedge\nu}) + \gamma(n\wedge\nu))$$

is a super-martingale. Indeed, if $(\mathcal{F}_n)$ is the natural filtration associated to the sequence (A_n), on the event $\{\nu > n\}$ the Markov property gives the equality

$$\begin{aligned}\mathbb{E}(Z_{n+1}/\mathcal{F}_n) - Z_n &= \mathbb{E}_{(B(T_n),A_n)}(\log(1 + A_1)) - \log(1 + A_n) + \gamma \\ &\leq \log(1 + \mathbb{E}(A_1/A_n)) - \log(1 + A_n) + \gamma \leq 0,\end{aligned}$$

by Jensen's inequality and the relations (5) and (6). Consequently $\mathbb{E}(Z_n) \leq Z_0$, hence $\gamma\mathbb{E}(n\wedge\nu) \leq \mathbb{E}(Z_n) \leq Z_0$, by letting n go to infinity, we get

$$\mathbb{E}_x(\nu) \leq \frac{\log(1 + \|x\|)}{\gamma}. \tag{7}$$

For $n \geq 1$, the bin is always full between T_n and T_{n+1}, except during some integrable period whose expected value is bounded with respect to the size of the initial state. By Wald's formula, the contribution of the ν cycles in the integral defining D is bounded by $K\mathbb{E}_x(\nu) \leq K\log(1+\|x\|)/\gamma$, for some constant K.

Since $\lambda\mathbb{E}(S_1) < 1$, the proposition 6 of Dantzer *et al.* [9] shows that the system is ergodic. Consequently, starting from the state $(B(T_\nu), A_\nu)$ $(\leq a_0))$, the hitting time of the empty state $\emptyset$ is integrable and with an expected value bounded with respect to $\|x\|$. Therefore, the expected value of the contribution of this period in the integral defining D is bounded with respect to $\|x\|$. The lemma is proved. □

Now we consider the general case with three sizes. The condition $\lambda\mathbb{E}(S_1) < 4$ turns out to be sufficient for ergodicity when $\lambda p > 1$.

Proposition 4. *If $\lambda\mathbb{E}(S_1) < 4$ and $\lambda p > 1$, then $(X(t))$ is an ergodic Markov process.*

Proof. Let $(x_n) = (b_n, l_n)$ a sequence of $\mathcal{S}$ whose norm converges to infinity. Since the number of configurations in the bin is finite, by taking subsequences we can suppose that the sequence of the initial states in the bin (b_n) is constant, hence $(x_n) = (b, l_n)$. Using Proposition 5 of Dantzer *et al.* [9], we can assume that for the states (x_n) the bin is not full. Consequently (l_n) does not contain any item 1, it is a sequence of strings of items 2 and 3.

We denote by τ the first time when the bin is not full after all the initial items 2 and 3 have left the system; τ is clearly a (possibly infinite) stopping time. If D is the duration of time between time 0 and τ during which the bin is not full, we claim that D is integrable and, moreover,

$$\lim_{n\to+\infty} \frac{\mathbb{E}_{x_n}(D)}{\|x_n\|} = 0.$$

If our assertion is true, between 0 and τ the load of the system is decreased at rate 4, except during some periods of total duration D, i.e. for $t \geq 0$ we have

$$\sum_{i=1}^{\|b\|} b_i\sigma_i^0 + \sum_{i=\|b\|+1}^{\|l_n\|+\|b\|} l_{n,i}\sigma_i^0 + \sum_{i=1}^{\mathcal{N}_\lambda(]0,t\wedge\tau])} S_i\sigma_i - 4(t\wedge\tau - D) \geq 0,$$

the sequences (σ_i) and (σ_i^0) are the respective service times of the arriving items and of the initial items. These variables are independent and exponentially distributed with parameter 1. Taking the expectation of the two members of this inequality, we get the relation

$$\mathbb{E}_{x_n}(t\wedge\tau)(4-\lambda\mathbb{E}(S_1)) \leq \|x_n\| - 4\mathbb{E}_{x_n}(D).$$

By letting t go to infinity, according to our assumption on $(\mathbb{E}_{x_n}(D))$ we obtain the inequality

$$\limsup_{n\to+\infty} \frac{\mathbb{E}_{x_n}(\tau)}{\|x_n\|} \leq \frac{1}{4-\lambda\mathbb{E}(S_1)}. \tag{8}$$

In the same manner, the following inequality holds,

$$\mathbb{E}_{x_n}(\|X(t\wedge\tau)\|) \leq \|x_n\| + (\lambda\mathbb{E}(S_1)-4)\mathbb{E}_{x_n}(\tau\wedge t) + 4\mathbb{E}_{x_n}(D);$$

Fatou's lemma and Lebesgue's Theorem give when t goes to infinity

$$\mathbb{E}_{x_n}(\|X(\tau)\|) \leq \|x_n\| + (\lambda\mathbb{E}(S_1)-4)\mathbb{E}_{x_n}(\tau) + 4\mathbb{E}_{x_n}(D),$$

since all the initial items 2 and 3 are served at time τ, at most two of the initial items can be served at the same time, hence,

$$\mathbb{E}_{x_n}(\tau) \geq \frac{\|x_n\|}{2},$$

consequently

$$\limsup_{n\to+\infty} \frac{\mathbb{E}_{x_n}(\|X(\tau)\|)}{\|x_n\|} \leq \frac{\lambda \mathbb{E}(S_1)}{4} < 1. \tag{9}$$

The relations (8) and (9) and Theorem 2 show that the Markov process $(X(t))$ is ergodic.

All we have to prove now is that $(\mathbb{E}_{x_n}(D))$ is negligible with respect to $\|x_n\|$ when n is large. There are several possibilities for b, the common content of the bin for the initial states (x_n). We discuss the different cases, throughout this discussion, we shall say that random variable H is a "bounded integrable variable" if the sequence $(\mathbb{E}_{x_n}(H))$ is bounded with respect to $\|x\|$.

(1) If b is $(1,1,1)$ or $(1,1,1,1)$. As long as there is at least an item 1 in the queue, all the other items are ignored. The condition $\lambda \mathbb{E}(S_1) < 4$ implies that $\lambda p < 4$. From the point of view of the items 1, the system is a stable $M/M/4$ queue. Hence the first time there will be at least two empty places is a bounded integrable variable. At that time, an item 2 will be inserted in the bin. Notice that for this period, the duration of time during which the bin is not full is a bounded integrable variable.

(2) If b has at least an item 2. The items 3 are not taken into account as long as the bin contains at least an item 2 or two items 1. In this case the items 3 are ignored. Consequently, a string of the initial items 3 builds up at the beginning of the queue. Since the condition $\lambda \mathbb{E}(S_1) < 4$ implies

$$\lambda p + 2q < 4,$$

the system with the items 1 and 2 is stable (see [9]). The lemma 3 shows that until an item 3 enters the bin the wasted space is negligible compared to the number of initial items 2.

(3) If b contains a 3. Clearly one can assume that l_n is a string of items 3. Otherwise, if at some moment 2 enters in the bin, the cases considered above show that the 1 and 2 will be cleared from the system until an item 3 is in the bin. The condition $\lambda p > 1$ implies that the residual space in the bin left by the items 3 is saturated by the items 1. Consequently the duration of time the bin is not full is a bounded integrable variable.

This discussion shows that the assertion is proved and consequently, the proposition. □

The result of the above proposition is fairly easy to understand: under the condition $\lambda p > 1$, basically there is no waste of space so that the natural condition $\lambda \mathbb{E}(S_1) < 4$ is sufficient for the ergodicity of $(X(t))$. Notice however that the proof of this intuitive result (Lemma 3 and Proposition 4) has required the detailed analysis of the possible evolutions starting from a given initial state. As we shall see, the situation is more delicate in the case $\lambda p < 1$.

