1991

O.J. Boxma

Analysis and optimization of polling systems

Department of Operations Research, Statistics, and System Theory Report BS-R9102 January

CWI, nationaal instituut voor onderzoek op het gebied van wiskunde en informatica

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Research (N.W.O.).

Analysis and Optimization of Polling Systems

O.J. Boxma
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands;
Faculty of Economics, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

The basic polling system is a system of multiple queues, attended to by a single server in a cyclic order. Polling systems arise naturally in the modelling of many communication, computer and production networks where several users compete for access to a common resource. Such applications also give rise to several variants of the basic polling system, like periodic polling; here the server visits the queues in a fixed order specified by a polling table in which each queue occurs at least once.

The theory of polling systems is going through a period of feverish activity. The purpose of this paper is to stimulate the discussion on what are the really important problems, both from a theoretical and an applied point of view. It is argued that optimization of polling systems is one of those problems. Part of the paper is devoted to the analysis of a polling optimization problem, viz., the determination of that polling table in a periodic polling model that minimizes a certain weighted sum of the mean waiting times.

1980 Mathematics Subject Classification: 60K25, 68M20.
Key Words & Phrases: cyclic polling, polling table, optimization.
This invited paper will appear in the proceedings of the Workshop on Stochastic Modelling of the 13th International Teletraffic Congress (Copenhagen, June 1991).

1. Introduction

In computer-communication networks one frequently encounters a situation where several users compete for access to a common transmission channel. Conflict-free access protocols successively allow each user access; the transmission right is often granted in a cyclic order, cyclic resource allocation either arising naturally or being considered "fair".

A well-suited model for the performance of such multiple-access protocols is a multiple-queue single-server model in which the server moves from queue to queue in a cyclic order. The latter model is called the basic *polling model.* Generalizations to non-cyclic visit orders of the queues by the server, and to multiple servers, are also usually denoted as polling models. The characteristic feature of polling models is that the server is moving between queues (which possibly requires switchover times), implying that the priority of the queues is *dynamically* (e.g., cyclically) changing. This sharply contrasts with classic *static* priority queueing models, and resembles Synchronous Time Division Multiplexing (STDM) models. The main difference with STDM is that in STDM each queue is attended to by the server for a fixed period of time, *regardless* of whether or not there are customers in that queue to be served.

Recently two surveys on applications of polling models to computer-communication networks have appeared (Grillo [24], Takagi [34]). Outside communications, polling models also arise naturally in many situations where several users compete for access to a common resource; examples abound in computer, production and road traffic networks.

The polling literature has grown explosively over the last few years. In the most recent update [33] of his series of surveys "Queueing analysis of polling models", Takagi lists 455 references, of which 337 have appeared in the last decade and 242 in the last five years. In part, questions raised by exciting new application areas (e.g., token passing networks) have contributed to this rapid growth. Also, interesting new queueing theoretic problems have arisen which stimulated research in various directions. In such a period of feverish activity in a small and specialized research area, it is important to

reflect now and then on what has been accomplished so far, and in which direction the research has been (or should be) moving. The temptation to fill one of the many small gaps in the rapidly growing polling edifice, while contributing to the coherence of the theory, may distract one from the really important problems in polling. The purpose of this paper is to stimulate the discussion on what are those problems - both from a theoretical and an applied point of view. We discuss five polling issues that we feel worthy of special attention:
(A) polling systems with multiple servers;
(B) non-Poisson arrivals;
(C) reservation mechanisms and global policies;
(D) fairness;
(E) optimization of polling systems.

The paper is organized as follows. In Section 2 we reflect on the enormous variety of polling models, and on the complexity of their exact analysis. Only few references are given; detailed references to the various polling variants can be found in [33]. The five important polling issues (A)-(E) mentioned above are discussed in Section 3. In Section 4 we consider a polling optimization problem - with several fairness-related aspects - in some detail. It concerns the determination of the polling table in a non-cyclic ('periodic') polling model that minimizes a certain weighted sum of the mean waiting times. We present a mean waiting time approximation for periodic polling models; this approximation is sufficiently simple to allow explicit minimization of the weighted sum of the mean waiting times w.r.t. occurrence frequencies of the queues in the polling table. We close the paper with a further discussion of fairness, based on that mean waiting time approximation.

2. Polling models and their analysis

Range of models

Assume a single server S visits N queues $Q_1, \ldots, Q_N$ in cyclic order. Usually the buffer size of each queue is assumed unbounded; in some studies unit buffers are considered. Each of the queues has its own arrival and service request processes; these processes are usually assumed to be independent stochastic processes, but there are also examples in which it is more realistic to assume that the arrival processes at the various queues are correlated. The time it takes S to switch from one queue to the next is usually considered to be a stochastic variable, independent of the arrival processes, service times and other switchover times; again, exceptions do occur - like dependence of the switchover time on whether the visited queue is empty or not. Motivated by token ring applications, it is nowadays mostly assumed that an idle server keeps moving along the empty queues; in early polling studies an empty server usually did not move until somewhere a customer arrived.

It is assumed in this paper that when S visits a queue, say Q_i, he serves its customers in order of arrival. The *service policy* at this queue prescribes the further behaviour of S. A plethora of service policies has been considered, ranging from *1-limited* (serve just one customer, if any) via *gated* (serve only the customers present in Q_i upon the arrival of S) to *exhaustive* (continue to serve Q_i until it becomes empty). It seems that there are two major classes of policies: the *exhaustive-type* policies and the *gated-type* policies. In an exhaustive-type policy, all customers present upon the arrival of S at the queue plus all those arriving during his visit are *candidates for service.* In a gated-type policy, only the customers present upon the arrival of S at the queue are candidates for service. Whether a candidate is really served may depend on various deterministic or probabilistic rules. The 1-limited service policy is a special deterministic example of both a gated-type and an exhaustive-type policy. It is generalized by the *probabilistically-limited* policy of Leung [28], in which with probability a_i^j at most j customers are served in a visit of Q_i. This policy has a gated-type and exhaustive-type version. The introduction of the probabilistically-limited policy and the division into exhaustive-type and gated-type policies are useful steps towards a much needed unification and classification of the various service policies.

