Portugal. Math. (N.S.)
Vol. 65, Fasc. 4, 2008, 535–568

**Portugaliae Mathematica**


# Study of the maximal throughput of multiclass queueing systems

Faiza Belarbi and Amina Angelika Bouchentouf

(Communicated by Raul Cordovil)

**Abstract.** The stability properties of the bandwidth allocation algorithm First Fit are analyzed for the distributions concentrated on three sizes for the requests and the bin equal to 5. To analyze these processes we introduce the notion of a smooth initial state. Starting from a smooth initial state the fluid limits of these systems are investigated.

**Mathematics Subject Classification (2000).** 37A25.

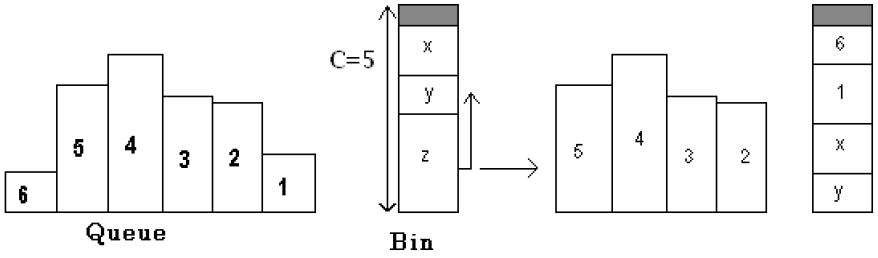
**Keywords.** Bin packing algorithms, ergodicity, fluid limits, multi-class queueing systems, bandwidth allocation.

## 1. Introduction

Our study is about a queueing model of storage and transmission bandwidth in a computer and communication systems. The model which 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 throughput, i.e., variable portions of the offered bandwidth $C$ of the network. The sum of throughput 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 throughput 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 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 $\lambda$ 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 stings). The problems investigated in this paper concern the stability properties of this bandwidth allocation problem: Under some probabilistic assumptions on the sizes of the items, what is maximum input rate under which the size of the queue converges in distribution?




A departure for the First Fit algorithm

**Related models.** A similar problem has been analyzed by Kipnis and Robert [16] with the FIFO algorithm: an item enters the bin only if all the items arrived before it have left 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, with the exception of the first, are i.i.d. random variables with distribution $\mu$. 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 $\mu$. 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 [5] analyzed the stability of the algorithms First Fit and Best Fit (Best Fit algorithm: the largest message that fits is transmitted) when the

services are constant equal to 1. In this setting the problem is related to static bin packing problems. They prove that the natural condition $\lambda \mathbb{E}(S_1) < C$ is sufficient for the stability in the case of a symmetrical distribution of the sizes; in Coffman et al. [4] the sufficiency for stability of the condition $\lambda \mathbb{E}(S_1) < C$ 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. [8]. In that paper the stability condition has been established and, more interesting, a curious transient behaviour has been analyzed. The present paper is a generalization of this work. The case analyzed here requires a much more detailed analysis of the evolution of the string structure than it was necessary in [8].

**An overview.** In this paper we give a necessary and sufficient condition under which the size of the queue converges in distribution (Theorem 2 and 3). If this condition has some interesting features, it is expressed as 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 in general difficult to analyze. The paper proposes an approach to analyze such processes. To keep the presentation simple, the simplest of these complicated models has been chosen.

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.

String valued processes can travel in infinitely many direction. 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 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 Section 4 where the ergodicity condition is quite intuitive, but its proof requires some discussion on the possible bifurcation of the system. This situation seems unavoidable, especially when the ergodicity condition is not natural at all (see Section 6).

The present 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 et al. [8] 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.

Next we study the fluid limits of the distributions of the process. The fluid limits can be described by piecewise linear processes in $\mathbb{R}_+^2$. The associated dynamical system turns out to be a product of random $2 \times 2$ matrices in $\mathbb{R}_+^2$; its stability properties are analyzed. These results are then used to derive the ergodicity and transience conditions for the Markov processes.

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

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

Since $X(0) = x$, $X_x(0) = 1$. The time variable are scaled by factor $\|x\|$. The following theorem is the combination of two results, one due to Filonov [13] and the other due to Rybko and Stolyar [24]. It provides an ergodicity criterion.

**Theorem 1.** *If there exist an integrable stopping time $U$, constants $K$ and $\epsilon > 0$ such that*

$$\limsup_{\|x\| \to +\infty} \frac{\mathbb{E}_x\big(\|X(U)\|\big)}{\|x\|} \leq 1 - \epsilon, \tag{2.1}$$

$$\limsup_{\|x\| \to +\infty} \frac{\mathbb{E}_x(U)}{\|x\|} \leq K, \tag{2.2}$$

*the Markov process $\big(X(t)\big)$ 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.*,

$$\limsup_{\|x\|\to+\infty} \frac{\mathbb{E}_x(H^2)}{\|x\|^2} < +\infty \tag{2.3}$$

The condition (2.1) requires that at some random time, $U/\|x\|$, the norm of the rescaled process $\big(X_x(t)\big)$ is, in average, below its initial value. This suggests the analysis of the sequences of processes $\big(X_x(t)\big)$, 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 integrability argument, Theorem 1 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 [15] for the classical results). Has'minskii [14] 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 [1], [2], Dumas [9], Lu and Kumar [19], Kumar and Seidman [17] and Rybko and Stolyar [24] among others—gave an impulse to the studies in this domain recently. Chen and Mandelbaum [3] used fluid limits to study Jackson networks. Dai [6] set a framework to apply these methodes to prove Harris ergodicity for some queueing networks. Concerning transience criteria, Dai [7], Meyn [20] and Puhalskii and Rybko [22] 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 [10] and the references therein).

Relations (2.1) and (2.2) 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 considerd the process when some of the coordinates of the initial state are large. In general, Theorem 1 can then be applied when all the possibilities for the large coordinates have been considered. Applying Theorem 1 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 "patterns". We establish that, starting from any large arbitrary state, the process will eventally travel along some smooth random states.

## 3. The string valued Markov process

The items arrive according to a Poisson process $\mathcal{N}_\lambda$ with parameter $\lambda$; 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 5.

The size $(S_i)$ 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 $\delta_x$ 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 $\mathscr{T} = \{1, 2, 3\}$ and $\mathscr{T}^{(\mathbb{N})}$ is the set of finite vectors with coordinates in $\mathscr{T}$ if $x \in \mathscr{T}^{(\mathbb{N})}$, $\|x\|$ 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 element $X$ of the state space $\mathscr{S}$ of the Markov process describing the storage process can be written as $X = (B, L)$, where $L$ and $B$ are elements of $\mathscr{T}^{(\mathbb{N})}$, the set of finite vectors with coordinates in $\mathscr{T}$. The vector $B = (b_j; j = 1, \ldots, \|B\|)$ 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 $L = (l_i, i = 1, \ldots, \|L\|)$ 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 $i = 1, \ldots, \|L\|$ the following inequality holds

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

Since $C = 5$, the possible values of $B$ are the following

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

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 and 3.

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

- Arrival. At rate $\lambda$ 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 $L(t)$.
- 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 $\big(X(t)\big)$ is an irreducible Markov process on $\mathscr{S}$. We shall say that the model is stable when $\big(X(t)\big)$ is an ergodic Markov process on $\mathscr{S}$. In Dantzer et al. [8] it has been proved that the condition $\lambda\mathbb{E}(\mathrm{S}_1) \leq C$ is necessary for the stability of the system, i.e., the Markov process $\big(X(t)\big)$ is transient if $\lambda\mathbb{E}(\mathrm{S}_1) > C$.

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

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

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

Notice that the load of the system increases at rate $\lambda\mathbb{E}(\mathrm{S}_1)$ in average and decreases at most at rate 5.

