HAL
open science

On the Stability of a Bandwidth Packing Algorithm

Jean-François Dantzer, Mostafa Haddani, Philippe Robert

To cite this version:

Jean-François Dantzer, Mostafa Haddani, Philippe Robert. On the Stability of a Bandwidth Packing Algorithm. [Research Report] RR-3605, INRIA. 1999. inria-00073075

HAL Id: inria-00073075

https://inria.hal.science/inria-00073075

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

On the stability of a bandwidth packing algorithm

Jean-François Dantzer, Mostafa Haddani, and Philippe Robert

N° 3605

THÈME 2

Rapport de recherche

ISSN 0249-6399

On the stability of a bandwidth packing algorithm

Jean-François Dantzer, Mostafa Haddani, and Philippe Robert

Thème 2 — Génie logiciel
et calcul symbolique
Projet ALGO

Rapport de recherche n°3605 — — 21 pages

Abstract: The stability properties of the bandwidth allocation algorithm First Fit are analyzed for some distributions on the sizes of the requests. Fluid limits are used to get the ergodicity results. When there are two possible sizes, the description of the transient behavior involves a finite Markov chain on the exit states of a transient Markov chain on a countable state space. The explicit expression of this exit matrix is given.

Key-words: Bin Packing Algorithms. Ergodicity. Transient Behavior. Fluid Limits. Multi-class Queueing Systems. Bandwidth Allocation.

(Résumé : tsvp)

This work has been partially supported by the ESPRIT Project 25046 - CATSERVER

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
Téléphone : 01 39 63 55 11 - International : +33 1 39 63 55 11
Télécopie : (33) 01 39 63 53 30 - International : +33 1 39 63 53 30

Sur la stabilité d'un algorithme d'allocation de bande passante

Résumé : Les propriétés stabilité de l'algorithme d'allocation de bande passante First Fit sont analysées pour certaines distributions sur la taille des requêtes. Les limites fluides sont utilisées pour obtenir les conditions d'ergodicité. Quand il n'y a que deux tailles de requêtes possibles, la description du cas instable fait intervenir une chaîne de Markov à espace d'états fini sur les états de sortie d'une chaîne de Markov transiente. L'expression explicite de la matrice de sortie est donnée.

Mots-clé : Algorithmes de bin packing. Ergodicité. Transience. Limites fluides.

1. Introduction

The model we consider here is a simplified description of a bandwidth allocation scheme in a network: a stream of messages arrive at a service facility working at rate C. The messages are of different nature, it can be video, voice or data, each of them requires a portion of the bandwidth C to be transmitted. A video stream will of course require a significant portion of the capacity of the server, at the opposite data messages do not usually need high transmission rates. 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, if there is enough room, messages in the queue can be transmitted. At this point several algorithms are available:

1. the FIFO policy. A message is transmitted if it is the oldest in the queue.
2. First Fit (FF) algorithm. A message in the queue is transmitted if it fits and none of the other messages before it in the queue is transmitted.
3. Best Fit (BF) algorithm. The largest message that fits is transmitted.

The list is of course not exhaustive. The FF and BF algorithms have the advantage of reducing the wasted bandwidth, none of the messages in the queue can be transmitted without violating the capacity condition; this is not the case for the FIFO algorithm. In this paper, we study the problem of stability of the First Fit algorithm: for a given distribution on the sizes of the items, what is the maximal rate under which the server can accommodate all the requests without an explosion of the queue ?

For convenience we shall use the bin packing terminology: the server is a box of size C, messages are items and the bandwidth required by a message is the size of the item. Hence we have a stream of items which require a service delivered by a bin. 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). Some general results for which a simple stability condition holds are established. The case where items have two possible sizes is analyzed and the stability condition established. This simple case already exhibits the difficulty of working in the state space mentioned above; a simple result (Proposition 7) reduces the difficulty.

We also analyze the transient behavior of the system, i.e. when the system is not stable. This is an important topic, since a generalization to more general distributions requires such an analysis. The main result concerning the transient behavior is the following: the process is basically a two dimensional process exploding linearly, except at instants where, during a short time, the complexity of the process shows up. This short period of time determines the two dimensional process that will drive the process until the next change of direction. The paper finishes with the explicit expression of the transition from a direction in the plane to next one. The interesting behavior of the transient case is the main point of this paper. Usually in most of the known examples a saturated node has a regular behavior for the other nodes, hence it is replaced by an equivalent node with different parameters. In the model considered here when a node of a network is saturated, it may causes oscillations to the rest of the network; it cannot be reduced to another equivalent node.

Assumptions. The items arrive according to a Poisson process with parameter λ. Their sizes S_i, $i \in \mathbb{N}$ form an i.i.d. sequence with a common distribution μ on $[0, C]$, an item of size s will also be called an item s. It is assumed that the range of the S_i's is some finite subset T of $]0, C]$. Once in the bin, each item is served for an exponentially distributed amount of time with parameter 1. The capacity of service of this system is C and the average load $\lambda E(S_1)$. Throughout this paper, $\mathcal{N}_\xi$ denotes a Poisson process with parameter ξ and $\mathcal{N}_\xi(]0, t])$ is the number of points of this process between 0 and t, all the Poisson processes considered are independent.

The markovian representation. The description of this system as a Markov process involves two components, $L = (l_j)$ and $B = (b_i)$ elements of the set $T^{(\mathbb{N})}$ of finite vectors with components in the set of T of the possible sizes of the items. The vector L describes the queue and B the state of the bin, they satisfy the following conditions: the components of B fit in the bin,

$$\sum b_i \leq C, \tag{1}$$

and none of the items in the queue fit in the bin, i.e. for any index j

$$l_j + \sum_i b_i > C, \tag{2}$$

$l_1(t)$ is the size of the older item still waiting at time t; $l_j(t)$ is the size of the j-th oldest. The state space $\mathcal{S}$ of the Markov process is the set of the $x = (L, B) \in T^{(\mathbb{N})} \times T^{(\mathbb{N})}$ satisfying (1) and (2. The norm $|x|$ is the sum of the components of the vector x, if $(L, B) \in \mathcal{S}$, $|(L, B)| = \sum b_i + \sum l_j$. When the component B of (L, B) in $\mathcal{S}$ is 0 (the empty vector), the queue is empty and $(0, 0)$ corresponds to an empty system. Since there are only a finite number of possible sizes, the dimension of the B-component is upper bounded. This is not the case for the L-component. Its dimension (the number of items in the queue) can be arbitrarily large. Our state space is a countable but embedded in an infinite dimensional vector space. Since the First Fit algorithm scans all the queue at each event, the infinite dimension really plays a role (this is not the case for example for the FIFO algorithm). This complicates significantly the analysis of the stability since the number of ways to go to infinity is an order of magnitude greater compared to the finite dimensional case.

The Markov process $(X(t)) = ((L(t), B(t))$ describing the first fit algorithm has 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.
2. *Departure.* At rate 1, each item in the bin leaves the system. In the case of a departure, the first element of the queue that fits is added, 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.

