Optimal Open-Loop Control of Vacations, Polling and Service Assignment

Eitan Altman, Bruno Gaujal, Arie Hordijk

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Optimal Open-Loop Control of Vacations, Polling and Service Assignment

Eitan ALTMAN Bruno GAUJAL and Arie HORDIJK

N° 3261

September, 1997

THÈME 1

ISSN 0249-6399

INRIA
SOPHIA ANTIPOLIS

Optimal Open-Loop Control of Vacations, Polling and Service Assignment

Eitan ALTMAN [*] Bruno GAUJAL [**] and Arie HORDIJK [***]

Thème 1 — Réseaux et systèmes
Projet Mistral, Sloop

Rapport de recherche n° 3261 — September, 1997 — 26 pages

Abstract: We consider in this paper the optimal open-loop control of vacations in queueing systems. The controller has to take actions without state information. We first consider the case of a single queue, in which the question is when should vacations be taken so as to minimize, in some general sense, workloads and waiting times. We then consider the case of several queues, in which service of one queue constitutes a vacation for others. This is the optimal polling problem. We solve both problems using new techniques from [2, 4] based on multimodularity.

Key-words: Multimodular functions, balanced sequences, control of vacations, polling, service assignment.

(Résumé : tsvp)

[*] 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

[**] INRIA/UNSA/CNRS(I3S), BP 93, 2004 Route des Lucioles, 06902 Sophia Antipolis Cedex, France. E-mail: gaujal@sophia.inria.fr. Bruno Gaujal is a member of a common project between CNRS, UNSA and INRIA.

[***] Dept. of Mathematics and Computer Science, Leiden University, P.O.Box 9512, 2300RA Leiden, The Netherlands. E-mail: hordijk@wi.leidenuniv.nl. The research of Arie Hordijk was 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.

Contrôle optimal en boucle ouverte de vacances, polling et de service

Résumé : Dans cet article, nous étudions le contrôle optimal en boucle ouverte des vacances dans un système de files d'attente. Le contrôleur doit agir sans information sur l'état du système. Nous considérons d'abord le cas d'une seule file d'attente, la question étant de déterminer quand les vacances du serveur doivent être prises de façon à minimiser, dans un sens général, la charge et les temps d'attente. Nous considérons ensuite le cas de plusieurs files, dans lequel un service dans une file constitue une vacance pour toutes les autres. C'est le problème du polling optimal. Nous résolvons les deux problèmes en utilisant de nouvelles techniques présentées dans [2, 4] qui reposent sur la multimodularité.

Mots-clé : Multimodularité, convexité, contrôle de vacances et de service, polling, suites équilibrées.

1 Introduction

We consider in this paper the control of vacations in several queueing settings. Vacations are time periods during which the server does not serve customers, even when there are some in the system.

We consider in this paper three types of vacation models:
(i) vacations driven by service completions,
(ii) vacations driven (triggered) by arrivals, in which both the beginning as well as the end of a vacation are related to arrivals instants, and
(iii) the potential vacation times form a renewal process and the arrival epochs are stationary subsequence of this renewal process.

We shall consider two types of problems. In both cases, we consider open-loop control where the controller has no information on the state of the system.

In the first problem, there is a single infinite FIFO queue. Some *vacation opportunities* are presented (depending on the type of vacations, these opportunities are triggered by service, or arrivals or by some other mechanism). The server should go on vacation during a fraction of at least p of these opportunities. The goal of the control is to minimize the average workload or waiting time (or any increasing convex functions of these).

The second problem concerns a polling model. There are several infinite queues; when serving one queue, the server is unavailable for other queues. We wish to minimize some linear combination of the average workloads in the different queues (or of waiting times, or of increasing convex functions of these).

The solution is based on (i) establishing the multimodularity of the cost (e.g. of the expected waiting times and workloads), (ii) We use the theoretic results presented in [2, 4] to establish the optimality of policies known as regular policies (that have already been used in [10] in the context of optimal admission control).

This approach has recently been used in the optimal admission control of queues [3] and in the optimal routing into single buffer queues [5]. It has also been used for the optimal control of polling into single buffer queues [5], where the objective was to minimize losses, or maximize the throughputs.

We conclude the introduction by describing the structure of the paper. In Section 2 we formulate the two type of control problems: the one of the single queue (P1), and the one of optimal polling of several queues (P2). We then formulate the four generic type of results obtained for these problems. In the following sections we then present and analyze the different models for the vacations and derive the appropriate results for the control. In Section 3 we analyse the case where vacations are triggered by service completion. In Sections 4 and 5 we consider an arrival driven vacations. Finally, in 6 and 7 we consider models where the vacations are a renewal process.

2 The generic control models and main results

We consider two generic control problems in this paper. We shall formulate these problems in an abstract setting, and then focus in the following sections on specific assumptions.

Constrained model:

- Customers (or some demand for service) arrive to a single G/G/1 queue (or to a network) according to some given stationary process.
- There is a single server at some output of the service facility (queue or network) that may be either active in providing service, or may be absent for vacation periods. Some "vacation opportunities" are presented, in which that server can decide whether to take a vacation or not. These opportunities would depend on the model we consider: they could be related to arrivals, to end of services, or to be a renewal process.
- At the nth vacation opportunity, the controller chooses a control a_n that determines the number $num(a_n)$ of vacations to be taken at the nth vacation opportunity. Let a be the control sequence $(a_1, a_2, ...)$.
- Performance measures and objectives:
  – Let $h : \mathbb{R} \to \mathbb{R}$ be a convex increasing function.
  – Given some fraction p, consider the class $\Pi(p)$ of all policies that satisfy the constraint:

$$\underline{\lim}_{s\to\infty} \frac{1}{s} \sum_{n=1}^{s} num(a_n) \geq p. \tag{1}$$

Consider the following 2 subproblems of

(P1) the vacation control for one queue

(P1a): Let W_n be the waiting time of the nth arriving customer.
Define the average expectation of the function h of the waiting time:

$$g(a) = \overline{\lim}_{s\to\infty} \frac{1}{s} \sum_{n=1}^{s} Eh(W_n(a_1, ..., a_n)), \tag{2}$$

The objective is to minimize $g(a)$ over $a \in \Pi(p)$.

(P1b): Let V_n be the workload in the system at some special time instants T_n.
Define the average expectation of the function h of the workload:

$$g'(a) = \overline{\lim}_{s\to\infty} \frac{1}{s} \sum_{n=1}^{s} Eh(V_n(a_1, ..., a_n)). \tag{3}$$

The objective is to minimize $g'(a)$ over $a \in \Pi(p)$.

Next we describe the nonconstrained problem (P2):

(P2) Polling of several queues

- There are K queues to which a server is allocated. When serving one queue, the server is unavailable for other queues.
- Again, some "vacation opportunities" (or "switching opportunities") are presented, in which the server can decide to stop serving one queue and start serving another one.
- At the nth vacation opportunity, the controller chooses a control $a_n = (a_n^1, ..., a_n^K)$; for each n, all components of a_n are 0 except one component that may be 1 or 0, $a_n^i = 1$ will mean that the server is assigned to queue i at the nth opportunity instant.
  Performance measures and objectives:
  Let $h_i : \mathbb{R} \to \mathbb{R}$ be a convex increasing functions, $i = 1, ..., K$.

