On the cut-off phenomenon in some queueing systems

Panagiotis Konstantopoulos, François Baccelli

UNITÉ DE RECHERCHE
INRIA-SOPHIA ANTIPOLIS

Rapports de Recherche

N° 1290

Programme 3
Réseaux et Systèmes Répartis

ON THE CUT-OFF PHENOMENON IN SOME QUEUEING SYSTEMS

Panagiotis KONSTANTOPOULOS
François BACCELLI

Octobre 1990

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

RR_1290

Sur les phénomènes de coupure dans les sytèmes de files d'attente

Panagiotis Konstantopoulos* et François Baccelli

Septembre 1990

INRIA, Sophia Antipolis
2004 Route des Lucioles
06565 Valbonne Cedex
FRANCE

Résumé

Un certain nombre de systèmes de files d'attente possèdent une intéressante propriété connue sous le nom de "phénomène de coupure": après une normalisation adéquate, la distance en variation totale entre le processus transitoire et le processus stationnaire converge vers une fontion de saut lorsque la charge initiale tend vers l'infini. Le but de cet article est de prouver que cette propriété est une conséquence directe du couplage entre ces deux processus, et qu'elle est donc généralisable à des systèmes sans structure markovienne.

Mots-clés: files d'attente stationnaires et ergodiques, distance en variation totale, temps de relaxation

*Le travail de cet auteur a été en partie financé par la "National Science Foundation" (grant ASC 88-8802764)

On the cut-off phenomenon in some queueing systems

P. Konstantopoulos*and F. Baccelli

INRIA, Sophia Antipolis
2004 Route des Lucioles
06565 Valbonne Cedex
FRANCE

Abstract

A number of stochastic queueing systems exhibit an interesting phenomenon, termed as the cut-off phenomenon. A properly scaled version of the distance between the transient process and the stationary one, converges to a step function as the initial load converges to infinity. The purpose of this paper is to promote the idea that this phenomenon is a direct consequence of the coupling between the two processes, being thus generalizable to systems lacking any kind of Markovian structure.

STATIONARY AND ERGODIC QUEUEING SYSTEMS, TOTAL VARIATION DISTANCE, SETTLING TIME

1 Introduction

The cut-off phenomenon in a stochastic system, refers, loosely speaking, to the existence of a time such that, before this time the system is far from stationarity, while, after this time the system is very close to stationarity. The above statement has to be interpreted asymptotically, as a parameter N of the system converges to infinity. To be more specific, we give a few examples of systems that have been studied in the literature:
Aldous [1] has looked into random walks on finite groups. For these processes he has shown that the distance between the transient process and its stationary version is maximal before the settling time and minimal after. The parameter that is taken to converge to infinity here is the total number of states. Anantharam [2] has extended these ideas in order to identify the settling time of a closed Jackson Network, asymptotically as the total number of customers N converges to infinity. Stamoulis and Tsitsiklis [9] treated a non-Markov-chain case, that of the GI/GI/1 queue. This system was parametrized by the initial population N.

The purpose of our paper is to show, by means of specific examples, first that the cut-off phenomenon is valid even for systems without any Markovian structure at all and second that it is essentially based on coupling. In Section 2 we study the general single-server queue with stationary and ergodic input and in Section 3 we study a multi-dimensional system that has applications, for instance, to queues with locking customers.

*This work was supported in part by the National Science Foundation under grant ASC 88-8802764

2 The cut-off phenomenon for the single-server queue

We start with this simple model so as to make the ideas of the proof transparent. These will be extended to the more general system of Section 3. Consider a single-server queue with general service discipline, input rate λ, and service rate μ. Assuming that the queue is stable, i.e., that the state of the queue at time t converges to a steady-state as $t \to \infty$, we are interested in a numerical estimate for the time that the queue "reaches stationarity" (the so-called settling time) under the condition that the queue has a large initial load.

Let $d_N(t)$ be the total variation distance between the distribution of the population X_t^N of the queue at time t when the initial load is N (i.e., when $X_0^N = N$) and its steady-state distribution π. This is defined as

$$d_N(t) = \sup_A |P(X_t^N \in A) - \pi(A)|, \tag{2.1}$$

where the supremum ranges over all subsets A of the non-negative integers. Then the following happens: There exists a positive constant, say α, such that $d_N(t)$ is approximately equal to its maximum value (that is, 1) when $t < N\alpha$ and approximately equal to its minimum value (that is, 0) when $t > N\alpha$, asymptotically as $N \to \infty$. This statement (and the conditions under which it holds) will be made more precise later. It will also be shown that the value of α is equal to $1/(\mu - \lambda)$.