Instead of limitations on the number of customers to be served, as in Leung [28], one can also pose

time limitations on the visit to a queue (or on the whole cycle). Such *timer-limited* policies have received less attention than their practical interest, e.g. in FDDI, would justify. Coffman et al. [18] present an interesting mathematical analysis of a 2-queue model with exponentially distributed visit time limits.

Next to the large variety of service policies describing the behaviour of the server at a queue, there are also several possibilities for the server movement *between queues.* The basic cyclic polling model has been generalized in the following directions. *(i) Probabilistic polling:* the server visits the queues according to a probabilistic routing mechanism (e.g., according to a Markov routing chain). *(ii) Periodic polling:* the server visits the queues in a fixed order specified by a *polling table* in which each queue occurs at least once. The latter generalization seems particularly interesting from an applied point of view; some examples are provided by the token bus protocol in local area networks, and by star polling at a computer with multidrop terminals (polling table [1,2,1,3,...,1,N]). In many multiple-access protocols, it is possible to implement a scheme that allows some users more frequent right of transmission than other users. This raises the need for rules to build a good polling table. Section 4 is devoted to this topic.

Mathematical analysis

Various combinations of service policies lead to an astounding range of mathematical complexity - even in the case of independent Poisson arrival processes, to which we restrict ourself here. Models with only purely gated and exhaustive policies can be exactly analysed in detail, for an arbitrary number of queues. A 2-queue model with exhaustive service at Q_1 and 1-limited service at Q_2 is also almost trivial to analyse (and with zero switchover times, it reduces to a model with two priority classes and nonpreemptive priority for the first class). But a 2-queue model with gated service at Q_1 and 1-limited service at Q_2 is yet unsolved! Determination of the joint queue length distribution here seems to be a challenging mathematical problem, solution of which may also be of some practical interest; Bisdikian [1] uses a minor modification of this model for representing a station in a DQDB (Distributed Queue Dual Bus) network.

A 2-queue model with 1-limited service at both queues can be completely analysed by using the theory of Riemann-Hilbert boundary value problems of mathematical physics. This holds regardless of whether there are switchover times or not, and of whether the service time distributions are exponential or not. See [4] for a further discussion of the application of this technique to 2-queue polling models. Unfortunately, extension of this theory to higher dimensions has not yet been possible, making it highly unlikely that in the near future a detailed exact analysis will be presented for models with three or more queues with some limited service policy.

The observation that a detailed exact analysis of polling models is only possible in exceptional cases has stimulated research into three directions: (i) (pseudo-)conservation laws, (ii) waiting time approximations, and (iii) numerical algorithms.

(i) (Pseudo-)conservation laws

In single-server multi-queue systems with a work-conserving service discipline at all queues, the total amount of work $\mathbf{V}$ in the whole system equals the amount of work $\mathbf{V}_{FCFS}$ in a system with the same traffic characteristics but with a global FCFS service discipline:

$$\mathbf{V} = \mathbf{V}_{FCFS}. \tag{2.1}$$

Kleinrock (cf. [5,26]) has exploited this principle of work conservation to derive an exact expression for a weighted sum of the mean waiting times $E\mathbf{W}_n$ at the queues of a single-server multi-queue system. Under the conditions of independent Poisson arrival processes, and a non-preemptive service discipline in which the server only uses information about the current state and the past of the queueing process in making scheduling decisions, the following holds:

$$\sum_{n=1}^{N} \rho_n E\mathbf{W}_n = \rho \frac{\sum_{n=1}^{N} \lambda_n \beta_n^{(2)}}{2(1-\rho)}. \tag{2.2}$$

Here λ_n is the arrival rate at Q_n, β_n, $\beta_n^{(2)}$ denote the first and second moments of the service times, and $\rho_n = \lambda_n \beta_n$, $\rho = \sum \rho_n$. Kleinrock has called (2.2) a *conservation law* to indicate that a change in the scheduling discipline (under the above restrictions) does not lead to a change in $\sum \rho_n E\mathbf{W}_n$. The conditions for (2.2) to hold are fulfilled for an ordinary cyclic polling system with Poisson arrivals. However, in a polling system with switchover times of the server between queues, the principle of work conservation is violated in the sense that at a switchover the service process is interrupted although work may still be present at some of the queues. In [5,6] it has been shown that, under rather weak conditions, a simple extension of the work conservation principle holds, viz. a *work decomposition principle:*

$$\mathbf{V} \overset{D}{=} \mathbf{V}_{FCFS} + \mathbf{Y}, \tag{2.3}$$

$\mathbf{V}_{FCFS}$ and $\mathbf{Y}$ being independent. In this relation $\mathbf{V}$ is the steady-state amount of work in the system with switchover times, $\mathbf{Y}$ is the steady-state amount of work in that system at an arbitrary switchover epoch, and $\mathbf{V}_{FCFS}$ is the steady-state amount of work in the corresponding system *without* switchover times (hence a work conserving system). Here $\overset{D}{=}$ denotes equality in distribution, and 'the corresponding system without switchover times' indicates a single-server multi-queue system with exactly the same arrival and service demand process as the system under consideration, but without switchover times.