  (P2a): Let W_n^i be the waiting time of the nth arrival to queue i. Define

$$g(a) \stackrel{\triangle}{=} \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{i=1}^{K} \left(\sum_{n=1}^{N} f_n^i(a^i) \right),$$

  where $f_n^i \stackrel{\triangle}{=} Eh_i(W_n^i(a)) = Eh_i(W_n^i(a^i)), i = 1, ..., K$. (The last equality expresses the fact that W_n^i will depend on a only through a^i.)
  The objective is to minimize $g(a)$.

  (P2b): Let V_n^i be the workload in the ith queue at some special time instants T_n. Define

$$g'(a) \stackrel{\triangle}{=} \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{i=1}^{K} \left(\sum_{n=1}^{N} \tilde{f}_n^i(a^i) \right),$$

  where $\tilde{f}_n^i \stackrel{\triangle}{=} Eh_i(V_n^i(a)), i = 1, ..., K$.
  The objective is to minimize $g'(a)$.

Next we present four generic results that will be established in the following sections for different models.

Let p and θ be two positive reals. We define the *balanced sequence* $\{a_k^p(\theta)\}$ with rate p and initial phase θ as,

$$a_k^p(\theta) = \lfloor kp + \theta \rfloor - \lfloor (k-1)p + \theta \rfloor, \tag{4}$$

where $\lfloor x \rfloor$ is the largest integer smaller than or equal to x.

In the different models that we study in the next sections, we shall show the following for both problems (P1a) and (P1b):

Result 2.1. *There exists some rate p^* such that for any θ, the sequence $a_k^{p^*}(\theta)$ is optimal.*

In Section 3, we shall establish Result 2.1 and show that $p^* = p$, where p is given in constraint (1). In all other sections where (P1) is considered, we shall have $p^* = 1 - p$. The difference is simply due to different definitions of the control in different models that we study.

Next, we consider problem (P2). For any vector $\theta \in \mathbb{R}^{\mathrm{K}}$ (which is called a phase vector) and a rate vector $p \in [0,1]^K$, we define the vector valued sequence $a(p,\theta)$ by

$$a_n^i(p,\theta) = \lfloor np_i + \theta_i \rfloor - \lfloor (n-1)p_i + \theta_i \rfloor. \tag{5}$$

Note that $a(p,\theta)$ need not correspond to a policy since it may have more than one component that equals to 1 for the same n. In that case we say that it is not feasible, if it defines a policy, we say that it is feasible.

We cite some generic results for problems (P2a) and (P2b) [4].

Result 2.2. *Assume that $K = 2$. There exist some p^* and θ such that $a(p^*,\theta)$ is a feasible policy and is optimal.*

Result 2.3. *Consider an arbitrary K. Suppose costs and service disciplines are symmetric for all queues. Then the round robin policy is optimal for (P2a) (resp. for (P2b)).*

There are settings for $K > 2$ queues other than the symmetric one in which there exists some feasible $a(p,\theta)$ which defines an optimal policy. We shall not specify these, but instead, will give a sketch of the general results in [2] (from which Results 2.2 and 2.3 were derived).

Result 2.4. *The sequence of functions f_n allows to construct some convex function $\overline{f} : \mathbb{R}^{\mathrm{K}} \to \mathbb{R}$ [2]. Let p^* be the vector that minimizes this function. Assume that there is a sequence of numbers $\{i_n\}_n$, where $i_n \in \{1,...,K\}$ such that for every $k \in 1,...,K$ the sequence $a_n^k = 1\{i_n = k\}$ is balanced with rate p_k^* (for some θ that may depend on k). Then $\{a_n^k\}$ are optimal for (P2a) and (P2b).*

The main tool for obtaining the above results is by establishing the multimodularity of some sequence of functions $f_n : \overline{\mathbf{Z}}^n \to \mathbb{R}$, where $\overline{\mathbf{Z}}^n$ is some convex subset of $\mathbf{Z}^n$, the set of n-dimensional vectors of integers. We thus recall the defition of multimodularity.

Let $e_i \in \mathbb{N}^m$, $i = 1, ..., m$ denote the vector having all entries zero except for a 1 in its ith entry. Define $d_i = e_{i-1} - e_i$, $i = 1, ..., m$ (for an integer i taking values between 0 and m, we understand throughout $i - 1 = m$ for $i = 0$).

Let $\mathbb{F} = \{-e_1, d_2, ..., d_m, e_m\}$.

Definition 2.1 ([10, 2]). *A function f on $\overline{\mathbf{Z}}^m$ is multimodular with respect to $\mathbb{F}$ if for all $x \in \overline{\mathbf{Z}}^m$, $v, w \in \mathbb{F}$, $v \neq w$,*

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

whenever $x + w, x + v, x + v + w \in \mathbb{F}$.

We shall also use the fact [10] that a function is multimodular if and only if we replace in the above definition $\mathbb{F}$ by the set $\mathbb{F}' = \{-v : v \in \mathbb{F}\}$. Both $\mathbb{F}$ and $\mathbb{F}'$ are called a basis.

3 A single queue with service driven vacations

Consider a single G/G/1 queue (Problem (P1)). The nth customer arrives at time T_n, bringing a workload of σ_n to the system. Customers are served according to the FIFO order. The arrival process will be assumed to be a point process throughout the paper, unless otherwize stated, and we assume that $T_0 < 0 \leq T_1$.

Let $\tau_n = T_{n+1} - T_n$ denote the inter-arrival times. When a service of a customer is completed, the server is allowed to go on vacation. We consider the, so called, repeated vacation model, where on each completion of a vacation, another vacation can be initiated.

In this model, "vacation opportunities" are thus triggered by the end of a service or of a vacation.

Let $a = (a_1, a_2...)$ be the server's policy, where $a_i \in \mathbb{N}$ has the interpretation of the number of vacations to be taken after the ith service time completion. (In terms of the notation of Section 2, we have $num(a_n) = a_n$.)

Let $v_n, n = 0, 1, 2, ...$ be the duration of the nth vacation period. Let $m(n)$ denote the number of vacations that occur till the $n + 1$st service starts. We set $m(0) = 0$. Denote by

$$S_n \stackrel{\triangle}{=} \sigma_n + \sum_{j=m(n-1)}^{m(n)-1} v_j \tag{7}$$

the total delay related to the nth customer. It is the sum of its service time, plus the vacations that will take place after its service. The waiting time W_n of the nth customer is given recursively by

$$W_{n+1} = (W_n - \tau_n + S_n)^+. \tag{8}$$

In particular, assume that the system is initially empty. If no vacations are taken before the service of the second customer then $m(1) = 0$, and the waiting time of the second customer is

$$W_2 = (W_1 - \tau_1 + S_1)^+ = (-\tau_1 + \sigma_1)^+.$$

If, instead, the 1st vacation is taken just after the service of the 1st customer, then $m(1) = 1$ and

$$W_2 = (W_1 - \tau_1 + S_1)^+ = (-\tau_1 + \sigma_1 + v_0)^+.$$

Let $V_n = V_n(a)$ be the virtual workload in the system immediately after the nth arrival; it is defined to be the total time required by the server to serve all the customers actually in the system (including the one that arrives at time T_n) plus all the vacation times that will elapse from the arrival instant T_n until customer $n+1$ is served. V_n is given by

$$V_n = W_n + S_n. \tag{9}$$

Fix some integer N. $W(a) \stackrel{\triangle}{=} W_{N+1}(a)$ can be written as

$$W(a) = \max(0, w_1 + x, w_2, ..., w_N), \text{ where } w_i = \sum_{j=i}^{N} (S_j - \tau_j). \tag{10}$$

Here, $x = 0$ is the initial workload in the system. Define $V(a) \stackrel{\triangle}{=} V_N(a)$.

