No. 584

Queueing Analysis of Polling Models: Progress in 1990–1993

Hideaki Takagi

May 11, 1994

A chapter in: *Frontiers in Queueing: Models, Methods and Problems*, edited by J. H. Dshalalow, CRC Press, 1994.

Queueing Analysis of Polling Models: Progress in 1990–1993

Hideaki Takagi
Institute of Socio-Economic Planning, University of Tsukuba
1-1-1 Tennoudai, Tsukuba-shi, Ibaraki 305, Japan

Polling models generally refer to systems of multiple queues served by nondedicated servers with rules that allocate the servers to the queues. Started with studies of basic models, which consist of separate queues with independent Poisson arrivals served by a single server in cyclic order, over 30 years ago, some significant progress in the analysis of these and extended systems have enriched queueing theory as well as contributed to the system performance evaluation techniques in such fields as computers, communications, manufacturing, and transportation. Quite active research efforts are still on with both theoretical and application-oriented motivations, as surveyed in this chapter with a focus on the last few years.

0. Introduction

A basic *polling model* is a system of multiple queues attended by a single server in cyclic order. The term *polling* originates in the *polling* data link control scheme in which the central computer interrogates each terminal on a multidrop communication line to find whether it has data to transmit. The addressed terminal transmits data, and the computer then examines the next terminal. Here, the server represents the computer and a queue corresponds to a terminal. This was an application of a polling model studied in the early 1970s. Situations represented by polling models and their variations appear not only in computers and communications but also in other fields of engineering such as manufacturing (e.g., a patrolling machine repairman, assembly work on a carrousel, and an automated guided vehicle) and transportation systems (e.g., traffic lights at an intersection, mail delivery, and elevators). The ubiquitous application is not surprising because the cyclic allocation of the server (resource) is a natural and simple way for fair arbitration to multiple queues (requesters). Therefore, polling models in various settings have been studied by many researchers since the late 1950s, focusing on the applications to technologies emerging at each period. The reader is referred to a monograph [Takagi 1986] and surveys [Boxma 1991; Campbell 1991; Grillo 1990; Levy and Sidi 1990; Rubin and Baker 1990; Szpankowski 1990; Takagi 1988, 1990a, 1991b] for the developments in the analysis, (an early stage of) optimization, and applications of polling models prior to 1990. Recently, a special issue (Volume 11, No.1–2, 1992, edited by O. J. Boxma and H. Takagi) of *Queueing Systems* is devoted to polling models, and Volume 35 for stochastic modeling of telecommunication systems (1992, edited by P. Nain and K. W. Ross) of *Annals of Operations Research* assigns a section to polling models. Conti et al. [1993a] survey some latest applications of polling models to the performance evaluation of metropolitan area networks.

The aim of this chapter is to highlight some progress in the analysis and optimization of polling models witnessed during the recent few years. This is done in Section 2, which follows

Section 1 for a description of basic polling models and a summary of their analysis results available before 1990. In Section 3, possible research topics in the future are suggested. The references include original papers published in journals and conference proceedings as well as those in the stage of preprints and technical reports, dated during 1990–1994 (up to March), that happen to have come under my notice.

1. Basic Models and Summary of Progress Before 1990

Let us describe basic polling models and summarize the progress in their analysis before 1990. To do so, we introduce some notation common to all models. The number of queues in the system is denoted by N. Queues are indexed by i, $i = 1, 2, \cdots, N$, in the order of server movement. In continuous-time systems, we assume a Poisson process for the arrival of customers at rate λ_i for queue i. The Laplace-Stieltjes transform (LST) of the distribution function (DF), the mean, and the second moment of the service time of a customer at queue i are denoted by $B_i^*(s)$, b_i, $b_i^{(2)}$, respectively. The total load offered to the system is then given by

$$\rho = \sum_{i=1}^{N} \rho_i \quad ; \quad \rho_i := \lambda_i b_i \tag{1}$$

The LST of the DF, the mean, and the variance of the time needed by the server to switch from queue i to queue $i+1$ are denoted by $R_i^*(s)$, r_i, and δ_i^2, respectively. The switchover times are independent of the arrival and service processes. The mean and the variance of the total switchover time are then given by

$$R := \sum_{i=1}^{N} r_i \quad ; \quad \Delta^2 := \sum_{i=1}^{N} \delta_i^2 \tag{2}$$

The mean polling cycle time, generally independent of the queue index i, is denoted by $E[C]$. The LST of the DF and the mean of the waiting time W_i of a customer at queue i are denoted by $W_i^*(s)$ and $E[W_i]$, respectively. A system in which the arrival and service processes in all queues and switchover times are independent of queue index i is called *symmetric*. For a symmetric system, we let $\rho_i = \rho/N = \lambda b$, and omit subscript i from other variables.

In the presentation of this section, we restrict ourselves to basic models and also leave out references for the sake of conciseness. Readers who are interested in the variations of the basic models and the references are referred to Takagi [1986, 1988, 1990a].

1.1. Single-Buffer Systems

A system in which each queue can accommodate at most one customer at a time is called a *single-buffer system*. Those customers that arrive to find the buffer occupied are lost. This system can model a patrolling machine repairman and an interactive transaction processing system in a computer shared by multiple users.

For a symmetric single-buffer system in which the service time is a constant b, closed-form expressions for the mean cycle time $E[C]$ and the mean waiting time $E[W]$ are available:

$$E[C] = R + E[Q]b \tag{3}$$

$$E[W] = (N-1)b - \frac{1}{\lambda} + \frac{NR}{E[Q]} \tag{4}$$

where

$$E[Q] = \frac{N \sum_{n=0}^{N-1} \binom{N-1}{n} \prod_{j=0}^{n} \left[e^{\lambda(R+jb)} - 1 \right]}{1 + \sum_{n=1}^{N} \binom{N}{n} \prod_{j=0}^{n-1} \left[e^{\lambda(R+jb)} - 1 \right]} \tag{5}$$

is the mean number of customers served in a polling cycle.

For an asymmetric single-buffer system, we define the *station time* ω_k for the kth visited queue as a time interval consisting of the switchover time from queue $k-1$ to queue k and the possible service time at queue k, where the visit number k is incremented one by one at every visit to all queues. The LST of the DF for the joint distribution of N successive station times $\omega_{k-N+1}, \cdots, \omega_k$ is defined by

$$\Omega_k^*(s_1, \cdots, s_N) := E\left[\exp\left(-\textstyle\sum_{j=1}^{N} \omega_{k-N+j} s_j\right)\right] \tag{6}$$

which satisfies the equations

$$\begin{aligned} \Omega_k(s_1, \cdots, s_N) &= R_{k-1}^*(s_N + \lambda_k)[1 - B_k^*(s_N)]\Omega_{k-1}^*(0, s_1 + \lambda_k, \cdots, s_{N-1} + \lambda_k) \\ &\quad + R_{k-1}^*(s_N) B_k^*(s_N) \Omega_{k-1}^*(0, s_1, \cdots, s_{N-1}) \qquad k = 1, 2, \cdots \end{aligned} \tag{7}$$

The LST of the DF and the mean of the waiting time at the kth visited queue are then given by

$$W_k^*(s) = \frac{\lambda_k [I_k^*(s) - I_k^*(\lambda_k)]}{(\lambda_k - s)[1 - I_k^*(\lambda_k)]} \quad ; \quad E[W_k] = \frac{E[I_k]}{1 - I_k^*(\lambda_k)} - \frac{1}{\lambda_k} \tag{8}$$

where $I_k^*(s)$ is the LST of the DF of the intervisit time I_k for the kth visited queue, which is given by

$$I_k^*(s) = \Omega_{k-1}^*(0, s, \cdots, s) R_{k-1}^*(s) \tag{9}$$

From (7), we can get a set of $N(2^{N-1} - 1)$ linear equations for the same number of unknowns that are used to obtain $W_k^*(s)$. This number has been reduced to $2^N - 1$ by appropriately augmenting parameters in $\Omega_k^*(\cdot)$.

Taking the limit $N \to \infty$ with ρ and R fixed at finite values, we obtain a *continuous polling model*. In this limit, the server travels around a circle on which customers arrive uniformly. If the service time is constant, we have

$$I^*(s) = e^{-sR} \left(\frac{1-\rho}{1 - \rho e^{-sR}} \right)^{R/b} \quad ; \quad W^*(s) = \frac{1 - I^*(s)}{E[I]s} \tag{10}$$

The following mean values are available for systems with generally distributed service times:

$$E[C] = E[I] = \frac{R}{1-\rho} \quad ; \quad E[W] = \frac{R}{2(1-\rho)} + \frac{\rho b^{(2)}}{2b(1-\rho)} \tag{11}$$

1.2. Infinite-Buffer Systems

In a system in which any number of customers can be accommodated without loss at each queue, we usually deal with four basic disciplines with respect to the rule by which the server leaves the queue. In an *exhaustive service* system, the server continues to serve each queue until it empties. Customers that arrive at the queue being served are also served in the current service period. In a *gated service* system, the server serves only those customers that were found in a queue when it visited the queue. Those that arrive at the queue during its service period are set aside to be served in the next round of polling. In a *k-limited service* system, each queue is served until either it empties or k customers are served, whichever occurs first. (The k-limited service system is subdivided into *exhaustively limited service* system and the *gatedly limited service* system depending on whether the served k messages include those that arrive during the service period.) In a *k-decrementing service* system, each queue is served until either it empties or the queue size decreases to k less than that found at the polling instant.