The intuition (at least for an M/M/1 queue) is as follows: It takes time $N/(\mu - \lambda)$, on the average, for the queue to get rid of its initial load N (this is due to the fact that the first time that the queue reaches the empty state is the sum of N i.i.d. random variables with the same mean $1/(\mu - \lambda)$). If N is large then the steady state version of the queue will have emptied before time $N/(\mu - \lambda)$ with high probability. After hitting the zero state we can couple the two Markov chains without changing their marginal distributions. This is an intuitive explanation of the conjecture that $d_N(t)$ is almost 0 after $N/(\mu - \lambda)$ and almost 1 before. The result holds true for a queue with general stationary and ergodic input. The idea of the proof is not far from the above intuitive explanation for the M/M/1 queue, as it is indeed based on coupling.

Let now A_t be the arrival process and S_t the service process of the queue. We assume that the processes are jointly stationary and ergodic. Thus both $\lambda := \lim_{t\to\infty} A_t/t$ and $\mu := \lim_{t\to\infty} S_t/t$ exist and are assumed to be finite. We also assume that the stability condition $\lambda < \mu$ is satisfied. Let X_t be the population (number of customers in the queue) at time t. The initial population X_0 is a finite random variable. We denote by X_t^N the population process with $X_0 = N$. It is known that there exists a finite random variable $\tilde{X}_0$ for the initial population such that the resulting process $\tilde{X}_t$ is stationary and ergodic. Furthermore, any population process X_t couples with $\tilde{X}_t$ in finite time. As a result, $d(X_t, \tilde{X}_t) \to_{t\to\infty} 0$. Here $d(X,Y)$ denotes the total variation distance between two random variables (or, more precisely, between their distributions) X and Y:

$$d(X,Y) := \sup_A |P(X \in A) - P(Y \in A)|.$$

For a proof of these results we refer to the original paper of Loynes [8]. Now let $d_N(t)$ be as in (2.1) with $\pi(A) := P(\tilde{X}_t \in A)$. The following will be shown in the next section.

Theorem 1 *Under the above-mentioned hypotheses,*

$$\lim_{N\to\infty} d_N(Nt) = \begin{cases} 1 & \text{if } 0 \le t < \alpha \\ 0 & \text{if } t > \alpha, \end{cases} \tag{2.2}$$

where $\alpha := 1/(\mu - \lambda)$.

Before proceeding to the proof Theorem 1 we first show the following lemma:

Lemma 1 *If* $T_N := \inf\{t > 0 : X_t^N = 0\}$ *then*

$$\lim_{N\to\infty} \frac{T_N}{N} = \frac{1}{\mu - \lambda}, \ a.s.$$

Proof Consider the process defined by

$$Z_t := A_t - S_t.$$

Clearly, the piece of the process X_t^N for $0 \le t \le T_N$ is the same as the piece of the process $N + Z_t$ for t between 0 and the first time at which Z_t hits $-N$. That is, $T_N := \inf\{t > 0 : Z_t = -N\}$. Clearly then, $T_N \to \infty$ and $Z_t/t \to \lambda - \mu$. Hence, $Z_{T_N}/T_N \to \lambda - \mu$, and since $Z_{T_N} = -N$, we have $T_N/N \to 1/(\mu - \lambda)$. Note that all the above convergences are in the almost sure sense. This proves Lemma 1. □

Proof of Theorem 1

Suppose first that $t > \alpha := 1/(\mu - \lambda)$. From the triangle inequality we have

$$d_N(Nt) = d(X_{Nt}^N, \tilde{X}_{Nt}) \le d(X_{Nt}^N, X_{Nt}^0) + d(X_{Nt}^0, \tilde{X}_{Nt}), \tag{2.3}$$

where X_t^0 is the population process that starts from zero. By what has been mentioned above, this process couples with the stationary process $\tilde{X}$ and so the last term of (2.3) tends to 0 as $N \to \infty$. For the first term we have the usual coupling estimate

$$d(X_{Nt}^N, X_{Nt}^0) \le P(T_N > Nt)$$

which, in view of the result of Lemma 1 and the assumption that $t > \alpha$, tends to zero as well.

Suppose next that $0 \le t < \alpha$. Observe that

$$d_N(Nt) = d(X_{Nt}^N, \tilde{X}_{Nt}) = d(\frac{X_{Nt}^N}{N}, \frac{\tilde{X}_{Nt}}{N}). \tag{2.4}$$

From the fact that $\tilde{X}_t$ can be written as $\tilde{X}_0 + A_t - D_t$, where D_t is a stationary and ergodic departure process with the same rate λ as the arrival process A_t, we get

$$\lim_{N\to\infty} \frac{\tilde{X}_{Nt}}{N} = 0, \text{ a.s.} \tag{2.5}$$

On the other hand, the fact that $X_{Nt}^N = N + A_{Nt} - S_{Nt}$ for $Nt < T_N$ together with the result of Lemma 1 and the convergences $A_{Nt}/N \to_{N\to\infty} \lambda t$, $S_{Nt}/N \to_{N\to\infty} \mu t$, yields

$$\lim_{N\to\infty} \frac{X_{Nt}^N}{N} = 1 - (\mu - \lambda)t, \text{ a.s.} \tag{2.6}$$