Denote E_v the expectation over the v's. Below we shall establish the multimodularity of $E_v h(W(a))$, where h is any nondecreasing convex function. The dynamics of vacation model resembles the one of the admission control model in [3].

Property 3.1. *The following holds for* $0 < i < N$. *If* $a_i \geq 1$ *then*

$$w_i(a - d_i) = w_i(a) + v_{m(i)}, \qquad w_j(a - d_i) = w_i(a) \quad \textit{for } j \neq i.$$

Note that $(-d_i)$ *corresponds to adding a vacation after the end of service of the ith customer, and delete the last vacation from the* $i-1$*st one.*

Property 3.2. *Consider the vacation sequence* $v = (v_0, ...)$, *and the shifted sequence:* $v' = (v'_0, v'_1, ...) = (v_1, v_2, ...)$. *Let* w'_i *be defined as* w_i *in (10) with the sequence* v' *replacing the original one. The following holds for* $0 < i < N$.

$$w'_j(a + e_i) = w_i(a) \qquad j > i,$$

$$w'_j(a + e_i) = w_i(a) + v_{m(j-1)} \qquad j \leq i.$$

Lemma 3.1. *Assume that v_n are stationary with respect to the 1-step shift. Let $h : \mathbb{R} \to \mathbb{R}$ be a nondecreasing convex function. Then $E_v h(W(a))$ and $E_v h(V(a))$ are multimodular in a.*

Proof. We consider the basis $\mathbb{F} = (e_1, -d_2, ..., -d_{N-1}, -e_m)$ and check the condition $h(W(a-v)) + h(W(a-w)) \geq h(W(a)) + h(W(a-v-w))$, $v \neq w$.

Case 1: we check for $d_i, d_j, i \neq j$.

$$W(a-d_i) = \max(W(a), w_i + v_{m(i)}), \qquad W(a-d_j) = \max(W(a), w_i + v_{m(j)}),$$

$$W(a-d_i-d_j) = \max(W(a), w_i + v_{m(i)}, w_j + v_{m(j)}), \quad j \neq i.$$

If $W(a)$ is maximizing in the above equation, then

$$h(W(a-d_i)) + h(W(a-d_j)) = 2h(W(a)) = h(W(a)) + h(W(a-d_i-d_j))$$

and the condition is satisfied. It is then also satisfied for $V(a)$ since $V(a) = W_{N-1}(a) + S_N(a)$, and $S_N(a)$ is the same for $a, a-d_i, a-d_j$ and $a-d_i-d_j$. If the maximizer is $w_i + v_{m(i)}$, then $W(a-d_i-d_j) = W(a-d_i)$, and the condition follows from the monotonicity of h. By symmetry we obtain the argument for j instead of i. The same argument holds for $V(a)$. Since this inequality holds samplewise, it also holds in expectation.

Case 2: we check for the first term of the basis. It corresponds to adding an additional vacation v_0 after the service of the first customer. In order to check the inequality for the expectation, we consider the following coupling. We consider a second system defined on the same probability space. Quantities in the new system will be denoted with an over-line. We let $\overline{v}_{n+1} = v_n$ for all n.

We compute $W(a)$ and $W(a-d_i)$ in our original system, and compare them to $\overline{W}(a+e_1)$ and $\overline{W}(a-d_i+e_1)$ in our new system.

$$\overline{W}(a+e_1) = \max(W(a), w_1 + v_0), \quad \overline{W}(a+e_1-d_i) = \max(W(a), w_1 + v_0, w_1 + v_0 + v_{m(i)}),$$

and $W(a-d_i) = \max(W(a), w_1 + v_{m(i)})$. The condition for the multimodularity holds for both $h(W)$ and $h(V)$ by arguments as in Case 1. Since this inequality holds for any sample, it holds in expectation.

Case 3: we check for the last term of the basis. It corresponds to removing the last vacation $v_{m(N)}$.

$$W(a-e_N) = (W(a) - v_{m(N)})^+ \tag{11}$$

$$W(a-e_N-d_i) = \max(W(a) - v_{m(N)}, w_i + v_m(i) - v_m(N))^+. \tag{12}$$

If the argmax of the last maximization is 0, then $W(a-e_N) = W(a-e_N-d_i)$ and the multimodularity condition is seen to hold (since h is nondecreasing). If it is not 0, then $W(a-d_i) - W(a-e_N-d_i) = v_{m(N)}$. Hence

$$W(a) - W(a-e_N) \leq W(a-d_i) - W(a-e_N-d_i) = v_{m(N)}.$$

Since h is convex nondecreasing and $W(a-e_N-d_i) \geq W(a-e_N)$,

$$\begin{aligned} h(W(a-d_i)) - h(W(a-e_N-d_i)) &= h(W(a-e_N-d_i)+v_{m(n)}) - h(W(a-e_N-d_i)) \\ &\geq h(W(a-e_N)+v_{m(n)}) - h(W(a-e_N)) \\ &\geq h(W(a) - h(W(a-e_N)). \end{aligned}$$

Next, we check this case for the workload. We have

$$V(a-e_N) = V(a) - v_{m(N)} \tag{13}$$

$$V(a-e_N-d_i) = V(a-d_i) - v_{m(N)} \tag{14}$$

Since h is convex nondecreasing, and since $V(a-d_i) \geq V(a)$, this implies that

$$\begin{aligned} h(V(a)) - h(V(a-e_N)) &= h(V(a)+v_{(n)}) - h(V(a)) \leq h(V(a-d_i)+v_{(n)}) - h(V(a-d_i)) \\ &= h(V(a-d_i)) - h(V(a-e_N-d_i)). \end{aligned}$$

Thus the multimodularity condition holds for $h(V)$ as well.

Again, the condition for the multimodularity holds samplepathwise, and thus in expectation. ∎

Using Theorem 4.1 in [2] we get:

Theorem 3.1. *Consider problem (P1), where the expectation E in (2) is with respect to all random sequences (and is denoted by $E_{\tau,\sigma,v}$), and where $num(a_n) = a_n$. Assume that*

- *the inter-arrival and service times (τ_n, σ_n) are stationary with respect to the 1-step shift, and are independent of the sequence v_n. (They may however depend on each other.)*
- *the v sequence is stationary with respect to the 1-step shift,*
- *The following stability condition holds: the queue is in a stationary regime at time 0, corresponding to the policy that does not take vacations (see more details in Remark 3.1 below).*

Then Result 2.1 holds for (P1a) and (P1b) where $p^* = p$ is the fraction given in the constraint (1).

Proof. The proof is based on Theorem 3.3 in [2]. We check the following conditions for that Theorem hold. in [2]. We have to show that for any integer n,

$$f_n(a) \text{ is nondecreasing in } a_i,\ i = 1, ..., n \tag{15}$$

where $f_n(a) = E_{\tau,\sigma,v} h(W_n(a_1, ..., a_n))$, that it satisfies

$$f_k(a_1, ..., a_k) = f_m(\underbrace{0, ..., 0}_{m-k}, a_1, ..., a_k),\ k < m, \tag{16}$$

(this implies conditions $< 2 >$ and $< 3 >$ in [2]) and that it is multimodular (condition $< 1 >$ in [2]).

The monotonicity condition follows from Property 3.2, the stationarity of the vacation times, and the fact that the vacations are independent of the interarrival and service times.

