

Balanced Sequences and Optimal Routing

EITAN ALTMAN

INRIA, Sophia Antipolis, Cedex, France

BRUNO GAUJAL

LORIA, Villers-des-Nancy, France

AND

ARIE HORDIJK

Leiden University, Leiden, The Netherlands

Abstract. The objective pursued in this paper is two-fold. The first part addresses the following combinatorial problem: is it possible to construct an infinite sequence over n letters where each letter is distributed as “evenly” as possible and appears with a given rate? The second objective of the paper is to use this construction in the framework of optimal routing in queuing networks. We show under rather general assumptions that the optimal deterministic routing in stochastic event graphs is such a sequence.

Categories and Subject Descriptors: B.4.4 [**Input/Output and Data Communications**]: Performance Analysis and Design Aids; B.8.2 [**Performance and Reliability**]: Performance Analysis and Design Aids

General Terms: Performance

Additional Key Words and Phrases: Balanced sequences, multimodularity, optimal control, stochastic event graphs

The research of A. Hordijk was partially done while he was on sabbatical leave at INRIA, Sophia-Antipolis; it has been partially supported by the Ministère Français de l'Éducation Nationale et de l'Enseignement Supérieur et de la Recherche and by the cooperation between France and the Netherlands, Van Gogh N.98001.

Authors' addresses: E. Altman, INRIA, BP 93, 2004 Route des Lucioles, 06902 Sophia Antipolis Cedex, France. E-mail: altman@sophia.inria.fr. URL: <http://www.inria.fr:80/mistral/personnel/Eitan.Altman/me.html>; B. Gaujal, LORIA, 615 route du jardin botanique, BP 101, F-54606 Villers-les-Nancy, France. E-mail: Bruno.Gaujal@loria.fr.; A. Hordijk, Dept. of Mathematics and Computer Science, Leiden University, P.O. Box 9512, 2300RA Leiden, The Netherlands. E-mail: hordijk@wi.leidenuniv.nl.

Permission to make digital/hard copy of part or all of this work for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery (ACM), Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee.

© 2000 ACM 0004-5411/00/0700-0752 \$05.00

1. Introduction

It is a rather general problem to consider a system with multiple resources and tasks. Tasks can be performed by any resource and arrive in the system sequentially. The problem is to construct a *routing* of the tasks to the resources to minimize a given cost function. Such models are common in multiprocessor systems and communication networks, where the cost function may be the combined load in the resources.

In this paper, we show that under rather general assumptions, the optimal routing policy in terms of expected average workload in each resource is given by a *balanced sequence*, that is, a sequence in which the option to route towards a given resource, is taken in an evenly distributed fashion. To show the relevance of this type of problem, let us briefly review the recent literature on load balancing. Load balancing in a distributed multiprocessor computer system has become an important issue to improve their performance. Many papers have been devoted to the load balancing problem, and for an overview, we refer to Boel and Van Schuppen [1989] and Hariharan et al. [1990]. Let us assume there is a centralized controller, and the information available to the controller determines the type of control which can be used. In dynamic feedback control (or close loop control), the information on the system (e.g., queue sizes) increases as time runs. In static or open loop control, the central controller only knows his past actions. Clearly, the closed loop models have a better performance than the open loop ones.

In the dynamic control setting, it is well known that under various assumptions and cost structures, the “join the shortest queue” policy is optimal for homogeneous processors (i.e., all processors are stochastically identical). For nonhomogeneous processors, an optimal load balancing policy has not been found. Only partial optimality results are available (see Hordijk and Koole [1992] for results on shortest queue policies and Altman and Stidham [1995] for results on the monotonicity structure of optimal policies).

For open loop control, probabilistic routing and pattern allocation have been studied. Again, in the case of homogeneous processors, the optimality of equal probabilities for probabilistic routing is established in Chang et al. [1990] and Koole [1996]. The round robin routing was proved optimal in the case of pattern allocation in Liu and Righter [1998]. For the nonhomogeneous case, the problem of finding the optimal pattern allocation is generally considered difficult. Approximations of optimal allocations are found in Bar-Noy et al. [1998], Combé and Boxma [1994], and Hordijk et al. [1994].

In this paper, we are concerned with the open-loop control and more particularly with optimal pattern allocation problem. The main feature of the paper is to relate this problem with the combinatorial problem of finding balanced sequences and with the multimodularity analysis of the cost function, done in previous papers. The main advantages of this approach is to extend the existing results into several directions.

- In the homogeneous case, the optimality of round robin is shown for a certain class of networks of queues. Up to the authors’ knowledge, this problem was only addressed for single queues in the literature.
- The stochastic assumptions that are used here are very general. Renewal assumptions on the arrival process are replaced by mere stationarity assump-

tions. Also, independence and identical distributions for the service times are replaced by stationarity as well.

- Finally, up to the knowledge of the authors, the structure of the optimal policy in the nonhomogeneous case was not known for the special models discussed in Section 4 and for models with optimal rates which are balanceable. Also our proof uniformizes the cases of homogeneous processors with the special cases of nonhomogeneous processors.

The first part of the paper is essentially based on word combinatorics and uses its specific vocabulary. This part relies heavily on results in Graham [1973] and Morse and Hedlund [1940]. The problem of balanced sequences and exactly covering sequences has been studied using combinatorial as well as arithmetic techniques.¹

However, in these studies, the analysis of balanced sequences was done per se and did not present any motivations or applications. In particular, the use of these constructions in discrete event dynamic systems may not have been discovered before the seminal work of Hajek [1985], where he solves the one dimensional case problem, further generalized in Altman et al. [1997a]. The goal of this paper is to extend the same type of results to a multidimensional case, which is surprisingly more difficult and of different nature as the one-dimensional case.

This is done in the second part of this paper, where we will show an application of balanced sequences for routing problems. This part uses results from convex analysis, mainly from Hajek [1985] and Altman et al. [1997a; 1997b]. The main results that are used here are of two different kinds. First, we use the fact that the workload as well as the waiting time of customers entering a $(\max, +)$ linear system are multimodular functions, under fairly general assumption (stationarity of the arrival process and of the service times) [Altman et al. 1997a]. We pose the problem of routing in terms of time average of multimodular sequences. Then, we develop the required theory for the case where the performance for each subsystem is stochastic but its expectation is multimodular. This extends the deterministic analysis that we did in Altman et al. [1997b], that proves that multimodular functions are minimized by regular sequences. The superposition of several regular sequences being a balanced sequence, this is the basis of the main result of this paper.

It is interesting to exhibit this link between balanced sequences and scheduling problems, such as routing among several systems.

More precisely, the paper is structured as follows: In the second section, we introduce a formal definition of *balanced sequence* and we present an overview on their properties. Section 3 shows the link between the notion of balanced sequences and the optimal scheduling in networks. It is also used to establish the optimality of balanced sequences for routing customers in a multiple queue system. Section 4 presents special cases for which the optimal rates can be computed.

¹ See, for example, Tijdeman [1996], Wilf [1994], Newman [1971], Fraenkel [1973], and Morikawa [1982–1995].

2. Balanced Sequences

In this section, we will present the notion of balanced sequences, which is closely related to the notion of *Sturmian* sequences [Lothaire 1991] as well as exactly covering sequences. This presentation is not exhaustive and many other related articles can be consulted for further investigation on this topic.² Although the section is self-contained and presents several results that are of interest by their own, we mostly focus on the *rate problem* (see Problem 2), which will be used in the application section (Sect. 3). The main new results established in this section are those concerning problem P2. The proof of Theorem 2.20 is improved from previous presentations and Theorem 2.21 is new. All the propositions given in Section 2.8 are also new. As a general fact, whenever a result is not original, this will be mentioned in the paper.

2.1. PRELIMINARIES. Let \mathcal{A} be a finite alphabet and $\mathcal{A}^{\mathbb{Z}}$ the set of sequences³ defined on \mathcal{A} .

If $u \in \mathcal{A}^{\mathbb{Z}}$, then a *word* W of u is a finite subsequence of consecutive letters in u : $W = u_a u_{a+1} \cdots u_{a+k-1}$. The integer k is the *length* of W and will be denoted $|W|$.

If $a \in \mathcal{A}$, $|W|_a$ is the number of a 's in the word W .

Definition 2.1. The sequence $u \in \mathcal{A}^{\mathbb{Z}}$ is *a-balanced* if for any two words W and W' in u of same length,

$$-1 \leq |W|_a - |W'|_a \leq 1.$$

The sequence u is *balanced* if it is *a-balanced* for each $a \in \mathcal{A}$.

If $a \in \mathcal{A}$, we also define the *indicator* in u of the letter a as the function $\delta_a(u): \mathbb{Z} \rightarrow \{0, 1\}^{\mathbb{Z}}$ by, $\delta_a(u)_i = 1$ if $u(i) = a$ and 0 otherwise. The *support* in u of the letter a is the set $S_a = \{i \in \mathbb{Z} : \delta_a(u)_i = 1\}$.

For any real number x , $\lfloor x \rfloor$ will denote the largest integer smaller than or equal to x and $\lceil x \rceil$ will denote that smallest integer larger than or equal to x .