4. Smoothing the initial state

This section studies the properties of the dynamic on the string structure of the queue (the L-component of the Markov process). From now on, we shall assume that $\lambda p < 1$ and that the initial states are strings of items 2 and 3 (see Proposition 5 of Dantzer *et al.* [9]). Even if the L-component of the initial states have only items 2 and 3, they are still too complicated to use fluid limits starting from these states. Loosely speaking the main idea is the following: after some time the initial disorder of the L-component is smoothed, i.e. it will have some regularity properties. If the initial state is a "smooth" state, a fluid analysis of the Markov process is then possible. This will be done in the next section. Compared to our previous analysis [9], this smoothing procedure is an important additional step. It is discussed in a more general context in Robert [27].

Definition 5. *For $X(0) = x \in \mathcal{S}$ and $t \geq 0$, if $X(t) = (B(t), L(t))$ and $L(t) = (l_i(t))$*

$$\begin{aligned}\nu_{x,1}(t) &= \inf\{k \geq 1/l_k(t) = 1\},\\ \nu_{x,2}(t) &= \inf\{1 \leq k < \nu_{x,1}(t)/l_k(t) = 2\},\\ \nu_{x,3}(t) &= \inf\{1 \leq k \leq \nu_{x,2}(t)/l_k(t) = 3\},\end{aligned}$$

with the convention $\inf\emptyset = +\infty$. If the initial state is without ambiguity, the subscript x is omitted; in the same way, the notation ν_a is used for $\nu_a(0)$.

The next definition formalizes the notion of "smooth" state, in fact the notion of a smooth distribution on $\mathcal{S}$.

Definition 6. *For integers l, m, n we define the distribution $R_{l,m,n}(dx)$ on $\mathcal{T}^{(\mathbb{N})}$ by*

$$R_{l,m,n}(dx) = \delta_3(du)^{(l)} \otimes F_{2,3}(du)^{(m)} \otimes F(du)^{(n)}, \tag{10}$$

where $G(dx)^{(n)}$ is the n-th power of the distribution $G(dx)$ and $F_{2,3}(dx)$ is the conditional distribution $F_{2,3}(dx) = (q\,\delta_2 + r\,\delta_3)/(q+r)$.

A distribution μ on $\mathcal{S}$ is smooth *if its L-component is in the convex hull of the $R_{l,m,n}$, l, m, $n \in \mathbb{N}$ i.e. if there exists a probability distribution (q_i) on $\mathbb{N}^3$ such that*

$$\mu(L \in dx) = \sum_{i\in\mathbb{N}^3} q_i R_i(dx).$$

The distribution $R_{l,m,n}$ is the distribution of the concatenation of several i.i.d. strings. The L-component of a distribution of type $R_{0,0,n}(dx)$ is just an i.i.d. string of length n with distribution F.

Proposition 7. *If $\lambda\mathbb{E}(S_1) < 4$, for any stopping time U greater than the first time when all the initial items have left the queue, the distribution of $X(U)$ is smooth.*

Proof. We denote by $M(t)$, the number of initial items in the queue at time t. A tag is inserted after the last initial item in the queue; $M(t)$ is in fact the position of the tag at time t, $(M(t))$ remains constant equal to 0 after it has reached 0. We first give a rough picture of the evolution of the queue. After time 0, the new items arrive behind the tag at rate λ. Recall that the queue of our initial state has no item 1. As long as some initial items 2 are in the queue, the First Fit algorithm picks (possibly) only items 1 after the tag. The departure of some of the items 1 builds a string of 2 and 3's after the tag. In the case where all the initial items 2 are processed and that some initial items 3 remain, the next items 2 are picked after the tag. In this case, a string of items 3's will build up behind the tag and before the string of 2's and 3's.

The notation $\tilde{\nu}_a$, $a \in \mathcal{T}$ is analogous to the definition 5 except that it concerns only the portion of the queue after the tag,

$$\begin{aligned}
\tilde{\nu}_1(t) &= \inf\{k > M(t)/l_k(t) = 1\},\\
\tilde{\nu}_2(t) &= \inf\{M(t) < k < \tilde{\nu}_{x,1}(t)/l_k(t) = 2\},\\
\tilde{\nu}_3(t) &= \inf\{M(t) < k \leq \tilde{\nu}_{x,2}(t)/l_k(t) = 3\}.
\end{aligned}$$

Notice that if $\tilde{\nu}_3(t)$ is finite, then necessarily $\tilde{\nu}_3(t) = M(t) + 1$. The variable $\tilde{L}(t)$ is the sub-string at the end of the queue consisting of the items located after the tag, $\tilde{L}(t)$ is the string $L(t)$ shifted $M(t)$ times. Consequently, if $U \leq t$ then $\tilde{L}(t) = L(t)$ and $\tilde{\nu}_a = \nu_a$ for $a \in \mathcal{T}$.

Assertion. *If τ is a stopping time, then conditionally on $\tilde{\nu}_a(\tau)$, $a \in \mathcal{T}$ and $\|L(\tau)\|$, the distribution of $\tilde{L}(\tau)$ is given by (10) for some convenient l, m and $n \in \mathbb{N}$.*

Since $L(U) = \tilde{L}(U)$, (the initial items are served at time U), the proposition will be then proved if the assertion is.

To show this claim, we shall assume that all the $\tilde{\nu}_a(\tau)$, $a = 1, 2, 3$ are finite. The analysis for the other cases is analogous. The string $\tilde{L}(\tau)$ is thus the concatenation of three strings $\tilde{L}(\tau) = (H_3, H_2, H_1)$, with

$$\begin{aligned}
H_3 &= (3, \ldots, 3), & \|H_3\| &= \tilde{\nu}_2(\tau) - 1;\\
H_2 &= (2, l_{\tilde{\nu}_2(\tau)+1}, \ldots, l_{\tilde{\nu}_1(\tau)-1}), & \|H_2\| &= \tilde{\nu}_1(\tau) - \tilde{\nu}_2(\tau);\\
H_1 &= (1, l_{\tilde{\nu}_1(\tau)+1}, \ldots, l_{|L(\tau)|}), & \|H_1\| &= \|L(\tau)\| - \tilde{\nu}_1(\tau).
\end{aligned}$$

For the rest of the proof, all the probabilistic statements are supposed to be conditioned by the values of the $\tilde{\nu}_a(\tau)$ and $|L(t)|$. Between time 0 and τ the First Fit algorithm never scanned the queue after the position $\tilde{\nu}_1(\tau)$, otherwise the item 1 located there would have been taken in the bin. The string H_1 is thus independent of H_3 and H_2. The first item 1 of H_1 is followed by the $(|L(\tau)| - \tilde{\nu}_1(\tau) - 1)^+$ items which arrived after that 1, hence it is an i.i.d. sequence with distribution $F(du)$.

In the same way for the string H_2, the First Fit algorithm never scanned the queue in search of a 2 after the position $\tilde{\nu}_1(\tau)$. Consequently the string H_2 consists of all the items arrived between the items located at the positions $\tilde{\nu}_2(\tau)$ and $\tilde{\nu}_1(\tau)$, with all the items 1 removed. the first item 2 in H_2 is followed by an i.i.d. string of length $(\tilde{\nu}_1(\tau) - \tilde{\nu}_2(\tau) - 1)^+$ and distribution $F(du|u \geq 2)$. The assertion is proved. □

Proposition 8. *If $\lambda\mathbb{E}(S_1) < 4$, $\lambda p < 1$ and U_0 is the first time t after all the initial items have left the queue that $B(t) = (2,1)$ then*

$$\sup_{x \in \mathcal{S}_1} \mathbb{E}_x\left(\left(\frac{U_0}{\|x\|}\right)^2\right) < +\infty \text{ and } \sup_{x \in \mathcal{S}_1} \mathbb{E}_x\left(\left(\frac{\|X(U_0)\|}{\|x\|}\right)^2\right) < +\infty, \tag{11}$$

where $\mathcal{S}_1$ is the subset of the states of $\mathcal{S}$ for which the bin is not full,

$$\mathcal{S}_1 = \{x = (b, l) \in \mathcal{S}/|b| < 4\}.$$

Proof. The initial state $X(0)$ is given by $x = (B, L)$ with $L = (l_1, \dots, l_p)$ for some $p \geq 1$. We denote by T_2 [resp. T_3] is the time when all the initial items 2 [resp. 3] have left the queue. The variable T is the first time when all the initial items have left the queue, T is clearly stopping time bounded by $T_2 + T_3$.