For a broad class of infinite-buffer systems, including the four basic models mentioned above, the mean cycle time is simply given by

$$E[C] = \frac{R}{1-\rho} \tag{12}$$

For symmetric systems, we have the following results for the mean waiting time:

$$E[W]_{\text{exhaustive}} = \frac{\delta^2}{2r} + \frac{N\lambda b^{(2)} + r(N-\rho)}{2(1-\rho)} \tag{13a}$$

$$E[W]_{\text{gated}} = \frac{\delta^2}{2r} + \frac{N\lambda b^{(2)} + r(N+\rho)}{2(1-\rho)} \tag{13b}$$

$$E[W]_{1-\text{limited}} = \frac{\delta^2}{2r} + \frac{N\lambda b^{(2)} + r(N+\rho) + N\lambda\delta^2}{2(1-\rho-N\lambda r)} \tag{13c}$$

$$E[W]_{1-\text{decrementing}} = \frac{\delta^2}{2r} + \frac{N\lambda b^{(2)}(1-\lambda r) + (r+\lambda\delta^2)(N-\rho)}{2[1-\rho-\lambda r(N-\rho)]} \tag{13d}$$

For asymmetric systems, the exact computation of the mean waiting time at each queue is available for exhaustive and gated service systems. For an exhaustive service system, the mean waiting time $E[W_i]$ at queue i can be expressed in terms of the moments of the intervisit time I_i as

$$E[W_i] = \frac{\lambda_i b_i^{(2)}}{2(1-\rho_i)} + \frac{E[(I_i)^2]}{2E[I_i]} \tag{14}$$

We have

$$E[I_i] = \frac{R(1-\rho_i)}{1-\rho} \quad ; \quad E[(I_i)^2] = (E[I_i])^2 + \delta_{i-1}^2 + \frac{1-\rho_i}{\rho_i}\sum_{\substack{j=1 \\ (j\neq i)}}^{N} r_{ij} \tag{15}$$

where r_{ij} is the covariance of the station times ω_i and ω_j for queues i and j, respectively. For the exhaustive service system, the station time ω_i is defined as the time interval between successive instants at which the server leaves queue $i-1$ and queue i. The set $\{r_{ij}; i,j = 1,2,\cdots,N\}$ is computed by solving the following set of equations:

$$r_{ij} = \frac{\rho_i}{1-\rho_i}\left(\sum_{m=i+1}^{N} r_{jm} + \sum_{m=1}^{j-1} r_{jm} + \sum_{m=j}^{i-1} r_{mj}\right) \qquad j < i \tag{16a}$$

$$r_{ij} = \frac{\rho_i}{1-\rho_i}\left(\sum_{m=i+1}^{j-1} r_{jm} + \sum_{m=j}^{N} r_{mj} + \sum_{m=1}^{i-1} r_{mj}\right) \qquad j > i \tag{16b}$$

$$r_{ii} = \frac{\delta_{i-1}^2}{(1-\rho_i)^2} + \frac{\lambda_i b_i^{(2)} E[I_i]}{(1-\rho_i)^3} + \frac{\rho_i}{1-\rho_i}\sum_{\substack{j=1 \\ (j\neq i)}}^{N} r_{ij} \tag{16c}$$

Note that eq. (14) has a form of *decomposition property* for an M/G/1 with *generalized vacations.* Although there are $O(N^2)$ equations in (16), an alternative set of $O(N)$ equations to yield the mean waiting times is also available. A similar set of equations holds for a gated service system. An exact analysis of a 1-limited or 1-decrementing service system with two queues is possible by reducing the determination of unknowns to a boundary value problem on a complex plane. A number of approximation techniques have been proposed for other limited and decrementing service systems. While the individual mean waiting times may not be available, their load-weighted sum, called the *pseudoconservation law,* has been derived for systems with the four basic disciplines and a mixture thereof. The extensions of the pseudoconservation law in the recent years are surveyed in Section 2.3.

Major variations of the above-mentioned models, which needed certain ingenious treatment, were: (1) systems with zero switchover times, (2) systems with non-cyclic polling order, (3) tandem queues served by a single server, and (4) discrete-time systems. In particular, non-cyclic polling order included *deterministic* as in elevators, *probabilistic* (*random* and *Markovian routing*), and *state-dependent* (e.g., a *greedy server*). Many minor variations were also brought in by incorporating batch Poisson arrivals, modified service disciplines, priorities within each queue, and so on.

2. Progress in 1990–1993

In this section, we overview several topics investigated recently.

2.1. Single-Buffer Systems

In a single-buffer system, each queue can accommodate only one customer at a time. A customer that arrives to find the queue nonempty is blocked and lost. One may add a feature of Bernoulli feedback such that the customer whose service is completed returns to the queue with probability p_i, where $0 \leq p_i < 1$, without losing analytical tractability. If the system is asymmetric and the service times are generally distributed, an exact analysis of this system requires the solution to a set of $O(2^N)$ linear equations. Such an analysis was given by Takine et al. [1989] for a system with nonzero switchover times, and by Takine et al. [1990] and Takine and Hasegawa [1992] for a system with zero switchover times. Bunday et al. [1992] analyze a scanning polling system with constant service and switchover times with Bernoulli feedback; see also Bunday and Sztrik [1992]. Chung et al. [1992] analyze a single-buffer system with Markovian server routing.

An interesting variation of the single-buffer system is a system consisting of many single-buffer queues and an infinite- or finite-buffer queue. This system was used to model a token ring network with a gateway or a bridge that connects the network to an external network [Murata and Takagi 1992; Takine et al. 1990a].

Another interesting variation is brought in by considering two classes of customers (the priority customer and the ordinary customer) that are treated differently according to the length of the preceding cycle time. While the priority customer is always served when the server arrives, the ordinary customer is served only if the preceding cycle time does not exceed the prescribed limit. This system was introduced as a model of a timed-token protocol that handles both synchronous and asynchronous traffic by Takagi [1990b], who analyzed a symmetric system with constant service times. The analysis was extended to an asymmetric system with general service times by Nakamura et al. [1990]. Woodward [1994 (sec. 7.2.2)] treats a symmetric system by the equilibrium point analysis (EPA) technique.

There is some controversy over the characteristics of the mean waiting times at lightly even-loaded queues located between two heavily loaded queues. While Takine et al. [1988, 1990b] and Jung [1991] claim that the mean waiting times at the lightly loaded queues increase in the direction of polling, Mukherjee et al. [1990] present numerical results with opposite behavior, even though they consider the same model.

2.2. Continuous Polling Models

The service time was assumed to be a constant in the previous analysis of a continuous polling model on a circle. Bisdikian and Merakos [1992] study the output process of this model. Coffman and Gilbert [1986], who derived eq. (10), pointed out a difficulty in obtaining the waiting time distribution when the service time has general distribution. Recently, however, Kroese and Schmidt [1992] successfully analyzed a continuous polling model on a circle with general service times by means of point processes and regenerative processes in combination with stochastic integration theory.

Studies of other continuous polling models have also been tried; Most of them are still in preprint form, such as polling on a graph [Altman and Foss 1993; Coffman and Stolyar 1992; Kroese and Schmidt 1993] and noncyclic polling in two and more dimensional space [Altman and Levy 1994; Betsimas and Van Ryzin 1991].

2.3. Extension of Pseudoconservation Laws

A weighted sum of the mean waiting times $\{E[W_i]; 1 \leq i \leq N\}$, called the *pseudoconservation law*, was derived by Boxma and Groenendijk [1987] for a continuous-time system in which each queue has a Poisson arrival process and one of exhaustive, gated, 1-limited, and 1-decrementing service disciplines by unifying the results obtained before by others for similar systems with a single service discipline. It was later extended to systems with compound Poisson processes by Boxma and Groenendijk [1988] and Boxma [1989]. An error in this extension for a queue with decrementing service discipline was corrected by Chiarawongse and Srinivasan [1991] who assumed that all the customers contained in each arriving batch belong to the same class.

Combining these results and correcting the previous errors, we can get the most general result as follows. We assume that the system has a Poisson arrival process at rate λ such that each arrival contains G_i customers for queue $i, 1 \leq i \leq N$, simultaneously. The pseudoconservation law is then given by

$$\sum_{i\in E,G} \rho_i E[W_i] + \sum_{i\in L} \rho_i \left(1 - \frac{\lambda_i R}{1-\rho}\right) E[W_i] + \sum_{i\in D} \rho_i \left[1 - \frac{\lambda_i(1-\rho_i)R}{1-\rho}\right] E[W_i]$$

$$= \frac{\lambda \sum_{i=1}^{N} \left[\rho g_i b_i^{(2)} + g_i^{(2)} b_i^2\right] + 2\lambda \sum_{i=2}^{N} b_i \sum_{j=1}^{i-1} g_{ij} b_j}{2(1-\rho)}$$

$$+\rho\frac{\Delta^2}{2R} + \frac{R\left(\rho - \sum_{i=1}^{N}\rho_i^2\right)}{2(1-\rho)} + \frac{R\sum_{i\in G,L}\rho_i^2}{1-\rho} + \frac{\lambda R \sum_{i\in L} g_i^{(2)} b_i}{2(1-\rho)} \tag{17}$$

$$+\frac{\lambda R \sum_{i\in D}\left[(1-2\rho_i) g_i^{(2)} b_i - (\lambda g_i)^2 g_i b_i b_i^{(2)}\right]}{2(1-\rho)}$$

where $g_i = E[G_i]$, $g_{ij} = E[G_iG_j]$ for $i \neq j$, $g_i^{(2)} = E[G_i(G_i - 1)]$, $\rho_i = \lambda g_i b_i$, and E, G, L, and D stand for the index sets of queues with exhaustive, gated, limited, and decrementing service disciplines, respectively. Some special cases are as follows. If each arrival contains only customers for a single queue, we have $g_{ij} = 0$ for $i \neq j$, $1 \leq i, j \leq N$. Furthermore, if each arrival contains a single customer, we have $g_i = 1$ and $g_i^{(2)} = 0$ for $1 \leq i \leq N$. If the numbers of customers contained in each arrival are independent for different queues, we have $g_{ij} = g_ig_j$ for $i \neq j$, $1 \leq i, j \leq N$. If all the switchover times are zero, eq. (17) reduces to the *conservation law* for multiclass $M^X/G/1$ systems given in Takagi [1991c (sec. 3.5)].

For discrete-time systems, the pseudoconservation law first derived by Boxma and Groenendijk [1988] must be corrected in two ways. First, Bisdikian [1993] points out an error in the terms not related to switchover times. Second, the term for queues with decrementing service discipline is in error as shown by Chiarawongse and Srinivasan [1991]. Consequently, if the numbers Λ_i of customers that arrive at queue i in each slot are independent slot by slot, the corrected pseudoconservation law is given by

$$\sum_{i\in E,G} \rho_i E[W_i] + \sum_{i\in L} \rho_i \left(1 - \frac{\lambda_i R}{1-\rho}\right) E[W_i] + \sum_{i\in D} \rho_i \left[1 - \frac{\lambda_i(1-\rho_i)R}{1-\rho}\right] E[W_i]$$

$$= \frac{\rho \sum_{i=1}^{N} \lambda_i b_i^{(2)} + \sum_{i=1}^{N} b_i \sum_{j=1}^{N} \lambda_{ij} b_j - \rho^2}{2(1-\rho)}$$

$$+\rho\frac{\Delta^2}{2R} + \frac{R\left(\rho - \sum_{i=1}^{N}\rho_i^2\right)}{2(1-\rho)} + \frac{R\sum_{i\in G,L}\rho_i^2}{1-\rho} + \frac{R \sum_{i\in L} \lambda_i^{(2)} b_i}{2(1-\rho)} \tag{18}$$

$$+\frac{R \sum_{i\in D}\left[(1-2\rho_i) \lambda_i^{(2)} b_i - \rho_i \lambda_i^2 b_i^{(2)}\right]}{2(1-\rho)}$$

where $\lambda_i = E[\Lambda_i]$, $\lambda_{ij} = E[\Lambda_i\Lambda_j]$ for $i \neq j$, $\lambda_{ii} = E[\Lambda_i(\Lambda_i - 1)]$, and $\rho_i = \lambda_i b_i$.

