INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Optimal Routing in Deterministic Queues in Tandem

Bruno Gaujal — Emmanuel Hyon

N° 4393

Mars 2002

THÈME 1

ISSN 0249-6399 ISRN INRIA/RR--4393--FR+ENG

Optimal Routing in Deterministic Queues in Tandem

Bruno Gaujal* , Emmanuel Hyon †

Thème 1 — R´eseaux et systèmes
Projet TRIO

Rapport de recherche n° 4393 — Mars 2002 — 29 pages

Abstract: In this paper we address the problem of routing a stream of customers in two parallel networks of queues in tandem with deterministic service times in order to minimize the average response time of the whole system. We show that the optimal routing is a Sturmian word which density depends on the decomposition in continuous fraction of the maximum service time on each route. In order to do this we study the output process of deterministic queues when the input process is Sturmian.

Key-words: Routing, Deterministic queues, Sturmian words, Continued Fraction, Networks of tandem queues.

* INRIA/LIP, ENS Lyon, 46, All´ee d'Italie, F-69364 Lyon, France. E-mail Bruno.Gaujal@ens-lyon.fr
† LORIA UMR 7513, 615 route du jardin botanique, BP 101, F-54606 Villers-les-Nancy, France. E-Mail Emmanuel.Hyon@loria.fr

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,
615, rue du Jardin Botanique, BP 101, 54602 Villers-Lès-Nancy (France)
Téléphone : +33 3 83 59 30 00 — Télécopie : +33 3 83 27 83 19

Politique de routage optimal dans des réseaux de files d'attentes déterministes en tandem

Résumé : Dans ce papier nous nous int´eressons au routage de flux de clients dans deux r´eseaux parallèles de files d'attente d´eterministes en s´eries. Notre but est de minimiser le temps de r´eponse moyen du système. Nous montrons que la politique optimale est un mot de Sturm dont la densit´e d´epend de la d´ecomposition en fractions continues du temps de service maximal sur chaque route. Dans ce but nous consacrons une large part de notre ´etude au processus de sortie d'une file d´eterministe dont le processus d'entr´ee est sturmien.

Mots-clés : Routage, Files d'attente d´eterministes, Mots Sturmien, Fractions Continues, R´eseaux de files en tandem.

1 Introduction

In this paper the following problem is considered. Someone wants to send a stream of customers through a network to a distant destination where two different routes (that do not share any node) are available. This problem occurs for instance under MPLS-OMP routing strategies, where several tunnels (or virtual circuits) are constructed for each connection [18, 19]. The optimal policy should make use of the two resources (or tunnels) in order to minimize some cost function of the average response time of the network. Rather surprisingly, the optimal policy is highly discontinuous with the parameters of the problem and has a fractal behavior when the total load is close to one. We show how to compute exactly the optimal policy when the service time in each node is deterministic. This policy is characterized by a Sturmian sequence which density can be computed from the decomposition in ceiled continuous fractions of the service times of all the nodes involved. The case with two queues in tandem is studied in full detail.

A similar problem where the network of queues in tandem is replaced by single queues is addressed in [9]. The main difference here is the need to study the output process, which is an essential component to the global study of a network of queues. Here we focus on the output process of a deterministic queue when the input is Sturmian.

We also show how to compute the average waiting time of the packets in any queue when the input process in the system is sturmian. Another novelty is the characterization of the stream of customers after a large number of deterministic queues. This stream converges to a normalized stream with inter arrival times bounded from below by the maximal service time encountered. This is to be compared with the case of exponential service times ([4]) where the stream converges to a Poisson process. In a more general case with independent service times, it has been showed in [16] that a fixed point of the output stream exists.

More precisely we are interested in the routing in two parallel systems Q^1 and Q^2, each one composed by $./D/1/\infty/FIFO$ queues in tandem. When they arrive the customers are routed in one of the two system. The problem is to find the optimal routing policy minimizing the weighted average response time of a customer in the system. The optimal policy will be given as a binary infinite word m where $m(i) = 1$ (resp. $m(i) = 0$) means that the ith customer is sent to system Q^1 (resp. Q^2).

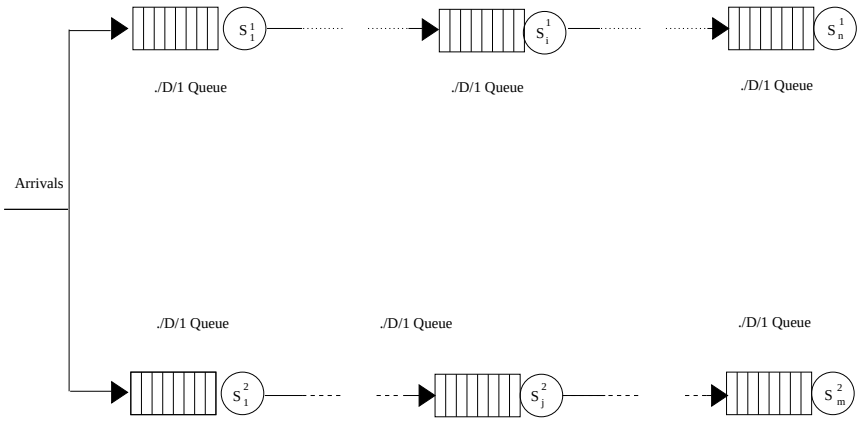

Figure 1: Admission control in N queues in tandem

Figure 1 displays such a system. All the queues have constant service times denoted by S_i^j for the ith queue of the jth system. The time unit is chosen such that all the inter-arrival times before the routing are equal to one.

The paper is structured as follows. The second section gives several results concerning deterministic queues in tandem. Section 3 introduced mechanical words and ceiled continuous fractions. It also shows how a system of queues in tandem behaves when the input is such a sequence. Section 4 shows how the previous computations can be used to derive the optimal routing policy in two parallel networks of queues in tandem whereas Section 5 presents the exhaustive analysis for some examples.

2 Deterministic queues in tandem

In this part we study a network made of deterministic FIFO queues with infinite buffers in tandem. Let Q a network made of N queues in tandem. Let us denote by $T_i(n)_{n\in\mathbb{N},i\in\{1,\cdots,N\}}$ with $n \geq 1$ the arrival process in the ith queue ($T_i(n)$ is the arrival time of the n-th customer in queue i). The input process of the first queue is the input process of the whole system: $T_1 = T$.

All the queues have constant service times denoted by S_i for the ith queue.

2.1 Average response time

We also recall now the stability condition in a given queue with service time S:

$$\underline{\lim}_{n\to\infty} n/T(n) \leq \frac{1}{S}. \tag{1}$$

For a system of N queues, the stability condition is

$$\underline{\lim}_{n\to\infty} n/T(n) \leq \frac{1}{\overline{S}}, \tag{2}$$

where $\overline{S}$ is the maximal service time among the N queues *i.e.* $\overline{S} = \max_{1\leq j\leq N} S_j$.

The workload $w(t)$ which denotes the amount of service (in time units) remaining to be done by the server at epoch t is given by the Lindley's formula for general G/G/1 queues:

$$w(t) = \Big(w(T(n-1)_-) + S - (t - T(n-1))\Big)^+, \text{ for } T(n-1) \leq t < T(n),$$

where $x_- = \lim_{y\uparrow x} y$ and $(\cdot)^+ = \max(\cdot, 0)$. The waiting time of the kth customer in the first queue is given by

$$W_1(k) = w(T(k)_-).$$

As for N queues in tandem, the waiting time of the kth customer in the ith queue will be denoted by $W_i(k)$.

Definition 1 (Average sojourn time). *The average sojourn time in queue i is given by*

$$V_i = \overline{\lim}_{n\to\infty} \frac{1}{n} \sum_{k=0}^{k=n} V_i(k), \tag{3}$$

where $V_i(k)$ is the sojourn time (waiting time plus service time) of the kth customer in the queue i.

Under rather general assumptions on the ergodicity of the input process and the service times, the limit sup above is often a limit.

Definition 2 (Average response time). *The response time for the kth customer of the first n queues is $R_n(k) = \sum_{i=1}^n V_i(k)$. As for the average response time, $R_N = \sum_{i=1}^N V_i$.*

2.2 Output process

In order to study systems composed by series of queues, we need to study the output process of a deterministic queue for a given input process.

Proposition 3. *The output process of a ./D/1 queue with an initial load equal to w_0 is $\{O(k)\}_{k\geq 1}$ where $T(0) := w_0 - S$ and for all $k \geq 1$,*

$$O(k) = \max_{0\leq i\leq k}(T(i) + (k-i+1)S). \tag{4}$$

Proof. We have $O(1) = (T(1)+S) \vee w_0 + 2S = \max_{0\leq i\leq 1}(T(i)+(k-i+1)S)$.

Assume Formula (4) is true up to $k-1$. The kth departure $O(k)$ satisfies $O(k) = T(k) + W_k + S$, hence with the Lindley's Formula and the induction assumption, it comes

$$\begin{aligned} O(k) &= T(k) + S + \max(W_{k-1} + S - (T(k) - T(k-1)), 0) \\ &= \max(O(k-1) + S, T(k) + S) \\ &= \max(\max_{0\leq i\leq k-1}(T(i) + (k-i)S) + S, T(k) + S) \\ &= \max_{0\leq i\leq k}(T(i) + (k-i+1)S). \end{aligned}$$

□

In tandem queues, the output process of queue i is the input process of queue $i+1$: $O_i(k) = T_{i+1}(k)$ for all $k \geq 0$.

Proposition 4 (General output process). *The output process O_N of the system is given by*

$$O_N(k) = \Sigma_{j=1}^{N} S_j + \max_{0\leq l\leq k}(T(l) + \overline{S}(k-l)). \tag{5}$$

Proof. This proof is made for a system composed of two queues the generalization to an arbitrary number of queues follows easily. By definition $T_2(k) = O_1(k)$.

From Proposition 3 it comes

$$\begin{aligned} O_2(k) &= \max_{0\leq v\leq k}\big(O_1(v) + (k+1-v)S_2\big) \\ &= \max_{0\leq v\leq k}\big(\max_{0\leq u\leq v}(T(u) + (v+1-u)S_1) + (k+1-v)S_2\big) \\ &= \max_{0\leq u\leq k}\Big(T(u) + (1-u)S_1 + (k+1)S_2 + \max_{u\leq v\leq k}\big(v(S_1 - S_2)\big)\Big). \end{aligned}$$

When $S_1 \leq S_2$, $\max_{u\leq v\leq k}\big(v(S_1-S_2)\big)$ is reached when $v = u$ and

$$O_2(k) = S_1 + \max_{0\leq u\leq k}\big(T(u) + (k+1-u)S_2\big).$$

When $S_1 \geq S_2$, the maximum is obtained when $v = k$ and

$$O_2(k) = S_2 + \max_{0\leq u\leq k}\big(T(u) + (k+1-u)S_1\big).$$

□

2.3 Commutative properties

Using the fact that the output process at queue n only depends on the initial input and the maximal service time in all the queues up to queue n, one can get commutative properties of the response time of the system.

Let $\overline{S}_n$ be the maximal service time among the first n queues *i.e.* $\overline{S}_n = \max_{1\leq j\leq n} S_j$. Let $R_n(k)$ be the response time for the k-th customer of a system made of n queues in tandem. Let us modify the system by permuting the queues according to an arbitrary permutation σ. We denote by $R_{\sigma,n}(k)$ the response time of the new system.

We have

Lemma 5. *For any permutation σ of $\{1.\cdots,n\}$. $R_n(k) = R_{\sigma,n}(k)$.*

Proof. By definition, $R_n(k) = O_n(k) - T(k)$. Using Proposition 4,

$$\begin{aligned} R_n(k) &= \Sigma_{j=1}^n S_j + \max_{0\leq v\leq k} (T(v) + \overline{S}_n(k-v)) - T(k)) \\ &= R_{\sigma,n}(k). \end{aligned}$$

The last equality comes from the fact that the formula for $O_n(k)$ does not depend on the order of the queues. □

Corollary 6. *If the sum of the service times is fixed to Σ, then the response time is minimal when all the service times are equal.*

Proof. Let σ be a permutation of the service times such that $S_{\sigma(1)} \geq \cdots \geq S_{\sigma(n)}$. Then $R_n(k) = R_{\sigma,n}(k) = \hat{W}_1(k) + \Sigma$, where $\hat{W}_1(k)$ is the waiting time of the kth customer in the first queue of the modified system with the largest service time in the first queue. The quantity Σ being fixed, $\hat{W}_1(k)$ is minimal when the largest service time is as small as possible. This is achieved when all service times are equal. □

2.4 Multimodularity

Multimodularity is a kind of convex property of discrete functions first defined in [12] which is useful for optimization (see [1]).

Here, we consider a slotted bursty input process with arrival *opportunities* $D(k), k = 1,\cdots$. The number of arrivals $a_k \in \mathbb{N}$ is the number of customers arriving at time $D(k)$. Note that a_k can also be 0. This means that no customer arrives at time $D(k)$. We denote by b_h the number of time slots between arrival $h-1$ and arrival h.

