Monotonicity and stability of periodic polling models

Christine Fricker, M.R. Jaibi

Monotonicity and stability of periodic polling models

Christine Fricker[1], M. Raouf Jaïbi[2]

Abstract.

This paper deals with the stability of periodic polling models with mixed service policies. The interarrivals to all queues are independent and exponentially distributed and the service and the switch-over times are independent with general distributions. The necessary and sufficient condition for the stability of such polling systems is established. The proof is based on the stochastic monotonicity of the state process at the polling instants. The stability of only a subset of the queues is also analyzed, and, in case of heavy traffic, the order of explosion of the queues is given.

[1]INRIA, domaine de Voluceau
B.P. 105, 78153 Le Chesnay Cédex, France
[2]Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Monotonie et stabilité des systèmes de polling

Christine Fricker[1], M. Raouf Jaïbi[2]

Résumé.

Ce papier étudie les propriétés de stabilité des modèles de polling périodiques avec des politiques de service hétérogènes. Les processus d'arrivée aux files d'attente sont de Poisson. Les services et les temps de passage d'une file à l'autre sont des variables aléatoires indépendantes de distribution générale. La condition nécessaire et suffisante de stabilité de ces modèles est établie. La preuve se base sur la propriété de monotonie de ces modèles aux instants d'arrivée du serveur aux files. La stabilité d'un sous ensemble des files d'attente est étudiée et dans le cas d'un modèle avec haut trafic, la vitesse d'explosion de la taille des files d'attente est donnée.

[1]INRIA, domaine de Voluceau
B.P. 105, 78153 Le Chesnay Cédex, France
[2]Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

MONOTONICITY AND STABILITY OF PERIODIC POLLING MODELS

C. Fricker [1] , M.R. Jaïbi [2,‡]

[1] INRIA, Domaine de Voluceau
BP.105, 78153 Le Chesnay Cedex, France

[2] Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Abstract

This paper deals with the stability of periodic polling models with mixed service policies, where the interarrivals to all queues are independent and exponentially distributed, and where the service and the switch-over times are independent with general distributions. The necessary and sufficient condition for the stability of such polling systems is established. The proof is based on the stochastic monotonicity of the state process at the polling instants. The stability of only a subset of the queues is also studied, and, in case of heavy traffic, the order of explosion of the queues is given.

Keywords : polling system, stability, markov chain, stochastic monotonicity, heavy traffic.

1 Introduction

This paper deals with periodic polling systems with mixed service policies and occurrence of switch-over times. In such systems, the server attends to

‡This work was supported in part by a Fellowship of the Netherlands Organization for Scientific Research NWO-ECOZOEK.

the queues according to a polling table in a cyclic way. The queues may be served at different stages in a cycle. Each stage is ruled by a service policy, not necessarily the same for all the stages in a cycle. Particularly, the same queue may be served according to different policies at different stages. We consider general service policies satisfying some properties specified later; these properties are satisfied by the main service policies, like the exhaustive, the gated and the semi-exhaustive policies in their pure and stochastically limited versions, and the time-limited policy without preemption.

The stability condition for such systems is known for a long time (Eisenberg [1972], Kuehn [1979]). Recently, Georgiadis & Szpankowski [1992] addressed the stability of a strictly cyclic polling model served by the l-limited gated policy at all queues. However, no complete proof has been provided up to now, at least for periodic systems with mixed service policies.

The polling system is said to be stable if it admits a stationary regime with integrable cycle time. The necessary and sufficient condition for the stability of the system is established straightforwardly. For the sufficient part, the proof is essentially based on the stochastic monotonicity of the Markov chain representing the state of the system at the polling instants. This property is interesting in itself, and, to our knowledge, has not been noticed up to now. Our main result is the following necessary and sufficient condition for the stability of the polling system:

$$\rho + \max(\lambda_j / G_j^*)\ S < 1$$

where S is the mean of the total switch-over time in a cycle, $\rho = \sum_j \lambda_j \sigma_j$ is the total traffic load of the system, λ_j is the arrival rate to queue j, and G_j^* is the mean of the maximal number of customers that may be served per cycle in queue j ($\lambda_j / G_j^* = 0$ if $G_j^* = \infty$; see Sections 3 and 4 for more details). This analysis allows to give the stability condition for only a subset of the queues, when the whole system is not stable. In particular, in case of heavy traffic, the order in which the queues become unstable is given, providing some insight in the working of the polling system. Our method extends to non deterministic routing of the server between the queues, like the markovian routing for example (Fricker & Jaïbi [1992])

The paper is organized as follows. In the next Section, we describe the model. Section 3 is devoted to a formal definition of service policies and to a classification of them. In Section 4, the crucial stochastic monotonicity of

the Markov chain representing the state of the system at the polling instants is proved and dominant sub-systems are defined. In Section 5, the necessary and sufficient condition for the stability of the system is established and the stability of only a subset of the queues is studied.

2 Model Description

We consider the following model. A polling system with c infinite buffer queues, indexed by $j \in \{1, 2, \ldots, c\}$, is served by a single server. The server attends to the queues in a repeating sequence of a stages, defined by *the polling table*

$$t : \{1, \ldots, a\} \to \{1, \ldots, c\}$$

where $t(i)$ is the queue attended to by the server at stage i (see e.g. Eisenberg [1972]). A *stage* is the period of time during which the server works continuously on a single queue and a *cycle* is the period of time needed to accomplish a consecutive stages. Queue k is attended to by the server a_k times in a cycle, at the successive stages $k_1 < k_2 < \ldots < k_{a_k}$ ($a_k > 0$ for all k and $\sum_{k=1}^{c} a_k = a$). Stage i of cycle n will be referred to by stage (n, i) or stage (n, k_l) when it brings the l-th visit to queue k during cycle n. The succession in time of the stages is expressed by the lexicographic order on (n, i). Stage $(n, i+1)$ for $i = a$ means stage $(n+1, 1)$, and stage $(n, i-1)$ for $i = 1$ means stage $(n-1, a)$.

A service policy is attached to each stage (for all cycles) irrespective of the queue which is served, and not necessarily the same for all stages. It determines the number of customers who are or may be served during the stage, depending on the length of the queue at the polling instant. General service policies which satisfy the properties stated in the next section are allowed.

At completion of stage (n, i), or if queue $t(i)$ is empty upon arrival of the server, the switch of the server to queue $t(i+1)$ (resp. $t(1)$ when $i = a$) requires a switch over time $s_{n,i} \geq 0$. The sequences $(s_{n,i})_n$, $1 \leq i \leq a$, are a independent sequences of i.i.d. random variables, having for each i a general distribution with finite mean $S_i > 0$. The total switch over time in a cycle has mean $S = \sum_{i=1}^{a} S_i$.

Finally, the arrival processes N_k , $1 \leq k \leq c$, to the queues are c independent Poisson processes with intensities $\lambda_k > 0$ respectively; $N_k(u, v]$ denotes the number of arrivals to queue k in the time interval $(u, v]$. The service times required are c independent sequences $(\sigma_k^m)_m$, $1 \leq k \leq c$, of i.i.d random variables, having for each k a general distribution with finite mean σ_k. The traffic load of queue k is $\rho_k = \lambda_k \sigma_k$ and the total traffic load is $\hat{\rho}_c = \sum_{k=1}^{c} \rho_k$. We assume $\rho_k < 1$ for all k to ensure the stability of each queue when it is operating as a classical M/G/1 queue in isolation.

3 Service policies

3.1 Definition and required properties

A service policy determines which customers should be served during a stage. The service policies that we consider are required to satisfy the next four properties:

- **P1** The service policies do not depend on the past history of the service process, for example the number of customers being already served or the time spent serving them.
- **P2** The selection of a customer for service is independent of the required service time and of possible future arrivals.
- **P3** The server serves at constant rate. He leaves immediately a queue which is or becomes empty, but provides service with a positive probability once there is "enough" customers in the queue.
- **P4** The service policies are assumed to be monotonic as defined below in assumption A4.

Property P1 is a lack of memory property. P2 is similar to the lack of anticipation property of Wolf [1982]. The first part of P3 is the work-conservation property of Levy et al. [1990]; the threshold on the number of customers in the second part of P3 is usually one, but other possibilities are not excluded.

For a formal definition, consider a queue with Poisson arrival process N, i.i.d. service times $(\sigma_i)_i$ with mean σ, and traffic load ρ. Suppose a stage begins at time 0 according to a service policy while x customers are waiting in the queue. Call:

$f(x)$ the (random) number of customers that are served during the stage,
$v(x)$ the duration of the stage,
$\varphi(x)$ the number of customers left in the queue at the end of the stage.
The three random functions (f, v, φ) characterize the service policy. Moreover, from property P3, v and φ are related to f for all x by

$$v(x) = \sum_{i=1}^{f(x)} \sigma_i \tag{1}$$

$$\varphi(x) = x - f(x) + N(0, v(x)] \tag{2}$$

Therefore, we associate any service policy with *the random function*

$$f : \Omega \times N \to N$$

induced by the service policy, where $f(., x) = f(x)$ is defined above, and we refer to the service policy as policy f.