For a fixed $k \in \{1, \dots, p-1\}$, we define $\check{x} = (B, \check{L})$ where $\check{L}$ is the same string as L except the components k and $k+1$ are permuted. For $1 \leq i < p$, the quantities τ_i, $\check{\tau}_i$ denote respectively the waiting time necessary for the i-th item l_i to enter the bin when the initial state is respectively x, $\check{x}$. We assume that for these two initial states, the arrival stream and the services associated with the items are the same. there are two cases:

- If $\tau_{k+1} < \tau_k$, in both systems the item l_k will enter the bin at time τ_k, thus $\check{\tau}_k = \tau_k$.
- Otherwise, when $\tau_{k+1} \geq \tau_k$ and the initial state is $\check{x}$, at time τ_k the First Fit algorithm checks if the item l_{k+1} fits in the bin and after the item l_k is checked.

Hence, in any case $\tau_k \leq \check{\tau}_k$. By induction, the quantity $\mathbb{E}_x(T_2)$ is thus bounded by $\mathbb{E}_{x'}(T_2)$ where $x' = (B, L')$ is the initial state given by $L' = (3, \dots, 3, 2, \dots, 2)$, L' is a permutation of L with all the items 3 at the beginning. Similarly, $\mathbb{E}_x(T_2) \leq \mathbb{E}_{x''}(T_2)$, T_2 is the time to get rid of the initial 2's and $x'' = (B, L'')$, where L'' is a permutation of the L-component of x when all the items 3 are at the head of the queue. To bound $\mathbb{E}_x(T^2)$ it is sufficient to give an upper bound for $\mathbb{E}_{x''}(T_2^2)$ and $\mathbb{E}_{x'}(T_3^2)$.

- The items 2 are at the beginning. We can assume that the bin does not contain a 3 at time 0 (otherwise, as soon as it leaves it is replaced by an item 2). As long as an item 2 is at the head of the queue, the system works only with items of size 1 and 2. When the system without items 3 has at most one item 1 in the bin, an item 3 enters in the bin, then all the initial items are served consecutively. The estimation of T_3 is thus reduced to the estimation of the time to empty the system without the items 3. Since the condition $\lambda\mathbb{E}(S_1) < 4$ implies that $\lambda(p + 2q)/(p + q) < 4$, the system without the items 3 is ergodic. Using the ergodicity result of [9] and the inequality (4) of Theorem 2, we get that $\mathbb{E}_{x'}(T_3^2) \leq A_1\|x\|^2$ (notice that $\|x'\| = \|x\|$), for some constant A_1.
- The items 3 are at the beginning.
  - If the initial state of the bin has an item 3, all the initial 3's are served consecutively and then the initial 2's are served. For a convenient constant A_2, one easily gets that $\mathbb{E}_{x''}(T_2^2) \leq A_2\|x\|^2$.
  - If there is a 2 in the bin and at least a 1, the situation is more interesting. At the difference of the previous case, an item 3 can enter the bin before some of the initial items of size 2. if at some moment the state of the bin is $(2, 1)$ (there an empty place of size 1) and that no new item 1 arrives before a departure form the bin. If the item 2 leaves before the item 1 then the item 3 at the head of the queue enters the bin, and then all the other initial items 3; otherwise if the item 1 leaves first, an additional 2 enters the bin, then all the initial items 2 are processed. Since $\lambda p < 1$, if there are sufficiently many 2's in the queue, one of this two cases will occur with probability 1 (if it

is not the case, the 3's occupy the bin and it is finished). We thus get a constant A_3 such that $\mathbb{E}_{x''}(T_2) \leq A_3\|x\|$.

At time T all the initial items have left the queue. Since

$$X(T) \leq \|x\| + \mathcal{N}_\lambda(]0,T]),$$

Wald's formula and the above estimation show that $\|X(T)\|$ is bounded by a constant times $\|x\|$.

Now we have to estimate $\bar{T}$ the first time when the state of the bin is $(2,1)$. It is sufficient to prove that, if the initial state is x, $\bar{T}$ has a second moment of the order $\|x\|^2$. The first step is to get rid of the items 3. If there is one in the bin and if some of them are located at the head of the queue and one has to process these ones until an additional item 1 or 2 enters the bin. There are two possibilities:

- The bin has at least 1 item 1. Since $\lambda p < 1$, after some time the queue will not have any item 1 and the bin will have two items 1.
  * If, at that time, there are sufficiently many items 2 in the queue the state of the bin will reach the state $(2,1)$ with probability 1.
  * If not. All the items 3 in the queue at that time are served. When this is finished, the condition $\lambda p < 1$ implies that the number of items 1 is tight (as a family of random variables indexed by x, the initial state). The items 2 accumulated during that time are served, consequently with probability 1, the state of the bin will be $(2,1)$.
- The state of the bin is $(2,2)$ and the initial items 2 are served at rate 2. When this is finished, with probability 1, two items 1 enters in the bin. This is the situation of the previous case.

It is easily seen that each of the steps we have described has a duration with a second moment of the order $\|x\|^2$. The proposition is proved. The last inequality is a consequence of Wald's formula applied to the stopping time U_0.

□

5. A random dynamical system in $\mathbb{R}_+^2$

In this section we assume that (μ_n) is a sequence of smooth distributions on $\mathcal{S}$ (see Definition 6) such that

$$\mu_n(B=(2,1), L \in dx) = \mathbb{E}\left(R_{a_n,b_n,0}(dx)\right), \tag{12}$$

i.e. if f is a non negative measurable function on $\mathcal{T}^{(\mathbb{N})}$

$$\mathbb{E}_{\mu_n}\left(f(L(0))1_{\{B(0)=(2,1)\}}\right) = \mathbb{E}\left(\int_{\mathcal{T}^{(\mathbb{N})}} f(x) R_{a_n,b_n,0}(dx)\right), \tag{13}$$

where a_n and b_n are random variables such that the convergence

$$\lim_{n\to+\infty} \frac{a_n}{n} = a \quad \text{and} \quad \lim_{n\to+\infty} \frac{b_n}{n} = b,$$

holds in L_1. We assume that a and b are non negative integrable random variables and $\mathbb{P}(a+b>0)=1$. The B-component of μ_n is $(2,1)$ and the L-component of the distribution μ_n does not have an item 1 in the queue, it is the concatenation of a_n items 3 followed by an i.i.d. string of length b_n of 2's and 3's with respective probability $q/(q+r)$ and $r/(q+r)$.

Definition 9. *A sequence (X_n) of random variables is equivalent to (α_n) if the sequence (X_n/α_n) converges to 1 in $L_1(\mathbb{P})$.*

A random transition of the fluid model

If the initial distribution is given by μ_n, the initial state of the bin is $(2,1)$. If there is a departure before a new arrival,

1) with probability $1/2$ this is the item 2 and then an item 3 enters the bin and then all the other $a_n - 1$ items 3.
2) with probability $1/2$ this is the item 1, an other item 2 enters the bin, and then all the other $b_n - 1$ items 2.

We remark that the dynamic of the system is influenced by the fact that either the 2 leaves first or not, this is also true at the fluid level as we shall see. A similar phenomenon has been already encountered in the model analyzed in Dantzer [9]. Here the randomness remains because of this 1/2-1/2 transition and not because there are many possibilities for the content of the bin as it is the case in [9].

If the distribution of $X(0)$ is given by μ_n, then $\|X(0)\|$ is equivalent to $((a+b)n)$. The next proposition shows that, up to a linear transformation, the distribution of X at a stopping time has a property similar to the identity (12).