The work decomposition formula (2.3) is in particular valid for cyclic and even periodic polling systems. Similar to (2.2), formula (2.3) leads to the following exact expression for a weighted sum of the mean waiting times in a polling system with switchover times:

$$\sum_{n=1}^{N} \rho_n E\mathbf{W}_n = \rho \frac{\sum_{n=1}^{N} \lambda_n \beta_n^{(2)}}{2(1-\rho)} + E\mathbf{Y}. \tag{2.4}$$

This has been called a *pseudoconservation law:* a change in the visit order or service policy at a queue generally *does* lead to a change in $E\mathbf{Y}$, and hence in the lefthand side of (2.4). $E\mathbf{Y}$ can generally be determined quite easily, for cyclic polling and even for periodic polling. Only one term sometimes presents difficulties: the mean amount of work in a queue at the departure epoch of the server from that queue. This term only depends on the service policy at that queue. For many service policies it is trivial to determine this term (like for exhaustive service, in which it equals zero); in other cases it is a yet unsolved problem. The cases of either exhaustive or gated service at all queues yield pseudoconservation laws first obtained by Ferguson and Aminetzah [21]. The pseudoconservation law for 1-limited service was first discovered by Watson [35]; the contrast between the simplicity and beauty of its expression and the complexity of its derivation triggered the research that led to the above-described work decomposition result. The reader is referred to [5] for a lengthy discussion of the concepts of work decomposition and pseudoconservation law.

(ii) Waiting time approximations

Pseudoconservation laws have recently been used in several papers to develop approximations for the individual mean waiting times in polling systems. See [12,20,25]; further references and discussions are given in Section 4 of [5]. Pseudoconservation laws are also being used to test approximations and simulations.

(iii) Numerical algorithms

Blanc [2,3] and Leung [28] are developing interesting procedures that enable one in principle to determine polling performance measures with any required accuracy. Blanc uses a power-series algorithm that can be applied to a large class of multi-queue systems with a multi-dimensional birth-and-death process structure. It is based on power-series expansions of the state probabilities and moments of

the joint queue length distributions as functions of the load of the system in light traffic. Leung proposes an iterative numerical solution, based on discrete Fourier Transforms, for cyclic-service systems with a probabilistically-limited service policy. An essential difficulty is that the computational complexity of their procedures is quite large (memory and CPU time being exponential functions of the number of queues) as a result of which a restriction to fairly small and simple models must be made; still, considerable progress should be possible in the near future.

3. Some polling topics for further research

In this section we discuss a number of global polling issues that we consider to be important from either a theoretical or an applied point of view. Of course the view presented here is a biased one, of someone who has little direct involvement with the application area; but ITC 13 seems a suitable forum to open a discussion on what are the really important polling problems lying ahead. To avoid repetition, we have not deemed it necessary to discuss again some topics that were designated in Section 2 as interesting theoretical developments (numerical procedures, pseudoconservation laws). We start with brief remarks on polling systems with multiple servers, on non-Poisson arrival processes and on reservation mechanisms and global policies. Some of the points raised there are influenced by the development of Broadband ISDN. The extremely high speeds in B-ISDN and projected high-speed local area networks ask for simple slotted transmission mechanisms (like in ATM = Asynchronous Transfer Mode). Accordingly, discrete-time polling systems will receive more attention. Furthermore, as there are far more slots in a high-speed local area network than in a low-speed local area network with the same configuration, there will be a need for polling models with not just one server (or a few), but several thousands of servers. This brings us to the first topic on our list.

A. Polling models with multiple servers

So far, hardly any exact results on polling models with multiple servers have been obtained. Such polling models appear not only in the performance analysis of slotted ring systems but also, as observed in Takagi [33, p. 299], in the modelling of token ring networks with multiple channels, multiprocessor computers, multiple machine repairmen and multiple elevators. It would be most interesting to study the concepts of work conservation, work decomposition and (pseudo)conservation laws for multi-server multi-queue systems. In the case of a very large number of servers, an asymptotic waiting time analysis may yield useful results. Browne et al. [16,17] present an interesting discussion of a vacation queue with an infinite number of servers. The servers work in parallel and serve customers in stages. The customers being served in a stage are those present at the beginning of that stage. When no customers are present at the end of a stage, the servers take a vacation. Determination of the steady-state distribution of the number of customers served during a stage requires the solution of a Fredholm integral equation of the second kind. The results have bearing upon a polling model with an infinite number of servers that always operate together.

B. Non-Poisson arrivals

The assumption of Poisson arrivals is not always realistic; e.g., in B-ISDN some traffic sources will present traffic of a very bursty character, and in an ATM environment there is also the possibility of dependent interarrival times since messages at a higher layer usually create several ATM blocks. Extensions of single Poisson arrivals in polling models to compound Poisson arrivals (with correlated batch sizes) have been studied in a number of papers; see Takagi [33, p. 290] for some references. In discrete-time models also some non-Poisson arrival processes have been taken into consideration. The numerical procedures of Blanc [2,3] and Leung [28] seem to offer possibilities for allowing more general arrival processes. It would be interesting to try and extend the concepts of work decomposition and pseudoconservation law to the case of general interarrival time distributions.

C. Reservation mechanisms and global policies

Future metropolitan area networks and high-speed local area networks will provide capacities beyond

one Gigabit per second and geographical coverage of hundreds of kilometers. Such networks require access schemes that are simple, flexible in handling heterogeneous traffic, and sufficiently efficient to provide high throughput and small access delay (Nassehi [31]). Token access schemes satisfy these requirements for present-size local area networks, but for larger networks the token propagation time will seriously affect performance. Some recently proposed access schemes use a reservation mechanism. The Distributed Queue Dual Bus access scheme DQDB uses a reservation approach for slotted unidirectional bus systems [32]. Each user can reserve only one slot. This scheme can achieve an aggregate throughput near the bus capacity. In [31] another reservation access scheme is proposed: Cyclic Reservation Multiple Access (CRMA). It is also based on a slotted unidirectional bus structure, and uses cyclic reservation access to provide throughput efficiency. The cyclic reservation structure is realized by the headend periodically issuing reserve commands.