The second condition is satisfied due to the stationarity assumptions. Indeed, since the system is assumed to be in a stationary regime at time 0, corresponding to the policy that does not take vacations, the Palm probability P_N^0 (of the process seen at the times T_n) is invariant under the shift Θ^{τ_1} (see, e.g. [7] p. 19). In particular, if we do not take vacations till time n, then $W_k(a) = W_1(a)$, $k < n$ in distribution. Hence, the distribution of W_n under the policy a is the same as the distribution of W_{n+j} under the policy $a' = (\underbrace{0, ..., 0}_{j}, a_1, ..., a_k, ...)$, for any $j \geq 0$. This implies (16).

The multimodularity condition was established in Lemma 3.1. ∎

Remark 3.1. A sufficient condition for the stability condition in Theorem 3.1 is that

- (τ_n, σ_n) is stationary ergodic (with respect to the one step shift), and
- $E\sigma_1 < E\tau_1$

(see [8, 9]). (This sufficient condition also implies coupling to the stationary regime from any initial state, provided that we do not take vacations.)

Extension to an arbitrary network

Let queue i be one of several possible output queues of an arbitrary network. Assume that every customer that is served in that queue leaves the system. Then the waiting time of the nth customer equals to its sojourn time till it arrives to that queue, plus its waiting time in queue i. Since customers served at queue i are not rerouted, the first component of the waiting time does not depend on the polling strategy. The total waiting time of a customer is thus the sum of a part that does not depend on the control, plus the waiting time in a GG1 queue which is influenced by the controller of the vacation. It is thus multimodular, due to the results of the first part. Hence the optimality of balanced policy for the total average waiting time is directly obtained.

4 An arrival-driven vacation model

Consider a single G/G/1 queue (problem (P1)). The nth customer arrives at times T_n, bringing a workload of σ_n to the system. Let $\tau_n = T_{n+1} - T_n$ denote the inter-arrival times. Immediately after an arrival occurs, the server may go on vacation, that lasts till the next arrival occurs. Then it may go back serving, or take another vacation, etc. A vacation policy $a = (a_1, a_2, ...)$ indicates, for each n, whether the server goes on vacation ($a_n = 0$) of continues serving ($a_n = 1$) immediately after the nth arrival. (In terms of the notation of Section 2, we have $num(a_n) = 1 - a_n$.)

We call this system an *arrival driven* vacation model since the beginning and end of vacations are initiated by arrivals.

In this section, and in all following sections that deal with problem (P1), the constraint (1) translates to the following one, directly on the rates of a_n:

$$\overline{\lim_{s\to\infty}} \frac{1}{s} \sum_{n=1}^{s} a_n \leq 1 - p. \tag{17}$$

The waiting time $W_n(a)$ of the nth arriving customer is given recursively by

$$W_{n+1}(a) = (W_n(a) + \sigma_n - a_n\tau_n)^+, \tag{18}$$

or explicitly by

$$W_{n+1}(a) = \max(0, w_1, w_2, ..., w_n), \text{ where } w_i = \sum_{j=i}^{n} (\sigma_j - a_j\tau_j). \tag{19}$$

The workload in the system immediately after the nth arrival is given by $V_n = W_n + \sigma_n$. Note that is satisfies the recursion

$$V_{n+1}(a) = (0, V_n(a) - a_n\tau_n)^+ + \sigma_{n+1}. \tag{20}$$

The equation (18) seems dual to the dynamics of the admission control in [3]. Therefore it seems natural to expect to obtain the same type of multimodularity results, and therefore also the optimization results. In order to obtain multimodularity in [3] it is necessary to let σ_n, the nth service time, be the service of the nth customer actually accepted. Thus, the service time of a customer that is rejected is not defined. Then stationarity conditions are assumed on this σ_n sequence, rather than on the sequence of service times of all customers (both the ones accepted as well as the ones rejected).

We thus proceed similarly, and define τ'_n to be the duration of the nth slot during which a vacation *was not* taken. These are the effective interarrival times, since, as we see in (18), only these have influence on the dynamics of W_n and V_n.

Let $k(n) := \sum_{i=1}^{n} a_i$. Then

$$W_{n+1} = \max(0, W_n + \sigma_n - a_n \tau'_{k(n)}). \tag{21}$$

To see why (21) holds, we first note that it agrees with (18) for those n's for which $a_n = 0$. On the other hand, if $a_n = 1$ then $\tau'_{k(n)} = \tau_n$, so (21) again agrees with (18).

If we now assume that τ'_n (and not τ_n) are stationary, we could expect to obtain results dual to those of the admission control.

However, this does not seem natural: it would mean that the vacation control decisions influence the actual interarrival times. In the case of i.i.d. interarrival times, however, both τ'_n and τ_n are stationary. We shall therefore use the i.i.d. assumption below.

Let

$$S_n = \sigma_n - a_n \tau'_{k(n)}.$$

The waiting time of the $n+1$st customer is given by $W_{n+1} = (W_n + S_n)^+$. The workload just after this arrival is $V_n = W_n + \sigma_n$.

Remark 4.1. *Since σ_n does not depend on a, V_n is multimodular if and only if W_n is.*

$W(a) := W_{N+1}(a)$ is given explicitely by

$$W(a) = \max(0, w_1, w_2, ..., w_N), \text{ where } w_i = \sum_{j=i}^{N} S_j. \tag{22}$$

We can now obtain the multimodularity of the expected waiting time as we did in the first section. The corresponding properties are:

Property 4.1. *The following holds for $0 < i < N$. If $a_i \geq 1$ then*

$$w_i(a + d_i) = w_i(a) + \tau'_{k(i)}, \qquad w_j(a + d_i) = w_i(a) \quad \textit{for } j \neq i.$$

Note that d_i corresponds to removing a vacation at time T_{i-1} and adding the vacation at T_i.

Property 4.2. *Consider the sequence of effective interarrival times $\tau' = (\tau'_0, ...)$, and the shifted sequence: $\Theta\tau' = (\tau'_1, \tau'_2, ...)$. Let w'_i be defined as w_i in (22) with the sequence $\Theta\tau'$ replacing the original one. The following holds for $0 < i < N$.*

$$w'_j(a - e_i) = w_i(a) \qquad j > i,$$

$$w'_j(a - e_i) = w_i(a) + \tau'_{k(j-1)} \qquad j \leq i.$$

Denote E_τ the expectation over the effective interarrival times.

Lemma 4.1. *Assume that τ_n are i.i.d. Let $h : \mathbb{R} \to \mathbb{R}$ be a nondecreasing convex function. Then $E_\tau h(W(a))$ and $E_\tau h(V(a))$ are multimodular in a.*

Proof. The proof for the expected waiting time is the same as the one of Lemma 3.1, with $\tau'_{k(i)}$ replacing $v_{m(i)}$. The proof for the expected workload follows from Remark 4.1. ∎

Theorem 4.1. *Assume that*

- *(i) the service times (σ_n) are stationary with respect to the 1-step shift,*
- *(ii) the interarrival times (τ_n) are i.i.d. and independent of (σ_n),*
- *(iii) the queue is initially (at time 0) in a unique stationary regime corresponding to the policy that never takes vacations.*

Then Result 2.1 holds where $p^* = 1 - p$, and where p is the fraction given in the constraint (1) (or (17)).