Hence, the arrival of the kth customer, $T(k) = \sum_{i=1}^{b_1+\cdots+b_k} D(i)$. The gap process b_k defines perfectly the arrival process as soon as the arrival opportunity process is given.

Let $e_i \in \mathbb{N}^m$, $i = 1,\cdots,m$ denote the vector having all entries zero except for a 1 in its ith entry. Define $s_i = e_i - e_{i+1}$, $i = 1,\cdots,m-1$ and $s_0 = -e_1, s_m = e_m$.

Let $F = \{s_0, s_2, \cdots, s_m\}$, F will be called a multimodular base of $\mathbb{Z}^m$.

Definition 7 (Multimodular function). *A real-valued function $f : \mathbb{Z}^m \longrightarrow \mathbb{R}$ is multimodular with respect to F if for all $x \in \mathbb{Z}^m$, v and w in F, $v \neq w$, the following holds:*

$$f(x+v) + f(x+w) \geq f(x) + f(x+v+w). \tag{6}$$

For a system of N queues in tandem with gap input sequence $(b_1, \cdots, b_k, \cdots)$, we consider the sojourn time of the k-th customer in queue i as a function of $(b_1, \cdots, b_k)$. With a slight abuse of the notations defined previously, $V_i(b_1, \cdots, b_k) := V_i(k)$.

Theorem 8. *Let Q be a system of N empty queues in tandem with gap input sequence $(b_i)_{i\in\mathbb{N}}$. Then the sojourn time of the k-th customer in queue i, $V_i(b_1, \cdots, b_k)$ is a multimodular function in $(b_1, \cdots, b_k)$.*

Proof. For the first queue, the proof can be found in [1]. In the following, we denote by B the vector $(b_1, \cdots, b_k)$. For the n-th queue, we construct the arrival opportunities process $D_i(k)$ up to k as the ordered sequence of all the points in

$$\{\sum_{i=1}^{n} S_i + \max_{j=1}^{k}\Big(\sum_{h=1}^{\beta_1+\cdots+\beta_j} D(h) + (n-j)\overline{S}\Big), \beta_i \in \{b_i, b_i+1\}, \beta_1 + \cdots + \beta_n \leq b_1 + \cdots + b_n + 2\}.$$

Then, the gap input process in queue i corresponding to the global input B is denoted $C = (c_1, \cdots, c_k)$. Now, let us consider the gap sequence for queue i corresponding to a modified gap input $B^1 = B + s_x$. Since the order of the customers in $B + s_x$ is the same as with B and since the queue is FIFO, then the order of the customers for the modified input process in queue i is the same as with the original case.

Examining the modifications of C induced by adding s_i to B we get

$$\begin{aligned} b_1^1 + \cdots + b_j^1 &= b_1^1 + \cdots + b_j^1 \quad \text{if } j \neq x, \\ b_1^1 + \cdots + b_x^1 &= b_1^1 + \cdots + b_x^1 + 1 \quad \text{otherwise.} \end{aligned}$$

Since customers cannot overtake each other, this means that $C^1 \in \{C, C + s_x\}$.

The same kind of modification occurs when two different customers are shifted ($B + s_x + s_y$) inducing a gap process $C^{12} \in \{C, C + s_x, C + s_y, C + s_y + s_x\}$.

Consider queue 1 with a modified service time $S_1 = S_i$, with input opportunities $D_i(k)$ and initial gap process C, we get $V_i(B + s_x) = V_1(C^1)$, $V_i(B + s_y) = V_1(C^2)$ and $V_i(B + s_x + s_y) = V_1(C^{12})$. Using the multimodularity of queue 1 we get $V_i(B+s_x)+V_i(B+s_y) \geq V_i(B+s_x+s_y)+V_i(B)$. □

3 Sturmian inputs

In the rest of the paper, we will focus on special input sequences, namely Sturmian words, which will happen to be closely related to the optimal routing policies. In other several papers [1, 2, 3, 9], it was shown that Sturmian sequences have a particular interest for the optimization when used an input sequences in queues.

The customers arrive according to an exogenous process $\{T(n)\}_{n\in\mathbb{N}}$ ($T(n)$ denote the time of arrival of the nth customer). This process is deterministic and follows an *upper mechanical word (or a Sturmian word) with slope* α if $T(n) = \lfloor n/\alpha \rfloor$.

Note that all the arrival times are integer valued. Therefore, another way (less compact) to describe the arrival sequence is by a binary word m such that $m(i) = 1$ if i is an arrival (*i.e.* belongs to the set $\{T(n)\}_{n\in\mathbb{N}}$) and $m(i) = 0$ otherwise.

3.1 Mechanical sequences and ceiled continued fractions

Let $A = \{0, 1\}$ be the binary alphabet. The free monoid A^* is the set of the finite *words* on A. An infinite word is an element of $A^{\mathbb{N}}$. The empty word is denote ε

Definition 9 (Slope). *The slope of a finite nonempty word m is the number :*

$$\alpha(m) = \frac{|m|_1}{|m|},$$

where $|m|_1$ is the number of letters equal to one in m and $|m|$ is the length of m.

Let $m[n], \forall\, n \geq 1$ be the prefix of length n of an infinite word m. If the sequence $\alpha(m[n])$ converges when $n \to \infty$ the limit is called the slope of m.

If m is a word on A (either finite or infinite), a *factor* of m is a word f such that there exists two words u and v (possibly empty) such that $m = u \cdot f \cdot v$.

A word m is factorized in $(f_1, \cdots, f_n)$ if $m = f_{i_1}^{n_1} \cdot f_{i_2}^{n_2} \cdots$. If m is finite, the number of times a factor f_j appears in this factorization is denoted $|m|_{f_j}$. Note that this number depends on the factorization. This is not the number of times the factor f_j appears in m. However in the following no confusion will ever be possible.

For a real number x, we denote by $\lfloor x \rfloor$ (resp. $\lceil x \rceil$) the largest (resp. smallest) integer smaller (resp. larger) than x.

Definition 10 (Mechanical word). *The upper mechanical word with slope α is the infinite word $\overline{m}_\alpha$ where the n^{th} letter, with $n \geq 0$, is :*

$$\overline{m}_\alpha(n) = \lceil (n+1) \times \alpha \rceil - \lceil n \times \alpha \rceil.$$

The lower mechanical word with slope α is the infinite word $\underline{m}_\alpha$ where the n^{th} letter, with $n \geq 0$, is :

$$\underline{m}_\alpha(n) = \lfloor (n+1) \times \alpha \rfloor - \lfloor n \times \alpha \rfloor.$$

If α is a rational number ($\alpha = \frac{p}{q}$), then $\overline{m}_\alpha, \underline{m}_\alpha$ are periodic of period q. If α is an irrational number, then $\overline{m}_\alpha, \underline{m}_\alpha$ are all aperiodic. These results are proved for example in [15].

In the following, by a slight abuse of notation when an infinite word m is periodic we also denote by m its shortest period.

The *ceiled* continued fraction expansion (see for example [7]) of a number α with $0 < \alpha < 1$ is given by :

$$\left\{ \begin{array}{ccccc} \alpha = \frac{1}{a_1 - \alpha_1} & ; & a_1 = \lceil \alpha^{-1} \rceil & ; & \\ \alpha_n = \frac{1}{a_{n+1} - \alpha_{n+1}} & ; & a_{n+1} = \lceil \alpha_n^{-1} \rceil & ; \forall\, n \geq 1 & \end{array} \right\}. \tag{7}$$

A *ceiled* continued fraction expansion of a number α, with $0 < \alpha < 1$, is denoted here by $\langle a_1, a_2, \ldots, a_n \rangle$ when the expansion is finite and by $\langle a_1, a_2, \ldots, a_n, \ldots \rangle$ when the expansion is infinite. We also denote by $\langle a_1, a_2, \ldots, a_n - \alpha_n \rangle$ a partial expansion of order n of α, where the rest at order n is α_n with $0 < \alpha_n < 1$.

Theorem 11 ((x,y)-factor decomposition,[9]). *Let $0 < \alpha < 1$ be such that $\alpha = \langle a_1, a_2, \ldots, a_n, \ldots \rangle$. We define two sequences, $\{x_i(\alpha)\}_{i \geq 0}$ and $\{y(i)(\alpha)\}_{i \geq 0}$, as follow :*

$$\begin{array}{llll} x(0)(\alpha) = 1, & x(i)(\alpha) = x(i-1)(\alpha)\big(y(i-1)(\alpha)\big)^{a_i - 2}, & \text{for } i \geq 1, \\ y(0)(\alpha) = 0, & y(i)(\alpha) = x(i-1)(\alpha)\big(y(i-1)(\alpha)\big)^{a_i - 1}, & \text{for } i \geq 1. \end{array}$$

For all nonnegative i, the upper mechanical word $\overline{m}_\alpha$ can be factorized by using the two factors $x(i)(\alpha)$ and $y(i)(\alpha)$. These two sequences are called (x-y)-factor decomposition sequences associated with the ceiled continued fraction decomposition of α.

When α is rational then the (x-y)-factor decomposition sequence is finite. In this case at the last step, the word $\overline{m}_\alpha$ is in fact equal to the last factor of type y, as it is shown in the example below.

Example 12 (Example of a (x-y) factor decomposition). *Assume that $\alpha = \frac{5}{12}$. The ceiled continued fraction expansion of α obtained with the use of Equations* (7) *is $\langle 3, 2, 3\rangle$ and*

$$\overline{m}_{5/12} = 101010010100.$$

The (x-y)-factor decomposition gives us $x_1 = 10$, $y_1 = 100$, $x_2 = x_1$, $y_2 = x_1 y_1$, et $y_3 = x_2 y_2 y_2$.

$$\overline{m}_{5/12} = \overbrace{\underbrace{10}_{x_2}\underbrace{10}_{y_2}\underbrace{100}_{y_2}}^{y_3}\underbrace{10}_{y_2}\underbrace{100}_{y_2}.$$

Definition 13 (Ceiled convergents). *Let s be a real number such that $0 \leq s \leq 1$ which ceiled continued fraction is $s = \langle a_1, \ldots, a_n, \ldots\rangle$. We then define the sequence of rational numbers $r(n)(s)$ by $r(n)(s) = \langle a_1, a_2, \ldots, a_n\rangle$. These rational numbers are called ceiled convergents of s.*

Since the $r(n)(s)$ are rational numbers they can be written as $r(n)(s) = \frac{p(n)(s)}{q(n)(s)}$, with $p(n)(s) = |\overline{m}_{r(n)(s)}|_1$ and $q(n)(s) = |\overline{m}_{r(n)(s)}|$. The numbers $p(n)(s)$ and $q(n)(s)$ can be computed by

$$\begin{aligned} p(n)(s) &= a_n p(n-1)(s) - p(n-2)(s) \\ q(n)(s) &= a_n q(n-1)(s) - q(n-2)(s), \end{aligned} \tag{8}$$

with $p(0)(s) = 0$, $p(1)(s) = 1$, $q(0)(s) = 1$ and $q(1)(s) = a_1$ as initial conditions.

It should be noticed that the sequence $r(n)(s)$ forms an increasing sequence which goes to s when n goes to infinity. It also could point out that $\overline{m}_{r(n)(s)} = y(n)(s)$.

Lemma 14. *Under the foregoing notation, let α be another real number such that $r(n)(s) \leq \alpha \leq r(n+1)(s)$ then*
i) The ceiled continuous fraction of α has the same n first coefficients as s: $\alpha = \langle a_1, \ldots, a_n - \alpha_n\rangle$.
ii) The word $\overline{m}_\alpha$ can be factorized in $y(n+1)(s)$ and $y(n)(s)$.
iii) The rest α_n satisfies

$$\alpha = \frac{p(n) - \alpha_n p(n-1)}{q(n) - \alpha_n q(n-1)}. \tag{9}$$

Proof.
The proof is postponed in Appendix 7 □

This result about the factorization of $\overline{m}_\alpha$ with $\overline{m}_{r(n)}$ and $\overline{m}_{r(n+1)}$ for any α in $[r(n), r(n+1)]$ will be used several times in the following. Another proof of Lemma 9 *iii* can be found in [7] using combinatorial properties of continued fraction.

In addition to this, we can relate α_n with the ratio between α and $\frac{p(n+1)(s)}{q(n+1)(s)}$. We define the coefficient d in $[0, 1]$ such that

$$\alpha = (1-d)\cdot\frac{p(n)(s)}{q(n)(s)} + d\cdot\frac{p(n+1)(s)}{q(n+1)(s)}.$$

Using Equation (9), one gets

$$d = \frac{\alpha_n q(n+1)(s)}{\alpha_n q(n+1)(s) + (1-\alpha_n a(n+1))q(n)(s)}.$$

This also means that d can be viewed as the ratio of letters in $\overline{m}_\alpha$ belonging to $y(n+1)(s)$ factors in the so called $(y(n+1)(s), y(n)(s))$ factorization of $\overline{m}_\alpha$.