Definition 1. *For $i \in T$ the variable $L_i(t)$ [resp. $B_i(t)$] denotes the number of items of size i in the queue [resp. in the bin]. An additional index $x \in \mathcal{S}$, $L_{x,i}(t)$ $B_{x,i}(t)$ indicates the initial state x.*

Related models. A similar problem has been analyzed by Kipnis and Robert [14] with the FIFO algorithm. 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 size of the items in the queue are unlikely to be 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 throughput 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 [4] analyzed the stability of the algorithms First Fit and Best Fit when the services are constant equal to 1. They prove that the natural condition $\lambda E(S) < C$ is sufficient for the stability in the case of a symmetrical distribution of the sizes; in Coffman *et al.* [3] the sufficiency for stability of the condition $\lambda E(S) < C$ is considered in a more complex communication network.

Markov processes on strings occur also naturally in the multi-class queueing networks. Rybko and Stolyar [18], Bramson [2, 1] analyzed such Markov processes. In their work, due to clever arguments, the string structure is not really taken into account in their analysis. Dumas [9] presented an analysis of the fluid equations for the string structure of Bramson's networks.

In the same vein, Malyshev *et al.* [13, 16] 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.

Further properties of the systems considered here are presented in [8].

Before analyzing in detail the ergodicity of $(X(t))$, we recall the main results and ideas concerning fluid limits.

2. Fluid limits

In this section $(X(t))$ is an irreducible Markov process on some countable space $\mathcal{S}$ and f a positive function on $\mathcal{S}$. The following classical criterion is a generalization of Foster's result due to Filonov (see Filonov [12] or [11]).

Proposition 2. *If there exists an integrable stopping time τ, a finite set $F \subset \mathcal{S}$, $\gamma > 0$ and a positive function f such*

a) $E_x(f(X(1))) < +\infty$ *for all* $x \in \mathcal{S}$;

b) $E_x(f(X(\tau)) - f(x)) \leq -\gamma E_x(\tau)$ *if* $x \notin F$;

and if $\{x/f(x) < K\}$ *is finite for all* $K > 0$*, then* $(X(t))$ *is an ergodic Markov process.*

The origin of this criterion is the Lyapounov stability test of ordinary differential equations adapted to the stochastic case. The function f is usually called a Lyapounov function. The main problem with this kind of criterion is (of course !) to find an appropriate function f. If the transitions of the Markov process are not sufficiently homogeneous, finding a function that satisfies b) can be cumbersome, even in the "obvious" cases.

To overcome this problem, the rescaled process $(\bar{X}_f(t))$ is introduced

$$\bar{X}_x^f(t) = \frac{f(X(tf(x))}{f(x)}, \tag{3}$$

if $X(0) = x$; notice that $\bar{X}_x^f(0) = 1$. The time variable and the space variable are scaled by a factor $f(x)$. The main result concerning the ergodicity is the following theorem. This is a slight generalization of a theorem due to Rybko and Stolyar [18].

Theorem 3. *If there exist an integrable stopping time* τ*, constants* A*,* $\varepsilon > 0$ *and a positive function* f *on* $\mathcal{S}$ *such that*

$$\limsup_{f(x)\to+\infty} E\left(\frac{f(X(\tau))}{f(x)}\right) \leq 1 - \varepsilon, \tag{4}$$

$$\limsup_{f(x)\to+\infty} \frac{E_x(\tau)}{f(x)} \leq A, \tag{5}$$

if the subset $\{x \in \mathcal{S}/f(x) \leq K\}$ *is finite for any* $K \in \mathbb{N}$*, the Markov process* $(X(t))$ *is ergodic.*

Apparently the situation did not change with this theorem, there is still a function f to find out. But in practice, if $\mathcal{S}$ is embedded in a vector space, $f(x) = |x|$, the norm of x, is a good candidate in most of the cases. The condition (4) requires that at some random time the norm of the process $(\bar{X}_x^f(t))$ is, in average, below its initial value. This suggests then the analysis of the process $(\bar{X}_x^f(t))$, when $|x| \to +\infty$. If one can prove that it converges almost surely to 0 after some time T, then up to an integrability argument, the theorem 3 can be applied.

Usually it is easier to analyze the behavior of the sample paths of $(\bar{X}_x^f(t))$ as $x \to +\infty$ rather than playing with the coefficients of the transition matrix to have condition b) of Proposition 2. These scaling ideas are difficult to trace back. Similar methods are used in the analysis of infinite particle systems but in a different context (see Comets [5] for example).

The discovery of some unexpected phenomena for the stability of queueing systems (Bramson [1, 2], Lu and Kumar [15], Rybko and Stolyar [18] among others) gave an impulse to the studies in this domain recently. Dai [6] set a framework to apply Rybko and Stolyar's scaling method to prove Harris ergodicity for queueing networks; for the transience, Dai [7] and Meyn [17] obtained partial counterparts to the ergodicity results.

3. A natural ergodicity condition

With our assumptions the system has a capacity of service C and the offered load of the flow of the items is $\lambda E(S)$. A natural condition for stability is thus $\lambda E(S) < C$. The following proposition shows that it is at least necessary for stability.

Proposition 4. *If the Markov process $(X(t))$ is ergodic then $\lambda E(S_1) \leq C$.*

Proof. The process can be viewed as a rectangle packing in an infinite horizontal strip of height C. The length of a rectangle is the time required by an item and the height is its size. The x-coordinate of the bottom left corner of a rectangle is the time at which the item has been accepted in the bin. Obviously, the rectangles do not overlap; if the process is ergodic, the area covered by the rectangles arrived between 0 and t must be less than $Ct + o(t)$ for t sufficiently large. The law of large numbers shows that this implies $\lambda E(S_1) \leq C$. □

With the same argument, it is easy to prove that the condition $\lambda E(S_1) < C$ is sufficient for ergodicity under the condition that the bin is full whenever the size of the queue is sufficiently large. The gap between the "optimal" condition $\lambda E(S_1) < C$ and the real one is due the waste of space in the bin. The reference [8] studies this problem.

Theorem 5. *If there exist an integrable stopping time τ and a random variable D such that*

1. *the items present at time 0 have been served at time τ ;*
2. *the bin is full during $\tau - D$ between 0 and τ and*

$$\lim_{|x| \to +\infty} \frac{E_x(D)}{|x|} = 0,$$

the Markov process $(X(t))$ is ergodic if $\lambda E(S_1) < C$.