Proof. The proof is based on Theorem 3.4 in [2]. We have to check the following conditions for that Theorem hold. We have to show that

- for any integer n,

$$f_n(a) \text{ is nonincreasing in } a_i,\ i = 1, ..., n \tag{23}$$

  where $f_n(a) = E_{\tau,\sigma} h(W_n(a_1, ..., a_n))$,

- the following holds:

$$f_k(a_1, ..., a_k) \geq f_{k-1}(a_2, ..., a_k), \forall k > 1 \tag{24}$$

  (this is condition $< 2 >$ in [2]),

- $$f_k(a_1, ..., a_k) = f_m(\underbrace{1, ..., 1}_{m-k}, a_1, ..., a_k),\ k < m, \tag{25}$$

  (this implies condition $< 3 >$ in [2]) and

- $f_k(a)$ is multimodular (condition $< 1 >$ there).

The monotonicity (23) follows directly from the explicit solution to the Lindley equations (19). (24) and (25) follow from Property 4.2, the assumption that the initial state is initially in the stationary regime corresponding to no vacations, the fact that τ_n are i.i.d. (and thus stationary), and the fact that they do not depend on σ_n. For (24) we also make use of the monotonicity property established in (23). ∎

In Theorem 4.1, we assume that the queue is initially (at time 0) in a unique stationary regime corresponding to the policy that never takes vacations. Let β_0 be the corresponding distribution of the initial waiting time W_1. We next show that the results hold for other distributions β as well.

Lemma 4.2. *(Relaxing the assumption on the initial condition)*
Consider any initial distribution β of W_1. Assume that conditions (i) and (ii) of Theorem 4.1 hold and instead of condition (iii), the following is satisfied:

- *β is stochastically larger than the β_0,*
- *σ_n and τ_n are stationary ergodic under the k-shift for any integer k,*
- *the stability condition $E\sigma < (1-p)E\tau$ holds.*

Then the results of Theorem 4.1 still hold.

Proof. For any policy a, both the waiting time as well as the workload at any time n are strictly increasing in W_1 as can be seen from (21). By the definition of stochastic ordering, the expectation of any increasing function of W_1 is larger for the initial distribution β, than for β_0. This implies that the average expected costs $g(a)$ and $g'(a)$ are larger for the initial distribution β. In order to establish the theorem it suffices thus to show that for the optimal policy, the average expected costs do not depend on the initial state.

If p (in the constraint (1)) is rational, then the (candidate for the) optimal policy $a_k^{p^*}(\theta)$ (with $p^* = 1-p$) is periodic. In that case, the process corresponding to the optimal policy starting from any two different initial states couple in a time that is finite with probability one (see Section 6 of [1]). We may couple now the initial states, i.e. construct a common probability space where the initial state corresponding to β is larger than the one corresponding to β_0. It follows that the convergence of the difference between W_n corresponding to β and to β_0 is *monotone* decreasing. This implies that the difference between $f_n(a) = E_{\tau,\sigma}h(W_n(a_1,...,a_n))$, starting at the different initial distributions of W_1, converges to 0. Hence the expected average cost under the two initial distributions is the same under the optimal policy.

The same convergence (and hence the same conclusion) is obtained for p irrational. Indeed, if p is irrational, then the (candidate for the) optimal policy $a^{p^*}(\theta)$ is aperiodic. This

cost obtained by that policy is unchanged we replace θ by a random variable Θ, uniformly distributed in $[0,1]$ (this follows from [2] Theorems 3.1, 3.2, and equation (28) there). The policy $a^{p^*}(\Theta)$ is stationary ergodic with respect to the 1-step shift. Then we can use the theory of (noncontrolled) stochastically recursive sequences by Borovkov [8] pp. 260-272 and [9] to obtain the same convergence results as above. ∎

Remark 4.2. *The assumption that β is stochastically larger than the β_0, is not really restrictive. Indeed, if β_t^u is the distribution of the state at time t, then one can show using [1] that for any policy u, $\underline{\lim}_{t\to\infty} \beta_t^u$ is stochastically larger than the β_0. Hence the assumption is suitable for the case where the system has operated for a sufficiently long time under an arbitrary policy.*

5 Arrival-driven polling model

We now analyze problem (P2). Consider K queues, each of which behaves like the one in the previous section. The service period for one queue constitutes a vacation for the others.

The nth customer arrives at time T_n, and brings K jobs to the K queues: a workload of σ_n^i arrives to queue i, $i = 1, ..., K$. These components are processed according to the FIFO order in each queue.

Service beginning and vacations are synchronized with arrivals. More precisely, T_n is also the time at which the nth potential service begins; it may be in any one of the queues. The service time duration τ_n is the difference between consecutive interarrival times: $T_{n+1} = T_n + \tau_n$. If queue i is the nth to be served and is empty then we assume that the server still remains τ_n time at that queue.

Let W_n^i be the waiting time of the nth job arriving to queue i, and V_n^i denotes the workload at queue i just after the arrival of the n customer. The evolution of the waiting time $W_n^i(a)$ in queue i is given by:

$$W_{n+1}^i(a) = \max(0, W_n^i(a) + \sigma_n^i - a_n^i \tau_n),$$

where $a_n^i = 1$ if queue i is served at the nth period, and is, otherwise, zero.

A policy is a sequence $a = (a_1, a_2, ...,)$, where $a_n = \{a_n^k, k = 1, ..., K\}$, as defined in Section 2. We consider the further constraint that for every integer j, only one of the components $a_j^i, i = 1, ..., K$ may be different than 0. Denote $a^i = (a_1^i, a_2^i, ...)$ the actions corresponding to queue i.

The following is a consequence of Section 4 in [2] and the properties established for a single queue in the previous section.

Theorem 5.1. *Consider problems (P2a) and (P2b). Assume that*

- *the inter-arrival times are i.i.d., and independent of the service times,*
- *the service times in each queue is stationary with respect to the* 1*-step shift,*
- *for each* $i = 1, ..., K$*, queue* i *is initially at a unique stationary regime that corresponds to the policy that never takes vacations at that queue.*

Then Results 2.2, 2.3 and 2.4 hold.

Again, we may relax the assumption on the initial distribution, as we did in Lemma 4.2.

6 The potential vacation times are a renewal process

6.1 A single queue

Let u_n be an increasing random sequence of potential switching times. Immediately after u_n, the server may decide to go on vacation till the next instant u_{n+1}. As in the previous section, a vacation policy $a = (a_1, a_2, ...)$ indicates, for each n, whether the server goes on vacation ($a_n = 0$) of continues serving ($a_n = 1$) at time u_n.

Let s_n be the sequence of differences between consecutive potential switching times. Thus $u_{n+1} = u_n + s_{n+1}$.

The kth customer arrives at time $T_k = u_{n_k}$, where n_k is some increasing sequence of positive integers. Thus, u_n can be viewed as basic time epochs to which both arrivals and vacations are related. However, unlike the model in the previous section, where arrivals occurred at beginning of each vacation slots, arrivals are only synchronized with u_n, and need not occur at every period. This will allow us to handle dependent arrival times, and more precisely the case of stationary interarrival arrivals.

Customer k brings a workload (request for service time) of σ_k^*. Hence, the amount σ_n of workload that arrives at time u_n is given by

$$\sigma_n = \begin{cases} 0 & \text{if } n \neq n_k, \forall k \\ \sigma_k^* & \text{if for some } k, n = n_k. \end{cases} \tag{26}$$

The waiting time W_n of the (possibly virtual) customer that arrives at time u_n in the system can now be computed using the following recursion:

$$W_{n+1}(a) = \max(0, W_n(a) + \sigma_n - a_n s_n).$$

The workload V_n at the nth time epoch (i.e. immediately after u_n) is $W_n(a) + \sigma_n$. It can also be given recursively as

$$V_{n+1}(a) = \max(0, V_n(a) - a_n s_n) + \sigma_{n+1}. \tag{27}$$

We are now back to the model described by (18) of section 4, and therefore, the multimodularity results in Lemma 4.1, and the optimality results in Theorem 4.1 hold.