Since the limit in (2.5) is not equal to the limit in (2.6) (the latter is positive, by the assumption $t < \alpha$), it follows easily that (2.4) converges to 1. To see this just choose an $\epsilon > 0$ such that $1 - (\mu - \lambda)t > \epsilon$. Then $d(X_{Nt}^N/N, \tilde{X}_{Nt}/N) \ge |P(-\epsilon < X_{Nt}^N/N < \epsilon) - P(-\epsilon < \tilde{X}_{Nt}^N/N < \epsilon)| \to 1 - 0 = 1$. This finishes the proof of Theorem 1. □

Settling times for other processes associated with the system

We have investigated the settling time of the queue-length process $\{X_t, t \in R\}$. Consider now the queue length process $X_n := X_{t_n-}$, $n \in Z$, just before the arrival times t_n. It is not difficult to see that the settling time (settling index) for the latter process is proportional to $\lambda/(\mu - \lambda)$. In other words, if π^0 denotes the steady-state distribution of $\{X_n\}$, then $d(X^N_{Nn}, \pi^0)$ converges to 1 [resp. 0] if $n < \lambda/(\mu - \lambda)$ [resp. $n > \lambda/(\mu - \lambda)$], as $N \to \infty$. We should, perhaps, note at this point that the customer-stationary distribution π^0 is related to the time-stationary distribution π by means of a Palm transformation.

Another process that is of interest is the workload process. The settling time for this process can be seen to be proportional to $\mu/(\mu - \lambda)$ (provided that the parameter converging to infinity is the initial workload). Finally, the process of the workload as seen by the arriving customers (which is equal to the waiting time if the service discipline is first-come first-serve), has a settling time proportional to $\lambda\mu/(\mu - \lambda)$.

Let us finally note that the results presented in this section generalize the results of [9], as they get rid of the independence assumptions between the inter-arrival or service times.

3 The cut-off phenomenon for more general systems

We turn now to the generalization of the ideas developed in the previous section and consider more general systems. The class of systems for which our analysis is applicable is the one that can be described by equations that generalize the classical Lindley's equation (see (3.2) below). Examples of such systems are: Queues with locking customers (see the definition below), queueing networks with a certain service discipline (discussed in reference [4]), a tandem network with manufacturing blocking and unit-size buffers (see [4]) and, more generally, the class of Petri nets that are *event graphs* (these are discussed in [3]). However, for concreteness, we choose to give the proofs of a system with locking.

Consider the following computer system with s servers: Customers arrive at times t_n. Customer n demands service from one or more servers. Let π_n be the set of servers required by customer n. This is the case, for instance, when a task is split into sub-tasks for execution in parallel processors. For each j in π_n we have a corresponding service time σ^j_n. Furthermore, customer n locks the servers after service. In other words, it does not leave the system before all of its sub-parts have completed service. This means that at times, some of the servers may be idle while there are other customers awaiting for service. Note that we do not require the sub-parts to synchronize when service begins. This system differs from the fork-join queue (see [5] and [7]) because in the latter system the sub-parts do have to synchronize after service, but this synchronization takes place in an infinite-size buffer and so the finished parts do not block the server. Let W^j_n be the virtual waiting time in server j at time t_n- (just before the arrival of customer n). The idea is that W^j_n is the actual waiting time of part j if $j \in \pi_n$. It is then not difficult to see that we have the following recursion:

$$W^i_{n+1} = \begin{cases} \max_{j \in \pi_n}(W^j_n + \sigma^j_n - \tau_n)^+ & \text{if } i \in \pi_n \\ (W^i_n - \tau_n)^+ & \text{otherwise.} \end{cases} \tag{3.1}$$

Here $\tau_n = t_{n+1} - t_n$ and $x^+ = \max(x, 0)$. It can be easily seen that this equation is a generalization of Lindley's equation for a single-server queue. The cutoff phenomenon for the latter system was presented in the previous section. Our argument there was based on the continuous-time evolution of the system but we could have easily taken a discrete-time point of view as well.

In the present section we prefer to work in discrete time. The treatment of the multi-dimensional system of this section is somewhat more technical than that of the single-server queue. However, the essential ideas remain unchanged. This is why we presented the simpler case first. The statistical assumptions for the system are as follows:

(i) The sequence $\{\tau_n, \pi_n, (\sigma_n^i; i \in \pi_n)\}_{n \in Z}$ is a stationary and ergodic sequence under some probability measure P.