Proof. With the rectangle packing analogy, we get the following inequality,

$$\sum x_i \sigma_i^0 + \sum_{i=1}^{\mathcal{N}_\lambda[0,\tau]} S_i \sigma_i \geq C\tau - DC,$$

where the x_i are the components of x, the σ's are the service times (with an upper index 0 for the initial items) and $\mathcal{N}_\lambda$ is the arrival process. Taking the expectation in the above inequality, we get with the help of Wald's formula

$$E_x(\tau) \leq \frac{|x| + E_x(D)C}{C - \lambda E(S_1)},$$

hence,

$$\limsup_{|x| \to +\infty} \frac{E_x(\tau)}{|x|} \leq \frac{1}{C - \lambda E(S_1)}. \tag{6}$$

If (y_i) is the vector of the sizes of the items present at time τ in the system and (σ_i') the vector of their residual services,

$$\sum \sigma_i' y_i \leq \sum x_i \sigma_i^0 + \sum_{i=1}^{\mathcal{N}_\lambda[0,\tau]} S_i \sigma_i - C(\tau - D),$$

taking the expected values, we get

$$E(|X(\tau)|) \leq |x| + (\lambda E(S_1) - C)E(\tau) + CE_x(D);$$

since the initial items are served we have $CE(\tau) \geq |x|$, hence

$$\frac{E(|X(\tau)|)}{|x|} \leq \frac{\lambda E(S_1)}{C} + C\frac{E_x(D)}{|x|},$$

$$\limsup_{|x|\to+\infty} \frac{E(|X(\tau)|)}{|x|} \leq \frac{\lambda E(S_1)}{C} < 1. \tag{7}$$

The inequalities (6) and (7) and the theorem 3 show the ergodicity of $(X(t))$. □

The following corollary is deduced directly from the inequalities (6) and (7), it gives upper bounds on the behavior the first time where the bin is not full.

Corollary 6. *Under the condition $\lambda E(S_1) < C$, if T_1 is the first time the bin is not completely used,*

$$T_1 = \inf\{t/|B(t)| < C\},$$

there exists $A > 0$ such that

$$E_x(T_1) \leq A|x|, \text{ and } E_x(|X(T_1)|) \leq |x|, \tag{8}$$

for any $x \in \mathcal{S}$.

We can now give a simplified form of Theorem 3. This result reduces the size of the set of initial states to consider in the analysis of ergodicity.

Proposition 7. *If $\mathcal{S}_0$ is the subset of elements of $\mathcal{S}$ with the bin not completely full, $\mathcal{S}_0 = \{x = (B, L) \in \mathcal{S}/|B| < C\}$, and if there exist constants $\varepsilon > 0$, $A \in \mathbb{R}$ and a stopping time τ such that*

$$\limsup_{|x|\to+\infty, x\in\mathcal{S}_0} E\left(\frac{|X(\tau)|}{|x|}\right) \leq 1 - \varepsilon, \tag{9}$$

$$\limsup_{|x|\to+\infty, x\in\mathcal{S}_0} \frac{E_x(\tau)}{|x|} \leq A, \tag{10}$$

under the condition $\lambda E(S_1) < C$ the Markov process $(X(t))$ is ergodic.

Proof. According to our hypothesis, for $\varepsilon_0 < \varepsilon$ and $A_0 > A$, there exists $K > 0$ such that

$$E_x\left(\frac{|X(\tau)|}{|x|}\right) < 1 - \varepsilon_0, \quad \text{and} \quad \frac{E_x(\tau)}{|x|} \leq A \tag{11}$$

if $|x| \geq K$ and $x \in \mathcal{S}_0$. The random variable T is defined as follows:

1. if $X(0) \in \mathcal{S}_0$, $T = \tau$;
2. if $X(0) \notin \mathcal{S}_0$ and $|X(T_1)| \leq (1-\varepsilon_0)K$, $T = T_1$;
3. $T = T_1 + \tau$, otherwise;

as before T_1 is the hitting time of $\mathcal{S}_0$. The variable T is clearly a stopping time. If $x \in \mathcal{S}$ is such that $|x| \geq K_1 = K/(1-\varepsilon_0)$, then $E_x(X(T)) \leq (1-\varepsilon_0)|x|$ when $x \in \mathcal{S}_0$. Otherwise when $x \notin \mathcal{S}_0$, the strong Markov property of $(X(t))$ gives

$$E_x(|X(T)|) = E_x(|X(T_1))|1_{\{|X(T_1)|<(1-\varepsilon_0)|x|\}}) + E_x(E_{X(T_1)}(|X(\tau)|)1_{\{|X(T_1)|\geq(1-\varepsilon_0)|x|\}}),$$

the inequality (11) shows that

$$E_x(|X(T)|) \leq E_x(|X(T_1))|1_{\{|X(T_1)|<(1-\varepsilon_0)|x|\}}) + (1-\varepsilon_0)E_x(|X(T_1)|1_{\{|X(T_1)|\geq(1-\varepsilon_0)|x|\}}),$$

hence

$$E_x(|X(T)|) \leq E_x(|X(T_1)|) - \varepsilon_0 E_x(|X(T_1)|1_{\{|X(T_1)|\geq(1-\varepsilon_0)|x|\}}).$$

Using (8), we get finally

$$E_x(|X(T)|) \leq |x|(1-\varepsilon_0(1-\varepsilon_0)),$$

whenever $|x| \geq K_1$. With similar arguments, it is easy to prove that there exist K_2, $A_2 > 0$ such that $E_x(T) \leq A_2|x|$ when $|x| \geq K_2$. Theorem 3 shows that the Markov process $(X(t))$ is ergodic. □

4. The case of distributions with two sizes : the ergodicity condition

The possible sizes of the items are 1 and $a \in \mathbb{N}$, the capacity C is an integer greater than a and the probability distribution μ is given by

$$\mu = p\delta_1 + (1-p)\delta_a,$$

where $p \in [0,1]$ and δ_x is the Dirac measure in x. The condition $\lambda E(S_1) < C$ is $\lambda(p + a(1-p)) < C$. The capacity C is decomposed as $C = ka + r$, with $0 \leq r < a$; at most k items of size a are served at the same time and r items of size 1 can be served simultaneously in any case.

Assumptions. To avoid messy details, we discard some singular cases; we assume that λp is not r or $C - \lfloor C/a \rfloor a$, and that $(C-\lambda p)/a$ and $C(1-p)/E(S_1)$ are not integers.

A priority mechanism. If at some moment, the bin contains a number of items 1 greater than $C-(a-1)$ so that an item a cannot be accommodated, the bin will be then filled by the items 1 if there are many of them. As long as there will be an item 1 in the queue, the first fit algorithm will serve only the items 1 (an item 1 is replaced by an item 1). In this case, the first fit algorithm is similar to a priority queue.

Notice that this phenomenon occurs with positive probability if the initial state has sufficiently many long strings of items 1 in the queue. If an item 1 is at the head of the queue when an item a leaves the bin, then a items 1 will fill the a empty places. The items 1 may occupy all the bin if this situation happens sufficiently often. However, if there is a long string of items a at the head of the queue, an item a leaving the bin is replaced by another item a.

From a heuristic point of view, the ergodicity condition can be easily guessed. If $\lambda p \geq r$, there will be no waste of space even when a long string of items a is served, the items 1 will saturate the empty space in the bin; the condition $\lambda E(S) < C$ is thus sufficient. Otherwise $\lambda p < r$, the items 1 will be transparent since they are served in the empty space left by the items a, hence the condition should be $\lambda(1-p) < k$.