It is useful to present conditions directly on the original service sequence σ^* (instead of the sequence σ which are used in Theorem 4.1) for the optimality results.

In order to make general and yet useful probabilistic assumptions on σ_n (i.e. on the marks of the arrival process), we use the stochastic point process formalism. The sequence n_k, which we used in (26), defines a *discrete time* point process $(\mathcal{N}, \vartheta, P)$ (where ϑ is the θ_1 of [7] p. 43):

$$\mathcal{N}(\omega, C) = \sum_{k \in \mathbf{Z}} \delta_{n_k(\omega)}(C).$$

Thus,

$$\text{for } w \in \Omega, \text{ an arrival occurs at time } u_n(\omega) \text{ if } \mathcal{N}(\omega, \{n\}) = 1. \tag{28}$$

We associate to the process $\mathcal{N}$ the marks σ_k^*. We assume that $\mathcal{N}$ is compatible with the ϑ flow, i.e.

$$n_k(\omega) = n_0(\vartheta_k \omega).$$

Assume that the interarrival times are a strictly stationary sequence, i.e.

$$P^0_{\mathcal{N}}(\vartheta_{n_k} \in \cdot) = P^0_{\mathcal{N}}(\cdot), \qquad k \in \mathbf{Z}, \tag{29}$$

where $P^0_{\mathcal{N}}$ is the Palm probability related to $\mathcal{N}$. Then there exists a probability measure P for which $\mathcal{N}$ is stationary (w.r.t. (ϑ, P)). This follows from the inverse construction of Slivnyak (see p. 27 in [7]) in a discrete-time version (which follows from p. 44 in [7]).

Define

$$\overline{\sigma}(l) = \sigma_k^* \quad for \quad n_k \leq l < n_{k+1}.$$

Then $(\mathcal{N}, \overline{\sigma})$ are jointly stationary (with respect to (ϑ, P)) as follows from the argument in [7] p. 13-14. Since

$$\sigma_n = \overline{\sigma}(n) \times \mathcal{N}(\{n\}),$$

it then follows that σ_n are stationary w.r.t. (ϑ, P). Indeed,

$$\sigma_n(\omega) = \overline{\sigma}(n, \omega) \times \mathcal{N}(\omega, \{n\}) = \overline{\sigma}(n-1, \vartheta\omega) \times \mathcal{N}(\vartheta\omega, \{n-1\}) = \sigma_{n-1}(\vartheta\omega). \tag{30}$$

We conclude that if we assume that the original process σ_n^* is stationary, then there exists a probability measure under which $\mathcal{N}$ is stationary (w.r.t. (ϑ, P)) (thus in particular, the process σ_n will be stationary). Note that, in general, there may other nonstationary processes $\mathcal{N}$ that have a stationary Palm distribution.

If we assume that the service times are independent of the vacation opportunities times s_n and of the sequence n_k, then a simple argument shows that the stationarity of σ_n^* implies that σ_n are stationary too. Indeed, assume as above, that σ_n^* are the marks of the process $\mathcal{N}$, and assume that (29) holds. Fix some integer j and let $S_1, S_2, ..., S_j$ be some Borel sets in $\mathbb{R}$. Then

$$\begin{aligned} P(\sigma_1^* \in S_1, ...\sigma_j^* \in S_j) &= \sum_{k=-\infty}^{\infty} P(\sigma_1^* \in S_1, ...\sigma_j^* \in S_j | n_1 = k) P(n_1 = k) \\ &= \sum_{k=-\infty}^{\infty} P(\sigma_1^* \in S_1, ...\sigma_j^* \in S_j | n_1 = 0) P(n_1 = k) \\ &= P(\sigma_1^* \in S_1, ...\sigma_j^* \in S_j | n_1 = 0) = P_{\mathcal{N}}^0(\sigma_1^* \in S_1, ...\sigma_j^* \in S_j). \end{aligned}$$

Hence, the stationarity of σ_n^* under $P_{\mathcal{N}}^0$ implies that it is stationary under P, and if u_n are i.i.d. then the process $\mathcal{N}$ is stationary (w.r.t. (ϑ, P)). Thus, as in (30), we conclude that σ_n are stationary.

We summerize this in the following Theorem.

Theorem 6.1. *Assume that*

- *the inter-potential vacation times $\{s_n\}$ are i.i.d. and hence $\{u_n\}_n$ is a renewal process,*
- *arrivals occur at u_{n_k}, where $\{n_k\}_k$ defines a point process $\mathcal{N}$,*
- *the service times (σ_n^*) are the marks of the point process $\mathcal{N}$,*
- *σ_n^* are stationary,*
- *the duration of the potential vacations s_n do not depend on the service durations and on the sequence n_k,*
- *the queue is initially (at time 0) in a unique stationary regime corresponding to the policy that never took vacations.*

Then Result 2.1 holds where $p^* = 1 - p$, and where p is the fraction given in the constraint (1) (or (17)).

Proof. We show that Theorem 4.1 can be applied. As we showed above, we can consider an equivalent model where arrivals occur at each time u_n instead of the original ones. The service time for this new model are σ_n, which are stationary. Due to the independence between s_n, n_k, and the service duration, the interarrival times in the new model are independent of the service times. The conditions of Theorem 4.1 are thus satisfied. (Note that the fact that in the new model, arrivals occur at times u_n which are independent of other quantities, allows to have dependence between the n_k sequence.) ∎

Note that we may relax the assumption on the initial distribution, as we did in Lemma 4.2.

6.2 The polling control problem

Having seen that the setting described in the previous subsection for a single queue can be embedded into the one in Section 5, we can obtain the corresponding results for the optimal control problem (P2) of polling to several queues.

Theorem 6.2. *Consider problems (P2a) and (P2b). Assume that*

- *the potential switching times* $\{u_n\}$ *are a renewal process,*
- *the service times* $(\sigma_n^{*,i})$ *of the nth customer at queue* $i, i = 1, ..., K$ *are stationary with respect to the* 1*-step shift,*
- *arrivals to queue* i *occur at times* $u_{n_k(i)}, i = 1, ..., K$*, where* $n_k(i)$ *is a stationary point process,*
- *the duration of the slots* $s_{n+1} = u_{n+1} - u_n$ *do not depend on the service durations and on the sequences* $n_k(i)$*,*
- *for each* $i = 1, ..., K$*, queue* i *is initially at a unique stationary regime that corresponds to the policy that never takes vacations at that queue.*

Then Results 2.2, 2.3 and 2.4 hold.

Note that the fact that service times in different queues were allowed to be dependent in Theorem 5.1 allows us to have dependence between the $n_k(i), i = 1, ..., K$ sequences in different queues.

We may relax the assumption on the initial distribution, as we did in Lemma 4.2 and Remark 4.2.