Consider now a stage starting at (stopping) time T while Q customers are waiting and D customers have already been served, and call
F the number of customers who are served in the stage,
V the duration of the stage,
Φ the number of customers left in the queue at the end of the stage.

Let $\mathcal{F}_T$ be any (stopped) σ-field containing the history of the service process up to random time T, but which is independent of the process $N(T, T + \cdot]$ of arrivals after T (which starts afresh after T as an independent Poisson process) and of the service times $(\sigma_{D+i})_{i>0}$ of the customers that have not been served up to time T. The four properties can be formally stated as the following assumptions:

• **A1** *(F, V, Φ) is conditionally independent of $\mathcal{F}_T$ given Q, and has the distribution of $(f(Q), v(Q), \varphi(Q))$ where the random functions (f, v, φ) are taken independent of Q.*
• **A2** *(F, V, Φ) is independent of $((\sigma_{D+F+i})_{i>0}\ ,\ N(T + V, T + V + \cdot])$.*
• **A3** *Equations (1) and (2) hold, $f(0) = v(0) = \varphi(0) = 0$ and there exists $x > 0$ such that $E(f(x)) > 0$.*
• **A4** *$(f(x), v(x), \varphi(x))$ is $\leq_d$-monotone in x.*

Let us recall briefly the definition of the $\leq_d$-monotonicity for random vectors, called also stochastic monotonicity or monotonicity in distribution. The (partial) order $\leq$ on $\mathbb{R}^n$ is given by $x \leq y$ if $x_i \leq y_i$ for all i. A real function h defined on $\mathbb{R}^n$ is said to be $\leq$-monotone when $x \leq y$ implies $h(x) \leq h(y)$. The stochastic order for multidimensional distributions and random vectors, denoted by $\leq_d$, is then defined as follows: for two distributions P_1 and P_2 on $\mathbb{R}^n$, $P_1 \leq_d P_2$ if $\int h \, dP_1 \leq \int h \, dP_2$ for all $\leq$-monotone functions h for which the integrals are well-defined. For random vectors, $X_1 \leq_d X_2$ if their distributions satisfy $P_1 \leq_d P_2$; a sequence $(X_n)_n$ is $\leq_d$-monotone if $X_n \leq_d X_{n+1}$ for all n. For more details, we refer to Stoyan [1983]. For the other assumptions, A1 reflects P1 by the fact that apart from Q, (F, V, Φ) is function only of the arrival process $N(T, T + \cdot]$ after T and of the service times $(\sigma_{D+i})_{i>0}$ of the customers that have not been served up to time T; but these are independent of $\mathcal{F}_T$, and in particular of Q. A2 and A3 reflect P2 and P3 by similar arguments.

Remarks. It follows from A2 that the distribution of (F, V, Φ) is insensitive to the order of service of the customers. Hence the numbering of the customers in (1) do not need to be in the order of arrivals. For deterministic policies, when $f(x)$ is a fixed integer for all x, a service policy satisfies A4 if and only if it is monotonic and contractive in the sense of Levy et al. [1990]. For stochastic policies, the stochastic monotonicity of $f(x)$ alone implies that of $(f(x), v(x))$ but not of $\varphi(x)$.

V and Φ are related to F by

$$V = \sum_{i=D+1}^{D+F} \sigma_i \tag{3}$$

$$\Phi = Q - F + N(T, T + V] \tag{4}$$

A2 allows the use of Wald's identity to calculate the expectations of V and of $N(T, T + V] = \sum_{i=1}^{F} N(\sum_{l=1}^{i-1} \sigma_{D+l}, \sum_{l=1}^{i} \sigma_{D+l}]$ in (3) and (4), yielding:

$$E(V) = E(F)\sigma \tag{5}$$

$$E(N(T, T + V]) = E(F)\rho \tag{6}$$

Expectations in (5) and (6) may be infinite. Nevertheless, this will not be the case when Q is integrable. Indeed, let ν be the random function (the particular f) induced when the queue is working as a classical M/G/1 queue, or equivalently if the service policy is the pure exhaustive policy: $\nu(x)$ -respectively $\nu(Q)$- is the number of customers that are served during a busy period initiated by x -respectively by Q- customers. The fact that the server serves continuously during a stage, required in P3 and expressed in (1)-(2), implies that $f(x) \leq \nu(x)$. Hence, from A1, it holds that $F \leq_d \nu(Q)$, and in particular

$$E(F) \leq E(\nu(Q)) = E(Q)(1-\rho)^{-1} \tag{7}$$

Therefore, if Q is integrable, so is F and, from (5), V.

3.2 Limited and unlimited types of policies

Let f be a service policy. By the monotonicity assumption, as x goes to ∞, $(f(x), v(x))$ converges in distribution to a (possibly degenerate) random vector (F^*, V^*). When F^* and V^* have proper distributions, these are the least upper bounds in the sense of the $\leq_d$-ordering for the number of customers that may be served during a stage and the duration of a stage ruled by policy f respectively, whatever the state of the queue to be served is. More precisely, F^* would be the number of customers that are served in a stage if there is an infinite number of customers waiting in the queue, and V^* would be the duration of the stage. On the other hand,

$$0 < \lim_{x\to\infty} E(f(x)) = E(F^*) \leq \infty$$

and, from (5),

$$\lim_{x\to\infty} E(v(x)) = E(V^*) = E(F^*)\sigma$$

The integrability or non-integrability of F^*, and accordingly of V^*, will play an important role in the analysis. Therefore we introduce the classification:

Definition 1 *The policy f is said of limited type when F^* is integrable and of unlimited type otherwise. In the first case, F^* and V^* are called the bounds of policy f.*

For the derivation of the necessary and sufficient condition for the stability of system, in Section 5, we need the next technical lemma.

Lemma 1 *Let $(Q_n)_n$ be a sequence of random variables converging in distribution to a (possibly degenerate) integer-valued random variable Q. Let (f, v, φ) be random functions independent of this sequence and such that $(f(x), v(x), \varphi(x))$ is $\leq_d$-monotone in x. Then $(Q_n, f(Q_n), v(Q_n), \varphi(Q_n))_n$ converges in distribution to $(Q, f(Q), v(Q), \varphi(Q))$. If moreover $(Q_n, n \geq 1)$ is $\leq_d$-monotone, then*
i) $lim_{n\to\infty} E(f(Q_n)) = E(f(Q))$.
ii) When $E(F^*) < \infty$, $E(f(Q)) < E(F^*)$ if and only if there exists y such that $P\{Q \leq y\} > 0$ and $E(f(y)) < E(F^*)$.
iii) When $E(F^*) = \infty$, $E(f(Q)) < \infty$ implies that Q has a proper distribution, i.e. that $lim_{x\to\infty} P\{Q \leq x\} = 1$.
4i) When $E(F^*) < \infty$, and if Q has a defective distribution, so is the limiting distribution of $Q_n - f(Q_n)$.

Proof: The first assertion of the lemma is obvious because (f, v, φ) is independent of $(Q_n)_n$ and by the convergence in distribution of Q_n to the integer-valued$Q$. When this convergence is $\leq_d$-monotonic, the convergence of the expectations in i) follows. Moreover, from A1 and for all $x \in I\!N$,

$$\begin{aligned} E(f(Q_n)) &\geq E(f(Q_n) I_{\{Q_n \geq x\}}) \\ &\geq \sum_{k \geq x} E(f(k)) P\{Q_n = k\} \\ &\geq E(f(x)) P\{Q_n \geq x\} \end{aligned}$$

Taking limits in n,

$$\begin{aligned} E(f(Q)) &= \lim_{n\to\infty} E(f(Q_n)) \\ &\geq E(f(x)) P\{Q \geq x\} \end{aligned} \tag{8}$$