So far, very few performance studies of these interesting reservation mechanisms have been published. See Potter and Zukerman [32] for an analysis of a discrete shared processor model for DQDB. In [11] an exact analysis is presented of a simple queueing model that catches some essential features of CRMA. In this model, a collector moves along the queues gathering customers in a batch and bringing them to a server who serves arriving batches in FIFO order. We expect reservation mechanisms as used in DQDB and CRMA, with additional refinements including priorities and backpressure cancellation [31], to give rise to interesting queueing theoretic models that are closely related to polling models.

In DQDB and CRMA some performance measures of the stations are dependent on the position of the stations on the bus. In [11] a new service policy for cyclic polling is introduced and analysed which also has a cyclic reservation mechanism, and has the feature of position dependence of the stations. It is called Globally Gated (GG). GG uses a time stamp mechanism for its cyclic reservation: the server moves cyclically along the queues, and uses the instant of cycle beginning as reference point. When he reaches a queue, he serves there all customers present at the cycle beginning. This strategy can be implemented by marking all customers with a time stamp denoting their arrival time. GG resembles the purely gated policy, and leads to a simpler mathematical model in which exact expressions for the waiting time distributions at all queues can be obtained [11].

D. Fairness

Most people consider the 1-limited service policy in cyclic polling models as a fair policy, because at each cycle each queue has the chance of having exactly one customer served. The gated service policy has another fairness property: each customer has to wait only one cycle before being taken into service. The Globally Gated policy mentioned under C has similar properties as the gated policy, and also has yet another fairness property: all customers arriving in the n-th cycle are served before all customers arriving in the $(n+1)$-th cycle. An unfair aspect of Globally Gated is that the position of the queues has an effect on the mean waiting times. This unfairness can be removed by first putting the global gate at Q_1, at the next cycle putting it at Q_2, etc. [M. Zukerman, private communication]. Variants like this form the subject of a future study. See Yechiali [36] for an 'elevator-type' variant of GG that leads to identical mean waiting times at all queues.

As opposed to 1-limited service, exhaustive service is considered to be unfair because one heavy-traffic queue can dominate the system, keeping the server occupied for a very long time. Fairness considerations may lead one to implement the extremely inefficient 1-limited policy instead of the exhaustive policy, the most efficient policy [30] in the sense that the workload at any time t in a system with exhaustive service is at most equal to the workload at t under any other policy. But is exhaustive service really so much less fair than 1-limited service? And how fair are the gated and GG policies? To answer these questions, one needs a criterion for fairness in multiple-queue systems. We are not aware of any generally accepted criterion for fairness, although many designers and researchers consider fairness an important performance issue. One useful criterion might be the percentage with which the largest mean waiting time exceeds the smallest mean waiting time, or equivalently, the ratio FR between the largest and smallest mean waiting times.

Periodic polling, instead of cyclic polling, might be used to reduce the fairness ratio FR. In Section

4 we use rough mean waiting time approximations for a mixture of exhaustive, gated and 1-limited service policies, for (i) estimating *FR* in cyclic polling systems, and (ii) choosing a polling table for which *FR* is below a certain predetermined level.

Finally we'd like to refer to Greenberg and Madras [23] for a quite different discussion of fairness in multiple-queue systems. They state that head-of-the-line processor sharing provides an appealing paradigm for the fair sharing of a server. According to this discipline, the customer at the head of each non-empty queue is served in processor sharing fashion. In [19] a new queueing discipline is introduced, termed fair queueing, which attempts to emulate head-of-the-line processor sharing but never preempts a service. Greenberg and Madras [23] analyse two variants of the fair queueing discipline, and show that these variants indeed strongly resemble head-of-the-line processor sharing.

E. Optimization of polling systems

The ultimate goal of performance modelling and analysis is performance improvement and optimization. The extensive research on cyclic-service systems has been useful for performance evaluation, but it has only partly led to an ability to control the systems under consideration and to affect their design. Modern developments in computer and communication technology enable the use of more sophisticated scheduling disciplines, while the need to control complex networks makes the use of such disciplines imperative. Recently a few studies have appeared which open up possibilities for optimization (see Levy and Sidi [29] for some references); much more research seems needed here.

In Section 4 we shall touch upon some static optimization problems regarding the routing of a single server in a periodic polling system (in which the route of the server is 'once and for all' determined, before operation). We refer to Yechiali [36] for a discussion and references on (semi-)dynamic server routing (during operation); see also [13], [14], and the recent report [15] which considers the optimal semi-dynamic routing of two servers which together move from queue to queue.

Next to server routing one may also regulate the length of the visit period of S at a queue. E.g., one can try to choose a time limit for visit periods in such a way that a certain function is optimized under given constraints. In the case of probabilistic policies, like Leung's [28] probabilistically-limited policy, one can even try to choose the limit probabilities optimally.

It obviously makes sense to combine consideration of service policies and server routing strategies. For example, instead of including a queue several times in the polling table and serving 1-limited, it may be better to visit it only once and provide exhaustive service.

Finally we'd like to remark that the very heterogeneous characteristics of the various traffic types in B-ISDN will give rise to hybrid switching systems that support both circuit switching and packet switching. The existing systems FDDI and DQDB have such a hybrid character; part of the (server) capacity is reserved for isochronous channels (circuit switching, uniframe technique). The assignment of this capacity may lead to interesting optimization problems.