Proposition 10. *If U_1 is the first time when all the initial items 2 have left the queue, the initial items 2 in the bin have been served and the state of the bin is $(2,1)$, there exist $\mathcal{F}_{U_1}$-measurable random variables A_n, B_n such that the following relation holds*

$$\mathbb{P}_{\mu_n}(B(U_1) = (2,1), L(U_1) \in dx) = \mathbb{E}_{\mu_n}\left(R_{A_n,B_n,0}(dx)\right),$$

and a random matrix M such that the convergence

$$\lim_{n\to+\infty} \frac{1}{n}(A_n, B_n) = M \cdot (a,b) \tag{14}$$

holds almost surely and in L_1. The random matrix M has two possible values with equal probability

$$m_1 = \begin{pmatrix} 1 & \dfrac{1-p-q}{1-p} \\ 0 & \dfrac{2\lambda q}{4-\lambda p} \end{pmatrix} \quad \textit{and} \quad m_2 = \begin{pmatrix} \lambda(1-p-q) & \dfrac{1-p-q}{1-p} \\ \dfrac{2\lambda^2(1-p)q}{4-\lambda p} & \dfrac{2\lambda q}{4-\lambda p} \end{pmatrix}; \tag{15}$$

M is independent of (a,b) if $\mathbb{P}(a>0, b>0) = 1$.

Proof. Using the Skorohod's representation Theorem (See Ethier and Kurtz [13]), with a change of the probability space we can assume that the sequences (a_n/n) and (b_n/n) converge almost surely (since they converge in L_1, they converge in distribution).

If $\mathbb{P}(a>0, b>0) = 1$. The state of the bin is $(2,1)$ at time 0. If a new item 1 arrives, the bin is full and, during that time, the 2 in the bin is replaced by the initial items 2. So, the state of the bin will come back to the state $(2,1)$ after an integrable duration of time (the time for an $M/M/2$ queue starting with two customers to have only one customer). In this manner, a finite number of initial items 2 in the queue are served in the bin before a significant change occurs. If there is a departure when the state of the bin is $(2,1)$, another item 2 may enter if the item 1 leaves, otherwise an item 3 enters. Hence, after this event the state of the bin will be $(2,2)$ or $(3,1)$ with probability $1/2$. Since (b_n) converges almost surely to infinity, there will be an item 2 in the queue with probability 1 at the occasion of such a departure. We conclude that the fact that an item 3 or a second item 2 enters the bin is independent of the the limit of $(a_n, b_n)/n$, as long as a and b are positive with probability 1.

Throughout this discussion, we shall ignore small strings in our statements, i.e. strings with an integrable length independent of the initial state. At the fluid level, most of them do not play a role (but not all of them!). As we already noticed:

1) With probability $1/2$ the item 2 leaves first. In this case the first item 3 enters the bin and then all the other $a_n - 1$ items 3 will follow it in the bin. During that time, since $\lambda p < 1$, the items 1 are processed by the empty space in the bin. It is easily checked that the time τ_1 to get rid of the initial items 3 is equivalent to $a_n \sim an$.

   At time τ_1 the head of the queue is the original string of items 2 and 3 followed by another string of 2 and 3 built up during the service of the items 3. Consequently, using again the law of large numbers, the length of the queue is thus equivalent to $(b + \lambda(q+r)a)n$ (Lemma 16 of the appendix). Very quickly an item of size 2 is in the bin, it is easy to check after an integrable amount of time the state of the bin will be $(2,2)$. Starting from that time, all the initial items 2 are served consecutively: a string of items 3 builds up at the head of the queue followed by a shrinking string of 2's and $3'$. At the end of the queue the new items arrived during that time form a string (since the bin is full the items 1 are not served during this phase). The time τ_2 to serve all the items 2 arrived before the state of the bin reaches $(2,2)$ is equivalent to the quantity
$$(b + \lambda(q+r)a)\frac{q}{2(q+r)}n.$$
   At time $\tau_1 + \tau_2$ there is a string of 3's at the head of the queue of length equivalent to
$$(b + \lambda(q+r)a)\frac{r}{q+r}n, \tag{16}$$
   followed by an i.i.d. string with distribution $F(du)$ whose length is equivalent to the quantity
$$\lambda(b + \lambda(q+r)a))\frac{q}{2(q+r)}n.$$
   If there is a departure of an item 2 it is immediately replaced by an item 2 or 2 items 1; the items 3 cannot be served at that moment. Due to the i.i.d. structure of the queue at that time, it is then easily seen that after an integrable amount of time, the bin will be in the state $(1,1,1,1)$. From that time all the 1 will be served at rate 4. The time τ_3 it takes to empty the queue of the items 1 and to have exactly a 2 and a 1 in the bin is equivalent to
$$\frac{\lambda p(b + \lambda(q+r)a))q}{2(q+r)(4-\lambda p)}n,$$

(take $\xi = \lambda p$ and $\mu = 4$ in Lemma 16 of the appendix). At time $\tau_1+\tau_2+\tau_3$ there is a string of 3's of length whose length is equivalent to (16), followed by a string of 2's and 3's of length equivalent to

$$
\begin{aligned}
(17)\quad \lambda(b+\lambda(q+r)a))&\frac{q}{2(q+r)}(q+r)\,n+\lambda(q+r)\frac{\lambda p(b+\lambda(q+r)a))q}{2(q+r)(4-\lambda p)}\,n\\
&=\frac{\lambda q}{2}(b+\lambda(q+r)a))\left(1+\frac{\lambda p}{4-\lambda p}\right)=\frac{2\lambda q}{(4-\lambda p)}(b+\lambda(q+r)a)).
\end{aligned}
$$

For this case the distribution of $L(U)$ is given by $\mathbb{E}_{\mu_n}\left(R_{A_n,B_n,0}(dx)\right)$ and (A_n,B_n) satisfies the relation(14) with the matrix $M=m_2$.

The uniform integrability of the sequences (Z_n) and (n/Z_n) can be proved following the same discussion.

2) With probability $1/2$ this is the item 1, an other item 2 is in the bin, then all the other b_n-1 items 2 will be served. The method is the same as in the previous case. It is slightly simpler since the initial items 3 are not served at time U_1.

Finally, the discussion is similar on the set $\{a=0, b\neq 0\}\cup\{a\neq 0, b=0\}$, the difference is that the duration of some transitions described above are negligible in this case. ☐

The next proposition gathers some facts and estimations which will be used in the sequel. Its proof, which is not difficult, follows the discussion of the above proof. It is skipped.

Proposition 11. *With the same notations as in Proposition 10, there exists a constant K_0 such that*

$$
\limsup_{n\to+\infty}\mathbb{E}\left(\frac{U_1}{n}\right)<K_0\mathbb{E}(a+b). \tag{18}
$$

If a and b are deterministic, positive and $Z_n=(A_n+B_n)/(a_n+b_n)$, the sequences (Z_n) and $(1/Z_n)$ are uniformly integrable.

The main result on the ultimate behavior of the fluid limits is contained in the following proposition.

Proposition 12. *If (M_n) is an i.i.d. sequence of random matrices with the same distribution as M in Proposition 10 and $P_n=M_n\cdot M_{n-1}\cdots M_1$, there exist α, $\beta>0$ and a function η on $\mathbb{R}_+$ such that for any $n\in\mathbb{N}$ and $x\in\mathbb{R}^2_+$,*

$$
\mathbb{E}(\langle(\alpha,\beta),P_n\cdot x\rangle)=\eta(\lambda)^n\,\langle(\alpha,\beta),x\rangle\,, \tag{19}
$$

where $<\cdot,\cdot>$ is the usual scalar product in $\mathbb{R}^2$. If

$$
\lambda^*=\frac{4-3p-2q-\sqrt{(4-3p-2q)^2-16p(1-p-q)}}{2p\,(1-p-q)}, \tag{20}
$$

then $\eta(\lambda)<1$ if $\lambda<\lambda^*$ and $\eta(\lambda)>1$ if $\lambda^*<\lambda<4/p$.