LEMMA 2.2. *If a sequence $u \in \mathcal{A}^{\mathbb{Z}}$ is balanced, then for any $a \in \mathcal{A}$, there exists a real number p_a , such that*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^n \delta_a(u)_i = \lim_{n \rightarrow -\infty} \frac{1}{n} \sum_{i=n}^0 \delta_a(u)_i = p_a.$$

p_a is called the *rate* of $\delta_a(u)$.

PROOF. Let us define $s_n = \sum_{i=0}^n \delta_a(u)_i$ and remark that $s_n + s_m - 1 \leq s_{n+m} \leq s_n + s_m + 1$. The rest of the proof is classical by subadditivity arguments. The proof for $\{\delta_a(u)_n\}_{n \leq 0}$ is similar. The fact that both limits coincide is obvious. \square

² See, for example, Vuillon [1998], Hubert [1996], Arnoux et al. [1994], Baryshnikov [1995], Tijdeman [1996], and Fraenkel [1973].

³ Such sequences with domain \mathbb{Z} are often called *bisequences*.

Note that the sum of the rates for all letters in a sequence u is one.

$$\sum_{a \in \mathcal{A}} p_a = 1.$$

Now, we can present the main result which is the foundation of the theory of balanced sequences. We follow the presentation given in Tijdeman [1998b].

THEOREM 2.3 (MORSE AND HEDLUND [1940]). *A sequence $d \in \{0, 1\}^{\mathbb{Z}}$ is balanced with asymptotic rate $0 < p = 1/\alpha$ for letter 1 if and only if the support S_1 of d satisfies one of the following cases.*

(a) (irrational case) p is irrational and there exists $\phi \in \mathbb{R}$ such that

$$S_1 = \{\lfloor i\alpha + \phi \rfloor\}_{i \in \mathbb{Z}} \quad \text{or} \quad S_1 = \{\lceil i\alpha + \phi \rceil\}_{i \in \mathbb{Z}}.$$

(b) (periodic case) $p \in \mathbb{Q}$ and there exists $\phi \in \mathbb{Q}$ such that

$$S_1 = \{\lfloor i\alpha + \phi \rfloor\}_{i \in \mathbb{Z}}.$$

(c) (skew case) $p \in \mathbb{Q}$ ($p = k/n$, $k, n \in \mathbb{N}$) and there exists $m \in \mathbb{Z}$ such that

$$S_1 = \left\{ \left\lfloor \frac{in}{k} + m \right\rfloor \right\}_{i < k} \cup \left\{ \left\lfloor \frac{in}{k} - \frac{1}{k} + m \right\rfloor \right\}_{i > 0}$$

or

$$S_1 = \left\{ \left\lfloor \frac{in}{k} + m \right\rfloor \right\}_{i > 0} \cup \left\{ \left\lfloor \frac{in}{k} - \frac{1}{k} + m \right\rfloor \right\}_{i < k}.$$

As for the case $p = 0$, it includes two balanced sequences, namely the sequence where S_1 is the empty set and the sequence where S_1 is a singleton.

Sequences for which the support is of the form $S_1 = \{\lfloor i\alpha + \phi \rfloor\}_{i \in \mathbb{Z}}$ are called *bracket sequences*. This theorem shows the relation that exists between balanced sequences and *bracket sequences*. The irrational case is the easiest case and can be characterized (see Theorem 2.18). The two rational cases are more difficult to study.

Note that the skew sequences are not periodic but are ultimately periodic. In our applications (see Sections 3 and 4), we will be using (ultimately) regular sequences, as defined below.

Definition 2.4. A sequence $d \in \{0, 1\}^{\mathbb{Z}}$ is *regular* (respectively, *ultimately regular*) if there exist two real numbers θ and p (and an integer k , respectively) such that for all n ($n \geq k$), $d_n = \lfloor (n+1)p + \theta \rfloor - \lfloor np + \theta \rfloor$.

Note that $d_n = \lfloor (n+1)p + \theta \rfloor - \lfloor np + \theta \rfloor$ is equivalent to $S_1 = \{\lceil n \frac{1}{p} + \phi \rceil\}_{n \in \mathbb{Z}}$ with $\phi = -1 - \theta/p$.

THEOREM 2.5. Let $u \in \mathcal{A}^{\mathbb{Z}}$.

- (i) If $\delta_a(u)$ is regular for all $a \in \mathcal{A}$, then u is balanced.
- (ii) If u is balanced, then $\delta_a(u)$ is ultimately regular for all $a \in \mathcal{A}$.

PROOF

- (i) is straightforward.
- (ii) is a direct consequence of Theorem 2.3, since in all three cases, the sequence $\delta_a(u)$ is ultimately regular. An elementary proof of (ii) that does not use Theorem 2.3 can be found in Tijdeman [1995]. \square

In Loeve [1995], the sequence u is said to have the reduction property if $\delta_a(a)$ is regular for all $a \in \mathcal{A}$. Corollary 2.5 shows that u has the reduction property (ultimately) if and only if it is balanced.

2.2. CONSTANT GAP SEQUENCES. Constant gap sequences are strongly balanced sequences, in the following sense.

Definition 2.6. A sequence G is *constant gap* if for any letter a , $\delta_a(G)$ is periodic, with a period of the form $0 \cdots 010 \cdots 0$.

Note that this explains the fact that G is said to have constant gaps for the letter a , since each a is separated from the next a in G by a constant number of letters.

PROPOSITION 2.7. *Constant gap sequences are balanced.*

PROOF. For each letter a , $\delta_a(G)$ is of the form $(0^n 10^m)^\omega$. Therefore, $\delta_a(G)$ is regular with $p = 1/(m + n + 1)$ and $\theta = m/(m + n + 1)$. Using the characterization of balanced sequences given in Theorem 2.5, this shows that G is balanced. \square

PROPOSITION 2.8. *Constant gap sequences are periodic.*

PROOF. For each letter a , $\delta_a(G)$ is periodic with period p_a . The period of G is $\text{lcm}(p_a, a \in \mathcal{A})$. \square

In the next lemma, we give a characterization of constant gap sequences that stresses the fact that constant gap is some kind of strong balance.

PROPOSITION 2.9. *G is constant gap if and only if, for any two finite words, W and W' included in G with $\|W\| - \|W'\| \leq 1$, then for each letter a , $\|W\|_a - \|W'\|_a \leq 1$.*

PROOF. Let a be a letter in the alphabet.

First, assume that G is constant gap. If $\|W\|_a - \|W'\|_a \geq 2$, then, necessarily, $\|W\| - \|W'\| \geq 2$.

Conversely, let $W = aUa$ and $W' = aU'a$ be any two words in G with no a in the subwords U and U' . If $|U| \geq |U'| + 1$, then we have $\|U\| - \|U'\| \leq 1$ and $\|U\|_a - \|U'\|_a = 2$. This is a contradiction. Therefore, $|U| = |U'|$ and G is constant gap. \square

Since a constant gap sequence is balanced, each letter appears with a given rate in the sequence. Note however that since a constant gap sequence is necessarily periodic, the rate of each letter is rational.

As we will do in Section 2.4 for the case of balanced sequences, we now address the following question:

PROBLEM 1. *Given a set (p_1, \dots, p_N) , with $p_1 + \dots + p_N = 1$, is it possible to construct a constant gap sequence on N letters with rates (p_1, \dots, p_N) ?*

We will not solve this problem for a general N , since it is NP-complete (see Bar-Noy et al. [1998]).

REMARK 2.10. *As mentioned by a referee, one can use the following facts to show that the set of rates solving Problem 1 is finite. Let (p_1, \dots, p_N) be a set of rates solving problem P_1 . We order the rates so that $1 > p_1 \geq \dots \geq p_N$. Then necessarily,*

—For all i , p_i^{-1} is an integer.

—For all i ,

$$p_{i-1} \geq p_i \geq \frac{1 - (p_1 + \dots + p_{i-1})}{N - i + 1}.$$

Note that combining both items leaves only a finite number of possibilities for (p_1, \dots, p_N) .

We will now give some properties of the set (p_1, \dots, p_N) which will be useful in the following. A characterization of such (p_1, \dots, p_N) is given in Wilf [1994], under the name *exact covering sequences*, but it does not provide an algorithm.

Definition 2.11. The set of couples $\{(\theta_i, g_i), i = 1 \dots N\}$ is called an *exact covering sequence* if for every nonnegative integer n , there exists one and only one $1 \leq i \leq N$ such that $n = \theta_i \bmod g_i$.

As a general remark, note that (p_1, \dots, p_N) are rational numbers of the form $p_i = k_i/d$, with d the smallest period of G . Therefore, we have, $\sum_i k_i/d = 1$ and for each i , k_i divides d . By definition of the rates, we also have $p_a = 1/q_a$ for all letters.

We have the following result.

PROPOSITION 2.12. *The rates (p_1, \dots, p_N) are constant gap if there exist N numbers called phases, $\theta_1, \dots, \theta_N$ such that the couples $\{(\theta_i, q_i), i = 1 \dots N\}$ form an exact covering sequence.*

PROOF. This property is a simple rewriting of the fact that each letter a_i in a constant gap sequence appears every $\{\theta_i + kq_i, k \in \mathbb{N}\}$. \square

Now, suppose that $\{(\theta_i, q_i), i = 1 \dots N\}$ is an exact covering sequence. Then in the series

$$S(x) \stackrel{\text{def}}{=} \sum_{i=1}^k \sum_{k \geq 0} x^{\theta_i + kq_i},$$

the coefficient of x^n in this series is equal to 1, $\forall n \geq 0$. Therefore, we have

$$S(x) = \sum_{i=1}^k \frac{x_i^{\theta_i}}{1 - x^{q_i}} = \frac{1}{1 - x}.$$

Using this characterization we have the following interesting property, which was proved in Newman [1971].