(ii) $E\tau_n < \infty$ and $E\sigma_n^i < \infty$ for all $i = 1, \ldots, s$.

(iii) For each i, j there is a sequence $i = i_0, i_1, \ldots, i_k = j$ (for some finite $k \in \{1, 2, \ldots\}$) such that $P(\{i_{m-1}, i_m\} \subseteq \pi_n) > 0$ for all $m = 1, \ldots, k$. (For instance, if π_n is equal to the whole set $\{1, \ldots, s\}$ with positive probability, then this last requirement is satisfied.)

Some remarks about these assumptions: The first two are natural and we have nothing to add. The last one is a kind of "irreducibility" assumption in the sense that if we define a graph with a vertex set $\{1, \ldots, s\}$ and an edge from i to j whenever $P(\{i, j\} \subseteq \pi_0) > 0$, then **(iii)** says that this graph is strongly connected. This assumption will be later used in the proofs of Lemmas 3 and 4.

The problem that we want to analyze here is again the form of the settling time for the system. In other words, does there exist a constant α such that (2.2) holds? First let us see what d_N is. Let $W_n(N)$ be the solution of (3.1) with initial condition $W_0(N) = N.a$, where $a = (a^1, \ldots, a^s) \in R_+^s$ with $a^i > 0$ for all i. When there is no possibility of confusion we will denote $W_n(N)$ simply by W_n. Assuming that there is a steady-state, $d_N(n)$ is defined as the total variation distance between the steady-state distribution and the distribution of $W_n(N)$.

But first, let us examine the existence of such a steady-state. We will begin by writing (3.1) in a more convenient form. Let us introduce some extra notation: Define the random subsets $\phi_n(i)$, for $i = 1, \ldots, s$, of $\{1, \ldots, s\}$ by

$$\phi_n(i) = \begin{cases} \pi_n & \text{if } i \in \pi_n \\ \{i\} & \text{otherwise.} \end{cases}$$

Define the random variables ξ_n^i, for $i = 1, \ldots, s$, by

$$\xi_n^i = \begin{cases} \sigma_n^i - \tau_n & \text{if } i \in \pi_n \\ -\tau_n & \text{otherwise.} \end{cases}$$

Then (3.1) takes the form

$$W_{n+1}^i = \max_{j \in \phi_n(i)} (W_n^j + \xi_n^j)^+. \tag{3.2}$$

Observe that the random sequence $\{[\phi_n(i), \xi_n^i; 1 \leq i \leq s]\}_{n \in Z}$ is still stationary and ergodic under P. Finally, for $m < n$, define the following quantities

$$S_{m,n}(i_m, \ldots, i_{n-1}) = \xi_m^{i_m} + \xi_{m+1}^{i_{m+1}} + \ldots + \xi_{n-1}^{i_{n-1}},$$

$$U_{m,n} = \max_{1 \leq i \leq s} \max S_{m,n}(i_m, \ldots, i_{n-1}),$$

The second maximum in the latter is taken over all indices $i_m, \ldots, i_{n-1}$ such that $i_m \in \phi_m(i_{m+1}), \ldots, i_{n-1} \in \phi_{n-1}(i)$. This will save some space. It is essential to notice this convention carefully in order to avoid possible confusion later in the paper. These quantities appear when one writes down the solution of the recursion (3.2). See also (3.4) below.

Observe now that the process $U_{m,n}$ is subadditive:

$$l < m < n \Rightarrow U_{l,n} \leq U_{l,m} + U_{m,n}.$$

This inequality holds pathwise and is easy to deduce from the very definition of $U_{m,n}$. Furthermore, $U_{m,n}$ is stationary in the following sense:

$$\{U_{m,n}, m \leq n\} = \{U_{m+k,n+k}, m \leq n\} \text{ in distribution for each integer } k.$$

Hence, by Kingman's theorem (see [6]), the linear rate of growth of $U_{m,n}$, either as a function of m or as a function of n, exists, and is a deterministic constant that will be denoted by γ:

$$\lim_{n\to\infty} \frac{U_{m,n}}{n} = \gamma; \lim_{m\to-\infty} \frac{U_{m,n}}{-m} = \gamma, \text{ a.s.} \tag{3.3}$$

We are now ready to prove the following theorem, concerning the existence of the steady-state:

Theorem 2 *If $\gamma < 0$ then there is a steady-state, in the sense that there is a finite, stationary and ergodic process that satisfies (3.2).*

Proof The idea of the proof is classical. It is based on the fact that the right-hand-side of (3.2) is non-decreasing in W_n^j. Our proof follows closely the proof of Baccelli and Liu [4].