Proposition 8 (Ergodicity condition). *If the distribution of the size of the items is given by $\mu = p\,\delta_1 + (1-p)\,\delta_a$, with $p \in [0,1]$, $a \in \mathbb{N}$ and the capacity of the bin is $C = ka + r$ with $0 \leq r < a$, under the condition*

$$\max(r - \lambda p, 0) + \lambda(p + (1-p)a) < C, \tag{12}$$

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

Proof. To apply Proposition 7, we take a sequence of initial states $(x_N) = (l_N, b_N)$ in the set $\mathcal{S}_0$ such that $|x_N|$ converges to infinity as $N \to +\infty$; we denote by $(X_N(t)) = (l_N(t), b_N(t))$ the process with initial state x_N. All we have to prove is the existence of a stopping time τ such that

$$\limsup_{N\to+\infty} \frac{E_{x_N}\left(|X_N(\tau)|\right)}{|x_N|} < 1-\varepsilon, \tag{13}$$

$$\limsup_{N\to+\infty} \frac{E_{x_N}\left(\tau\right)}{|x_N|} < C, \tag{14}$$

for some positive constants C and ε not depending of the sequence (x_N).

An element $((b_i),(l_j))$ of $\mathcal{S}_0$ satisfies $\sum b_i < C$ (by definition) and (l_j) is a string of a, $l_j = a$ for all j (since an item 1 would fit). Two cases occur, depending on the value of λp.

Case $\lambda p < r$. The condition (12) is $\lambda(1-p) < k$. Since there are always r places for the 1 in the bin, the number of 1 in the system $(L_1(t))$ is bounded by the number of clients in a stable $M/M/r$ queue with input rate λp, service rate 1 and C initial customers. The first time when this $M/M/r$ queue is empty is integrable, consequently the same property holds for the first time T_2 when the number of 1 in the system is 0. At time T_2, there will k items of size a in the bin; using the same method as in the proof of Proposition 7 (T_2 playing the role of T_1), it is easily seen that we assume that for all the x_N, there are k items of size a in the bin.

The items a arrive at rate $\lambda(1-p)$ and are served at rate $k > \lambda(1-p)$. If τ is the first time when all the items a initially in the queue are served, it satisfies

$$\lim_{N\to+\infty} \frac{E_{x_N}(\tau)}{|x_N|} = \frac{1}{k},$$

by the law of large numbers. Because of the stability of the $M/M/r$ mentioned above, the expected value of the total number of items 1 is $O(|x_N|)$, and the number of items a is bounded by $\mathcal{N}_\lambda[0,\tau]$, consequently

$$\lim_{N\to+\infty} \frac{E_{x_N}(X(\tau))}{|x_N|} = \frac{\lambda(1-p)}{k} < 1.$$

Case $\lambda p > r$. The number of items a in the bin for the initial state $x_N = (b_N, l_N) \in \mathcal{S}_0$ is denoted by k_N. At time 0 less than $C - k_N a$ items 1 are in the bin and l_N is a string of items a.

If $\lambda p < C - k_N a$ then within an integrable amount of time all the items 1 will have left the bin, hence k items a will occupy it. Without loss of generality we assume that the inequality $\lambda p > C - k_N a$ holds for all the initial states. The relation $\lambda p > C - k_N a$ implies that after a finite time, the items 1 will saturate the places left by the items a and according to the condition (12) it implies also that $\lambda(1-p) < k_N$. Consequently the bin is full up to time T when the initial string of items a has been served, clearly

$$\lim_{N\to+\infty} \frac{E_{x_N}(T)}{|x_N|} = \frac{1}{ak_N} \leq 1,$$

moreover

$$\lim_{N\to+\infty} \frac{E_{x_N}(|X(\tau)|)}{|x_N|} \leq \frac{\lambda(1-p)}{\bar{k}} < 1,$$

with $\bar{k} = \inf\{i/\lambda(1-p) < i\}$. Our proposition is proved. □

Notice the advantage of starting from a state in $\mathcal{S}_0$, i.e. the queue is a string of items a, it amounts to identify (x_N) with $(|x_N|)$ for the analysis of the ergodicity. The complexity of the numerous patterns of 1 and a is ruled out with this method.

5. The case of distributions with two sizes: the transient behavior

We describe here the way the process $(X(t))$ goes to infinity when it is transient. It is important to describe the transient behavior of such a process. Indeed, if for example there are three possible sizes for the items, 1, a and b with $1 < a < b$, the analysis of the stability in this case necessarily contains the study of the following case: a large string of b at the beginning of the queue and k_0 items b in the bin; if there are sufficiently many items a and 1, the rest of the bin is occupied by them. For a while there is a bin of size $C - k_0 b$ with an arrival stream of items 1 and a and this system may be transient. When the string of items b at the head of the queue is processed, to describe the future behavior of the system, it is necessary to know how many items 1 and a remain in the queue at that time and how

they are distributed. This is why a precise description of the transient behavior is a minimal requirement to get some insight on more general distributions.

Proposition 9. *If the distribution of the size of the items is given by* $\mu = p\,\delta_1 + (1-p)\,\delta_a$, $p \in [0,1]$ *and the capacity of the bin is* $C = ka + r$, *with* $0 \le r < a$, *under the condition*

$$\max(r - \lambda p, 0) + \lambda(p + a(1-p)) > C, \tag{15}$$

the Markov process $(X(t))$ *is transient.*

Proof. The proof is straightforward: if $\lambda p < r$, the condition (15) is $\lambda(1-p) > k$, the number of items a is exploding at least at rate $\lambda(1-p) - k$. Otherwise if $\lambda p > r$, the condition (15) is simply $\lambda E(S_1) > C$; The rectangle packing view of the algorithm gives, almost surely,

$$\lim_{t \to +\infty} \frac{|X(t)|}{t} \ge \lambda E(S_1) - C > 0,$$

the process $(X(t))$ is transient. □

More interesting is the way the shape of queue evolves. We describe now the paths to infinity in the transient case. Throughout this section the condition (15) holds.

Definition 10. *For* $t \in \mathbb{R}_+$, $\nu(t)$ *is the rank of the first item of size* 1 *in the queue,*

$$\nu(t) = \inf\{i / l_i(t) = 1\},$$

if $X(t) = ((l_i(t)), (b_i(t)))$, *and with the convention* $\inf \emptyset = +\infty$. *The variable* τ *is the first time when an item of size* 1 *is at the head of the queue and* η *is the first time when the queue is a string of items* a,

$$\tau = \inf\{t / \nu(t) = 1\} \quad \textit{and} \quad \eta = \inf\{t / \nu(t) = +\infty\}.$$

The variables τ_x, η_x *and* $\nu_x(t)$ *denote respectively the variables* τ, η *and* ν *associated with the initial state* $X(0) = x$.