LEMMA 2.13. Assume $\{(\theta_i, q_i), i = 1 \cdots N\}$ is an exact covering sequence and that $P = \max_i q_i$. Then P appears at least twice in the set q_1, \dots, q_N .

PROOF. The proof given here is similar to the discussion in Wilf [1994] on exact covering sequences. Let $w \stackrel{\text{def}}{=} \exp^{2i\pi/r}$ for some integer $r > 1$. By definition, w is a primitive r th root of one. We have:

$$(w - x)S(x) = \sum_{i=1}^k \frac{(w - x)x_i^{\theta_i}}{1 - x^{q_i}} = \frac{w - x}{1 - x}.$$

Let $x \rightarrow w$. This yields

$$\sum_{i:r|q_i} \frac{-w^{\theta_i}}{-q_i w^{q_i-1}} = 0. \quad (1)$$

Now, take $r = P$. The set $\{i: P|q_i\}$ is exactly the set $\{i: P = q_i\}$. Equation (1) specified for $r = P$ can be written

$$\sum_{i:q_i=P} w^{\theta_i} = 0.$$

This implies that the set $\{i: P = q_i\}$ cannot be reduced to a single point since w is not zero. \square

To give some concrete examples, we consider the cases where N is small. First, note that in the case where the k_i are not all equal, (assume k_1 is the largest of all), we have

$$\sum_{i \neq 1} \frac{k_i}{k_1} = \frac{d}{k_1 - 1} = l_1,$$

where l_1 is the gap between two letters a_1 . This implies,

$$l_1 \leq N - 2. \quad (2)$$

PROPOSITION 2.14. There exists a constant gap sequence G with rates $(p, 1 - p)$ if and only if $p = 1/2$.

PROOF. Let a be a letter in G with gap l . Since the alphabet contains only two letters, $l = 1$. This means $p = 1/2$. \square

PROPOSITION 2.15. There exists a constant gap sequence G with rates (p_1, p_2, p_3) if and only if the following holds: $(p_1, p_2, p_3) = (1/3, 1/3, 1/3)$ or $(1/2, 1/4, 1/4)$ (up to a permutation).

PROOF. Assume that $(p_1, p_2, p_3) \neq (1/3, 1/3, 1/3)$ (otherwise, $G = (abc)^\omega$ is constant gap). Using Inequality (2), $l_1 = 1$ and $p_1 = 1/2$. Therefore, the sequence obtained from G when removing all the letters a_1 is constant gap. Applying Lemma 2.14 shows that $p_2 = p_3$. The only solution is $p_1 = 1/2$, $p_2 = 1/4$ and $p_3 = 1/4$. The associated constant gap sequence is $(a_1 a_2 a_1 a_3)^\omega$. \square

PROPOSITION 2.16. *There exists a constant gap sequence G with rates (p_1, p_2, p_3, p_4) if and only if $\{p_1, p_2, p_3, p_4\}$ belongs to the set (up to a permutation),*

$$\{(1/4, 1/4, 1/4, 1/4), (1/2, 1/4, 1/8, 1/8), (1/2, 1/6, 1/6, 1/6), (1/3, 1/3, 1/6, 1/6)\}.$$

PROOF. We give a sketch of an elementary proof of this fact. If the rates are all equal, then $(p_1, p_2, p_3, p_4) = (1/4, 1/4, 1/4, 1/4)$. Now, note that Eq. 2 implies that if the rates are not all equal $l_1 \leq 2$. We consider any letter a_i , $i \neq 1$, assume that the number of a_1 's in between two a_i 's is not constant and takes values m and n , $m > n$. Then we have $l_i \geq (n-1)l_1 + n$ on one hand and $l_i \leq n(l_1) - 3$ on the other hand. This is impossible since $l_1 \leq 2$. Therefore, the number of a_1 's in between two a_i 's is constant. This is true for all i . The sequence formed by removing all a_1 's is still constant gap. It has rates of the form $(1/3, 1/3, 1/3)$ or $(1/2, 1/4, 1/4)$. From this point, a case analysis shows that the original sequence has rates $(1/2, 1/4, 1/8, 1/8)$, $(1/2, 1/6, 1/6, 1/6)$ or $(1/3, 1/3, 1/6, 1/6)$ by inserting the letter a_1 in a constant gap sequence over the letters a_2, a_3, a_4 . \square

These few examples of constant gap sequences illustrate the fact that there are very few rates that can be achieved by constant gap sequence.

2.3. CHARACTERIZATION OF BALANCED SEQUENCES. Several studies have been recently done on balanced (or bracket) sequences.⁴ In Graham [1973] and Hubert [1996], a characterization involving constant gap sequences is given.

The proof of the following results, 2.17 and 2.18 was given by Graham [1973] for bracket sequences. An independent later proof can be found in Hubert [1996] for balanced sequences. The relation between balanced and bracket sequences given in Theorem 2.5 makes both proofs more or less equivalent.

PROPOSITION 2.17. *Let U be a balanced sequence on the alphabet $\{0, 1\}$. Construct a new sequence S by replacing in U , the subsequence of zeros by a constant gap sequence G on an alphabet \mathcal{A}_1 , and the subsequence of ones by a constant gap sequence H on a disjoint alphabet \mathcal{A}_2 . Then S is balanced on the alphabet $\mathcal{A}_1 \cup \mathcal{A}_2$.*

PROOF. We give a proof similar to Hubert's proof [Hubert 1996]. Let a be a letter in \mathcal{A}_1 (the proof is similar for a letter in \mathcal{A}_2). Let W and W' be two words of S of the same length. Then, the corresponding words X and X' in U verify $\|W\|_0 - \|W'\|_0 \leq 1$ since U is balanced. If we keep only the 0's in X and X' , then the corresponding Z and Z' words in G satisfy $\|Z\| - \|Z'\| \leq 1$. Since G is constant gap, and using Lemma 2.9, $\|Z\|_a - \|Z'\|_a \leq 1$. We end the proof noting that the construction of Z and Z' implies $|Z|_a = |W|_a$ and $|Z'|_a = |W'|_a$. \square

Conversely, we have the following theorem:

THEOREM 2.18. *Let $u \in \mathcal{A}^{\mathbb{Z}}$ be balanced and non ultimately periodic. Then there exists a partition of \mathcal{A} into two sets \mathcal{A}_1 and \mathcal{A}_2 such that the sequence v defined*

⁴ See, for example, Loeve [1995], Tijdeman [1995; 1996; 1998b], and Morikawa [1985–1993; 1982–1995].

by:

$$v_n = 1 \quad \text{if } u_n \in \mathcal{A}_1, \quad (3)$$

$$v_n = 0 \quad \text{if } u_n \in \mathcal{A}_2, \quad (4)$$

is balanced. Furthermore, the sequences z_1 and z_2 constructed from u by keeping only the letters from \mathcal{A}_1 and \mathcal{A}_2 respectively have constant gaps.

2.4. RATES IN BALANCED SEQUENCES. Let us formulate precisely the problem, which we will study in this section.

PROBLEM 2. Given a set (p_1, \dots, p_N) , is it possible to construct a balanced sequence on N letters with rates (p_1, \dots, p_N) ?

We will see in the following that this construction is not possible for all the values of the rates (p_1, \dots, p_N) . If an N -tuple (p_1, \dots, p_N) makes the construction possible, such a tuple is said to be *balanceable*. A similar problem has been addressed in Graham [1973], Morikawa [1982–1995], Tijdeman [1996], and Fraenkel [1973], where relations between the rates in balanced sequences are studied.

2.5. THE CASE $N = 2$. This case is well known and balanced sequences with two letters have been extensively studied (see, for example, Berstel and Seebold [1994] and Lothaire [1991]). The following result is known even if it is often given under different forms.

THEOREM 2.19. For all p , $0 \leq p \leq 1$, the set of rates $(p, 1 - p)$ is balanceable.

PROOF. The proof is similar to the proof of the first part of Theorem 2.5. We construct a sequence S as the support of the function $s(n) = \lfloor pn \rfloor - \lfloor p(n - 1) \rfloor$. S is a balanced sequence because the interval $]k, k + m]$ contains exactly $e = \lfloor pk + pm \rfloor - \lfloor pk \rfloor$ elements of S . Now, $\lfloor pm \rfloor + \lfloor pk \rfloor - \lfloor pk \rfloor \leq e \leq \lfloor pm \rfloor + \lfloor pk \rfloor - \lfloor pk \rfloor$. This shows the value of e can differ by at most one when k varies so S is a balanced sequence. If S' is the complementary set of S , then it should be clear that S' has asymptotic rate $1 - p$ and S' is balanced because $S'_{]k, k+m]}$ contains $m - e$ elements. \square

Note that the proof of Theorem 2.19 also gives a construction of a balanced sequence with the given rates.