Proof. We denote by $\mathbb{E}(P_n)$ the expected value of the matrix P_n, i.e. the matrix of the expected values of the coefficients of P. The i.i.d. property of the M_n's gives the relation $\mathbb{E}(P_n)=\mathbb{E}(M_1)^n$. The positive matrix $\mathbb{E}(M_1)$ has two positive eigenvalues, $\eta(\lambda)$ denotes the largest of them and (α,β) is the corresponding right eigenvector; α and β can be chosen strictly positive (see Seneta [29] for example). Consequently, we get

$$\begin{aligned}\mathbb{E}\left(\langle(\alpha,\beta),P_n\cdot x\rangle\right) &= \langle(\alpha,\beta),\mathbb{E}(P_n)\cdot x\rangle = \langle(\alpha,\beta),\mathbb{E}(M_1)^n\cdot x\rangle\\ &= \eta(\lambda)^n\,\langle(\alpha,\beta),x\rangle\,.\end{aligned}$$

It is easily seen that $\eta(\lambda)$ can be expressed as

$$\eta(\lambda)=\max\left\{\frac{\langle\mathbb{E}(M_1),x\rangle}{\langle x,1\rangle}\Big/ x\in\mathbb{R}_+^2\right\},$$

since the components of $\mathbb{E}(M_1)$ are increasing with respect to λ if $\lambda p<4$, the same property is true for the largest eigenvalue $\eta(\lambda)$. The smallest root of the equation $\eta(\lambda)=1$ is given by $\lambda=\lambda^*$. (Routine calculations show that the term under the square root in (20) is non negative if $p+q\leq 1$ and that $\lambda^* p<4$). The proposition is proved. $\square$

Corollary 13. *With the notations of Proposition 20, if $\lambda<\lambda^*$ for any $\gamma<\eta(\lambda)$ and $x\in\mathbb{R}_+^2$ the sequence $(\gamma^{-n}P_n\cdot x)$ converges almost surely and in $L_1(\mathbb{P})$ to $(0,0)$.*

Proof. Using the identity (19) for $n=1$, it is easily seen that

$$(Z_n)=\left(\langle(\alpha,\beta),\eta(\lambda)^{-n}\,P_n\cdot x\rangle\right)$$

is a martingale. The sequence (Z_n) being non negative, it converges almost surely to some finite limit Z_∞. Since α and β are positive and all the coefficients of P_n are non negative, we deduce that the sequence $(\gamma^{-n}P_n\cdot x)$ converges almost surely to 0 for any $\gamma<\eta(\lambda)$. The L_1-convergence follows from the fact that $\mathbb{E}(P_n\cdot x)=\mathbb{E}(M_1)^n\cdot x$ and the fact that the eigenvalues of $\mathbb{E}(M_1)$ are in the interval $[0,1[$. $\square$

Our study here is simplified because all our matrices are non negative. For general results on the product of arbitrary random matrices, see Bougerol and Lacroix [1] for example.

6. Ergodicity

Now we have all the necessary ingredients to get the ergodicity result. Starting from some arbitrary large initial state, the process "hits" some smooth state whose size has the same order of magnitude of the initial state. From that time, under the appropriate condition, the Markov process shrinks with a factor γ at each cycle described in the previous section. After some fixed number of such cycles, the size of the process will be a fraction of the original size. The ergodicity will be proved then.

Theorem 14. *If the arrival rate of the items is λ and the distribution of their sizes is given by*

$$F(dx)=p\delta_1+q\delta_2+(1-p-q)\delta_3,$$

when the size of the bin is 4, if

$$\lambda<\lambda_{FF}\stackrel{def}{=}\frac{4-3p-2q-\sqrt{(4-3p-2q)^2-16p(1-p-q)}}{2p\,(1-p-q)}\wedge\frac{4}{3-2p-q}, \tag{21}$$

the Markov process $(X(t))$ describing the First Fit algorithm is ergodic.

Proof. If $\lambda p > 1$, the proposition 4 shows that the condition $\lambda \mathbb{E}(S_1) < 4$, i.e. $\lambda(3-2p-q) < 4$ is sufficient for the ergodicity of $(X(t))$. One can check that in that case

$$\frac{4}{3-2p-q} < \frac{4-3p-2q-\sqrt{(4-3p-2q)^2-16p(1-p-q)}}{2p\,(1-p-q)}.$$

We assume that the conditions (21) and $\lambda p < 1$ are satisfied. According to Theorem 2, to prove the ergodicity it is sufficient to prove that there exists a stopping time V such that for any sequence $(x_n)=(b_n,l_n)$ of $\mathcal{S}_1$ with $\|x_n\|=n$, the following inequalities hold

$$\limsup_{n\to+\infty}\frac{\mathbb{E}_{x_n}(\|X(V)\|)}{n} \leq 1-\varepsilon, \qquad \limsup_{n\to+\infty}\frac{\mathbb{E}_{x_n}(V)}{n} \leq K, \tag{22}$$

where $K>1$ and $\varepsilon>0$ are constants independent of the sequence (x_n). The symbol K for the constant is used throughout this proof, to avoid subscripts we keep the same letter.

According to the propositions 7 and 8 if $X(0)=x_n$ there exists a stopping time U_0 such that

- the distribution of $L(U_0)$ is given by $\mathbb{E}(R_{a_n,b_n,0}(dx))$, where a_n and b_n are some random variables and $B(U_0)=(2,1)$;
- the following relations hold

$$\limsup_{n\to+\infty}\frac{\mathbb{E}_{x_n}U_0}{n} \leq K, \tag{23}$$

$$\limsup_{n\to+\infty}\mathbb{E}_{x_n}\left(\frac{a_n+b_n}{n}\right)^2 \leq \mathbb{E}_{x_n}\left(\frac{\|X(U_0)\|}{n}\right)^2 \leq K. \tag{24}$$

According to the inequality (24), the sequence of random variables $(a_n/n, b_n/n)$ is tight for the convergence in distribution. By taking a subsequence, we can suppose that they jointly converge in distribution to some random variable (a,b). The relation (24) shows that the sequence $(a_n/n,b_n/n)$ is uniformly integrable; consequently, it converges in L_1. In particular the following convergence holds

$$\lim_{n\to+\infty}\mathbb{E}_{x_n}\left(\frac{a_n+b_n}{n}\right) = \mathbb{E}(a+b) \leq K, \tag{25}$$

the last inequality is a consequence of the relation (24). Using again the Skorohod's representation Theorem (See Ethier and Kurtz [13]), with a change of the probability space we can assume that the sequences (a_n/n) and (b_n/n) converge almost surely to a and b respectively.

On the event $\{\|X(U_0)\|\leq\|X(0)\|/4\}$ one sets $V=U_0$, so that

$$\mathbb{E}_{X(0)}\left(\frac{\|X(V)\|}{\|X(0)\|}1_{\{\|X(U_0)\|\leq\|X(0)\|/4\}}\right) \leq \frac{1}{4} \tag{26}$$

We have to determine V on the event $\{\|X(U_0)\|>\|X(0)\|/4\}$. The proposition 10 shows that, on the on the event $\{a+b>0\}$ there exist a stopping time U_1, random variables $(A_{n,1})$, $(B_{n,1})$ and a matrix M_1 independent of (a,b) such that

$$\mathbb{P}_{\mu_n}(B(U_1)=(2,1), L(U_1)\in dx) = \mathbb{E}_{\mu_n}\left(R_{A_{n,1},B_{n,1},0}(dx)\right),$$

where μ_n is the distribution of $X(U_0)$ when $X(0)=x_n$, and the relation

$$\lim_{n\to+\infty}\frac{1}{n}(A_{n,1},B_{n,1})=M_1\cdot(a,b)$$

holds almost surely and in L_1.

From now on, until further notice, we work on the set $\{a+b>0\}$. We denote by $(\theta_t\ ; t\geq 0$ the time-shift for the Markov process, if we iterate, we get the existence of a random variables $(A_{n,2})$, $(B_{n,2})$ and a matrix M_2 such that

$$\mathbb{P}_{X(U_1)}\left(B(U_1\circ\theta_{U_1})=(2,1),L(U_1\circ\theta_{U_1})\in dx\right) \\ =\mathbb{E}_{X(U_1)}\left(R_{A_{n,2},B_{n,2},0}(dx)\right), \tag{27}$$

and

$$\lim_{n\to+\infty}\frac{1}{n}(A_{n,2},B_{n,2})=M_2M_1(a,b),$$

almost surely and in L_1.