Hence

$$E(f(Q)) \geq \lim_{x\to\infty} E(f(x)) P\{Q \geq x\} \tag{9}$$

ii) When $E(F^*) < \infty$ the right hand side of (9) is $E(F^*) \lim_{x\to\infty} P\{Q \geq x\}$, and $E(f(Q)) < E(F^*)$ implies $\lim_{x\to\infty} P\{Q \geq x\} < 1$. Hence, there exists

y_0, the smallest one, such that $P\{Q \geq y_0 + 1\} < 1$ and $P\{Q \geq y_0\} = 1$; y_0 satisfies $P\{Q \leq y_0\} = 1 - P\{Q \geq y_0 + 1\} > 0$ and from (8) $E(f(y_0)) \leq E(f(Q)) < E(F^*)$. Conversely,

$$\begin{aligned} E(f(Q)) &\leq E(f(y))P\{Q \leq y\} + E(F^*)P\{Q > y\} \\ &< E(F^*) \end{aligned}$$

as soon as y having the properties stated in ii) exists.
iii) When $E(f(Q)) < \infty = E(F^*) = \lim_{x\to\infty} E(f(x))$, (9) implies that $\lim_{x\to\infty} P\{Q \geq x\} = 0$ and that Q has a proper distribution.
4i) When F^* is integrable, the smallest integer z such that $P\{F^* < z\} > 0$ is well defined. Because $f(x) \leq_d F^*$ for all $x > 0$, it holds:

$$\begin{aligned} P\{Q_n - f(Q_n) > x\} &= \sum_{y=x+1}^{\infty} P\{f(x) < y - x\}P\{Q_n = y\} \\ &\geq \sum_{y=x+1}^{\infty} P\{F^* < y - x\}P\{Q_n = y\} \\ &\geq P\{F^* < z\}P\{Q_n \geq x + z\} \end{aligned}$$

by simple monotonicity arguments. Taking limits first as $n \to \infty$ then as $x \to \infty$ in the last inequality shows that if the limiting distribution of Q_n is defective, so is that of $Q_n - f(Q_n)$. □

3.3 Main service policies

Here we follow Levy et al. [1990] to review the main service policies appearing in the literature. The associated random functions f and the types in view of definition 1 are indicated. In the following, $\nu(x)$ still denotes the total number of customers served in a busy period initiated by x customers in an M/G/1 queue. L denotes a positive and integer-valued random variable, independent of the arrival process and of the service times. For any (pure) service policy, the L-limited policy refers to the policy of service which is similar to the pure policy except that the server does not serve more than L customers

• **Gated policies**: Only customers that are present at the beginning of the stage are considered for service.

- For the pure gated policy, by which all present customers at the beginning of the stage are served, $f(x) = x$, $F^* \equiv \infty$ and the type is unlimited.
- For the L-limited gated policy, $f(x) = \min(x, L)$ and $F^* \stackrel{d}{=} L$. Examples of limited type are the l-limited gated for deterministic $L \equiv l$ and the Bernouilli-gated for L having a geometric distribution.
- For the Binomial-gated, by which every present customer is served with some probability p, $f(x)$ has a Binomial distribution with parameters (x, p) and the type is unlimited ($F^* \equiv \infty$).

• **Exhaustive policies**: Customers that are present at the beginning of the stage and customers arriving while service is provided are considered for service.
- For the pure exhaustive policy, by which the server continues serving until the queue is emptied, $f(x) = \nu(x)$, $F^* \equiv \infty$ and the type is unlimited.
- For the L-limited exhaustive policy, $f(x) = \min(\nu(x), L)$ and $F^* = L$. For example, the Bernouilli-exhaustive policy is of limited type.
- The Binomial-exhaustive policy, by which every present or arriving customer is served with probability p, has unlimited type ($F^* \equiv \infty$).
- The **time limited policy without preemption** sets a limit in time to the duration of stage, and if this limit is reached, the service of the customer under service (if any) is resumed, but no more customers are served. It is a particular L-limited exhaustive policy. Indeed, let τ be the (random) limit in time and define $L = R(\tau) + 1$, where R is the counting measure of a zero-delayed renewal process having the distribution of the service times as interarrival distribution, and is independent of τ. Then $f(x) = \min(\nu(x), L)$, $F^* = R(\tau) + 1$; the type is limited if τ is integrable and unlimited otherwise.

• **Decrementing policies**: They are similar to the exhaustive policy except that the server continues serving until the number of customers present in the queue vanishes or is reduced by a prespecified (random) number L of customers. These policies are also called semi-exhaustive policies. For these policies, we have $f(x) = \nu(\min(x, L))$, $F^* \stackrel{d}{=} \nu(L)$ and $E(F^*) = E(L)(1 - \rho)^{-1}$.

All these policies are allowed in our model, as shown by the next lemma.

Lemma 2 *All the policies that are quoted above satisfy assumptions A1-A4.*

Proof: We establish the lemma only for the L-exhaustive policy, for which $f(x) = \min(\nu(x), L)$. Because L is independent of the arrival process and of the service times, and by the classical properties of the M/G/1 queue, it is easy to see that A1, A2 and A3 are satisfied. The monotonicity in A4 is easily proved by a coupling argument on the sample paths. Suppose $x+1$ customers are present upon arrival of the server, with service times $\sigma_0, \ldots, \sigma_x$. Let the server start serving customers $1, \ldots, x$ and all arriving customers (the offsprings of the x customers, in the terminology of Fuhrmann and Cooper [1984]) until the limit L is reached or only customer 0 remains in the queue: the total number of customers that are served is thus $f(x)$. Then serve the remaining customer 0 and his offspring until L is reached or the queue is emptied: the total number of customers that are served is now $f(x+1)$. Clearly, $f(x+1) = f(x)$ if $\nu(x) \geq L$ and $f(x+1) \geq f(x)+1$ otherwise. Therefore $f(x+1) \geq f(x)$ and obviously $v(x+1) \geq v(x)$. It remains to compare $\varphi(x)$ and $\varphi(x+1)$: when the first is positive, $f(x) \geq L$ and $\varphi(x+1) = \varphi(x)+1$; otherwise, the first is 0 and the second non-negative. Because the distribution of $(f(x), v(x), \varphi(x))$ is not affected by the order in which customers are served, the lemma is established for the L-exhaustive policy. Similar proofs hold for the other policies. □

4 The mathematical model

Let f_i be the service policy ruling stage i at which queue $t(i)$ is visited, and let F_i^* and V_i^* be the bounds of policy f_i. For further needs, we distinguish the queues that are served by a policy of unlimited type *at least at one* stage from the queues that are served by policies of limited type *at all* stages. This is done by assuming that *the c queues in the system system are numbered such that*[1]*:*

- queues $1, \ldots, b$ are ruled by a policy of unlimited type at least at one stage
- queues $b+1, \ldots, c$ are ruled by policies of limited type at all stages and are numbered such that λ_k / G_k^* is non-decreasing in $k \in \{b+1, \ldots, c\}$,

[1]If necessary, the queues are renumbered in this way and the polling table is adapted accordingly to preserve the route of the server.

where

$$G_k^* := \sum_{l=1}^{a_k} E(F_{k_l}^*) \tag{10}$$

is the mean of the maximal number of customers that may be served during a cycle in queue k.
Note that $G_k^* = \infty$ when $k \leq b$, but is finite otherwise. We do not exclude in the model the case $b = 0$ where all the involved policies have limited type, or the case $b = c$ where every queue is ruled by a policy of unlimited type once at least.

4.1 The embedded Markov chains

We describe the system by the lengths of the queues at the polling instants. At time t, these are represented by the random vector

$$M(t) = (Q_1(t), \ldots, Q_c(t))$$

where $Q_k(t)$ is *the length of queue k*. $D_k(t)$ is *the cumulative number of customers that have been served in queue k up to time t.*

Stage $(1,1)$ starts at time 0. Call $T_{n,i}$ *the polling instant of stage* (n,i). Then, by the periodic strategy of the server,

$$0 = T_{1,1} \leq \cdots \leq T_{n,i} \leq T_{n,i+1} \leq \cdots \leq T_{n,a} \leq T_{n+1,1} \leq \cdots$$

We introduce the following notations at the polling instant $T_{n,i}$: $M_{n,i}$ for $M(T_{n,i})$, $Q_{n,i}$ for the length $Q_{t(i)}(T_{n,i})$ of the queue $t(i)$ to be served and $D_{n,i}$ for $D_{t(i)}(T_{n,i})$. Moreover we call
$F_{n,i}$ *the number of customers served in stage* (n,i),
$V_{n,i}$ *the duration of stage* (n,i),
$\Phi_{n,i}$ *the number of customers left in queue $t(i)$ at the end of stage* (n,i).
Note that

$$(F_{n,i}, V_{n,i}, \Phi_{n,i}) \stackrel{d}{=} (f_i(Q_{n,i}), v_i(Q_{n,i}), \varphi_i(Q_{n,i})) \tag{11}$$

with (f_i, v_i, φ_i) independent of $Q_{n,i}$. For the queue $k = t(i)$ served at stage (n,i), it holds

$$Q_k(T_{n,i+1}) = \Phi_{n,i} + N_k(T_{n,i} + V_{n,i}, T_{n,i} + V_{n,i} + s_{n,i}] \tag{12}$$