2.6. THE CASE $N = 3$. The case $N = 3$ is essentially different from the case $N = 2$. In the case $N = 2$, all possible rates are balanceable while when $N = 3$, there is only one set of distinct rates which is balanceable. This result, when formulated under this form, was partly proved and conjectured in Loeve [1995], and proved in Tijdeman [1996]. In earlier papers by Morikawa [1982–1995] a similar result is proved for bracket sequences. If Theorem 2.5 is used, then the result of Morikawa can be used directly to prove the following theorem:

THEOREM 2.20. A set of rates (p_1, p_2, p_3) is balanceable if and only if,

$$(p_1, p_2, p_3) = (4/7, 2/7, 1/7)$$

or two rates are equal.

PROOF. The proof of Morikawa is very technical since it does not use the balanced property for bracket sequences. If the balanced property is used, then the proof becomes easier. We give a proof slightly simpler than the proof in Tijdeman [1996]. First, assume that $p_1 = p_2$. Then, let S be a balanced sequence with two letters $\{a, b\}$ constructed with the rates $(p_1 + p_1, p_3)$. In S , replace alternatively the “ a ”s by the letters a_1, a_2 , we get a sequence S' on alphabet $\{a_1, a_2, b\}$ with rates (p_1, p_1, p_3) . Let us show that S' is balanced. Since S is balanced, the number of “ a ”s in an interval of length m is k or $k + 1$, for some k . Now, for S' , the number of “ a_1 ”s (respectively “ a_2 ”s) in such an interval is either $(k - 1)/2$ or $(k + 1)/2$ if k is odd and $k/2$ or $k/2 + 1$ if k is even. This proves that S' is balanced.

Now, assume that (p_1, p_2, p_3) are three different numbers. We assume that $p_1 > p_2 > p_3$. We will try to construct a sequence W with these respective rates on the alphabet $\{a, b, c\}$.

Step (1). The sequence “ aca ” must appear in W .

There exists a pair of consecutive “ a ” with no “ b ” in between since $p_1 > p_2$. This means that a sequence “ aa ” or “ aca ” appears. If “ aa ” appears, then a “ c ” is necessarily surrounded by two “ a ”s.

Step (2). The sequences “ $baab$ ” and “ $abaaba$ ” must appear in W .

There exist a pair of consecutive “ b ” with no “ c ” in between. This sequence is of the form “ $ba^n b$ ”. Now, $n \leq 1$ is not possible because of the presence of “ aca ” and b -balance. $n \geq 3$ implies the existence of “ $a^{n-1}ca^{n-1}$ ” by a -balance which is incompatible with “ $ba^n b$ ” because of b -balance. Therefore, $n = 2$. Note that this also implies the existence of “ aa ” and of “ $abaaba$ ”.

Step (3). The sequence “ $abacaba$ ” appears in W .

The sequence W must contain a “ c ”. This “ c ” is necessarily surrounded by two “ a ”s since “ aa ” exists by Step (2). This group is necessarily surrounded by two “ b ”s since “ $baab$ ” exists, and consequently, necessarily surrounded by two “ a ”s because “ $abaaba$ ” exists. We get the sequence “ $abacaba$ ”.

Last Step. $W = (abacaba)^\omega$.

No letter around this word can be a “ c ” because “ $baab$ ” exists. None can be a “ b ” since “ aca ” exists. Therefore, they have to be two “ a ”s. Then note that the two surrounding letters cannot be “ c ” (because of the existence of “ $abaaba$ ”), nor “ a ” (because of the existence of “ bac ”) so they are “ b ”, then followed necessarily by “ a ” (because “ aa ” exists). At this point, we have the sequence “ $\star abaabacabaaba \star$ ”. Both \star s are necessarily “ c ”s.

To end the proof, note that we have obtained the configuration around every “ c ” and this determines the whole sequence. The sequence W is periodic of the form $(abacaba)^\omega$. \square

2.7. THE CASE $N = 4$. For distinct rates, the case $N = 4$ is very similar to the case $N = 3$. However, when two rates are equal, this case is more complicated. Again, it a similar result for bracket sequences is contained in Morikawa [1982–1995] under a weaker form since the proof is only done for rates of the form $(a_1/c, \dots, a_4/c)$. The following theorem gives the result in its

full generality. Again, using of the balanced property helps to keep the proof rather elementary.

THEOREM 2.21. *A rate tuple (p_1, p_2, p_3, p_4) with four distinct rates is balanceable if and only if $(p_1, p_2, p_3, p_4) = (8/15, 4/15, 2/15, 1/15)$.*

PROOF. We suppose that $p_1 > p_2 > p_3 > p_4$ and we show that there is only one balanced sequence with frequencies $p_1 > p_2 > p_3 > p_4$ and those frequencies are $(8/15, 4/15, 2/15, 1/15)$.

As a preliminary remark, note that if $p_i > p_j$, then there exists at least one word " $a_i \cdots a_i$ " that does not contain any a_j . This fact will be used several times in the following arguments.

The proof involves different steps.

Step (1). W contains the words " aca " or " ada " or " $acda$ " or " $adca$ ".

There exist two consecutive " a "s with no " b " in between because $p_1 > p_2$. Therefore, either " aa " or " aca " or " ada " or " $acda$ " or " $adca$ " exist. If " aa " exists, then, a " c " is surrounded by two " a "s.

Step (2). W contains the word " $baab$ ".

First, we show that if a word " $ba^n b$ " exists, then $n = 2$. Indeed, Step (1) makes " bb " and " bab " impossible. On the other hand, if $n \geq 3$, the existence of " $a^{n-1}ca^{n-1}$ " is necessary by a -balance and is incompatible with the existence of " $ba^n b$ " because of b -balance.

Now, if no word of the form " $ba^n b$ " exists, then there exist two consecutive " b "s with one " d " and no " c " in between. Such a word is of the form: " $ba^i da^k b$ " and is surrounded by some " a "s, so that we get a word called s_1 and equal to " $a^i ba^i da^k ba^l$ ". Note that the numbers i, j, k, l may be equal to zero but $j + k \geq 1$ by b -balance.

There also exist two consecutive " c "s with no " d " in between. In between those two " c " we must have some " a "s and some " b "s. In fact we have exactly one " b " since no word " $ba^n b$ " exists by assumption. We pick such a word called s_2 which is of the form: " $ca^m ba^n c$ ". Note that i, j, k, l, m, n are integers that can differ by at most one. As for the length of s_1 , we have $|s_1| \geq \max(4, 2(n + m) + 1) \geq n + m + 3 = |s_2|$. This is impossible by c -balance.

Step (3). The word " $abacabaabacaba$ " exists in W .

There exist two consecutive " c "s with no " d " in between. From Step (3) in the proof of Lemma 2.20, we know that a " c " is necessarily surrounded by the word " aba ". Moreover, from Step (4) in the proof of Lemma 2.20, we have: " $abacabaabaUcaba$ ", where U is a word that contains no " d " and no " c ". U cannot start with an " a " (because of " $bacab$ ") and cannot start with a " b " (because of " aca "). Therefore, U has to be empty.

Step (4). W is uniquely defined and is periodic of period " $abacabadabacaba$ ".

Somewhere, W contains a " d ". From this point on, we can extend the word uniquely as: " $abacabadabacaba$ " around this " d ", and the word " $abacabaaba-caba$ " has to be surrounded by two " d "s. This ends the proof. \square

To complete the picture, it is not difficult to see that,

PROPOSITION 2.22. *If the tuple (p_1, p_2, p_3, p_4) is made of no more than two distinct numbers, then it is balanceable.*

PROOF. First, if the rates are all equal, they are obviously balanceable. If three of them are equal, say $p_1 = p_2 = p_3$, then, we can construct a balanced sequence with rates $(3p_1, p_4)$ and we construct a balanced sequence with rates (p_1, p_1, p_1, p_4) by using Proposition 2.17 (where we take G the constant gap sequence $(a_1 a_2 a_3)^\omega$ and $H \stackrel{\text{def}}{=} (a_4)^\omega$). If two pairs of rates are equal, say $p_1 = p_2$ and $p_3 = p_4$, then we construct a balanced sequence with rates $(2p_1, 2p_3)$ and we apply Proposition 2.17. \square

If the tuple (p_1, p_2, p_3, p_4) is made of exactly three distinct numbers, then this is a more complex case which is not studied here.

2.8. THE GENERAL CASE. In this section, we are interested in the case of arbitrary N . First, note that Proposition 2.22 easily generalizes to any dimension.

PROPOSITION 2.23. *If the tuple $(p_1, p_2, p_3, \dots, p_N)$ is made of less than two distinct numbers, then it is balanceable.*

PROOF. The proof is similar to the proof of Proposition 2.22. \square

PROPOSITION 2.24. *If (p_1, p_2, \dots, p_N) is balanceable, then*

$$\underbrace{(p_1/k, \dots, p_1/k)}_k, p_2, \dots, p_N$$

is balanceable.

PROOF. The proof is very similar to that of Proposition 2.17. If W is a balanced sequence with letters $\{a_1, \dots, a_N\}$, consider the sequence W' constructed starting from W and replacing each “ a_1 ” by an element of (b_1, \dots, b_k) in a cyclic way. Note that W' has the following set of rates, $(p_1/k, \dots, p_1/k, p_2, \dots, p_N)$.