4. Static optimization of polling systems: visit frequencies

Consider the following polling model. A single server, S, serves N infinite-capacity queues (stations) $Q_1, \ldots, Q_N$, switching from queue to queue. Customers arrive at all queues according to independent Poisson processes. The arrival intensity at Q_i is λ_i, $i = 1,...,N$. Customers arriving at Q_i are called class-i customers. The service times of class-i customers are independent, identically distributed stochastic variables. Their distribution $B_i(.)$ has first moment β_i and second moment $\beta_i^{(2)}$, $i = 1,...,N$. The offered traffic load, ρ_i, at Q_i is defined as $\rho_i := \lambda_i \beta_i$, $i = 1,...,N$, and the total offered load, ρ, as $\rho := \sum_{i=1}^{N} \rho_i$. The service policy at each queue is 1-limited. With a visit to Q_i we associate a switch time (swap-in or swap-out time) S_i with mean s_i and second moment $s_i^{(2)}$; the switchover times are mutually independent, and also independent of the arrival and service time processes. It is assumed that all involved stochastic processes (of waiting times, queue lengths, etc.) possess an equilibrium distribution. S visits the queues according to a periodic - not necessarily cyclic - polling scheme. As observed in Section 2, periodic polling systems represent the behaviour of token bus protocols in local

area networks; see, e.g., the Manufacturing Automation Protocol (MAP) which is becoming the standard for communication in automated manufacturing. Suppose we still have the freedom to choose the polling table. Our goal is to determine that polling table that solves the following problem:

$$Min \sum_{i=1}^{N} C_i E\mathbf{X}_i = \sum_{i=1}^{N} C_i \lambda_i E\mathbf{W}_i. \tag{4.1}$$

Here $C_i \geqslant 0$ is a weight factor associated with Q_i, (e.g., the cost associated with waiting one unit of time at Q_i), $E\mathbf{X}_i$ is the mean number of customers in Q_i and $E\mathbf{W}_i$ is the mean waiting time of an arbitrary class-i customer. The choice of C_i gives us freedom to attach, in a sense, priorities to particular queues. This choice may also reflect some *fairness* criterion. For $C_i = 1$, $i = 1,...,N$ the problem amounts to minimizing the average mean waiting time of an arbitrary customer in the system; for $C_i = \beta_i$, $i = 1,...,N$ the problem amounts to minimizing the mean workload in the system, cf. [8,9].

REMARK 4.1
[8] and [9] consider the minimization of $\sum \rho_i E\mathbf{W}_i$ (hence $C_i \equiv \beta_i$) for the case of *exhaustive* and *gated* service policies. This amounts to minimization of the workload in the system; and for those policies, an explicit expression for $\sum \rho_i E\mathbf{W}_i$ is given by the pseudoconservation law [7], which is a great help for solving the minimization problem. Actually, there and here we do not try to find the absolute minimum, reasoning that as long as no restriction on the table size is put, one can generally improve upon some obtained value of the sum by taking a bigger table with slightly different visit frequencies. Instead, in [8,9] we have developed a method for finding 'good' visit frequencies. Subsequently we have determined a table size for which such frequencies (almost) can be realized; finally we have developed a procedure for generating a table with given size and visit frequencies, in which the visits to each particular queue are spread out evenly. For example, in a 3-queue case for which we found the visit frequencies 0.52, 0.32 and 0.16, we approximated the visit frequencies by 1/2, 1/3 and 1/6, and we took the following table of size 6: [1,2,1,3,1,2].

The visit frequencies f_i^{exh} and f_i^{gated}, which we have derived for exhaustive and gated service policies, are [9]:

$$f_i^{exh} = \frac{\sqrt{\rho_i(1-\rho_i)/s_i}}{\sum_{j=1}^{N} \sqrt{\rho_j(1-\rho_j)/s_j}}, \tag{4.2}$$

$$f_i^{gated} = \frac{\sqrt{\rho_i(1+\rho_i)/s_i}}{\sum_{j=1}^{N} \sqrt{\rho_j(1+\rho_j)/s_j}}. \tag{4.3}$$

Numerical tests show that the proposed rules perform extremely well.

In tackling problem (4.1) with not all C_i equal to β_i we no longer can use the pseudoconservation law. We now look for an approximation for $E\mathbf{W}_i$ that is sufficiently simple to allow explicit solution of (4.1), and yet is reasonably accurate. Details and numerical examples will be presented in [10]; here we sketch the approach.

Starting point is a mean waiting time approximation for purely cyclic polling models, that has been developed by Everitt [20] for exhaustive and gated service, by Boxma and Meister [12] for 1-limited service, and by Groenendijk [25] for a mixture of those. For all three policies, the mean waiting time $E\mathbf{W}_i$ is related to the mean residual time $E\mathbf{RC}_i$ of a cycle of the server, starting and ending at Q_i:

$$exhaustive: \quad E\mathbf{W}_i = (1-\rho_i) E\mathbf{RC}_i, \tag{4.4}$$

$$\textit{gated}: \quad E\mathbf{W}_i = (1+\rho_i)E\mathbf{RC}_i, \tag{4.5}$$

$$1-\textit{limited}: \quad E\mathbf{W}_i \approx \frac{1-\rho+\rho_i}{1-\rho-\lambda_i\sum_{j=1}^{N}s_j}E\mathbf{RC}_i. \tag{4.6}$$

Formulas (4.4) and (4.5) are respectively exact when the cycle starts with a departure of S from Q_i, respectively an arrival of S at Q_i. Subsequently it is argued that $E\mathbf{RC}_i$ is approximately the same ($E\mathbf{RC}$) for all queues, after which $E\mathbf{RC}$ is determined by substituting all $E\mathbf{W}_i$ approximations in the pseudoconservation law.