We note that Chang and Sandhu [1990, 1992a] show a pseudoconservation law for a continuous-time system in which each queue has a Poisson arrival process and one of exhaustive-limited, gated-limited, and general-decrementing service disciplines. However, their result contains undetermined constants. A similar pseudoconservation law for a system with such service disciplines and compound Poisson arrival processes can also be derived [Baba 1991; Zhigang 1993].

Pseudoconservation laws have also been extended to a continuous-time system in which each queue has Poisson arrival processes of multiple priority classes of customers and one of exhaustive, gated, or 1-limited service disciplines [Fournier and Rosberg 1991; Shimogawa and Takahashi 1992], and to a similar discrete-time system [Takahashi and Krishna Kumar 1994].

2.4. Time-Limited Service Systems

Limited service systems have gained much attention recently in view of the application to the performance modeling of timed-token passing protocols employed in the medium access control (MAC) layer of local and metropolitan area networks standards such as IEEE 802.4 token passing bus and ANSI/IEEE Fiber Distributed Data Interface (FDDI). In timed-token protocols, the time during which each station on the network can continue to transmit packets is limited according to a certain rule that depends on the congestion of the network. Polling models with limited service can be used to approximate the operation of timed-token protocols for which exact performance analysis seems hopeless. We note, however, that the exact result for the mean waiting time in limited service polling systems is only available for a symmetric system with 1-limited service as given in eq. (13c).

Several approximate formulas for the mean waiting time in limited service polling systems were published before 1990; see Takagi [1990a]. Chang and Sandhu [1991] propose an approximation technique for the mean waiting time in a polling system in which the number of customers served at queue i per visit of the server is exhaustively limited by K_i, $1 \leq i \leq N$. They use the known result for a single-queue system with server vacations

$$E[W_i] = \frac{\lambda_i b_i^{(2)}}{2(1-\rho_i)} + \frac{E[V_i^2]}{2E[V_i]} \tag{19}$$

which holds if the vacation times (corresponding to the intervisit times in the polling system) satisfy certain conditions [Cooper 1970; Fuhrmann and Cooper 1985]. While $E[V_i] = R(1-\rho_i)/(1-\rho)$ is known as in eq. (15), $E[V_i^2]$ is estimated from the assumption that V_i is exponentially distributed [Chang and Sandhu 1992a] or it has Gamma distribution in which the two parameters are determined by the approximate analysis of cycle times [Chang and Sandhu 1992b].

Leung [1991] takes a different approach to consider a system in which the parameter K_i is an independent random variable (*probabilistically-limited service*). He uses a numerical technique based on the discrete Fourier transform to determine the queue size distribution (truncated with sufficient accuracy) at service completion times in each queue, and calculates the mean waiting time. LaMaire [1991] studies a single-queue system with server vacations in which K_i is a random variable and the vacation time can be correlated with the value of K_i that was used for the preceding service period. When using this result in the framework of a polling model, he uses simulation to determine the distribution of K_i and the dependence of the vacation time.

In the timed-token protocol, the maximum length of service period is limited not by the number of served customers but by the sum of the service times since the start of the service period as well as during the preceding cycle time. Therefore, for the approximation of the timed-token protocol by a limited service system, the time limit M_i is somehow connected with parameter K_i. Chang and Sandhu [1992b] simply approximate

$$K_i = \frac{M_i + \hat{B}_i(M_i)}{b_i} \quad . \tag{20}$$

where $\hat{B}_i(M_i)$ denotes the mean residual service time of a customer when M_i expires. This form assumes that the service of a customer is continued to completion even if M_i expires in the middle of a service time; this option is called *asynchronous overrun* in FDDI. Karvelas and Leon-Garcia [1991, 1993] show a new approach to the delay analysis of symmetric timed-token networks by which the formula (13a) for the exhaustive service system is used with appropriately inflated service times. de Souze e Silva et al. [1993] propose a numerical technique for an asymmetric time-limited service system with exponentially distributed service times and constant switchover times. The stability and delay performance of timed-token protocols are also studied by Altman [1991], Altman and Kofman [1993], Altman and Liu [1994], Conti et al. [1992, 1993b], Genter and Vastola [1990], Rubin and Wu [1992], and Tangeman and Sauer [1990, 1991]. A survey of these studies is given by Conti et al. [1993a].

Although some interesting techniques have been proposed recently to study a single-queue, time-limited service system with server vacations, their extension to polling systems remains much unexplored.

2.5. Reservation Schemes

In systems with exhaustive, gated, limited, and decrementing service disciplines, the number of messages served at a queue is determined only when and after the server visits the queue. Recently, a few service disciplines according to which this number is determined prior to the time of visit have been analyzed. Let us call such a discipline a *reservation scheme.*

A simple reservation scheme, described by Bertsekas and Gallager [1992 (sec. 3.5.2)], is that when the server visits queue i, it serves only those messages that were there when the server left queue $i-1$. Thus, the switchover time from queue $i-1$ to queue i is supposed to be used for scheduling the services at queue i. The analysis of a polling system with this discipline is similar to that of a polling system with gated service discipline.

Boxma et al. [1992] study two reservation schemes: the globally-gated discipline and the cyclic reservation multiple access (CRMA). In a polling system with *globally-gated discipline,* there is a special queue which we designate queue 1. When the server visits queue i, only those messages that were present there when the server visited queue 1 most recently are served. Customers who may arrive at queue i afterwards (even before the server reaches queue i) will be served in the next cycle. Explicit expressions for the mean waiting time and the pseudoconservation law for a polling system with globally-gated discipline and Poisson arrival processes are available as follows:

$$E[W_i] = \frac{1}{\rho(1+\rho)}\left(1 + 2\sum_{j=1}^{i-1}\rho_j + \rho_i\right)\left[\frac{\rho}{2(1-\rho)}\sum_{j=1}^{N}\lambda_j b_j^{(2)} + \rho\frac{\Delta^2}{2R} + \frac{\rho(1+\rho)}{2(1-\rho)}R\right] + \sum_{j=1}^{i-1} r_i \tag{21a}$$

$$\sum_{i=1}^{N}\rho_i E[W_i] = \frac{\rho}{2(1-\rho)}\sum_{j=1}^{N}\lambda_j b_j^{(2)} + \rho\frac{\Delta^2}{2R} + \frac{\rho(1+\rho)}{2(1-\rho)}R + \sum_{i=2}^{N}\rho_i\sum_{j=1}^{i-1} r_i \tag{21b}$$

This model is extended to a system with server interruptions by Boxma et al. [1993].

Altman et al. [1992] discover a peculiar result when the globally-gated discipline is applied to a scanning polling system. In this system, the server first visits the queues in one direction, that is, in the order 1, 2, $\cdots, N$, serving only those customers that were present when the service to queue 1 was started. Then the server reverses its orientation, and visits the queues in the opposite direction, namely in the order N, $N-1, \cdots$, 1, serving only those customers that were present when the service to queue N was started upon the orientation reversal, and so on. The mean waiting time for queue i in this system is given by

$$E[W_i] = \frac{\rho}{2(1-\rho)} \sum_{j=1}^{N} \lambda_j b_j^{(2)} + \frac{\Delta^2}{2R} + \frac{R}{1-\rho} \tag{22}$$

which is independent of i. Thus the mean waiting times are identical for all queues, exhibiting the fairness. Higher-order moments of the waiting times differ generally.

The *cyclic reservation multiple access* (CRMA) was originally proposed for controlling the access to high-speed local and metropolitan area networks based on a slotted unidirectional folded bus architecture. For modeling the CRMA, Boxma et al. [1992] assume that the single server controls the system by sending out "collectors" periodically for a cyclic tour among the queues to collect their current service requests. After a completion of a tour by a collector, the requests are served (once the services from previous tours have been completed) in the order collected. The mean waiting time for each queue can be calculated.

Khamisy et al. [1992] generalize the globally-gated discipline by assuming that there are n $(1 \leq n \leq N)$ special queues each of which has a disjoint set of subordinate queues (these are logical relations independent of the queue indices). When the server visits a subordinate queue, only those messages that were present there when the server visited its master queue most recently are served. This scheme is called the *synchronized gated discipline.* An extreme case in which $n = 1$ is the globally-gated discipline, while another case in which $n = N$ is the ordinary gated service discipline. The mean waiting times as well as the pseudoconservation law are derived.

Lee and Sengupta [1992a, 1992b] analyze a polling system with limited service and reservations to model the *pipeline polling protocol* for satellite communications. Let K be a parameter, and let $k_{i,j}$ be the number of messages in queue i when the server leaves queue i in the jth cycle. Then, the maximum number of messages served at queue i in the $j+1$th cycle is given by

$$\max[1, \min(K, k_{i,j})] \tag{23}$$

At each cycle, at least one message and at most K messages can be served for each queue. An exact analysis for a single-queue system with server vacations and an approximate analysis for a polling system are provided. The stability of this protocol is discussed by Chang [1992].

2.6. Extended Analysis

Various new developments in the analysis of individual systems as well as generic methodologies have continued to appear. Let us mention a few of them. Systems of two queues, one with exhaustive service and the other with limited or decrementing service are studied by Ibe [1990], Katayama [1992], Lee [1994], Ozawa [1990], and Weststrate [1990]. Kubat and Servi [1991] derive the optimal server scheduling for systems with two queues and zero service times. Levy and Sidi [1991] analyze exhaustive and gated service systems with compound Poisson arrivals. Approximation techniques for the mean waiting times in limited and decrementing service systems, based on pseudoconservation laws, are proposed by Balsamo [1990], Casares-Giner [1991], and Chang and Hwang [1991, 1993]. Systems in which a setup time (also called a switch-in time) is needed before starting service at each queue are dealt with by Altman et al. [1994], Altman and Yechiali [1993], and Gupta and Srinivasan [1993] (while systems with zero switchover times in which a setup time is needed before starting service when a message arrives in an empty system are treated by Fuhrmann and Moon [1990] and Takagi [1987]). Stochastic ordering of polling systems can be found in Altman et al. [1992], Boxma and Kelbert [1994], Levy et al. [1990], and Liu et al. [1992]. Resing [1993] characterizes polling systems with multitype branching processes; he notes that he was unaware of an early work by Fuhrmann [1981]. An idea of ancestral line is also employed for numerical computation by Konheim and Levy [1992] and Konheim et al. [1992]. Tsai and Rubin [1992] analyze the mean waiting time in a symmetric polling system with two priority classes of customers. Zhang and Acampora [1992] propose and analyze a modified polling scheme for the indoor wireless local area networks.