Next, we show that W' is balanced. Since W is balanced, for an arbitrary integer m , the number of “ a_1 ”s in an interval of length m is n or $n + 1$, for some n . Now, for W' , the number of “ b_i ”s in such an interval is either $\lfloor (n - 1)/k \rfloor$ or $\lfloor (n + 1)/k \rfloor$. This proves that W' is balanced. \square

For the general case and distinct rates, it is natural to give the following conjecture (due to Fraenkel for bracket sequences):

CONJECTURE 2.25. *A set of distinct rates $\{p_1, \dots, p_N\}$ with $N > 2$ is balanceable if and only if*

$$\{p_1, \dots, p_N\} = \{2^{N-1}/(2^N - 1), \dots, 2^{N-i}/(2^N - 1), \dots, 1/(2^N - 1)\}.$$

We have not been able to prove this fact. Morikawa has also given some insight in this problem. Very recently, in Tijdeman [1998a], the cases $N = 5$ and $N = 6$ are proven using techniques inspired by those introduced here. However, it seems clear that a different approach is needed in order to complete the proof in the general case. Here, we only have partial results given in the following propositions.

PROPOSITION 2.26. *The rates $(2^{N-1}/(2^N - 1), \dots, 2^{N-i}/(2^N - 1), \dots, 1/(2^N - 1))$ are balanceable, for all $N \in \mathbb{N}$.*

This proposition is the “if” direction of the conjecture.

PROOF. We construct a balanced sequence W_N in the following inductive way. $V_1 = a_1$, $V_N = V_{N-1}a_NV_{N-1}$ and $W_N = (V_N)^\omega$. First note that W_N has rates $(2^{N-1}/(2^N - 1), \dots, 2^{N-i}/(2^N - 1), \dots, 1/(2^N - 1))$. Then, we show that W_N is balanced by induction. In the sequence W_N , any letter (say letter j) appears 2^{N-j} times in one period and is of the form of $2^{N-j} - 1$ intervals of the same length (2^j) and one of length $2^j - 1$.

By construction of W_{N+1} , these properties still hold and therefore, W_{N+1} is balanced. \square

PROPOSITION 2.27. *Let $N > 2$ and W be balanced with rates $p_1 > \dots > p_N$, then, W is ultimately periodic. In particular, this means that $p_i \in \mathbb{Q}$, $\forall 1 \leq i \leq N$.*

PROOF. If W is not ultimately periodic, Theorem 2.18 says that W is composed of two constant gap sequences. At least one of these sequences has at least two letters, and therefore two letters have rates which are equal by Lemma 2.13. Therefore, the rates in W of these two letters are also equal. \square

PROPOSITION 2.28. *Let W be balanced with rates $p_1 > \dots > p_N$, with the following property: for any $1 \leq i \leq N$, there exists two consecutive letters “ a_i ” with no a_j in between, with $j > i$. Then, $(p_1, \dots, p_N) = (2^{N-1}/(2^N - 1), \dots, 2^{N-i}/(2^N - 1), \dots, 1/(2^N - 1))$ and W is uniquely defined, up to a shift.*

This proposition is a partial “only if” result for the conjecture.

PROOF. The proof holds by induction. Let V_k denote the period of the balanced sequence with rates $(2^{k-1}/(2^k - 1), \dots, 2^{k-i}/(2^k - 1), \dots, 1/(2^k - 1))$ given in Proposition 2.26. We recall that according to the construction in the proof of Proposition 2.26, $V_k = V_{k-1}a_kV_{k-1}$. We will prove by induction that W is periodic with period V_N .

We prove by induction on k that W contains the word $V_{k-1}a_kV_{k-1}V_{k-1}a_kV_{k-1}$ and that for $j \geq k$, each letter in W , “ a_j ”, is surrounded by V_{k-1} ’s, for all possible $1 < k \leq N$. For the first step of the induction ($k = 1$), note that according to the property on W , W contains the word “ a_1a_1 ” which is the same as “ V_1V_1 ”. Therefore, any other letter is surrounded by two “ a_1 ”s. This also implies the existence of “ $a_1a_2a_1a_1a_2a_1$ ” by using a similar argument as Step (2) in the proof of Theorem 2.21, which ends the case $k = 1$.

For the general case, by the induction assumption, a_k is surrounded by V_{k-2} , and W contains the word “ $V_{k-2}a_kV_{k-2}$ ”. The existence of “ $V_{k-2}a_{k-1}V_{k-2}V_{k-2}a_{k-1}V_{k-2}$ ” proves that this word is surrounded by two “ a_{k-1} ”s. Therefore, two consecutive a_k form the word

$$V_{k-2}a_{k-1}V_{k-2}a_kV_{k-2}a_{k-1}V_{k-2}UV_{k-2}a_{k-1}V_{k-2}a_kV_{k-2}a_{k-1}V_{k-2},$$

where U does not contain any letter “ a_j ”, $j \geq k$.

—If U does not contain “ a_{k-1} ” and contains a letter “ a_i ” with $i < k - 1$, then U is reduced to this letter, because of a_{k-1} -balance and the existence of the

word $a_{k-1}V_{k-2}V_{k-2}a_{k-1}$ (induction assumption). But now the construction of V_{k-2} implies the presence of " $a_iV_{i-1}a_i$ " and contradicts the existence of " $a_iV_{i-1}V_{i-1}a_i$ ".

—If U contains " a_{k-1} ", then by a_{k-1} -balance, U must contain $V_{k-2}a_{k-1}V_{k-2}$. Therefore, U is of the form $XV_{k-2}a_{k-1}V_{k-2}Y$. The arguments used for U can be applied to X and Y . If they are not empty, they must both contain $V_{k-2}a_{k-1}V_{k-2}$. Eventually, we have the existence of the word $V_{k-2}V_{k-2}a_{k-1}V_{k-2}V_{k-2}$ which contradicts the existence of $V_{k-2}a_{k-1}V_{k-2}a_kV_{k-2}a_{k-1}V_{k-2}$ by a_1 -balance.

Therefore, U is empty and we obtain the existence of $V_{k-1}a_kV_{k-1}V_{k-1}a_kV_{k-1}$.

Now, we finish the proof by noticing that the letter a_N is surrounded by V_{N-1} and by noting that $V_{N-1}V_{N-1}$ is necessarily surrounded by " a_N ". \square

PROPOSITION 2.29. *The projection W' of a sequence W over the alphabet $\mathcal{A} - \{a\}$ is W where all a 's have been removed. If $p_a \geq 0.5$, then W is balanced implies that W' is balanced.*

PROOF. Choose two words V'_1, V'_2 of length n in W' . Let V_1 and V_2 be any two words in W whose projections over the alphabet $\mathcal{A} - \{a\}$ are V'_1 and V'_2 , respectively. Assume, furthermore, that the first and last letters in V_1 and V_2 are not a . Let $k = |V_1| - n$ and $l = |V_2| - n$ denote the number of appearances of the letter a in V_1 and V_2 , respectively.

Step (1). If $l = k$, then the difference in the number of occurrences of any letter $b \neq a$ in V'_1 and in V'_2 is at most 1, since W is balanced, and since the number of b 's in V_1 (respectively, V_2) is the same as its number in V'_1 (respectively, V'_2).

Step (2). Assume that $l > k + 1$.

—Let \hat{V}_2 be the word obtained from V_2 by truncating the first and last letter. Then $|\hat{V}_2| = n + l - 2$, and the number of a 's in \hat{V}_2 is l .

—Let \hat{V}_1 be the word obtained from V_1 by concatenating to it the next $m = l - k - 2$ letters that appear after V_1 in the sequence W . Then $|\hat{V}_1| = n + l - 2 = |\hat{V}_2|$, and the number of a 's in \hat{V}_1 is not larger than $k + m = l - 2$. This is a contradiction with the fact that W is balanced.

Step (3). It remains to check the case $l = k + 1$. Add to V_1 the next letter that occurs in W to its right, to form the new word \bar{V}_1 . If it is not a , then we have two successive letters that are not a , which contradicts the fact that a has an asymptotic frequency of at least $1/2$. If it is a , then \bar{V}_1 and V_2 have the same number of a 's. We can now apply the same argument as in Step (1) and conclude that the number of occurrences of any letter b in V'_1 and in V'_2 differs by at most 1.

Combining the above steps, we conclude that W' is balanced. \square

2.9. EXTENSIONS OF THE ORIGINAL PROBLEM. So far we have only analyzed the case where all the rates add up to one. The different results tend to prove that very few sets of rates are balanced.

Now let us look at a generalization when all the rates do not add up to one. Assume that S is a sequence on the alphabet $\{a_1, a_2, \dots, a_k, *\}$. We only require that S is balanced for the letters a_1, \dots, a_k , but not for the special letter $*$.

On a more practical point of view, the question can be viewed as whether this allows more possibilities for rates to be balanced when “losses” are allowed (represented by the letter $*$). Then again, in general, the rates are not balanced, even if the total sum is very small as illustrated by the following proposition:

PROPOSITION 2.30. *For an arbitrary $\epsilon > 0$, there exists two real numbers p_1 and p_2 such that $p_1 + p_2 < \epsilon$ and such that there is no sequence S on the alphabet $\{a, b, *\}$ with asymptotic rate p_1 for letter a and p_2 for letter b which is balanced for a and b .*

PROOF. Choose two irrational numbers p_1 and p_2 with $p_1 + p_2 < \epsilon$ such that p_1 , p_2 and 1 are not linearly dependent on \mathbb{Z} . Now assume that there exists a sequence S on $\{a, b, *\}$ with asymptotic rate p_1 for letter a and p_2 for letter b which is balanced for a and b . By Theorem 2.5, then there exists two real numbers x, y such that $\delta_a(S)(n) = 1$ if $x + p_1 n \bmod 1 \in [1 - p_1, 1]$ and 0, otherwise. $\delta_b(S)(n) = 1$ if $y + p_2 n \bmod 1 \in [1 - p_2, 1]$ and 0 otherwise. In the cube $[0, 1]^2$, the set of points $(x, y) + n(p_1, p_2) \bmod (1, 1)$ is dense (see for example, Weyl’s ergodic theorem Sinai [1976]) and therefore hits the rectangle $[(1 - p_1, 1 - p_2), (1, 1)]$. This is not possible. \square

More on this kind of problems can be found in Tijdeman [1995].

