Probability in the Engineering and Informational Sciences, **14**, 2000, 57–79. Printed in the U.S.A.

ON THE STABILITY OF A BANDWIDTH PACKING ALGORITHM

JEAN-FRANÇOIS DANTZER
Département de Mathématiques
Bâtiment Fermat Université de Versailles
78035 Versailles, France
E-mail: Jean-Francois.Dantzer@inria.fr

MOSTAFA HADDANI AND PHILIPPE ROBERT
INRIA
Domaine de Voluceau
78153 Le Chesnay Cedex, France
E-mail: Mostafa.Haddani@inria.fr; Philippe.Robert@inria.fr

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.

1. INTRODUCTION

The model we consider here is a simplified description of a bandwidth allocation scheme in a communication 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 requiring 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, whereas 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:

© 2000 Cambridge University Press 0269-9648/00 $12.50

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

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. For these algorithms the unused portion of the bandwidth is not sufficient for the transmission of any message in the queue. 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 unlimited growth 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). A general result for which a simple stability condition holds is 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 5) 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 growing linearly, except at instants where, during a short time, the complexity of the process appears. 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 the 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 can be replaced by an equivalent node with different parameters. In the model considered here, when a node of a network is saturated, it may cause oscillations to the rest of the network; it cannot be reduced to another equivalent node. We conclude the paper with some comments of the behavior of the Best Fit algorithm.

1.1. 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 $]0, C]$. To simplify the presentation, we suppose each item in the bin is served at rate 1. The main results of the paper can be generalized without any difficulty to the case where the service rate depends on the size. The capacity of service

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

of this system is C and the average load is $\lambda E(S_1)$. 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.

1.2. The Markovian Representation

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

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

and none of the items in the queue fit in the bin; that is, for any index j

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

where $l_1(t)$ is the size of the older item still waiting at time t and $l_j(t)$ is the size of the jth oldest. The state space $\mathcal{S}$ of the Markov process is the set of the $x = (\underline{L}, \underline{B}) \in T^{(\mathbb{N})} \times T^{(\mathbb{N})}$ satisfying (1) and (2). The norm $|x|$ is the sum of the components of the vector x: if $(\underline{L}, \underline{B}) \in \mathcal{S}$, $|(\underline{L}, \underline{B})| = \sum b_i + \sum l_j$. When the component $\underline{B}$ of $(\underline{L}, \underline{B})$ 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 $\underline{B}$ component is upper bounded. This is not the case for the $\underline{L}$ component. Its dimension (the number of items in the queue) can be arbitrarily large. Our state space is countable but embedded in an infinite-dimensional vector space. Since the First Fit algorithm scans all the queues 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 $(X(t)) = ((\underline{L}(t), \underline{B}(t))$ 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 $\underline{L}$.
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 $x \in \mathcal{S}$, $L_{x,i}(t)$ [resp. $B_{x,i}(t)$] indicates the initial state x.*

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

1.3. 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 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. Further properties of the systems considered in this paper are presented in [8].

Coffman and Stolyar [4] analyzed the stability of the First Fit and Best Fit algorithms when the services are constant equal to 1. Because of the hypothesis on the services, the problem is related to static bin packing problems. They prove that the natural condition $\lambda E(S_1) < C$ is sufficient for the stability in the case of a symmetrical distribution of the sizes; in Coffman et al. [3], the sufficiency for stability of the condition $\lambda E(S_1) < C$ is considered in a more complex communication network.

Markov processes on strings occur also naturally in the multiclass queueing networks. Rybko and Stolyar [19] and Bramson [1,2] 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, Gaĭrat et al. [11] and Malyshev [17] 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.

Before analyzing in detail the ergodicity of $(X(t))$, we recall the main result 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 rescaled process $(\overline{X}_f(t))$ is defined by

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

since $X(0) = x$ and $\overline{X}_x^f(0) = 1$. The time variable and the space variable are scaled by a factor $f(x)$. The following theorem is a slight generalization of a theorem due to Rybko and Stolyar [19]; it gives an ergodicity criterion.

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

THEOREM 2: *If there exist an integrable stopping time U, constants $A, \varepsilon > 0$, and a positive function f on $\mathcal{S}$ such that*

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

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

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.

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. Condition (4) requires that at some random time, the norm of the process $(\overline{X}_x^f(t))$ is, on average, below its initial value. This suggests the analysis of the sequences of processes $(\overline{X}_x^f(t))$ when $|x|$ tends to infinity; the limit of one of its converging subsequences is called a *fluid limit*. If one can prove that every fluid limit converges almost surely to 0 after some time T, then, up to an integrability argument, Theorem 2 can be applied.