2.7. Stability

Stability issues in polling models are not straightforward as they should be analyzed in the framework of multidimensional Markov chains; see Szpankowski [1990] for stability criteria in multidimensional stochastic processes. Recently, Georgiadis and Szpankowski [1992, 1993] prove the following necessary and sufficient condition for cyclic polling systems with k_i-limited service discipline at queue i:

$$\rho < 1 \quad \text{and} \quad \lambda_i < \frac{k_i(1-\rho)}{R} \quad \text{for all } i \in \{1, 2, \cdots, N\} \tag{24}$$

While this condition was "known" for years, a formal proof has been lacking. The proof is based on stochastic dominance techniques and application of Loynes' stability criteria for an isolated queue. Altman et al. [1992] also provide a sufficient condition under which the vector of queue sizes is ergodic. In addition, they show that the queue sizes, station times, intervisit times, and cycle times are stochastically increasing in arrival rates, in service times, and in switchover times. Furthermore, the mean cycle time, the mean intervisit time, and the mean station time are invariant under general service disciplines and general arrival and service processes. Altman and Spieksma [1992] give necessary and sufficient conditions for the existence of all the moments of station times, and show the geometric convergence. Sufficient conditions for the central limit theorem and the law of iterated logarithm for the moments of station times are also given. Fricker and Jaibi [1994] and Massoulié [1993] also deal with the stability.

2.8. Numerical Algorithms

The exact mean waiting times in an asymmetric exhaustive service system can be computed by solving a set of $O(N^2)$ equations in (16a)–(16c), which is due to Ferguson and Aminetzah [1985], or another set of $O(N)$ equations derived by Sarkar and Zangwill [1989] (implementation of the Sarkar-Zangwill algorithm are given in Garner [1988] and Tayur and Sarkar [1988]). Using the standard method for solving a set of linear simultaneous equations, even the latter approach requires $O(N^3)$ operations. The situation is the same for a gated service system. Later, Srinivasan [1991] and Srinivasan et al. [1991] propose an algorithm that requires $O(N^2)$ operations to compute the mean waiting time at a given queue. Iterative algorithms that require $O(N \log_\rho \epsilon)$ operations, where ϵ is the accuracy required, are invented by Konheim and Levy [1992], Konheim et al. [1992], and Levy [1991]. Federgruen and Katalan [1993] show an efficient numerical method to compute the steady state queue size distributions for exhaustive and gated service systems.

An exact computation of the mean waiting time in an asymmetric limited service system has been unavailable. Thus, Blanc [1990a, 1990b, 1991, 1992a, 1992b] and Blanc and van der Mei [1993] developed the *power-series algorithm* which can be applied to systems with Coxian distributed service and switchover times. Leung [1990a, 1990b, 1991] used discrete Fourier transforms to approximate unknown functions that appear in the analysis of limited service systems.

2.9. Networks of Queues

An open network of queues served by a single server in cyclic order is considered by Sidi and Levy [1990] and Sidi et al. [1992]. This is a generalization of the basic polling system to the case in which customers, after their service completion at queue i, move to queue j with probability p_{ij}, or leave the network with probability $1 - \sum_{j=1}^{N} p_{ij}$. Levi and Sidi [1990] mention the modeling of selective-repeat ARQ protocol and distributed algorithm on token ring networks and robotics systems with staged jobs as potential applications of this network. Not only the queue size and the waiting time at each queue but also the expected delay of customers who follow a specific route in the network can be calculated for gated and exhaustive service systems. Altman and Yechiali [1994] analyze a closed network of queues served by a single server in cyclic order. Katayama [1991] analyzes a tandem of two queues (with Bernoulli feedback at the first queue) for customers of multiple classes served cyclically by a single server with zero switchover times. Katayama [1994] also considers a tandem of two queues with gated service and vacations for the first queue and exhaustive service for the second queue.

Walrand [1990] describes dynamic optimization problems to determine, at each service completion time, which queue should be served next by the single server so as to minimize the given cost, for example,

$$E\left[\sum_{i=1}^{N} c_i L_i\right] \tag{25}$$

where L_i is the number of customers present at queue i. For networks without external arrivals, an optimization problem of this kind can be viewed as a *multi-armed bandit problem* because the system state evolves only at the queue that is attended by the server.

2.10. Finite Systems

The state space of a polling system is finite if the number of customers involved in the system is finite or if the capacity of the system is finite. Let us call the former a *finite-population system*, and call the latter a *finite-capacity system.*

A finite-population system is a closed system in which each customer alternates a period of being in the queue (including a period of being served) and a period of being in the source. From the analogy with the finite-population systems in priority queues [Jaiswal 1968 (sec. III.1); Takagi 1993 (sec. 4.7)], we may consider two models of the polling system with a finite population that differ with respect to the population constraint: a multiple finite-source model and a single finite-source model. In the *multiple finite-source model*, the source of customers for each queue is associated only with that queue; a customer whose service has been completed always returns to its original source. The single-buffer system discussed in Sections 1.1 and 2.1 is a special case of the multiple finite-source model in which the population size for each queue is one. In the *single finite-source model*, there is only a single source of customers each of which selects one of the queue at random when it leaves the source.

The multiple finite-source model and the single finite-source model may be associated with the *flow control* and *congestion avoidance* mechanisms in computer communications networks. Namely, the multiple finite-source model in which the population size is fixed for each queue corresponds to the *window flow control* [Davies et al. 1979 (sec. 4.2)]. The single finite-source model in which the total population size is a single constraint corresponds to the *isarithmic congestion avoidance* scheme [Davies et al. 1979 (sec. 4.3)].

The multiple finite-source model was analyzed by Choi and Trivedi [1992] and Ibe and Trivedi [1990] using a technique called *stochastic Petri nets* under the assumption (which is essential in this technique) that both service and switchover times are exponentially distributed. A similar model with generally distributed service and switchover times was analyzed by Takagi [1992] by using the analysis of an M/G/1//N system with server vacations.

A finite-capacity polling system with exhaustive, gated, and 1-limited service is analyzed by Takagi [1991a] by considering the Markov chain for the set of queue sizes embedded at the times when the server visits and leaves each queue. The server vacation time for each queue is calculated by conditioning on this Markov chain. The queue size at each queue is then obtained by utilizing the existing analysis of an M/G/1/K queue with server vacations. An independence assumption implicitly involved in Takagi [1991a] for the vacation time and the busy period is removed by Kofman [1991]. Jung [1991] and Jung and Un [1991] provide a different technique, called *virtual buffering*, to analyze a finite-capacity polling system with exhaustive service. They introduce a virtual buffer of an infinite capacity for each queue when the server is on vacation for that queue. When the server comes to the queue, the messages that exceed the capacity of the real buffer are removed from the queue. A Markov chain for the set of queue sizes at the time when the server visits each queue is studied based on this model, and the duration of the vacation is obtained. Tran-Gia [1992] gives an approximate analysis of a discrete-time, 1-limited service system with a finite capacity and a renewal arrival process. Lang and Bosch [1991] present a similar technique for the analysis of a continuous-time, k-limited service system with a finite capacity and a Poisson arrival process.

2.11. Multiserver Systems

Polling systems with multiple servers are studied in a series of papers by Ajmone Marsan et al. [1990–1994] using a technique of *generalized stochastic Petri nets* (GSPNs). In this technique, once a model is described in terms of GSPN primitives, all possible marking states are mapped into a state space of a Markov chain, whose steady-state distribution is computed numerically. The performance measures are then calculated. While several earlier studies of multiserver systems were all approximate, the method of Ajmone Marsan et al. can provide exact numerical values for the customer waiting times. The restrictions include exponential distributions for the customer interarrival times, service times, and the server switchover times, finite-capacity queues so as to obtain finite state spaces, and the limited size of the models (in terms of the numbers of queues, servers, and buffers) that can be solved with acceptable time and space complexity. A unique parameter in multiserver systems is the maximum number of servers that can attend the same queue simultaneously. Some servers may pass others at some queues (just like an empty bus on a street passes another at a crowded bus stop). These features that are difficult to handle analytically can be easily incorporated in the GSPN model.

The numerical results presented by Ajmone Marsan et al. [1990–1994] show that, when the total service rate is fixed, the mean waiting times can be greatly reduced by using multiple servers, particularly when the switchover times are long and the system is loaded (this situation corresponds to the large distances and high data rates in metropolitan area networks). For some values of the system parameters, the same can also happen for the customer response time (waiting time plus service time), but in a less sizable manner, due to the fact that the increase in the number of servers at a fixed service rate implies an increase in the mean service time. An interesting behavior of an asymmetric system with two servers observed by Ajmone Marsan et al. [1994] is that a heavily loaded queue monopolizes one server while lightly loaded queues share the other server.

Multiserver systems are also studied by Borst [1994] and by van der Mei and Borst [1994] using a different approach.

2.12. Effects of Switchover Times

Traditionally, the analysis of systems with zero switchover times has been separate from that of systems with nonzero switchover times [Cooper 1970]. A reason for technical difficulty in relating the two was said to be as follows. In a system with zero switchover times, the server executes an infinite number of cycles in any finite period during which the system is empty, which implies that the mean cycle time is zero. On the other hand, the analysis of systems with nonzero switchover times is based on the evaluation of variables averaged over a cycle time. Recently, however, a few works have appeared to relate the analysis for the two systems.

As a special case, if all the switchover times are constant, we can derive from eqs. (16a)–(16c) another set of equations with respect to $\hat{r}_{ij} := r_{ij}/R$ which can be used to calculate the ratio $E[(I_i)^2]/E[I_i]$ for $E[W_i]$ in eq. (14). We can then obtain a set of equations for a system of zero switchover times by letting $R \to 0$. This idea was given by Choudhury [1990] and Levy and Kleinrock [1991]. More recently, Cooper et al. [1992] found the following results. For an exhaustive service system, we have

$$E[W_i] = E[W_i^0] + \frac{R(1-\rho_i)}{2(1-\rho)} \tag{26a}$$

where W_i^0 is the waiting time in the corresponding system with zero switchover times and with the modified service-time variances

$$\tilde{b}_i^{(2)} = b_i^{(2)} + \frac{\delta_{i-1}^2}{R}\frac{1-\rho}{\lambda_i} \tag{26b}$$

For a gated service system, we have

$$E[W_i] = E[W_i^0] + \frac{R(1+\rho_i)}{2(1-\rho)} \tag{27a}$$

where W_i^0 is the waiting time in the corresponding system with zero switchover times and with the modified service-time variances

$$\tilde{b}_i^{(2)} = b_i^{(2)} + \frac{\delta_i^2}{R}\frac{1-\rho}{\lambda_i} \tag{27b}$$

The results in (26a) and (27a) were also derived by Fuhrmann [1992] in a special case in which the switchover times are constant. We note that the modification of the service-time variances as in (26b) and (27b) was used by Ferguson and Aminetzah [1985] when they derived the pseudoconservation laws. The work of Fuhrmann [1992] and Cooper et al. [1992] has been generalized to transforms (and therefore to higher moments) of the waiting times by Srinivasan et al. [1993].