## 4. The natural condition is sufficient for ergodicity

We study a case where it is not necessary to know much about the structure of $L$-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 1.** *Under the conditions $\lambda\mathbb{E}(\mathrm{S}_1) < 5$, if $r = 0$ (only items 1 and 2 arrive) and*

$$\tau = \inf\{t \leq 0 : \|L(t)\| = 0\} \qquad \text{and} \qquad D = \int_0^\tau \mathbb{1}_{\{b_1(t)+\cdots+b_{\|B(t)\|}(t)<5\}}\, dt,$$

*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 $x = (l, b) \in \mathscr{S}$.*

*Proof*. The variable $D$ is the duration of time during which the bin is not full over a busy period. Notice first that there is no waste of space as long as there are items 1 in the $L$-component of $\big(X(t)\big)$. 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 $\big(B(t)\big)$ of the bin is $(2,2)$, $(2,1,1)$ or $(1,1,1,1)$ (an empty space of size 1).

We set $A_0 = l$ and $T_0 = 0$ and by induction we define

$$T_{n+1} = \inf\{t > T_n : C(t-) = 5, C(t) < 5 \text{ and all the items 2 present at time } T_n \text{ are served at time } t\}$$

with $C(s) = b_1(s) + \cdots + b_{\|B(s)\|}(s)$ 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 $\big(B(T_n), A_n\big)$ is clearly a Markov chain.

If $b$ the initial state of the bin is $(1,1,1,1)$ (an empty space of size 1). As long there is at least an item 1 in the queue, because of the First Fit discipline, the items 2 are ignored. Since $\lambda p \leq \lambda \mathbb{E}(\mathrm{S}_1) < 5$, 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 $M/M/5$ queue (5 servers)with parameter $\lambda p$ for the input rate and 1 for the service rate.

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

(1) $\lambda p < 2$. This condition clearly implies that, with probability 1, at some time there will be no item 1 in the system and, consequently, the 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): "$(2,1,1) \Rightarrow (2,1) \Rightarrow (1,\cdot) \Rightarrow (1,2) \Rightarrow (2,2) \Rightarrow (2,\cdot) \Rightarrow (2,2)$".

When the initial items 2 have been served the queue is an i.i.d. strings of items 2 and 1. At that moment an item 1 will enter the bin, then two items 1: "$(1,2,2) \Rightarrow (1,1,2)$".

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 three places in the bin, "$(2,1,1) \Rightarrow (2,1,1,1)$". In this case, the number of items 1 is the number of customers of transient

$M/M/3$ queue starting with two customers (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 $M/M/3$ 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 three items 1. The condition $\lambda\mathbb{E}(\mathrm{S}_1) < 5$ implies that $\lambda q < 1$ ($\lambda p > 2$); therefore, with probability 1, after some period of time the system will not contain any items 2. At that moment the state of the bin will be $(1,1,1,1,1)$. (Recall that $\lambda p \leq \lambda\mathbb{E}(\mathrm{S}_1) < 5$.) 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 $(2,1,1)$, this is the starting situation.

b) The queue reaches the empty state. After an integrable amount of time two items 2 occupy the bin: "$(2,1,1) \Rightarrow (2,1,1,1) \Rightarrow (2,2,1) \Rightarrow (2,2,\cdot)$". 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 Proposition 16 of the appendix of Dantzer et al. [8], it is easy to check that there exists some constant $c > 0$ such that the following convergence holds in $L_1$ and almost surely:

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

(the 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):

$$\mathbb{E}_x\big(W\big(X(T_1)\big)\big) \leq \mathbb{E}_x\big(W(x)\big) + \mathbb{E}\Big(\sum_{i=1}^{\mathcal{N}_\lambda([0,T_1])} \mathrm{S}_i\Big) - 5\mathbb{E}_x(T_1 - I).$$

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

$$\mathbb{E}_x\big(W\big(X(T_1)\big)\big) \leq \mathbb{E}_x\big(W(x)\big) + \big(\lambda\mathbb{E}(\mathrm{S}_1) - 5\big)\mathbb{E}_x(T_1) + 5\mathbb{E}_x(I).$$

Since there are no items 1 in the queue at 0 and $T_1$ (let us return to the case to a)), we have

$$\mathbb{E}_x\big(W\big(X(0)\big)\big) = 3\|x\| \quad \text{and} \quad \mathbb{E}_x\big(W\big(X(T_1)\big)\big) = 4 + 3\mathbb{E}_x(A_1).$$

The quantity $\mathbb{E}_x(I)$ being bounded with respect to $\|x\|$, it follows that

$$3\limsup_{\|x\|\to+\infty}\frac{\mathbb{E}_x(A_1)}{\|x\|}=\limsup_{\|x\|\to+\infty}\frac{\mathbb{E}_x\big(W\big(X(T_1)\big)\big)}{\|x\|}\leq 3+c\big((\lambda\mathbb{E}(\mathrm{S}_1)-5\big),$$

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

$$\mathbb{E}_x(A_1)\leq\alpha\|x\|, \tag{4.1}$$

where $\alpha=1+\frac{C}{3}\big(\lambda\mathbb{E}(\mathrm{S}_1)-5\big)$ and

$$\gamma=-\log\left(\frac{1+\alpha\|x\|}{1+\|x\|}\right)>0. \tag{4.2}$$

If we set

$$v=\inf\{n\geq 1: A_n\leq a_0\},$$

the sequence

$$(Z_n)=\big(\log(1+A_{n\wedge v})+\gamma(n\wedge v)\big).$$

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

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

by Jensen's inequality and the relations (4.1) and (4.2). Consequently $\mathbb{E}(Z_n)\leq Z_0$, hence $\gamma\mathbb{E}(n\wedge v)\leq\mathbb{E}(Z_n)\leq Z_0$. By letting $n$ go to infinity, we get

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

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 size of the initial state. By Wald's formula, the contribution of the $v$ cycles in the integral defining $D$ is bounded by $K\mathbb{E}_x(v)\leq K\log(1+\|x\|)/\gamma$ for some constant $K$.

Since $\lambda\mathbb{E}(\mathrm{S}_1)<1$, the proposition 6 of Dantzer et al. [8] shows that the system is ergodic. Consequently, starting from the state $\big(B(T_v),A_v\big)(\leq a_0)\big)$, 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 $\|x\|$.

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

1) $\lambda p < 2$. After an integrable amount of time, the item 2 will enter the bin. Then all the other items 2 will be served consecutively. The expected value of this duration of time is easily seen to be bounded with respect to $\|l\|$. Starting from that time, the queue will be an i.i.d. string of items 2 and 1. At that moment an item 1 will enter the bin "$b = (1, 2, 2)$"; then, with probability 1, two items 1 will enter the bin "$b = (2, 1, 1)$", so we will come back to the previous case "$b = (2, 1, 1)$ and $\lambda p < 2$" and will follow the same discussion.
2) $\lambda p > 2$. This condition implies that the arriving items 1 will saturate three places in the bin "$(2, 1, 1, 1)$". In this case, the number of items 1 is the number of customers of transient $M/M/3$ queue. After amount of time, with probability 1 the bin will be $(2, 1, 1)$, so here we will come back to the case where "$b = (2, 1, 1)$ and $\lambda p > 2$" and will follow the same discussion.

The lemma is proved. □

Now we are going to introduce a lemma which gives an estimation of the wasted space when there are only two possible sizes: 1 and 3.

**Lemma 2.** *Under the conditions* $\lambda \mathbb{E}(S_1) < 5$, *if* $q = 0$ (*only items* 1 *and* 3 *arrive*) *and*

$$\tau = \inf\{t \le 0 / \|L(t)\| = 0\} \quad \text{and} \quad D = \int_0^\tau \mathbb{1}_{\{b_1(t)+\cdots+b_{\|B(t)\|}(t)<5\}}\, dt,$$

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

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

*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 $\big(X(t)\big)$. We can therefore assume that $l'$ is a string of items 3. In this context the only possibility to waste space with a non empty queue is when the state $\big(B(t)\big)$ of the bin is $(1, 1, 1, 1)$ or $(3, 1)$ (an empty space of size 1).