For $m \leq n$, let $W_{m,n}$ be the R_+^s-valued process that, as function of n, satisfies (3.1) (or, equivalently, (3.2)) with initial condition at time $n = m$ zero: $W_{m,m} = 0$. Observe that $W_{m,n}^i \leq W_{m-1,n}^i$ for all m, n and i. Define

$$\tilde{W}_n^i = \lim_{m\to-\infty} W_{m,n}^i.$$

It is easy to see that the latter (i) satisfies the recursion (3.2) (by its definition and the fact that $n \mapsto W_{m,n}$ also satisfies (3.2)), and (ii) that is stationary and ergodic. We only need to prove then that $\tilde{W}_n^i$ is finite with probability one. To this end, observe, after some algebra, that

$$\max_{1\leq i\leq s} W_{m,n}^i = U_{m,n} \vee U_{m+1,n} \vee \ldots \vee U_{n-1,n} \vee 0.$$

Since (3.3) holds and since $\gamma < 0$, it follows that

$$\sup_{m\leq n} \max_{1\leq i\leq s} W_{m,n}^i < \infty, \text{ a.s.}$$

The existence of the steady-state has been shown by construction. □

The rest of this section is devoted to the investigation of the cutoff phenomenon. We will throughout assume that $\gamma < 0$. The main result is Theorem 3 below. Before proceeding to it we need some technical lemmata.

First let

$$L_N = \inf\{n \geq 0 : W_n(N) = W_n(0)\}.$$

This is the first time that the process $W_n(N)$ meets the process $W_n(0)$. A priori, this time might be infinite with positive probability. Among other things, the next lemma shows that this is not the case, as long as $\gamma < 0$.

Lemma 2 *For any N there is a finite sequence of indices $\{k_n; 0 \leq n < L_N\}$, such that $\phi_{n-1}(k_n) \ni k_{n-1}, \ldots, \phi_0(k_1) \ni k_0$ and $W_n^{k_n} = W_0^{k_0} + S_{0,n}(k_0, \ldots, k_{n-1})$. Furthermore, $L_N < \infty$ with probability one.*

Proof Let first $n = L_N - 1$. Then there is an index $k_n \in \{1, \ldots, s\}$ such that $W_n^{k_n} > W_n^{k_n}(0)$, by the very definition of L_N. But this means that $W_n^{k_n}$ is not zero and hence, from (3.2), there is a previous index $k_{n-1} \in \phi_{n-1}(k_n)$ such that $W_n^{k_n} = W_{n-1}^{k_{n-1}} + \xi_{n-1}^{k_{n-1}}$. It is easy to see that $W_{n-1}^{k_{n-1}} > W_{n-1}^{k_{n-1}}(0)$, as well. Propagating the argument (backwards) we prove the first part of the theorem.
In particular, we showed that

$$\begin{aligned} P(L_N > n) &\leq P(W_n^{k_n} = W_0^{k_0} + S_{0,n}(k_0, \ldots, k_{n-1})) \\ &= P(W_n^{k_n}/n = W_0^{k_0}/n + S_{0,n}(k_0, \ldots, k_{n-1})/n), \end{aligned}$$

which implies that

$$\limsup_{n\to\infty} P(L_N > n) \leq P(\limsup_{n\to\infty} W_n^{k_n}/n = \gamma).$$

But $\gamma < 0$ and hence the latter probability is zero. □

Note that the indices k_n depend on N, but we omit to write it explicitly for simplicity reasons.

This lemma showed that as long as n is before the first time that $W_n(N)$ meets with $W_n(0)$, there is some component of the vector $W_n(N)$ that can be expressed in terms of some component $W_0^{k_0}$ of the initial condition without having to use the operator $(\ )^+$ of the equation (3.2). The next lemma is concerned with the behavior of the index k_0 (recall that it depends on N) as N tends to ∞. It is shown that, for all realizations of the process, if N is suitably large, then k_0 is equal to the index that achieves the maximum of W_0^i over $1 \leq i \leq s$.

We should also mention at this point that the proof of Lemma 2 also works for showing that any process W_n that satisfies (3.2) (starting from a finite random initial condition W_0) couples with the stationary process $\tilde{W}_n$, that was constructed in Theorem 2, in finite time.

Recalling that $W_0^i = N \cdot a^i$, by definition, we will show the following:

Lemma 3 *Let* $\beta = \max\{a^1, \ldots, a^s\}$. *Then* $\lim_{N\to\infty} a^{k_0} = \beta$.