2.13. Optimal Server Routing/Action

When we have control over the server movement, we can optimize its routing (a sequence of choices of the queues to serve) and its action (service discipline) in each queue at given decision epochs in order to minimize a given objective function. Such *stochastic optimization problems* may be classified into static, semi-dynamic, and dynamic ones.

In the *static optimization*, decisions are made once and for all before the operation of the system by selecting controllable parameters. For the static optimization of service disciplines when the polling order is given (not necessarily cyclic), Levy et al. [1990] prove that the exhaustive service discipline dominates all other disciplines (under the assumption that the server does not wait idling at a queue) with respect to the total amount of unfinished work in the system at any time. Boxma et al. [1990, 1991, 1993] (see also Boxma [1991]) investigate the statically optimized polling order table (ratio of occurrences of all queues in the table, size of the table, and the order within the table) that minimizes

$$\sum_{i=1}^{N} \rho_i E[W_i] \tag{28}$$

for an exhaustive or gated service system. Interestingly, they show that the *golden ratio policy* provides a good heuristic for determining a good visit order in the table. An optimization that specifies not only the polling order but also the starting time of each polling is considered by Borst et al. [1992].

In the *semi-dynamic optimization*, the decision for the polling order is made at the beginning of each polling cycle based on the knowledge of the system process at most up to that time. Browne and Yechiali [1989] formulate a *Markov decision process* to consider the minimization of the cycle time given the set of queue sizes $\{L_1, \cdots, L_N\}$ at the beginning of a polling cycle. According to them, the expected cycle time is minimum if the server visits queues in the order of increasing values of L_i/λ_i (this rule is independent of the service times) in gated and exhaustive service systems with zero switchover times. This result is extended by Fabian and Levy [1993, 1994] for the cycle time maximization. A similar optimization is considered for single-buffer systems by Browne and Yechiali [1991]. Yechiali [1991] summarizes the results including those for systems with binomial-gated and Bernoulli-gated disciplines.

Liu et al. [1992] address *dynamic optimization* problems that determine the server's routing and action so as to stochastically minimize the unfinished work and the number of customers in the system at all times. They find the following results: When the server is at a nonempty queue, it should neither idle nor switch until that queue is empty (exhaustive and greedy discipline). For symmetric systems, (i) when the system has become empty, the server should stay idling at the last visited queue (patient policy); and (ii) when the server has emptied a queue, it should subsequently visit the queue with the largest queue size (stochastically largest queue policy). (The claim (ii) is also shown by Miyoshi et al. [1993].) The cyclic routing is optimal when the only available information is the previous decisions. (Prior to this work, Liu and Nain [1992] considered a particular polling system arising from the videotex system, and identified optimal scheduling policies depending on the amount of information available to the controller. Towsley et al. [1991] proved that the largest queue policy minimizes the number of customers lost when the queues have finite and equal buffer capacities.) A similar dynamic optimization is studied by Ajmone Marsan et al. [1993] for systems with multiple servers. One of their findings is that the cyclic polling order provides a very good performance, often superior to some idealistic schemes, in particular when systems with a large number of queues and a small number of servers are heavily loaded (high arrival rates and long switchover times). This is because the cyclic polling follows the minimum distance Hamiltonian tour of queues. For single-buffer systems, Harel and Stulman [1990] find the optimal parameter d of a *horizon server* such that it acts like a greedy server within the d nearest queues and acts like a cyclic server when all the d nearest queues are empty.

The optimal buffer capacity has been studied by Birman et al. [1991], Gail et al. [1992], and Sasaki [1993] as *deterministic optimization problems*, in which the gradual (i.e., non-instantaneous) arrival process is given and the server routing is *longest queue first*, *least time to reach bound*, or *partially gated* (only those messages that have completed arrival by the polling instant are served).

3. Future Research Directions

A strong thrust for the progress in mathematical theory comes from the need by application. One of the recent application fields that motivated the study of polling models was media access protocols in local and metropolitan area networks. Another incentive is the establishment of mathematical foundations, such as stability, stochastic decomposition, and boundary value problems. Also, there can be innumerable results by the "Cartesian product" of solved models (combinations of variations). Certain developments made in the recent years are overviewed in Section 2. It is likely that many more works will be done along the existing lines.

Some new directions, though challenging, may be suggested as my personal view. First, we can investigate real systems and build models with new features. For example, the server in switchover may slow down if there are customers to serve in the next queue, as usual in transportation and other systems accompanying mechanical movement. If there are no customers to serve in the whole system, the server may go to a home base like elevators or stop at the last served queue (a *homing server* or a *patient server*, as opposed to a *roving server* in the conventional model). Such models have been studied recently by Ajmone Marsan et al. [1994], Borst [1993], Eisenberg [1994], and Srinivasan and Gupta [1993]. Note that the implementation of the homing or patient server requires the information about the whole system, which may not be easy to obtain for certain applications. Polling models with non-Poisson arrival processes should also be paid attention in view of vigorous modeling efforts for broadband ISDN systems. They include modulated Markov Poisson processes, linear input models [Daganzo 1991; Lee 1993], gradual input processes [Birman et al. 1991, Gail et al. 1992; Sasaki 1993], and "Cruz-type" processes [Altman et al. 1993, 1994]. Embedding polling models in a network of queues or in other total system models is also important. Optimization of the server routing/action is much unexplored because of analytical intractability. As an alternative, Matsumoto [1994] proposes a method of controlling the server by a *neural network* in an asymmetric system so that the mean waiting time is minimized.

Acknowledgments

This work is supported in part by the Telecommunications Advancement Foundation.

The author wants to thank Dr. E. Altman and Professors O. J. Boxma, R. B. Cooper, and M. M. Srinivasan for their valuable comments on the draft version of this text.

References

[1] Ajmone Marsan, M., de Moraes, L. F., Donatelli, S., and Neri, F. 1990. Analysis of symmetric nonexhaustive polling with multiple servers. *IEEE INFOCOM '90*, pp.284-295, San Fransisco, June 3-7, 1990.

[2] Ajmone Marsan, M., de Moraes, L. F., Donatelli, S., and Neri, F. 1992. Cycles and waiting times in symmetric exhaustive and gated multiserver multiqueue systems. *IEEE INFOCOM '92*, pp.2315-2324, Florence, Italy, May 6-8, 1992.

[3] Ajmone Marsan, M., Donatelli, S., and Neri, F. 1989. GSPN models of multiserver multiqueue systems. *The Third International Workshop on Petri Nets and Performance Models*, pp.19-28, December 11-13, 1989, Kyoto, Japan.

[4] Ajmone Marsan, M., Donatelli, S., and Neri, F. 1990. GSPN models of Markovian multiserver multiqueue systems. *Performance Evaluation*, Vol.11, No.4 (November), pp.227-240.

[5] Ajmone Marsan, M., Donatelli, S., and Neri, F. 1991. Multiserver multiqueue systems with limited service and zero walk time. *IEEE INFOCOM '91*, pp.1178-1188, Bal Harbour, Florida, April 9-11, 1991.

[6] Ajmone Marsan, M., Donatelli, S., Neri, F., and Fantini, G. 1994. Some numerical results on finite-buffer Markovian multiserver polling systems. Preprint.

[7] Ajmone Marsan, M., Donatelli, S., Neri, F., and Rubino, U. 1991a. GSPN models of random, cyclic, and optimal 1-limited multiserver multiqueue systems. *ACM SIGCOMM '91 Conference, Communications Architectures & Protocols*, pp.69-80, Zürich, Switzerland, September 3-6, 1991.

[8] Ajmone Marsan, M., Donatelli, S., Neri, F., and Rubino, U. 1991b. On the construction of abstract GSPNs: an exercise in modeling. *Proceedings of the Fourth International Workshop on Perti Nets and Performance Models*, pp.2-17, Melbourne, Australia, December 2-5, 1991.

[9] Ajmone Marsan, M., Donatelli, S., Neri, F., and Rubino, U. 1993. Good and bad dynamic polling orders in symmetric single buffer Markovian multiserver multiqueue systems. *IEEE INFOCOM '93*, pp.176-185, San Francisco, California, March 30 - April 1, 1993.

[10] Altman, E., 1991. Analysis of timed-token ring protocols. Report RR 1570, INRIA, December 1991. To appear in ITC 14, Antibes, June 1994.

[11] Altman, E., Blanc, H., Khamisy, A., and Yechiali, U. 1994. Gated-type polling systems with walking and switch-in times. To appear in *Stochastic Models.*

[12] Altman, E., and Foss, S. 1993. Polling on a graph with general arrival and service time distribution. Report No. 1992, INRIA.

[13] Altman, E., Foss, S., Riehl, E. R., and Stidham, S. 1993. Performance bounds and pairwise stability for generalized vacation and polling systems. Research Report UNC/OR TR93-8, University of North Carolina at Chapel Hill.

[14] Altman, E., Foss, S., Riehl, E. R., and Stidham, S. 1994. Sample path analysis of token rings. To appear in ITC 14, Antibes, June 1994.

[15] Altman, E., Khamisy, A., and Yechiali, M. 1992. On elevator polling with globally gated regime. *Queueing Systems*, Vol.11, No.1-2 (July), pp.85-90.

[16] Altman, E., and Kofman, D. 1993. Sample path bounds for performance measures of token rings. Research Report Telecom Paris, 1993.

[17] Altman, E., Konstantopoulos, P., and Liu, Z. 1992. Stability, monotonicity and invariant quantities in general polling systems. *Queueing Systems,* Vol.11, No.1-2 (July), pp.35-57.

[18] Altman, E., and Levy, H. 1994. Queueing in space. To appear in *Journal of Applied Probability.*

[19] Altman, E., and Liu, Z. 1994. Improving the stability characteristics of asynchronous traffic in FDDI token rings. *High Speed Networks and Their Performance,* H. G. Perros and Y. Viniotis (editors), pp.441-460, North-Holland.