For $p\in\mathbb{N}$, we define the variable $U_{p+1}=U_p+U_1\circ\theta_{U_p}$, U_p is clearly a stopping time. The relation (27) gives the following equality

$$\mathbb{P}_{\mu_n}(B(U_2)=(2,1),L(U_2)\in dx)=\mathbb{E}_{\mu_n}\left(R_{A_{n,2},B_{n,2},0}(dx)\right),$$

By induction, it is easily seen that there exist random variables $(A_{n,p})$ and $(B_{n,p})$, independent matrices M_p, $p\geq 2$ such that

$$\mathbb{P}_{\mu_n}(B(U_p)=(2,1),L(U_p)\in dx)=\mathbb{E}_{\mu_n}\left(R_{A_{n,p},B_{n,p},0}(dx)\right),$$

and

$$\lim_{n\to+\infty}\frac{1}{n}(A_{n,p},B_{n,p})=M_pM_{p-1}\cdots M_2M_1(a,b),$$

almost surely and in L_1. According to Proposition 12,

$$\lim_{n\to+\infty}\frac{1}{n}\left\langle(\alpha,\beta),(\mathbb{E}_{\mu_n}(A_{n,p}),\mathbb{E}_{\mu_n}B_{n,p})\right\rangle=\gamma^p\,(\alpha a+\beta b),$$

Since $\|X(U_p)\|=2+A_{n,p}+B_{n,p}$,

$$\limsup_{n\to+\infty}\frac{1}{n}\mathbb{E}_{\mu_n}(\|X(U_p)\|)\leq\frac{1}{\alpha\wedge\beta}\lim_{n\to+\infty}\frac{1}{n}\left\langle(\alpha,\beta),(\mathbb{E}_{\mu_n}(A_{n,p}),\mathbb{E}_{\mu_n}B_{n,p})\right\rangle \\ =\gamma^p\,\frac{\alpha\mathbb{E}(a)+\beta\mathbb{E}(b)}{\alpha\wedge\beta}\leq\gamma^p\,\frac{\alpha\vee\beta}{\alpha\wedge\beta}\mathbb{E}(a+b)\leq\gamma^p\,\frac{\alpha\vee\beta}{\alpha\wedge\beta}K, \tag{28}$$

according to the inequality (25). Since the condition $\lambda<\lambda_{FF}$ implies that $\gamma<1$, we choose $p\in\mathbb{N}$ such that

$$\gamma^p<\frac{1}{4}\frac{\alpha\wedge\beta}{(\alpha\vee\beta)K}.$$

On the event $\{\|X(U_0)\|>\|X(0)\|/4\}$, the variable V is defined as $U_0+U_p\circ\theta_{U_0}$. For $n\in\mathbb{N}$, by the strong Markov property,

$$\mathbb{E}_{x_n}\left(\|X(V)\|1_{\{\|X(U_0)\|>\|x_n\|/4\}}\right)=\mathbb{E}_{x_n}\left(\mathbb{E}_{\mu_n}(\|X(U_p)\|)1_{\{\|X(U_0)\|>\|x_n\|/4\}}\right) \\ =\mathbb{E}_{x_n}\left(\mathbb{E}_{\mu_n}((A_{n,p}+B_{n,p}+2)1_{\{\|X(U_0)\|>\|x_n\|/4\}}\right).$$

We can assume that $\mathbb{P}(a+b=1/4)=0$, otherwise we replace the constant $1/4$ by some real r less than $1/4$ such that $\mathbb{P}(a+b=r)=0$.

$$\left|\mathbb{E}_{x_n}\left(\mathbb{E}_{\mu_n}\left(\frac{\|X(U_p)\|}{n}\right)\left(1_{\{\|X(U_0)\|>\|x_n\|/4\}}-1_{\{a+b>1/4\}}\right)\right)\right|$$
$$\leq C_0\mathbb{E}_{x_n}\left|1_{\{\|X(U_0)\|>\|x_n\|/4\}}-1_{\{a+b>1/4\}}\right|+2\mathbb{E}_{x_n}\left(\frac{\|X(U_p)\|}{n}1_{\{\|X(U_p)\|/n\geq C_0\}}\right).$$

Due to the L_1-convergence of $\|X(U_p)\|/n=(A_{n,p}+B_{n,p})/n$, the second term of the right hand side is arbitrarily small uniformly on n for some $C_0>0$. The first term converges to 0 since $\|X(U_p)\|/n$ converges almost surely to $a+b$ and $\mathbb{P}(a+b=1/4)=0$. Hence it is enough to consider the quantity

$$\mathbb{E}_{x_n}\left(\mathbb{E}_{\mu_n}\left(\frac{\|X(U_p)\|}{n}\right)1_{\{a+b>1/4\}}\right).$$

The relation (28) implies that

$$\limsup_{n\to+\infty}\mathbb{E}_{x_n}\left(\frac{\|X(V)\|}{n}1_{\{\|X(U_0)\|>\|x_n\|/4\}}\right)\leq\gamma^p\frac{\alpha\vee\beta}{\alpha\wedge\beta}K\leq\frac{1}{4} \tag{29}$$

The inequalities (23) and (18) show that there exists some constant K such that

$$\limsup_{n\to+\infty}\frac{\mathbb{E}_{x_n}(V)}{\|X(0)\|}\leq K,$$

and the relations (26) and (29) give

$$\limsup_{n\to+\infty}\mathbb{E}_{x_n}\left(\frac{\|X(V)\|}{\|X(0)\|}\right)\leq\frac{1}{2}.$$

The proof of the proposition is completed. □

Remark. We used the independence of the matrices M_2, M_3, ... in our proof. Since one may have $a=0$ or $b=0$, M_1 is not necessarily independent of (a,b). It is nevertheless true after step 2 for M_p, for $p\geq 2$ (see the definition of m_1 and m_2 in Proposition 10.

7. Transience

Theorem 15. *If the arrival rate of the items is λ and the distribution of their sizes is given by*

$$F(dx)=p\delta_1+q\delta_2+(1-p-q)\delta_3,$$

when the size of the bin is 4 if

$$\lambda>\lambda_{FF}=\frac{4-3p-2q-\sqrt{(4-3p-2q)^2-16p(1-p-q)}}{2p\,(1-p-q)}\wedge\frac{4}{3-2p-q}, \tag{30}$$

the Markov process $(X(t))$ describing the First Fit algorithm is transient,

Proof. We assume that the initial distribution of $(L(t))$ is given by $R_{a,b,0}$ and the initial state of the bin is $(2,1)$.

With the notation of Proposition 10, U_1 is the first time when all the initial items 2 have left the queue, the items 2 in the bin have been served and the state of the bin is $(2,1)$. As in the proof of Proposition 14, we define the sequence of stopping times (U_p) by

$$U_{p+1}=U_p+U_1\circ\theta_{U_p}.$$

The variable U_{p+1} is the first moment when all the items 2 present at time U_p have left the queue and the state of the bin is $(2,1)$. Clearly enough, the sequence $(L(U_p))$ is an homogeneous irreducible Markov chain on $\mathcal{T}^{(\mathbb{N})}$.

The distribution of $X(U_1)$ is represented by

$$\mathbb{P}_{R_{a,b,0}}(X(U_1)\in dx)=\mathbb{E}(R_{A_{a,b},B_{a,b},0}(dx)),$$

almost surely U_1 is a finite stopping time. The propositions 10 and 12 show that there exist constants α, β such that

$$\lim_{a+b\to+\infty}\frac{\alpha A_1+\beta B_1}{\alpha a+\beta b}=\gamma>1, \tag{31}$$

almost surely.