Proof We start with an explicit expression for $W_n^{k_n}$. This can be obtained from the recursion (3.2):

$$W_n^{k_n} = \max\{W_0^{j_0} + S_{0,n}(j_0, \ldots, j_{n-1})\} \vee \max\{S_{1,n}(j_1, \ldots, j_{n-1})\} \vee \ldots \vee 0. \tag{3.4}$$

Again, we employ our space-saver convention that the first maximum is taken over all $j_0, \ldots, j_{n-1}$ such that $j_0 \in \phi_0(j_1), \ldots, j_{n-1} \in \phi_{n-1}(k_n)$, the second one is taken over all $j_1, \ldots, j_{n-1}$ such that $j_1 \in \phi_1(j_2), \ldots, j_{n-1} \in \phi_{n-1}(k_n)$, and so forth and so on. (It should be observed at this point that the index k_n that appears explicitly on the l.h.s. of (3.4) also appears on the r.h.s. implicitly under the first max.) Fix now n. As $L_N \to \infty$, we can choose an N so that $L_N > n$. Then

$$W_0^{k_0} + S_{0,n}(k_0, \ldots, k_{n-1}) \geq \max\{W_0^{j_0} + S_{0,n}(j_0, \ldots, j_{n-1})\}, \tag{3.5}$$

from Lemma 2 and eq. (3.4) above. Dividing by N and letting N tend to ∞ we get

$$\liminf_{N\to\infty} a^{k_0} \geq \liminf_{N\to\infty} \max_{j_{n-1}\in\phi_{n-1}(k_n)} \ldots \max_{j_0\in\phi_0(j_1)} \{a^{j_0}\}. \tag{3.6}$$

For any i, j let $A_n^{i,j}$ be the event

$$A_n^{i,j} = \{i \in \phi_n(j)\}.$$

If i, j are such that $P(\{i, j\} \subseteq \pi_0) > 0$ then $P(A_n^{i,j}) > 0$ and hence $P(A_n^{i,j}, \text{infinitely often}) = 1$, by ergodicity. Otherwise, for arbitrary i, j we can find $i = i_0, \ldots, i_k = j$ for some finite $k \in \{1, 2, \ldots\}$, such that

$$P(\{i_{m-1}, i_m\} \subseteq \pi_0) > 0, \text{ for all } m = 1, \ldots, k.$$

(This is due to assumption **(iii)**.) Hence

$$P(A_n^{i_{m-1},i_m}, \text{infinitely often}) = 1, \text{ for all } m = 1,\ldots,k. \tag{3.7}$$

This, together with the fact that $i \in \phi_n(i)$ (by definition), shows that the set of j_0's under the maximum of (3.6) is eventually equal to $\{1,\ldots,s\}$. That is,

$$\{j_0 : \exists j_1,\ldots,j_{n-1} \text{ s.t. } j_0 \in \phi_0(j_1),\ldots,j_{n-1} \in \phi_{n-1}(k_n)\} = \{1,\ldots,s\}.$$

(To give an intuitive explanation of this, let us say that the points reached one step backwards at time n by some given node j are those points in the set $\phi_n(j)$; those reached in 2 steps backwards are the points k with $k \in \phi_{n-1}(i)$, for some $i \in \phi_n(k)$; and so forth and so on... These sets are non-decreasing (because i is always in $\phi(i)$) and they eventually include any node in $\{1,\ldots,s\}$ due to (3.7).)

Hence, the right hand side of (3.6) is equal to β. But, on the other hand, $a^{k_0} \leq \beta$. We conclude that

$$\lim_{N\to\infty} a^{k_0} = \beta.$$

This finishes the proof of Lemma 3. □

Our final lemma in this section has the same goal as Lemma 1, namely it shows that the coupling time L_N between $W_n(N)$ and $W_n(0)$ has a linear rate of growth as $N \to \infty$.

Lemma 4 *The coupling time L_N of $\{W_n(N)\}$ with $\{W_n(0)\}$ satisfies*

$$\lim_{N\to\infty} \frac{L_N}{N} = \frac{\beta}{|\gamma|}.$$

Proof First observe that $\lim_{N\to\infty} L_N = \infty$. Choose a subsequence $\{n_N\}$ such that $n_N \to \infty$ as $N \to \infty$ and $n_N < L_N$ for all N. By Lemma 2 we have

$$W_{n_N}^{k_{n_N}} = a^{k_0} N + S_{0,n_N} \tag{3.8}$$

for some sequence of indices $k_0, k_1, \ldots, k_{L_N-1}$. Here S_{0,n_N} is an abbreviation for $S_{0,n_N}(k_0,\ldots,k_{n_{N-1}})$. Proceeding as in the proof of Lemma 3 we get (by comparing the expression (3.8) with (3.4))

$$\max S_{0,n_N}(k_0, j_1, \ldots, j_{n_N-1}) \leq S_{0,n_N} \leq U_{0,n_N}, \tag{3.9}$$