In order to keep things simple we denote in the following $r(n)(s), p(n)(s), q(n)(s)$ by $r(n), p(n)$, $q(n)$ respectively when no confusion is possible.

Some others results about the combinatorial properties of mechanical words can be found in [14, 15] and concerning the continued fractions in [6, 8].

3.2 Output process with a Sturmian input

When the input in the system is Sturmian, then the output process of each queue can be described more precisely. In particular, we can estimate when the maximum in Equations (4) and (5) is reached. Nevertheless it requires to show that the structure of the output process induced by a mechanical input in the first queue is kept after passing through several queues.

We consider the case with N queues with respective service times S_i, $i = 1, \cdots, N$. From now on the ceiled expansion of S_i^{-1} is written as follows $S_i^{-1} = \langle l_i(1), l_i(2), \ldots, l_i(n) - s_i(n)\rangle$.

We associate with each queue the following sequences.

Definition 15. *Let $S_i^{-1} = \langle l_i(1), l_i(2), \ldots, l_i(n) - s_i(n)\rangle$ be the partial ceiled continued fraction expansion of $1/S_i$. We now define the sequence of rational numbers $r_i(n)$ of ceiled convergents computed by*

$$r_i(n) = \langle l_i(1), \ldots, l_i(n)\rangle = \frac{p_i(n)}{q_i(n)}. \tag{10}$$

They are computed according to the recurrence equation and the initial conditions given by Equation (8).

We also define the sequence of terms $d_i(n,1)$ and $d_i(n,2)$ defined by

$$\begin{aligned} d_i(n,1) &= \big(p_i(n) - p_i(n-1)\big)S_i - \big(q_i(n) - q_i(n-1)\big), \\ d_i(n,2) &= p_i(n)S_i - q_i(n). \end{aligned} \tag{11}$$

Let us recall that $\overline{S}_n$ is the largest service time among the first$n$ service times : $\overline{S}_n = \max_{1\le j\le n} S_j$.

Lemma 16 (Output process of ith queue).
-i) If $S_i \le \overline{S}_{i-1}$ then the output process of the ith queue is the output process of the previous queue shifted by S_i.
If $S_i > \overline{S}_{i-1}$ then
-ii) If $\alpha \le \lceil S_i\rceil^{-1}$ then the output process is a mechanical word of slope α shifted by $\Sigma_{j=1}^{i} S_j$.
-iii) If $r_i(n) \le \alpha \le r_i(n+1)$ then the inter-departure times can only take three values: S_i, $S_i + |d_i(n+1,2)|$ and $S_i + |d_i(n,2)|$. The inter-departure times with value $S_i + |d_i(n+1,2)|$ (respectively $S_i + |d_i(n,2)|$) occur between the departure of the last customer of a factor $\overline{m}_{r_i(n+1)}$ (resp. $\overline{m}_{r_i(n)}$) and the departure of next admission.
-iv) If $\alpha = r_i(n)$ then the inter-departure times can only take two values: S_i and $S_i + |d_i(n,2)|$. The inter-departure times with value $S_i + |d_i(n,2)|$ occur after the last departure of a factor $\overline{m}_{r_i(n)}$.
-v) If α is equal to S_i then the queue is fully loaded and the time elapsed between two departures is always equal to S_i.

Proof. -i) From Proposition 4, the output process is given by $\overline{S}_i$. Hence if $S_i \leq \overline{S}_{i-1}$ then $S_i \neq \overline{S}_i$ and the result follows.

When $S_i > \overline{S}_{i-1}$ then the output process depends mainly on S_i.
-ii) The condition $\alpha \leq \lceil S_i \rceil^{-1}$ implies $\frac{j}{\alpha} \geq j\lceil S_i \rceil$ which yields $\lfloor \frac{j}{\alpha} \rfloor \geq j\lceil S_i \rceil \geq jS_i$. This means by Proposition 3 that the departure of the jth customer of the queue i takes place at the epoch $T_i(j) + S_i$ without any condition on j. Since $S_i = \overline{S}_i$ the same result holds for all preceding queues. Henceforth the output process of the ith queue is a mechanical word.

The characterization of the epochs where the maximum of Equation (4) is achieve by the term $T_i(\cdot) + S_i$ can be obtained using the $(x - y)$ factor decomposition. Indeed the only customer which find an empty queue when they arrive are these one which belong to a factor which immediately follows either an $y_i(n + 1)$ or an $y_i(n)$ when the queue is not fully loaded. This shows *iii)* while *iv)* comes from the fact that when $\alpha = r_i(n)$ only the $y(i)(n)$ factors appear in the decomposition of $\overline{m}_\alpha$. When the queue is fully loaded there is not any idle period, this for *v)*.

Let us precise now the proof of the fact that the only customer which find an empty queue are these one which belong to a factor which immediately follows a factor either $y_i(n + 1)$ or $y_i(n)$. For a single queue this is due to Lemma 14 of [10]. Since Lemma 20 shows that the properties of the $(x-y)$ factorization are kept after a passage in a queue provided that the input process is described by this Lemma. Henceforth the output process can be described by this Lemma for all the queues. □

We give now an example of an output process of a single queue. It can be said without loss of generality by Proposition 4 that such a process is general for all queue of a tandem queue network with a Sturmian input.

Example 17 (Output process of a single queue). *In this example we study the output process of a system made by single queue. We explicit this output process when the admission word is a Sturmian word. We assume that the mechanical word has a slope equal to* $5/12$ *hence* $\overline{m}_{5/12} = 101010010100$. *The service time is assumed to be equal to* $S = 9/4$. *The computations of the ceiled convergents of* $4/9$ *by Definition 13 gives us* $r(1) = 1/3$, $r(2) = 2/5$, $r(3) = 3/7$, $r(4) = 4/9$. *First note that we have* $2/5 < 5/12 < 3/7$, $\overline{m}_{2/5} = 10100$ *and* $\overline{m}_{3/7} = 1010100$, *moreover* $\overline{m}_{5/12} = \overline{m}_{3/7}\overline{m}_{2/5} = y(3)y(2)$. *Thus the inter departure process is* $(S, S, S + |d(3, 2)|, S, S + |d(2, 2)|)$, *with* $|d(3, 2)| = 0.25$ *and* $|d(2, 2)| = 0.5$.

On Figure 2 the departures are represented by arrows pointing down, $|d(3, 2)|$ *by* a *and* $|d(2, 2)|$ *by* b.

It could be noticed that the only cases for which the output process is composed by equal inter-departure occur either when the queue is fully loaded ($\alpha = 1/\overline{S}$) or when the mechanical word has a slope α such that $\alpha = \frac{1}{l}$ with $l \in \mathbb{N}$ such that $l = l(1)(\overline{S})$. We can also stress the following degenerated case : if $\overline{S} \leq 1$ then the output process is always a Sturmian word.

Lemma 16 gives corollary below.

Corollary 18 (Periodicity in rational cases). *Let* α *be a rational number with* $\alpha = \frac{p}{q}$, *such that* $r(n) \leq \alpha \leq r(n + 1)$. *Since* q *is the period of the word* $\overline{m}_\alpha$, *then*
- i) The output process is periodic of period q.
- ii) The time between the last departures of two consecutive $\overline{m}_\alpha$ *is* q.

Proof. The proof is made for the output process of a single queue. The generalization comes from Lemma 16.
i) Since the input process is periodic then the sub-words $\overline{m}_{r_1(n)}$ and $\overline{m}_{r_1(n+1)}$ appear periodically and always in the same order in $\overline{m}_\alpha$. Since moreover all the time between two departures are equal except

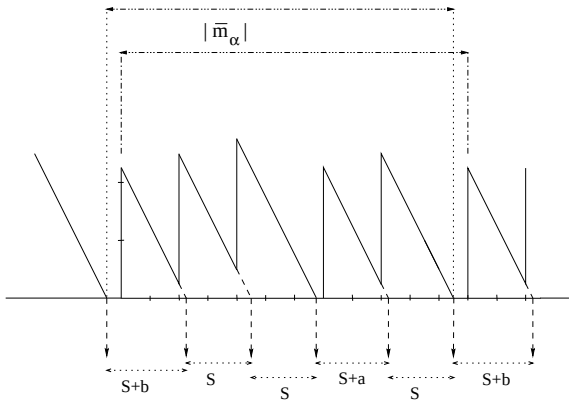

Figure 2: Output process from $S = 9/4$ and $\overline{m}_{5/12}$

when the sub-words $\overline{m}_{r_1(n)}$ and $\overline{m}_{r_1(n+1)}$ finish and since the load is null at the end of these sub-words, therefore at the end of one $\overline{m}_\alpha$ the state of the queue is identical to the state at the beginning of $\overline{m}_\alpha$ and the output process is periodic of period q.
ii) We can notice that $|d(n,2)|$ is the time which separates the last departure of $\overline{m}_\alpha$ and the following admission. We can also notice that the length q is the time between the first admission of the word and the first admission of the following word. Hence the time between the last departures of two consecutive $\overline{m}_\alpha$ can be decomposed as follows : time until the beginning of the following word : $|d(n,2)|$ added by the time until the last departure of the word that is $q - |d(n,2)|$. Hence the time between the last departure is q.

The same arguments holds for $w = \overline{m}_{r(n)}$ and $\overline{m}_{\overline{S}^{-1}}$. □

On the top of Figure 2 these different periods, presented in above corollary, are represented each one with a particular departure. In addition it can be said that for any $\alpha \in [r_k, r_{k+1}]$ the intensity of the output sub-processes y_k and y_{k+1} are p_k/q_k and p_{k+1}/q_{k+1} respectively.

Note that when $\alpha = r_i(n)$, the numbers $p_i(n)$ and $q_i(n)$ are the number of customer admitted in the queue i and the time spent since the first admission of the period respectively.

3.2.1 Workload Properties

Although a well-known result insure the conservation of the intensity from the input to the output process through a stable queue (as presented for example in [5, 17]), it is necessary here to refine this and to investigate about the behavior of the queue during the special factors $y_i(n)$ and $x_i(n)$ which appear during the output process of a queue i. We point out here the special role played by the terms $r_i(n)$ and the factors $\overline{m}_{r_i(n)}$. Which is one of the main points of this section.

Definition 19. *Let us assume that $\overline{S}_{i-1}$ is achieved in the jth queue and that $\overline{S}_{i-1} \leq S_i$. Let $\overline{l}_i$ be the greatest integer such that $\forall\, k \leq \overline{l}_i$, $l_j(k) = l_i(k)$ and $\forall\, k > \overline{l}_i$, $l_j(k) \neq l_i(k)$.*

Here are some particular cases: if $i = 1$ then $\overline{l}_1 = 0$. If $S_i \leq \lceil (\overline{S}_{i-1})^{-1} \rceil$ then $\overline{l}_i = 0$.

We can add that if $\overline{l}_i \geq 1$ then $r_j(\overline{l}_i) \leq 1/S_i \leq r_j(\overline{l}_i + 1)$ and that $\forall\, 0 \leq k \leq \overline{l}_i$, $r_j(\overline{l}_i - k) = r_i(\overline{l}_i - k)$.

The workload in queue i at epoch t (denoted by $w_i(t)$) is given by

$$w_i(t) = \Big(w_i(T_i(n-1)_-) + S_i - (t - T_i(n-1)) \Big)^+, \text{ for } T_i(n-1) \leq t < T_i(n).$$

The following Lemma allows us to show the applicability of the framework introduced in [9] and the correctness of the generalization of the use of the expansion of $1/S_i$.

Lemma 20. *Let $(x(i)(n))_{n\geq 0}$ and $(y(i)(n))_{n\geq 0}$ be the sequences of factors computed according to Theorem 11 using the expansion of $1/S_i$. Assume that the input process is given by Lemma 16. If the initial workload in queue i is null then using $x(i)(\cdot)$ as input sequence, the workload during $x(i)(\cdot)$ is never null and remains non negative at the end of the sequence. If the initial workload in queue i is null then using $y(i)(\cdot)$ as input sequence, the workload during $y(i)(\cdot)$ is never null until the last letter of $y(i)(\cdot)$ (which is always a 0) and is null at the end of the sequence. Moreover the workload increase due to $x(i)(\cdot)$ equals $d_i(\cdot, 1)$ and the maximal workload decrease due to $y(i)(\cdot)$ equals $d_i(\cdot, 2)$.*

Proof. The proof is postponed in Appendix 8. □

3.3 Average values in tandem queues under a mechanical input in the system

We are interested in this section in the computation of the average waiting time, the average sojourn time and the average response time in one queue of a system of queues in tandem which input process is a mechanical word with slope α (denoted by $\alpha = p/q$ when the slope is rational). In this part, we improve, simplify and generalize for tandem queues the formula given in [9] for the waiting time in a single queue when the arrival sequence is Sturmian. Most of the properties shown in [9] can be easily deduced from the new formulas given here.