We assume that the Markov process $(X(t))$ is recurrent. In particular it visits the state $y_0=(\emptyset,(2,1))$ infinitely often, i.e. with probability 1 the queue will be empty and the state of the bin will be $(2,1)$. The first time the process $(X(t))$ visits the state y_0, is necessarily at one of the moments U_p, $p\geq 1$. Consequently, the Markov chain $(L(U_p))$ visits the state y_0 with probability 1.

We now define a Lyapounov function on the state space of $(L(U_p))$ by

$$f(l)=\log(1+\alpha p+\beta(\|l\|-p)),$$

if $l=(l_i)$ and $p=\inf\{k-1/l_k\neq 2\}$. With the notations defined above, we have $f(L(U_1))=\log(1+\alpha A_{a,b}+\beta B_{a,b})$, consequently

$$\mathbb{E}_{R_{a,b,0}}(f(L(U_1))-f(L(U_0)))=\mathbb{E}_{R_{a,b,0}}\left(\log\left(\frac{1+\alpha A_{a,b}+\beta B_{a,b}}{1+\alpha a+\beta b}\right)\right).$$

According to Proposition 10, the random variables

$$(1+\alpha A_{a,b}+\beta B_{a,b})/(1+\alpha a+\beta b)$$

and their inverse are uniformly integrable. Consequently, the elementary inequality

$$|\log x|\leq x+\frac{1}{x},$$

for $x>0$, the convergence (31) and Lebesgue's theorem show that

$$\lim_{a+b\to+\infty}\mathbb{E}_{R_{a,b,0}}(f(L(U_1))-f(L(U_0)))=\log\gamma>0.$$

Hence there exists some constant K_0 such that if $a+b\geq K_0$,

$$\mathbb{E}_{R_{a,b,0}}(f(L(U_1))-f(L(U_0)))\geq(\log\gamma)/2. \tag{32}$$

In the same way, the following inequality holds

$$\mathbb{E}_{R_{a,b,0}}\left(|f(L(U_1))-f(L(U_0))|^2\right)\leq 6\log(\alpha\vee\beta)^2 \\ +4\mathbb{E}_{R_{a,b,0}}\left(\log^2\left(\frac{1+A_{a,b}+B_{a,b}}{1+\alpha a+\beta b}\right)\right).$$

The elementary inequality

$$\log^2 x\leq\frac{4}{e^2}\left(x+\frac{1}{x}\right),$$

for $x>0$ and the uniform integrability argument give the following inequality

$$\sup_{a,b:a+b>K}\mathbb{E}_{R_{a,b,0}}\left(|f(L(U_1))-f(L(U_0))|^2\right)<+\infty \tag{33}$$

A theorem by Lamperti [21], see also [14] or [24] states that if the two relations (33) and (32) are satisfied then the Markov Chain $(L(U_p))$ is transient. In particular this implies that there exists an initial state such the chain will never visit the state y_0 with positive probability. This contradicts our assumption on the recurrence of $(X(t))$. The theorem is proved. □

The case of symmetrical distributions. The distribution F is symmetrical if

$$F(dx) = p\delta_1 + (1-2p)\delta_2 + p\delta_3,$$

for $p \in [0, 1/2]$. Since the expected value of the size of the items is $1/2$ for all these distributions, the value λ_{FF} of the corresponding critical λ cannot exceed 2.

According to Theorem 14 the critical value of λ for the First Fit algorithm is given by

$$\lambda_{FF} = \frac{2 + p - \sqrt{4 + 4p - 15p^2}}{2p^2}$$

and for the FIFO policy, it is given by (see Kipnis and Robert [19])

$$\lambda_{FIFO} = \frac{12}{6 + 12p - 24p^2 + 10p^3 - p^4}$$

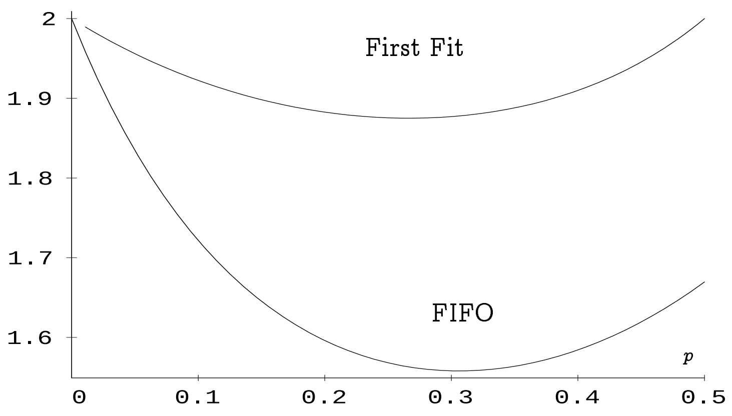

FIGURE 2. The effective bandwidth of the FIFO and First Fit policies for symmetrical distributions on $\{1, 2, 3\}$.

Coffman and Stolyar [6] considered the stability of the algorithms First Fit and Best Fit when the services are constant equal to 1. In that setting they have proved that the natural condition $\lambda E(S_1) < C$ is sufficient for the stability when the variable S_1 is uniformly distributed on $\{1/n, 2/n, \ldots, (n-1)/n\}$. This property is not true when the services are exponentially distributed (see figure 2).

The figure 3 shows that the First Fit algorithm is much more efficient than the FIFO policy. This comparison is intuitive since the First Fit algorithm reduces the wasted space in the bin. The minimal value of λ_{FF} is $15/8 = 1.875$ which is close to 2, the optimal value. Notice that in the case where there are only 1's and 3's equally likely, λ_{FF} is 2, hence is optimal; this is not the case for the FIFO policy.

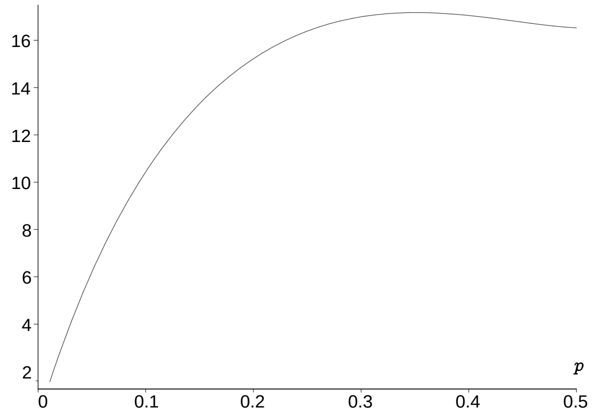

FIGURE 3. First Fit compared to FIFO for symmetrical distributions on $\{1, 2, 3\}$: Percentage of increase of the effective bandwidth.

The fact that, even in the case of symmetrical distributions, the constant λ_c is not $1/\mathbb{E}(S_1) = 2$ in general can be (roughly) explained as follows. The only time when there is some potentially wasted space is when an item 3 is in the bin. During that time some of the items 1 are served in the empty space left by the 3's. If all of them were served in this way, it is easy to see that it would imply $\lambda_c = 2$. This is not the case in fact. Indeed, in the description of the cycle of the previous section, we have seen that during some time 4 items 1 are in the bin. In particular these items 1 will not help to fill the empty spaces left by the 3's. Consequently, the condition $\lambda < 2$ is, in general, not sufficient to ensure ergodicity.

8. APPENDIX

The following elementary lemma is constantly used (some times implicitly) in the sections 3 and 5.

Lemma 16. *If $\mathcal{N}_\xi$, $\mathcal{N}_\mu$ are independent Poisson processes on $\mathbb{R}_+$ with respective parameters $0 < \xi < \mu$, and $(\nu(t))$ is a càdlàg (i.e. right continuous functions having a limit on the left at any point), non decreasing process on $\mathbb{R}_+$, independent of these Poisson processes, such that $(\nu(t)/t)$ converges to ν almost surely and in L_1, then*

$$Z(t) = \frac{1}{t}\Big((\mathcal{N}_\xi - \mathcal{N}_\mu)\,(]0, \nu(t)])\Big) - (\xi - \mu)\nu$$

converges to 0 almost surely and in L_1. If

$$\tau(t) = \inf\{s \geq 0 / \nu(t) + (\mathcal{N}_\xi - \mathcal{N}_\mu)(]0, s]) \leq 0\},$$

then $\tau(t)/t$ converges to $\nu/(\mu - \xi)$ almost surely and in L_1.