More generally, for any stage (n,i) and any queue j:

$$\begin{aligned} T_{n,i+1} &= T_{n,i} + V_{n,i} + s_{n,i} && (13)\\ Q_j(T_{n,i+1}) - Q_j(T_{n,i}) &= N_j(T_{n,i}, T_{n,i} + V_{n,i} + S_{n,i}] - F_{n,i}\,\delta_{j,t(i)} && (14)\\ D_j(T_{n,i+1}) &= D_j(T_{n,i}) + F_{n,i}\,\delta_{j,t(i)} && (15) \end{aligned}$$

where $\delta_{j,l}$ is the Kronecker symbol. Summing up equations (13) and (14) over a whole cycle, we get for all n and all k:

$$\begin{aligned} T_{n+1,1} - T_{n,1} &= \sum_{i=1}^{a} V_{n,i} + s_{n,i} && (16)\\ Q_k(T_{n+1,1}) - Q_k(T_{n,1}) &= N(T_{n,1}, T_{n+1,1}] - \sum_{l=1}^{a_k} F_{n,k_l} && (17) \end{aligned}$$

Similar relations hold when any other stage is taken as reference for the beginning of a cycle. The next proposition and corollary establish the Markovian behavior of the system at the polling instants. The history of the service process up to time $T_{n,i}$ is given by the stopped σ-field $\mathcal{F}_{n,i}$ generated by the arrival processes to all queues up to $T_{n,i}$, by the service times $(\sigma_k^m)_{1\le m\le D(T_{n,i})}$ of all customers that have already been served by time $T_{n,i}$, and by the switch-over times $(s_{m,l})$ for $(m,l) \le (n,i-1)$.

Proposition 1 *The sequence $(M_{n,i})_{n,i}$ is a Markov chain.*

Proof: At the random instant $T_{n,i}$, the server starts serving queue $t(i)$ (if not empty, otherwise he starts switching to queue $t(i+1)$) according to policy f_i while the state of all queues is given by $M_{n,i}$. The arrival processes after $T_{n,i}$ are Poisson and are independent of $\mathcal{F}_{n,i}$; the service times and the switch-over times involved after $T_{n,i}$ are also independent of $\mathcal{F}_{n,i}$. Because these quantities are mutually independent, it follows that given $M_{n,i}$, the evolution of the system after $T_{n,i}$ is independent of $\mathcal{F}_{n,i}$ which ensures the Markov property of the sequence $(M_{n,i})_{n,i}$. □

This Markov chain is in general not homogeneous because its transitions depend on (n,i) through i by the policy f_i and the queue $t(i)$ which is served. But for i fixed, they do not depend on n.

Corollary 2 *For all i fixed in $\{1,\ldots,a\}$, $(M_{n,i})_n$ is an homogeneous, irreducible and aperiodic Markov chain with state space $\mathbb{N}^c$.*

Proof: Let i be fixed. $(M_{n,i})_n$ is a subsequence of the Markov chain $(M_{n,i})_{n,i}$ and is thus also a Markov chain which is homogeneous because i is now fixed. It is irreducible because all states communicate. Indeed, $m = (m_1, \ldots, m_b)$ can be reached in one step from the state $(0, \ldots, 0)$: this is realized when first no arrivals occur to all queues during the whole cycle but the last switch-over time $s_{n,i-1}$ (such a cycle consists of switch-over times), and then the last switch-over time is positive and $(m_1, \ldots, m_b)$ arrivals occur during it, all this having a positive probability (in particular because the arrival processes are Poisson). On the other hand, $(0, \ldots, 0)$ is reached in (possibly) many steps from any state $(m_1, \ldots, m_c)$ with positive probability too: this is realized when there are no arrivals until it happens (the time needed to clear the totality of the work induced by the $(m_1, \ldots, m_c)$ customers present is integrable). By the same arguments, the state $(0, \ldots, 0)$ is aperiodic and so is the (irreducible) Markov chain. □

4.2 Monotonicity of the model

Call π_i *the transition operator of the Markov chain* $(M_{n,i})_{n,i}$ *for each given* i, defined by

$$\pi_i h(m) = E\left(h(M_{n,i+1}) \mid M_{n,i} = m\right)$$

for any $m = (m_1, \ldots, m_c) \in I\!N^c$ and any real function h defined on $I\!N^c$ for which the expectation exists. *The transition operator* $\tilde{\pi}_i$ *of the Markov chain* $(M_{n,i})_n$ is then the product:

$$\tilde{\pi}_i = \pi_{i-1} \cdots \pi_1 \pi_a \cdots \pi_{i+1} \pi_i$$

π_i is derived from equations (11)- (14). Having those in mind and for ease of notation, we express π_i in a few steps for tensor product functions $h = \otimes_{l=1}^{c} h_l$ (this class of functions characterizes completely π_i, extension to general h is immediate). Let $m^{\neq k} = (m_1, \ldots, m_{k-1}, m_{k+1}, \ldots, m_c)$ be the $(c-1)$-tuple obtained by removing the k-th component from m. Define on $I\!R_+ \times I\!N^c \times I\!R_+$ the function H_i by

$$\begin{aligned} H_i(u, r, m^{\neq t(i)}, s) \;=\; & E(\; h_{t(i)}(r + N_{t(i)}(T_{n,i} + u, T_{n,i} + u + s]) \\ & \Pi_{l \neq t(i)} h_l(m_l + N_l(T_{n,i}, T_{n,i} + u + s])\;) \end{aligned} \tag{18}$$

which does not depend on $T_{n,i}$ by the properties of the Poisson process. Because the switch-over time $s_{n,i}$ is independent of $T_{n,i}$ and of the arrivals

processes, the integral in s of the function H_k with respect to the distribution of $s_{n,i}$ (which does not depend on n) is the function K_i given by:

$$K_i(u, r, m^{\neq t(i)}) = E\left(H_i(u, r, m^{\neq t(i)}, s_{n,i})\right) \tag{19}$$

The arrivals to the queue $t(i)$ after $T_{n,i} + V_{n,i}$ and to the other queues after $T_{n,i}$ are independent of $(M_{n,i}, V_{n,i}, \Phi_{n,i})$. Given $M_{n,i} = m$, the conditional distribution of $(V_{n,i}, \Phi_{n,i})$ is the distribution of $(v_i(m_{t(i)}), \varphi_i(m_{t(i)}))$. Hence (11)-(14) lead to the following expression of π_i:

$$\pi_i h(m) = \int_{u,r} K_i(u, r, m^{\neq t(i)})\, dP_{v_i(m_{t(i)}),\varphi_i(m_{t(i)})}(u, r) \tag{20}$$

An operator π is $\leq_d$-monotone if for all distributions $P_1 \leq_d P_2$, $\pi P_1 \leq_d \pi P_2$. This holds whenever πh is $\leq$-monotone for any $\leq$-monotone function $h = \otimes_{l=1}^{c} h_l$ (Stoyan [1983] pages 27 and 63).

Lemma 3 *For all i, π_i and $\tilde{\pi}_i$ are $\leq_d$-monotone.*

Proof: We first prove the assertion for π_1. For ease of notation and without loss of generality, suppose $t(1) = 1$ and put $T = T_{n,i}$. Let $h = \otimes_{l=1}^{c} h_l$ be a $\leq$-monotone function. The random vector

$$(r + N_1(T + u, T + u + s], m_2 + N_2(T, T + u + s], \ldots, m_c + N_c(T, T + u + s])$$

has independent components, and all of them are all $\leq_d$-monotone in $(u, m^{\neq 1}, s)$; the vector is thus $\leq_d$-monotone in $(u, r, m^{\neq 1}, s)$. For any $h = \otimes_{l=1}^{c} h_l$ $\leq$-monotone, H_1 given by (18), expectation of the function h of the vector above, is $\leq$-monotone too. Similarly, K_1, given by (19), is $\leq$-monotone. Finally, from (20), $\pi_i h$ is also $\leq$-monotone: it is in $m^{\neq 1}$ by the monotonicity of K_1 for m_1 fixed, and in m_1 by the monotonicity assumption on the service policies. Hence π_1 is $\leq_d$-monotone. The same proof holds for π_i when $i \neq 1$. The product of $\leq_d$-monotone operators being also $\leq_d$-monotone, the $\leq_d$-monotonicity of $\tilde{\pi}_i$ follows.□

The crucial monotonicity property of state process at the polling instants is a consequence of the previous lemma:

Proposition 3 *Suppose $M_{1,1} = (0, \ldots, 0)$. Then for all i fixed, $M_{n,i}$ is $\leq_d$-monotone in n. In particular, $F_{n,i}$ and $V_{n,i}$ are $\leq_d$-monotone in n.*

Proof: Let $P_{n,i}$ be the distribution of $M_{n,i}$ when the initial distribution $P_{1,1}$ is Dirac at $(0, \ldots, 0)$. Because all components of $M_{2,1}$ are non-negative, $P_{1,1} \leq_d P_{2,1}$ and by lemma 5, for all $i > 0$,

$$P_{1,i} = \pi_{i-1} \cdots \pi_1 P_{1,1} \leq_d \pi_{i-1} \cdots \pi_1 P_{2,1} = P_{2,i}$$

By immediate induction, $P_{n,i} \leq_d P_{n+1,i}$ for all (n, i), and $M_{n,i}$ is $\leq_d$-monotone in n for i fixed. This implies the $\leq_d$-monotonicity of all components of $M_{n,i}$ and in particular of $Q_{n,i}$. The monotonicity of the service policies implies then the $\leq_d$-monotonicity of $(F_{n,i}, V_{n,i})$ in n for i fixed.□

4.3 Dominant sub-systems

When a huge number customers are waiting, the duration of a stage depends strongly on the type of the service policy which is used.