Definition 21 (Average response and waiting time). *We denote by $V_i(\overline{m}_\alpha)$ the average sojourn time and $W_i(\overline{m}_\alpha)$ the average waiting time of a customer after its admission in the i-th queue when the input process in the first queue is a mechanical word $\overline{m}_\alpha$. Similarly the response time for queue i is denoted by $R_i(\overline{m}_\alpha)$.*

We denote by $K_i(m, w)$ the sum of the waiting times admitted in the queue i during the finite sequence m with an initial workload equal to w, then the average waiting time over m in an empty queue $W_i(m)$ is given by

$$W_i(m) = \frac{K_i(m, 0)}{|m|_1}.$$

When the word m is infinite we consider then $m[n]$ the prefix of m which length is n. The average waiting time is given by

$$W_i(m) = \lim_{n\to\infty} \frac{K_i(m[n], 0)}{|m[n]|_1}.$$

We recall now the stability condition in the system of queues which is

$$\alpha \leq \frac{1}{\overline{S}},$$

since by the preceding section the intensity of the input process is α and $\overline{S}$ is the maximal service time in all the queues.

3.3.1 Computation the average waiting time in the ith queue

This part of the work is dedicated to the effective computations of the average waiting time in queue i. A closed formula for the average waiting time is obtained. The formulas of the average sojourn

time and of the average response time are directly deduced. We first assume that $S_i \geq \overline{S}_{i-1}$ indeed all other cases could be considered as degenerated cases and are treated later.

Since the workload is null at some special epochs the sum of the waiting time during $\overline{m}_\alpha$ is a linear combination of the sum of the waiting time during the factors $y_i(\cdot)$ and $y_i(\cdot+1)$. Moreover for all number α which ceiled expansion is common with those of S_i until order k, then the composition of $\overline{m}_\alpha$ in $y_n(S_i)$ and $y_{n+1}(S_i)$ can be expressed using the coefficients of the ceiled expansion of α after order k. This result allows to precise the appearance of the $(x-y)$ factors in $\overline{m}_\alpha$. This yield to Theorem 22.

Theorem 22 (Waiting time and response time). *Let α be the number such that $r_i(n) \leq \alpha \leq r_i(n+1)$ then the average waiting time $W_i(\overline{m}_\alpha)$ is given by*

$$\begin{aligned} W_i(\overline{m}_\alpha) \ &= q_i(n)K_i(\overline{m}_{r_i(n+1)},0) - q_i(n+1)K_i(\overline{m}_{r_i(n)},0) \\ &\quad - \frac{p_i(n)K_i(\overline{m}_{r_i(n+1)},0) - p_i(n+1)K_i(\overline{m}_{r_i(n)},0)}{\alpha}. \end{aligned} \tag{12}$$

Where $K_i(0)=0$ and $K_i(1)=0$ and $\forall n \geq 2$ we have
if $n \leq \bar{l}_i$ then

$$K_i(\overline{m}_{r_i(n)},0) = p_i(n)\big(p_i(n)-1\big)\frac{(S_i-\overline{S}_{i-1})}{2}, \tag{13}$$

if $n = \bar{l}_i + 1$, with the convention that $\overline{S}_{i-1} = S_{i-1}$, then

$$\begin{aligned} K_i(\overline{m}_{r_i(n)},0) &= l_i(n)K_i(\overline{m}_{r_i(n-1)},0) - K_i(\overline{m}_{r_i(n-2)},0) + \\ &\frac{S_i}{2}\big(l_i^2(n)p_i^2(n-1) - l_i(n)p_i^2(n-1) - 2l_i(n)p_i(n-1)p_i(n-2) + 2p_i^2(n-2)\big) - \\ &\frac{S_{i-1}}{2}\big(l_{i-1}^2(n)p_{i-1}^2(n-1) - l_{i-1}(n)p_{i-1}^2(n-1) - 2l_{i-1}(n)p_{i-1}(n-1)p_{i-1}(n-2) + 2p_{i-1}^2(n-2)\big) \\ &- \big(l_i(n) - l_{i-1}(n)\big)\, p_{i-1}(n-1)\big(\frac{1}{2}q_{i-1}(n-1)(l_i(n)+l_{i-1}(n)-1) - q_{i-1}(n-2)\big), \end{aligned} \tag{14}$$

and if $n \geq \bar{l}_i + 2$ then

$$\begin{aligned} K_i(\overline{m}_{r_i(n)},0) &= l_i(n)K_i(\overline{m}_{r_i(n-1)},0) - K_i(\overline{m}_{r_i(n-2)},0) + \big(p_i(n-1) - p_i(n-2)\big)\, d_i(n-1,1) \\ &\quad + p_i(n-1)\,\big(l_i(n)-2\big)\,\Big(d_i(n-1,1) + \frac{l_i(n)-1}{2}d_i(n-1,2)\Big). \end{aligned} \tag{15}$$

Proof. The proof is postponed in Appendix 9. □

It could be noticed that the cases which correspond to a direct sturmian input in the queue are these where $\bar{l}_i = 0$ (equivalently $S_i \leq \lceil \overline{S}_{i-1} \rceil$) and we found Formula of [9].

But it remains some degenerated cases where the framework of the $(x-y)$ factorization is not required.

Lemma 23 (Degenerated cases). *If $S_i \leq \overline{S}_{i-1}$ then the average waiting time is null.*

Proof. When $S_i \leq \overline{S}_{i-1}$ then all the inter-arrival times are larger than S_i and no customer has to wait before entering in queue i. □

3.3.2 Properties and examples

Lemma 24. *If α is such that $r_i(n) \leq \alpha \leq r_i(n+1)$, then*
i) the function $\alpha \to W_i(\overline{m}_\alpha)$ is continuous, increasing, and concave in α.
ii) the function $\alpha \to \alpha W_i(\overline{m}_\alpha)$ is continuous, increasing, and linear in α.

Proof. It suffices to notice that in Equation (12) the coefficient of the term divided by α is non negative and that the coefficient of the constant term is non negative as it is shown by Lemma 33. □

Hence the rational numbers $r_i(n)$ form the sequence of all the cusps of the functions $\alpha \to W_i(\overline{m}_\alpha)$ and $\alpha \to \alpha W_i(\overline{m}_\alpha)$.

Remark 25 (Linearity of the average number of customer). *Although the fact that the linearity of the average number of customer can be easily deduced from the lemma above and Little's Formula (see for example in [11]). We should present another proof (presented for a single queue in [13]) which helps to understand the behavior of the queue. This proof is made for a system of two queues and can be easily generalized.*

Let α be the number such that $\alpha = \mu r(n+1) + (1-\mu) r(n)$ where μ is a number in $[0,1]$, let $N(\overline{m}_\alpha)$ be the average number of customer in the second system (the server and the queue) under the admission sequence $\overline{m}_\alpha$ in the first queue.

Let us compute $N(\overline{m}_\alpha)$. Since $\overline{m}_\alpha$ can be factorized with $\overline{m}_{r(n+1)}$ and $\overline{m}_{r(n)}$ where the number of customer in the system at the end of these sub-words is equal to zero, then

$$N(\overline{m}_\alpha) = \lambda N(\overline{m}_{r(n+1)}) + (1-\lambda) N(\overline{m}_{r(n)}),$$

where λ is the fraction of time spent in a $\overline{m}_{r(n+1)}$ word and $1-\lambda$ the fraction of time spent in a $\overline{m}_{r(n)}$ word. As it is shown before $\mu = \lambda$.

In the example below we represent the function $\alpha \to W_2(\overline{m}_\alpha)$. One wants stress the dynamical changes of this curve according the different values of the first service time.

Example 26. *In this example the number of queues in the network is chosen to be equal to 2. We represent on Figure 3 the values of $W_2(\overline{m}_\alpha)$ when α varies for different values of S_1 while the second service time S_2 is fixed to $34/21$. The ceiled expansion of S_2 is $\langle 2,3,3,3\rangle$. Hence the cusps of $34/21$ are $1/2 = \langle 2\rangle$, $3/5 = \langle 2,3\rangle$ and $8/13 = \langle 2,3,3\rangle$. The service times in the first queue take the following values: $S_1 = 1$, $S_1 = 16/15$, $S_1 = 3/2$, $S_1 = 77/55$ and $S_1 = 8/5$.*

When $S_1 = 1$ we are in a case such that $\overline{l}_2 = 0$ and the values of $W_2(\overline{m}_\alpha)$ are the same as for the average waiting time of a classical mechanical input (meaning $W_2(\overline{m}_\alpha) = W_1(\overline{m}_\alpha)$). The values of $W_2(\overline{m}_\alpha)$ in the cusps are all computed with (15).

When $S_1 = 16/15 = \langle 2,2,13/14\rangle$ and when $S_1 = 3/2 = \langle 2,2\rangle$ the ceiled expansions of S_1 and S_2 have just their first coefficient in common. This means that the value in the first cusp is computed by (14) *and in the following cusps with* (15).

When $S_1 = 77/51 = \langle 2,3,24/25\rangle$ or $S_1 = 8/5 = \langle 2,3,2\rangle$ the ceiled expansions of the service times are identical until the second coefficient. The value of the average time in $3/5$ is computed by (13), *while the value in $8/13$ is computed* (14) *and finally the value in $21/34$ with* (15).

On Figure 3, starting from the top, the curves appear in the increasing order of the first service time (1, $16/15$, $3/2$, $77/55$, $8/5$). The part of each curve where the computations differ for the mechanical input are stressed. Although it could be thought that the shape of the curves where the computations are done using (13) *and* (14) *rather than* (15) *will be different, rather surprisingly this is not the case.*

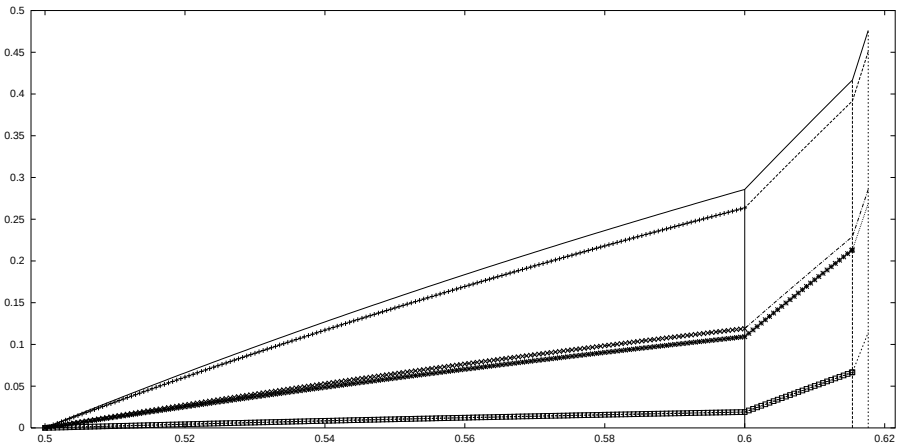

Figure 3: Average waiting time for 5 different first service times

4 Optimal routing for two parallel systems of queues in tandem

This is the main section of the paper. We are interested in routing in two parallel systems Q^1 and Q^2 composed by N^1 (resp. N^2) $./D/1/\infty/FIFO$ queues in tandem. The global inter-arrival times are all the same and the time unit is chosen (by scaling) such that the inter-arrival times before the routing are equal to one.

The routing operates as follows. Let m be a binary infinite word (periodic or not). If $m(n)$ (the n^{th} letter of m) is one then the customer is routed in system Q^1 otherwise the customer is sent in the system Q^2. If the word m has a slope, it represents the ratio of customers sent in system Q^1 and it is denoted by α while the ratio of customers sent in Q^2 is $1-\alpha$.

Our aim is to find a policy which minimizes some cost function denoted by g of the average total response time. We will assume that the cost function g is of the following form:

$$g(m) = \overline{\lim}_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} g(m,k), \tag{16}$$

where $g(m,k)$ is a weighted response time of the kth customer under routing m. If customer k is sent to system Q^j, $j \in \{1,2\}$, then

$$g(m,k) = c_0^j + \Sigma_{i=1}^{N^j} c_i^j V_i^j(k_m^j), \quad j = 1,2 \tag{17}$$

where m is the input sequence, c_i^j is a non-negative coefficient, k_m^j is the number of customers sent to system Q^j among the first k customers under routing m and $V_i^j(k_m^j)$ is the sojourn time in node i of system Q^j of the k_m^j-th customer.

The problem is to find an optimal allocation sequence in the two systems and to compute the optimal ratio of customers sent in each system $(\alpha, 1-\alpha)$ associated with this allocation sequence.

4.1 Optimal policies

In order to compute an optimal allocation sequence we first prove the following theorem.