To end this short overview on balanced sequences, we must mention on the positive side that “usual” rates, such as $(1/k, 1/k, \dots, 1/k)$ are often balanceable. In Appendix 4.3, some examples of balanced sequences and their rates are given.

3. Routing of Customers in Multiple Queues

The notion of balanced sequences gives birth to a large set of elegant properties in word combinatorics. However, they are rarely used in operations research. The balanced sequences in dimension $N = 2$ have been used for discrete line drawing [Lothaire 1991] and also for scheduling optimization in Hajek [1985] and Gaujal [1997].

However, the case $N = 2$ behaves differently from the cases of higher dimension (as illustrated by Theorem 2.19) and does not really grasp all the complexity of the model.

Here, we present an application of balanced sequences in arbitrary dimensions to scheduling optimization.

We consider a system where a sequence of tasks have to be executed by several processing units. The tasks arrive sequentially and each task can be processed by any server. The *routing* control consists in assigning to each task a server on which it will be processed. The routing is optimal if it minimizes some cost function that measures the performance of the system.

These kinds of models have been used to study load balancing within several processors in parallel processing problems as well as for efficient network utilization in telecommunication systems, as presented in the introduction.

3.1. PRESENTATION OF THE MODEL. In this section, we consider a more precise queuing model of the system that we described. Customers enter a multiple queue system composed of K subsystems. Each subsystem is made of several queues, initially empty, which form an *event graph*. Event graphs are a subset of Petri nets with no more than one input transition and one output

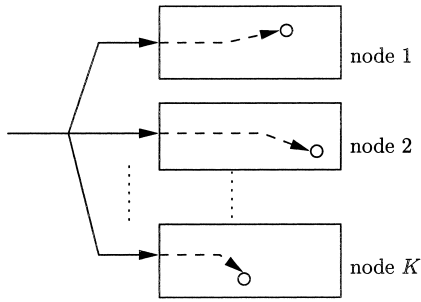


FIG. 1. Illustration of the routing of customers in a K node system.

transition per place. More details on the dynamics of event graphs can be found in Baccelli et al. [1992]. In particular, their dynamical behavior is linear in the so-called $(\max, +)$ algebra [Baccelli et al. 1992]. Many queuing networks can be represented as event graphs, as long as there is no inside choice for the route followed by the customers (see Altman [1997a] for a more complete discussion on this issue).

The routing of customers to the different subsystems is controlled by a sequence of vectors $\{a_n\}$, with a_n is in $\{0, 1\}^K$ and $a_n^i = 1$ means that the n th customer is routed to subsystem i . Note that a is a *feasible* admission sequence as long as for all n , $\sum_i a_n^i = 1$.

The link between a feasible routing policy and an infinite sequence on a finite alphabet as used in the first part, comes from choosing the alphabet \mathcal{A} composed by the letters

$$\{(1, 0, \dots, 0), (0, 1, 0, \dots), \dots, (0, \dots, 0, 1)\}.$$

Using this alphabet on K letters, a feasible routing policy can be viewed as an infinite sequence on \mathcal{A} .

Figure 1 shows an illustration of the system we are considering.

We denote by T_n the epoch when the n th customer enters the system. We assume that $T_1 = 0$. The interarrival time sequence is $\{\delta_k\} \stackrel{\text{def}}{=} \{T_{i+1} - T_i\}$. Finally, $\sigma_n^{i,j}$ will denote the service time of the n th customer entering the j th queue in node i .

The sequences $\{\delta_k\}$ and $\{\sigma_k^{i,j}\}$ will be considered as random processes. We also make stochastic assumption on these sequences. The interarrival time of the customers and the service times form stationary processes, and we assume that the interarrival times are independent of the service times.

3.2. OPTIMAL ROUTING SEQUENCE. In each subsystem i , we pick an arbitrary server s_i (which may be the last server in the subsystem for example). The immediate performance criterion for subsystem i will be the traveling time to server s_i of a virtual customer that would enter subsystem i at time T_n . Under a given routing policy, this quantity only depends on the values of the n first routing choices. From the routing sequence a , we can isolate the routing decision for node i : if $a_n^i = 1$, then the customer is admitted in node i and, if $a_n^i = 0$, then the customer is rejected (for node i). We denote the traveling time at time T_n by $W_n^i(a_1^i, \dots, a_n^i)$. We will be more particularly interested in the expected value of the traveling time with respect to the service times in all the servers

contained in node i and with respect to the interarrival times: $\mathbf{W}_n^i(a_1^i, \dots, a_n^i) \stackrel{\text{def}}{=} \mathbf{E}_{\sigma, f}(W_n^i(a_1^i, \dots, a_n^i))$, where $f: \mathbb{R} \rightarrow \mathbb{R}$ is any convex increasing function.

If we focus on a single node i , the function $\mathbf{W}_n^i(a_1^i, \dots, a_n^i)$ has been studied in Altman et al. [1997a]. Its most remarkable property is the fact that this function is *multimodular*.

Here, we merely give the definition of multimodular functions.

Let (e_1, \dots, e_n) be the canonical base of \mathbb{R}^n . Define $d_i = e_i - e_{i+1}$, $i = 1, \dots, m-1$. The set $\{d_0 = -e_1, d_1, \dots, d_{m-1}, d_m = e_m\}$ is called the *base of multimodularity*, and f is *multimodular* if $f(a + d_i) + f(a + d_j) \geq f(a) + f(a + d_i + d_j)$ for all $i \neq j$.

The reader is referred to Hajek [1985] and Altman et al. [1997b] for a precise exposition of the properties of multimodular functions. One of the main properties is that for any multimodular function $f: \mathbb{Z}^n \rightarrow \mathbb{R}$, one can construct a convex function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, which coincides with f on \mathbb{Z}^n .

THEOREM 3.1. [ALTMAN ET AL. 1997A]. *Under the foregoing assumptions, the function $\mathbf{W}_n^i(a_1^i, \dots, a_n^i)$ satisfies the following properties.*

- (1) $\mathbf{W}_n^i(a_1^i, \dots, a_n^i)$ is multimodular,
- (2) $\mathbf{W}_n^i(a_1^i, \dots, a_n^i)$ is increasing in a_k^i , $1 \leq k \leq n$,
- (3) $\mathbf{W}_m^i(0, \dots, 0, a_1^i, \dots, a_n^i) = \mathbf{W}_n^i(a_1^i, \dots, a_n^i)$, if $m > n$.

PROOF. These properties are shown in full detail in Altman et al. [1997a]. \square

Using these properties, we will derive as in Altman et al. [1997b], a lower bound denoted $B_i(\alpha, p)$ for any routing a^i , for the following discounted cost. This function $B_i(\alpha, p)$ is increasing in α and in p and lower-continuous.

Let us fix some arbitrary integer, N . We define $p_\alpha^i = (1 - \alpha) \sum_{k=1}^{\infty} \alpha^{k-1} a_k^i$. Now, using assumptions (1), (2), (3) of Theorem 3.1, we have,

$$\begin{aligned}
 & \sum_{n=1}^{\infty} (1 - \alpha) \alpha^{n-1} \mathbf{W}_n^i(a_1^i, a_2^i \dots a_n^i) \\
 & \geq \sum_{n=1}^N (1 - \alpha) \alpha^{n-1} \mathbf{W}_N^i(0 \dots 0, a_1^i, a_2^i \dots a_n^i) \\
 & \quad + \sum_{n=N+1}^{\infty} (1 - \alpha) \alpha^{n-1} \mathbf{W}_N^i(a_{n-N+1}^i \dots a_n^i) \\
 & = \sum_{n=1}^N (1 - \alpha) \alpha^{n-1} \tilde{\mathbf{W}}_N^i(0 \dots 0, a_1^i, \dots a_n^i) \\
 & \quad + \sum_{n=N+1}^{\infty} (1 - \alpha) \alpha^{n-1} \tilde{\mathbf{W}}_N^i(a_{n-N+1}^i \dots a_n^i) \\
 & \geq \tilde{\mathbf{W}}_N^i \left(\sum_{n=1}^N (1 - \alpha) \alpha^{n-1} (0 \dots 0, a_1^i \dots a_n^i) \right.
 \end{aligned}$$

$$+ \sum_{n=N+1}^{\infty} (1 - \alpha) \alpha^{n-1} (a_{n-N+1}^i \cdots a_n^i) \quad (5)$$

$$= \tilde{\mathbf{W}}_N^i(\alpha^N p_\alpha^i, \alpha^{N-1} p_\alpha^i, \cdots, p_\alpha^i), \quad (6)$$

where (5) follows from Jensen's inequality, since the function $\tilde{\mathbf{W}}_N^i$ is convex, and since the coefficients $(1 - \alpha) \alpha^{n-1}$ are nonnegative and sum to 1. Define

$$B_i(\alpha, p) = \sup_N \tilde{\mathbf{W}}_N^i(\alpha^N p, \alpha^{N-1} p, \cdots, p). \quad (7)$$

Note that B_i is defined for a fixed sequence $\{a_k^i\}$. Also note that $B_i(\alpha, p)$ is lower semicontinuous in α and in p .