Suppose a queue, served by a policy of unlimited type *at least at one* stage is *saturated*: there is an infinite number of customers waiting at time 0 in the queue. At the beginning of such a stage, the queue is still saturated and the duration of the stage is non-integrable, if not infinite. This will be seen to exclude any stationary behavior of the system.

Consider now a queue, say queue k, which is served by policies of limited type at *all* the stages, say f_{k_l} at stage k_l for $1 \leq l \leq a_k$. Suppose the queue is saturated at time 0. Then, the queue stays saturated for ever because at each stage (n, k_l) the number $F_{n,k_l} \stackrel{d}{=} F^*_{k_l}$ of customers that are served is integrable. But the polling system still "works" because the duration $V_{n,k_l} \stackrel{d}{=} V^*_{k_l}$ of these stages are integrable. The sequences V_{n,k_l} constitute a_k independent sequences of i.i.d. random variables. Moreover, these sequences are independent of all the quantities relative to the other queues in the system, and of all the switch-over times. Therefore, for the service of the other queues, the saturation of queue k is equivalent to the replacement of the stages devoted to it by additional switch-over times having the same properties as the original switch-over times. In this way, a sub-system where the number of queues

is reduced by 1 is obtained. This procedure of saturation may be applied to several queues. Referring to the numbering of the queues, we suppose in the following $b < c$.

Let $\mathcal{S}$ be the initial polling system. For $e \in \{b, b+1, \ldots, c\}$, *call* $\mathcal{S}^e$ *the polling (sub-)system of the queues* $\{1, \ldots, e\}$ resulting from the saturation of the queues $\{e+1, \ldots, c\}$, served according to the same polling table and to the same corresponding service policies as in $\mathcal{S}$. For $\mathcal{S}^e$, when stage i brings a visit to one of the e first queues ($t(i) \leq e$), queue $t(i)$ is served like in $\mathcal{S}$, according to the policy f_i attached to stage i; on the other hand, when $t(i) > e$, no queue of $\mathcal{S}^e$ is served but the server becomes unavailable for a period of time distributed like the bound V_i^* for the (limited type) policy f_i, followed by the switch-over time $s_{.,i}$. To facilitate the comparison of these sub-systems, we keep these periods of unavailability apart of the switch-over times. But for $\mathcal{S}^e$, they are included in the time during which the server is not serving in a cycle, or total "switch-over time", whose mean is now:

$$S^e := S + \sum_{j=e+1}^{c} \sigma_j G_j^* \tag{21}$$

The polling system $\mathcal{S}^e$ is similar to $\mathcal{S}$ and all our previous results apply to it. The state of $\mathcal{S}^e$ is described by the sequence $M_{n,i}^e = (Q_1^e(T_{n,i}^e), \ldots, Q_e^e(T_{n,i}^e))$ at the "polling" instants $T_{n,i}^e$ with $(n,i) \in I\!N^* \times \{1, \ldots, a\}$. In particular, for each i, $(M_{n,i}^e)_n$ is a Markov chain and is $\leq_d$-monotone in n when the initial state is empty (here we mean that queues $1, \ldots, e$ are empty at instant 0). Let us specify the transitions π_i^e of this Markov chain. Let $h^e = \otimes_{l=1}^e h_l$ and $m^e = (m_1, \ldots, m_e) \in I\!N^e$ (we also write m^e for the e first components of $m \in I\!N^c$):
- when $t(i) \leq e$, put $h_l \equiv 1$ for $l > e$ in (18)- (20). Then the functions H_i^e, K_i^e and $\pi_i h^e$ depend on m only through m^e. Repeating the arguments which led from (11)- (14) to (20), the operator π_i defined by (20) is seen to give the transitions of $M_{n,i}^e$ too:

$$\begin{aligned} \pi_i^e h^e(m^e) &:= E(h^e(M_{n,i+1}^e) \mid M_{n,i}^e = m^e) \\ &= \pi_i h^e(m^e) \ \ when \ \ t(i) \leq e \end{aligned}$$

- when $t(i) > e$, stage (n,i) corresponds to a period of unavailability of the

server with duration $V^*_{n,i}$. Then only new arrivals may occur to $\mathcal{S}^e$. Adapting the arguments, we get:

$$\pi_i^e h^e(m^e) = \int_u K_i^e(u, m^e)\, dP_{V_i^*}(u) \quad when \quad t(i) > e$$

It is easy to see that the proof of lemma 3 extends to π_i^e to establish its $\leq_d$-monotonicity.

The systems $\mathcal{S}^e$ satisfy a dominance property as we shall demonstrate in the next lemma. By $M^{g|e}$, we denote the e first components of a vector M having $g > e$ components.

Lemma 4 *For all $e < g$ both in $\{b, \ldots, c\}$, $\mathcal{S}^e$ dominates $\mathcal{S}^g$ in the sense that if $M_{1,1}^{g|e} \leq_d M_{1,1}^e$ then $M_{n,i}^{g|e} \leq_d M_{n,i}^e$ for all (n, i).*

Proof: It is enough to compare $\mathcal{S}$ and $\mathcal{S}^e$. Since $\mathcal{S}^e$ may be considered as a sub-system of $\mathcal{S}^{e+1}$, the assertion of the lemma follows by transitivity. Because the sequence $M_{n,i}^{c|e}$ of the e first components of $M_{n,i}$ is not a Markov chain, we need some calculations. We proceed by induction. Let first $i = 1$ and suppose $M_{1,1}^{c|e} \leq_d M_{1,1}^e$. Let $h^e = \otimes_{l=1}^e h_l$ be $\leq$-monotone; $\pi_i h^e$ and $\pi_i^e h^e$ are then $\leq$-monotone. Suppose now $M_{n,1}^{c|e} \leq_d M_{n,1}^e$.
- When $t(1) \leq e$, say $t(1) = 1$, $\pi_1 h^e(m)$ given by (14) depends only on m^e and coincide with $\pi_1^e h^e$. Thus,

$$\begin{aligned}
E(h^e(M_{n,2}^{c|e})) &= \sum_m P\{M_{n,1} = m\}\ \pi_1 h^e(m) \\
&= \sum_m P\{M_{n,1} = m\}\ \pi_1 h^e(m^e) \\
&= \sum_{m^e} \pi_1 h^e(m^e) \sum_{m_{e+1},\ldots,m_c} P\{M_{n,1} = m\} \\
&= \sum_{m^e} \pi_1 h^e(m^e)\ P\{M_{n,1}^{c|e} = m^e\} \\
&\leq \sum_{m^e} \pi_1 h^e(m^e)\ P\{M_{n,1}^e = m^e\} \\
&= E(h^e(M_{n,2}^e))
\end{aligned}$$

the inequality resulting from the $\leq$-monotonicity of $\pi_i h^e$ and from that $M_{n,1}^{c|e} \leq_d M_{n,1}^e$.