These scaling ideas are difficult to trace back. The origin of this criterion is the Lyapounov stability test of ordinary differential equations adapted to the stochastic case (see Hirsch and Smale [13] for the classical results and Has'minskiĭ [12] for the results concerning the stability of stochastic differential equations). The function f is usually called a Lyapounov function. 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 [16], Kumar and Seidman [15], and Rybko and Stolyar [19], 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; concerning transience criteria, Dai [7] and Meyn [18] obtained partial counterparts to the ergodicity results.

3. A NATURAL ERGODICITY CONDITION

The system has a capacity of service C, and the offered load of the flow of the items is $\lambda E(S_1)$. A natural condition for stability is thus $\lambda E(S_1) < C$. The following proposition shows that it is at least necessary for stability.

PROPOSITION 3: *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 arriving between 0 and t must be less than Ct plus the area due to the items in the bin at time t. The law of large numbers shows that this implies $\lambda E(S_1) \leq C$. ■

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

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 natural condition $\lambda E(S_1) < C$ and the real one is due to the waste of space in the bin.

PROPOSITION 4: *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

$$E_x(T_1) \leq A|x| \quad \textit{and} \quad E_x(|X(T_1)|) \leq |x|, \tag{6}$$

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

PROOF: With the rectangle packing analogy, we get the following inequality:

$$\sum x_i \sigma_i^0 + \sum_{i=1}^{\mathcal{N}_\lambda[0,T_1]} S_i \sigma_i \geq CT_1,$$

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, with the help of Wald's formula we get

$$E_x(T_1) \leq \frac{|x|}{C - \lambda E(S_1)}.$$

If (y_i) is the vector of the sizes of the items present at time T_1 in the system and (σ_i') the vector of their residual services,

$$\sum \sigma_i' y_i \leq \sum x_i \sigma_i^0 + \sum_{i=1}^{\mathcal{N}_\lambda[0,T_1]} S_i \sigma_i - CT_1,$$

taking the expected values, we get

$$E_x(|X(T_1)|) \leq |x| + (\lambda E(S_1) - C)E(T_1) \leq |x|.$$

The proposition is proved. ■

We can now give a simple ergodicity criterion. This result reduces the size of the set of initial states.

PROPOSITION 5: *If $\mathcal{S}_1$ is the subset of elements of $\mathcal{S}$ with the bin not completely full, $\mathcal{S}_1 = \{x = (B, L) \in \mathcal{S}/|B| < C\}$, and if there exist constants $0 < \varepsilon < 1$ and $A_0 \in \mathbb{R}$ and a stopping time U such that*

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

$$\limsup_{|x|\to+\infty,\, x\in\mathcal{S}_1} E\left(\frac{|X(U)|}{|x|}\right) \le 1-\varepsilon, \tag{7}$$

$$\limsup_{|x|\to+\infty,\, x\in\mathcal{S}_1} \frac{E_x(U)}{|x|} \le A_0, \tag{8}$$

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 $A_1 > A_0$, there exists $K > 0$ such that

$$E_x\left(\frac{|X(U)|}{|x|}\right) < 1-\varepsilon_0, \tag{9}$$

$$\frac{E_x(U)}{|x|} \le A_1, \tag{10}$$

if $|x| \ge K$ and $x \in \mathcal{S}_1$. If, as before, T_1 is the hitting time of the set $\mathcal{S}_1$ by $(X(t))$, the random variable V is defined as follows:

$V = U$ if $X(0) \in \mathcal{S}_1$;

$V = T_1$ if $X(0) \notin \mathcal{S}_1$ and $|X(T_1)| \le K$;

$V = T_1 + U$, otherwise;

the notation $T_1 + U$ stands for T_1 plus the variable U for the initial state $X(T_1)$. If (θ_t) is the time shift associated to the Markov process, it should be written as $T_1 + U \circ \theta_{T_1}$.