Theorem 27 ([2]). *The cost function $g(m)$ is minimized for a routing sequence which is a lower mechanical in sequence.*

Proof. The cost $g(m)$ is the average of a positive linear combination of multimodular functions (see Theorem 8). Since a positive linear combination of multimodular functions is obviously multimodular, then $g(m)$ is the average of multimodular functions. As a direct consequence of the multi-criteria optimization results found in [2], applied to the present case, g is minimized by a lower mechanical sequence. □

It can be shown (see [3]) that the average response time under an arrival process of the form $\underline{m}_\alpha$ is the same as the average response time under the arrival process $\overline{m}_\alpha$. Furthermore when the input sequence in the first system is an upper mechanical word with a slope α, then the input sequence in the second system is a lower mechanical word of slope $1-\alpha$, (see [15]).

The previous theorem says the optimal routing sequence is a Sturmian sequence, however it does not provide any method to compute the slope α_{opt} of the optimal Sturmian sequence.

4.2 Optimal cost

This part is devoted to the computation of the slope of the optimal Sturmian input. We do this by minimizing $g(\overline{m}_\alpha)$ over $\alpha \in [0,1]$.

Let us precise which slopes are acceptable or possible for our problem, that is the slopes (or ratios) for which the system is stable. The condition of stability in a queue is

$$\alpha \leq \frac{1}{S_i^j}, \forall i \in \{1,\ldots,N_j\}, j \in \{1,2\}.$$

But the stability of the whole system is also necessary yielding

$$\rho = \frac{1}{\frac{1}{\max_{1\leq i\leq N_1} S_i^1} + \frac{1}{\max_{1\leq i\leq N_2} S_i^2}} \leq 1.$$

This gives the interval of stability, denoted by I_s, in which α varies, with

$$I_s = \left[1 - \min_{1\leq i\leq N_2} \frac{1}{S_i^2}, \min_{1\leq i\leq N_1} \frac{1}{S_i^1}\right] \cap [0,1],$$

with the convention that $1/(\max_{1\leq i\leq N_1} S_i^1) \geq 1/(\max_{1\leq i\leq N_2} S_i^2)$.

If the system is not stable (the interval of stability is reduced to the empty set) then the cost is infinite for all routings. This case will no longer be considered.

Let $g(m)$ be the total weighted average response time of the two systems Q^1 and Q^2 under the input sequence m. Assuming that the routing sequences are Sturmian with slope α and conditioning over the choices, the weighted average response time of the customer in the system is given by

$$g(\overline{m}_\alpha) = \alpha g^1(\overline{m}_\alpha) + (1-\alpha) g^2(\underline{m}_{1-\alpha}), \tag{18}$$

where $g^1(\overline{m}_\alpha)$ is defined by Equation 16 with $g^1(\overline{m}_\alpha, k)$ computed by (17) for all routing in the system Q^1. Therefore we are interested to find the optimal ratio α_{opt} defined by :

$$\alpha_{opt} = \arg\min_{\alpha\in I_s} g(\overline{m}_\alpha).$$

4.2.1 Easy cases

Here are some cases where the framework developed in Section 3 is not necessary.

Lemma 28 (Fully loaded system). *If $1/\overline{S}^2 + 1/\overline{S}^1 = 1$ then I_s is reduced to a single point and there exists only one admission ratio*

$$\alpha_{opt} = \frac{1}{\overline{S}^1}. \tag{19}$$

This is the only case where the optimal ratio could be an irrational number as it will be seen later.

Lemma 29 (Service times smaller than one). *If the service times S_i^j are all smaller than one then an optimal policy is to send all the customers in the queue which satisfy $\min_{j=1,2} \Sigma_{1\le i\le N^j}(c_i^j S_i^j) + c_0^j$.*

Proof. In the queues, services are smaller than inter-arrival times. Therefore, the sojourn time is the service time. □

Lemma 30 (c_i decreasing in i). *If (c_i^1) and (c_i^2) are non-increasing finite sequences then the computation of the cost is the same as for a sum of simpler systems, each of them with a single queue in the first system and a single queue in the second system. For system k, the respective service times are $\overline{S}_k^j = \max_{i=1}^k(S_i^j)$, and coefficients $\hat{c}_0^j = (c_k^j - c_{k+1}^j) + \sum_{i,S_i^j\neq\overline{S}_j}^k S_i^j$ and $\hat{c}_1^j = (c_k^j - c_{k+1}^j)$ for all $j = 1, 2$.*

Proof. If c_i^j is non-increasing then $g^j(\overline{m}_\alpha)$ can be rewrited under the following form

$$g^j(\overline{m}_\alpha) = c_0^j + \Sigma_{i=1}^{N^j-1}(c_i^j - c_{i+1}^j)R_i^j + c_{N^j}^j R_{N^j}^j.$$

With the exchange arguments proved in Lemma 5 applied on the sum of the waiting time it follows

$$g^j(\overline{m}_\alpha) = c_0^j + \Sigma_{k=1}^{N^j}(c_k^j - c_{k+1}^j)\big(\hat{W}_{1,k}^j + \sum_{i=1}^k S_i^j\big),$$

where $\hat{W}_{1,k}$ is the waiting time in the first queue when the largest of the first k service times is placed in first position. □

Note that when the coefficients c_i^j are all equal, then the cost function $g(m)$ is equal to the average response time of the whole system multiplied by c^j.

4.2.2 General case

We consider here the case where the system is not fully loaded and the cost involves arbitrary coefficients c_i^1 and c_i^2. We are interested to find the optimal ratio α_{opt} defined by :

$$\alpha_{opt} = \arg\min_{\alpha\in I_s} g(\overline{m}_\alpha).$$

Theorem 31 (Periodicity). *For any real numbers S_i^1, S_j^2 with $i \in \{1..N^1\}$ and $j \in \{1..N^2\}$, the optimal ratio is a cusp of $g(\overline{m}_\alpha)$. Since cusps are rational numbers, the optimal policy associated with α_{opt} will always be periodic.*

Proof. Let us define r_k^*, $k \in \mathbb{Z}$ as the sequence of all the cusps of $g(\overline{m}_\alpha)$. Since $g(\overline{m}_\alpha)$ is the sum of two piece-wise linear functions, it is a piece-wise linear function whose set of cusps is the union of the set of cusps of the two functions. We order them in the increasing order such that r_0^* is the rational with the smallest denominator in I_s and such that r_{-n}^* with $n \in \mathbb{N}$ is a cusp of $(1-\alpha)g^2(\overline{m}_{1-\alpha})$ and r_n^* with $n \in \mathbb{N}$ is a cusp of $\alpha g^1(\overline{m}_\alpha)$.

The function $g(\overline{m}_\alpha)$ is linear for $\alpha \in [r_k^*, r_{k+1}^*]$. This is due to the linearity in α of the functions $\alpha g^1(\overline{m}_\alpha)$ and $(1-\alpha)g^2(\overline{m}_{1-\alpha})$.

Therefore the set of possible $\arg\min$ is the set of all the cusps r_k^* union $\{\max_{1\le i\le N^1}(S_i^1)^{-1}, \max_{1\le i\le N^2}(S_i^2)^{-1}\}$. Lemma 33 excludes the points $\max_{1\le i\le N^1}(S_i^1)^{-1}$ and $\max_{1\le i\le N^2}(S_i^2)^{-1}$ when these terms are not rational. Hence the optimal ratio is a rational number and by this way an optimal policy is periodic. □

Remark 32 (Double cusp). *The rational number with smallest denominator in I_s, namely r_0^*, is a cusp for both $\alpha \to \alpha g^1(\overline{m}_\alpha)$ and $\alpha \to (1-\alpha)g^2(\overline{m}_{1-\alpha})$. This is the only common cusp thus we call it the double cusp. Moreover in I_s all the cusps of $\alpha \to \alpha g^1(\overline{m}_\alpha)$ are larger than the double cusp and all the cusps of $(1-\alpha)g^2(\overline{m}_{1-\alpha})$ are smaller than the double cusp.*

This comes from the fact that the convergents of order n : $r(n)(\overline{S})$ is the rational number with the smallest denominator in $(r(n-1)(\overline{S}), \overline{S}^{-1})$. (This result can be proved using combinatorial properties of Sturmian word ([9]) or the characterization (presented for example in [13]) that for all $n \ge 1$ and for all i the intervals $(r_i(1), r_i(n))$ and $(r_i(n), S_i^{-1})$ are Farey's intervals).

4.3 Algorithm and computational issues

On Figure 4 we present an algorithm to compute the optimal ratio α_{opt}.

Find double cusp r_0^*
current-cusp := r_0^*
Compute the next-cusp-right of r_0^*
while $g(\overline{m}_{\text{current-cusp}}) > g(\overline{m}_{\text{next-cusp-right}})$ **do**
    current-cusp:= next-cusp-right
    Compute the next-cusp-right of current-cusp
endwhile
Compute the next-cusp-left of r_0^*
while $g(\overline{m}_{\text{current-cusp}}) > g(\overline{m}_{\text{next-cusp-left}})$ **do**
    current-cusp:= next-cusp-left
    compute the next-cusp-left of current-cusp
endwhile
return current-cusp

Figure 4: Algorithm computing α_{opt}

4.3.1 Correctness

Some preliminary results are required to get the correctness of the algorithm.

Lemma 33. *For all $j \in \{1..N\}$ the functions $i \to \frac{K_j(\overline{m}_{r(i)},0)}{p_j(i)}$ and $i \to \frac{K_j(\overline{m}_{r(i)},0)}{q_j(i)}$ -i) are increasing,*

-ii) are convex,
-iii) have a growth rate which tends to infinity with i.

The proof is detailed in Appendix 10.

Theorem 34 (Convergence of the algorithm). *The algorithm converges and finds α_{opt} in a finite number of steps.*

Proof. Correctness. Since by Lemma 33 the function $i \to \frac{K(\overline{m}_{r(i)},0)}{q(i)}$ is convex, then the function $k \to g(\overline{m}_{r_k^*})$ is also convex. The ratio α_{opt} being in $\{r_k^*\}$ this yields the correctness of the algorithm.

Finiteness. We introduce $n \in \mathbb{Z}$. According to Lemma 33 the growth rate of the function $n \to g^1(\overline{m}_{r_n^*})$, tends to infinity when $n \to +\infty$. Similarly the growth rate of $n \to g^2(\overline{m}_{r_n^*})$, tends to infinity when $n \to -\infty$. Therefore the integer numbers n_0^+ and n_0^- such that

$$\forall n \geq n_0^+, \quad g(\overline{m}_{r_{n+1}^*}) - g(\overline{m}_{r_n^*}) > 0,$$

and such that

$$\forall n \leq n_0^-, \quad g(\overline{m}_{r_{n-1}^*}) - g(\overline{m}_{r_n^*}) > 0,$$

are finite. □

5 Numerical experiments

This section is dedicated to the presentation of several runs of the algorithm in order to shown how the optimal policy (or equivalently the ratio of the optimal policy) behaves with respect to the parameters of the system, namely, the service times as well as the inter-arrival time.

The algorithm presented above has been implemented in Maple in order to keep exact values for all the rational numbers involved in the computations.

The computations are made for a network of systems composed by 2 deterministic queues in tandem as shown on Figure 5.

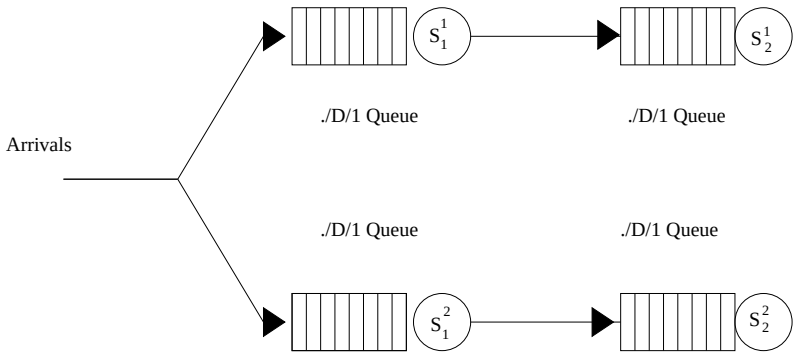

Figure 5: Admission control in networks of 2 queues in tandem

The first service times in the two systems are equal and fixed to $6/5$ that means $S_1^1 = S_1^2 = 6/5$ while the inter arrivals are fixed to one, but they can be modified by scaling the time units. We assume that $c_0^1 = c_0^2 = 0$, $c_1^1 = c_1^2 = 1$ and $c_2^1 = c_2^2 = 2$. We let the inverse of the second service times $1/S_2^1$ and $1/S_2^2$ vary. We restrict our investigations to the domain of stability namely I_s.

The figure 6 displays the zones where the values of α_{opt} remain the same: each cell represents all the couples $(1/S_2^1, 1/S_2^2)$ with the same optimal ratio.