Remark 6.1. *The asumptions of Theorem 6.2 contain as a special case the following exponential model. Suppose the arrival process are independent Poisson process with rates* λ_j *for queue* $j = 1, ..., K$*. The service time in queue* j *is exponential with parameter* $\mu_j, i = 1, ..., K$*.*

The potential switching times form a Poisson process with parameter $\nu \geq (\lambda_1 + ... + \lambda_K)$, this is a natural assumption on ν as in case we want to uniformize all processes, ν is taken as $\sum_{i=1}^{K}(\lambda_i + \mu_i)$. Also as approximation of continuous-time polling control we may take ν large. Theorem 6.2 now shows the regularity of the optimal polling control for the exponential model with $K = 2$ and for the symmetrical model with $K > 2$. This shows Property 1 in paper [11], where an algorithm for computing optimal policies is given.

7 1-gated service

We describe in this section models that have stationary arrival processes which may be more general than point processes. The vacations opportunities in the following models will be a periodic process, independent of inter-arrival times or service times.

7.1 A single queue

We now consider a vacation as in the previous section, but with a "continuous" arrival into a single queue. We assume that the total workload that arrives during the interval $(u_n, u_{n+1}]$ is σ_n. This workload might arrive in a single batch, or continuously, or at several distinct instants in that interval. If the server does not go on vacation on time u_n, then the amount of service given to the queue till $u_{n+1} = u_n + s_{n+1}$ is the minimum between s_{n+1} and V_n (the workload present just before the nth potential switching-interval). Thus, only workload present in the gating epoch u_n is candidate to be served during the interval $(u_n, u_{n+1}]$. We assume that if the server is not on vacation at the beginning of the nth slot, then it remains in that queue till time u_{n+1}, even if there is no workload to be served during a part (or all) the interval.

We shall require that the process σ_n be stationary in n (i.e. w.r.t. the shift ϑ_1). The recursion (20) for the workload in the system at gating instants holds in our case, so we could obtain again optimality of the balanced policy (as in Theorem 4.1).

In order for the conditions of Theorem 4.1 to hold we need, however, that σ_n be independent of s_n. This is impossible in general, unless s_n are identical. (For example, assume Poisson arrivals with rate λ, where each customer requires a unit of workload. Then the expected amount of workload arriving during a period v_n, conditioned on s_n, is λs_n. Hence it is not independent of s_n). Note that this problem did not occur in previous sections, since the service time was not related to the arrival instants, but to the order of arrival.

Consider an underlying probability space $(\Omega, \mathcal{F})$. Define $\{\theta_t\}, t \in \mathbb{R}$ to be a measurable flow on $(\Omega, \mathcal{F})$ (see [7] p. 8 and Remark 7.1 below). We define $\{\vartheta_n\}, n \in \mathbf{Z}$ to be another measurable flow on $(\Omega, \mathcal{F})$ (see [7] p. 44); ϑ_n will be related to shifts in discrete time.

Consider a general random measure Z describing the arriving workload; in particular, if C is an interval in $\mathbb{R}$, then $Z(\omega, C)$ has the interpretation of the amount of workload arriving during that interval for a realization ω. This includes in particular the case where the arrival process is a point process. Assume that Z is stationary with respect to (θ_t, P) (θ_t is the continuous time shift). Then the amount of workload σ_n that arrives during the deterministic periods s_n is stationary in n (i.e. w.r.t to (ϑ_1, P)) due to the stationarity of Z w.r.t. (θ_t, P). Hence the conditions of Theorem 4.1 hold for any stationary arrival process (not necessarily a point process).

Remark 7.1. *Consider a more general model for the potential vacation process. Let (N, θ_t, P) be a stationary point process corresponding to the potential vacations: associated with N there is given the random sequence u_n, $n = 1, 2, ...$ We have*

$$N(\omega, C) = \sum_{n \in \mathbf{Z}} \delta_{u_n(\omega)}(C), \tag{31}$$

where we assume $-\infty \leq ... \leq u_{-1} \leq u_0 \leq 0 \leq u_1 \leq u_2 ... \leq \infty$ and δ_x is the Dirac measure at x. σ_n can then be considered as marks of the point process N:

Let Z be stationary with respect to (θ_t, P). Then

$$\sigma_n(\omega) \stackrel{\triangle}{=} Z(\omega, [u_n, u_{n+1}))$$

satisfies

$$\sigma_n(\omega) = \sigma_0(\theta_{u_n}\omega) = \sigma_0(\vartheta_n\omega).$$

Hence, $((N, \sigma), \theta_t, P)$ is a stationary marked point process (see [7] p. 10) and σ_n is stationary in n (see Section 1.3.2 in [7]), i.e. w.r.t. ϑ.

To summerize, we have:

Theorem 7.1. *Assume that*

- *the potential vacation durations (s_n) are constant,*
- *The amount of workload σ_n arriving arriving during the nth slot are stationary,*
- *the queue is initially (at time 0) in a unique stationary regime corresponding to the policy that never took vacations.*

Then Result 2.1 holds where $p^* = 1 - p$, and where p is the fraction given in the constraint (1) (or (17)).

We may relax the assumption on the initial distribution, as we did in Lemma 4.2 and Remark 4.2.

7.2 The case of several queues

Consider K queues, each of which behaves like the one in the queue in Subection 7.1. The service period for one queue constitutes a vacation for the others.

More precisely, let u_n, $n = 1, 2, ...$ be the time at which the nth potential service ends; it may be in either one of the queues. The service time duration s_n is the difference between consecutive inter-switching times: $u_{n+1} = u_n + s_{n+1}$. If queue i is the nth to be served and is empty then we assume that the server still remains s_n time at that queue.

Let σ_n^i be the amount of workload that arrives to queue i during the interval $(u_n, u_{n+1}]$. As in Subsection 7.1, we assume a 1-gated regime, where, only workload that is present at time u_n is candidate to being served during the interval $(u_n, u_{n+1}]$, and not workload that arrives during that interval.

The evolution of the waiting time $W_n^i(a)$ in queue i is given by:

$$W_{n+1}^i(a) = \max(0, W_n^i(a) + \sigma_n^i - a_n^i s_n),$$

where $a_n^i = 1$ if queue i is served at the nth period, and is, otherwise, zero. Here, W_n^i has the interpretation of the waiting time of a customer that would arrive at time u_n, and V_n^i is the workload in the system just after time u_n.

From the discussion in Subsection 7.1, we obtain the following results (with notation similar to those in Theorem 5.1):

Theorem 7.2. *Assume that*

- *the arriving workloads σ_n^i are stationary with respect to the 1-step shift of n for each i,*
- *s_n are constant.*
- *for each $i = 1, ..., K$, queue i is initially at a unique stationary regime that corresponds to the policy that never takes vacations at that queue.*

Then Results 2.2, 2.3 and 2.4 hold.

We may relax the assumption on the initial distribution, as we did in Lemma 4.2 and Remark 4.2.

Remark 7.2. *Note that we allow in this model for different (dependent or independent) arrival streams (and thus interarrival times) to different queues, unlike the model in Section 5. The restriction in Section 5 to a single sequence T_n that defines the time of arrivals, that occur simultaneously to all queues, was due to the fact that it was the arrival times that triggered the polling (the vacation opportunities). In this section, arrival can be more general. (Note that, even if the arrival is a point process, the interarrival times $\{T_n^i\}$ in queue i do not appear explicitly anymore in the evolution equations, due to the gating.)*

We illustrate the usefulness of the previous result in the following optimal scheduling control problem in an telecommunication (ATM) switch.