The variable V is clearly a stopping time. We take $x \in \mathcal{S}$ such that $|x| \ge 2K/\varepsilon_0$; then, the following hold:

- If $x \in \mathcal{S}_1$, the inequalities (9) and (10) show that $E_x(X(V)) \le (1-\varepsilon_0)|x|$ and $E_x(V) \le A_0|x|$.
- If $x \notin \mathcal{S}_1$, the strong Markov property of $(X(t))$ gives

  $$E_x(|X(V)|) = E_x(|X(T_1)|1_{\{|X(T_1)|\le K\}}) + E_x(E_{X(T_1)}(|X(U)|)1_{\{|X(T_1)|>K\}});$$

  using, again, relation (9) gives

  $$E_x(|X(V)|) \le K + (1-\varepsilon_0)E_x(|X(T_1)|);$$

  hence,

  $$\frac{E_x(|X(V)|)}{|x|} \le \frac{K}{|x|} + (1-\varepsilon_0)\,\frac{E_x(|X(T_1)|)}{|x|}$$

  the assumption on $|x|$ and inequality (6) give

  $$\frac{E_x(|X(V)|)}{|x|} \le \frac{\varepsilon_0}{2} + 1 - \varepsilon_0 = 1 - \frac{\varepsilon_0}{2}.$$

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

The estimation of the first moment of U is done in a similar way:

$$E_x(V) = E_x(T_1) + E_x(E_{X(T_1)}(U)1_{\{|X(T_1)|>K\}}).$$

Relations (6) and (10) give

$$E_x(V) \leq A|x| + A_1 E_x(|X(T_1)|) \leq A(1 + A_1)|x|.$$

The stopping time U satisfies the assumptions of Theorem 2; hence, 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, where a is an integer greater than 2. 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.

4.1. Assumptions.

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

4.2. A Priority Mechanism

If, at some moment, the bin contains only items 1, 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 $\lambda p > r$, as long as the queue is nonempty there will be no waste of space; even when a long string of items a is served with k items a in the bin, the items 1 will saturate the empty space in the bin. The natural condition $\lambda E(S_1) < C$ must be sufficient. Otherwise, for $\lambda p < r$, the items 1 will be transparent, since they can be served in the empty space which cannot be used by the items a; hence, only the items a matter. The ergodicity condition should be $\lambda(1-p) < k$.

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

PROPOSITION 6 (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*

$$\max(r - \lambda p, 0) + \lambda(p + (1-p)a) < C, \tag{11}$$

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

PROOF: To apply Proposition 5, we take a sequence of initial states $(x_N) = (l_N, b_N)$ in the set $\mathcal{S}_1$ such that $|x_N|$ converges to infinity as $N \to +\infty$; we denote by $(X_N(t)) = (\underline{L}_N(t), \underline{B}_N(t))$ the process with initial state x_N. All we have to prove is the existence of a stopping time U such that

$$\limsup_{N\to+\infty} \frac{E_{x_N}(|X_N(U)|)}{|x_N|} < 1 - \varepsilon, \tag{12}$$

$$\limsup_{N\to+\infty} \frac{E_{x_N}(U)}{|x_N|} < C, \tag{13}$$

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

An element $((b_i),(l_j))$ of $\mathcal{S}_1$ 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 $\lambda p < r$: Condition (11) is $\lambda(1-p) < k$. Since there are always r places for the 1 in the bin, the number of 1's in the system $(L_1(t))$ is bounded by the number of clients in a stable M/M/r queue with input rate λp and service rate 1. 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's in the system is 0. At time T_2, there will be k items of size a in the bin (or the queue is empty, but this is a trivial case); using the same method as in the proof of Proposition 5 (T_2 playing the role of T_1), it is easily seen that we may assume that for all the initial states x_N, the queue is a string of items a and that there are k items of size a in the bin.

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

$$U = \inf\left\{t \geq 0 \,|\, \mathcal{N}_k([0,t]) = \frac{|x_N|}{a}\right\},$$

$\mathcal{N}_k$ is the Poisson process with rate k describing the departure of the items a as long as k items a are in the bin; by the law of large numbers, we have

$$\lim_{N\to+\infty} \frac{E_{x_N}(U)}{|x_N|} = \frac{1}{ka}.$$

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

At time U, the number of items 1 is bounded by $Q(U)$, where $Q(t)$ is the number of jobs at time t of an independent stable M/M/r queue with input rate λp, service rate 1, and $Q(0) \leq C$. It is easily seen that

$$\lim_{N\to+\infty} \frac{E(Q(U))}{|x_N|} = 0.$$

At time U, the number of items a is less than $\mathcal{N}_{\lambda(1-p)}([0,U])$; hence, Lebesgue's theorem gives the identity