5.1. **A saturated model.** At time $t = 0$, the state of the system is defined as follows:

1. the bin contains k_0 items of size a, $k_0 \le k$ and $C - k_0 a$ items of size 1;
2. the queue is an infinite vector $(L_N) = (l_i)$ such that
   (a) $l_i = a$ for $i = 1, \dots, N$;
   (b) $\{l_{i+N}, i \ge 1\}$ is an infinite string of i.i.d. random variables with distribution $\mu = p\delta_1 + (1-p)\delta_a$.

Since there is an infinite number of items 1 in the queue, the bin remains full forever. Moreover the process $(Z(t)) = (\nu(t), B_a(t))$ has the Markov property. As we shall see, the process $(Z(t))$ is transient in some cases; its exit states will give the key of the transient behavior of the process $(X(t))$. The following lemma gives the condition under which $(Z(t))$ is transient, or equivalently when the string of items a at the head of the queue grows or shrinks. This will be useful to classify the paths to infinity of the process $(X(t))$.

Lemma 11. *If $k_0 < C(1-p)/E(S_1)$, the string of items a grows: almost surely there exists some $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$ and $t \in \mathbb{R}_+$,*

$$B_{x_N,a}(t) = k_0,$$

$$\nu_{x_N}(t) > 0 \text{ and } \lim_{t\to+\infty} \frac{\nu_{x_N}(t)}{t} = \frac{C(1-p) - k_0 E(S_1)}{p}.$$

If $k_0 > C(1-p)/E(S_1)$, the string of items a shrinks: for $N \in \mathbb{N}$ sufficiently large, the variable τ_{x_N} is almost surely finite and

$$\lim_{N\to+\infty} \frac{\tau_{x_N}}{N} = \frac{p}{k_0 E(S_1) - C(1-p)}.$$

Proof. The proof is fairly simple: as long as the first item 1 has not hit the head of the queue, its rank decreases by 1 when an item a leaves the bin, i.e. at rate k_0. When an item 1 leaves the bin, the item 1 located at $\nu(t)$ enters the bin; the i.i.d. assumption shows that the next item 1 in the queue is located at $\nu(t)+G$ where G is random variable geometrically distributed with parameter $1-p$. The process $(\nu(t \wedge \tau_{x_N}))$ has the same distribution as

$$\left(n + \sum_{i=1}^{\kappa(t\wedge\tau_{x_N})} G_i - \mathcal{N}_{k_0}[0, t \wedge \tau_{x_N}] \right),$$

with $\kappa(t) = \mathcal{N}_{C-k_0 a}[0,t]$, (G_i) is an i.i.d. sequence with a geometric distribution of parameter p; all the random variables are assumed to be independent. The proof of the lemma is then easy to complete with the law of large numbers. □

5.2. **A Markov chain at infinity.** Roughly speaking, the transient behavior of the process $(X(t))$ depends on the location of the number $B_a(0)$ of items a in the bin among the subintervals $\mathcal{I}_0$, $\mathcal{I}_1$ or $\mathcal{I}_2$ of $\{0, 1, \ldots, k\}$.

Definition 12.

$$\mathcal{I}_0 = \left[0, \frac{(C-\lambda p)^+}{a}\right[\cap \mathbb{N}, \quad \mathcal{I}_1 = \left[\frac{(C-\lambda p)^+}{a}, \frac{C(1-p)}{E(S_1)}\right[\cap \mathbb{N},$$

$$\mathcal{I}_2 = \left[\frac{C(1-p)}{E(S_1)}, \frac{C}{a}\right] \cap \mathbb{N}.$$

The process with some specified large initial state is now investigated. We study separately the cases $B_a(0) \in \mathcal{I}_2$ and $B_a(0) \in \mathcal{I}_0$.

5.2.1. *Starting from $\mathcal{I}_2$.* For $N \in \mathbb{N}$, the initial condition x_N is the following

1. the bin is full and contains $Y_0 = k_0 \in \mathcal{I}_2$ items of size a;
2. an item 1 is at the head of the queue, it is followed by a string of α_N i.i.d. random variables with distribution μ, α_N is equivalent to αN with $\alpha > 0$.

If N is sufficiently large, an item of size 1 in the bin is replaced by an item 1. The function $t \to B_a(t)$ is thus a non increasing function, hence $Y_1 = \lim_{t\to+\infty} B_a(t)$ exists and the limit is reached in finite time by $(B_a(t))$.

If $Y_1 > C(1-p)/E(S_1)$ Lemma 11 shows that with probability 1 there will be an item of size 1 at the head of the queue an infinite number of times. Each time there is such a configuration, with positive probability an item a in the bin is replaced by a items of size 1 in the bin. Hence with probability 1, $B_a(t)$ will be decreased at least by 1 as long as its value is greater than $C(1-p)/E(S_1)$; consequently the limit Y_1 satisfies the inequality $Y_1 < C(1-p)/E(S_1)$, equivalently Y_1 is in the set $Y_1 \in \mathcal{I}_0 \cup \mathcal{I}_1$. We examine now the two possibilities for Y_1, either in $\mathcal{I}_0$ or $\mathcal{I}_1$.

a) If $Y_1 \in \mathcal{I}_0$, when Y_1 is reached by $(B_a(t))$, the state of the bin does not change as long as items 1 are in the queue, that is until η_{x_N} the first time when the queue contains only items a. The condition $\lambda p < C - aY_1$ and Lemma 11 show that η_{x_N} is almost surely finite. The initial number $L_{x_N,1}(0)$ of items 1 in the queue is equivalent to $p\alpha$ and the number of items 1 is decreased at rate $\lambda p - (C - Y_1 a)$, hence almost surely

$$\tau(\alpha) = \lim_{N\to+\infty} \frac{\eta_{x_N}}{N} = \frac{p\alpha}{C - Y_1 a - \lambda p},$$

and moreover for $0 \le t < \tau(\alpha)$,

$$\begin{aligned} \lim_{N\to+\infty} \frac{L_{x_N,a}(Nt)}{N} &= \alpha(1-p) + (\lambda(1-p) - Y_1)t, \\ \text{and} \quad \lim_{N\to+\infty} \frac{L_{x_N,1}(Nt)}{N} &= \alpha p + (\lambda p - (C - aY_1))t, \end{aligned}$$

in particular,

$$\lim_{N\to+\infty} \frac{L_{x_N,a}(\eta_{x_N})}{N} = \alpha \frac{C(1-p) - Y_1 E(S_1)}{C - Y_1 - \lambda p}.$$

b) If $Y_1 \in \mathcal{I}_1$. Since $\lambda p > C - Y_1 a$, the number of items 1 is converging to infinity, hence Lemma 11 shows that for N sufficiently large $\eta_{x_N} = +\infty$ a.s. and for all $t \ge 0$,

$$\begin{aligned} \lim_{N\to+\infty} \frac{L_{x_N,a}(Nt)}{N} &= \alpha(1-p) + (\lambda(1-p) - Y_1)t, \\ \text{and} \quad \lim_{N\to+\infty} \frac{L_{x_N,1}(Nt)}{N} &= \alpha p + (\lambda p - (C - aY_1))t. \end{aligned}$$

Here the number of items 1 *and* the number of items a explode linearly.