Proof. To prove the first part, it is enough to show that

$$Z_1(t) = \frac{1}{t}\Big((\mathcal{N}_\xi - \mathcal{N}_\mu)\,(]0, \nu(t)]) - (\xi - \mu)\nu(t)\Big)$$

converges to 0 almost surely and in L_1. For $t \geq 0$, using the independence properties, we get

$$\begin{aligned}\mathbb{E}(Z_1(t)^2) &= \frac{1}{t^2}\int_0^{+\infty} \mathbb{E}\left(\left(\left(\mathcal{N}_\xi - \mathcal{N}_\mu\right)(]0,x]) - (\xi-\mu)x\right)^2\right)\mathbb{P}(\nu(t)\in[x,x+dx])\\ &= \frac{1}{t^2}\int_0^{+\infty}(\xi+\mu)x\mathbb{P}(\nu(t)\in[x,x+dx]) = (\xi+\mu)\frac{\mathbb{E}(\nu(t))}{t^2},\end{aligned}$$

hence the convergence to 0 in L_1.

On the set $\{\nu > 0\}$, the law of large numbers for Poisson processes gives immediately the almost sure convergence of $(Z(t))$. If $\nu > 0$, for $\varepsilon > 0$ and t sufficiently large we have $\nu(t) \leq \varepsilon t$, hence

$$|Z(t)| \leq \frac{1}{t}\left(\left(\mathcal{N}_\xi + \mathcal{N}_\mu\right)(]0,\varepsilon t])\right),$$

consequently

$$\limsup_{t\to+\infty}|Z(t)| \leq (\xi+\mu)\varepsilon,$$

we get the almost sure convergence on $\{\nu = 0\}$ by letting ε go to 0.

Using the law of large numbers for Poisson processes, the almost sure convergence of $\tau(t)/t$ to $\nu/(\mu - xi)$ is straightforward to obtain. By definition of $\tau(t)$ we have,

$$\nu(t) + \left(\mathcal{N}_\xi - \mathcal{N}_\mu\right)(]0,\tau(t)]) \leq 0 \leq \nu(t) + \left(\mathcal{N}_\xi - \mathcal{N}_\mu\right)(]0,\tau(t)]) + 1,$$

taking the expected value of this inequality and using Wald's formula, we obtain,

$$\frac{1+\mathbb{E}(\nu(t))}{t} \geq (\mu-\xi)\frac{\mathbb{E}(\tau(t))}{t} \geq \frac{\mathbb{E}(\nu(t))}{t}.$$

Hence $\mathbb{E}(\tau(t))/t$ converges to $\mathbb{E}(\nu)/(\mu-\xi)$, this is enough to ensure the L_1-convergence of $(\tau(t))/t$ since it already converges almost surely. The lemma is proved. □

References

1. Philippe Bougerol and Jean Lacroix, *Products of random matrices with applications to Schrödinger operators*, Birkhäuser Boston Inc., Boston, Mass., 1985.
2. Maury Bramson, *Instability of FIFO queueing networks*, Annals of Applied Probability **4** (1994), no. 2, 414–431.
3. ———, *Instability of FIFO queueing networks with quick service times*, Annals of Applied Probability **4** (1994), no. 3, 693–718.
4. H. Chen and A. Mandelbaum, *Discrete flow networks: bottleneck analysis and fluid approximations*, Mathematics of Operation Research **16** (1991), no. 2, 408–446.
5. E.G. Coffman, Anja Feldman, Nabil Kahale, and Bjorn Poonen, *Computing call admission capacities in linear networks*, 1997, preprint.
6. E.G. Coffman and Alexander L. Stolyar, *Bandwidth packing*, 1998, preprint.
7. J. G. Dai, *On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models*, Annals of Applied Probability **5** (1995), no. 1, 49–77.
8. ———, *A fluid limit model criterion for instability of multiclass queueing networks*, Annals of Applied Probability **6** (1996), no. 3, 751–757.
9. Jean-François Dantzer, Mostapha Haddani, and Philippe Robert, *On the stability of a bandwidth packing algorithm*, Probability in the Engineering and Informational Sciences **14** (2000), no. 1, 57–79.
10. Vincent Dumas, *A multiclass network with non-linear, non-convex, non-monotonic stability conditions*, Queueing Systems, Theory and Applications **25** (1997), no. 1-4, 1–43.
11. ———, *Diverging paths in FIFO fluid networks*, IEEE Transactions on Automatic Control **44** (1999), no. 1, 191–194.

12. P. Dupuis and R.J. Williams, *Lyapounov functions for semimartingale reflecting brownian motions*, Annals of Applied Probability **22** (1994), no. 2, 680–702.
13. Stewart N. Ethier and Thomas G. Kurtz, *Markov processes*, John Wiley & Sons Inc., New York, 1986, Characterization and convergence.
14. G. Fayolle, V. A. Malyshev, and M. V. Men'shikov, *Topics in the constructive theory of countable Markov chains*, Cambridge University Press, Cambridge, 1995.
15. Y. Filonov, *A criterion for the ergodicity of discrete homogeneous Markov chains*, Akademiya Nauk Ukrainskoi SSR. Institut Matematiki. Ukrainskii Matematicheskii Zhurnal **41** (1989), no. 10, 1421–1422.
16. A.S. Gaĭrat, V.A. Malyshev, M. V. Men'shikov, and K.D. Pelikh, *Classification of Markov chains describing the evolution of random strings*, Russian Mathematical surveys **50** (1995), no. 2, 237–255.
17. R. Z. Has'minskiĭ, *Stochastic stability of differential equations*, Sijthoff & Noordhoff, Alphen aan den Rijn, 1980, Translated from the Russian by D. Louvish.
18. Morris W. Hirsch and Stephen Smale, *Differential equations, dynamical systems, and linear algebra*, Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1974, Pure and Applied Mathematics, Vol. 60.
19. C. Kipnis and Ph. Robert, *A dynamic storage process*, Stochastic Processes and their Applications **34** (1990), 155–169.
20. P.R. Kumar and T.I. Seidman, *Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems*, IEEE Transactions on Automatic Control **35** (1990), no. 3, 289–298.
21. J. Lamperti, *Criteria for the recurrence or transience of stochastic process. I*, Journal of Mathematical Analysis and Applications **1** (1960), 314–330.
22. S.H. Lu and P.R. Kumar, *Distributed scheduling based on due dates and buffer priorities*, IEEE Transactions on Automatic Control **36** (1991), 1406–1416.
23. V. A. Malyshev, *Stabilization laws in the evolution of a random string*, Problems on Information Transmission **30** (1994), no. 3, 79–95.
24. S. Meyn and R. Tweedie, *Markov chains and stochastic stability*, Communications and control engineering series, Springer, 1993.
25. Sean P. Meyn, *Transience of multiclass queueing networks via fluid limit models*, Annals of Applied Probability **5** (1995), no. 4, 946–957.
26. A.A. Puhalskii and A. Rybko, *Non-ergodicity of queueing networks under non-stability of their fluid models*, November 1998.
27. Philippe Robert, *Smooth initial distributions and fluid limits for multi-class queueing systems*, 2000, In preparation.
28. A. N. Rybko and A. L. Stolyar, *On the ergodicity of random processes that describe the functioning of open queueing networks*, Problems on Information Transmission **28** (1992), no. 3, 3–26.
29. E. Seneta, *Nonnegative matrices and Markov chains*, second ed., Springer-Verlag, New York, 1981.
30. Richard Serfozo, *Introduction to stochastic networks*, Springer-Verlag, New York, 1999.

(J.-F. Dantzer) Département de Mathématiques Bâtiment Fermat Université de Versailles-St-Quentin 45, Avenue des États-Unis 78035 Versailles
E-mail address: dantzer@fermat.math.uvsq.fr

(Ph. Robert) INRIA, domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France
E-mail address: Philippe.Robert@inria.fr
URL: http://algo.inria.fr/robert/