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

Case $\lambda p > r$: The initial state $x_N = (b_N, l_N)$ is in the set $\mathcal{S}_1$ and k_N is the number of items a in the bin for this state. 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 there are sufficiently many items a and $\lambda p < C - k_N a$, as long as $B_a(t) = k_N$, the number of items 1 is bounded by the number of customers in a stable M/M/$(C - ak_N)$; hence, within an integrable amount of time, there will be a "hole" of a items 1 in the bin. At that time, an item a enters the bin and the same phenomenon will occur again if $\lambda p < C - (k_N + 1)a$, and so on. Without loss of generality, we may thus assume that the inequality $\lambda p > C - k_N a$ holds for all the initial states. The relation $\lambda p > C - k_N a$ shows that after a finite time, the items 1 will saturate the places left by the items. As long as the initial items a are not served, the duration of time during which the bin is not full (i.e., the items 1 do not still saturate the queue) is clearly integrable and thus negligible at the fluid limit level. The variable U is defined as the time when the initial items a are served; clearly

$$\lim_{N\to+\infty} \frac{E_{x_N}(U)}{|x_N|} \leq \frac{1}{a\lambda(1-p)}.$$

Since the bin is almost always full between 0 and U,

$$\lim_{N\to+\infty} \frac{E_{x_N}(|X(U)|)}{|x_N|} = \lim_{N\to+\infty} \frac{(|x_N| + \lambda E(S_1) - C)E(U)}{|x_N|}$$

$$\leq 1 + \frac{\lambda E(S_1) - C}{\lambda(1-p)a} < 1.$$

Our proposition is proved. ■

Notice the advantage of starting from a state in $\mathcal{S}_1$ (i.e., the queue is a string of items a); it amounts to identifying (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.

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

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 subcase: 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 subsystem 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 description of the transient behavior is required to get some insight into more general distributions.

PROPOSITION 7: *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*

$$\max(r - \lambda p, 0) + \lambda(p + a(1-p)) > C, \tag{14}$$

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

PROOF: The proof is straightforward: If $\lambda p < r$, condition (14) is $\lambda(1-p) > k$, the number of items a is growing at least at rate $\lambda(1-p) - k$. Otherwise, if $\lambda p > r$, condition (14) is simply $\lambda E(S_1) > C$. The rectangle packing view of the algorithm gives, almost surely,

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

and the process $(X(t))$ is transient. ■

More interesting is the way the composition of the queue evolves. We now describe the paths to infinity in the transient case. Throughout this section, condition (14) holds.

The items in the queue after an item 1 form an i.i.d. (finite) sequence. An item 1 always fits in the bin if it is not full; hence, the First Fit algorithm does not alter the i.i.d. structure of the items after an item 1. When the process $(X(t))$ is transient, the string describing the queue is large. These large queues can thus be described by a finite string of items a followed by a finite i.i.d. string. All we have to describe is the behavior of these two strings: Does the string of items a "swallow" the i.i.d. string (i.e., the items 1 disappear)? Conversely, does the i.i.d. string swallow the string of items a (i.e., an item 1 arrives at the head of the queue)? In both cases, what is going on afterward? The interplay between these two cases is the main feature of the

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

analysis of the transient behavior. We shall first study the case where the i.i.d. string is infinite. The following definition will be used in the sequel.

DEFINITION 8: *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 $\inf \varnothing = +\infty$. *The variable τ is the first time an item of size* 1 *is at the head of the queue and η is the first time 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

We now study the case when the bin is full and the queue is composed of a string of items a followed by an infinite sequence of i.i.d. items 1 and a. As we shall see, the variable $\nu(t)$ (the number of items a at the beginning of the queue) will either increase when t is large or reach 0 with probability 1. This result is quite simple, but it will give the key to understand the transient behavior of the model.

At time $t = 0$, the initial state x_N of the system is defined as follows:

1. The bin contains k_0 items of size a, $k_0 \leq k$, and $C - k_0 a$ items of size 1 (the bin is full).
2. The queue is an infinite vector (l_i) such that
   (a) $l_i = a$ for $i = 1, \ldots, N$;
   (b) $\{l_{i+N}, i \geq 1\}$ 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.

As long as the number $B_a(t)$ of items a in the bin do not change the process, $\nu_{x_N}(t)$ is clearly a Markov process. The following lemma gives the condition under which $(\nu_{x_N}(t))$ is transient or, equivalently, when the string of items a at the head of the queue grows or shrinks.