We set $A_0 = l'$ and $T_0 = 0$ and by induction we define

$$T_{n+1} = \inf\{t > T_n : C(t-) = 5, C(t) < 5 \text{ and all the items 3 present at time } T_n \text{ are served at time } t\}$$

with $C(s) = b_1(s) + \cdots + b_{\|B(s)\|}(s)$ and $A_n = \|L(T_n)\|$ for $n \geq 1$. Notice that $L(T_n)$ is necessarily a (possibly empty) string of items 3. The sequence $\big(B(T_n), A_n\big)$ is clearly a Markov chain.

If $b$ the initial state of the bin is $(1,1,1,1)$ (an empty space of size 1). As long there is at least an item 1 in the queue because of the First Fit discipline, the items 3 are ignored. Since $\lambda p \leq \lambda \mathbb{E}(\mathrm{S}_1) < 5$, after an integrable amount of time not depending on $\|l'\|$, at least two places will be vacant in the bin and consequently an item 3 will enter the bin "$(1,1,1,1) \Rightarrow (1,1,\cdot,\cdot) \Rightarrow (3,1,1)$".

In this situation the number of items 1 is the number of customers of an $M/M/5$ queue (5 servers) with parameter $\lambda p$ for the input rate and 1 for the service rate.

If $b = (3,1)$. 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, a second item 3 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 3 are served as long as the initial items 3 are present (since these items are located at the beginning of the queue, the First Fit algorithm selects them): "$(3,1) \Rightarrow (3,\cdot) \Rightarrow (3,\cdot) \Rightarrow (3,\cdot)$".

When the initial items 3 have been served the queue is an i.i.d. string of items 3 and 1. At that moment an item 1 will enter the bin, then two items 1 "$(1,3,\cdot) \Rightarrow (1,1,3)$".

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, "$(3,1) \Rightarrow (3,1,1)$". In this case, the number of items 1 is the number of customers of 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 3 and two items 1. The condition $\lambda \mathbb{E}(\mathrm{S}_1) < 5$ implies that $\lambda r < 1$ ($\lambda p > 2$). Therefore, with probability 1 after some period of time the system will not contain any items 3. At that moment the state of the bin will be $(1,1,1,1,1)$. (Recall that $\lambda p \leq \mathbb{E}(\mathrm{S}_1) < 5$.) It is easily seen that, with probability 1, the total number of items 1 will be less than 2. An item 3 will be in the bin at that time $(3,1)$, this is the starting situation.

b) The queue reaches the empty state. After an integrable amount of time item 3 and two items 1 occupy the bin. The initial items 3 and 1

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 $\|I'\|$).

Using Proposition 16 of the appendix of Dantzer et al. [8], it is easy to check that there exists some constant $c > 0$ such that the following convergence holds in $L_1$ and almost surely:

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

(the 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):

$$\mathbb{E}_x\big(W\big(X(T_1)\big)\big) \le \mathbb{E}_x\big(W(x)\big) + \mathbb{E}\Big(\sum_{i=1}^{\mathcal{N}_\lambda([0,T_1])} \mathrm{S}_i\Big) - 5\mathbb{E}_x(T_1 - I).$$

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

$$\mathbb{E}_x\big(W\big(X(T_1)\big)\big) \le \mathbb{E}_x\big(W(x)\big) + \big(\lambda\mathbb{E}(\mathrm{S}_1) - 5\big)\mathbb{E}_x(T_1) + 5\mathbb{E}_x(I).$$

Since there are no items 1 in the queue at 0 and $T_1$ (returned to a)), we have

$$\mathbb{E}_x\big(W\big(X(0)\big)\big) = 2\|x\| \quad \text{and} \quad \mathbb{E}_x\big(W\big(X(T_1)\big)\big) = 4 + 3\mathbb{E}_x(A_1).$$

The quantity $\mathbb{E}_x(I)$ being bounded with respect to $\|x\|$, it follows that

$$2\limsup_{\|x\|\to+\infty} \frac{\mathbb{E}_x(A_1)}{\|x\|} = \limsup_{\|x\|\to+\infty} \frac{\mathbb{E}_x\big(W\big(X(T_1)\big)\big)}{\|x\|} \le 2 + c\big(\lambda\mathbb{E}(\mathrm{S}_1) - 5\big),$$

where $W(\cdot)$ is the load (see Definition 1). The rest of the proof is the same as the previous lemma. □

Now we consider the general case with three sizes. The condition $\lambda\mathbb{E}(\mathrm{S}_1) < 5$ turns out to be sufficient for ergodicity when $\lambda p > 2$.

**Proposition 1.** *If $\lambda\mathbb{E}(\mathrm{S}_1) < 5$ and $\lambda p > 2$, then $\big(X(t)\big)$ is an ergodic Markov process.*

*Proof.* Let $(x_n) = (b_n, l_n)$ be a sequence of $\mathscr{S}$ 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.* [8], 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 $\tau$ the first time when the bin is not full after all the initial items 2 and 3 have left the system; $\tau$ is clearly a (possibly infinite) stopping time. If $D$ is the duration of time between time 0 and $\tau$ 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 $\tau$ the load of the system is decreased at rate 5, except during some periods of total duration $D$, i.e., for $t \geq 0$ we have

$$\sum_{i=1}^{\|b\|} b_i \sigma_i^0 + \sum_{i=\|b\|+1}^{\|l_n\|+\|b\|} l_{n,i}\sigma_i^0 + \sum_{i=1}^{\mathscr{N}_\lambda([0,t\wedge\tau])} \mathrm{S}_i\sigma_i - 5(t\wedge\tau - D) \geq 0.$$

The sequences $(\sigma_i)$ and $(\sigma_i^0)$ 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

$$\mathbb{E}_{x_n}(t\wedge\tau)\big(5 - \lambda\mathbb{E}(\mathrm{S}_1)\big) \leq \|x_n\| + 5\mathbb{E}_{x_n}(D).$$

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

$$\limsup_{n\to+\infty} \frac{\mathbb{E}_{x_n}(\tau)}{\|x_n\|} \leq \frac{1}{5 - \lambda\mathbb{E}(\mathrm{S}_1)}. \tag{4.3}$$

In the same manner, we have

$$\mathbb{E}_{x_n}\big(W\big(X(t\wedge\tau)\big)\big) \leq W(x_n) + \big(\lambda\mathbb{E}(\mathrm{S}_1) - 5\big)\mathbb{E}_{x_n}(\tau\wedge t) + 5\mathbb{E}_{x_n}(D);$$

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

$$\mathbb{E}_{x_n}\big(W\big(X(\tau)\big)\big) \leq W(x_n) + \big(\lambda\mathbb{E}(\mathrm{S}_1) - 5\big)\mathbb{E}_{x_n}(\tau) + 5\mathbb{E}_{x_n}(D).$$

Since all the initial items 2 and 3 are served at time $\tau$, at most two of the initial items can be served at the same time (see the papers by Haddani, Dantzer and Robert), hence

$$\mathbb{E}_{x_n}(\tau) \geq \frac{\|x_n\|}{2} \geq \frac{W(X_n)}{8}.$$

Thus

$$\limsup_{n\to+\infty} \frac{\mathbb{E}_{x_n}\big(W\big(X(\tau)\big)\big)}{W(x_n)} \leq 1 + \frac{\lambda\mathbb{E}(\mathrm{S}_1) - 5}{8} < 1. \tag{4.4}$$

By using the fact that $W(\cdot)$ is an equivalent norm to $\|\cdot\|$ on $\mathscr{S}$, the relations (4.3) and (4.4) and Theorem 1 of Dantzer et al. [8] show that the Markov process $\big(X(t)\big)$ is ergodic.

All we have to prove now is that $\big(\mathbb{E}_{x_n}(D)\big)$ 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 the random variable $H$ is a "bounded integrable variable" if the sequence $\big(\mathbb{E}_{x_n}(H)\big)$ is bounded with respect to $\|x\|$.