5.2.2. *Starting from* $\mathcal{I}_0$. For $n \in \mathbb{N}$, the initial condition x_N is the following

1. the bin contains $Y_0 = k_0 \in \mathcal{I}_0$ items of size a;
2. the queue is a string of α_N items of size a, α_N is equivalent to αn with $\alpha > 0$.

If N is sufficiently large, an item a in the bin is replaced by another item a; the number of a, $B_a(t)$ is non decreasing. The condition $k_0 \in \mathcal{I}_0$ is equivalent to $\lambda p < C - B_a(0)a$ and as long as $\lambda p < C - B_a(t)$, the number of 1 in the queue remains finite and hits 0. When this happens, $L_1(t) = 0$, there is a positive probability that at least a items of size 1 leaves the bin before the arrival of new item 1, hence that an additional item a enters the

bin. Consequently when N is large and t goes to infinity, $B_a(t)$ converges to Y_1 such that $\lambda p > C - Y_1 a$. Again, Y_1 is either in $\mathcal{I}_1$ or $\mathcal{I}_2$

a) If $Y_1 \in \mathcal{I}_2$. The lemma 11 shows that the variable τ_{x_N} is finite and

$$\tau(\alpha) = \lim_{N\to+\infty} \frac{\tau_{x_N}}{N} = \frac{p\alpha}{Y_1 E(S_1) - C(1-p)}, \tag{16}$$

and for $0 \le t < \tau(\alpha)$,

$$\lim_{N\to+\infty} \frac{L_{x_N,a}(Nt)}{N} = \alpha + (\lambda(1-p) - Y_1)t, \tag{17}$$

$$\text{and} \quad \lim_{N\to+\infty} \frac{L_{x_N,1}(Nt)}{N} = (\lambda p - (C - Y_1 a))t, \tag{18}$$

and at time τ_{x_N}, an item 1 is at the head of the queue, followed by a string of i.i.d. random variables with distribution μ. The length of this string is given by

$$\lim_{N\to+\infty} \frac{L_{x_N,1}(\tau_{x_N}) + L_{x_N,a}(\tau_{x_N})}{N} = \alpha \frac{\lambda p - (C - aY_1)}{Y_1 E(S_1) - C(1-p)}. \tag{19}$$

b) If $Y_1 \in \mathcal{I}_1$. This case is similar to the case b) of Section 5.2.1, with the following modification due to the initial state: for all $t \ge 0$,

$$\lim_{N\to+\infty} \frac{L_{x_N,a}(Nt)}{N} = \alpha + (\lambda(1-p) - Y_1)t,$$

$$\text{and} \quad \lim_{N\to+\infty} \frac{L_{x_N,1}(Nt)}{N} = (\lambda p - (C - aY_1))t.$$

5.2.3. *Starting from $\mathcal{I}_1$.* When $B_a(0) \in \mathcal{I}_1$, we set $Y_1 = B_a(0)$.

Definition 13. *The probability matrix (Q_{ij}) on $\{0, \dots, k\}$ is defined by*

$$Q_{ij} = P(Y_1 = j / Y_0 = i). \tag{20}$$

Notice that the Markov process associated with this matrix is not necessarily aperiodic: if $\mathcal{I}_0$ and $\mathcal{I}_2$ are non empty and $\mathcal{I}_1$ is empty, the Markov chain jumps from $\mathcal{I}_0$ to $\mathcal{I}_2$ and conversely.

The stochastic matrix Q is associated with a component (the state of the bin) of a transient process at infinity. The variable Y_1 is in some sense the exit state of this transient process.

5.3. **Random fluid limits and paths to infinity.** To describe the fluid limit in the transient case we need the following definition.

Definition 14. *The process (Y_n) is a Markov chain with transition matrix Q with $Y_0 = k_0 \in \mathcal{I}_1 \cup \mathcal{I}_2$. The hitting time of $\mathcal{I}_1$ by (Y_n) is denoted by H. The non-decreasing sequences (s_n), (t_n) and the sequences (u_n), (v_n) are defined by induction, $s_0 = 0$, $u_0 = 1$,*

$$v_{n+1} = u_n \frac{p(\lambda p - (C - aY_{2n}))}{Y_{2n} E(S_1) - C(1-p)}, \qquad t_{n+1} = s_n + u_n \frac{p}{Y_{2n} E(S_1) - C(1-p)},$$

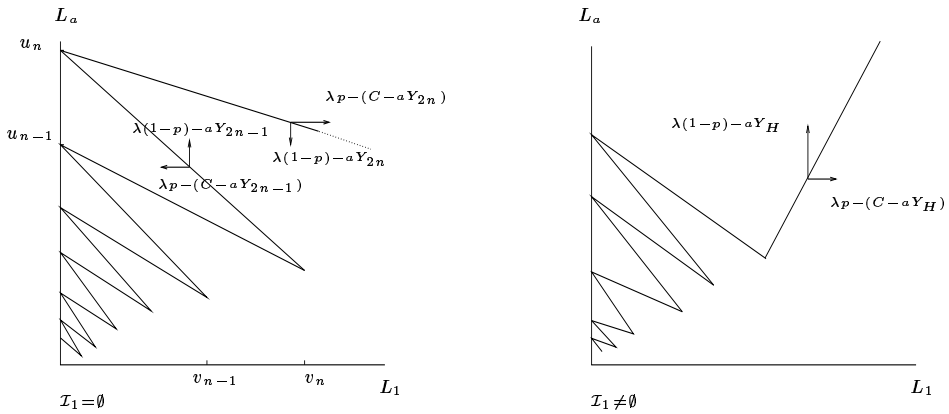

FIGURE 1. Some paths to infinity in the transient case

if $2n < H$, $t_{n+1} = +\infty$ *otherwise;*

$$u_{n+1} = v_{n+1}\frac{C(1-p) - Y_{2n+1}E(S)}{p(C - aY_{2n+1} - \lambda p)}, \qquad s_{n+1} = t_{n+1} + v_n\frac{1}{C - aY_{2n+1} - \lambda p},$$

if $2n+1 < H$, $s_{n+1} = +\infty$ *otherwise.*

The sequences (s_n) and (t_n) are possibly constant and equal to $+\infty$ after some finite rank, and for $n \in \mathbb{N}$, $s_n < t_{n+1} < s_{n+1}$. As long as $\mathcal{I}_1$ is not reached by the Markov chain, $Y_{2n} \in \mathcal{I}_2$ and $Y_{2n+1} \in \mathcal{I}_0$, hence the denominators in the above definition are positive.