LEMMA 9: *With the above assumptions, there are two cases.*

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

$$B_{x_N,a}(t) = k_0, \qquad \nu_{x_N}(t) > 0,$$

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

and the rate of growth is given by

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

- *If $k_0 > C(1-p)/E(S_1)$, the string of items a shrinks*: *For an integer 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: Notice that our assumption on the initial state implies that the string of items located behind the position $\nu(t)$ is an i.i.d. sequence with distribution μ. The proof of the lemma is fairly simple: As long as the first item 1 has not hit the head of the queue, its rank decreases by 1 at rate k_0 (when an item a leaves the bin). 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 a geometrically distributed random variable with parameter $1-p$. Consequently, the process $(\nu(t \wedge \tau_{x_N}))$ has the same distribution as

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

where $\kappa(t) = \mathcal{N}_{C-k_0 a}[0,t]$ is the counting measure of the departures of items 1 and (G_i) is an i.i.d. sequence with a geometric distribution of parameter $1-p$. 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. ■

The above lemma gives an indication on the interplay of the string of items a and the i.i.d. string in the queue. It can be very roughly described as follows:

If there are k_0 items a and $C - k_0 a$ items 1 in the bin and $k_0 < C(1-p)/E(S_1)$, then the following hold:

- If $\lambda p < C - k_0$, the number of items 1 decreases; hence, the string of items a grows and swallows the i.i.d. string.
- Otherwise, $\lambda p > C - k_0$, so the number of items a at the head of the queue grows as well as the number of items 1.

If $k_0 < C(1-p)/E(S_1)$, then the string of items a vanishes and an item 1 reaches the head of the queue. The next subsection gives a detailed description of these cases.

5.2. The Transient Cycle

The transient behavior will be defined with the intervals $\mathcal{I}_0$, $\mathcal{I}_1$, and $\mathcal{I}_2$ defined as follows:

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

DEFINITION 10:

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

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

The process with some specified large initial state is now investigated. In all the cases, we assume that the bin is initially full. The variable $B_a(0)$, the number of items a in the bin at time 0, determine the different cases.

5.2.1. Starting from $\mathcal{I}_2$. The initial condition x_N is given by the following:

1. The bin contains $B_a(0) = k_0 \in \mathcal{I}_2$ items of size a (and thus $C - k_0 a$ items 1).
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, where α_N is equivalent to αN for some $\alpha > 0$.

If $B_a(t)$ is in $\mathcal{I}_2$, using the definition of $\mathcal{I}_2$ and the condition $\lambda E(S_1) > C$, then

$$C(1-p) < B_a(t)E(S_1) < \frac{B_a(t)C}{\lambda};$$

hence, $\lambda(1-p) < B_a(t)$, so $\lambda p > C - aB_a(t)$. If N is sufficiently large, the number of items 1 is growing. Consequently, as long as $B_a(t)$ is in $\mathcal{I}_2$,

- an item of size 1 in the bin is replaced by an item 1;
- according to Lemma 9, an item 1 reaches the head of the queue with probability 1. At that time, if an item a leaves the bin (and this occurs with positive probability), then it will be replaced by this item 1 followed with $a-1$ others.

The function $t \to B_a(t)$ may decrease as long as an item 1 reaches the head of the queue. This is the case with probability 1 when $B_a(t) \in \mathcal{I}_2$. Hence, in a finite time, $(B_a(t))$ will leave this interval. When $B_a(t)$ is in $\mathcal{I}_2^c$, $B_a(t) < C(1-p)/E(S_1)$ and as long as items 1 are in the queue, $t \to B_a(t)$ is nonincreasing. Lemma 9 shows in this situation that if the number of items a in the queue is sufficiently large (and because of the stochastic fluctuations, this happens with probability 1), the items 1 will not reach the head of the queue any longer. Consequently, after a while, the value of $(B_a(t))$ will be stabilized at some value Y_1 until the items 1 leave the queue completely. This value is clearly reached in finite time by $(B_a(t))$. We have two cases $Y_1 \in \mathcal{I}_0$ or $Y_1 \in \mathcal{I}_1$.

(a) If $Y_1 \in \mathcal{I}_0$, then $\lambda p < C - aY_1$. The initial number $L_{x_N,1}(0)$ of items 1 in the queue is equivalent to $p\alpha N$ and the number of items 1 is decreased at a rate

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

of $\lambda p - (C - Y_1 a)$. The time η_{x_N} when the queue is only a string of items a satisfies, almost surely

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

and since the number of items a and 1 increases respectively at rates $\lambda(1-p) - Y_1$ and $\lambda - (C - aY_1)$ until that time, for $0 \leq t < \tau(\alpha)$,

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

and

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

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 $Y_1 \in \mathcal{I}_1$, then $\lambda p > C - Y_1 a$. The number of items 1 is almost surely converging to infinity and the length of the string of items a at the head of the queue grows linearly (see Lemma 9). For N sufficiently large, $\eta_{x_N} = +\infty$ a.s. and, for all $t \geq 0$,

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

and

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

In this case, the number of items 1 *and* the number of items a grow linearly.