(1) $b$ is $(1,1,1,1)$ or $(1,1,1,1,1)$.
As long as there is at least an item 1 in the queue, all the other items are ignored. The condition $\lambda\mathbb{E}(\mathrm{S}_1) < 5$ implies that $\lambda p < 5$. From the point of view of the items 1, the system is a stable $M/M/5$ queue. Hence the first time there will be at least two empty places is a bounded integrable variable (since $\lambda p > 2$). 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) $b$ has at least an item 2.
   a) $b = (2,\cdot)$ (the bin contains only one item 2). At that time the items 3 are selected by the First Fit algorithm. Since the condition $\lambda\mathbb{E}(\mathrm{S}_1) < 5$ implies that $\lambda(2q + 3r) < 5$, the system with the items 2 and 3 is stable (see Dantzer et al. [8]). After the amount of time, $\lambda p > 2$ implies that the residual space in the bin left by the items 2 is saturated by the items 1. Consequently the duration of time the bin is not full is a bounded integrable variable.
   b) $b = (2,2,\cdot)$ (the bin contains two items 2). In this case the items 3 are ignored. Consequently, a string of items 3 builds up at the beginning of the queue. Since $\lambda\mathbb{E}(\mathrm{S}_1) < 5$ implies that $\lambda(p + 2q) < 5$, the system with the items 1 and 2 is stable (see Dantzer et al. [8]), then until an item 3 enters the bin the wasted space is negligible compared to the number of initial items 2.

(3) $b$ contains a 3.

a) $b = (3, \cdot)$. All the other items 3 are ignored. At that time the items 2 are selected by the First Fit algorithm. Since $\lambda\mathbb{E}(S_1) < 5$ implies that $\lambda(2q + 3r) < 5$, the system with the items 2 and 3 is stable (see Dantzer et al. [8]). Then the condition $\lambda p > 2$ implies that the duration of time the bin is not full is a bounded integrable variable.

b) If $b = (3, 1)$. The items of size 2 are ignored. Since $\lambda\mathbb{E}(S_1) < 5$ implies that $\lambda(p + 3r) < 5$, the system with the items 1 and 3 is stable (see Dantzer et al. [8]). After amount of time, $\lambda p > 2$ implies that the residual space in the bin left by the items 2 and 3 is saturated by the items 1. Consequently the duration of time the bin is not full is a bounded integrable variable.

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

The result of the above proposition is fairly easy to understand: under the condition $\lambda p > 2$, basically there is no waste of space so that the natural condition $\lambda\mathbb{E}(S_1) < 5$ is sufficient for the ergodicity of $\big(X(t)\big)$. Notice however that the proof of this intuitive result (Lemma 1 and Proposition 1) has required the detailed analysis of the possible evolution starting from a given initial state. As we shall see, the situation is more delicate in the case $\lambda p < 2$.

## 5. Smoothing the initial state

In this section we shall assume that $\lambda p < 2$ and that the initial states are strings of items 2 and 3.

**Definition 2.** For $X(0) = x \in \mathscr{S}$ and $t \geq 0$ if $X(t) = \big(B(t), L(t)\big)$ and $L(t) = \big(l_i(t)\big)$, let

$$v_{x,1}(t) = \inf\{k \geq 1 : l_k(t) = 1\},$$
$$v_{x,2}(t) = \inf\{1 \leq k < v_{x,1}(t) : l_k(t) = 2\},$$
$$v_{x,3}(t) = \inf\{1 \leq k \leq v_{x,2}(t) : l_k(t) = 3\},$$

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

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

**Definition 3.** For integers $l$, $m$, $n$ we define the distribution $R_{l,m,n}(dx)$ on $\mathscr{T}^{(\mathbb{N})}$ by

$$R_{l,m,n}(dx) = \delta_3(du)^{(l)} \otimes F_{2,3}(du)^{(m)} \otimes F(du)^{(n)}, \tag{5.1}$$

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

A distribution $\mu$ on $\mathscr{S}$ is smooth if its $L$-component is in the convex hull of the $R_{l,m,n}^{l\times m\times n}$, $l,m,n \in \mathbb{N}$, 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 2.** *If $\lambda\mathbb{E}(S_1) < 5$, 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 tag at time $t$, $\big(M(t)\big)$ 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 $\lambda$. Recall that the queue of our initial state has no more than one item 1 "at most one item 1". As long as some initial items 2 are in the queue, the First Fit algorithm picks (possibly) these items. Once all initial items 2 are served the First Fit algorithm looks for items 1 which are just after the tag to not lose much time in searching for items 2, so the departure of some of the items 1 builds a string of 2, 3 after the tag. In the case where all the initial items 2 are processed and that some initial items 3 remain, the next items 3 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 $\tilde{v}_a$, $a \in \mathscr{T}$ is analogous to the definition 2 except that it concerns only the portion of the queue after the tag,

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

Notice that if $\tilde{v}_3(t)$ is finite, then necessarily $\tilde{v}_3(t) = M(t) + 1$. 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 $U \le t$ then $\tilde{L}(t) = L(t)$ and $\tilde{v}_a = v_a$ for $a \in \mathscr{T}$.

*Assertion.* If $\tau$ is a stopping time, then conditionally on $\tilde{\nu}_a(\tau)$, $a \in \mathscr{T}$, and $\|L(\tau)\|$, the distribution of $\tilde{L}(\tau)$ is given by (5.1) for some convenient $l, m, n \in \mathbb{N}$.

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 the latter, we 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

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

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 $\tau$ the First Fit algorithm never scanned the queue after the position $\tilde{\nu}_1(\tau)$, otherwise the items 1 located there would have been taken in the bin. The string $H_1$ is thus independent of $H_2$ and $H_3$. 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 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 $\big(\tilde{\nu}_1(\tau) - \tilde{\nu}_2(\tau) - 2\big)^+$ and distribution $F(du/u \geq 2)$. The assertion is proved. □

**Proposition 3.** *If $\lambda\mathbb{E}(S_1) < 5$, $\lambda p < 2$ and $U_0$ is the first time $t$ after all the initial items have left the queue that $B(t) = (3, 1)$ then*

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

*where $(\mathscr{S}_1)$ is the subset of the states of $S$ for which the bin is not full,*

$$\mathscr{S}_1 = \{x = (b, l) \in \mathscr{S} : b_1 + \cdots + b_{\|b\|} < 5\}.$$

*Proof.* The initial state $X(0)$ is given by $x = (B, L)$ with $L = (l_1, \ldots, l_p)$ 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 $k \in \{1, \ldots, p-1\}$, 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 $\tau_i$, $\check{\tau}_i$ 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 $\tau_k$, thus $\check{\tau}_k = \tau_k$.
- Otherwise, when $\tau_{k+1} \geq \tau_k$ and the initial state is $\check{x}$, at time $\tau_k$ the First Fit algorithm checks whether 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 $L' = (3, \ldots, 3, 2, \ldots, 2)$, $L'$ is a permutation of $L$ with all the items 3 at the beginning.