(the inequality on the left is actually an equality) where the maximum is taken over the set $j_1 \in \phi_1(j_2),\ldots,j_{n_N-1} \in \phi_{n_N-1}(k_{n_N})$. The term on the right satisfies

$$\lim_{N\to\infty} \frac{U_{0,n_N}}{n_N} = \gamma,$$

as in (3.3). It is not difficult to see that the same is true for the term on the left:

$$\lim_{N\to\infty} \frac{1}{n_N} \max S_{0,n_N}(k_0, j_1, \ldots, j_{n_N-1}) = \gamma \tag{3.10}$$

To see this, we just have to show that the quantity $Z_n^k := \max_{j_1\in\phi_1(j_2)} \ldots \max_{j_{n-1}\in\phi_{n-1}(k)} \{\xi_0^{k_0} + \xi_1^{j_1} + \ldots + \xi_{n-1}^{j_{n-1}}\}$ satisfies $\lim Z_n^k/n = \gamma$ as $n \to \infty$. Observe that

$$Z_n^k = \max_{j_{n-1}\in\phi_{n-1}(k)} (Z_{n-1}^{j_{n-1}} + \xi_{n-1}^{j_{n-1}}) = \max_{j_{n-1}\in\phi_{n-1}(k)} \max_{j_{n-2}\in\phi_{n-2}(j_{n-1})} (Z_{n-2}^{j_{n-2}} + \xi_{n-2}^{j_{n-2}} + \xi_{n-1}^{j_{n-1}}) = \ldots$$

We continue the iteration until the first time, say, m, such that the index j_{n-m} below the nested maxima will range over the whole set $\{1,\ldots,s\}$. This is true for sufficiently large n because of our assumption **(iii)** (see also the proof of Lemma 3). Noting that m (which depends on n) is of order $o(n)$, by stationarity, we readily conclude that $\lim Z_n^k/n = \lim U_{0,n-m}/n = \gamma$, i.e., its rate of growth does not depend on k. Hence (3.10) is true.

Looking now at inequality (3.9) we conclude that, along the subsequence $\{n_N\}$,

$$S_{0,n_N}/n_N \to \gamma. \tag{3.11}$$

Another limit that we need in the sequel is

$$\lim_{n\to\infty} \frac{1}{n} W_n(0) = 0. \tag{3.12}$$

This follows easily from the fact that $W_n(0)$ converges to $\tilde{W}$ and hence the rate at which work exits the system is the same as the rate that work enters the system.

From the definition of L_N and the recursion (3.2) we get

$$W_{L_N}^i(0) = W_{L_N}^i = \max_{j\in\phi_{L_N-1}(i)} (W_{L_N-1}^j + \xi_{L_N-1}^j)^+. \tag{3.13}$$

Combining (3.13) with (3.8) we get

$$W_{L_N}^i(0) \geq a^{k_0} N + S_{0,L_N-1} + \xi_{L_N-1}^{k_{L_N-1}}.$$

Divide by N, use (3.10), (3.12) and Lemma 3 to get

$$\liminf_{N\to\infty} \frac{L_N}{N} \geq \frac{\beta}{|\gamma|}. \tag{3.14}$$

On the other hand,

$$W_{L_N-1}^{k_{L_N-1}} \geq W_{L_N-1}^{k_{L_N-1}}(0) \geq 0.$$

This, together with (3.8), implies

$$a^{k_0} N + S_{0,L_N-1} \geq 0.$$

Dividing, as usual, by N and using the result of Lemma 3 again, we get another inequality, for the limsup this time:

$$\limsup_{N\to\infty} \frac{L_N}{N} \leq \frac{\beta}{|\gamma|}. \tag{3.15}$$

Combining (3.14) and (3.15) we obtain the proof of Lemma 4. □

We are now ready to state and prove the main theorem.

Theorem 3 *Provided that the constant γ (defined by (3.3)) is strictly negative, we have*

$$\lim_{N\to\infty} d(W_{Nn}(N), \tilde{W}) = \begin{cases} 1 & \text{if } n < \beta/|\gamma| \\ 0 & \text{if } n > \beta/|\gamma|, \end{cases}$$

where $\tilde{W}$ is the steady-state that was constructed in the proof of Theorem 2.

Proof

Suppose first that $n > \beta/|\gamma|$. Using a triangle inequality, as in (2.3), we get

$$d(W_{Nn}, \tilde{W}_{Nn}) \leq d(W_{Nn}, W_{Nn}(0)) + d(W_{Nn}(0), \tilde{W}_{Nn}).$$

The last term on the right converges to 0, by coupling (Lemma 2). The first term is dominated above by $P(L_N > Nn)$ which converges to 0 also (Lemma 4 and the assumption $n > \beta/|\gamma|$).