5.2.2. Starting from $\mathcal{I}_1$. If $B_a(0) \in \mathcal{I}_1$ and the initial state for the queue is the same as for the previous case, it is easily seen that for N sufficiently large, $(B_a(t))$ is constant and equal to $Y_1 = B_a(0)$. As in the case (b), almost surely

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

and

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

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

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

(1) The bin contains $B_a(0) = k_0 \in \mathcal{I}_0$ items of size a and $C - k_0 a$ items of size 1.
(2) The queue is a string of α_N items a; α_N is equivalent to αN with $\alpha > 0$.

As long as the string of items a at the head of the queue is nonempty, at its departure an item a in the bin is replaced by the first item a of the queue. If $B_a(t) \in \mathcal{I}_0$, then $\lambda p < C - aB_a(t)$; as long as $B_a(t) \in \mathcal{I}_0$ with probability 1, the queue will not contain any 1 at some instants. At these times, with positive probability a, items 1 may leave the bin; in this case, they will be replaced by an item a (if N is sufficiently large). During that period, the function $t \to B_a(t)$ will be nondecreasing. Hence, with probability 1, $(B_a(t))$ will leave $\mathcal{I}_0$. At that time, $\lambda p > C - aB_a(t)$; thus, a string of i.i.d. random variables appears at the end of the string. The variable $\nu_{x_n}(t)$ will be always finite after some time t_0. We denote by Y_1 the value of $B_a(t)$ at time t_0. As long as the string of items a is not empty, the value of $B_a(t)$ remains constant after that time. We also have to consider the two cases $Y_1 \in \mathcal{I}_1$ or $Y_1 \in \mathcal{I}_2$.

(a) If $Y_1 \in \mathcal{I}_2$, then $Y_a > C(1-p)/E(S_1)$. Lemma 9 shows that the variable τ_{x_N} is finite and almost surely,

$$\tau(\alpha) = \lim_{N \to +\infty} \frac{\tau_{x_N}}{N} = \frac{p\alpha}{Y_1 E(S_1) - C(1-p)}; \tag{15}$$

as before it is easily seen that for $0 \le t < \tau(\alpha)$,

$$\lim_{N \to +\infty} \frac{L_{x_N,a}(Nt)}{N} = \alpha + (\lambda(1-p) - Y_1)t, \tag{16}$$

and

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

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

$$\lim_{N \to +\infty} \frac{L_{x_N,1}(\tau_{x_N}) + L_{x_N,a}(\tau_{x_N})}{N} = \alpha \frac{\lambda p - (C - aY_1)}{Y_1 E(S_1) - C(1-p)}. \tag{18}$$

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

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

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

and

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

A Markov chain

We have defined a variable Y_1 for all the possible $B_a(0)$; this gives the following stochastic matrix.

DEFINITION 11: *The probability matrix* $Q = (q_{ij})$ *on* $\{0,\dots,k\}$ *is defined by*

$$q_{ij} = P(Y_1 = j \,|\, B_a(0) = i).$$

Notice that the Markov process associated with this matrix is not necessarily aperiodic: If $\mathcal{I}_0$ and $\mathcal{I}_2$ are nonempty and $\mathcal{I}_1$ is empty, the Markov chain jumps from $\mathcal{I}_0$ to $\mathcal{I}_2$ and conversely. If $\mathcal{I}_1$ is nonempty, any state in this interval is absorbing.

The choices of the three specified initial states respectively in $\mathcal{I}_0$, $\mathcal{I}_1$, and $\mathcal{I}_2$ has led to two transient Markov behaviors: Basically, there is a transient Markov process for which the string of items a is growing, and for the other Markov process, the string of i.i.d. items is growing. The process $B_a(\cdot)$ is a component of this transient Markov process and its limiting value is Y_1. The possible values of Y_1 represent the different points at infinity of the Markov process. The matrix Q can be interpreted as an exit matrix of a transient Markov process.

5.3. Random Fluid Limits and Paths to Infinity

We have all the necessary ingredients to describe the transient paths of this Markov process.

DEFINITION 12: *The process* (Y_n) *is a Markov chain with transition matrix* Q *with* $Y_0 = k_0 \in \mathcal{I}_2$. *The hitting time of the interval* $\mathcal{I}_1$ *by* (Y_n) *is denoted by* H. *The nondecreasing sequences* (s_n) *and* (t_n) *and the sequences* (u_n) *and* (v_n) *are defined by induction*, $s_0 = 0$, $u_0 = 1$:

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

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

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

and 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 Definition 12 are positive.