Similarly, the relation $\mathbb{E}_x(T_2) \leq \mathbb{E}_{x''}(T_2)$ holds if $T_2$ 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 so on 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 "the state of the bin will be $(2, 2, 1)$ (without items 3)". When the system without items 3 has at most one item 1 in the bin, an item 2 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 $\lambda \mathbb{E}(S_1) < 5$ implies that $\lambda(p + 2q)/(p + q) < 5$, the system without the items is ergodic. Using the ergodicity result of Dantzer [8] and inequality (2.3) of Theorem 1, 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 "$b = (3, \cdot)$", all the initial 3's are served consecutively and then the initial 2's and 3's are served "First Fit algorithm selects all the initial items 2". 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 3 in the bin and at least a 1 "$b = (3, 1)$", the situation is more interesting. In contrast to 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 $(3, 1)$ (there is an empty space of size 1) no new item 1 arrives before a departure from the bin. If the item 3 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 and an item 2 enters the bin, then all the initial items 2 are processed. Since $\lambda p < 2$, if

there are sufficiently many 2's in the queue, one of these 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 $\overline{T}$ the first time when the state of the bin is $(3,1)$. It is sufficient to prove that if the initial state is $x$, $\overline{T}$ 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, one has to process these ones until an additional item 1 or 3 enters the bin. There are two possibilities:

- The bin has at least one item 1 "$b = (1, \cdot)$". Since $\lambda p < 2$, after some time the queue will not have any item 1 and the bin will have two items 1 "$b = (1, 1, \cdot)$".
  - If, at that time, there are sufficiently many items 3 in the queue, the state of the bin will reach the state $(3, 1, 1)$; then with probability 1 the state of the bin will be $(3, 1)$.
  - If not, all the items 2 in the queue at that time are served "$b = (1, 1, 2)$". When this is finished, the condition $\lambda p < 2$ 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 $(3, 1)$. $((1, 1, 2) \Rightarrow (1, 1, 2) \cdots \Rightarrow (1, 2) \Rightarrow (1, 2) \Rightarrow (1, 3) \Rightarrow (1, 3) \ldots)$.
- The state of the bin is $(3, 2)$.
  - If item 2 is served before item 3, at that moment the First Fit algorithm selects all the initial items 2. When this is finished, with probability 1, two items 1 enter in the bin, $(3, 1, 1)$. The condition $(\lambda p < 2)$ implies that, with probability 1, the bin will reach the state $(3, 1)$.
  - If item 3 is served before item 2, the First Fit algorithm selects all the initial items 3, "$(3, 2)$". Then all the initial items 2 will be served consecutively at rate 2, $((2, \cdot) \Rightarrow (2, 2, 1) \Rightarrow (2, 2, 1))$. When this is finished and 2 goes out from the bin, with probability 1, an item 1 enters the bin. This is the situation of the previous case "$(2, 1, 1)$".

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 $U_0$. □

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

In this section we assume that $(\mu_n)$ is a sequence of smooth distributions on $\mathscr{S}$ (see Definition 3) such that

$$\mu_n\big(B=(3,1), L \in dx\big) = \mathbb{E}\big(R_{a_n,b_n,0}(dx)\big), \tag{6.1}$$

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

$$\mathbb{E}_{\mu_n}\big(f\big(L(0)\big)\mathbb{1}_{\{B(0)=(3,1)\}}\big) = \mathbb{E}\Big(\int_{\mathscr{T}^{(\mathbb{N})}} f(x)R_{a_n,b_n,0}(dx)\Big),$$

where $a_n$, $b_n$ 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 $\mu_n$ is $(3,1)$ and the $L$-component of the distribution $\mu_n$ 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 probabilities $q/(q+r)$ and $r/(q+r)$.

**Definition 4.** A sequence $(X_n)$ of random variables is equivalent to $(\alpha_n)$ if the sequence $(X_n/\alpha_n)$ converges to 1 in $L_1(\mathbb{P})$.

**A random transition of the fluid model.** If the initial distribution is given by $\mu_n$, the initial state of the bin is $(3,1)$. If there is a departure before a new arrival with

1) probability $1/2$, this is the item 3, then an item 3 enters the bin, and then all the other $a_n-1$ items 3 (it remain an empty space of size 1);
2) probability $1/2$, this is the item 1, another item 2 enters the bin, and then all the other initial items 2 will be served consecutively.

We remark that the dynamic of the system is influenced by the fact that either the 3 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 et al. [8]. Here the randomness remains because of this 1/2-1/2 transition and not because there are many possibilities for the contents of the bin. In Robert [23], it is shown that this random bifurcation may depend on the current state; this is not the case here.

If the distribution of $X(0)$ is given by $\mu_n$, then $\|X(0)\|$ is equivalent to $\big((a+b)n\big)$. The next proposition shows that, up to a linear transformation, the distribution of $X$ at a stopping time has a property similar to identity (6.1).

**Proposition 4.** *If $U_1$ is the first time when all the initial items 3 have left the queue, the initial items 3 in the bin have been served and the state of the bin is $(3,1)$. There exist $\mathscr{F}_{U_1}$-measurable random variables $A_n$ and $B_n$ such that*

$$\mathbb{P}_{\mu_n}\big(B(U_1)\big) = \big((3,1), L(U_1) \in dx\big) = \mathbb{E}_{\mu_n}\big(R_{A_n,B_n,0}(dx)\big),$$

*holds and there is a random matrix $M$ such that the convergence*

$$\lim_{n\to+\infty} \frac{1}{n}(A_n, B_n) = M \cdot (a,b) \tag{6.2}$$

*is true almost surely and in $L_1$. The random matrix $M$ has two possible values with equal probability*

$$m_1 = \begin{pmatrix} 1 & \frac{1-p-q}{2(1-p)} \\ 0 & \lambda\Big(q + \frac{\lambda p(1-p-q)}{2(5-\lambda p)}\Big) \end{pmatrix}$$

*and*

$$m_2 = \begin{pmatrix} \frac{\lambda}{2}(1-p-q) & \frac{1-p-q}{2(1-p)} \\ \lambda^2(1-p)\Big(q + \frac{\lambda p(1-p-q)}{2(5-\lambda p)}\Big) & \lambda\Big(q + \frac{\lambda p(1-p-q)}{2(5-\lambda p)}\Big) \end{pmatrix}.$$

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

*Proof*. Using Skorohod's representation Theorem (See Ethier and Kurtz [11]), 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 $(3,1)$ at time 0. If a new item 1 arrives, the bin is full, and during that time the 3 in the bin is replaced by the initial items 3. So the state of the bin will come back to the state $(3,1)$. In this manner, a finite number of initial items 3 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 $(3,1)$, another item 2 may enter if the item 1 leaves. Hence, after this event the state of the bin will be $(3,2)$ 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 an

item 2 enters the bin is independent of 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 3 leaves first. In this case the first item 3 enter the bin and all the other $a_n - 1$ items 3 will follow it in the bin.

During that time, since $\lambda p < 2$, the items 1 are processed by the empty space in the bin. It is easily checked that the time $\tau_1$ to get rid of the initial items 3 is equivalent to $a_n \sim an$.

At time $\tau_1$ the head of the queue is the original string of items 2 and 3 followed by another strings 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 $\big(b + \lambda(q+r)a\big)n$ (Lemma 16 of the appendix of Dantzer [8]).

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 $(3, 2)$.

$$\begin{aligned}&\text{“}(3,1) \Rightarrow (\cdot,1) \Rightarrow (3,1) \Rightarrow (\cdot,1) \Rightarrow (3,1) \Rightarrow (\cdot,1) \Rightarrow (3,1)\\ &\qquad \Rightarrow (3,\cdot) \Rightarrow (3,2) \Rightarrow (3,\cdot) \Rightarrow (3,2)\text{”}\end{aligned}$$

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 strings of 2's and 3's. At the end of the queue the new items arriving during that time form a string (since the bin is full the items 1 are not served during this phase).