[20] Altman, E., and Spieksma, F. 1992. Polling systems: moment stability of station times and central limit theorems. Research report No. TW-92-09, Leiden University, The Nethelands.

[21] Altman, E., and Yechiali, U. 1993. Cyclic Bernoulli polling. *Zeitschrift für Operations Research,* Vol.38, No.1, pp.55-76.

[22] Altman, E., and Yechiali, U. 1994. Closed polling systems. To appear in *Probability in the Engineering and Informational Sciences.*

[23] Baba, Y. 1991. Analysis of batch arrival cyclic service multiqueue systems with limited service discipline. *Journal of the Operational Research Society of Japan,* Vol.34, No.1 (March), pp.93-104.

[24] Balsamo, S. 1990. Analysis of asymmetric polling systems with limited service: waiting time approximation. Preprint.

[25] Bertsekas, D., and Gallager, R. 1992. *Data Networks.* Second edition, Prentice-Hall, Englewood Cliffs, New Jersey (First edition: 1987).

[26] Bertsimas, D. J., and Van Ryzin, G. 1991. A stochastic and dynamic vehicle routing problem in the Eucledian plane. *Operations Research,* Vol.39, No.4 (July-August), pp.601-615.

[27] Birman, A., Gail, H. R., Hantler, S. L., Rosberg, Z., and Sidi, M. 1991. An optimal service policy for buffer systems. IBM Research Report RC 16641, March 5, 1991.

[28] Bisdikian, C. 1993. A note on the conservation law for queues with batch arrivals. *IEEE Transactions on Communications,* Vol.41, No.6 (June), pp.832-835.

[29] Bisdikian, C., and Merakos, L. 1992. Output process from a continuous token-ring local area network. *IEEE Transactions on Communications,* Vol.40, No.12 (December), pp.1796-1799.

[30] Blanc, J. P. C., 1990a. A numerical approach to cyclic-service queueing models. *Queueing Systems,* Vol.6, No.2 (April), pp.173-188.

[31] Blanc, J. P. C., 1990b. Cyclic polling systems: limited service versus Bernoulli schedules. Technical report FEW 442, Faculty of Economics, Tilburg University, P. O. Box 90153, 5000 LE Tilburg, The Netherlands.

[32] Blanc, J. P. C., 1991. The power-series algorithm applied to cyclic polling systems. *Communications in Statistics - Stochastic Models,* Vol.7, No.4, pp.527-545.

[33] Blanc, J. P. C., 1992a. Performance evaluation of polling systems by means of the power-series algorithm. *Annals of Operations Research,* Vol.35, No.1-4 (April), pp.155-186.

[34] Blanc, J. P. C., 1992b. An algorithmic solution of polling models with limited service discipline. *IEEE Transactions on Communications,* Vol.40, No.7 (July), pp.1152-1155.

[35] Blanc, J. P. C., and van der Mei, R. D. 1993. Optimization of polling systems: computation of derivatives by means of power-series algorithm. *Performance '93,* G. Iazeolla and S. Lavenberg (editors), pp.385-398, September 29 - October 1, 1993, Roma, Italy.

[36] Borst, S. C. 1993. A pseudo-conservation law for a polling system with a homing server. Preprint.

[37] Borst, S. C. 1994. Polling systems with multiple coupled servers. CWI Report BS-R9408, The Netherlands.

[38] Borst, S. C., Boxma, O. J., and Huitema, G. B. 1992. Optimization of fixed time polling schemes. CWI Report BS-R9237, The Netherlands. To appear in *Telecommunication Systems.*

[39] Boxma, O. J. 1989. Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems,* Vol.5, No.1-3 (November), pp.185-214.

[40] Boxma, O. J. 1991. Analysis and optimization of polling systems. *Queueing, Performance and Control in ATM (ITC-13),* J. W. Cohen and C. D. Pack (editors), pp.173-183, Elsevier Science Publishers B.V. (North-Holland), Amsterdam.

[41] Boxma, O. J., and Groenendijk, W. P. 1987. Pseudo-conservation laws in cyclic-service systems. *Journal of Applied Probability,* Vol.24, No.4 (December), pp.949-964.

[42] Boxma, O. J., and Groenendijk, W. P. 1988. Waiting times in discrete-time cyclic-service systems. *IEEE Transactions on Communications,* Vol.COM-36, No.2 (February), pp.164-170.

[43] Boxma, O. J., and Kelbert, M. 1994. Stochastic bounds for a polling system. *Annals of Operations Research,* Vol.48.

[44] Boxma, O. J., Levy, H., and Weststrate, J. A. 1990. Optimization of polling systems. *Performance '90,* P. J. B. King, I. Mitrani, and R. J. Pooley (editors), pp.349-361, Elsevier Science Publishers B. V. (North-Holland), Amsterdam.

[45] Boxma, O. J., Levy, H., and Weststrate, J. A. 1991. Efficient visit frequencies for polling tables: minimization of waiting cost. *Queueing Systems,* Vol.9, No.1-2 (October), pp.133-162.

[46] Boxma, O. J., Levy, H., and Weststrate, J. A. 1993. Efficient visit orders for polling systems. *Performance Evaluation,* Vol.18, No.2 (September), pp.103-123.

[47] Boxma, O. J., Levy, H., and Yechiali, U. 1992. Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Annals of Operations Research,* Vol.35, No.1-4 (April), pp.187-208.

[48] Boxma, O. J., Weststrate, J. A., and Yechiali 1993. A globally gated polling system with server interruptions, and applications to the repairman problem. *Problems in Engineering and Information Science,* Vol.7, pp.187-208.

[49] Browne, S., and Yechiali, U. 1989. Dynamic priority rules for cyclic-type queues. *Advances in Applied Probability,* Vol.21, No.2 (June), pp.432-450.

[50] Browne, S., and Yechiali, U. 1991. Dynamic scheduling in single server multi-class service systems with unit buffers. *Naval Research Logistics,* Vol.38, pp.383-396.

[51] Bunday, B. D., and Sztrik, J. 1992. The maintenance of bi-directionally patrolled machines. *I. M. A. Journal of Mathematics Applied in Business and Industry,* Vol.3, pp.377-386.

[52] Bunday, B. D., Sztrik, J., and Tapsir, R. B. 1992. A heterogeneous scan service polling model with single-message buffer. *Performance of Distributed Systems and Integrated Communication Networks,* T. Hasegawa, H. Takagi, and Y. Takahashi (editors), pp.99-111, Elsevier Science Publishers B.V. (North-Holland), Amsterdam.

[53] Campbell, G. M. 1991. Cyclic queueing systems. *European Journal of Operational Research,* Vol.51, No.2 (March), pp.155-167.

[54] Casares-Giner, V. 1991. A pseudoconservation law based approximate analysis for cyclic service queuing system with decrementing service discipline. *Teletraffic and Datatraffic in a Period of Change, ITC-13,* A. Jensen and V. B. Iversen (editors), pp.749-754, Elsevier Science Publishers B.V. (North-Holland), Amsterdam.

[55] Chang, C.-J., and Hwang, L.-C. 1991. Analysis for a general service nonexhaustive polling system with pseudoconservation law. Technical report, National Chiao Tung University, Taiwan, December 1991.

[56] Chang, C.-J., and Hwang, L.-C. 1993. Analysis of a general service nonexhaustive polling system using a heuristic combination method and pseudoconservation law. Preprint.

[57] Chang, K. C. 1992. Stability conditions for a pipeline polling scheme in satellite communications. Preprint.

[58] Chang, K. C., and Sandhu, D. 1990. Pseudo-conservation laws in cyclic-server, multiqueue systems with a class of limited service policies. *IEEE INFOCOM '90*, pp.260-267, San Francisco, June 3-7, 1990.

[59] Chang, K. C., and Sandhu, D. 1991. Mean waiting time approximations in cyclic-service systems with exhaustive limited service policy. *IEEE INFOCOM '91*, pp.1168-1177, Bal Harbour, Florida, April 9-11, 1991. Also in *Performance Evaluation*, Vol.15, No.1, pp.21-40, April 1992.

[60] Chang, K. C., and Sandhu, D. 1992a. Pseudo-conservation laws in cyclic-service systems with a class of limited service policies. *Annals of Operations Research*, Vol.35, No.1-4 (April), pp.209-229.

[61] Chang, K. C., and Sandhu, D. 1992b. Delay analysis of token-passing protocols with limited token holding times. *IEEE INFOCOM '92*, pp.2299-2305, Florence, Italy, May 6-8, 1992.

[62] Chiarawongse, J., and Srinivasan, M. M. 1991. On pseudo-conservation laws for the cyclic server system with compound Poisson arrivals. *Operations Research Letters*, Vol.10, No. 8 (November), pp.453-459.

[63] Choi, H., and Trivedi, K. S. 1992. Approximate performance models of polling systems using stochastic Petri nets. *IEEE INFOCOM '92*, pp.2306-2314, Florence, Italy, May 6-8, 1992.

[64] Choudhury, G. L. 1990. Polling with a general service order table: Gated service. *IEEE INFOCOM '90*, pp.268-276, San Francisco, June 3-7, 1990.

[65] Chung, H., Un, C. K., and Jung, W. Y. 1992. Performance analysis of Markovian polling systems with single buffers. To appear in *Performance Evaluation*.

[66] Coffman, E. G. Jr., and Gilbert, E. N. 1986. A continuous polling system with constant service times. *IEEE Transactions on Information Theory*, Vol.IT-33, No.4 (July), pp.584-591.

[67] Coffman, E. G., Jr., and Stolyar, A. 1992. Continuous polling on graphs. Preprint.

[68] Conti, M., Gregori, E., and Lenzini, L. 1992. A vacation model for interconnected FDDI networks. *ACM 1992 Computer Science Conference*, pp.9-16, Kansas City, March 3-5, 1992.

[69] Conti, M., Gregori, E., and Lenzini, L. 1993a. Metropolitan area networks (MANs): protocols, modeling and performance evaluation. In: *Performance Evaluation of Computer and Communication Systems, Joint tutorial papers of Performance '93 and Sigmetrics '93*, L. Donatiello and R. Nelson (editors), pp.81-120, Lecture Notes in Computer Science 729, Springer-Verlag, Berlin.

[70] Conti, M., Gregori, E., and Lenzini, L. 1993b.. Design and analysis of a medical communication system based on FDDI. *Microcomputer Applications, (Special Issue on Medical Applications)*, Vol.12, No.3, pp.118-127.

[71] Cooper, R. B. 1970. Queues served in cyclic order: Waiting times. *The Bell System Technical Journal*, Vol.49, No.3 (March), pp.399-413.

[72] Cooper, R. B., Niu, S.-C., and Srinivasan, M. M. 1992. A decomposition theorem for polling models: the switchover times are effectively additive. To appear in *Operations Research*.