- When $t(1) > e$, say $t(1) = c$, $\pi_1 h^e(m)$ depends only on (m^e, m_c). Remember that $Q_{n,1} := Q_{t(1)}(T_{n,1}) = Q_c(T_{n,1})$ here. Then,

$$\begin{aligned} E(h^e(M_{n,2}^{c|e})) &= \sum_m P\{M_{n,1} = m\}\ \pi_1 h^e(m) \\ &= \sum_m P\{M_{n,1} = m\}\ \pi_1 h^e(m^e, m_c) \\ &= \sum_{m^e, m_c} \pi_1 h^e(m^e, m_c) \sum_{m_{e+1}, \ldots, m_{c-1}} P\{M_{n,1} = m\} \\ &= \sum_{m^e, m_c} \pi_1 h^e(m^e, m_c)\ P\{M_{n,1}^{c|e} = m^e, Q_{n,1} = m_c\} \end{aligned}$$

But K_1^e, defined by (13) for h^e, does not depend on r and is increasing in u; thus

$$\begin{aligned} \pi_1 h^e(m^e, m_c) &= \int_{u,r} K_1^e(u, m^e)\ dP_{v_1(m_c), \varphi_1(m_c)}(u, r) \\ &= \int_u K_1^e(u, m^e)\ dP_{v_1(m_c)}(u) \\ &\leq \int_u K_1^e(u, m^e)\ dP_{V_1^*}(u) \\ &= \pi_1^e h^e(m^e) \end{aligned}$$

because $v_1(m_c) \leq_d V_1^*$. The last term does not depend on m_c any more. Therefore,

$$\begin{aligned} E(h^e(M_{n,2}^{c|e})) &\leq \sum_{m^e, m_c} P\{M_{n,1}^{c|e} = m^e, Q_{n,1} = m_c\} \int_u K_1(u, m^e)\ dP_{V_1^*}(u) \\ &= \sum_{m^e} P\{M_{n,1}^{c|e} = m^e\} \pi_1^e h^e(m_c) \\ &\leq \sum_{m^e} P\{M_{n,1}^e = m^e\} \pi_1^e h^e(m_c) \\ &= E(h^e(M_{n,2}^e)) \end{aligned}$$

This completes the proof of the fact that $M_{n,2}^{c|e} \leq_d M_{n,2}^e$ as soon as $M_{n,1}^{c|e} \leq_d M_{n,1}^e$. Repeating the proof for $i \neq 1$ and by immediate induction, the assertion of the lemma follows. □

5 Stability of the polling model

The polling model is said to be stable when the lengths of the queues at the polling instants admit proper stationary distributions and when the stationary cycle time has finite expectation. Only when both conditions are satisfied, one can construct a stationary model on a probability space. In particular, the integrability of the stationary cycle time ensures the existence of integrable regeneration points of the system, like for example the polling instants $T_{n,1}$ at which the Markov chain $M_{n,1}$ returns to the empty state (Asmussen [1987]).

5.1 The sufficient condition

We suppose here the system empty at time $T_{1,1} = 0$; because we deal with Markov chains, this is not restrictive because the existence or not of a stationary distribution does not depend on the initial distribution. The assumption $\rho_k < 1$ for all (the queues) k ensures that for all (n, i) and all k, the polling instant $T_{n,i}$, the lengths $Q_k(T_{n,i})$ of the queues, $F_{n,i}$, and $V_{n,i}$ are integrable. Indeed, from (7), the integrability of $Q_{n,i}$ implies the integrability of $F_{n,i}$ and of $V_{n,i}$, which imply the integrability of $T_{n,i+1}$ and of $Q_k(T_{n,i+1})$ for all k, and so on.

Let $G_{n,k}$ be *the mean of the number of customers that are served at queue k during cycle n*:

$$G_{n,k} := \sum_{l=1}^{a_k} E(F_{n,k_l})$$

Taking expectations in (16)-(17) and using (5)-(6), we obtain

$$\begin{aligned} E(T_{n+1,1} - T_{n,1}) &= \sum_{j=1}^{c} \sigma_j G_{n,j} + S \qquad (22) \\ E\left(Q_k(T_{n+1,1}) - Q_k(T_{n,1})\right) &= \lambda_k E(T_{n+1,1} - T_{n,1}) - G_{n,k} \\ &= \lambda_k \left(\sum_{j=1}^{c} \sigma_j G_{n,j} + S \right) - G_{n,k} \qquad (23) \end{aligned}$$

By the stochastic monotonicity at the polling instants, established in proposition 3, all relevant quantities are $\leq_d$-monotone. Their expectations are

then non-decreasing and in particular,

$$E(Q_k(T_{n+1,1}) - Q_k(T_{n,1})) \geq 0 \tag{24}$$

Inserting the last inequality in (23) leads to the system of inequalities:

$$G_{n,k} \leq \lambda_k \left(\sum_{j=1}^{c} \sigma_j G_{n,j} + S \right) \quad , \quad 1 \leq k \leq c \tag{25}$$

$G_{n,k}$ is also non-decreasing in n and is bounded by G_k^*, defined in (10) and finite for $k > b$; the limit

$$G_k = \lim_{n \to \infty} G_{n,k}$$

is thus finite for $k > b$, but may be infinite for $k \leq b$. The condition of the next lemma excludes this. Define for $b \leq k \leq c$:

$$\hat{\rho}_k := \sum_{j=1}^{k} \rho_j$$

Lemma 5 *If $\hat{\rho}_b < 1$, then $G_j < \infty$ for all $k \leq b$ and*

$$\sum_{j=1}^{b} \sigma_j G_j \leq \frac{\hat{\rho}_b}{1 - \hat{\rho}_b} \left(\sum_{j=b+1}^{c} \sigma_j G_j + S \right) \tag{26}$$

Proof: Multiplying (25) by σ_k and summing up,

$$\begin{aligned} \sum_{k=1}^{b} \sigma_k G_{n,k} &\leq \hat{\rho}_b \left(\sum_{j=1}^{c} \sigma_j G_{n,j} + S \right) \\ (1 - \hat{\rho}_b) \sum_{k=1}^{b} \sigma_k G_{n,k} &\leq \hat{\rho}_b \left(\sum_{j=b+1}^{c} \sigma_j G_{n,j} + S \right) \end{aligned} \tag{27}$$

The expression on the right hand side is positive and bounded by

$$\hat{\rho}_b (\sum_{j=b+1}^{c} \sigma_j G_j^* + S) < \infty$$

Therefore

$$\lim_{n \to \infty} \sum_{k=1}^{b} \sigma_k G_{n,k} = \sum_{k=1}^{b} \sigma_k G_k = \infty$$

implies $1 - \hat{\rho}_b \leq 0$. Thus $\hat{\rho}_b < 1$ implies the finiteness of $\sum_{k=1}^{b} \sigma_k G_k$ and consequently of G_k for $1 \leq k \leq b$. Moreover, (26) follows by taking limits as $n \to \infty$ in (27). $\square$

Suppose now that $\hat{\rho}_b < 1$. Then, by the previous lemma, G_k is finite for all k. Taking limits in (25), we get for all $1 \leq k \leq c$

$$G_k \leq \lambda_k \left(\sum_{j=1}^{c} \sigma_j G_j + S \right) \tag{28}$$

Inserting (26) in (28) leads, after straightforward calculations, to the system of inequalities:

$$(1 - \hat{\rho}_b) G_k \leq \lambda_k \left(\sum_{j=b+1}^{c} \sigma_j G_j + S \right) \quad , \quad b < k \leq c \tag{29}$$

By a triangularisation procedure, (29) implies the system of inequalities

$$(1 - \hat{\rho}_k) G_k \leq \lambda_k \left(\sum_{j=k+1}^{c} \sigma_j G_j + S \right) \quad , \quad b < k \leq c \tag{30}$$

as shown in the appendix.

Up to now in this sub-section, we have only considered the initial polling system $\mathcal{S}$. Similar definitions and (in-)equalities hold for the (dominant) sub-systems $\mathcal{S}^e$, and are obtained by adding to all involved quantities the superscript e or $*$ according to $k \leq e$ or $k > e$, respectively. In particular, lemma 4 holds: $\hat{\rho}_b < 1$ implies $G_k^e < \infty$ for all $k \leq b$. The subsequent inequalities relative to $\mathcal{S}^e$ involve the e variables G_k^e for $k \leq e$ while for $j > e$, $G_j = G_j^*$ is fixed. For $\mathcal{S}^e$, the system of inequalities (30) reads:

$$(1 - \hat{\rho}_k) G_k^e \quad \leq \quad \lambda_k \left(\sum_{j=k+1}^{e} \sigma_j G_j^e + S^e \right) \quad , \quad b < k \leq e \tag{31}$$

where S^e has been defined in (21).

For $b \leq e \leq c$, let $\mathcal{C}^e$ be the condition

$$\mathcal{C}^e : \qquad \hat{\rho}_e + \frac{\lambda_e}{G_e^*} S^e < 1 \tag{32}$$

It is easy to seen that $\mathcal{C}^{e+1}$ implies $\mathcal{C}^e$, because the queues are numbered such that λ_j/G_j^* is non-decreasing in j. With the convention $\lambda_j/G_j^* = 0$ if $G_j^* = \infty$, or equivalently $j \leq b$, $\mathcal{C}^e$ is

$$\hat{\rho}_e + \max_{1 \leq j \leq e} (\lambda_j/G_j^*) S^e < 1$$

In particular, $\mathcal{C}^b$ reduces to $\hat{\rho}_b < 1$, and $\mathcal{C} \equiv \mathcal{C}^c$ is

$$\mathcal{C}: \qquad \hat{\rho}_c + \max_{1 \leq j \leq c} (\lambda_j/G_j^*)\ S < 1 \tag{33}$$

We have the following:

Lemma 6 *Condition $\mathcal{C}^e$ implies that $G_k^e < G_k^*$ for all $b < k \leq e$.*

Proof: The proof is immediate from (31). Indeed, for all $b < k \leq e$, it holds

$$\begin{aligned} G_k^e &\leq (1-\hat{\rho}_k)^{-1}\lambda_k \left(\sum_{j=k+1}^{e} \sigma_j G_j^e + S^e \right) \\ &\leq (1-\hat{\rho}_k)^{-1}\lambda_k \left(\sum_{j=k+1}^{e} \sigma_j G_j^* + S^e \right) \\ &= (1-\hat{\rho}_k)^{-1}\lambda_k S^k \end{aligned}$$

But from $\mathcal{C}^k$ which is implied by $\mathcal{C}^e$,

$$(1-\hat{\rho}_k)^{-1}\lambda_k S^k < G_k^* \qquad \square$$

The next theorem provides the sufficient condition for the stability of the polling system $\mathcal{S}$.