The time $\tau_2$ to serve all the items 2 arriving before the state of the bin reaches $(3, 2)$ is equivalent to the quantity

$$\big(b + \lambda(q+r)a\big)\frac{q}{(q+r)}n.$$

At time $\tau_1 + \tau_2$ there is a string of 3's at the head of the queue of length equivalent to

$$\big(b + \lambda(q+r)a\big)\frac{r}{2(q+r)}n, \tag{6.3}$$

followed by an i.i.d. string with distribution $F(du)$ whose length is equivalent to the quantity

$$\lambda\big(b + \lambda(q+r)a\big)\frac{q}{q+r}n.$$

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 $(1,1,1,1,1)$. From that time all the 1's will be served at rate 5. The time $\tau_3$ it takes to empty the queue of the items 1 and to have exactly a 3 and a 1 in the bin is equivalent to

$$\frac{\lambda p\big(b+\lambda(q+r)a\big)r}{2(q+r)(5-\lambda p)}n.$$

At time $\tau_1+\tau_2+\tau_3$ there is a string of 3's of length whose length is equivalent to (6.3), followed by a string of 2's and 3's of length equivalent to

$$\begin{aligned}&\lambda\big(b+\lambda(q+r)a\big)\frac{q}{q+r}n(q+r)+\lambda(q+r)\frac{\lambda p\big(b+\lambda(q+r)a\big)r}{2(q+r)(5-\lambda p)}n\\&\quad=\lambda\big(b+\lambda(q+r)a\big)\left(q+\frac{\lambda rp}{2(5-\lambda p)}\right)\end{aligned}$$

For this case the distribution of $L(U)$ is given by $\mathbb{E}_{\mu_n}\big(R_{A_n,B_n,0}(dx)\big)$ and $(A_n,B_n)$ satisfies the relation (6.2) 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 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 5.** *With the same notations as in Proposition* 4, *there exists a constant* $K_0$ *such that*

$$\limsup_{n\to+\infty}\mathbb{E}\left(\frac{U_1}{n}\right)<K_0\mathbb{E}(a+b). \tag{6.4}$$

*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 6.** *If $(M_n)$ is an i.i.d. sequence of random matrices with the same distribution as $M$ in Proposition 4 and $P_n = M_n \cdot M_{n-1} \dots M_1$, there exist $\alpha, \beta > 0$ and a function $\eta$ on $\mathbb{R}_+$ such that for any $n \in \mathbb{N}$ and $x \in \mathbb{R}_+^2$,*

$$\mathbb{E}(\langle(\alpha,\beta), P_n \cdot x\rangle) = \eta(\lambda)^n \langle(\alpha,\beta), x\rangle) \tag{6.5}$$

*where $\langle \cdot, \cdot \rangle$ is the usual scalar product in $\mathbb{R}^2$. If*

$$\lambda^* = \frac{1}{4pq}(5 - 3p + 5q - \sqrt{(25 - 30p + 50q + 9p^2 - 110pq + 25q^2)} \tag{6.6}$$

*then $\eta(\lambda) < 1$ if $\lambda < \lambda^*$ and $\eta(\lambda) > 1$ if $\lambda^* < \lambda < 5/p$.*

*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 $(\alpha, \beta)$ is the corresponding right eigenvector; $\alpha$ and $\beta$ can be chosen strictly positive.

Consequently, we get

$$\mathbb{E}(\langle(\alpha,\beta), P_n \cdot x\rangle) = \langle(\alpha,\beta), \mathbb{E}(P_n) \cdot x\rangle = \langle(\alpha,\beta), \mathbb{E}(M_1)^n \cdot x\rangle = \eta(\lambda)^n \langle(\alpha,\beta), x\rangle).$$

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

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

Since the components of $\mathbb{E}(M_1)$ are increasing with respect to $\lambda$ if $\lambda p < 5$, 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 (6.6) is non-negative if $p + q \leq 1$ and that $\lambda^* p < 5$.) The proposition is proved. □

**Corollary 1.** *With the notations of Proposition 6, if $\lambda < \lambda^*$ for any $\gamma > \eta(\lambda)$ and $x \in \mathbb{R}_+^2$ the sequence $(\gamma^{-n} P_n.x)$ converges almost surely and in $L_1(\mathbb{P})$ to $(0,0)$.*

*Proof.* Using identity (6.5) for $n = 1$, it is easily seen that

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

is a martingale. The sequence $(Z_n)$ being non-negative converges almost surely to some finite limit $Z_\infty$. Since $\alpha$ and $\beta$ are positive and all the coefficients of $P_n$ are non-negative, we deduce that the sequence $(\gamma^{-n}P_n.x)$ 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 $\mathbb{E}(M_1)$ belong to the interval $[0, 1[$. □

## 7. Ergodicity

**Theorem 2.** *When the arrival rate of the items is $\lambda$, the distribution of their sizes is given by*

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

*and the size of the bin is* 5. *If*

$$\lambda_{\mathrm{FF}} := \min\left\{\frac{1}{4pq}\left(5 - 3p + 5q - \sqrt{(25 - 30p + 50q + 9p^2 - 110pq + 25q^2)}\right)\right\} \tag{7.1}$$

*then the Markov process $\big(X(t)\big)$ describing the First Fit algorithm is ergodic when $\lambda < \lambda_{\mathrm{FF}}$.*

*Proof.* If $\lambda p > 2$, Proposition 1 shows that the condition $\lambda\mathbb{E}(S_1) < 5$, i.e., $\lambda(3 - 2p - q) < 5$ is sufficient for the ergodicity of $\big(X(t)\big)$. One can check that in that case

$$\frac{5}{3 - 2p - q} < \frac{1}{4pq}\left(5 - 3p + 5q - \sqrt{(25 - 30p + 50q + 9p^2 - 110pq + 25q^2)}\right).$$

We assume that condition (7.1) and $\lambda p < 2$ are satisfied. According to Theorem 1, to prove the ergodicity it is sufficient to show that there exists a stopping time $V$ such that for any sequence $(x_n) = (b_n, l_n)$ of $\mathscr{S}_\infty$ with $\|x_n\| = n$ the following inequalities hold

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

Here $K > 1$ 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 Propositions 2 and 3, if $X(0) = x_n$ there exists a stopping time $U_0$ such that

- the distribution of $L(U_0)$ is given by $\mathbb{E}\big(R_{a_n, b_n, 0}(dx)\big)$, where $a_n$ and $b_n$ are some random variables and $B(U_0) = (3, 1)$;
- the following relations hold

$$\limsup_{n \to +\infty} \frac{\mathbb{E}_{x_n}(U_0)}{n} \leq K, \tag{7.2}$$

$$\limsup_{n \to +\infty} \mathbb{E}_{x_n}\left(\frac{a_n + b_n}{n}\right)^2 \leq \mathbb{E}_{x_n}\left(\frac{\|X(U_0)\|}{n}\right)^2 \leq K. \tag{7.3}$$

According to inequality (7.3), 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 (7.3) shows that the sequence $(a_n/n, b_n/n)$ is uniformly integrable; consequently, it converges in $L_1$. In particular we have

$$\limsup_{n \to +\infty} \mathbb{E}_{x_n}\left(\frac{a_n + b_n}{n}\right) = \mathbb{E}(a + b) \leq K, \tag{7.4}$$

which is a consequence of the relation (7.3). Using again the Skhorohod representation theorem (see Ethier and Kurtz [11]), 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 $\{\|X(U_0)\| \leq \|X(0)\|/5\}$ one sets $V = U_0$, so that

$$\mathbb{E}_{X(0)}\left(\frac{\|X(V)\|}{\|X(0)\|} \mathbb{1}_{\{\|X(U_0)\| \leq \|X(0)\|/5\}}\right) \leq \frac{1}{5}. \tag{7.5}$$

We have to determine $V$ on the event $\{\|X(U_0)\| > \|X(0)\|/5\}$. Proposition 4 shows that 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

$$\mathbb{P}_{\mu_n}\big(B(U_1)\big) = (3, 1), L(U_1) \in dx) = \mathbb{E}_{\mu_n}\big(R_{A_{n,1}, B_{n,1}, 0}(dx)\big),$$

where $\mu_n$ 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

$$\mathbb{P}_{X(U_1)}\big(B(U_1 \circ \theta_{U_1}) = (3,1), L(U_1 \circ \theta_{U_1}) \in dx\big) = \mathbb{E}_{X(U_1)}\big(R_{A_{n,2},B_{n,2},0}(dx)\big) \quad (7.6)$$

and

$$\lim_{n\to+\infty} \frac{1}{n}(A_{n,2}B_{n,2}) = M_2.M_1(a,b)$$

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 (7.6) gives

$$\mathbb{P}_{\mu_n}\big(B(U_2) = (3,1), L(U_2) \in dx\big) = \mathbb{E}_{\mu_n}\big(R_{A_{n,2},B_{n,2},0}(dx)\big).$$