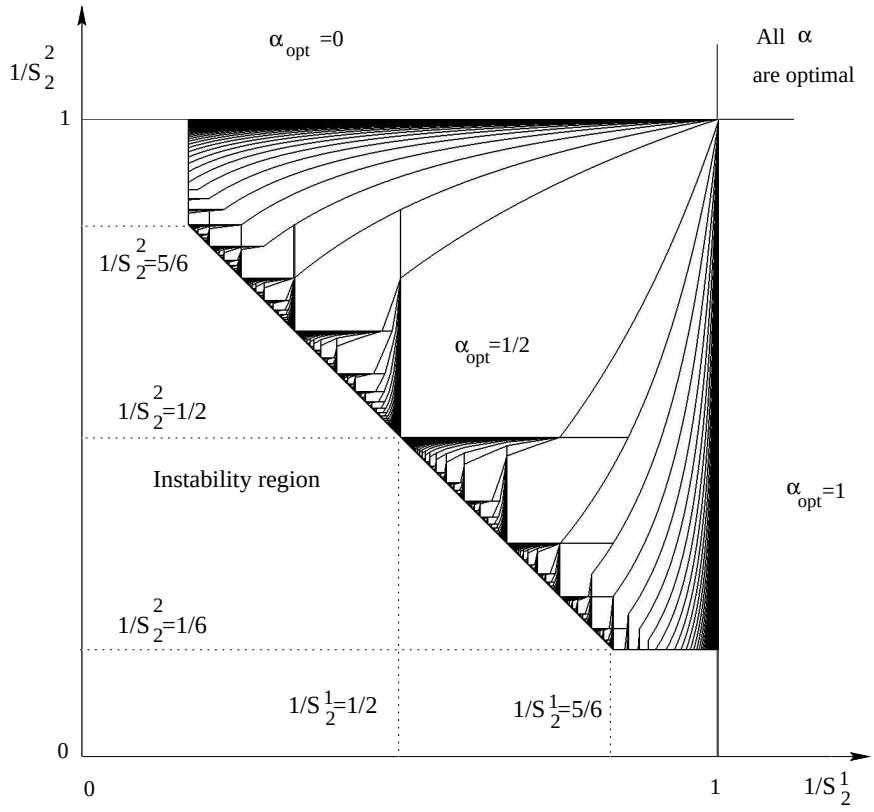

Figure 6: Optimal ratios when $1/S_2^1$ and $1/S_2^2$ vary

The larger cell corresponds to the area where the round robin policy ($\alpha_{opt} = 1/2$) is optimal. The vertical (resp. horizontal) border of I_s takes place at the point $1/S_2^1 = 1/6$ with $1/S_2^2 \geq 5/6$ (resp. $1/S_2^2 = 1/6$ with $1/S_2^1 \geq 5/6$). Indeed when $1/S_2^1 < 1/6$ and $1/S_2^2 \geq 5/6$ we have $S_2^1 \wedge S_1^1 = S_2^1 > 6/5$ and $1/S_2^2 \wedge S_1^2 = S_1^2 = 6/5$. Therefore the stability condition $\frac{1}{S_1^2} + \frac{1}{S_2^1} \geq 1$ is not satisfied since $1/S_1^2 < 1 - 5/6$. A similar argument explains the horizontal border.

When $S_2^i \leq 6/5$ the waiting time in the second queue of system i is null since we are in a case where $S_2^i \leq S_1^i$. Thus the part of the figure where $1/S_2^1 \geq 5/6$ and $1/S_2^2 \geq 5/6$ is a part where α_{opt} is computed similarly to the optimal ratio of the response time in one queue with a mechanical input. Furthermore this is the only case where the two waiting times are both null in the second queues.

Let us precise now the parts of Figure 6 where one can find a first waiting time null. From Lemma 23 in order to get a null waiting time in the first queue one only needs to be in the following situation

$$\frac{1}{S_2^i} \leq \frac{1}{\lceil S_1^i \rceil} = \frac{1}{2}.$$

Therefore the zones where $1/S_2^i \leq 1/2$ correspond to $W_1^i(\overline{m}_\alpha) = 0$.

In addition the part of the graph delimited by $1/2 \leq 1/S_2^2 \leq 5/6$ and $1/2 \leq 1/S_2^1 \leq 5/6$ is an area where none of the waiting times is null.

Finally when the ceiled convergents of the second service times get out of I_s this causes a sudden change of the optimal ratio. The horizontal or vertical straight lines which appear in the figure can be explained by this phenomena.

6 Conclusion

In this paper, the framework used in [9] has been extended and it has been shown that the combinatorial properties of the mechanical word are kept after the passage through a deterministic queue. The multimodularity results of [1, 2, 3] have been extended which allows to compute the optimal routing for several general cost function. The results presented here could be adapted for other protocols which require a constant bandwidth : applications using the CBR mode of ATM networks or RSVP for example. In a more theoretical way more general arrival process (with stochastic inter arrival) and more general services are currently investigated by the authors.

References

[1] E. Altman, B. Gaujal, and A. Hordijk. Admission control in stochastic event graphs. *IEEE Trans. Aut. Cont.*, 45(5):854–867, May 2000.

[2] E. Altman, B. Gaujal, and A. Hordijk. Balanced sequences and optimal routing. *J. Assoc. Comput. Mach.*, 47:752–775, 2000.

[3] E. Altman, B. Gaujal, and A. Hordijk. Multimodularity, convexity and optimization properties. *Mathematics of Operation Research*, 25:324–347, May 2000.

[4] F. Baccelli, A. Borovkov, and J. Mairesse. Asymptotic results on infinite tandem queueing networks. *Probability Theory and related Fields*.

[5] F. Baccelli and P. Bremaud. *Elements of queueing theory*. Springer, 1992.

[6] P. Flajolet. Combinatorial aspects of continued fractions. *Discrete Mathematics*, 32:125–161, 1980.

[7] P. Flajolet and F Guillemin. The formal theory of birth-and-death processes, lattice path combinatorics and continued fractions. *Advances in Applied Probability*, 32(3):750–778, 2000.

[8] Philippe Flajolet and Brigitte Vall´ee. Continued fraction algorithms, functional operators, and structure constants. *Theor. Comp. Sc.*, 194(1-2):1–34, March 1998.

[9] B. Gaujal and E. Hyon. Optimal routing policies in two deterministic queues. *Réseaux et systèmes répartis - Calculateurs Parallèles*, 13(6):601–633, 2001.

[10] B. Gaujal and E. Hyon. Routage optimal dans des r´eseaux de files d'attentes d´eterministes. In G. Juanole and R. Valette, editors, *Modélisation des systèmes réactifs*, pages 133–148. Hermès, 2001. In French.

[11] D. Gross and C.M. Harris. *Fundamentals of queueing theory*. Wiley, 2nd edition edition, 1985.

[12] B. Hajek. Extremal splittings of point processes. *Mathematics of Operation Research*, 10(4):543–556, 1985.

[13] A. Hordijk and D. van der Laan. On the linearity of the average number customers in a queue on a farey's interval. Private Communication.

[14] M. Lothaire. *Mots*, chapter Trac´e de droites, fractions continues et morphismes it´er´es. Hermes, 1991.

[15] M. Lothaire. *Algebraic Combinatorics on Words*, chapter Sturmian Words. Cambridge University Press, 2002.

[16] J. Mairesse and P. Prabhakar. On the existence of fixed points for the $./gi/1/\infty$ queue. Technical Report 99/25, LIAFA, 1999.

[17] P. Robert. *Réseaux et files d'attentes : Méthodes probabilistes*. Math´ematiques et applications. Springer, 2000.

[18] E. Rosen, A. Viswanathan, and R. Callon. A proposed architecture for multiprotocol label swicthing. Technical Report RFC 3031, IETF, 2001.

[19] C. Villamizar. Mpls optimized multipath (mpls-omp). Internet draft (work in progress), IETF, ftp://ftp.isi.edu/internet-drafts/draft-villamizar-mpls-omp-01, 1999.

7 Appendix : Proof of Lemma 14

i) Let $s = \langle a_1, \ldots, a_{n+1} - s_{n+1} \rangle$ be the partial ceiled expansion at order $n+1$ of s. Let us compute now the ceiled expansion at order n of α, we get $\alpha = \langle a_1, \ldots, a_{n-1}, a_n - \alpha_n \rangle$ since $\alpha \geq r(n)$. The condition $\alpha \geq r(n)$ also gives $\alpha_n \neq 0$, whereas the condition $\alpha \leq r(n+1)$ gives $\lceil (\alpha_n)^{-1} \rceil \geq a_{n+1}$.

ii) Using Theorem 11, $\overline{m}_\alpha$ can be factorized by $x(n+1)(\alpha)$ and $y(n+1)(\alpha)$. Since $\lfloor (\alpha_n)^{-1} \rfloor \geq a_{n+1}$ then it exists $h \geq 1$ such that $x(n+1)(\alpha) = x(n)(\alpha)\,(y(n)(\alpha))^{a_{n+1}+h-2}$ and $y(n+1)(\alpha) = x(n)(\alpha)\,(y(n)(\alpha))^{a_{n+1}+h-1}$. Since the coefficients of expansion of α and those of s are equal until order n then $\forall\, k \leq n$ we have $x_k(\alpha) = x_k(s)$ and $y_k(\alpha) = y_k(s)$. This implies $x(n+1)(\alpha) = y(n+1)(s)\,(y(n)(s))^{h-1}$ and $y(n+1)(\alpha) = y(n+1)(s)\,(y(n)(s))^{h}$.

iii) The total number of ones (or the total number of letters) in $\overline{m}_\alpha$ can be written using the terms $p(n)(s)$ and $p(n+1)(s)$ (or $q(n)(s)$ and $q(n+1)(s)$ respectively). Indeed noting that $x(n)(\alpha)$ is a factor of $y(n+1)(s)$ and not a factor of $y(n)(s)$, then when $x(n)(\alpha)$ appears in $\overline{m}_\alpha$ it corresponds to an appearance of $y(n+1)(s)$ in $\overline{m}_\alpha$. Let $|\overline{m}_\alpha|_{x(n)}$ and $|\overline{m}_\alpha|_{y(n)}$ denote the number of factor $x(n)$ and the number of factor $y(n)$ in $\overline{m}_\alpha$ respectively corresponding to the factorization in $y(n+1)(s)$ and $y(n)(s)$. By Lemma 11 of [9], one can show that $\alpha_n = |\overline{m}_\alpha|_{x(n)}/(|\overline{m}_\alpha|_{x(n)} + |\overline{m}_\alpha|_{y(n)})$ when $\overline{m}_\alpha$ is a finite word.

Hence when $\overline{m}_\alpha$ is a finite word, then

$$\frac{|\overline{m}_\alpha|_1}{|\overline{m}_\alpha|} = \frac{\alpha_n p(n+1) + [(1-\alpha_n) - \alpha_n(a_{n+1}-1)]\, p(n)}{\alpha_n q(n+1) + [(1-\alpha_n) - \alpha_n(a_{n+1}-1)]\, q(n)}. \tag{20}$$

Now, assume that $\overline{m}_\alpha$ is infinite.

Let $m[i]$ be an increasing sequence of prefixes of $\overline{m}_\alpha$ such that $m[i]$ finishes either by a factor $y(n+1)(s)$ or a factor $y(n)(s)$. The Lemma 11 of [9] for infinite words becomes

$$\alpha_n = \lim_{i \to \infty} \frac{|m[i]|_{x(n)}}{|m[i]|_{x(n)} + |m[i]|_{y(n)}}. \tag{21}$$

Therefore with Definition 9 we get

$$\alpha = \lim_{i\to\infty} \frac{|m[i]|_1}{|m[i]|} = \lim_{i\to\infty} \frac{p(n+1)|m[i]|_{x(n)} + (|m[i]|_{y(n)} - (a_{n+1}-1)|m[i]|_{x(n)})p(n)}{q(n+1)|m[i]|_{x(n)} + (|m[i]|_{y(n)} - (a_{n+1}-1)|m[i]|_{x(n)})q(n)}.$$

Multiplying both the numerator and the denominator by $1/(|m[i]|_{x(n)} + |m[i]|_{y(n)})$, since $\lim_{i\to\infty} \frac{|m[i]|_{x(n)}}{|m[i]|_{x(n)}+|m[i]|_{y(n)}}$ exists and is finite then with Equation (21) it follows (20) again :

$$\alpha = \frac{\alpha_n p(n+1) + [(1-\alpha_n) - \alpha_n(a_{n+1}-1)]\, p(n)}{\alpha_n q(n+1) + [(1-\alpha_n) - \alpha_n(a_{n+1}-1)]\, q(n)}.$$

By simplifying (20) and owing to Equation (8) this leads to

$$\alpha = \frac{-\alpha_n p(n-1) + p(n)}{-\alpha_n q(n-1) + q(n)}.$$

8 Appendix : Proof of Lemma 20

We begin to notice that $\forall n \geq 1$ the term $p_i(n)S_i - q_i(n) = d_i(n,2)$ is non positive. This immediately follows from Definition 15.