Theorem 4 *If condition $\mathcal{C}$ is satisfied, that is if*

$$\hat{\rho}_c + \max_{1 \leq j \leq c} (\lambda_j/G_j^*) S < 1$$

the polling system $\mathcal{S}$ is stable.

Proof: $\mathcal{C}$ implies that $\hat{\rho}_b < 1$ and that condition $\mathcal{C}^e$ is satisfied for each $b < e \leq c$. The strategy of proof is to show that all the dominant subsystems $\mathcal{S}^e$ are then stable, inductively on e. For all these systems, we suppose the initial state to be the empty state. By the monotonicity, this ensures the existence of limiting distributions for the lengths of the queues at the polling instants of each stage, but these may be degenerate (not proper) distributions.

First, consider the system $\mathcal{S}^b$. Each queue is served according to a policy with unlimited type at least at one stage, say stage r_k for queue k. When $\hat{\rho}_b < 1$, according to lemma 4, $G_k < \infty$ for all $1 \leq k \leq b$ and in particular, $\lim_{n\to\infty} E(F^b_{n,r_k}) < \infty$. By lemma 1-iii), Q^b_{n,r_k} converges in distribution to a proper random variable $Q^b_{r_k}$. This implies that for every queue k and for all k_l, Q^b_{n,k_l} converges in distribution to a proper random variable $Q^b_{k_l}$. Indeed, suppose the opposite, say Q^b_{n,k_1} has a defective limit in distribution; from lemma 1-4i), it follows that $Q^b_{n,k_1} - F^b_{n,k_1}$ has a defective limit too. But, because,

$$Q^b_{n,k_2} \geq Q^b_{n,k_1} - F^b_{n,k_1}$$

the limit of Q^b_{n,k_2} is also be defective; by induction, this happens for all k_l, and in particular to Q^b_{n,r_k} which is a contradiction. On the other hand, between the the polling instants of stage 1 and the first stage k_1 devoted to queue k in a cycle, only arrivals to queue k may occur (if $k_1 = 1$, these stages coincide); thus $Q_k(T_{n,1}) \leq_d Q_k(T_{n,k_1})$ and $Q_k(T_{n,1})$ admits also a proper limiting distribution. Thus all components of $(M^b_{n,1})_n$ have a proper limiting distribution; this implies that the Markov chain $(M^b_{n,1})_n$ is ergodic, and converges to its stationary distribution independently of the initial state. Indeed,

$$\lim_{n\to\infty} P\{M^b_{n,1} \leq m^b\} \geq 1 - \sum_{k=1}^{c} \lim_{n\to\infty} P\{Q^b_k(T^b_{n,1}) \geq m_k\}$$

where the right hand side is positive when all m_k are chosen large enough: this excludes transience or null-recurrence. It follows that the cycle times converge in distribution to the stationary cycle time, which turns out to be integrable because $G^b_k < \infty$ for $1 \leq k \leq b$. Hence $\mathcal{S}^b$ is stable. Moreover, for all $1 \leq i \leq a$, the Markov chain $(M^b_{n,i})_n$ is ergodic: if μ^b_1 is the invariant distribution of $M^b_{n,1}$, $\mu^b_i = \pi^b_{i-1} \cdots \pi^b_1 \mu^b_1$ is a probability and is invariant for $M^b_{n,i}$.

Let us now suppose that $\mathcal{S}^{e-1}$ is stable and impose $\mathcal{C}^e$ $(e > b)$. From lemma 4, $M_{n,i}^{e|e-1} \leq_d M_{n,i}^{e-1}$ for all (n,i). But because $(M_{n,i}^{e-1})_n$ is ergodic and has a proper limiting distribution, so has $(M_{n,i}^{e|e-1})_n$ for all i. On the other hand, for the last component $Q_e^e(T_{n,i}^e)$ of $M_{n,i}^e$, we know from lemma 6 that $G_e^e < G_e^*$. Thus there exists a stage e_l, say $e_l = r$, such that $\lim_{n\to\infty} E(F_{n,r}^e) < E(F_r^*)$. By lemma 1-ii) there exists y such that $\lim_{n\to\infty} P\{Q_e^e(T_{n,r}^e) \leq y\} > 0$. Like previously, it implies the ergodicity of the Markov chain $(M_{n,r}^e)_n$: for $m_1, \ldots, m_{e-1}$ chosen large enough,

$$\lim_{n\to\infty} P\{M_{n,r}^e \leq (m_1, \ldots, m_{e-1}, y)\} > 0$$

Moreover, the cycle time has then an integrable limiting distribution. Thus $\mathcal{S}^e$ is stable and, by induction, the proof is complete. □

Remark. When $b = c$, the previous proof reduces to the first two paragraphs, and the stability is established without having recourse to the dominant subsystems (which are then not defined).

5.2 Necessity of the sufficient condition

To establish the necessity of the condition of theorem 4 for the stability of the system, we need the following technical lemma (Neveu [1983])

Lemma 7 *Let $(Q_n)_n$ be a stationary sequence of non-negative random variables. If $Q_2 - Q_1$ is integrable, then $E(Q_{n+1} - Q_n) = 0$, even when the Q_n's are not integrable.*

Proof: Because $(Q_n)_n$ is stationary, there exists a shift θ on the canonical probability space of the sample paths of the process $(Q_n)_n$ which preserves the probability measure (see Billingsley [1965], p.19). To avoid additional notations, we suppose that our variables are defined on this canonical probability space. By the stationarity, it suffices to prove that $E(Q_2 - Q_1) = 0$. Let $\mathcal{I}$ be the σ-field of the invariant events. Put $Q = Q_1$, and define for any constant δ the integrable random variable $R_\delta = \min(Q, \delta) \circ \theta - \min(Q, \delta)$. Then, by the ergodic theorem,

$$E(R_\delta|\mathcal{I}) \stackrel{a.s.}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} R_\delta \circ \theta^k$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \left(\min(Q,\delta) \circ \theta^{k+1} - \min(Q,\delta) \circ \theta^{k} \right) \\
&= \lim_{n\to\infty} \frac{1}{n} \left(\min(Q,\delta) \circ \theta^{n} - \min(Q,\delta) \right) \\
&= 0
\end{aligned}
$$

Thus $E(R_\delta) = 0$ for all δ. But $|R_\delta| \leq |Q \circ \theta - Q|$ which is integrable. By the Lebesgue convergence theorem, for $c \to \infty$, we obtain

$$
\begin{aligned}
E(Q_2 - Q_1) &= E(Q \circ \theta - Q) \\
&= lim_{\delta\to\infty} E(R_\delta) \\
&= 0 \qquad\qquad \square
\end{aligned}
$$

The next theorem establishes the necessity of condition $\mathcal{C}$ for the stability of the polling system $\mathcal{S}$.

Theorem 5 *Condition $\mathcal{C}$ is necessary for the stability of the polling system $\mathcal{S}$.*

Proof: Suppose the polling system $\mathcal{S}$ is stable. Put for all i the stationary distribution as initial distribution of the Markov chain $(M_{n,i})_n$: these chains are then stationary and all states are positive-recurrent. Hence, $P\{Q_k(T_{n,i}) = 0\} > 0$ for all k and all (n,i). The cycle time being stationary and integrable, $G_k \sum_{l=1}^{a_k} E(F_{n,k_l})$ does not depend on n for all k, is finite for $k \leq b$, and by lemma 1-ii) (if part with $y = 0$), $G_k < G_k^*$ for $k > b$. In particular, $G_c < G_c^*$.