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

$$\mathbb{P}_{\mu_n}\big(B(U_p) = (3,1), L(U_p) \in dx\big) = \mathbb{E}_{\mu_n}\big(R_{A_{n,p},B_{n,p},0}(dx)\big)$$

and

$$\lim_{n\to+\infty} \frac{1}{n}(A_{n,p}, B_{n,p}) = M_p M_{p-1} \ldots M_2 M_1(a,b)$$

holds almost surely and in $L_1$. According to Proposition 6, we have

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

Since $\|X(U_p)\| = 2 + A_{n,p} + B_{n,p}$ it follows from (7.4) that

$$\begin{aligned}\limsup_{n\to+\infty} \frac{1}{n}\mathbb{E}_{\mu_n}\big(\|X(U_p)\|\big) &\leq \frac{1}{\alpha\wedge\beta}\lim_{n\to+\infty}\frac{1}{n}\big\langle(\alpha,\beta), \big(\mathbb{E}_{\mu_n}(A_{n,p}), \mathbb{E}_{\mu_n}(B_{n,p})\big)\big\rangle \\ &= \gamma^p \frac{\alpha\mathbb{E}(a) + \beta\mathbb{E}(b)}{\alpha\wedge\beta} \\ &\leq \gamma^p \frac{\alpha\vee\beta}{\alpha\wedge\beta}\mathbb{E}(a+b) \leq \gamma^p \frac{\alpha\vee\beta}{\alpha\wedge\beta}K. \end{aligned} \quad (7.7)$$

Since the condition $\lambda < \lambda_{\mathrm{FF}}$ implies that $\gamma < 1$, we choose $p \in \mathbb{N}$ such that $\gamma^p < \frac{1}{5}\frac{\alpha\wedge\beta}{(\alpha\vee\beta)K}$. On the event $\{\|X(U_0)\| > \|X(0)\|/5\}$, the variable $V$ is defined as $U_0 + U_p \circ \theta_{U_0}$.

For $n \in \mathbb{N}$, by the strong Markov property, we have

$$\begin{aligned}\mathbb{E}_{x_n}(\|X(V)\|\mathbb{1}_{\{\|X(U_0)\|>\|x_n\|/5\}}) &= \mathbb{E}_{x_n}\big(\mathbb{E}_{\mu_n}(\|X(U_p)\|)\mathbb{1}_{\{\|X(U_0)\|>\|x_n\|/5\}}\big)\\ &= \mathbb{E}_{x_n}\big(\mathbb{E}_{\mu_n}(A_{n,p}+B_{n,p}+2)\mathbb{1}_{\{\|X(U_0)\|>\|x_n\|/5\}}\big).\end{aligned}$$

We can assume that $\mathbb{P}(a+b=1/5)=0$. Otherwise we replace the constant $1/5$ by some real $r$ less than $1/5$ such that $\mathbb{P}(a+b=r)=0$. We have

$$\begin{aligned}&\left|\mathbb{E}_{x_n}\left(\mathbb{E}_{\mu_n}\left(\frac{\|X(U_p)\|}{n}\right)(\mathbb{1}_{\{\|X(U_0)\|>\|x_n\|/5\}}-\mathbb{1}_{\{a+b>1/5\}})\right)\right|\\ &\quad\le C_0\mathbb{E}_{x_n}|\mathbb{1}_{\{\|X(U_0)\|>\|x_n\|/5\}}-\mathbb{1}_{\{a+b>1/5\}}|+2\mathbb{E}_{x_n}\left(\frac{\|X(U_p)\|}{n}\mathbb{1}_{\{\|X(U_p)/n\ge C_0\}}\right).\end{aligned}$$

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 $\mathbb{P}(a+b=1/5)=0$. Hence it is enough to consider the quantity $\mathbb{E}_{x_n}\left(\mathbb{E}_{\mu_n}\left(\frac{\|X(U_p)\|}{n}\right)(\mathbb{1}_{\{a+b>1/5\}})\right)$. Relation (7.7) implies that

$$\limsup_{n\to+\infty}\mathbb{E}_{x_n}\left(\frac{\|X(V)\|}{n}\mathbb{1}_{\{\|X(U_0)\|>\|x_n\|/5\}}\right)\le\gamma^p\frac{\alpha\vee\beta}{\alpha\wedge\beta}K\le\frac{1}{5}. \tag{7.8}$$

Inequalities (7.2) and (6.4) show that there exists some constant $K$ such that

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

and relations (7.5) and (7.8) give

$$\limsup_{n\to+\infty}\mathbb{E}_{x_n}\left(\frac{\|X(V)\|}{\|X(0)\|}\right)\le\frac{2}{5}.$$

The proof is completed. □

## 8. Transience

**Theorem 3.** *When the arrival rate of the items is $\lambda$, the distribution of their sizes is given by*

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

*and the size of the bin is* 5, *the Markov process* $\big(X(t)\big)$ *describing the First Fit algorithm is transient when* $\lambda > \lambda_{FF}$, *where* $\lambda_{FF}$ *is defined by equation* (7.1).

*Proof*. We assume that the initial distribution of $\big(L(t)\big)$ is given by $R_{a,b,0}$ and the initial state of the bin is $(3,1)$. With the notation of Proposition 4, $U_1$ is the first time when all the initial items 3 have left the queue, the items 3 in the bin have been served and the state of the bin is $(3,1)$. As in the proof of Proposition 2, 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 3 present at time $U_p$ have left the queue and the state of the bin is $(3,1)$. Clearly the sequence $\big(L(U_p)\big)$ is an homogeneous irreducible Markov chain on $\mathscr{T}^{(\mathbb{N})}$.

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

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

Almost surely $U_1$ is a finite stopping time. Proposition 4 and 6 show that there exist constants $\alpha$, $\beta$ such that

$$\lim_{(a+b)\to+\infty} \frac{\alpha A_1 + \beta B_1}{\alpha a + \beta b} = \gamma > 1 \tag{8.1}$$

almost surely.

We assume that the Markov process $\big(X(t)\big)$ is recurrent. In particular it visits the state $y_0 = \big(\emptyset, (3,1)\big)$ infinitely often, i.e., with probability 1 the queue will be empty and the state of the bin will be $(3,1)$. The first time the process $\big(X(t)\big)$ visits the state $y_0$ is necessarily at one of the moments $U_p$, $p \geq 1$. Consequently, the Markov chain $\big(L(U_p)\big)$ visits the state $y_0$ with probability 1. We now define a Lyapounov function on the state space of $\big(L(U_p)\big)$ by

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

if $l = (l_i)$, $p_1 = \inf\{k - 1/l_k \neq 3\}$. With the notations defined above, we have $f\big(L(U_1)\big) = \log(1 + \alpha A_{a,b} + \beta B_{a,b})$. Thus

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

According to Proposition 4, 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

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

for $x > 0$, convergence (8.1) and Lebesgue's theorem show that

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

Hence there exists some constant $K_0$ such that if $(a+b) \ge K_0$, then

$$\mathbb{E}_{R_{a,b}}\big(f\big(L(U_1)\big) - f\big(L(U_0)\big)\big) \ge (\log\gamma)/2. \tag{8.2}$$

In the same way, we have

$$\begin{aligned}\mathbb{E}_{R_{a,b,0}}&\big(\big|f\big(L(U_1)\big) - f\big(L(U_0)\big)\big|^2\big)\\ &\le 11\log(\alpha\vee\beta)^2 + 5\mathbb{E}_{R_{a,b,0}}\left(\log^2\left(\frac{1+A_{a,b}+B_{a,b}}{1+\alpha a+\beta b}\right)\right).\end{aligned}$$

The elementary inequality

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

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

$$\sup_{a,b;\,a+b>k} \mathbb{E}_{R_{a,b,0}}\big(\big|f\big(L(U_1)\big) - f\big(L(U_0)\big)\big|^2\big) < +\infty. \tag{8.3}$$

A theorem by Lamperti [18] (see Fayolle et al. [12] or Meyn and Tweedie [21]) states that if the relations (8.3) and (8.2) are satisfied, then the Markov chain $\big(L(U_p)\big)$ is transient. In particular this implies that there exists an initial state such that the chain will never visit the state $y_0$ with positive probability. This contradicts our assumption on the recurrence of $\big(X(t)\big)$. The theorem is proved. □