These definitions have the following interpretation, for a moment Nx will mean a quantity equivalent to Nx when N is large. At time Nt_n, a string of Nu_n items a composes the queue and Y_{2n} items a are in the bin. Since $Y_{2n} \in \mathcal{I}_2$, the string of items a decreases and Nt_{n+1} is the first time since Ns_n when an item 1 is at the head of the queue. The queue is then a string of i.i.d. random variables with Nv_n items 1, and Y_{2n+1} items a are in the bin. The period between Nt_{n+1} and Ns_{n+1} is the opposite phase, the number of items 1 decreases and a string of items a grows at the head of the queue. The next theorem gives a formulation of this.

Theorem 15. *If the initial state x_N of the Markov process $(X_N(t))$ is a string of $N_0 = N$ items of size a, and the bin is full with k_0 items of size a, $k_0 \in \mathcal{I}_1 \cup \mathcal{I}_2$, then almost surely for $t \in [s_n, t_{n+1}]$,*

$$\lim_{N\to+\infty} \frac{L_{x_N,a}(Nt)}{N} = u_n + (\lambda(1-p) - Y_{2n})(t - s_n), \tag{21}$$

$$\text{and} \quad \lim_{N\to+\infty} \frac{L_{x_n,1}(Nt)}{N} = (\lambda p - (C - aY_{2n})(t - s_n). \tag{22}$$

and for $t \in [t_{n+1}, s_{n+1}]$,

$$\lim_{N\to+\infty} \frac{L_{x_N,a}(Nt)}{N} = v_n \frac{1-p}{p} + (\lambda(1-p) - Y_{2n+1})(t - t_n), \tag{23}$$

$$\text{and} \quad \lim_{N\to+\infty} \frac{L_{x_n,1}(Nt)}{N} = v_n + (\lambda p - (C - aY_{2n+1}))(t - t_n), \tag{24}$$

Proof. Everything has been already proved in the sections 5.2.1, 5.2.2 and 5.2.3. To illustrate this assertion we prove the first step between s_0 and t_1.

If Y_0 is in $\mathcal{I}_2$, then the string of items a at the head of the queue shrinks and disappears at time τ_{x_N}. Taking $u_0 = 1$, $s_0 = 0$, the variables t_1 and v_1 are defined by

$$t_1 = \lim_{N\to+\infty} \frac{\tau_{x_N}}{N} = u_0 \frac{p}{Y_0 E(S_1) - C(1-p)},$$

the last equality is given by the relation (16), and with (19) we define v_1 by

$$v_1 = \lim_{N\to+\infty} \frac{L_{x_N,1}(\tau_{x_N})}{N} = u_0 \frac{p(\lambda p - (C - aY_0))}{Y_0 E(S_1) - C(1-p)}.$$

Between s_0 and t_1 the behavior of the number of items 1 and a is given by (17) and (18), namely for $t \in [s_0, t_1[$,

$$\lim_{N\to+\infty} \frac{L_{x_N,a}(Nt)}{N} = u_0 + (\lambda(1-p) - Y_0)(t - s_0),$$

$$\text{and} \quad \lim_{N\to+\infty} \frac{L_{x_N,1}(Nt)}{N} = (\lambda p - (C - aY_0))(t - s_0).$$

At time τ_{x_n} there is an item 1 at the head the queue followed by an i.i.d. sequence of length $L_{x_N,1}(\tau_{x_N}) + L_{x_N,a}(\tau_{x_N})$. The system is then in the same caseas in Section 5.2.2.

If Y_0 is in $\mathcal{I}_1$, the results of Section 5.2.2, case b) apply. Taking $t_1 = +\infty$ for all $t \le t_1$,

$$\lim_{N\to+\infty} \frac{L_{x_N,a}(Nt)}{N} = u_0 + (\lambda(1-p) - Y_0)t,$$

$$\text{and} \quad \lim_{N\to+\infty} \frac{L_{x_N,1}(Nt)}{N} = (\lambda p - (C - aY_0))t.$$

□

Remark 16. *The above theorem shows that in the transient case the fluid limit is piecewise linear, the fluid equations are given by (21), (22), (23) and (24). The queueing network considered by Lu and Kumar* [15], *Rybko and Stolyar* [18] *has also this property. The additional and interesting feature of our model is the non deterministic nature of the slopes of these paths. This is mainly due to the "richness" of our state space. The scaling methods basically eliminate the stochastic noise around the limiting trajectories. But some randomness may remain, here at the instants (for the renormalized process)* (s_n) *and* (t_n) *when the process "chooses" the slope of the next piecewise path. A similar phenomenon occurs in the case of the particular model analyzed in* [10]. *In the case of queueing networks it is (perhaps) possible if the network has a significant number of nodes or classes.*

Theorem 15 considers a special initial condition: $k_0 \in \mathcal{I}_0$ items a in the bin and a string of N items a in the queue. We now give a somewhat informal description of the general case: all the transient paths will stick to the transient paths we described. If x is a "large" initial state, the initial items 1 will be served, hence after that time the size of the items after the first item 1 in the queue are i.i.d. with distribution μ. We can assume that this is the initial state. At this point we have several possibilities:

1. $L_{x,a}(0) \in \mathcal{I}_0$: in a finite time the queue will be a string of items a. This is the initial condition of the theorem.
2. $L_{x,a}(0) \in \mathcal{I}_2$: in a finite time the queue is an i.i.d string with distribution μ (Lemma 11). This is the state of the system at time t_1 in Theorem 15.
3. $L_{x,a}(0) \in \mathcal{I}_1$,
   (a) if $(B_{x,a}(t))$ remains forever in $\mathcal{I}_1$; this is the case described in the theorem after the hitting time of $\mathcal{I}_1$.
   (b) If $(B_{x,a}(t))$ visits $\mathcal{I}_0$, necessarily there will be an item 1 at the head of the queue followed by an i.i.d. string, otherwise the number of items a in the bin could not decrease. But then the number of items 1 in the system will reach 0, this is our case 1) here.
   (c) If $(B_{x,a}(t))$ visits $\mathcal{I}_2$, again at some time the queue will be a string of items a which will vanish with probability 1, this is the case 2) here.

5.4. **An explicit expression of the exit matrix Q.** With the above notations $\mathcal{I}_i$, $i = 0$, 1, 2 and $C = ka + r$, we define $-1 \leq n \leq k$, $-1 \leq m \leq k$ so that $\mathcal{I}_0 = \{0, \dots, n\}$ and $\mathcal{I}_2 = \{m+1, \dots, k\}$, with the convention that $\mathcal{I}_0$ [resp. $\mathcal{I}_2$] is empty if $n = -1$ [resp. $m = k$].

Proposition 17. *The coefficients of the matrix Q describing the transient behavior of the process $(X(t))$ are given by*

1. *if $i \in \mathcal{I}_0$ and $j \in \mathcal{I}_0^c = \{n+1, \dots, k\}$,*

$$Q(i,j) = f(j,a) \prod_{l=n+1}^{j-1} (1 - f(l,a)),$$

for $l > n$, $f(l, \cdot)$ is given by

$$f(l,a) = \sum_{i=C-(l+1)a+1}^{C-la} H(C-la, i) \prod_{j=i+1}^{C-la} (1 - H(C-la, j))$$

where the sequence $(H(K,i))$ is defined by induction

$$H(K,K) = 1 - \frac{K}{\lambda p},$$

$$H(K,i) = \frac{\lambda p H(K, i+1)}{\lambda p + i - 1 + H(K, i+1)} \tag{25}$$

for $K \in \mathbb{N}$ and $i < K$.