The previous analysis shows that

$$\sum_{n=1}^{\infty} (1 - \alpha) \alpha^{n-1} \mathbf{W}_n^i(a_1^i, a_2^i, \cdots, a_n^i) \geq B_i(\alpha, p_\alpha^i). \quad (8)$$

Also, for a given p , we define the regular sequence with rate p and arbitrary phase θ ,

$$a^p(\theta) \stackrel{\text{def}}{=} \lfloor np + \theta \rfloor - \lfloor (n-1)p + \theta \rfloor,$$

(see Definition 2.4). One can show as in Altman et al. [1997b] that $a^p(\theta)$ satisfies

$$\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{n=1}^m \mathbf{W}_n^i(a_1^p(\theta), \cdots, a_n^p(\theta)) = B_i(1, p). \quad (9)$$

Here, however, we are interested in the performance of all nodes together. Therefore, we choose as a cost function, the undiscounted average on n of some linear combination of the expected traveling time in all nodes.

Let h be any increasing linear function, $h: \mathbb{R}^K \rightarrow \mathbb{R}$. We consider the undiscounted average cost of a feasible routing sequence a ,

$$g(a) \stackrel{\text{def}}{=} \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N h(\mathbf{W}_n^1, \cdots, \mathbf{W}_n^K). \quad (10)$$

Our objective is to minimize $g(a)$.

THEOREM 3.2. *The following lower bound holds for all policies:*

$$g(a) \geq \inf_{p_1 + \cdots + p_K = 1} h(B_1(p_1), \cdots, B_K(p_K)).$$

PROOF. We adapt the method developed in Altman et al. [1997b] for our case. We introduce the following notation.

$$B_i(p_i) \stackrel{\text{def}}{=} \sup_{\alpha \leq 1} B_i(\alpha, p_i).$$

Due to Littlewood's and Jensen's inequalities as well as Eq. (8), we have

$$\begin{aligned}
 & \overline{\lim}_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N h(\mathbf{W}_n^1, \dots, \mathbf{W}_n^K) \\
 & \geq \overline{\lim}_{\alpha \rightarrow 1} (1 - \alpha) \sum_{n=1}^{\infty} \alpha^{n-1} h(\mathbf{W}_n^1, \dots, \mathbf{W}_n^K) \\
 & \geq \overline{\lim}_{\alpha \rightarrow 1} h \left((1 - \alpha) \sum_{n=1}^{\infty} \alpha^{n-1} \mathbf{W}_n^1, \dots, (1 - \alpha) \sum_{n=1}^{\infty} \alpha^{n-1} \mathbf{W}_n^K \right) \\
 & \geq \overline{\lim}_{\alpha \rightarrow 1} h(B_1(\alpha, p_1^1), \dots, B_K(\alpha, p_K^K))
 \end{aligned} \tag{11}$$

By definition, we note that $\sum_{i=1}^K p_{\alpha}^i = 1$. Hence, one may choose a sequence $\alpha_n \uparrow 1$ such that the following limits exist:

$$\lim_{n \rightarrow \infty} p_{\alpha_n}^i = p_i, \quad i = 1, \dots, K \tag{12}$$

and $\sum_{i=1}^K p_i = 1$. From the continuity of $B_i(\alpha, p_i)$ in p and α we get from (11)

$$\begin{aligned}
 g(a) & \geq h(B_1(p_1), \dots, B_K(p_K)) \\
 & \geq \inf_{p_1 + \dots + p_K = 1} h(B_1(p_1), \dots, B_K(p_K)).
 \end{aligned} \tag{13}$$

□

Note that there exists some p^* that achieves the infimum

$$\inf_{p_1 + \dots + p_K = 1} h(B_1(p_1), \dots, B_K(p_K)),$$

since $h(B_1(p_1), \dots, B_K(p_K))$ is continuous in $p = (p_1, \dots, p_K)$.

Consider $\theta = (\theta_1, \dots, \theta_K)$ and the routing policy $a^{p^*}(\theta)$ given for each i by

$$a_k^{i,p^*}(\theta_i) = \lfloor kp_i^* + \theta_i \rfloor - \lfloor (k-1)p_i^* + \theta_i \rfloor. \tag{14}$$

There are some p^* 's for which the condition of feasibility of the policy $a^{p^*}(\theta)$ is satisfied, that is, there exists some $\theta = (\theta_1, \dots, \theta_K)$, such that the regular policy $a^{p^*}(\theta)$ is feasible.

Using the correspondence between a routing policy and a sequence on the alphabet \mathcal{A} , these p^* 's correspond precisely to *balanceable* rates.

THEOREM 3.3. *Assume that p^* is balanceable. Then $a^{p^*}(\theta)$ is optimal for the average cost, that is, it minimizes $g(a)$ over all feasible policies.*

PROOF. The proof follows directly from Theorem 3.2 and Eq. (9). □

REMARK 3.4. *The previous theorem says that regular sequences are optimal routing policies. As for balanced sequences (which are ultimately regular, see Theorem 2.5), they are also optimal if the buffer in each node empties infinitely often. This situation occurs when the system is stable as shown in Baccelli et al. [1996]. In such cases, a finite prefix of the routing sequence does not alter its cost.*

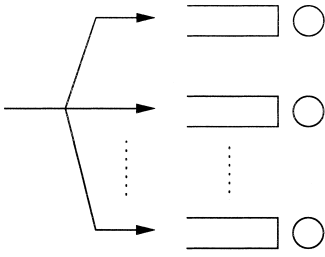


FIG. 2. Routing in homogeneous queues.

4. Study of Some Special Cases

The problem that remains to be addressed is to find in which cases, the rate vector p^* is balanceable. We will present several simple examples for which we can make sure that the optimal rate p^* is balanceable.

4.1. THE CASE $K = 2$. If $K = 2$, then, the optimal rate vector is of the form $p^* = (p_1^*, 1 - p_1^*)$. Theorem 2.19 says that p^* is always balanceable and therefore, the optimal routing sequence is given by an associated balanced sequence. Note that this approach does not give any direct way to compute the value of p^* , however, it gives the structure of the optimal policy.

4.2. THE HOMOGENEOUS CASE. Now let K be arbitrary and each node is made of a single server, all servers being identical. This model is displayed in Figure 2.

Also assume that the function h is symmetrical in all coordinates (e.g., just the sum of all waiting times). By symmetry and convexity in (p_1, \dots, p_K) of the function $h(V_1^1(p_1), \dots, V_K^1(p_K))$, then, $p^* = (1/K, \dots, 1/K)$, which is balanceable. The associated balanced sequence is the round robin routing scheme. Applying Theorem 3.3 yields the following result, which is new (to the best of the author's knowledge).

THEOREM 4.1. *The round robin routing to K identical $/G/1$ queues, minimizes the total average expected workload of all the queues over all admission sequences with no information on the state of the system.*

In Liu and Righter [1998], the round robin routing is proved to be optimal in separable-convex increasing order for K identical $/GI/1$ queues. Their method uses an intricate coupling argument, whereas our proof is a simple corollary of the general theory on multimodular functions.

To illustrate the advantage of our approach, we further generalize the result to a system composed of K identical $(\max, +)$ linear systems with a single entry. In this case, the symmetry argument used in the case of simple queues still holds. Then again, the round robin routing policy minimizes the traveling time in each system. This case includes models such as routing among several identical systems composed of queues in tandem, for example (see Figure 3).

4.3. TWO SETS OF IDENTICAL SERVERS. As a consequence of the two previous cases, we can consider a system composed of K_1 identical queues of type 1 and K_2 queues of type 2. Again, assume that h is symmetrical in the K_1 nodes of type 1 and symmetrical in the K_2 nodes of type 2. Then, by symmetry arguments, the

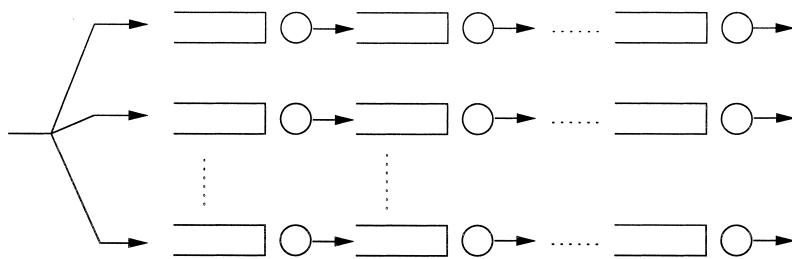


FIG. 3. Routing in queues in tandem.

optimal rate vector p^* is of the form

$$\left(\frac{p}{K_1}, \dots, \frac{p}{K_1}, \frac{1-p}{K_2}, \dots, \frac{1-p}{K_2} \right).$$

This rate vector is balanceable indeed. This implies that for the weighted total average expected workload, the optimal routing is of balanced type, if nodes of the same type have the same weight.