7.3 Application to an ATM switch

We consider an $M \times N$ switch with M inputs and N outputs, as depicted in Fig. 1. We assume that there are separate input queues for each output, so we do not have HOL (head-of-line) blocking. Each input is associated with N queues, one for each output. We denote by queue ij the queue for cells arriving to input i and destined for output j. We consider a slotted queueing model where in each time slot at most one cell can be transmitted from each of the M inputs, and at most one cell can be received by each of the N outputs. In ATM (Asynchronous Transfer Mode), indeed packet size are fixed and constant, so it is natural to consider time-slotted models. A scheduling mechanism decides at each time slot, from which inputs and to which output do we send a packet.

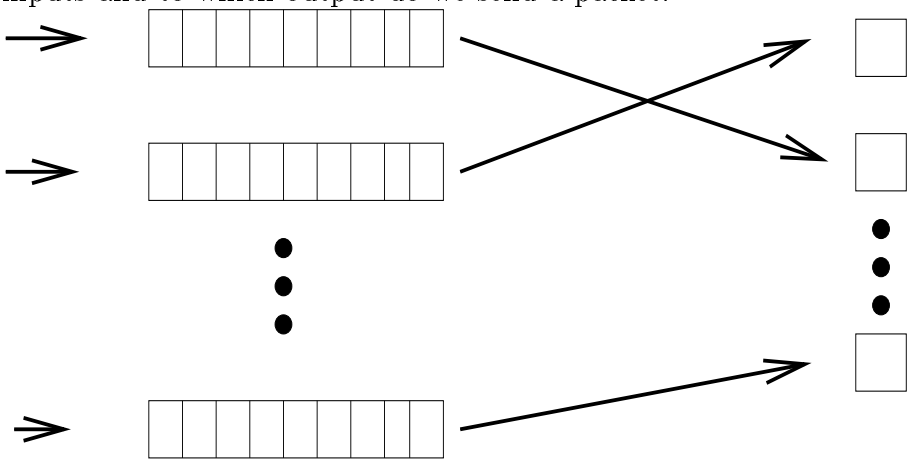

Figure 1: Application to an ATM switch: a feasible schedule

The scheduler may send simultaneously packets from different inputs to different outputs, as long as the following constraints are met:
C1: The scheduler cannot send more than one packet from the same input simultaneously, and
C2: it cannot send more than one packet to the same output, simultaneously.

We are interested here in open-loop scheduling policies, i.e. in scheduling that do not rely on queue length information, but only on the input rates (that will be detailed below).

A class of policies have been presented in [6], that achieve 100% throughput of the switch. A natural problem is whether one can obtain a policy that not only achieves the above goal, but also minimizes increasing convex functions of the workload in the system.

Note that a policy that minimizes the workload, maximizes the amount of workload that departs, and therefore, the throughput. Therefore, if any policy achieves maximum throughput (and stability), then so does the policy which minimizes the workload.

For all $1 \leq i \leq M$, $1 \leq j \leq N$, let $A_{ij}(n)$ be the number of cells that arrive at queue ij in time slot n. We assume that the arrival process $\{A_{ij}\}_n$ is stationary with rate λ_{ij} (i.e. the average number of cells arrived in each time slot). The arrival processes may be mutually dependent.

We consider the symmetric case below, i.e. we assume that the λ_{ij} do not depend on i (they may depend on j), and that $N = M$. For any $i = 1, ..., m$, define $j(i,t) = (i+t)mod(N) + 1$. Consider the round-robin scheduling policy u that sends at time t a packet (if there is any) from queue i to queue $j(i,t)$ for each i. This policy clearly meats the constraints C1 and C2 above. For each j, all queues $ij, i = 1, ..., N$ (having j as destination) receive a round-robin service, which, under the conditions of Theorem 7.2, is optimal among all (open-loop) scheduling policies (in fact, even among those that do not satisfy C1).

A scheduling policy is a sequence $a = (a_1, a_2, ...,)$, where $a_n = \{a_n^{ij}, i, j = 1, ..., N\}$. If the ijth component of a_n is 1, this means that the server polls queue ij at the nth time slot.

Let $h_{ij} = h$ be a convex nondecreasing function. Define for each j,

$$f_n^{ij}(a^{ij}) = Eh(W_n^{ij}(a^{ij})), i, j = 1, ..., N,$$

$$g^j(a) \stackrel{\triangle}{=} \overline{\lim_{s\to\infty}} \frac{1}{s} \sum_{i=1}^{N} \left(\sum_{n=1}^{s} f_n^{ij}(a^{ij}) \right),$$

and further define

$$g(a) \stackrel{\triangle}{=} \overline{\lim_{s\to\infty}} \frac{1}{s} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\sum_{n=1}^{N} f_n^{ij}(a^{ij}) \right).$$

The following Theorem is then a consequence of Theorem 7.2:

Theorem 7.3. *Consider the above $N \times N$ switch. Assume that*

- *the arriving workloads $\sigma_n^{ij} \stackrel{\triangle}{=} A_{ij}(n)$ are stationary with respect to the 1-step shift of n for each i. For each j, the distribution of the processes of arrivals of workload to the queues ij does not depend on i (in other words, $f_n^{ij}(a^{ij}) = f_n^{kj}(a^{kj})$ if $a^{ij} = a^{kj}$),*
- *the time slots s_n are constant,*
- *the workload initially in each queue corresponds to a unique stationary regime that would be obtained if this queue had always been served before.*

Then the round-robin policy u minimizes both $g(a)$ as well as $g^j(a)$, for $j = 1, 2, ..., N$.

Again, we may relax the assumption on the initial distribution, as we did in Lemma 4.2 and Remark 4.2.

References

[1] E. Altman and A. Hordijk, "Applications of Borovkov's renovation theory to non-stationary stochastic recursive sequences and their control", *Advances of Applied Probability* **29**, pp. 388-413, 1997.

[2] E. Altman, B. Gaujal and A. Hordijk, "Multimodularity, Convexity and Optimization Properties", INRIA research report No. 3181, 1997.

[3] E. Altman, B. Gaujal and A. Hordijk, "Admission Control in Stochastic Event Graphs", INRIA research report No. 3179, 1997.

[4] E. Altman, B. Gaujal and A. Hordijk, "Balanced Sequences and Optimal Routing", INRIA research report No. 3180, 1997.

[5] E. Altman, B. Gaujal, A. Hordijk and G. Koole, "Optimal admission and routing control: the case of no buffering" submitted, 1997.

[6] E. Altman, Z. Liu and R. Righter, "Scheduling of an input-queued switch to achieve maximal throughput", Research report 96-020, Leavey School of Business and Administration, Santa Clara, CA 95053, USA. Submitted, 1996.

[7] F. Baccelli and P. Brémaud, *Elements of Queueing theory*, Springer-Verlag, 1994.

[8] A. A. Borovkov, *Asymptotic Methods in Queueing Theory*, John Wiley & Sons, 1984 (translated from Russian).

[9] A. A. Borovkov and S. G. Foss, "Stochastically recursive sequences and their generalizations", *Siberian Advances in Mathematics*, **2**, No. 1, pp. 16-81, 1992.

[10] B. Hajek, "Extremal splitting of point processes", *Mathematics of Operations Research*, **10**, pp. 543-556, 1985.

[11] A. Hordijk and A. Loeve, "Optimal noncyclic server allocation in a polling model", to appear in *Proceedings of the 36th IEEE Conference on Decision and Control*, San-Diego, California, Dec. 1997.

[12] R. Loynes, "The stability of a queue with non-independent inter-arrival and service times", *Proc. Cambr. Phil. Soc.* **58**, No. 3, pp. 497-520, 1962.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - B.P. 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Lorraine : Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - B.P. 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 Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot St Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

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