2. *If* $i \in \mathcal{I}_1$, $Q(i,i) = 1$.
3. *If* $i \in \mathcal{I}_2$ *and* $j \in \mathcal{I}_2^c = \{0, \dots, m\}$,

$$Q(i,j) = q_j \prod_{l=j+1}^{m} (1 - q_l),$$

with, for $l \leq m$,

$$q_l = \frac{(C - l(a-1))^{a-1}(1-p) - lp^{a-1}(C - l(a-1))^{a-2}}{(C - l(a-1))^{a-1}(1-p) - lp^{a-1}(C - la)(C - l(a-1))^{a-3}}.$$

Proof. 1. If $i \in \mathcal{I}_0$. Initially the queue has an infinite number of items a and no item 1, the bin contains i items a and $C - ia$ items 1. By definition the arrival rate λp of the items 1 is strictly less than their maximal rate of service $C - ia$. If $\lambda p < (i+1)a$, the total number of items 1 (in the bin and in the queue) will reach $C - (i+1)a$ with probability 1, leaving a free places. At that time, an item a enters the bin and the items 1 are served at rate $C - (i+1)a$ at the maximum, consequently the number of items a in the bin will reach the set $\mathcal{I}_0^c$ at the value $n+1$.

If $i \in \mathcal{I}_0^c = \{0, \dots, k\} - \mathcal{I}_0$, $\lambda p > (C - ia)$, there is a positive probability denoted by $f(i,a)$ that the number of items 1 remains strictly greater than $C - (i+1)a$; on this event the total number of items a remains constant equal to i.

Considering an $M/M/K$ queue with arrival rate λp, service rate 1 and $\lambda p > K$, we have to express the probability that, starting with K initial customers, the number of customers of this queue will always be strictly greater than $K - a$.

For $i \leq K$, denote by $H(K,i)$ the probability of never reaching $i-1$ starting from i, clearly

$$\begin{aligned} H(K,K) =& 1 - \frac{K}{\lambda p}, \\ H(K,i) =& \frac{\lambda p}{\lambda p + i}(H(K,i+1) + (1 - H(K,i+1))H(K,i)), \end{aligned}$$

for $i \leq K$. The quantity $f(l,a)$ can be expressed as

$$f(l,a) = \sum_{i=C-(l+1)a+1}^{C-la} H(C - la, i) \prod_{j=i+1}^{C-la} (1 - H(C - la, j)).$$

The first part of the proposition is proved.

2. If $i \in \mathcal{I}_1$, by definition $Q(i,i) = 1$.
3. If $i \in \mathcal{I}_2$. An item 1 is at the head of the queue followed by an infinite i.i.d. sequence with distribution μ. The bin contains i items a and $C - ia$ items 1. When an item a leaves the bin and if there is an item 1 at the head of the queue, it enters in the bin and it is followed by $a-1$ items 1; the number of items a is thus decreased by 1 and the first item 1 in the queue is at the position $1 + \sum_{l=1}^{a-1} G_l$ in the queue, where the G_l's are i.i.d. random variables with a geometric distribution with parameter $1-p$. As long as the number of items a in the bin is in $\mathcal{I}_2$, the number of a at the head of the

queue will reach 0 and at these moments there will be an item 1 in the first position of the queue, hence the possibility of replacing an item a in the bin. The number $(B_a(t))$ of items a in the bin hits the set $\mathcal{I}_2^c$ in m.

The first time $(B_a(t))$ reaches the value $j \in \mathcal{I}_2^c$, there is a string of $1 + \sum_{l=1}^{a-1} G_l$ items a at the head of the queue and there is a positive probability q_j that $(B_a(t))$ remains in j forever,

$$q_j = 1 - E\left(h(j)^{1+\sum_1^{a-1} G_l}\right) + E\left(h(j)^{1+\sum_1^{a-1} G_l}\right) \frac{C - ja}{C - j(a-1)},$$

and $h(j)$ is the probability that starting with exactly one item a at the head of the queue, an item 1 will reach the head of the queue; by homogeneity $h(j)^u$ is the probability of the same event but with k items a initially at the head of the queue. A classical result (easily verified) shows that

$$h(j) = \frac{j}{(1-p)(C - j(a-1))},$$

trite calculations finish the proof.

□

The sequence $(f(l, \cdot)$ defined by (25) may be expressed of course more explicitly (with hypergeometric functions) ; in any case it is a complicated expression.

References

[1] Maury Bramson, *Instability of FIFO queueing networks*, Annals of Applied Probability **4** (1994), no. 2, 414–431.

[2] ———, *Instability of FIFO queueing networks with quick service times*, Annals of Applied Probability **4** (1994), no. 3, 693–718.

[3] E.G. Coffman, Anja Feldman, Nabil Kahale, and Bjorn Poonen, *Computing call admission capacities in linear networks*, 1997, preprint.

[4] E.G. Coffman and Alexander L. Stolyar, *Bandwidth packing*, 1998, preprint.

[5] F. Comets, *Limites hydrodynamiques*, Séminaire Bourbaki, Astérisque, 1991, pp. 167–192.

[6] 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.

[7] ———, *A fluid limit model criterion for instability of multiclass queueing networks*, Annals of Applied Probability **6** (1996), no. 3, 751–757.

[8] Jean-François Dantzer and Philippe Robert, *A stability analysis of a multi-class queueing system*, 1998, In preparation.

[9] Vincent Dumas, *Unstable cycles in fluid Bramson networks*, Tech. Report 2318, INRIA, Domaine de Voluceau, Rocquencourt B.P. 105 78153 Le Chesnay Cedex France, August 1994.

[10] G. Fayolle, I. A. Ignatyuk, V. A. Malyshev, and M. V. Men'shikov, *Random walks in two-dimensional complexes*, Queueing Systems, Theory and Applications **9** (1991), no. 3, 269–300.

[11] G. Fayolle, V.A. Malyshev, and M.V. Men'shikov, *Topics in the constructive theory of countable Markov chains*, Cambridge University Press, 1995.

[12] 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.

[13] 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.

[14] C. Kipnis and Ph. Robert, *A dynamic storage process*, Stochastic Processes and their Applications **34** (1990), 155–169.

[15] 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.

[16] V. A. Malyshev, *Stabilization laws in the evolution of a random string*, Problems on Information Transmission **30** (1994), no. 3, 79–95.

[17] Sean P. Meyn, *Transience of multiclass queueing networks via fluid limit models*, Advances in Applied Probability **5** (1995), no. 4, 946–957.

[18] 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.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399