[73] Daganzo, C. F. 1990. Some properties of polling systems. *Queueing Systems*, Vol.6, No.2 (April), pp.137-154.

[74] Davies, D. W., Barber, D. L. A., Price, W. L., and Solomonides, C. M. 1979. *Computer Networks and Their Protocols*. John Wiley and Sons, New York.

[75] de Souza e Silva, E., Gail, H. R., and Muntz, R. R. 1993. Polling systems with server timeout. Preprint.

[76] Eisenberg, M. 1994. The polling system with a stopping server. Preprint.

[77] Fabian, O., and Levy, H. 1993. Polling system optimization through dynamic routing policies. *IEEE INFOCOM '93*, pp.194-200, San Francisco, California, March 30 - April 1, 1993.

[78] Fabian, O., and Levy, H. 1994. Pseudo-cyclic policies for multi-queue single server systems. Annals of Operations Research, Vol.48.

[79] Federgruen, A. and Katalan, Z. 1993. Approximating queue size and waiting time distributions in general polling systems. To appear in *Queueing Systems.*

[80] Ferguson, M. J., and Aminetzah, Y. J. 1985. Exact results for nonsymmetric token ring systems. *IEEE Transactions on Communications,* Vol.COM-33, No.3 (March), pp.223-231.

[81] Fournier, L., and Rosberg, Z. 1991. Expected waiting times in cyclic service systems under priority disciplines. *Queueing Systems,* Vol.9, No.4 (October), pp.419-439.

[82] Fricker, C., and Jaibi, R. 1994. Monotonicity and stability of periodic polling models. *Queueing Systems,* Vol.15, pp.211-238.

[83] Fuhrmann, S. W. 1981. Performance analysis of a class of cyclic schedules. Bell Laboratories Technical Memorandum, 59531-810316.01TM, March 16, 1981.

[84] Fuhrmann, S. W. 1992. A decomposition result for a class of polling models. *Queueing Systems,* Vol.11, No.1-2 (July), pp.109-120.

[85] Fuhrmann, S. W., and Cooper, R. B. 1985. Stochastic decompositions in the M/G/1 queue with generalized vacations. *Operations Research,* Vol.33, No.5 (September-October), pp.1117-1129.

[86] Fuhrmann, S. W., and Moon, A. 1990. Queues served in cyclic order with an arbitrary start-up distribution. *Naval Research Logistics,* Vol.37, No.1 (February), pp.123-133.

[87] Gail, H. R., Grover, G., Guérin, R., Hantler, S. L., Rosberg, Z., and Sidi, M. 1992. Buffer size requirements under longest queue first. *Performance of Distributed Systems and Integrated Communication Networks,* T. Hasegawa, H. Takagi, and Y. Takahashi (editors), pp.413-424. Elsevier Science Publishers B.V. (North-Holland), Amsterdam; Also in *Performance Evaluation,* Vol.18, No.2, pp.133-140, September 1993.

[88] Garner, G. M. 1988. Mean and variance of waiting time for nonsymmetric cyclic queueing systems. Technical memorandum 59482-880600-00TM, AT&T Bell Laboratories.

[89] Genter, W. L., and Vastola, K. S. 1990. Delay analysis of the FDDI synchronous data class. *IEEE INFOCOM '90,* pp.766-773, San Fransisco, June 3-7, 1990.

[90] Georgiadis, L., and Szpankowski, W. 1992. Stability of token passing rings. *Queueing Systems,* Vol.11, No.1-2 (July), pp.7-33.

[91] Georgiadis, L., and Szpankowski, W. 1993. Stability criteria for yet another multidimensional distributed system. Revised version. Report No. 1996, INRIA.

[92] Grillo, D. 1990. Polling mechanism models in communication systems – some application examples. In: *Stochastic Analysis of Computer and Communication Systems,* H. Takagi (editor), pp.659-698, Elsevier Science Publishers B. V. (North-Holland), Amsterdam.

[93] Gupta, D., and Srinivasan, M. M. 1993. Polling systems with state-dependent setup times. Presentation at the ORSA/TIMS Phoenix 1993 meeting.

[94] Harel, A., and Stulman, A. 1990. Polling greedy and horizon servers on a circle. To appear in *Operations Research.*

[95] Ibe, O. C. 1990. Analysis of polling systems with mixed service disciplines. *Communications in Statistics – Stochastic Models,* Vol.6, No.4, pp.667-689.

[96] Ibe, O. C., and Trivedi, K. S. 1990. Stochastic Petri net models of polling systems. *IEEE Journal on Selected Areas in Communications,* Vol.8, No.9 (December), pp.1649-1657.

[97] Jaiswal, N. K. 1968. *Priority Queues.* Academic Press, New York.

[98] Jung, W. Y. 1991. Analysis of finite-capacity polling systems based on virtual buffering and lumped modeling. Ph.D. dissertation, Report CRT-T-9109, Communications Research Laboratory, Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Seoul, Korea, June 1991.

[99] Jung, W. Y., and Un, C. K. 1991. Analysis of a finite buffer polling system with exhaustive service based on virtual buffering. To appear in *IEEE Transactions on Communications.*

[100] Karvelas, D., and Leon-Garcia, A. 1991. A general approach to the delay analysis of symmetric token passing networks. *IEEE INFOCOM '91,* pp.181-190, Bal Harbour, Florida, April 9-11, 1991.

[101] Karvelas, D., and Leon-Garcia, A. 1993. Delay analysis of various service disciplines in symmetric token passing networks. *IEEE Transactions on Communications,* Vol.41, No.9 (September), pp.1342-1355.

[102] Katayama, T. 1991. Analysis of a multi-class service tandem queueing model with feedback attended by a single server. *Teletraffic and Datatraffic in a Period of Change, ITC-13,* A. Jensen and V. B. Iversen (editors), pp.743-748, Elsevier Science Publishers B.V. (North-Holland), Amsterdam.

[103] Katayama, T. 1992. Performance analysis for a two-class priority queueing model with general decrementing service. *IEICE Transactions on Communications,* Vol.E75-B, No.12 (December), pp.1301-1307.

[104] Katayama, T. 1994. Analysis of a cyclic-service tandem queue with a gate and vacations. *Symposium on Performance Models for Information Communication Networks,* pp.321-332, January 19-21, 1994, Hakone, Japan.

[105] Khamisy, A., Altman, E., and Sidi, M. 1992. Polling systems with synchronization constraints. *Annals of Operations Research,* Vol.35, No.1-4 (April), pp.231-267.

[106] Kofman, D. 1991. Blocking probability, throughput and waiting time in finite capacity polling systems. To appear in *Queueing Systems.*

[107] Konheim, A. G., and Levy, H. 1992. Efficient analysis of polling systems. *IEEE INFOCOM '92,* pp.2325-2331, Florence, Italy, May 6-8, 1992.

[108] Konheim, A. G., Levy, H., and Srinivasan, M. M. 1992. Descendant set: An efficient approach for the analysis of polling systems. To appear in *IEEE Transactions on Communications.*

[109] Kroese, D. P., and Schmidt, V. 1992. A continuous polling system with general service times. *The Annals of Applied Probability,* Vol.2, No.4, pp.906-927.

[110] Kroese, D. P., and Schmidt, V. 1993. Single-server queues with spatially distributed arrivals. Preprint.

[111] Kubat, P., and Servi, L. D. 1991. Cyclic service queues with very short service times. *European Journal of Operational Research,* Vol.53, No.2 (July), pp.172-188.

[112] LaMaire, R. O. 1991. An M/G/1 vacation model of an FDDI station. *IEEE Journal on Selected Areas in Communications,* Vol.9, No.2 (February), pp.257-264.

[113] Lang, M., and Bosch, M. 1991. Performance analysis of finite capacity polling systems with limited-M service. *Teletraffic and Datatraffic in a Period of Change, ITC-13,* A. Jensen and V. B. Iversen (editors), pp.731-735, Elsevier Science Publishers B.V. (North-Holland), Amsterdam.

[114] Lee, D.-S. 1994. Analysis of a cyclic server queue with the exhaustive and limited service discipline. Preprint.

[115] Lee, D.-S. , and Sengupta, B. 1992a. A reservation based cyclic server queue with limited service. *Performance Evaluation Review (1992 ACM SIGMETRICS and Performance '92),* Vol.20, No.1 (June), pp.70-77.

[116] Lee, D.-S. , and Sengupta, B. 1992b. An approximate analysis of a cyclic server queue with limited service and reservations. *Queueing Systems,* Vol.11, No.1-2 (July), pp.153-178.

[117] Lee, T. Y. S. 1993. Naive control of a queueing network. Preprint.

[118] Leung, K. K. 1990a. Waiting time distribution for token-passing systems with limited-one service via discrete Fourier transforms. *IEEE INFOCOM '90,* pp.1111-1118, San Francisco, June 3-7, 1990.

[119] Leung, K. K. 1990b. Waiting time distributions for token-passing systems with limited-k service via discrete Fourier transforms. *Performance '90,* P. J. B. King, I. Mitrani, and R. J. Pooley (editors), pp.333-347, Elsevier Science Publishers B.V. (North-Holland), Amsterdam.

[120] Leung, K. K. 1991. Cyclic-service systems with probabilistically-limited service. *IEEE Journal on Selected Areas in Communications,* Vol.9, No.2 (February), pp.185-193.

[121] Levy, H. 1991. A note on the complexity of Swartz's method for calculating the expected delay in non-symmetric cyclic polling systems. *Operations Research Letters,* Vol.10, No.6 (August), pp.363-368.

[122] Levy, H., and Kleinrock, L. 1991. Polling systems with zero switch-over periods: A general method for analyzing the expected delay. *Performance Evaluation,* Vol.13, No.2 (October), pp.97-107.

[123] Levy, H., and Sidi, M. 1990. Polling systems: Applications, modeling, and optimization. *IEEE Transactions on Communications,* Vol.38, No.10 (October), pp.1750-1760.

[124] Levy, H., and Sidi, M. 1991. Polling systems with simultaneous arrivals. *IEEE Transactions on Communications,* Vol.39, No.6 (June), pp.823-827.

[125] Levy, H., and Sidi, M., and Boxma, O. J. 1990. Dominance relations in polling systems. *Queueing Systems,* Vol.6, No.2 (April), pp.155-172.

[126] Liu, Z., and Nain, P. 1992. Optimal scheduling in some multi-queue single server systems. *IEEE Transactions on Automatic Control,* Vol.37, pp.247-252;

[127] Liu, Z., Nain, P., and Towsley, D. 1992. On optimal polling policies. *Queueing Systems,* Vol.11, No.1-2 (July), pp.59-83.