We can do something similar for the case of a polling table. Assume that Q_i occurs n_i times in a table, $i=1,...,N$. In the following we assume that these visits are spread as evenly as possible. Denoting by $E\mathbf{RSC}_i$ the mean residual time of a subcycle for Q_i, a subcycle being the interval between two successive server visits to Q_i, we propose the following approximation for $E\mathbf{W}_i$:

$$\textit{exhaustive}: \quad E\mathbf{W}_i \approx (1-\rho_i)E\mathbf{RSC}_i, \tag{4.7}$$

$$\textit{gated}: \quad E\mathbf{W}_i \approx (1+\rho_i)E\mathbf{RSC}_i, \tag{4.8}$$

$$1-\textit{limited}: \quad E\mathbf{W}_i \approx \frac{1-\rho+\rho_i}{1-\rho-(\lambda_i/n_i)\sum_{j=1}^{N}n_js_j}E\mathbf{RSC}_i. \tag{4.9}$$

The arguments leading to (4.7) and (4.8) are simple extensions of those leading to (4.4) and (4.5), (cf. p. 209 of [5] for a discussion of those arguments). The 1-limited approximation (4.9) is derived as follows. A customer arriving at Q_i has to wait a residual subcycle; when he meets $\mathbf{X}_i$ customers in Q_i upon his arrival, he also has to wait $\mathbf{X}_i$ subcycles $\mathbf{SC}_i^+$ - which differ from ordinary subcycles in the sense that each of these subcycles contains a service at Q_i:

$$E\mathbf{W}_i \approx E\mathbf{RSC}_i + E\mathbf{X}_iE\mathbf{SC}_i^+. \tag{4.10}$$

Using the PASTA property (which implies that $E\mathbf{X}_i$ equals the mean number of customers waiting in Q_i at an arbitrary epoch) and Little's formula, we can write $E\mathbf{X}_i = \lambda_iE\mathbf{W}_i$. Similar to the 1-limited cyclic polling approximation in [12], we use a traffic balance argument to write

$$E\mathbf{SC}_i^+ \approx \beta_i + \sum_{j=1}^{N} n_js_j/n_i + (\rho-\rho_i)E\mathbf{SC}_i^+, \tag{4.11}$$

yielding

$$E\mathbf{SC}_i^+ \approx \frac{\beta_i + \sum_{j=1}^{N} n_js_j \,/\, n_i}{1-\rho+\rho_i}. \tag{4.12}$$

Substitution in (4.10) leads to (4.9). Approximations (4.7), (4.8) and (4.9) can also be used for a polling table with a mixture of exhaustive, gated and 1-limited service policies.

We finally need a bold assumption to get rid of the unknown $E\mathbf{RSC}_i$ in those approximations. Noting that Q_i has n_i subcycles in a cycle, we assume that

$$E\mathbf{RSC}_i = \frac{A}{n_i}E\mathbf{C}, \quad i=1,...,N, \tag{4.13}$$

with A some unknown constant and EC the mean time to complete one cycle of the table; it is well-known that

$$EC = \frac{\sum_{j=1}^{N} n_j s_j}{1-\rho}. \tag{4.14}$$

We expect the accuracy of approximation (4.13) to increase with decreasing randomness of the cycle. In the case of low traffic, rather symmetric queues and 1-limited service (which has a relatively small coefficient of variation of the visit period of a queue), (4.13) can be expected to lead to good results.

Let us now return to the minimization problem (4.1). Substitution of (4.13) and (4.7) into (4.1) leads to the following minimization problem: Determine the numbers of visits n_i, $i=1,...,N$ in a table that minimizes

$$\sum_{i=1}^{N} \frac{C_i \lambda_i (1-\rho_i)}{n_i} \frac{\sum_{j=1}^{N} n_j s_j}{1-\rho}. \tag{4.15}$$

This is a homogeneous unconstrained minimization problem: if $(n_1^*, \cdots, n_N^*)$ yields a minimum, then the same holds for $(Kn_1^*, \cdots, Kn_N^*)$ for any $K \neq 0$. In the gated case, (4.8) leads to the same problem, apart from the fact that $1-\rho_i$ has to be changed into $1+\rho_i$.

The visit frequencies that solve (4.1) for the approximate mean waiting times given by (4.7) and (4.8) in the exhaustive and gated cases are now easily found to be

$$f_i^{exh} = \frac{\sqrt{C_i \lambda_i (1-\rho_i)/s_i}}{\sum_{j=1}^{N} \sqrt{C_j \lambda_j (1-\rho_j)/s_j}}, \tag{4.16}$$

$$f_i^{gated} = \frac{\sqrt{C_i \lambda_i (1+\rho_i)/s_i}}{\sum_{j=1}^{N} \sqrt{C_j \lambda_j (1+\rho_j)/s_j}}. \tag{4.17}$$

For $C_i \equiv \beta_i$, these frequencies reduce to those given by (4.2) and (4.3).

In the 1-limited case, substitution of (4.13) and (4.9) into (4.1) leads to the following minimization problem: Determine the numbers of visits n_i, $i=1,...,N$ in a table that minimizes

$$\sum_{i=1}^{N} \frac{C_i \lambda_i (1-\rho+\rho_i)}{n_i(1-\rho) - \lambda_i \sum_{j=1}^{N} n_j s_j} \frac{\sum_{j=1}^{N} n_j s_j}{1-\rho}. \tag{4.18}$$

A simple calculation shows that the optimal visit numbers, up to a multiplicative constant, are given by

$$n_i \sim \lambda_i + (1-\rho-\sum_{k=1}^{N} \lambda_k s_k) \frac{\sqrt{C_i \lambda_i (1-\rho+\rho_i)/s_i}}{\sum_{j=1}^{N} s_j \sqrt{C_j \lambda_j (1-\rho+\rho_j)/s_j}}. \tag{4.19}$$

Here $\rho+\sum\lambda_k s_k = \sum\lambda_k(\beta_k+s_k)<1$, as otherwise the ergodicity condition of the system would be violated. Scaling of the n_i gives the visit frequencies f_i^{1-lim} that sum up to one.