On the other hand, from equations (17), for all $k \in \{1, \ldots, c\}$,

$$
-\sum_{l=1}^{a_k} F_{1,k_l} \leq Q_k(T_{2,1}) - Q_k(T_{1,1}) \leq N_k(T_{1,1}, T_{2,1}]
$$

where both bounds are integrable. Thus, the previous lemma applies and $E(Q_k(T_{n+1,1}) - Q_k(T_{n,1})) = 0$ for all n and all k. Taking expectations in (17) leads now to an equality in (25), and (27) reads:

$$
(1 - \hat{\rho}_b) \sum_{j=1}^{b} \sigma_j G_j = \hat{\rho}_b \left(\sum_{j=b+1}^{c} \sigma_j G_j + S \right)
$$

It implies that $\hat{\rho}_b < 1$ because the right hand side is positive. Moreover, all inequalities in (26) and (28)-(30) become equalities; in particular (30) becomes:

$$(1-\hat{\rho}_k)G_k = \lambda_k \left(\sum_{j=k+1}^{c} \sigma_j G_j + S \right) \quad , \quad b < k \leq c \tag{34}$$

For $k = c$, it provides

$$G_c = (1-\hat{\rho}_c)^{-1}\lambda_c S < \ G_c^*$$

which is condition $\mathcal{C}$. □
The remark following the proof of theorem 4 holds here too.

The system of equations (34) is easy to solve; its determinant is $1-\rho$ and it has as unique solution

$$G_k = \frac{\lambda_k S}{1-\rho} \quad for \quad 1 \leq k \leq c \tag{35}$$

It is *the mean number of customers served per cycle in queue* k in stationary regime, already known by balance arguments.

5.3 Local stability condition

By local stability, we understand the stability of only a subset of the queues. It is clear from the two previous sub-sections that condition $\mathcal{C}^e$ is the necessary and sufficient condition for the stability of the (sub-)polling system $\mathcal{S}^e$ in that $\mathcal{S}^e$ is completely similar to $\mathcal{S}$. The point here is to suppose that the polling system $\mathcal{S}$ is not stable, that is condition $\mathcal{C}$ is violated, and to determine which of the queues are not stable. Nevertheless, it is clear from the previous analysis that all the queues $\{1, \ldots, b\}$ are simultaneously stable or unstable, according to $\hat{\rho}_b < 1$ or $\hat{\rho}_b \geq 1$ respectively, and that in the second case, the mean cycle time converges to ∞, excluding any stable behavior of the system. Thus we suppose $\hat{\rho}_b < 1$ to ensure the integrability of the cycle times, the queues $\{b+1, \ldots, c\}$ contributing for integrable stage durations anyway.

Hence, we focus on the behavior of the queues $\{b+1,\ldots,c\}$ in an instable polling system $\mathcal{S}$ starting with an empty system at time 0. By the monotonicity property, all lengths $(Q_k(T_{n,i}))_n$ of these queues converge in distribution as $n \to \infty$, but the limit may be defective. Define

$$\epsilon := \max\{j : \hat{\rho}_j + \frac{\lambda_j}{G_j^*} S < 1\}$$

The previous set is not empty because $\hat{\rho}_b < 1$, and ϵ is well-defined. Thus condition $\mathcal{C}^\epsilon$ is fulfilled, but for any $k > \epsilon$, $\mathcal{C}^k$ is not. It is easy to see that for any $k > \epsilon$,

$$\begin{aligned} \hat{\rho}_b + \rho_k + \frac{\lambda_k}{G_k^*} \sum_{b+1 \le j \le c, j \ne k} \sigma_j G_j^* &\ge \hat{\rho}_{\epsilon+1} + \frac{\lambda_{\epsilon+1}}{G_{\epsilon+1}^*} \sum_{\epsilon+2 \le j \le c} \sigma_j G_j^* \\ &\ge 1 \end{aligned}$$

By analogy with the sub-systems $\mathcal{S}^e$, the sub-system of the queues $\{1,\ldots,b,k\}$, obtained when all the other queues $j > b$, $j \ne k$, are saturated, is stable if and only if the term on the left hand side above is less than 1. Thus, this sub-system is instable while $\mathcal{S}^b$ is stable. Arguments as in the second part of the proof of theorem 4 show that the length of queue k must go to ∞ in distribution. It can be shown in the same way that *none of the queues $k > \epsilon$ can be stable in any sub-system containing it and queues* $\{1,\ldots,b\}$. On the other hand, $\mathcal{S}^\epsilon$ is stable, $(M_{n,i}^\epsilon)_n$ is ergodic and by lemma 4, $(M_{n,i}^{c|\epsilon})_n$ converges in distribution to a proper distribution for all i. Thus *all the queues $k \le \epsilon$ are stable*. In particular, this shows that condition $\mathcal{C}^\epsilon$ is the right condition to ensure the stability of queue ϵ. For example, the additional conditions given by Boxma & Groenendijk [1987] for the individual stability of queues ruled either by the 1-limited policy, or by the semi-exhaustive policy are sufficient but not necessary.

Finally, suppose that $\mathcal{S}$ is stable. Multiply the arrival rates to all queues by a common factor $\alpha \ge 1$ and let α increase. This leads eventually to a heavy traffic situation. From the previous lines, it is easy to see that *the order of explosion of the queues is the decreasing order of λ_k/G_k^** (in case of equality, explosion is simultaneous), that is the order in which the conditions $\mathcal{C}^k$ fail to hold.

Appendix

Here we prove that for all e, $b < e \leq c$, the system of the e first inequalities of (29) imply the system of the e first inequalities of (30). The two systems have the same first inequality: $(1-\hat{\rho}_{b+1})G_{b+1} \leq \lambda_{b+1}\left(\sum_{j=b+2}^{c}\sigma_j G_j + S\right)$
The second inequality of (29)is
$(1-\hat{\rho}_b-\rho_{b+2})G_{b+2} \leq \lambda_{b+2}\left(\sigma_{b+1}G_{b+1}+\sum_{j=b+3}^{c}\sigma_j G_j + S\right)$
Inserting the first inequality, the first two inequalities of (29) imply
$(1-\hat{\rho}_b-\rho_{b+2})G_{b+2} \leq \lambda_{b+2}\left(\frac{\rho_{b+1}}{1-\hat{\rho}_{b+1}}(\sum_{j=b+2}^{c}\sigma_j G_j + S)+\sum_{j=b+3}^{c}\sigma_j G_j + S\right)$
Rearranging, we get
$\left(1-\hat{\rho}_b-\rho_{b+2}\,\frac{1-\hat{\rho}_b}{1-\hat{\rho}_{b+1}})\right)G_{b+2} \leq \lambda_{b+2}\frac{1-\hat{\rho}_b}{1-\hat{\rho}_{b+1}}\left(\sum_{j=b+3}^{c}\sigma_j G_j + S\right)$
But
$\left(1-\hat{\rho}_b-\rho_{b+2}\frac{1-\hat{\rho}_b}{1-\hat{\rho}_{b+1}})\right)(1-\hat{\rho}_{b+1}) = (1-\hat{\rho}_b)(1-\hat{\rho}_{b+2})$
and the second inequality of (30) is obtained. Iterating these algebraic manipulations, it is shown exactly in the same way and by induction that the e first inequalities of (29) imply the e first inequalities of (30). The fact that only the first e variables and only the e first inequalities are involved to show this implication allows to do the same for the systems of inequalities associated with $\mathcal{S}^e$ and to obtain (31). □

Acknowledgment
We are grateful to Hans Blanc and to Philippe Robert for helpful suggestions.

References

Asmussen, S. [1987], *Applied Probabilities and Queues*, John Wiley & Sons.
Billingsley, P. [1965],*Ergodic Theory and Information*, John Wiley & Sons.
Boxma, O.J., Groenendijk, W.P. [1987], "Pseudo-conservation Laws in Cyclic Queues," *J. Appl. Prob.* **24**, 949-964.
Eisenberg, M. [1972], "Queues with Periodic Service and Changeover Time," *Oper. Res.* **20**, 440-451.
Fricker, C., Jaïbi, M.R. [1992], "Stability of Random Polling Models," in preparation.
Fuhrmann,S.W., Cooper, R.B. [1984], "Stochastic Decomposition in the M/G/1 Queue with Generalized Vacations," *Oper. Res.* **33**, 1117-1129.

GEORGIADIS, L. AND SZPANKOWSKI, W. [1992], "Stability of Token Passing Rings," *Report Department of Computer Science, Purdue University, West Lafayette, U.S.A.*

KUEHN, P.J. [1979], "Multiqueue Systems with Nonexhaustive Cyclic Service," *The Bell System Technical Journal* **58**, No.3, 672-698.

LEVY, H., SIDI, M. AND BOXMA, O.J. [1990], "Dominance Relations in Polling Systems," *Queueing Systems* **6**, 155-172.

NEVEU, J. [1983], "Contruction de Files d'Attente Stationnaires," *Lect. Notes on Control and Information Sciences* **60**, 31-41.

STOYAN, D. [1983], *Comparison Methods for Queues and Other Stochastic Models*, John Wiley & Sons.

WOLF, R.W. [1982], "Poisson Arrivals See Time Averages," *Oper. Res.* **30**, 223-231.