On the other hand $\forall n \geq 1$ the term $(p_i(n+1) - p_i(n))S_i - (q_i(n+1) - q_i(n)) = d_i(n,1)$ is positive. Since when $\alpha \in (r_i(n), r_i(n+1))$, we have, by Definition 15, $r_i(n) = \langle l_i(1), \ldots, l_i(n)\rangle < {S_i}^{-1} < \langle l_i(1), \ldots, l_i(n), l_i(n+1)-1\rangle = (p_i(n+1) - p_i(n))/(q_i(n+1) - q_i(n))$. This yields $(q_i(n+1) - q_i(n))/(p_i(n+1) - p_i(n)) < S_i$ which implies

$$(p_i(n+1) - p_i(n))S_i - (q_i(n+1) - q_i(n)) \tag{22}$$

$$> (p_i(n+1) - p_i(n))\frac{(q_i(n+1) - q_i(n))}{(p_i(n+1) - p_i(n))} - (q_i(n+1) - q_i(n)) = 0. \tag{23}$$

We assume in this proof that $\bar{l}_i > 0$ this means that the input process in queue i is given by *iii*, *iv* and *v* of Lemma 16. Although the input in queue i is not Sturmian, it will be shown that $w_i(t)$ the workload in the queue can be computed as a function of the factorization of $\overline{m}_\alpha$ using the (x,y) factorization sequence obtained by the ceiled expansion of S_i. On the other hand by Lemma 16 the input sequence is given by the $(y(n)(\overline{S}_{i-1}^{-1}), y(n+1)(\overline{S}_{i-1}^{-1}))$ factorization. More precisely in the input sequence of the ith queue the inter arrivals are equal except for the last factor $y(n+1)(\overline{S}_{i-1})^{-1})$. Therefore the only epochs where the queue i could be empty are at the end of $y(n+1)(\overline{S}_{i-1}^{-1})$. Hence we have to know the decomposition of the $y_i(n+1)$ and $x_i(n+1)$ factors as functions of the factors $y(n+1)((\overline{S}_{i-1})^{-1})$ and $y(n)((\overline{S}_{i-1})^{-1})$.

Let α such that $r_i(n) \leq \alpha \leq r_i(n+1)$ with $n+1 \leq \bar{l}_i$. By definition of $\bar{l}_i$ the coefficients of the ceiled expansion of $\overline{S}_{i-1}$ and those of the ceiled expansion S_i are equal until $n+1$. Thus the factors $y(n)(\overline{S}_{i-1})$ and $y_i(n)$ are identical as well as the factors $x(n)(\overline{S}_{i-1})$ and $x_i(n)$.

Corollary 18 implies that in the output process the number of exits during $y(n)(\overline{S}_{i-1})$ is $p(n)(\overline{S}_{i-1}) = p_i(n)$ and the time elapsed is $q(n)(\overline{S}_{i-1}) = q_i(n)$, similarly the number of exits during $x(n)(\overline{S}_{i-1})$ is $p(n)(\overline{S}_{i-1}) - p(n-1)(\overline{S}_{i-1}) = p_i(n) - p_i(n-1)$ and the time elapsed is $q(n)(\overline{S}_{i-1}) - q(n-1)(\overline{S}_{i-1}) = q_i(n) - q_i(n-1)$.

Thus the equality of the inter arrival and the fact that $S_i \geq \overline{S}_{i-1}$ imply that $w_i(t)$ the workload in queue i can only be null at the end of either $y_i(n)$ or $y_i(n+1)$. Lindley's Formula implies the

linearity of the workload which can be computed easily and we get that the workload after an $x_i(n)$ factor equals $d_i(n,1)$ and the maximal decrease equals $d_i(n,2)$.

From now on to facilitate the reading, the variables associated with $\overline{S}_{i-1}$ are denoted as if the service $\overline{S}_{i-1}$ were in queue $i-1$.

Let α be such that $r_i(\overline{l}_i) \leq \alpha \leq r_i(\overline{l}_i+1)$ We study the factors $x_i(\overline{l}_i+1)$ and $y_i(\overline{l}_i+1)$, these factors are composed as follows $x_i(\overline{l}_i+1) = y_{i-1}(\overline{l}_i+1)\left(y_{i-1}(\overline{l}_i)\right)^{l_i(\overline{l}_i+1)-l_{i-1}(\overline{l}_i+1)-1}$ and $y_i(\overline{l}_i+1) = y_{i-1}(\overline{l}_i+1)\left(y_{i-1}(\overline{l}_i)\right)^{l_i(\overline{l}_i+1)-l_{i-1}(\overline{l}_i+1)}$.

We have now to show that the workload at the end of sequences of the form $y_{i-1}(\overline{l}_i+1)y_{i-1}(\overline{l}_i)^k$, with $1 \leq k \leq l_i(\overline{l}_i+1) - l_{i-1}(\overline{l}_i+1)$, is positive.

These workloads are equal to $S_i(p_{i-1}(\overline{l}_i+1)+kp_{i-1}(\overline{l}_i)) - (q_{i-1}(\overline{l}_i+1)+kq_{i-1}(\overline{l}_i))$ until $k \leq k_0$ where k_0 is the smallest integer such that $S_i(p_{i-1}(\overline{l}_i+1)+k_0p_{i-1}(\overline{l}_i)) - (q_{i-1}(\overline{l}_i+1)+k_0q_{i-1}(\overline{l}_i)) \leq 0$. While $k \leq k_0$, the term $S_i(p_{i-1}(\overline{l}_i+1)+kp_{i-1}(\overline{l}_i)) - (q_{i-1}(\overline{l}_i+1)+kq_{i-1}(\overline{l}_i))$ can be rewritten as

$$S_i(p_i(\overline{l}_i+1) - (l_i(\overline{l}_i+1) - l_{i-1}(\overline{l}_i+1) - k)p_i(\overline{l}_i)) - (q_i(\overline{l}_i+1) - (l_i(\overline{l}_i+1) - l_{i-1}(\overline{l}_i+1) - k)q_i(\overline{l}_i)).$$

Since

$$\begin{aligned}
&S_i\Big(p_i(\overline{l}_i+1) - \big(l_i(\overline{l}_i+1) - l_{i-1}(\overline{l}_i+1) - 1\big)p_i(\overline{l}_i)\Big) - \Big(q_i(\overline{l}_i+1) - \big(l_i(\overline{l}_i+1) - l_{i-1}(\overline{l}_i+1) - 1\big)q_i(\overline{l}_i)\Big) \\
&\ldots \geq S_i\Big(p_i(\overline{l}_i+1) - (l_i(\overline{l}_i+1) - l_{i-1}(\overline{l}_i+1) - k)p_i(\overline{l}_i)\Big) - \Big(q_i(\overline{l}_i+1) - \big(l_i(\overline{l}_i+1) - l_{i-1}(\overline{l}_i+1) - k\big)q_i(\overline{l}_i)\Big) \\
&\ldots \geq S_i(p_i(\overline{l}_i+1) - p_i(\overline{l}_i)) - (q_i(\overline{l}_i+1) - q_i(\overline{l}_i)) > 0 \geq S_ip_i(\overline{l}_i+1) - q_i(\overline{l}_i+1).
\end{aligned}$$

Hence $k_0 = l_i(\overline{l}_i+1) - l_{i-1}(\overline{l}_i+1)$ and $y_i(\overline{l}_i+1)$ is indeed the shortest sequence for which the workload is null at the end of the sequence. Therefore the result is proved for $x_i(\overline{l}_i+1)$ and $y_i(\overline{l}_i+1)$.

When $n = \overline{l}_i + 2$, Theorem 11 yields the result. Indeed applying the (x-y) factor composition at order $\overline{l}_i+2$ gives

$$x_i(\overline{l}_i+2) = x_i(\overline{l}_i+1)y_i(\overline{l}_i+1)^{(\overline{l}_i+2)-2} \qquad \text{and} \qquad y_i(\overline{l}_i+2) = x_i(\overline{l}_i+1)y_i(\overline{l}_i+1)^{(\overline{l}_i+2)-1}.$$

Since the result holds up to $\overline{l}_i+1$ and since the composition of the factors $x_i(\overline{l}_i+2)$ and $y_i(\overline{l}_i+2)$ amounts to adding a load before the factors $x_i(\overline{l}_i+1)$ and $y_i(\overline{l}_i+1)$, then in order to show the result it suffices to check the non negativity of the workload after $x_i(\overline{l}_i+2)$ with a null initial load and the non positivity of the workload $y_i(\overline{l}_i+2)$ with a null initial load. They respectively equal $S_i\big(p_i(\overline{l}_i+2) - p_i(\overline{l}_i+1)\big) - (q_i(\overline{l}_i+2) - q_i(\overline{l}_i+1))$ and $\big(S_ip_i(\overline{l}_i+2) - q_i(\overline{l}_i+2)\big)^+ = 0$. Therefore the result still holds for the step $\overline{l}_i+1$.

Using the same inductive argument for order larger than $\overline{l}_i+2$ concludes the proof.

9 Appendix: Proof of theorem 22

Without loss of generality it is assumed in the following proof that $\overline{S}_{i-1} = S_{i-1}$ in order to facilitate the reading. Let us recall that w_i^t is the workload in the i-th queue at epoch t. All the following formulas come from the linearity of the workload until the queue becomes empty.

Case $n \leq \overline{l}_i$: the number of customers admitted is $p_i(n)$. The first customer does not wait and the k^{th} customer waits during $(k-1)(S_i - S_{i-1})$ before being treated. The sum of all the waiting times follows (13).

Case $n = \overline{l}_i + 1$: from Lemma 14 we get

$$\begin{aligned}
K_i(\overline{m}_{r_i(n)}, 0) &= K_i(y_{i-1}(n), 0) + K_i(y_{i-1}(n-1), w_i^{p_{i-1}(n)}) + K_i(y_{i-1}(n-1), w_i^{p_{i-1}(n)+p_{i-1}(n-1)})) + \\
&\quad \ldots + K_i(y_{i-1}(n-1), w_i^{p_{i-1}(n)+(l_i(n)-l_{i-1}(n)-1)p_{i-1}(n-1)}), \\
&= K_i(y_{i-1}(n), 0) + p_{i-1}(n-1) w_i^{p_{i-1}(n)} + K_i(y_{i-1}(n-1), 0) + \ldots \\
&\quad \ldots + p_{i-1}(n-1) w_i^{p_{i-1}(n)+(l_i(n)-l_{i-1}(n)-1)p_{i-1}(n-1)} + K_i(y_{i-1}(n-1), 0), \\
&= K_i(y_{i-1}(n), 0) + (l_i(n) - l_{i-1}(n)) K_i(y_{i-1}(n-1), 0) \\
&\quad + (l_i(n) - l_{i-1}(n)) p_{i-1}(n-1) \big(p_{i-1}(n) S_{i-1} - q_{i-1}(n)\big) \\
&\quad + p_{i-1}(n-1) \big(p_{i-1}(n-1) S_{i-1} - q_i(n-1)\big) (l_i(n) - l_{i-1}(n)) \frac{l_i(n) - l_{i-1}(n)}{2}.
\end{aligned}$$

After some simplifications we get

$$\begin{aligned}
K_i(\overline{m}_{r_i(n)}, 0) &= \frac{S_i - S_{i-1}}{2} (p_i^2(n) - p_i(n)) + (l_i(n) - l_{i-1}(n)) p_{i-1}(n-1) \times \\
&\Big[S_i \big(p_{i-1}(n) + \frac{1}{2} p_{i-1}(n-1) \, (l_i(n) - l_{i-1}(n) - 1)\big) - \big(q_{i-1}(n) + \frac{1}{2} q_{i-1}(n-1)(l_i(n) - l_{i-1}(n) - 1)\big) \Big].
\end{aligned}$$

Introducing $K_i(\overline{m}_{r_i(n-1)}, 0)$ and $K_i(\overline{m}_{r_i(n-2)}, 0)$ gives (14).
Case $n > \overline{l}_i + 1$.

$$\begin{aligned}
K_i(\overline{m}_{r_i(n)}, 0) &= K_i(x_i(n-1), 0) + K_i(y_i(n-1), w_i^{p_i(n-1)-p_i(n-2)}) + \ldots \\
&\quad \ldots + K_i(y_i(n-1), w_i^{p_i(n-1)+(l_i(n)-2)p_i(n-2)}), \\
&= l_i(n) K(\overline{m}_{r_i(n-1)}, 0) - K(\overline{m}_{r_i(n-2)}, 0) + \big(p_i(n-1)(l_i(n) - 1) \\
&\quad - p_i(n-2)\big) \Big(\big(p_i(n-1) - p_i(n-2)\big) S_i - \big(q_i(n-1) - q_i(n-2)\big)\big) \Big) + \\
&\quad p_i(n-1) \big(p_i(n-1) S_i - q_i(n-1)\big) (l_i(n) - 1) \frac{(l_i(n) - 2)}{2}.
\end{aligned}$$

Reordering yields (15).