[128] Massoulié, L. 1993. Stability of a non-Markovian polling systems. Preprint, Laboratoire des Signaux et Systèmes, CNRS.

[129] Matsumoto, Y. 1994. On optimization of polling policy represented by neural network. Preprint.

[130] Miyoshi, N., Takahashi, Y., and Hasegawa, T. 1993. Discount optimal server assignment policy for symmetric polling systems. *Performance Models for Information Communication Networks,* pp.152-162, January 21-23, 1993, Numazu, Japan.

[131] Mukherjee, B., Kwok, C. K., Lantz, A. C., and Melody Moh, W.-H. L. 1990. Comments on "Exact analysis of asymmetric polling systems with single buffers." *IEEE Transactions on Communications,* Vol.38, No.7 (July), pp.944-946.

[132] Murata, M., and Takagi, H. 1992. Performance of token ring networks with a finite capacity bridge. *Computer Networks and ISDN Systems,* Vol.24, No.1 (March), pp.45-64.

[133] Nakamura, K., Takine, T., Takahashi, Y., and Hasegawa, T. 1990. Analysis of an asymmetric polling model with cycle-time constraint. *Architecture and Performance Issues of High-Capacity Local and Metropolitan Area Networks,* G. Pujolle (editor), pp.493-508, NATO ASI Series F., Springer-Verlag, Berlin, 1990.

[134] Ozawa, T. 1990. Alternating service queues with mixed exhaustive and K-limited services. *Performance Evaluation,* Vol.11, No.3 (September), pp.165-171.

[135] Resing, J. A. C. 1993. Polling systems and multitype branching processes. *Queueing Systems,* Vol.13, No.4 (July), pp.409-426.

[136] Rubin, I., and Baker, J. E. 1990. Media access control for high-speed local area and metropolitan area communication networks. *Proceedings of the IEEE,* Vol.78, No.1 (January), pp.168-203; Reprinted in *Advances in Local and Metropolitan Area Networks,* W. Stallings (editor), pp.96-131, IEEE Computer Society Press, 1994.

[137] Rubin, I., and Wu, C.-H. 1992. Analysis of an M/G/1 queue with vacations and its application to FDDI asynchronous timed-token service system. *IEEE Global Telecommunications Conference, (GLOBECOM '86),* pp.1630-1634, Orlando, Florida, December 6-9, 1992.

[138] Sarkar, D., and Zangwill, W. I. 1989. Expected waiting time for nonsymmetric cyclic queueing systems - Exact results and applications. *Management Science,* Vol.35, No.12 (December), pp.1463-1474.

[139] Sasaki, G. H. 1993. Input buffer requirements for round robin polling systems. *Performance Evaluation,* Vol.18, No.3 (November), pp.237-261.

[140] Shimogawa, S., and Takahashi, Y. 1992. A note on the pseudo-conservation law for a multi-queue with local priority. *Queueing Systems,* Vol.11, No.1-2 (July), pp.145-151.

[141] Sidi, M., and Levy, H. 1990. Customers routing in polling systems. *Performance '90,* P. J. B. King, I. Mitrani, and R. J. Pooley (editors), pp.319-331, Elsevier Science Publishers B.V. (North-Holland), Amsterdam.

[142] Sidi, M., Levy, H., and Fuhrmann, S. W. 1992. A queueing network with a single cyclically roving server. *Queueing Systems,* Vol.11, No.1-2 (July), pp.121-144.

[143] Srinivasan, M. M. 1991. Waiting time variances in continuous-time polling systems. Technical Report No.91-37, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, Michigan.

[144] Srinivasan, M. M., and Gupta, D. 1993. When should a roving server be patient? Preprint.

[145] Srinivasan, M. M., Levy, H., and Konheim, A. G. 1991. The individual station technique for the analysis of continuous-time polling systems. Technical Report No.91-36, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, Michigan.

[146] Srinivasan, M. M., Niu, S.-C., and Cooper, R. B. 1993. Relating polling models with nonzero and zero switchover times. Preprint.

[147] Szpankowski, W. 1990. Towards computable stability criteria for some multidimensional stochastic processes. *Stochastic Analysis of Computer and Communication Systems,* H. Takagi (editor), pp.131-172, Elsevier Science Publishers B. V. (North-Holland), Amsterdam.

[148] Takagi, H. 1986. *Analysis of Polling Systems.* The MIT Press, Cambridge, Massachusetts.

[149] Takagi, H. 1987. Queueing analysis of polling systems with zero switchover times. TRL Research Report, TR87-0034, IBM Japan, Ltd., Tokyo.

[150] Takagi, H. 1988. Queuing analysis of polling models. *ACM Computing Surveys,* Vol.20, No.1 (March), pp.5-28.

[151] Takagi, H. 1990a. Queueing analysis of polling models: An update. In *Stochastic Analysis of Computer and Communication Systems,* H. Takagi (editor), pp.267-318, Elsevier Science Publishers B. V. (North-Holland), Amsterdam.

[152] Takagi, H. 1990b. Effects of the target token rotation time on the performance of a timed-token protocol. *Performance '90,* P. J. B. King, I. Mitrani, and R. J. Pooley (editors), pp.363-370, Elsevier Science Publishers B.V. (North-Holland), Amsterdam. Also, Research Report RT 0022, IBM Tokyo Research Laboratory, November 1989.

[153] Takagi, H. 1991a. Analysis of finite capacity polling systems. *Advances in Applied Probability,* Vol.23, No.2 (June), pp.373-387.

[154] Takagi, H. 1991b. Application of polling models to computer networks. *Computer Networks and ISDN Systems,* Vol.22, No.3 (October), pp.193-211.

[155] Takagi, H. 1991c. *Queueing Analysis: A Foundation of Performance Evaluation, Volume 1: Vacation and Priority Systems, Part 1.* Elsevier Science Publishers B. V. (North-Holland), Amsterdam.

[156] Takagi, H. 1992. Analysis of an M/G/1//N queue with multiple server vacations, and its applications to a polling model. *Journal of the Operations Research Society of Japan,* Vol.35, No.3 (September), pp.300-315.

[157] Takagi, H. 1993. *Queueing Analysis: A Foundation of Performance Evaluation, Volume 2: Finite Systems.* Elsevier Science Publishers B. V. (North-Holland), Amsterdam.

[158] Takahashi, Y., and Krishna Kumar, B. 1994. On the pseudo-conservation law for discrete-time multi-queue systems with priority disciplines. *Symposium on Performance Models for Information Communication Networks,* pp.178-188, January 19-21, 1994, Hakone, Japan.

[159] Takine, T., and Hasegawa, T. 1992. A cyclic-service finite source model with round-robin scheduling. *Queueing Systems,* Vol.11, No.1-2 (July), pp.91-108.

[160] Takine, T., Takagi, H., Takahashi, Y., and Hasegawa, T. 1990. Analysis of asymmetric single-buffer polling and priority systems without switchover times. *Performance Evaluation,* Vol.11, No.4 (November), pp.253-264.

[161] Takine, T., Takahashi, Y., and Hasegawa, T. 1988. Exact analysis of asymmetric polling system with single buffers. *IEEE Transactions on Communications,* Vol.36, No.10 (October), pp.1119-1127.

[162] Takine, T., Takahashi, Y., and Hasegawa, T. 1989. Average message delay of an asymmetric single-buffer polling system with round-robin scheduling of services. *Modelling Techniques and Tools for Computer Performance Evaluation,* D. Potier and R. Puigjaner (editors), pp.179-187, Plenum Publishing Corporation, New York.

[163] Takine, T., Takahashi, Y., and Hasegawa, T. 1990a. Modeling and analysis of a single-buffer polling system interconnected with external networks. *INFOR,* Vol.28, pp.160-177.

[164] Takine, T., Takahashi, Y., and Hasegawa, T. 1990b. Further results on asymmetric polling systems with single buffers. Preprint. Department of Applied Mathematics and Physics, Kyoto University, Kyoto, Japan.

[165] Tangemann, M., and Sauer, K. 1990. Performance analysis of the FDDI media access control protocol. *Fourth International Conference on Data Communication Systems and their Performance,* pp.32-43, Barcelona, June 20-22, 1990.

[166] Tangemann, M., and Sauer, K. 1991. Performance analysis of the timed token protocol of FDDI and FDDI-II. *IEEE Journal on Selected Areas in Communications,* Vol.9, No.2 (February), pp.271-278.

[167] Tayur, S., and Sarkar, D. 1988. Implementation of an efficient algorithm to calculate waiting times in a nonsymmetric cycle queueing system. Internal Memorandum 45313-880822-01IM, AT&T Bell Laboratories.

[168] Towsley, D., Fdida, and Santoso, H. 1991. Design ana analysis of flow control protocols for metropolitan area networks. *High-Capacity Local and Metropolitan Area Networks,* G. Pujolle (editor), pp.471-492, Springer.

[169] Tran-Gia, P. 1992. Analysis of polling systems with general input process and finite capacity. *IEEE Transactions on Communications,* Vol.40, No.2 (February), pp.337-344.

[170] Tsai, Z., and Rubin, I. 1992. Mean delay analysis for a message priority-based polling scheme. *Queueing Systems,* Vol.11, No.3 (September), pp.223-240.

[171] van der Mei, R. D., and Borst, S. C. 1994. Analysis of multiple-server polling systems by means of the power-series algorithm. CWI Report BS-R9410, The Netherlands.

[172] Walrand, J. 1990. Queueing networks. *Handbooks in Operations Research and Management Science, Volume 2: Stochastic Models,* D. P. Heyman and M. J. Sobel (editors), pp.519-603, Elsevier Science Publishers B.V. (North-Holland), Amsterdam.

[173] Weststrate, J. A. 1990. Waiting times in a two-queue model with exhaustive and Bernoulli service. Technical report FEW 437, Tilburg University, The Netherlands.

[174] Woodward, M. E. 1994. *Communication and Computer Networks. Modelling with Discrete-Time Queues.* IEEE Computer Society Press, Los Alamitos, California.

[175] Yechiali, U. 1991. Optimal dynamic control of polling systems. *Queueing, Performance and Control in ATM (ITC-13),* J. W. Cohen and C. D. Pack (editors), pp.205-217, Elsevier Science Publishers B.V. (North-Holland), Amsterdam.

[176] Zhang, Z., and Acampora, A. S. 1993. Performance of a modified polling strategy for broadband wireless LANs in a harsh fading environment. *Telecommunication Systems,* Vol.1, No.3 (February), pp.279-294.

[177] Zhigang, A. 1993. Mean waiting delay for a token ring network with batch arrivals and limited service policies. Preprint.