Formulas (4.16), (4.17) and (4.19) give simple engineering rules for determining visit frequencies in a polling table. Preliminary numerical experiments indicate that this approach is very promising: although the estimates of the mean waiting time ratios in a few cases (in particular with 1-limited service) show quite large errors, the polling tables suggested by (4.16), (4.17) and (4.19) even then lead to values of (4.1) that generally hardly differ from the optimal value. Extensive tests of the rules will be presented in the forthcoming paper [10]. See [22] for a more detailed mean waiting time approximation for a polling table with two types of - completely symmetric - stations.

REMARK 4.2
If we add the constraint $E\mathbf{C} = \sum n_j s_j/(1-\rho) = C^*$ to the unconstrained problem (4.1) for 1-limited service, thus prescribing the mean cycle time of the polling table, then the following minimization problem results:

$$Min \sum_{i=1}^{N} \frac{C_i\lambda_i(1-\rho+\rho_i)}{n_i-\lambda_i C^*} \tag{4.20}$$

$$Sub \frac{\sum_{j=1}^{N} n_j s_j}{1-\rho} = C^*. \tag{4.21}$$

The optimal solution of the homogeneous unconstrained problem can be scaled to satisfy (4.21). This implies that the visit numbers given by (4.19), after a scaling with the factor C^*, also solve the constrained minimization problem (4.20)-(4.21). Advantages of this approach are the simple structure of (4.20) and the close similarity with Kleinrock's linear Capacity Assignment problem (Section 5.7 of [26]), which enables one to translate Kleinrock's [26] solution (5.26) immediately into (4.19). Similar to Kleinrock's solution interpretation, one can interpret (4.19): Station Q_i should be visited at least $\lambda_i C^*$ times during a cycle, as this is the mean number of arrivals during one cycle; the remaining 'capacity' is allocated according to a square root assignment. The latter assignment is very similar to those obtained for exhaustive and gated service in (4.16) and (4.17).

REMARK 4.3
In the light traffic situation $\lambda_i \to 0$, $i = 1,...,N$, the visit frequencies for each service policy reduce to

$$f_i = \frac{\sqrt{C_i\lambda_i/s_i}}{\sum_{j=1}^{N}\sqrt{C_j\lambda_j/s_j}}; \tag{4.22}$$

indeed, note that (4.7)-(4.9) now all reduce to the same approximation. On the other hand, in heavy traffic (in particular when $1-\rho-\sum\lambda_k s_k$ becomes small) the visit frequencies of the 1-limited queues will be more or less linearly related to the arrival rates.

REMARK 4.4
In the derivation of (4.7)-(4.9) the assumption of Poisson arrivals plays a minor role. We conjecture that the visit frequencies given by (4.16), (4.17) and (4.19) give acceptable results even for non-Poisson arrival processes. This conjecture has not yet been investigated numerically. Results of Kruskal [27] for a deterministic arrival (and service) process and exhaustive or gated service give the same visit rules as (4.16) and (4.17), thus lending some support to our conjecture.

Let us finally return to the issue of fairness, and apply the fairness criterion *FR*, the ratio between

the largest and smallest mean waiting times, to a polling model with exhaustive, gated and 1-limited service policies. In the cyclic case we use the approximate mean waiting times for such a model, given by (4.4)-(4.6) with $E\mathbf{RC}_i \equiv E\mathbf{RC}$. Assuming that the mean waiting time is largest for Q_1 and smallest for Q_N, the fairness ratios FR_{exh}, FR_{gated} and FR_{1-L} are respectively given by:

$$FR_{exh} \approx \frac{1-\rho_1}{1-\rho_N}, \quad FR_{gated} \approx \frac{1+\rho_1}{1+\rho_N}, \tag{4.23}$$

$$FR_{1-L} \approx \frac{1-\rho+\rho_1}{1-\rho+\rho_N} \frac{1-\rho-\lambda_N \sum_{j=1}^{N} s_j}{1-\rho-\lambda_1 \sum_{j=1}^{N} s_j}.$$

(In this case we should have $\rho_1 \leqslant \rho_N$ for exhaustive service, and $\rho_1 \geqslant \rho_N$ for gated service.) FR_{gated} will usually be the smallest of these ratios; it depends on the choice of parameters whether FR_{1-L} is smaller than FR_{exh}.

In the case of periodic polling, with the various visits to the same queue as evenly spread as possible, combination of the approximations (4.7)-(4.9) with (4.13) suggests simple rules to keep FR close to one. E.g., if Q_i has exhaustive service and Q_j gated service, then one should take $n_i:n_j \approx (1-\rho_i):(1+\rho_j)$.

ACKNOWLEDGEMENT

The author has strongly benefited from many discussions with Hanoch Levy and Jan Weststrate. Stimulating conversations with Hans Blanc, J.W. Cohen, Hideaki Takagi and Uri Yechiali are also gratefully acknowledged.

REFERENCES

[1] Bisdikian, C. (1989). A queueing model with applications to bridges and the DQDB (IEEE 802.6) MAN. *Report IBM Thomas J. Watson Research Center, RC 15218, Yorktown Heights (NY).*

[2] Blanc, J.P.C. (1990). A numerical approach to cyclic-service queueing models. *Queueing Systems* **6**, 173-188.

[3] Blanc, J.P.C. (1990). The power-series algorithm applied to cyclic polling systems. *Report Tilburg University, FEW 445, Tilburg.*

[4] Boxma, O.J. (1986). Models of two queues: a few new views. In: *Teletraffic Analysis and Computer Performance Evaluation,* eds. O.J. Boxma, J.W. Cohen, H.C. Tijms, North-Holland Publ. Cy., Amsterdam, 75-98.

[5] Boxma, O.J. (1989). Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems* **5**, 185-214.

[6] Boxma, O.J., Groenendijk, W.P. (1987). Pseudo-conservation laws in cyclic-service systems. *J. Appl. Prob.* **24**, 949-964.