Suppose next that $n < \beta/|\gamma|$. We have

$$d(W_{Nn}, \tilde{W}_{Nn}) = d(\tfrac{1}{N} W_{Nn}, \tfrac{1}{N} \tilde{W}_{Nn}).$$

We claim that the right-hand-side converges to 1. From Lemma 2 we have

$$W_n^{k_n} = a^{k_0} N + S_{0,n}(k_0, \ldots, k_{n-1}) \text{ if } n < L_N.$$

Replace n by Nn in the above and divide by N

$$\tfrac{1}{N} W_{Nn}^{k_{Nn}} = a^{k_0} + \tfrac{1}{N} S_{0,Nn}(k_0, \ldots, k_{Nn-1}) \text{ if } n < L_N/N.$$

Now let N go to ∞ and use Lemma 3, Lemma 4, (3.11), and the assumption $n < \beta/|\gamma|$ to get

$$\lim_{N \to \infty} \tfrac{1}{N} W_{Nn}^{k_{Nn}} = \beta + \gamma n. \tag{3.16}$$

On the other hand,

$$\lim_{N \to \infty} \tfrac{1}{N} \tilde{W}_{Nn} = 0, \tag{3.17}$$

for the same reason that (3.12) holds. Consider now the obvious inequality

$$d(\tfrac{1}{N} W_{Nn}, \tfrac{1}{N} \tilde{W}_{Nn}) \geq d(\tfrac{1}{N} ||W_{Nn}||, \tfrac{1}{N} ||\tilde{W}_{Nn}||), \tag{3.18}$$

where $||x|| = (\sum_{i=1}^{s} x_i^2)^{1/2}$. Fix $n < \beta/|\gamma|$ and choose $\epsilon > 0$ such that $\beta + \gamma n > \epsilon$. The right-hand side of (3.18) is dominated below by

$$\left| P(\tfrac{1}{N} \, ||W_{Nn}|| < \epsilon) - P(\tfrac{1}{N} \, ||\tilde{W}_{Nn}|| < \epsilon) \right|.$$

But (3.16) implies that $P(||W_{Nn}||/N < \epsilon, \text{ infinitely often}) = 0$, while (3.17) implies that $P(||\tilde{W}_{Nn}||/N < \epsilon) \to 1$. Hence (3.18) converges to 1. This concludes the proof of Theorem 3. □

Remarks

The settling time for the process W_n has been shown to be proportional to $1/|\gamma|$ (normalize by setting $\beta = 1$). The problem is, as usual with the subadditive ergodic theorems, the actual value of the constant γ which was defined only as the rate of growth of $U_{m,n}$. There are almost no systems where one knows γ exactly, unless $U_{m,n}$ is additive, instead of being subadditive (in which case γ equals $EU_{0,1}$). Observe that, if $s = 1$, the system of Section 3 is actually the G/G/1 queue of Section 2. In this case, γ is trivially equal to $\mu^{-1} - \lambda^{-1}$. This leads to a settling time proportional to $\lambda\mu/(\mu - \lambda)$, in agreement with the remarks at the end of Section 2.

4 Conclusions

We have investigated the settling time for two non-Markovian systems: The single-server queue and the multi-server queue with locking with stationary and ergodic inputs. The essential idea in our development has been the coupling between the transient process and the stationary one. It has been shown that the settling time is proportional to the (linear) rate of growth of the coupling time when the initial state goes to infinity. It is expected that our methods are also applicable to other systems that admit coupling.

References

[1] Aldous, D. (1983). *Random walks on finite groups and rapidly mixing Markov chains.* Lect. Notes Math., **986**, Springer-Verlag, Berlin.

[2] Anantharam, V. (1988). The settling time of a closed Jackson network. Preprint.

[3] Baccelli, F. (1989). Ergodic theory of stochastic Petri nets. *Ann. Prob.*, to appear.

[4] Baccelli, F. and Liu, Z. (1990). On a class of stochastic recursive sequences arising in queueing theory. *Ann. Prob.*, to appear.

[5] Baccelli, F., Makowski, A. and Shwartz, A. (1989). The fork-join queue and related systems with synchronization constraints: Stochastic ordering, approximations and computable bounds. *Adv. Appl. Prob.*, **21**, 629-660.

[6] Kingman, J.F.C. (1973). Subadditive ergodic theory. *Ann. Prob.*, **1**, 883-909.

[7] Konstantopoulos, P. and Walrand, J. (1989). Stationarity and stability of fork-join networks. *J. Appl. Prob.*, **26**, 604-614.

[8] Loynes, R.M. (1962). The stability of a queue with non-independent inter-arrival and service times. Proc. Cambridge Phil. Soc., **58**, 497-520.

[9] Stamoulis G.D. and Tsitsiklis, J.N. (1989). On the settling time of the congested G/G/1 queue. Preprint.

Imprimé en France
par
. l'Institut National de Recherche en Informatique et en Automatique .

ISSN 0249-6399