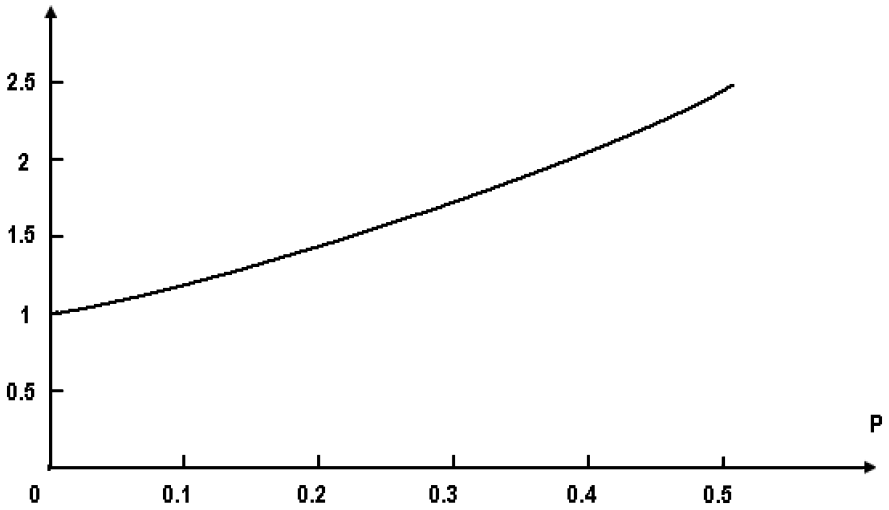
**The case of symmetrical distributions.** The distribution F is symmetrical if

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

for $p \in [0, 1/2[$. Since the expected value of size of the items is $1/2$ for all these distributions, the value $\lambda_{\mathrm{FF}}$ of the corresponding critical $\lambda$ cannot exceed $5/2$.

According to Theorem 2 the critical value of $\lambda$ for the First Fit algorithm is given by

$$\lambda_{\mathrm{FF}} = \frac{1}{4}\,\frac{-10 + 13p + \sqrt{100 - 340p + 329p^2}}{p(-1+2p)}.$$




The effective bandwidth of First Fit policie for symmetrical distribution on $\{1, 2, 3\}$

## 9. Conclusion

We have shown in this paper that in the situation $\lambda p > 2$ and $\lambda \mathbb{E}(S_1) < 5$ our system is stable. After that several conditions were established in the case $\lambda p < 2$ and in connection with the concept of "smooth random state". Then all these results were used to derive the ergodicity and transience conditions for the Markov process.

Finally, we have presented an example of symmetrical distributions.

## References

[1] M. Bramson, Instability of FIFO queueing networks. *Ann. Appl. Probab.* **4** (1994), 414–431; correction *ibid.* **4** (1994), 952. Zbl 0804.60079 MR 1272733; Zbl 0812.60080

[2] M. Bramson, Instability of FIFO queueing networks with quick service times. *Ann. Appl. Probab.* **4** (1994), 693–718. Zbl 0813.60087 MR 1284981

[3] H. Chen and A. Mandelbaum, Discrete flow networks: bottleneck analysis and fluid approximations. *Math. Oper. Res.* **16** (1991), 408–446. Zbl 0735.60095 MR 1106809

[4] E. G. Coffman, A. Feldman, N. Kahale and B. Poonen, Computing call admission capacities in linear networks. *Probab. Engrg. Inform. Sci.* **13** (1999), 387–406. Zbl 0969.90025 MR 1715420

[5] E. G. Coffman and A. L. Stolyar, Bandwidth packing. *Algorithmica* **29** (2001), 70–88. Zbl 0967.68167 MR 1887299

[6] J. G. Dai, On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Probab.* **5** (1995), 49–77. Zbl 0822.60083 MR 1325041

[7] J. G. Dai, A fluid limit model criterion for instability of multiclass queueing networks. *Ann. Appl. Probab.* **6** (1996), 751–757. Zbl 0860.60075 MR 1410113

[8] J.-F. Dantzer, M. Haddani and P. Robert, On the stability of a bandwidth packing algorithm. *Probab. Engrg. Inform. Sci.* **14** (2000), 57–79. Zbl 0969.90026 MR 1738273

[9] V. Dumas, A multiclass network with non-linear, non-convex, non-monotonic stability conditions. *Queueing Systems Theory Appl.* **25** (1997), 1–43. Zbl 0894.60082 MR 1458584

[10] P. Dupuis and R. J. Williams, Lyapunov functions for semimartingale reflecting Brownian motions. *Ann. Probab.* **22** (1994), 680–702. Zbl 0808.60068 MR 1288127

[11] S. N. Ethier and T. G. Kurtz, *Markov processes: Characterization and convergence*. John Wiley & Sons, New York 1986. Zbl 0592.60049 MR 0838085

[12] G. Fayolle, V. A. Malyshev and M. V. Menshikov, *Topics in the constructive theory of countable Markov chains*. Cambridge University Press, Cambridge 1995. Zbl 0823.60053 MR 1331145

[13] Y. P. Filonov, A criterion for the ergodicity of discrete homogeneous Markov chains. *Ukrain. Mat. Zh.* **41** (1989), 1421–1422; English transl. *Ukrain. Math. J.* **41** (1989), 1223–1225. Zbl 0709.60073 MR 1034693

[14] R. Z. Has'minskiĭ, *Stochastic stability of differential equations*. Sijthoff & Noordhoff, Alphen aan den Rijn 1980, Zbl 0441.60060 MR 0600653

[15] M. W. Hirsch and S. Smale, *Differential equations, dynamical systems, and linear algebra*. Pure Appl. Math. 60, Academic Press New York 1974. Zbl 0309.34001 MR 0486784

[16] C. Kipnis and P. Robert, A dynamic storage process. *Stochastic Process. Appl.* **34** (1990), 155–169. Zbl 0714.90038 MR 1039567

[17] P. R. Kumar and T. I. Seidman, Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Trans. Automat. Control* **35** (1990), 289–298. Zbl 0715.90062 MR 1044023

[18] J. Lamperti, Criteria for the recurrence or transience of stochastic process. I. *J. Math. Anal. Appl.* **1** (1960), 314–330. Zbl 0099.12901 MR 0126872

[19] S. H. Lu and P. R. Kumar, Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Automat. Control* **36** (1991), 1406–1416.

[20] S. P. Meyn, Transience of multiclass queueing networks via fluid limit models. *Ann. Appl. Probab.* **5** (1995), 946–957. Zbl 0865.60079 MR 1384361

[21] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Springer-Verlag, Berlin 1993. Zbl 0099.12901 MR 1287609

[22] A. A. Pukhal'skiĭ and A. N. Rybko, Nonergodicity of queueing networks when their fluid models are unstable. *Problemy Peredachi Informatsii* **36** (2000), 26–46; English transl. *Problems Inform. Transmission* **36** (2000), 23–41. MR 1746007

[23] P. Robert, Smooth initial distributions and fluid limits for multi-class queueing systems. Unpublished manuscript, 2002.

[24] A. N. Rybko and A. L. Stolyar, On the ergodicity of random processes that describe the functioning of open queueing networks. *Problemy Peredachi Informatsii* **28** (1992), 3–26; English transl. *Problems Inform. Transmission* **28** (1992), 199–220. Zbl 0768.60089 MR 1189331

Received January 31, 2007; revised October 28, 2007

F. Belarbi, Laboratoire de Mathématique, B.P. 89, Université Djillali Liabes, Sidi Bel Abbès 22000, Algeria

E-mail: faiza_belarbi@yahoo.fr

A. A. Bouchentouf, Laboratoire de Mathématique, B.P. 89, Université Djillali Liabes, Sidi Bel Abbès 22000, Algeria

E-mail: bouchentouf_amina@yahoo.fr