[7] Boxma, O.J., Groenendijk, W.P., Weststrate, J.A. (1991). A pseudoconservation law for service systems with a polling table. *IEEE Trans Commun.,* Vol. COM-**39**.

[8] Boxma, O.J., Levy, H., Weststrate, J.A. (1990). Optimization of polling systems. In: *Performance '90,* eds. P.J.B. King, I. Mitrani, R.J. Pooley, North-Holland Publ. Cy., Amsterdam, 349-361.

[9] Boxma, O.J., Levy, H., Weststrate, J.A. (1990). Efficient visit orders for polling systems. *Report BS-R9017, Centre for Mathematics and Computer Science, Amsterdam.*

[10] Boxma, O.J., Levy, H., Weststrate, J.A. (1991). Efficient visit frequencies for polling tables: minimization of waiting cost. *Report Tel-Aviv University.*

[11] Boxma, O.J., Levy, H., Yechiali, U. (1990). Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Report Tel-Aviv University.*

[12] Boxma, O.J., Meister, B. (1986). Waiting-time approximations for cyclic-service systems with switch-over times. *Performance Evaluation Review* **14**, 254-262.

[13] Browne, S., Yechiali, U. (1988). Dynamic scheduling in single server multi-class service systems with unit buffers. *Report Graduate School of Business, Columbia University (NY).*

[14] Browne, S., Yechiali, U. (1989). Dynamic priority rules for cyclic-type queues. *Adv. Appl. Prob.* **21**, 432-450.

[15] Browne, S. (1990). Dynamic priority rules when polling with two servers. *Report Graduate School of Business, Columbia University (NY).*

[16] Browne, S., Coffman, E.G.,Jr., Gilbert, E.N., Wright, P.E. (1990). Gated, exhaustive, parallel service. *Report Graduate School of Business Administration, Columbia University (NY).*

[17] Browne, S., Coffman, E.G.,Jr., Gilbert, E.N., Wright, P.E. (1990). The gated, infinite-server queue: uniform service times. *Report Graduate School of Business Administration, Columbia University (NY).*

[18] Coffman, E.G.,Jr., Fayolle, G., Mitrani, I. (1988). Two queues with alternating service periods. In: *Performance '87,* eds. P.-J. Courtois, G. Latouche, North-Holland Publ. Cy., Amsterdam, 227-239.

[19] Demers, A., Keshav, S., Shenker, S. (1989). Analysis and simulation of a fair queueing algorithm. In: *Proceedings Sigcomm '89 Symposium: Communications Architectures & Protocols.* ACM Press. Published as Computer Communications Review 19 (4).

[20] Everitt, D.E. (1986). Simple approximations for token rings. *IEEE Trans. Commun.,* Vol. COM-**34**, 719-721.

[21] Ferguson, M.J., Aminetzah, Y.J. (1985). Exact results for nonsymmetric token ring systems. *IEEE Trans. Commun.,* Vol. COM-**33**, 223-231.

[22] Giannakouros, N.P., Laloux, A. (1989). On the usefulness of the pseudoconservation law in the performance analysis of service systems with deterministic polling. *Report Telecommunications Laboratory, University of Louvain.*

[23] Greenberg, A.G., Madras, N. (1990). How fair is fair queueing? *Report AT&T Bell Laboratories, Murray Hill (NJ).* To appear in J. ACM.

[24] Grillo, D. (1990). Polling mechanism models in communication systems - some application examples. In: *Stochastic Analysis of Computer and Communication Systems,* ed. H. Takagi, North-Holland Publ. Cy., Amsterdam, 659-698.

[25] Groenendijk, W.P. (1989). Waiting-time approximations for cyclic-service systems with mixed service strategies. In: *Teletraffic Science for New Cost-Effective Systems, Networks and Services. Proc. 12th ITC,* ed. M. Bonatti, North-Holland Publ. Co., Amsterdam, 1434-1441.

[26] Kleinrock, L. (1976). *Queueing Systems, Vol. 2.* Wiley, New York.

[27] Kruskal, J.B. (1969). Work-scheduling algorithms: a nonprobabilistic queuing study (with possible application to No. 1 ESS). *Bell System Techn. J.* **48**, 2963-2974.

[28] Leung, K.K. (1990). Waiting time distributions for cyclic-service systems with probabilistically-limited service. *Report AT&T Bell Laboratories, Holmdel (NJ).*

[29] Levy, H., Sidi, M. (1990). Polling systems: applications, modelling and optimization. *IEEE Trans Commun.,* Vol. COM-**38**.

[30] Levy, H., Sidi, M., Boxma, O.J. (1990). Dominance relations in polling systems. *Queueing Systems* **6**, 155-171.

[31] Nassehi, M.M. (1989). CRMA: an access scheme for high-speed LANs and MANs. *Report IBM Research, Zürich Research Laboratory.*

[32] Potter, P.G., Zukerman, M. (1989). A discrete shared processor model for DQDB. *Report Telecom Australia Research Laboratories, Clayton.* In: ITC Specialist Seminar, Adelaide.

[33] Takagi, H. (1990). Queueing analysis of polling models: an update. In: *Stochastic Analysis of Computer and Communication Systems,* ed. H. Takagi, North-Holland Publ. Cy., Amsterdam, 267-318.

[34] Takagi, H. (1990). Application of polling models to computer networks. *Report IBM Research, Tokyo Research Laboratory, RT 0032, Tokyo.*

[35] Watson, K. S. (1985). Performance evaluation of cyclic service strategies - a survey. In: *Performance '84,* ed. E. Gelenbe, North-Holland Publ. Cy., Amsterdam, 521-533.

[36] Yechiali, U. (1991). Optimal dynamic control of polling systems. *Paper in this volume.*