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

These definitions have the following interpretation: For a moment, Nx will mean a quantity equivalent to Nx when N is large. At time Ns_n, a string of Nu_n i.i.d. items composes the queue and Y_{2n} items a are in the bin. Between Ns_n and Nt_n, a string of items a grows and the number of items 1 decreases. The instant Nt_n is the first time since Ns_n, when the queue is a string of Nv_n items a and Y_{2n+1} items a are in the bin. During the period between Nt_n and Ns_{n+1}, the number of items 1 increases and the string of items a decreases. At time $Ns_{n+1}, Y_{2(n+1)}$ items a are in the bin and the queue is a string of Nu_{n+1} i.i.d. items. This is a typical transient cycle of the system. The next theorem gives a formulation of this.

THEOREM 13: *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, $k_0 \in \mathcal{I}_2$, then almost surely for $t \in [s_n, t_{n+1}]$,*

$$\lim_{N\to+\infty} \frac{L_{x_N,a}(Nt)}{N} = u_n + (\lambda(1-p) - Y_{2n})(t - s_n) \tag{19}$$

and

$$\lim_{N\to+\infty} \frac{L_{x_n,1}(Nt)}{N} = (\lambda p - (C - aY_{2n}))(t - s_n); \tag{20}$$

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

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

and

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

PROOF: Everything has been proved already in Sections 5.2.1–5.2.3. 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 $u_0 = 1$ and $s_0 = 0$, the variable t_1 is 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 relation (15). With (18), we define v_1 by

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

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

Between s_0 and t_1, the behavior of the number of items 1 and a is given by (16) and (17); namely, for $t \in [s_0, t_1[$,

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

and

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

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 situation as in Section 5.2.3.

If Y_0 is in $\mathcal{I}_1$, the results of Section 5.2.3, case (b), apply. Taking $t_1 = +\infty$ for all $t \le t_1$,

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

and

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

Remark: Theorem 13 shows that in the transient case, the fluid limit is piecewise linear (see Figure 1); the fluid equations are given by (19)–(22). The queueing network considered by Lu and Kumar [16] and Rybko and Stolyar [19] also has this property. The additional and interesting feature of our model is the nondeterministic nature of the slopes (Y_n) of these piecewise linear paths. This is mainly due to the

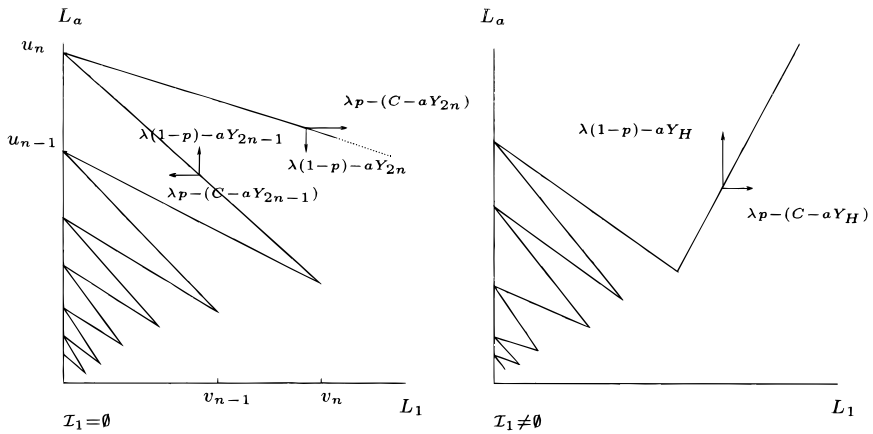


FIGURE 1. Some paths to infinity in the transient case.

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

"richness" of our state space. The scaling methods basically eliminate the stochastic noise around the limiting trajectories. However, 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 linear path. A related phenomenon has been observed in the case of the particular model [10]. In the case of queueing networks, it is (perhaps) possible if the network has a significant number of nodes or classes.

Theorem 13 considers a special initial condition: $k_0 \in \mathcal{I}_2$ 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 9). This is the state of the system at time t_1 in Theorem 13.
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).
   (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).

5.4. An Explicit Expression of the Exit Matrix Q

With the notations $\mathcal{I}_i$, $i = 0, 1, 2$, and $C = ka + r$, we define $-1 \le n \le k$ and $-1 \le m \le k$ 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 14: *The coefficients of the matrix Q describing the transient behavior of the process* $(X(t))$ *are given by the following*:

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*

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

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

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

$$H(K,K) = 1 - \frac{K}{\lambda p},$$
$$H(K,i) = \frac{\lambda p H(K,i+1)}{\lambda p + i - 1 + H(K,i+1)} \tag{23}$$

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* $j \in \mathcal{I}_2^c = \{0,\ldots,m\}$,

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

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. For $i \in \mathcal{I}_0$. Initially, the queue has an infinite number of items a and no item 1, and 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 a 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 $i \in \mathcal{I}_0^c = \{0,\ldots,k\} - \mathcal{I}_0$, $\lambda p > (C - ia)$, there is a positive probability denoted by $f(i,a)$ that the number of items 1 remains strictly greater than $C - (i+1)a$; for 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,

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

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

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 $i \in \mathcal{I}_1$, $Q(i,i) = 1$ by definition.
3. For $i \in \mathcal{I}_2$. 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:

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

   where $h(j)$ 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 (23) may be expressed, of course, more explicitly (with hypergeometric functions); in any case, it is a complicated expression.