From Lemma 14 and using Lemma 16 of [9] the waiting time could be computed in function of the waiting time during $\overline{m}_{r_i(n+1)}$ and $\overline{m}_{r_i(n)}$. Henceforth $W_i(\overline{m}_\alpha)$ is given by the equation

$$W_i(\overline{m}_\alpha) = \frac{\alpha_n K_i(\overline{m}_{r_i(n+1)}, 0) + (1 - \alpha_n l_i(n+1) K_i(\overline{m}_{r_i(n)}, 0)}{\alpha_n p_i(n+1) + (1 - \alpha_n l_i(n+1)) p_i(n)}.$$

A more detailed proof is given in [9].

We use Equation (9) to replace α_n by its value.

10 Appendix: Proof of Lemma 33

In order to simplify the reading of this proof since most of the variables used correspond to the i-th queue the index i of the queue is skipped when no confusion is possible. Further it is assumed that $S_{i-1} = \overline{S}_{i-1}$.

It suffices to show the result for one of the two functions the rest follows immediately. The growth rate of the function

$$n \to \frac{K(\overline{m}_{r(n)}, 0)}{q(n)}$$

is

$$\Big(\frac{K(\overline{m}_{r(n)},0)}{q(n)}-\frac{K(\overline{m}_{r(n-1)},0)}{q(n-1)}\Big)\Big(\frac{p(n)}{q(n)}-\frac{p(n-1)}{q(n-1)}\Big)^{-1}.$$

which equals $q(n-1)K(\overline{m}_{r(n)},0)-q(n)K(\overline{m}_{r(n-1)},0)$.

When $n\leq \bar{l}_i$, the term $q(n-1)K(\overline{m}_{r(n)},0)-q(n)K(\overline{m}_{r(n-1)},0)-q(n-2)K(\overline{m}_{r(n-1)},0)-q(n-1)K(\overline{m}_{r(n-2)},0)$ becomes using (13)

$$\frac{(S_i-S_{i-1})}{2}\Big(q(n-1)p^2(n)-q(n-1)p(n)-q(n)p^2(n-1)+q(n)p(n-1)-q(n-2)p^2(n-1)+ q(n-2)p(n-1)+q(n-1)p^2(n-2)-q(n-1)p(n-2)\Big).$$

Since $\forall n\geq 0$ we have $p(n)q(n-1)-p(n-1)q(n)=1$, this is equivalent to

$$\frac{(S_i-S_{i-1})}{2}q(n-1)\Big(p^2(n)-l(n)p^2(n-1)+p^2(n-2)\Big).$$

On the other hand, using (8) it comes

$$\begin{aligned}
&p^2(n)-l(n)p^2(n-1)+p^2(n-2)=p^2(n)+p^2(n-2)-p(n-1)\Big(p(n)+p(n-2)\Big)\\
&=\Big(p(n)-p(n-2)\Big)^2-p(n)\Big(p(n-1)-p(n-2)\Big)+p(n-2)\Big(p(n)-p(n-1)\Big)\\
&=\Big(p(n)-p(n-1)+p(n-1)-p(n-2)\Big)^2-p(n)\Big(p(n-1)-p(n-2)\Big)+p(n-2)\Big(p(n)-p(n-1)\Big)\\
&=\Big(p(n)-p(n-1)+p(n-2)\Big)\Big(p(n)-2p(n-1)+p(n-2)\Big)+2\big(p(n)-p(n-1)\big)\big(p(n-1)-p(n-2)\big)\\
&\geq 0.
\end{aligned}$$

The last inequality being obtained since $\forall n$ we have $p(n)-p(n-1)\geq 0$ and $p(n)-p(n-1)+p(n-2)\geq p(n)-2p(n-1)+p(n-2)\geq p(n)-l(n)p(n-1)+p(n-2)=0$. Moreover $q(0)\big(p^2(1)-p(1)\big)-q(1)\big(p^2(0)-p(0)\big)=0$. This gives the monotonicity and the convexity.

When $n=\bar{l}_i+1$ then using Equations (14) and (8) it comes

$$\begin{aligned}
&q_i(n-1)K_i(\overline{m}_{r_i(n)},0)-q_i(n)K_i(\overline{m}_{r_i(n-1)},0)=\\
&\qquad q_i(n-1)l_i(n)K_i(\overline{m}_{r_i(n-1)},0)-l_i(n)q_i(n-1)K_i(\overline{m}_{r_i(n-1)},0)\\
&\qquad -q_i(n-1)K_i(\overline{m}_{r_i(n-2)},0)+q_i(n-2)K_i(\overline{m}_{r_i(n-1)},0)+q_i(n-1)\times\\
&\qquad\Big[\frac{S_i}{2}\Big(l_i(n)p_i(n-1)\big(l_i(n)p_i(n-1)-p_i(n-1)-2p_i(n-2)\big)+2p_i^2(n-2)\Big)\\
&-\frac{S_{i-1}}{2}\big(l_{i-1}^2(n)p_{i-1}^2(n-1)-l_{i-1}(n)p_{i-1}^2(n-1)-2l_{i-1}(n)p_{i-1}(n-1)p_{i-1}(n-2)+2p_{i-1}^2(n-2)\big)\\
&\qquad -\big(l_i(n)-l_{i-1}(n)\big)p_{i-1}(n-1)\Big(\frac{1}{2}q_i(n-1)\big(l_i(n)+l_{i-1}(n)-1\big)-q_i(n-2)\Big)\Big],
\end{aligned}$$

which can be transformed, since $p_i(n-1) = p_{i-1}(n-1)$ and $p_i(n-2) = p_{i-1}(n-2)$, in

$$\begin{aligned}
q_i(n-1)K_i(\overline{m}_{r_i(n)},0) - q_i(n)K_i(\overline{m}_{r_i(n-1)},0) = \\
q_i(n-2)K_i(\overline{m}_{r_i(n-1)},0) - q_i(n-1)K_i(\overline{m}_{r_i(n-2)},0) + q_i(n-1) \\
\Bigg[\frac{S_i}{2}\big(l_i(n) - l_{i-1}(n)\big)p_i(n-1)\Big(\big(l_i(n)+l_{i-1}(n)\big)p_i(n-1) - 2p_i(n-2)\Big) \\
- \big(l_i(n) - l_{i-1}(n)\big)p_i(n-1)\Big(\frac{1}{2}q_i\big(n-1\big)(l_i(n)+l_{i-1}(n)\big) - q_i(n-2)\Big) \\
+ \frac{S_i - S_{i-1}}{2}\Big(p_i^2(n) - l_i(n)p_i^2(n-1) + p_i^2(n-2)\Big)\Bigg].
\end{aligned}$$

As shown before the term $\frac{S_i - S_{i-1}}{2}\big(p_i^2(n) - l_i(n)p_i^2(n-1) + p_i^2(n-2)\big)$ is non negative. Let us study now the sign of the term $S_i\Big(\big(l_i(n)+l_{i-1}(n)\big)p_i(n-1) - 2p_i(n-2)\Big) - \Big(q_i(n-1)\big(l_i(n) + l_{i-1}(n)\big) - 2q_i(n-2)\Big)$. This term can be rewritten and minorized knowing that when $n = \bar{l}_i + 1$ we have $l_i(n) - l_{i-1}(n) \geq 1$. Hence, we obtain

$$\begin{aligned}
2(q_i(n-2) - p_i(n-2)S_i) + (p_i(n-1)S_i - q_i(n-1))\big(l_i(n) + l_{i-1}(n) - 1\big) \\
\geq 2\big(q_i(n-2) - p_i(n-2)S_i\big) + \big(p_i(n-1)S_i - q_i(n-1)\big)(2l_i(n) - 1) \\
\geq 2\Big(S_i\big(p_i(n) - p_i(n-1)\big) - \big(q_i(n) - q_i(n-1)\big)\Big) > 0,
\end{aligned}$$

with the use of (8) and (23). Henceforth the growth rate in $n = \bar{l}_i + 1$ is greater than this one in $n = \bar{l}_i$.

When $n \geq \bar{l}_i + 2$, owing to (15) and (8) it comes

$$\begin{aligned}
q(n-1)K(\overline{m}_{r(n)},0) - q(n)K(\overline{m}_{r(n-1)},0) = \\
q(n-2)K(\overline{m}_{r(n-1)},0) - q(n-1)K(\overline{m}_{r(n-2)},0) + q(n-1)\Big[p(n-1)\big(l(n)-2\big) \\
\Big(\big(p(n-1) - p(n-2)\big)S - \big(q(n-1) - q(n-2)\big) + \frac{l(n)-1}{2}\big(p(n-1)S - q(n-1)\big)\Big) \\
+ \big(p(n-1) - p(n-2)\big)d(n-1,1).
\end{aligned}$$

If $l(n) = 2$ then $q(n-1)K(\overline{m}_{r(n)},0) - q(n)K(\overline{m}_{r(n-1)},0) - q(n-2)K(\overline{m}_{r(n-1)},0) - q(n-1)K(\overline{m}_{r(n-2)},0) = (p(n-1) - p(n-2))\, d(n-1,1)$. This last term is positive by (23).

When $l(n) \geq 3$ we focus on $d(n-1,1) + \frac{l(n)-1}{2}d(n-1,2)$ which is equivalent to

$$\frac{l(n)+1}{2}\big(p(n-1)S - q(n-1)\big) - \big(p(n-2)S - q(n-2)\big).$$

Since as soon as $l(n) \geq 3$ it comes $\frac{l(n)+1}{2} \leq l(n) - 1$, this implies that

$$\begin{aligned}
\frac{l(n)+1}{2}\big(p(n-1)S - q(n-1)\big) - \big(p(n-2)S - q(n-2)\big) \\
\geq \big((l(n)-1)p(n-1) - p(n-2)\big)S - \big((l(n)-1)\,q(n-1) - q(n-2)\big) \\
\geq d(n-1,1) > 0.
\end{aligned}$$

Therefore *i)* and *ii)* are proved.

Concerning *iii)*, We are only interested by the numbers which are greater than $\overline{l}_i + 2$ therefore we use Formula (15) and Formula (11). We focus on the minorization of $q(n-1)K(\overline{m}_{r(n)}, 0) - q(n)K(\overline{m}_{r(n-1)}, 0)$ by $\Sigma_{j=\overline{l}_i+2}^{n} q(j-1)\big(p(j) - p(j-1)\big)d(j,1)$.

Therefore it suffices to show that the terms of the series are minorized by a positive number. We consider there is an infinite number of cusps thus by the definition of the ceiled expansion $r(j) = \langle l(1), \ldots, l(j)\rangle < S^{-1} < \langle l(1), \ldots, l(j), l(j+1) - 1\rangle = (p(j+1) - p(j))/(q(j+1) - q(j))$. It follows that $(q(j+1) - q(j)) \,/\, (p(j+1) - p(j)) < S$, hence

$$\begin{aligned} d(j,1) &> \big(p(j) - p(j-1)\big)\frac{q(j+1) - q(j)}{p(j+1) - p(j)} - q(j) + q(j-1), \\ &> \frac{1}{p(j+1) - p(j)}\Big(q(j-1)p(j+1) - p(j-1)q(j+1)\Big) = \frac{l(j+1)}{p(j+1) - p(j)}. \end{aligned}$$

Therefore $q(j-1)\,(p(j) - p(j-1))\,d(j,1)$ can be minorized by

$$\begin{aligned} \frac{q(j-1)l(j+1)\big(p(j) - p(j-1)\big)}{p(j+1) - p(j)} &= \frac{q(j-1)l(j+1)\big(l(j)p(j-1) - p(j-2) - p(j-1)\big)}{(l(j+1)-1)\big(l(j)p(j-1) - p(j-2)\big) - p(j-1)} \\ &> \frac{q(j-1)l(j+1)\big(l(j)-2\big)p(j-1)}{l(j+1)l(j)p(j-1)} > \frac{q(j-1)(l(j)-2)}{l(j)}. \end{aligned}$$

Two cases may occurs either $l(j) = 2$ or $l(j) > 2$. If $l(j) \geq 3$, since the function $x \to (x-2)/(x)$ is increasing then

$$q(j-1)\,(p(j) - p(j-1))\,d(j,1) > q(j-1)/3 > \frac{2}{3}.$$

If $l(j) = 2$, since $(p(j) - p(j-1))\,d(j,1) = (p(j-1) - p(j-2))\,d(j-1,1)$ and since at the first step $(p(1) - p(0))\,d(1,1) = S - l(1) + 1 > 0$ then in the worst case

$$q(j-1)\,(p(j) - p(j-1))\,d(j,1) > q(j-1)(1-\alpha_1) > 0.$$

This leads to *iii)*.

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhˆone-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

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