Many other examples of this kind can be derived from these examples through similar constructions.

Appendix

In this appendix, we shall give a collection of balanceable set of rates that can be put into two classes. Some of them are *composite*: they can be constructed using Proposition 2.24 (once or several times) starting from a smaller balanceable set. The ones which cannot be constructed that way are called *primitive*.

We shall first give a list that contains some primitive balanceable rates with $N = 4$ (as well as other cases with different values of N).

- $(1/11, 2/11, 4/11, 4/11)$ is balanceable and $S = (abcababcabd)^\omega$.
- $(1/11, 2/11, 2/11, 6/11)$ is balanceable and $S = (abacaabacad)^\omega$.
- $(1/11, 1/11, 3/11, 6/11)$ is balanceable and $S = (acabaabadab)^\omega$.
- for all N , $((2^{N-1}/(2^N - 1), \dots, 2^{N-i}/(2^N - 1), \dots, 1/(2^N - 1)))$ is balanceable. The associated balanced sequence is constructed recursively as in Proposition 2.26.

Here are other balanceable sets of rates when $N = 4$, which are composite.

- $(1/14, 1/14, 4/14, 8/14)$ is balanceable and $S = (abacabaabadaba)^\omega$. It is composite since it comes from $(1/7, 2/7, 4/7)$ where the smallest rate is split into two.
- For each real number $0 < p \leq 1$, the rates $(1 - p, p/4, p/4, p/2)$ are balanceable, with a corresponding balanced sequence constructed from a regular sequence with rate p where all 1 are replaced in turn by the sequence $(abac)^\omega$ and each 0 by the letter d . It is composite, originating from $(1 - p, p)$ and split twice.
- $(1/k, \dots, 1/k)$ is balanceable. A balanced sequence is: $S = (a_1 a_2 a_3 \dots a_k)^\omega$. This is composite, coming from (1) split once into k rates.

— $(p, \dots, p, \beta, \dots, \beta)$ is balanceable. A balanced sequence with those rates is constructed in the following way: Choose a balanced sequence S on letters, (A, B) with rate $(p_1 = \sum_{i=1}^k p, p_2 = \sum_{i=1}^h \beta)$. In S replace all the A (respectively B) by a_1, a_2, \dots, a_k (respectively, b_1, \dots, b_h) in a round-robin fashion to get a balanced sequence with the required rates. This is also composite, coming from (p_1, p_2) , where p_1 is split into k rates and p_2 is split into h rates.

ACKNOWLEDGMENT. The authors would like to thank Alain Jean-Marie, Alex Heinis and Rob Tijdeman. Jean-Marie pointed out the Wilf [1994] reference and gave them a first proof of Lemma 2.13. Heinis gave us comments on an earlier version of the paper and Tijdeman told us the relation with the results of Morse and Hedlund, and mentioned the papers of Morikawa to us. Finally, anonymous referees helped us put our results in perspective and improve their presentation.

REFERENCES

- ALTMAN, E., GAUJAL, B., AND HORDIJK, A. 1997a. Admission control in stochastic event graphs. Tech. Rep. 3179. RUL-TW-97-06. INRIA, Sophia Antipolis Cedex, France, and Leiden University, Leiden, The Netherlands.
- ALTMAN, E., GAUJAL, B., AND HORDIJK, A. 1997b. Multimodularity, convexity and optimization properties. Tech. Rep. 3181. RUL-TW-97-07. INRIA, Sophia Antipolis Cedex, France, and Leiden University, Leiden, The Netherlands.
- ALTMAN, E., AND STIDHAM, S. 1995. Optimality of monotonic policies for two-action Markovian decision processes with applications to control of queues with delayed information. *Queueing Syst.: Theory Appl.* 21, 267–291.
- ARNOUX, P., MAUDUIT, C., SHIOKAWA, I., AND TAMURA, J. 1994. Complexity of sequences defined by billiards in the cube. *BSMF (Bull. Soc. Math. France)* 122, 1–12.
- BACCELLI, F., FOSS, S., AND GAUJAL, B. 1996. Free choice nets, an algebraic approach. *IEEE Trans. Autom. Cont.* 41, 12, 1751–1778.
- BACCELLI, F., GOHEN, G., OLSDER, G. J., AND QUADRAT, J.-P. 1992. *Synchronization and Linearity*. Wiley, New York.
- BAR-NOY, A., BHATIA, R., NAOR, J., AND SCHIEBER, B. 1998. Minimizing service and operation costs of periodic scheduling. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. ACM, New York.
- BARYSHNIKOV, Y. 1995. Complexity of trajectories in rectangular billiards. *Commun. Math. Phys.* 175, 43–56.
- BERSTEL, J., AND SÉÉBOLD, P. 1994. Morphismes de Sturm. *Bull. Belg. Math. Soc.* 1, 175–189.
- BOEL, R. K., AND VAN SCHUPPEN, J. H. 1989. Distributed routing for load balancing. *Proc. IEEE*, 210–221.
- CHANG, C. S., CHAO, X., AND PINEDO, M. 1990. A note on queues with Bernoulli routing. In *Proceedings of the 29th Conference on Decision and Control*.
- COMBÉ, M. B., AND BOXMA, O. 1994. Optimization of static traffic allocation policies. *Theoret. Comput. Sci.* 125, 17–43.
- FRAENKEL, A. 1973. Complementary and exactly covering sequences. *J. Combinat. Theory* 14, 8–20.
- GAUJAL, B. 1997. Optimal allocation sequences of two processes sharing a resource. *J. Disc. Event Dyn. Syst.* 7, 327–354.
- GRAHAM, R. 1973. Covering the positive integers by disjoint sets of the form $\{[n\alpha + \beta] : n = 1, 2, \dots\}$. *J. Combin. Theory* 15, 354–358.
- HAJEK, B. 1985. Extremal splittings of point processes. *Math. Oper. Res.* 10, 4, 543–556.
- HARIHARAN, R., HULHARNI, V. G., AND STIDHAM, S. 1990. A survey of research relevant to virtual circuit routing in telecommunication networks. Tech. Rep. WC/OR/TR 90-13. Univ. N. Carolina at Chapel Hill, Chapel Hill, N.C.
- HORDIJK, A., AND KOOLE, G. 1992. On the assignment of customers to parallel queues. *Probab. Eng. Inf. Sci.* 6, 495–511.

- HORDIJK, A., KOOLE, G., AND LOEVE, A. 1994. Analysis of a customer assignment model with no state information. *Probab. Eng. Inf. Sci.* 8, 419–429.
- HUBERT, P. 1996. Propriétés combinatoires des suites définies par le billard dans les triangles pavants. *Theoret. Comput. Sci.* 164, 1–2, 165–183.
- KOOLE, G. 1996. On the pathwise optimal Bernoulli routing policy for homogeneous parallel servers. *Math. Oper. Res.* 21, 469–476.
- LIU, Z., AND RIGHTER, R. 1998. Optimal load balancing on distributed homogeneous unreliable processors. *Journal of Operations Research* 46, 563–573.
- LOEVE, J. A. 1995. Markov decision chains with partial information. Ph.D. dissertation. Chap. 8: Regularity of words and periodic policies. Leiden Univ., Leiden, The Netherlands.
- LOTHAIRE, M. 1991. *Mots*, chapter Tracé de droites, fractions continues et morphismes itérés (J. Berstel). Hermes.
- MORIKAWA, R. 1985–1993. Disjoint sequences generated by the bracket function $i-v$. Tech. Reps. 26(1985), 28(1988), 30(1989), 32(1992), 34(1993). Bulletin of the Faculty of Liberal Arts, Nagasaki Univ., Nagasaki, Japan.
- MORIKAWA, R. 1982–1995. On eventually covering families generated by the bracket function $i-v$. Tech. Reps. 23(1982), 24(1983), 25(1984), 26(1985), 25(1985), 36(1995). Bulletin of the Faculty of Liberal Arts, Nagasaki Univ., Nagasaki, Japan.
- MORSE, M., AND HEDLUND, G. A. 1940. Symbolic dynamics II-Sturmian trajectories. *Amer. J. Math.* 62, 287–306.
- NEWMAN, M. 1971. Roots of unity and covering sets. *Math. Ann.* 191, 279–282.
- SINAI, Y. G. 1976. *Introduction to Ergodic Theory*. Princeton University Press, Princeton, N.J.
- TIJDEMAN, R. 1995. On disjoint pairs of Sturmian bisequences. Tech. Rep. W96-02. Leiden University, The Netherlands.
- TIJDEMAN, R. 1996. On complementary triples of Sturmian sequences. *Indag. Math.* 419–424.
- TIJDEMAN, R. 1998a. Fraenkel's conjecture for six sequences. Tech. Rep. W98-24. Leiden University, The Netherlands.
- TIJDEMAN, R. 1998b. Intertwined periodic sequences, Sturmian sequences and Beatty sequence. *Indag. Math. (N.S.)* 9, 113–122.
- UILLON, L. 1998. Combinatoire des motifs d'une suite sturmienne bidimensionnelle. *Theoret. Comput. Sci.* 209, 261–285.
- WILF, H. S. 1994. *Generatingfunctionology*. 2nd ed. Academic Press, Orlando, Fla.

RECEIVED MAY 1998; REVISED OCTOBER 1999; ACCEPTED OCTOBER 1999