6. THE BEST FIT ALGORITHM

The difference with First Fit in this case is that an item a is replaced by an item a in the bin if the queue contains at least an item a. This feature simplifies the state space. The Markovian description of the queue can be reduced to a two-dimensional component, the number of items 1 and a, respectively. It is easily seen that the ergodicity condition is the same as for First Fit. The description of the transient behavior is quite similar; the corresponding intervals in this case are

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press

$$\mathcal{I}_0 = \left[0, \frac{(C-\lambda p)^+}{a}\right[, \qquad \mathcal{I}_1 = \left[\frac{(C-\lambda p)^+}{a}, \lambda(1-p) \wedge k\right[,$$

$$\mathcal{I}_2 = \left[\lambda(1-p) \wedge k, \frac{C}{a}\right],$$

with $k = \lfloor C/a \rfloor$.

Acknowledgments

One of the authors (P.R.) wishes to thank Ed Coffman and Vincent Dumas for useful discussions on this paper.

This work has been partially supported by the Esprit Project 25046 - Catserver.

References

1. Bramson, M. (1994). Instability of FIFO queueing networks. *Annals of Applied Probability* 4(2): 414–431.
2. Bramson, M. (1994). Instability of FIFO queueing networks with quick service times. *Annals of Applied Probability* 4(3): 693–718.
3. Coffman, E.G., Feldman, A., Kahale, N., & Poonen, B. (1997). Computing call admission capacities in linear networks. Preprint.
4. Coffman, E.G. & Stolyar, A.L. (1998). Bandwidth packing. Preprint.
5. Comets, F. (1991). Limites hydrodynamiques. *Séminaire Bourbaki, Astérisque* 90/91(201–3):167–192.
6. Dai, J.G. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Annals of Applied Probability* 5(1): 49–77.
7. Dai, J.G. (1996). A fluid limit model criterion for instability of multiclass queueing networks. *Annals of Applied Probability* 6(3): 751–757.
8. Dantzer, J.-F. & Robert, P. (1998). A stability analysis of a multi-class queueing system. In preparation.
9. Dumas, V. (1999). Diverging paths in FIFO fluid networks. *IEEE Transactions on Automatic Control* AC-44(1), 191–194.
10. Fayolle, G., Ignatyuk, I.A., Malyshev, V.A., & Men'shikov, M.V. (1991). Random walks in two-dimensional complexes. *Queueing Systems, Theory and Applications* 9(3): 269–300.
11. Gaĭrat, A.S., Malyshev, V.A., Men'shikov, M.V., & Pelikh, K.D. (1995). Classification of Markov chains describing the evolution of random strings. *Russian Mathematical Surveys* 50(2): 237–255.
12. Has'minskiĭ, R.Z. (1980). Translator: D. Louvish, *Stochastic stability of differential equations*. Sijthoff & Noordhoff, Alphen aan den Rijn.
13. Hirsch, M.W. & Smale, S. (1974). Differential equations, dynamical systems, and linear algebra. New York and London: Academic Press.
14. Kipnis, C. & Robert, P. (1990). A dynamic storage process. *Stochastic Processes and Their Applications* 34: 155–169.
15. Kumar, P.R. & Seidman, T.I. (1990). Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Transactions on Automatic Control* AC-35(3): 289–298.
16. Lu, S.H. & Kumar, P.R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control* AC-36, 1406–1416.
17. Malyshev, V.A. (1994). Stabilization laws in the evolution of a random string. *Problems on Information Transmission* 30(3): 79–95.
18. Meyn, S.P. (1995). Transience of multiclass queueing networks via fluid limit models. *Advances in Applied Probability* 5(4): 946–957.
19. Rybko, A.N. & Stolyar, A.L. (1992). On the ergodicity of random processes that describe the functioning of open queueing networks. *Problems on Information Transmission* 28(3): 3–26.

https://doi.org/10.1017/S0269964800141063 Published online by Cambridge University Press